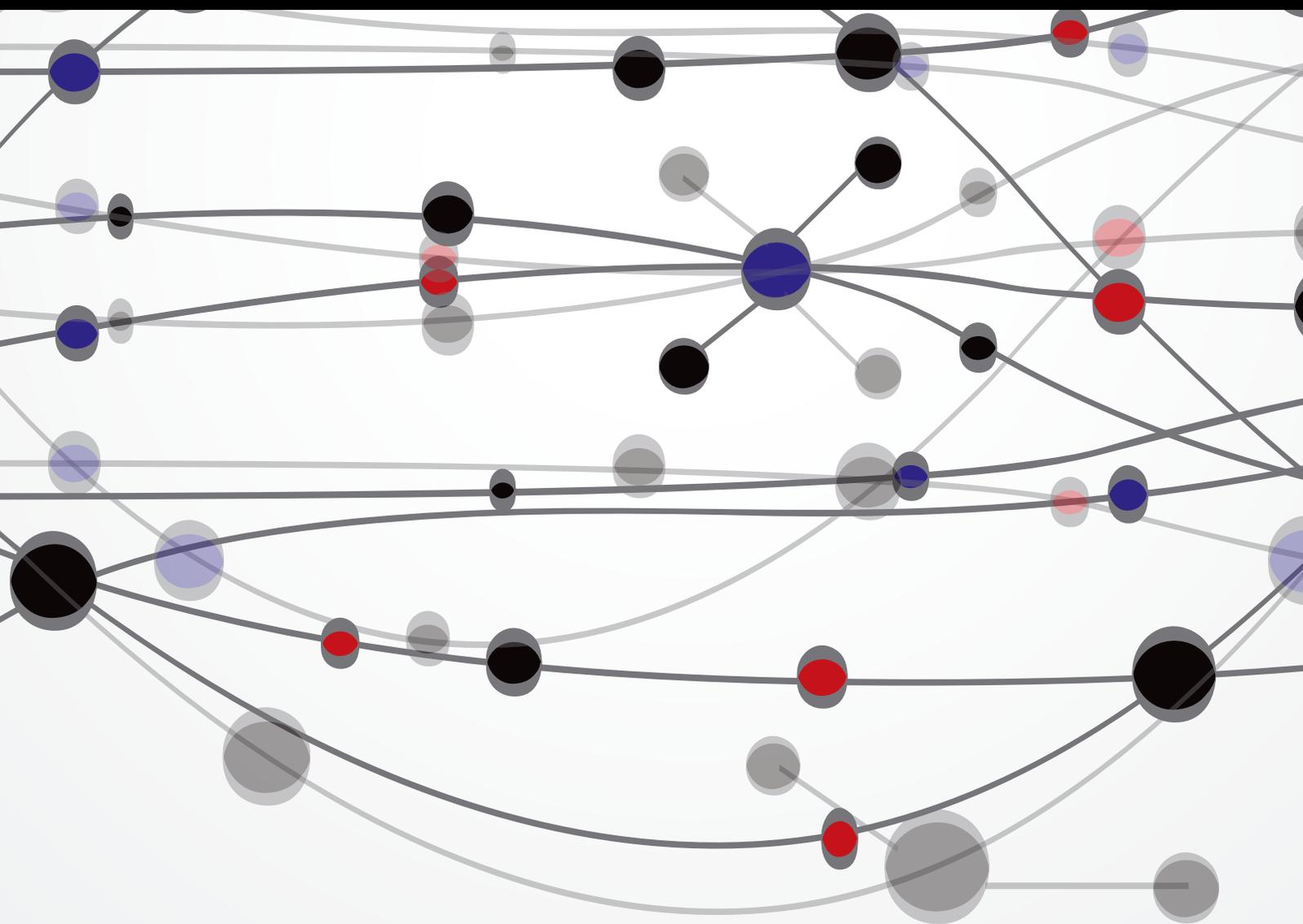


# Recent Advances in Communications and Networking

Guest Editors: Zhongmei Zhou, Yuxin Mao, Jaime Lloret,  
Xiaoxuan Meng, and Jingjing Zhou





---

# **Recent Advances in Communications and Networking**

The Scientific World Journal

---

## **Recent Advances in Communications and Networking**

Guest Editors: Zhongmei Zhou, Yuxin Mao, Jaime Lloret, Xiaoxuan Meng, and Jingjing Zhou



---

Copyright © 2014 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "The Scientific World Journal." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Contents

**Recent Advances in Communications and Networking**, Zhongmei Zhou, Yuxin Mao, Jaime Lloret, Xiaoxuan Meng, and Jingjing Zhou  
Volume 2014, Article ID 376260, 1 page

**Towards Internet QoS Provisioning Based on Generic Distributed QoS Adaptive Routing Engine**, Amira Y. Haikal, M. Badawy, and Hesham A. Ali  
Volume 2014, Article ID 694847, 29 pages

**Precoding Method Interference Management for Quasi-EVD Channel**, Wei Duan, Wei Song, Sang Seob Song, and Moon Ho Lee  
Volume 2014, Article ID 678578, 10 pages

**Spontaneous Ad Hoc Mobile Cloud Computing Network**, Raquel Lacuesta, Jaime Lloret, Sandra Sendra, and Lourdes Peñalver  
Volume 2014, Article ID 232419, 19 pages

**Indoor Positioning in Wireless Local Area Networks with Online Path-Loss Parameter Estimation**, Luigi Bruno, Paolo Addresso, and Rocco Restaino  
Volume 2014, Article ID 986714, 12 pages

**Describing the Access Network by means of Router Buffer Modelling: A New Methodology**, Luis Sequeira, Julián Fernández-Navajas, Jose Saldana, José Ramón Gállego, and María Canales  
Volume 2014, Article ID 238682, 9 pages

**Protection of HEVC Video Delivery in Vehicular Networks with RaptorQ Codes**, Pablo Piñol, Miguel Martínez-Rach, Otoniel López, and Manuel Pérez Malumbres  
Volume 2014, Article ID 619379, 9 pages

**A Secure 3-Way Routing Protocols for Intermittently Connected Mobile Ad Hoc Networks**, Ramesh Sekaran and Ganesh Kumar Parasuraman  
Volume 2014, Article ID 865071, 13 pages

**Design and Performance Evaluation of a Distributed OFDMA-Based MAC Protocol for MANETs**, Jaesung Park, Jiyoung Chung, Hyungyu Lee, and Jung-Ryun Lee  
Volume 2014, Article ID 708798, 15 pages

**A New Graph Drawing Scheme for Social Network**, Eric Ke Wang and Futai Zou  
Volume 2014, Article ID 930314, 9 pages

**Video Texture Synthesis Based on Flow-Like Stylization Painting**, Qian Wenhua, Xu Dan, Yue Kun, and Guan Zheng  
Volume 2014, Article ID 689496, 9 pages

**Reliable Multihop Broadcast Protocol with a Low-Overhead Link Quality Assessment for ITS Based on VANETs in Highway Scenarios**, Alejandro Galaviz-Mosqueda, Salvador Villarreal-Reyes, Hiram Galeana-Zapién, Javier Rubio-Loyola, and David H. Covarrubias-Rosales  
Volume 2014, Article ID 359636, 12 pages

**IDMA-Based MAC Protocol for Satellite Networks with Consideration on Channel Quality,**

Gongliang Liu, Xinrui Fang, and Wenjing Kang

Volume 2014, Article ID 181734, 16 pages

**A Mobile Anchor Assisted Localization Algorithm Based on Regular Hexagon in Wireless Sensor**

**Networks,** Guangjie Han, Chenyu Zhang, Jaime Lloret, Lei Shu, and Joel J. P. C. Rodrigues

Volume 2014, Article ID 219371, 13 pages

**Vehicle Density Based Forwarding Protocol for Safety Message Broadcast in VANET,** Jiawei Huang,

Yi Huang, and Jianxin Wang

Volume 2014, Article ID 584164, 9 pages

**Exploring a QoS Driven Scheduling Approach for Peer-to-Peer Live Streaming Systems with Network**

**Coding,** Laizhong Cui, Nan Lu, and Fu Chen

Volume 2014, Article ID 513861, 10 pages

**An Emergency Packet Forwarding Scheme for V2V Communication Networks,** Faika Hoque and

Sungoh Kwon

Volume 2014, Article ID 480435, 7 pages

**Towards Accurate Node-Based Detection of P2P Botnets,** Chunyong Yin

Volume 2014, Article ID 425491, 10 pages

**The Collaborative Search by Tag-Based User Profile in Social Media,** Haoran Xie, Xiaodong Li,

Jiantao Wang, Qing Li, and Yi Cai

Volume 2014, Article ID 608326, 7 pages

**The Application of Baum-Welch Algorithm in Multistep Attack,** Yanxue Zhang, Dongmei Zhao,

and Jinxing Liu

Volume 2014, Article ID 374260, 7 pages

**A New Seamless Transfer Control Strategy of the Microgrid,** Zhaoyun Zhang, Wei Chen, and Zhe Zhang

Volume 2014, Article ID 391945, 9 pages

**Reputation-Based Secure Sensor Localization in Wireless Sensor Networks,** Jingsha He, Jing Xu,

Xingye Zhu, Yuqiang Zhang, Ting Zhang, and Wanqing Fu

Volume 2014, Article ID 308341, 10 pages

**Supporting Seamless Mobility for P2P Live Streaming,** Eunsam Kim, Sangjin Kim, and Choonhwa Lee

Volume 2014, Article ID 134391, 8 pages

**Goodness-of-Fit Based Secure Cooperative Spectrum Sensing for Cognitive Radio Network,**

Hiep Vu-Van and Insoo Koo

Volume 2014, Article ID 752507, 6 pages

**Mobility-Assisted on-Demand Routing Algorithm for MANETs in the Presence of Location Errors,**

Trung Kien Vu and Sungoh Kwon

Volume 2014, Article ID 790103, 11 pages

**Effects of ADC Nonlinearity on the Spurious Dynamic Range Performance of Compressed Sensing**, Rongzong Kang, Pengwu Tian, and Hongyi Yu  
Volume 2014, Article ID 143693, 6 pages

**Modeling and Analysis of Mobility Management in Mobile Communication Networks**, Woon Min Baek, Ji Hyun Yoon, and Chesoon Kim  
Volume 2014, Article ID 250981, 11 pages

**Historical Feature Pattern Extraction Based Network Attack Situation Sensing Algorithm**, Yong Zeng, Dacheng Liu, and Zhou Lei  
Volume 2014, Article ID 473504, 8 pages

**Compressive Sensing Based Bayesian Sparse Channel Estimation for OFDM Communication Systems: High Performance and Low Complexity**, Guan Gui, Li Xu, Lin Shan, and Fumiyuki Adachi  
Volume 2014, Article ID 927894, 10 pages

**Code-Time Diversity for Direct Sequence Spread Spectrum Systems**, A. Y. Hassan  
Volume 2014, Article ID 146186, 15 pages

**Trusted Measurement Model Based on Multitenant Behaviors**, Zhen-Hu Ning, Chang-Xiang Shen, Yong Zhao, and Peng Liang  
Volume 2014, Article ID 384967, 12 pages

**Transition Characteristic Analysis of Traffic Evolution Process for Urban Traffic Network**, Longfei Wang, Hong Chen, and Yang Li  
Volume 2014, Article ID 603274, 9 pages

**Dynamic Resource Allocation in Hybrid Access Femtocell Network**, Afaz Uddin Ahmed, Mohammad Tariqul Islam, Mahamod Ismail, and Mohammad Ghanbarisabagh  
Volume 2014, Article ID 539720, 7 pages

**Analyses of Crime Patterns in NIBRS Data Based on a Novel Graph Theory Clustering Method: Virginia as a Case Study**, Peixin Zhao, Marjorie Darrah, Jim Nolan, and Cun-Quan Zhang  
Volume 2014, Article ID 492461, 8 pages

**Channel Estimation in DCT-Based OFDM**, Yulin Wang, Gengxin Zhang, Zhidong Xie, and Jing Hu  
Volume 2014, Article ID 813429, 6 pages

**Cooperative Search and Rescue with Artificial Fishes Based on Fish-Swarm Algorithm for Underwater Wireless Sensor Networks**, Wei Zhao, Zhenmin Tang, Yuwang Yang, Lei Wang, and Shaohua Lan  
Volume 2014, Article ID 145306, 10 pages

## Editorial

# Recent Advances in Communications and Networking

**Zhongmei Zhou,<sup>1</sup> Yuxin Mao,<sup>2</sup> Jaime Lloret,<sup>3</sup> Xiaoxuan Meng,<sup>4</sup> and Jingjing Zhou<sup>5</sup>**

<sup>1</sup>*Department of Computer Science and Engineering, Minnan Normal University, Zhangzhou 363000, China*

<sup>2</sup>*School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

<sup>3</sup>*Department of Communications, Universidad Politecnica de Valencia, 46730 Valencia, Spain*

<sup>4</sup>*VMware, 3401 Hillview Avenue, Palo Alto, CA 94304, USA*

<sup>5</sup>*School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

Correspondence should be addressed to Zhongmei Zhou; [zzm@zju.edu.cn](mailto:zzm@zju.edu.cn)

Received 18 November 2014; Accepted 18 November 2014; Published 18 December 2014

Copyright © 2014 Zhongmei Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The emerging communication and networking technologies and the way in which these are being integrated into the human, industrial, and social framework have made it evident that there are a number of related technical and socioeconomic areas whose understanding is still less than satisfactory and in which long-term research is needed. Meanwhile, a number of emerging concepts like cloud computing, Internet of Things, web 3.0, and green radio have been proposed in communication and networking. The research on emerging communication and networking technologies is considered as a global research challenge.

Therefore, we just hold this special issue for researchers, engineers, and practitioners to share their work on recent advances in communications and networking. The special issue focuses on the hottest and cutting-edge topics in communications and networking.

After peer-review, more than 30 papers from different affiliations and countries were accepted and published in this special issue. A feature of this special issue is that we have addressed the importance of wireless networking and accepted a number of papers related to this topic. We think that wireless networking will become more and more popular in both research and industry. Mobile network or MANET has been a hot topic during these years. Therefore, we have accepted a number of papers related to this topic in this special issue. As a special kind of MANET, VANET has also become a popular research topic recently. In this special issue, several papers are just talking about VANET, vehicular network, V2V, and urban traffic network. As an underlying

technology for Internet of Things, wireless sensor network has been studied a lot. Many research efforts have been proposed in this field. We have also addressed some new research in this special issue.

The rest of the papers in the special issue have talked about a wide range of emerging topics like cloud computing, social networks, satellite networks, microgrid, network security, and so forth.

We hope that readers of this journal, especially those from the communications subject area, will find in this special issue not only the new ideas, cutting-edge information, new technologies, and applications of communication and networking, but also a special emphasis on how to solve various emerging problems.

*Zhongmei Zhou  
Yuxin Mao  
Jaime Lloret  
Xiaoxuan Meng  
Jingjing Zhou*

## Research Article

# Towards Internet QoS Provisioning Based on Generic Distributed QoS Adaptive Routing Engine

**Amira Y. Haikal, M. Badawy, and Hesham A. Ali**

*Department of Computer Engineering & Control Systems, Mansoura University, Mansoura 35516, Egypt*

Correspondence should be addressed to M. Badawy; badawy\_mm@hotmail.com

Received 11 April 2014; Accepted 9 July 2014; Published 17 September 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 Amira Y. Haikal et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Increasing efficiency and quality demands of modern Internet technologies drive today's network engineers to seek to provide quality of service (QoS). Internet QoS provisioning gives rise to several challenging issues. This paper introduces a generic distributed QoS adaptive routing engine (DQARE) architecture based on OSPF<sub>x</sub>QoS. The innovation of the proposed work in this paper is its undependability on the used QoS architectures and, moreover, splitting of the control strategy from data forwarding mechanisms, so we guarantee a set of absolute stable mechanisms on top of which Internet QoS can be built. DQARE architecture is furnished with three relevant traffic control schemes, namely, service differentiation, QoS routing, and traffic engineering. The main objective of this paper is to (i) provide a general configuration guideline for service differentiation, (ii) formalize the theoretical properties of different QoS routing algorithms and then introduce a QoS routing algorithm (QOPRA) based on dynamic programming technique, and (iii) propose QoS multipath forwarding (QMPPF) model for paths diversity exploitation. NS2-based simulations proved the DQARE superiority in terms of delay, packet delivery ratio, throughput, and control overhead. Moreover, extensive simulations are used to compare the proposed QOPRA algorithm and QMPPF model with their counterparts in the literature.

## 1. Introduction

Increasing steadily gaining popularity of mobile phones, VoIP, IPTV, cloud computing, as well as sensor networks that interoperate with Internet creates a large demand for QoS support for future Internet applications [1]. The main motivation behind the design of the next generation Internet is convergence, that is to say, making the Internet the common carrier for all kinds of services. The Internet is destined to become the ubiquitous global communication infrastructure [2].

In the beginning, the Internet used the public switched telephone network (PSTN) telecommunications (TelCo) infrastructure. The major interest was on “*where issue*” which means where packets should deliver. Now the TelCo industry has started to use the Internet infrastructure as a backbone, and with the advent of multimedia applications people became aware of the “*how will issue*” (or, “quality of service

(QoS)”). Generally, QoS can be defined as the ability to create various traffic management mechanisms in the network to differentiate between different classes of services and to provide some level of assurance and performance optimization that can affect user perception [3]. QoS has become one of the most important issues in the next generation network (NGN) [4].

The Internet exists in order to transfer information from source nodes to destination nodes. Simultaneously, diversity of Internet services has become very competitive and end users are demanding very high quality services from their service providers. Accordingly, to accommodate service quality, Internet service providers (ISPs) have to provide interconnections more efficiently. Thus, one of the key issues in such a converged network is routing.

Routing is the process performed by routers to select the best path from the source node to a destination node in a network. The Internet traffic volume continues to grow at

a massive rate; there may be a time when networks start to be congested on a regular basis. This situation has been the major force for innovation and development of different QoS routing solutions. Future Internet will embrace QoS routing as a basic functionality for QoS provisioning.

QoS-effective routing scheme can be efficiently designed to allocate resources in the network, allowing user constraints to be met and maximizing operator benefits, taking into consideration properties of the underlying network. In general, routing involves two entities, namely, the routing protocol and the routing algorithm. Although there has been historically close tie between both entities, it is beneficial to decouple them. The routing protocol has the task of dynamically identifying and communicating topological information [5]. Although proposals for a QoS routing protocol for Internet exist, still there is no Internet QoS routing protocol in the Internet. A critical basis for routing is routing computation algorithm that calculates the shortest path (SP) at each router for every known destination based on current topological information.

Open Shortest Path First (OSPF) protocol [6] is perhaps the most famous link-state routing extensively deployed throughout the last decade. OSPF provides best-effort Internet routing relying on a single arbitrary metric for path computation. OSPF does not guarantee optimal network utilization of available network resources due to a single path/single metric routing, which may cause partial congestion of the network.

The notion of QoS is a guarantee by the network to satisfy a set of predetermined service performance constraints for the user in terms of the end-to-end delay statistics, available bandwidth, probability of packet loss, jitter, and so on. QoS-based routing must extend the current routing paradigm in four dimensions. First, routers need information about available network resources. Second, we calculate optimal paths that fulfill a set of constraints. Third, opportunistic routing must shift traffic from one path to another as soon as a "better" path is found. Fourth, optimal path forwarding algorithms must support multipath routing [7].

QoS extensions to OSPF (OSPFxQoS or QOSPF) [8] provide comprehensive mechanisms to support QoS. But it poses the following limitations.

- (i) OSPFxQoS used per-flow reservation via resource reservation protocol (RVSP) [9]. Resource reservation is not an appropriate method as overheads for setting up a reservation are simply too high.
  - (ii) OSPFxQoS considered limited QoS constraints in the routing process. OSPFxQoS considered only bandwidth as a metric. It did not fully capture the complete range of potential QoS requirements.
  - (iii) OSPFxQoS used precomputation routing algorithm that amortized the computational cost over multiple requests, but each computation instance is usually more expensive than in the on-demand case, as paths are computed to all destinations and for all possible bandwidth requests. Moreover, the accuracy of the selected paths may be lower.
  - (iv) OSPFxQoS gets feasible path (not the optimal one) with minimum number of hops and supports requested bandwidth.
  - (v) The major shortcoming of OSPF and OSPFxQoS is lack of self-optimization. The self-adaptation mechanism is static. Routers disseminate information only when topology changes.
  - (vi) OSPFxQoS is unable to readjust forwarding paths in order to lessen the impact of congestions or load-balance traffic to optimize the performance of the network.
- Although the literature is plentiful of numerous QoS architecture, routing plays essential role in QoS provisioning. The main intention of this paper is to describe QoS-based routing issues and identify the basic requirements of QoS intradomain routing. This paper introduces a general formulation that combines framework and approaches for QoS provisioning based on the knowledge necessary of the service with minimal impact to routing infrastructure established upon OSPFxQoS routing engine architecture. This paper presents the following theoretical and practical contributions.
- Theoretical contributions include the following.
- (i) First, exploring through discussion that regardless of QoS architecture, performance optimization inside autonomous systems (AS) is an important building block in the QoS provisioning.
  - (ii) Second, exploring that performance optimization can be achieved via proposing a generic distributed QoS adaptive routing engine (DQARE) architecture to overcome OSPFxQoS limitations. DQARE is a distributed software routing engine anticipated to exploit the full advantage of distributed hardware and improve scalability, overall performance, and resiliency. DQARE architecture is supplied with three unique features, namely, service differentiation, QoS routing, and traffic engineering (TE).
  - (iii) Third, we address factors that affect perceived QoS, study Internet applications and recognize their QoS requirements, and organize applications into classes. Then, we introduce a general guideline for marking packets with a distinct code in order to differentiate between different types of services. Consequently, no resource reservation is required.
- Practical contributions include the following.
- (i) First, developing an efficient hybrid QoS path computation scheme which compared with state-of-art QoS routing schemes is unique in providing minimal complexity and low error decision rate. Our proposed routing scheme assumes that the true state of the network is available to every node. The computation scheme involves combining precomputation and on-demand multiconstrained routing algorithms to retrieve multiple paths for a QoS request.
  - (ii) Second, we introduce a QoS multipath forwarding (QMPF) model. Multipath forwarding is a well-known approach to intradomain TE used to exploit

path diversity provided by the proposed routing algorithms and to provide an autonomic congestion management mechanism. QoS provisioning while load balancing is still a challenge. From our knowledge, no researches had been devoted to such area. Several problems such as supporting per class delay guarantees, packet reordering delay, and protection among different service classes are yet to be addressed.

The rest of this paper is organized as follows. Section 2 provides a general overview and discussion for QoS architectures. Section 3 explores how to achieve Internet QoS provisioning regardless of QoS architectures. Section 4 surveys related work. Section 5 presents the proposed DQARE architecture. In Section 6, a general configuration guideline for service differentiation in conjunction with Internet application QoS requirements is presented. Section 7 introduces the proposed routing framework and algorithms. The proposed QMPF model with algorithms is introduced in Section 8. In Section 9, we illustrate the applicability of the proposed solutions and present the experimental results.

## 2. QoS Architectures

There are three types of QoS, namely, perceived, assessed, and intrinsic QoS [10]. Perceived QoS (P-QoS) is a user-oriented QoS defined as the quality perceived by the users which depends on what the end points can do for the applications. It is mainly concerned with the software application industry not the network industry. Assessed QoS refers to the will of a user to keep on using a specific service. It is related to P-QoS and depends on marketing and commercial aspects.

Intrinsic QoS (I-QoS) is a network-oriented QoS concerned with what the networks can do for the applications. I-QoS may be described in terms of objective parameters such as delay, jitter, and loss. A key issue is how to provide efficient and fair routing so as to provide overall accepting service for most of the users. Most of QoS provision is offered in terms of I-QoS. It is mainly a technical problem dealt with by engineers, designers, and operators.

Supporting QoS in packet switching networks requires specialized infrastructure to be designed and developed that involves frameworks and new service models in addition to the existing best-effort service and resource management techniques that are based on the concept of dividing flows to traffic classes that are served with various QoS levels. This also requires traffic management schemes to be introduced, such as signaling, resource reservation, marking, packet classification, admission control, traffic conditioning, queuing, scheduling, and buffer management.

Two complementary basic approaches have been devised to guarantee QoS in data networks. The first is reservation based (state-oriented) approaches in which network resources are reserved in advance according to application's QoS request needs and subject to bandwidth management policy. The second is prioritization-based (stateless) approaches in which there is no resource reservation. Instead,

traffic is classified and network elements give preferential treatment to classifications identified as having more demanding requirements.

Although the literature is plentiful of different QoS architecture, none of them become dominant or widely implemented on the Internet. There are two well-known proposed QoS architectures, namely, integrated service (IntSer) and differentiated service (DiffSer). For a more comprehensive discussion about different architecture, we refer the reader to [11].

In early 1990s, the IETF standardized the service types that build the first complete model of QoS assurance IntSer framework [12]. IntSer framework was developed based on certain key theoretical assumptions established upon reservation-based approaches. For flow awareness and achieving end-to-end real guarantees, IETF specialized a signaling protocol called RSVP [9]. RSVP was used by hosts to request specific QoS from the network without any limitation on the number of classes of service. IntSer established the foundation for QoS architecture and flow awareness; however, it suffered from complexity and scalability problems. IntSer cannot be ignored, regardless of these drawbacks; its features allow assuring QoS for each flow.

Afterwards, some researchers followed the direction of IntSer to solve its drawbacks and others favored to forego a different direction. Connectionless approach [13] decided to follow IntSer and mitigate the scalability problem via automatic detection of QoS requirement on the fly rather than using signaling protocol. It used traffic conditioners that consist of a classifier, admission control, and scheduler. However, automatic detection not always work probably which affects the robustness of the architecture.

The DiffSer architecture [14, 15] was an alternative QoS model developed by IETF to cope with the scalability problem faced by IntSer. DiffSer deployed prioritization-based approaches, abandoned flow level, and worked at class level, where a class is an aggregate of many flows. DiffSer aims to guarantee QoS per traffic class. DiffSer architecture used resource allocation quite different from IntSer. DiffSer approach used six bits in the Type of Service (TOS) field in the IP packet header, which has been transformed into DiffSer field (or code point) to encode the forwarding treatment (technically called Forwarding Equivalence Class). DiffSer was contemplated using limited classes of services (CoS) that makes sense to add new complexity in small increments to the existing best-effort service. A router can offer to packets two aspects of preferential treatment which are expedited forwarding (EF) and assured forwarding (AF) in addition to unclassified service (i.e., best effort). DiffSer does not require the soft-state concept and thus avoids session-level scalability issue faced with RSVP.

Although DiffSer solved the scalability problem of IntSer, it failed to provide end-to-end QoS provisioning [16], and there is no performance guarantee. In reality, the DiffSer has been deployed by some ISPs. Li and Mao proposed a novel flow-based scheme [17] established upon DiffSer. Such scheme ensured a constant proportion between the P-QoS by flows in different classes, regardless of the current class loads. The scheme is furnished with an estimator for the

number of active flows, a dynamic weighted fair queuing (WFQ) scheduler, and a queue management mechanism. Nevertheless, this architecture still has a limited number of CoSs and complex operation.

Even if QoS is the subject of a great number of tutorials, papers, patents, recommendations, and standards, there is a question which still needs an answer: “does procuring an oversupply bandwidth (overprovisioning) will solve all QoS challenges?” Xiao [3] and Montanez [18] stated that (i) there is no need for QoS mechanisms and all you need to achieve QoS is sufficient capacity; (ii) overprovisioning works satisfactorily and the service differentiation will not be able to create perceivable differentiation either in normal or abnormal conditions; (iii) best effort can provide good enough performance for most applications in the developed countries; (iv) it is commercially difficult to install QoS in a network.

However, there are raised objections: (i) bandwidth guarantee is indeed a key component for offering QoS. But purchasing an oversupply of bandwidth will not solve all service-quality challenges; bandwidth optimization and possible future trends and requirements of new services must be considered; (ii) if utilization is observed to be higher than the acceptable threshold for a particular link type, so it is easy to trivially add bandwidth/capacity to the network. However, in large networks, where the utilization can be further impacted by routing, adding bandwidth is much more complex than this simple single-link network, and (iii) lack of differentiation among services leads to difficulty to sell QoS which is the fundamental cause of the commercial challenges.

Actually, an important lesson is learned from the above discussion. DiffSer traffic management is not the primary way to enable QoS, introducing DiffSer traffic management mechanisms alone cannot provide a satisfactory QoS solution, and too much complexity may have crept into the network. DiffSer traffic management is good for creating some service differentiation when there is congestion, but how often real-world networks are like that is unverified. In a more insistently QoS environments, service differentiation, QoS routing, and optimization using traffic engineering (TE) techniques have become the only viable alternative.

### 3. Proposed QoS Provisioning Methodology

The Internet today is going through new generations of innovative and fast applications that need high performance demands on the Internet infrastructure. Wójcik and Jajszczyk [11] stated that progress in access network capacities is far greater than in core networks and the bandwidth is always consumed. There may be a time when networks start to be congested on a regular basis and efficient and feasible QoS provisioning methodology might then be needed. To keep pace with the continuous demands, not only does the bandwidth need to be increased, but also the routers that power the Internet have to evolve architecturally and be furnished by congestion management schemes.

“How QoS is possible in Internet, and how can it be achieved in an efficient and reliable manner?”. At the core of

the answer, in a realistic sense, and to provide acceptable QoS performance, performance optimization inside autonomous systems is an important building block in the deployment of QoS. The majority of QoS publications focused on QoS architecture and traffic management schemes, although it only affects the performance of specific link of routers. The work presented in this paper takes a different approach. More effective Internet QoS provisioning can be achieved, as shown in Figure 1, via introducing a generic QoS routing engine architecture furnished with three intertwined traffic control schemes blocks that contribute to QoS provisioning.

The three-dimensional traffic control schemes contain service differentiation, unicast intradomain QoS routing algorithm, and traffic engineering (TE). These schemes must work in tandem for providing efficient services. These traffic control schemes can be used to reduce or prevent congestion and introduce a bigger impact on QoS than traffic management schemes as they affect traffic performance network wide. The coordination among these traffic control schemes is significant and utilizing service differentiation will assist in better perceived QoS.

*Service Differentiation.* ISPs are deploying more resources to handle the emerging applications. In order to lessen the amount of deployed network infrastructure and resources, differentiation of ISP offering services is needed. Service differentiation via setting the IP header TOS field must be designated in order to give better service when it is available. It is important to define a wide range of services, each with its own requirements. Accordingly, service differentiation has to enable deployment of scalable service discrimination in the Internet without the need for per-flow state and signaling protocols.

*QoS Routing Algorithm.* The problem of QoS routing has been the center of attention in both academic and industrial communities for some time. There is a need for additional capabilities in the IP routing world and performance management tools to determine optimal paths that satisfy a set of constraints. For QoS provisioning, the way of path selection should also be QoS-aware which means identifying an optimal route that meets multiple constraints which is more complex than best-effort routing.

*Traffic Engineering (TE).* Enhancing the performance of QoS routing at both traffic and resource levels is the major objectives of QoS intradomain TE. TE is thoughtful as an aspect of Internet network engineers. TE is the process of optimizing the operational performance of networks flows through better control, at both network flows and resource levels. One of the major intradomain TE techniques is multipath forwarding (MPF) using traffic splitting that is routing traffic in a way that can effectively maximize utilization of network resources, supplying path protection, reducing blocking capabilities, minimizing delays, and increasing throughput. Thus, TE regarding intradomain routing can be defined as supplementary to the routing infrastructure and QoS provisioning.

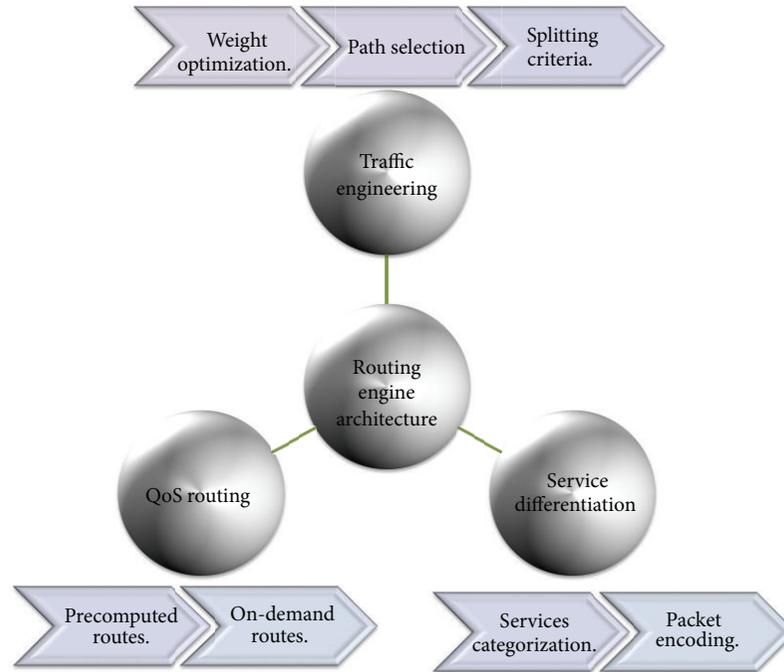


FIGURE 1: Internet QoS provisioning methodology.

#### 4. Related Works

QoS-based routing has been recognized as a missing piece in the evolution of QoS-based service offerings in the Internet. Special attention must be given to new powerful architectures for routers in order to fulfill the demanding critical role in QoS provisioning. This section covers a literature review and opens challenges for router architectures, service differentiation, QoS routing algorithms, and multipath forwarding mechanisms.

**4.1. Router Architecture.** The basic key functionalities in an IP router can be categorized into three functions (or planes): (i) route processing (how you direct bits), (ii) packet forwarding (how you move bits), and (iii) management services which includes management applications, protocol policies, and QoS. Router architectures have experienced three generations in terms of hardware and software [19, 20].

The first generation of IP router was built around conventional computer architecture [21]. Unfortunately, this simple architecture produced low performance as the three planes competing for the same processing unit.

In the second generation IP routers, improvement was introduced to increase the system throughput by distributing the packet forwarding operations by using multiple processors with on-demand lookup route caching. However, the frequent changes in network topology in the core of the Internet caused the cache entries to be invalidated frequently, resulting in smaller hits [22]. The third generation of routers introduced a hardware-forwarding engine and replaced shared bus by a high-speed crossbar switch with the aim to achieve higher throughput [23]. This architecture is still limited by the drawbacks of cache schemes.

Actually, structure of previously routers' architecture has an architectural limitation when it comes to meeting future requirements. The network performance degrades as the volume of traffic increases. These architectures relied on shared resource for all access and transfers and there is a centralized arbiter or scheduler responsible for granting access to the resource. The modularity of the hardware and the software is a key to the implementation of a modern router.

For QoS provisioning, next generation routers (NGRs) should be dependent upon fully distributed architectures in which partitioning the functions physically, logically or even both as clearly as possible to simplify system design and testing and achieve throughput and robustness as well as system availability. Distributed processing architecture is a combination of all the techniques discussed above.

A distributed routing engine comprises the modules running on different cards of a router. The main objective is to overcome the previous limitations of processing, memory bandwidth, and bus bandwidth via distributing overall processing and buffering capacity over the CPU and network interfaces equipped with processing power and buffer space. The functions of the forwarding engines are integrated into the interface cards in the distributed mode. Processing load gets distributed, ensuring faster and more reliable communication.

**Open Challenges.** NGRs are fitting QoS requirements that lie in the critical path of data flow. The software architecture for next generation routers should therefore be much more distributed in order to be scalable and to take full advantage of the distributed hardware platform entailed by the switch fabric.

**4.2. Service Differentiation.** Internet architecture is characterized by fairness which means that all kinds of applications are fairly shared network resources. Therefore, there is only one forwarding treatment deployed that cannot bear QoS oriented applications. QoS deployment needs a methodology to classify and differentiate between different applications. The Internet layer of TCP/IP stack on which routers operate must be able to distinguish between different classes of services.

Routers are able to distinguish between packets. The IP protocol provides a facility for upper layer protocols to convey hints to the Internet Layer about forwarding path behaviors. This facility is first addressed by the TOS field in IP header [24].

DiffSer standards [25] replaces the IP TOS field by 6-bits code points (also called differentiated services code point (DSCP)) and two bits currently unused to indicate the forwarding equivalence class (FEC). The DSCP identifies a specific traffic class and implies that all the packets identified with the same DSCP should receive the same treatment. Considerable debate took place on the allocation of these 6-bits code points. Following RFC2474 [15] and RFC4594 [25] general guidelines, the DSCP field can convey 64 distinct code points divided into three pools: 32 DSCPs are dedicated to standard recommended code points (Pool 1), 16 to be reserved for experimental and local use (Pool 2), and the other 16 (Pool 3) to be initially available for experimental and local use but may devote again to standard actions if Pool 1 is ever exhausted.

**Open Challenges.** DiffSer provided limited set of traffic classes and therefore a risk is the aggregation of nonhomogeneous traffic. Class based classification with a limited number of classes does not guarantee that flows classified with a higher priority will really observe a better quality of service than lower priority ones, due to the fact that the distribution of active flows in individual classes might be different. There is a crucial need for scalable service differentiation guidelines.

**4.3. QoS Routing Algorithms.** QoS-aware routing algorithms aim to find a path that obeys multiple constraints. Generally, traditional routing paradigm can be extended to support QoS by considering two important issues: first, choosing and distributing relevant QoS measures and, second, how to compute routes based on the information collected. Metric selection is very important in the sense that “the metrics must represent the basic network properties of interest.”

In QoS arena, routing metrics can be broadly divided into two classes. The first class is static (cumulative) metrics which value does not change over time. These metrics can be classified into (i) *additive parameters* (e.g., delay, hop count, and jitter), where the cost of a path is the sum of the individual link values along that path and (ii) *multiplicative parameters* (e.g., packet loss) which can be approximately transformed into additive by taking the logarithm of the multiplicative measures on each link. The second class of routing metrics is dynamic metrics (also referred to as a bottleneck or concave or min/max metrics), where metric’s value changes over

TABLE 1: A comparison between OQRA and PQRA.

Point of view	OQRA	PQRA
Computation	Every time a request initiated.	Precomputes paths from a source to all destinations.
Scalability	Limited.	Large-scale networks.
Suitability	When requests arise infrequently.	When requests arise very frequent.
Pros	Optimal routing during congestion state.	(i) Saves time. (ii) Better scalability. (iii) Improved load balancing.
Cons	Puts excessive time on packet processing.	(i) Path oscillation. (ii) Nonoptimal routing during congestion. (iii) Resources consumption. (iv) Complexity.

time with each request (e.g., bandwidth). Many researches considered only two additive QoS constraints [26].

The constraints associated with dynamic parameters can be handled by *postprocessing* which finds multiple paths from source to destination that satisfy set of static constraints and then select one path from these paths such that all the other dynamic parameters are satisfied [27]. Otherwise, preprocessing can be used for pruning from the graph all the links that do not satisfy constraints and then search for a feasible path [28]. In practice, the constraints on additive QoS metrics are more challenging, and, therefore, without loss of generality, the QoS metrics are assumed to be additive [29].

QoS multiconstrained path selection with additive parameters is an NP-complete problem that cannot be exactly solved in polynomial time [30]. However, Kuipers and Mieghem showed that the “worst-case” behavior is very unlikely to occur in practice and thus exact QoS routing algorithms seem feasible [31].

QoS-aware routing algorithms in the literature can be classified according to the path computation triggering criteria into precomputation QoS routing algorithms (PQRA) and on-demand computation QoS routing algorithm (OQRA). A comparison between the two paradigms is conducted in Table 1.

Most QoS routing algorithms presented in the literature used OQRA. OQRA may cause an insufferable computational overload in the high-speed next generation networks. OQRA needs to be called each time a new demand needs to be routed.

OQRA can be classified into four classes. The first class is heuristic algorithms [26, 32–35], where QoS routing is NP-complete; it demands heuristics in global optimization that help decide which one of a set of possible solutions is to be examined next. The second class is  $\epsilon$ -approximation algorithms [36–38] which give good approximate solutions to the problem which may not necessarily be exact. The third class is the exact algorithms [39, 40], where exactness can be reached by computing all possible paths between source

and destination where the exact path is guaranteed to be found. The last class is metaheuristics which used metaheuristics, such as ant colonies [41] and genetic algorithms [42].

TAMCRA [33] is a heuristic algorithm that is based on three concepts: (i) a nonlinear measure of the path length, (ii) a  $k$ -shortest path approach, and (iii) the principle of nondominated paths. TAMCRA aims only at finding a feasible path (not optimal). The major drawback is that if an intermediate node found at sub-path better than the stored  $k$  paths, it replaces stored paths even it has much longer post-path to destination. In other words, stored prepaths may misguide the search towards the shortest path. TAMCRA has a worst-case complexity of  $O(kN \log(kN) + k^3mM)$  per request.

To overcome the TAMCRA drawbacks, Korkmaz and Krunz [34] presented a heuristic algorithm called HMCOP, which tried to find an optimal path within the constraints by using the nonlinear path length function for feasibility. HMCOP searched for the path that not only is feasible but also minimizes the value of a primary QoS attribute. HMCOP executed two modified versions of Dijkstra's algorithm in the backward and forward directions. In the backward direction, HMCOP computed an estimate of how suitable the remaining subpaths are. In the forward direction, HMCOP used a modified version of Dijkstra's algorithm. This version heuristically determined complete path by concatenating the postpath from source to intermediate nodes and the estimated prepaths. HMCOP provided a preference rule for choosing paths. The drawback of HMCOP is that postpath may misguide the selection of prepaths. HMCOP had the worst-case complexity of  $O(N \log N + mM)$ .

SAMCRA [39] is the exact successor of TAMCRA to obtain multiconstraint optimal path. SAMCRA was based on four fundamental concepts: a nonlinear measure of the path length,  $k$ -shortest path approach with attainable bound for  $k_{\max}$ , the principle of nondominated paths, and the concept of lookahead to calculate attainable lower bounds and reducing search efforts. SAMCRA is considered an effective algorithm; however, the major drawback of SAMCRA resides in its complexity which is  $O(kN \log(kN) + k^2mM)$ .

Retrieving multiple paths subject to multiple constraints was addressed in [27]. The algorithm called  $A^*$  Prune finds not only one but also multiple shortest paths satisfying the constraints listed in order of increasing length.  $A^*$  Prune algorithm may be considered similar to the SAMCRA algorithm [39] except that it relied on linear path length. The worst-case complexity of  $A^*$  Prune is  $O(N!(m+h+N \log N))$ , where  $h$  is the number of hops of the retrieved path.

To compute feasible paths, Bellabas et al. [35] proposed two fast heuristic algorithms with less combinatorial complexity. The first heuristic algorithm is the hop count approach (HCA) that computes paths with the smallest hop count. The second is called metric linearization approach (MLA) that used a combination of QoS metrics. To store multiple paths at each intermediate node, they proposed a modification to Yen's algorithm [43] which was generalized by Lawler in [44]. HCA and MLA stop at the first feasible path they find thus reducing their execution time. However, they cannot obtain optimal paths.

Shin et al. proposed MPLMR [26] which is a heuristic multiconstraint QoS routing scheme. MPLMR used the same concepts as TAMCRA and HMCOP to store a limited number of subpaths between the source node and each intermediate node. MPLMR used an improved "lookahead" method to estimate full path length. Sanguankotchakorn et al. [45] proposed an algorithm (RMCOP) to find feasible path that somewhat satisfies multiple constraints. They proposed a relaxed lookahead algorithm. However, the computational complexity of the proposed algorithm is large.

Contrarily, precomputation schemes use an offline procedure. The overall computational load of such scheme is reduced, especially when the rate of QoS request arrivals is much higher than that of significant change in the network state. However, temporal conditions like congestion in the network and routing between two subsequent updates makes the routing decision be calculated based on inaccurate information resulting in nonoptimal path selection.

QoS routing mechanisms and OSPF extensions [8] focused on the algorithms used to compute QoS routes and the necessary modification to OSPF to support QoS. OSPFvQoS deployed a precomputing routing algorithm that amortized the computational cost over multiple requests with the motivation to get a feasible path with minimum number of hops (low resources) and support requested bandwidth for all possible QoS requests. The presented algorithm has a computational complexity compared with Bellman-Ford algorithm [46] but with limited QoS constraints consideration in the routing process.

An approach for pre-computation of multi-constrained path (PMCP) was proposed in [47]. It computed a number of QoS coefficients based on which linear QoS function was computed and then constructed different shortest path trees to compose the routing table. PMCP proposed algorithm has a complexity of  $O(B(m + n \log n + n))$ . Jin [48] proposed precomputation algorithm called limited selective flooding (LFS) routing algorithm. LFS considered an MCP problem with imprecise additive link state information. Authors in [49] presented  $O(Km + n \log n)$  time  $K$ -approximation precomputation algorithm. Afterwards, in [50], the authors reduced the computational complexity.

*Open Challenges.* NP-completeness of the QoS routing problem leads to only few exact algorithms proposed in the literature. The difficulty of QoS algorithm lies in its computational complexity and success rate. Designing QoS routing algorithms with low complexity, high performance, and high success rate is still an open issue.

*4.4. Multipath Forwarding (MPF).* Two major issues had drawn attention in recent years regarding TE. The first issue is to provide a TE mechanism to effectively develop a routing optimization that enhances network service capability without causing network congestion. The second issue is achieving resiliency via introducing a TE solution that minimizes the impact of nodes and link failure [51].

The first and foremost question of providing congestion control and QoS degradation mechanisms is "how to

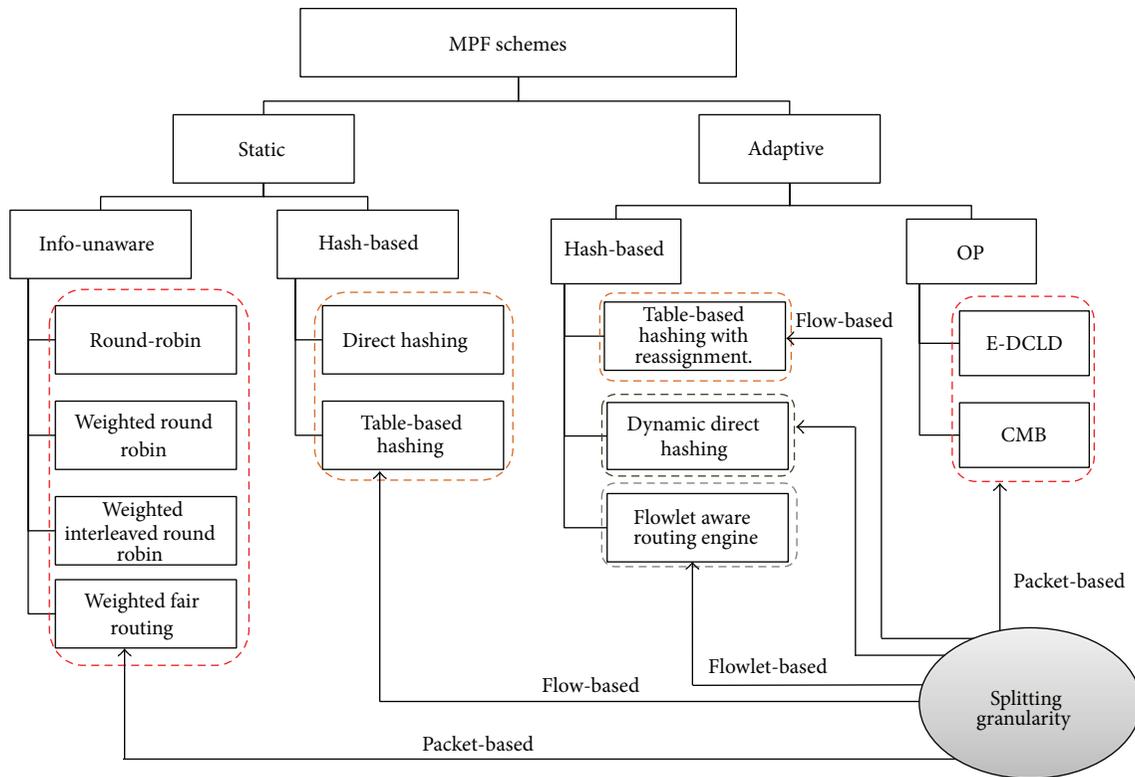


FIGURE 2: LDM classification.

explicitly control traffic distribution inside an AS??. At the core of the answer lies intradomain TE. Extensive deployment of intradomain routing protocols such as OSPF has drawn an ever increasing attention to Internet TE in recent years.

The major TE technique is multipath forwarding (MPF), also called multipath routing, using traffic splitting and path protection. Multipath intradomain forwarding can be used to handle traffic congestion inside a domain. Routers must be equipped with multipath forwarding mechanism to perform traffic forwarding. So, multipath load distribution is engineered by two key capabilities: (i) control-plane extension: deploying a routing algorithm to compute multiple paths and (ii) data-plane extension: providing a load distribution model to the forwarding process engineered by two key functionalities, namely, traffic splitting and path selection.

Although, Internet topology involves path diversity, it is underexploited, as 30–80 percent of the time an alternate path with lower loss or a smaller delay exists and is never exploited [52]. Fortz et al. reported that load balancing improves the network service capability by 50%–110% compared to single path routing [53–55]. MPF had attracted a large body of literature. A number of load distribution models had been proposed and studied. LDMs in the literature can be coarsely categorized, as shown in Figure 2, into the following two groups: static load distribution model (SLDM) and adaptive load distribution model (ALDM).

Information less models use packet as the traffic unit. It makes a raw decision on distributing traffic without

taking into account packet information [56–58]. Static hash-based models choose the path in terms of flows instead of packet-per-packet splitting. Such models process a flow as an allocation unit that must be traversing the same path. Such models calculate a hash over selected fields in the packet header. There are a variety of hash-based models such as direct hashing (DH), table-based hashing (TH) [59, 60], and fast switching (FS) [61].

ALDMs take into considerations network conditions such as average path delay, link utilization, packet interarrival time, and capacity in path selection. Chim and Yeung [62] proposed adaptive hash-based LDM named table-based hashing with reassignment (THR) that helps to redistribute the traffic load. THR improves the TH algorithm by combining actual load sharing statistics and dynamically reassigns some active flows (bin-to-path mapping) from the overutilized paths to underutilized paths. However, out-of-order packets delivery is still present.

Kandula et al. [63] proposed flowlet aware routing engine (FLARE) that operated on flowlet. In order delivery of flows can be achieved via assigning flows to any available path, if the time between two successive TCP packets is larger than the maximum delay difference between the parallel paths. However, the common limitation lies in the estimation process which may be inaccurate at high packet arrival rate and this may yield load imbalance.

Tian et al. [64] provided network-wide load balancing performance by introducing link-criticality-based ECMP

routing (LCER) algorithm. LCER selects path based on the link's average expected load, link capacity, and the path's length. LCER provides in-order packet delivery and the lowest average end-to-end packet delays.

Y. Wang and Z. Wang [65] considered multipath routing as an optimization problem (OP) with an objective function that minimizes the congestion of the most utilized link in the network. However, they did not consider the quality of the selected paths. Banner and Orda [66] proved through comprehensive simulations that multipath solutions obtained by optimal congestion reduction schemes are fundamentally more efficient than solutions obtained by heuristics. They formulated MPR problem as an optimization problem of minimizing network congestion. They established a polynomial time algorithm that approximates the optimal solution by a (small) constant approximation factor. However, this method is not a direct solution of potential of QoS necessities.

MPF can provide a unique solution to congestion problems by utilizing the available resources in an adaptive way to the dynamics of traffic demands. One way to prevent congestion is to control the delay within the network. Thus, network capacity and QoS provisioning need a new load distribution model aiming to minimize the difference among path delays, thereby reducing packet delay, jitter, and risk of packet reordering without additional network overhead.

Motivated by the scarceness of solutions to efficiently control packet delay for real-time traffic, Prabhavat et al. developed enhanced delay controlled load distribution model (EDCLD) [67]. EDCLD is an interesting packet-based delay-controlled LDM developed to strike the lower delay and packet ordering to utilize parallel paths for multimedia data transmission and real time applications. Prabhavat et al. formulated a delay-aimed problem model to figure the optimal load ratio and its corresponding path. EDCLD used iterative method to calculate optimal traffic-splitting vector so that maximum path delay can be minimized. The trick of EDCLD is to reduce the difference between path delays by using adaptive load adaptation algorithms that gradually, according to the number of paths, approach traffic-splitting vector among the paths. EDCLD decreased load assigned to the path with the largest delay and increased load by the same amount to the other path with the smallest delay. In the path selector, they implemented the SRR load sharing algorithm [68].

Li et al. [69] proved that the optimization problem of EDLCD is convex. They proposed a convex based method (CBM) that defines the optimal load ratio in one shot rather than gradually approaching algorithm used by EDLCD. Their proposed scheme outperforms EDCLD specifically with instability and large number of paths. Their proposed scheme also relied on the SRR load sharing algorithm.

Packet-based scheduling can achieve very accurate splitting percentages and adds very little extra overhead. However, it suffers from major problems which are packet reordering and TCP throughput degradation.

The major drawback of static hashing is a load imbalance problem due to an inability to deal with variation of the flow size distribution. Thus, one major challenge of TE is supplementing adaptive control capabilities that adapt quickly to significant changes in a network's state.

Majority of ALDMs is linked to TCP-traffic only and focus on load balancing efficiency and packet order preservation. These schemes are unsuitable for QoS oriented applications as they cannot guarantee low delay and packet ordering. In addition, Martin et al. [70] studied multistage network architecture. They discover that all pure hash-based algorithms have one serious problem, namely, traffic polarization effect (TPE). Also, Shi et al. [71] proves that pure hash-based algorithms cannot well balance load in the face of the highly skewed flow-size distributions in the Internet.

On the other hand, EDCLD and CBM are considered effective for real-time applications MPF. However, EDCLD used a gradually approaching method that needs several rounds, depending on the number of paths, to reach convergence and cannot handle paths instability. CBM limitation lies in solving a nonlinear optimization problem which can incur a significant computational overhead when performed on a per-packet basis.

*Open Challenges.* There is a crucial need for an effective MPF scheme that optimizes the network, with the joint goals of avoiding network congestion and ensuring QoS provisioning.

## 5. Proposed DQARE Architecture

The proposed DQARE architecture is based on extending the current Internet routing model of the OSPF<sub>x</sub>QoS routing engine to support strict QoS. DQARE architecture inherits the cons of the OSPF<sub>x</sub>QoS routing engine architecture explored in [8, 72] and nullifies the drawbacks of OSPF<sub>x</sub>QoS architectures. It is worthy to divide the control plane functionality among modules on different cards of router, namely, control card and line cards, exploiting the power of next generation routers to move some control functions to line cards.

Basically, as Figure 3 depicts, the proposed software architecture of a router is composed of three planes connected by interfaces: (i) forwarding plane (or data plane): the main task of this plane is to forward flows in a way that prevents congestion. So, forwarding plane is supplied with MPF model. (ii) Control plane hosts routing protocols that are responsible for establishing routes within an AS, routing table management, sending and receiving link state updates and computing shortest paths. The control plane computes forwarding information table (FIT) from one or several routing tables that are used by forwarding plane. Some of the functions of control plane are implemented on the control card and others on the line cards. (iii) Management plane handles network management applications, protocol policies, and QoS. This plane performs congestion and path management.

For QoS provisioning, the most significant parts in adoption are relying on prioritization approach instead of reserving resources as in OSPF<sub>x</sub>QoS, the changes to the routing algorithm in the control plane that computes diverse paths to the forwarding plane subject to two additive constraints, the implementation of load balancing, and protection algorithms to the forwarding plane.

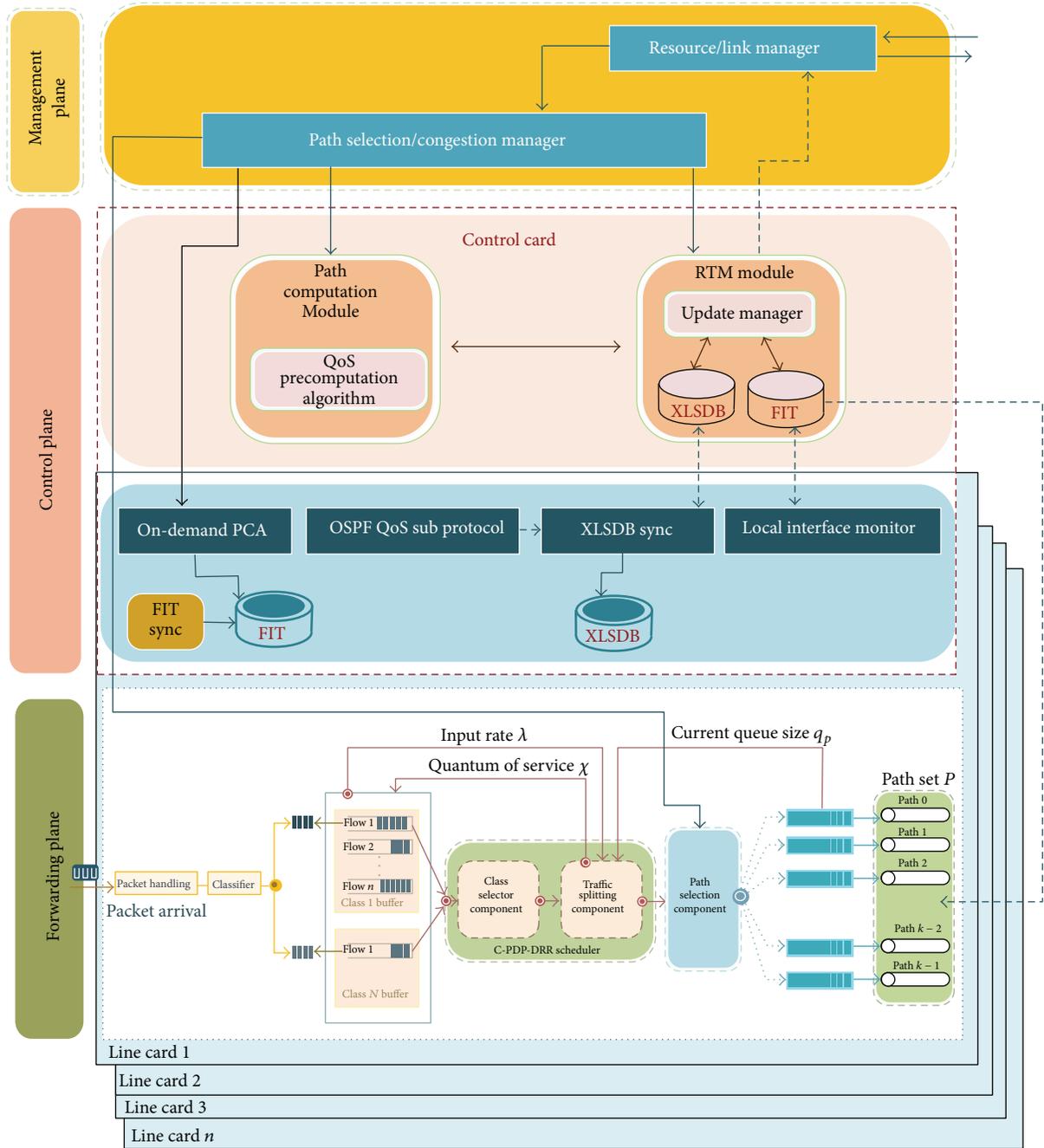


FIGURE 3: Proposed DQARE architecture.

The functional flow of the proposed framework is illustrated in the flowchart depicted in Figure 4. The functional blocks of the proposed architecture are as follows.

**Packet Handling.** Packet handling involves the following functions: *IP Packet Validation*: as a packet enters an ingress port, the forwarding logic verifies all layer 3 information (header length, packet length, protocol version, checksum, etc.) and *Route Lookup and Header Processing*: the router then performs on-demand path computation using the packet's destination address and QoS constraints to determine the

output of the egress port(s) and performs all IP forwarding operations (packet lifetime control, header checksum, etc.).

**Classifier.** QoS applications operate on packet flows so the routers must be able to classify individual traversing packets. Actually, it is very difficult to guarantee the delay bound to specific flows without flow isolation. DQARE architecture relies on marking packets using DSCP bits to identify the class of traffic. Flows are firstly grouped in traffic classes by the classifier. Classification of traffic provides more equitable management and more stability in the use of pure priorities.

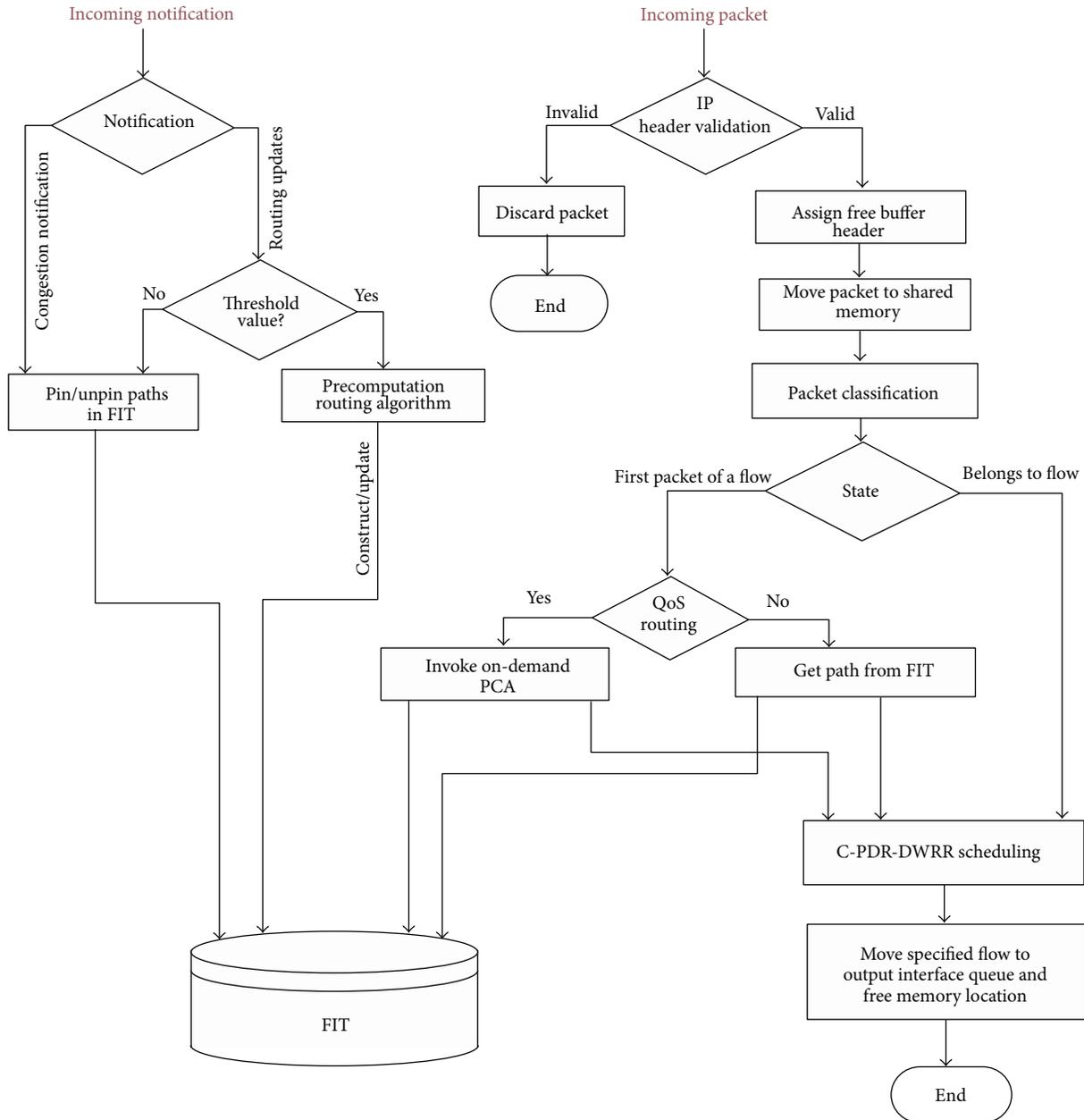


FIGURE 4: DQARE architecture functional flow.

We consider the objective of service differentiation to ensure low packet delay for streaming applications. All packets in the same class are treated equally by the *C-PDR-DWRR* scheduler.

*Path Computation Module.* Assuming the router maintains link state information of the entire domain, DQARE architecture uses two types of routing algorithms, firstly PQRA implemented in the control card. PQRA precomputes the routing paths from a node to all destinations subject to two additive metrics, prior to receiving the requests, and stores the QoS information in its routing table. When receiving a connection request with QoS requirements, OQRA implemented on line

cards computes a path from offline precomputed routing table and finds an optimal or a feasible path for this request if found. On-demand path computing can be running again in the congestion state with benefit of ensuring a strict bound on the computational load.

*QoS Load Balancing Module.* multipath Intra-domain forwarding is used to handle traffic congestion inside a domain. DQARE architecture is equipped with MPF mechanism to perform traffic forwarding. The load balancing module consists of combined proportional delay prioritization and dynamic weighted round robin (*C-PDR-DWRR* scheduler) and path selection component.

*C-PDP-DWRR* is work-conserving scheduler that works in two phases. In the first phase, class based priority scheduling is performed to select class to serve from multiple classes. Also, the first phase achieves proportional queuing delay differentiation protection among various classes. Proportional service differentiation, originally proposed by Dovrolis et al. [73], is perhaps the best known effort to enhance class-based services with relative guarantees.

Dovrolis et al. proposed proportional delay differentiation (PDD). The network traffic is grouped into  $N$  classes of service which are ordered, such that class  $i$  is better (or at least no worse) than class  $i - 1$  for  $1 < i \leq N$ , in terms of queuing delays.  $d_i$ ,  $\delta_i$  denote the average queuing delay and delay differentiation parameter (DDP) value. The PDD model aims to control the ratios of the average class queuing delays based on the delay differentiation parameters (DDPs)  $\{\delta_i : i = 1, 2, \dots, N\}$ . The PDD model requires that the ratio of average delays between two classes  $i$  and  $j$  is fixed to the ratio of the corresponding DDPs as follows:

$$\frac{d_i}{d_j} = \frac{\delta_i}{\delta_j}. \quad (1)$$

Higher classes provide better service, that is, lower queuing delays, and so  $\delta_1 > \delta_2 > \dots > \delta_N > 0$ .

In the second phase, the traffic splitting component defines the amount of traffic ( $\chi$ ) forwarded on the selected path. Within the selected priority class, flows are scheduled in DWRR manner [74]. DWRR, on the other hand, allows higher priority queues to send a predetermined amount of data during a service round. Each queue is configured with a quantum of service ( $\chi$ ) and a deficit counter (DC). The scheduler also has the task of path delay adaptation by decreasing quantum of service ( $\chi$ ) on the path having the largest estimated end-to-end delay and then increases the quantum of service ( $\chi$ ) on the path having the smallest estimated end-to-end delay by the same amount of the reduced load.

*Extended Link State Database (xLSDB)*. It consists of link state information. This information includes both static (delay, jitter, and loss rate) and dynamic metrics (current available bandwidth) of the whole topology.

*Path/Congestion Manager*. It selects a path for a request with particular QoS requirements and manages it once selected; that is, it reacts to link or reservation failures. It finds alternative paths by invoking OQRA again in the case that the used path becomes unavailable. It invokes the traffic splitting component to select a flow to be shifted during congestion. This module also manages congestion handling actions when a congestion notification delivered from designated router (DR) or any link becomes congested and its associated queues reach their threshold level.

*Routing Table Manager (RTM) Module*. The main task of the RTM is to build FIT that stores, if existing, multiple routes to the same destination from precomputation routing algorithms. It contains update manager module which determines when to advertise local link state updates and

when to perform QoS path precomputation and control the paths within routing table; that is, it activates paths that meet the requirements and deactivate routes that currently does not meet the requirements. It contains xLSDB that contains information about the current state of the network.

*Resource/Link Manager*. Its main role is to manage individual and bundled interfaces. It monitors the load on resources, handles the up/down status, and receive notifications from designated routers.

*Local Interface Monitor* in the control plane handles the up/down status of each router interface.

As shown in Figure 5, DQARE architecture exploits GANA [72] self-adaptation mechanism which is a typical example of a protocol-intrinsic control loop. The OSPF protocol acts as a virtual distributed decision element (DE) scattered all over the domain. GANA provides different types of basic network services such as autonomic routing and advanced services such as QoS management.

DQARE architecture provides the self-adaptive control loop to prevent instantaneous congestion, which involves invalidating congested routes in the FIT, using other available paths if they exist, precomputation of routes whenever a threshold of updates is reached, a global resynchronization of xLSDB, and recalculations of the routing tables. This scheme seems to make a comprehensive overhead; however, global, reactive response makes the convergence of the control loop acceptable because each flooding message is sent only once instead of by all interfaces, so it reduces the traffic in the network and the proposed framework enables direct communication among line cards belonging to the same OSPF area which makes improvement over traditional architectures which yields that the reaction to failures is not a lengthy process.

The autonomic management components involved in the control loop are (i) *QoS path selection and congestion management*: responsible for sending and receiving notifications to/from the designated router (DR), (ii) *monitoring entity*: monitoring the local interfaces, links, and notifications. It can be implemented via the Hello protocol, (iii) *QoS routing tables execution*: implemented via precomputation algorithm that can be executed periodically or after receiving  $n$ -updates, and (iv) *managed entities* which control loop effects. It consists of FIT and extended LSDB (xLSDB) with a predefined set of actions to be performed such as sanction and neutralizing routes in FIT.

## 6. Service Differentiation Guidelines

QoS provisioning for various applications requires: (i) studying Internet applications and recognizing their QoS requirements, (ii) organize applications into classes and (iii) differentiate between different types of services by marking packets with a distinct code so that they receive certain kinds of treatment from routers.

*6.1. Internet Traffic Classification*. Traffic flowing in a network can be divided into two groups, network-oriented traffic and

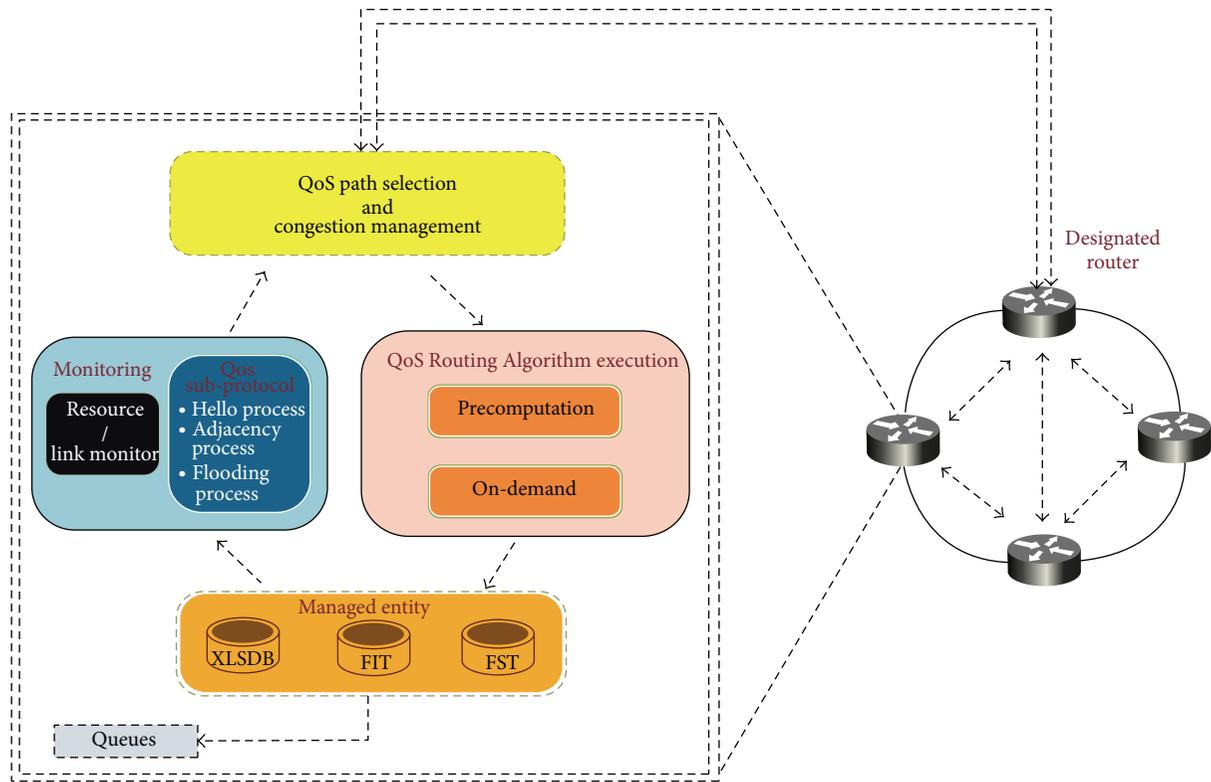


FIGURE 5: Self-adaptation mechanism.

user-oriented traffic [25]. The network-oriented traffic group is divided into three service classes, namely, network control function, OAM (operations, administration, and management) for network configuration, and management functions and signaling to control applications or user endpoints. User-oriented traffic can be broadly classified into three categories: elastic applications (data-oriented), tolerant real-time applications, and intolerant real-time applications.

Real-time applications as a class of applications need the data in each packet by a certain time and, if the data has not arrived by then, the data is essentially worthless, and elastic application as a class of applications will always wait for data to arrive. Elastic applications are applications that built on top of TCP protocol. TCP is a reliable transfer protocol that uses acknowledgements users working with applications based on symbolic data that can tolerate significant delays and loss. Such applications are considered non-real-time applications (NRT) that do not have stringent timing requirements and do not need any assurance from the network [75].

Data-oriented traffic can be further classified into (i) NRT asymmetric (NRTA) in which the requests are considerably using less resources than responses, such category can be further classified into interactive class and bulk-transfer class. Interactive class is suited for applications that use short packets, such as Telnet, web browsing, and enhanced web browsing. Bulk-transfer class is suited for store and forward applications that uses long packets, such as SMTP and FTP, and (ii) NRT symmetric (NRTS) in which requests and

responses use the same amount of resources such as Internet chatting applications [76].

On the other hand, real-time applications (media-oriented) inherently have more stringent QoS requirements due to the nature of real-time transmissions. To achieve user satisfaction, the transmission infrastructure should strongly considers delay and jitter requirements to maintain system timing and constant data rate. Real-time applications can be divided into two broad categories: (i) tolerant real-time asymmetric applications (TRTA) are real-time applications that are very sensitive to delay bounds. Timeliness is very important for these real-time applications.

These applications can tolerate moderate end-to-end delay, so it is called soft real time. However, it requires high throughput and very low error rate. Common TRTA applications include those which are conversational in nature, such as multimedia conferencing that includes video conferencing, group of participants in teleconferencing audio, and audiographics conferencing that enables participants to share workspace and telephony service that involves videophone conferencing and VOIP, and (ii) intolerant real-time asymmetric applications (IRTA) demand more stringent QoS from the network. Such applications must have precise bandwidth, delay, and jitter constraints, and if the timing constraints are not met, such applications suffer from high performance degradation so it is called hard real time. Common IRTS applications are audio and video broadcasting, interactive audio, and video on demand and streaming media.

TRTA and IRTS applications are built on top of UDP protocol. UDP is unreliable protocol that does not have acknowledgment or flow control. So, such applications need more concern in QoS scope. QoS has three attributes to measure the output performance of a process: timeliness, precision, and accuracy. Timeliness measures the time taken to produce the output of the process. Precision measures the amount or quantity of the produced output. Accuracy measures the correctness of the produced output.

**6.2. DSCP Assignment.** By following the classification proposed in [25], some DSCP values may be dedicated to administrative and control traffic. Per hop behaviors (PHB) mapped by a codepoint with a larger numerical value should receive better or equal forwarding treatment than the one with the lowest numerical value. Respecting these guidelines, DSCP assignment is reported in Table 2, together with possible examples of traffic types and possible ranges of QoS performance parameters. The selected path should be calculated based on each application of QoS metric's requirement.

## 7. Routing Computation Framework

As reviewed in Section 4.3, the previously proposed algorithms (OQRA and PQRA) suffer from excessive computational complexities or low performance. This section presents a routing computation framework that proposes network diversities, by applying PQRA and OQRA routing paradigms into a real network, with low-computational complexity and high routing performance. In that sense we use dynamic programming (DP) technique. DP solves a sequence of larger and larger instances, reusing the previously saved solutions for the smaller instances, until a solution is obtained for the given instance. This simple idea can sometimes transform exponential-time algorithms into polynomial-time algorithms. Unlike the majority of routing algorithms, which assume an adjacency-list representation of the graph, most of the algorithms which rely on DP use an adjacency matrix representation. The adjacency matrix is a memory efficient way of representing dense graphs while linked list is more efficient for sparse graph [77].

The basic idea behind the proposed routing framework addresses the following issues: firstly, how to find all precomputed pairs of the shortest paths with two additive constraints and, secondly, how on-demand routing computation takes place when new requests arrive. The main idea is explored in two-phase framework.

In the first phase, QoS optimal paths precomputation routing algorithm (QOPRA), recursively, computes all-pairs-shortest paths in a graph of nodes, as presented in [78], connected in the forward direction according to  $w_1$  and in the backward direction according to  $w_2$  (Algorithm 1).

Each node precomputes the routing from itself into a destination prior to receiving the requests and stores this information in its FIT. The precomputation algorithm overhead is shared between different requests. Then, an on-demand routing algorithm (OQRA) (Algorithm 2) activates

**Input:**  
G  
**Output:**  
 $D^M, \pi^M, i$   
(1) Let  $D^0$  and  $\pi^0$  be  $[M \times M]$   
(2)  $D^0 \leftarrow \text{INF}$   
(3)  $\pi^0 \leftarrow \text{Null}$   
(4) **Initialize**  
(5) **For**  $x = 1$  to  $N$   
(6)     **For**  $y = 1$  to  $N$   
(7)         **if**  $(x, y) \in E$   
(8)              $d^0(x, y) = w(x, y)$   
(9)              $\pi^0(x, y) = x$   
(10)         **Else**  
(11)              $d^0(x, y) = \infty$   
(12)              $\pi^0(x, y) = \text{Null}$   
(13)      $d^0(x, x) = 0$   
(14)      $\pi^0(x, x) = \text{Null}$   
(15) **For**  $p = 1$  to  $M$   
(16)     Let  $D^p = d_{xy}^p$  be a new  $n \times n$  matrix  
(17) **For**  $x = 1$  to  $M$   
(18)     **For**  $y = 1$  to  $M$   
(19)          $d_{xy}^p = \min(d_{xy}^{p-1}, d_{xp}^{p-1} + d_{py}^{p-1})$   
(20)         **If**  $(d_{xy}^{p-1} \leq d_{xp}^{p-1} + d_{py}^{p-1})$   
(21)              $\pi_{xy}^p = (\pi_{xy}^{p-1})$   
(22)         **Else**  
(23)              $\pi_{xy}^p = (\pi_{py}^{p-1})$

ALGORITHM 1: QOPRA.

**Input:**  
Source node  $x, D^M, \pi^M, l_a, l_b$   
**Output:**  
 $\delta_{(x,y)}, \text{path}[], \text{HC}$   
**Steps:**  
(1) **For**  $y = 1$  to  $N$   
(2)     **If**  $(d_{xy}^M/l_a > d_{yx}^M/l_b)$   
(3)          $\delta_{(x,y)} = d_{xy}^M/l_a$   
(4)         PrintPath( $x, y$ )  $\rightarrow$  Path[]  
(5)         HC = path.length  
(6)     **Else If**  $(d_{xy}^M/l_a < d_{yx}^M/l_b)$   
(7)          $\delta_{x,y} = d_{yx}^M/l_b$   
(8)         PrintPath( $y, x$ )  $\rightarrow$  Path[]  
(9)         HC = path.length  
(10)     **Else**  
(11)         PrintPath( $x, y$ )  $\rightarrow$  Path1[]  
(12)         HC1 = path.length  
(13)         PrintPath( $y, x$ )  $\rightarrow$  Path2[]  
(14)         HC2 = path.length  
(15)         HC = min(HC1, HC2)  
(16)     **If** (path1[].length < path2[].length)  
(17)         Path[] = path1[]  
(18)     **Else**  
(19)         Path[] = path2[]  
(20) Return ( $\delta_{x,y}$  path[], HC)

ALGORITHM 2: OQRA.

TABLE 2: DSCP Assignment for Internet traffic.

Service Group	Service Category	Major Service classes	Flow C/Cs	RT (sec)	Timeliness Delay (ms)	QoS Metrics			DSCP Value	Traffic class	Minor Service classes	
						B.W	Reliability	Jitter (ms)				
User oriented	Best Effort	Default	Unclassified	U	U	U	U	U	000000	CS0-Default	Everything else	
	Data oriented	Low priority	Symmetric	1 sec	<200	N/A	Elastic	Zero Loss	001000	CS1	Internet Chat	
		Interactive	(i) short lived	(i) short lived	2-5	250-400	N/A	Elastic	Zero Loss	010010	AF23	Telnet
			(ii) Low Latency	(ii) Low Latency								Web Browsing
	Bulk Transfer	(iii) Asymmetric	(iii) Asymmetric	2-5	Low-Medium	N/A	Elastic	Zero Loss	010110	AF21	Enhanced Web	
		(i) long-lived	(i) long-lived								E-Mail	
	Media oriented	Throughput	(ii) High	(ii) High	2-5	Low-Medium	N/A	Elastic	Zero Loss	001010	AF11	FTP
			(iii) Asymmetric	(iii) Asymmetric								Billing transfer
		Multimedia Conferencing	(i) interactive	(i) interactive	Rate adaptive	<150	<400	8 kbps-1 Mbps	Very low loss	100110	AF43	Audio
			(ii) group communication	(ii) group communication								Audio graphics video
Tolerant Real Time Symmetric	Telephony Service	(iii) Rate adaptive	(iii) Rate adaptive	<150	<100	8 kbps-1 Mbps	Very low loss	101100	EF	VOIP		
		CBR, fixed small packets, Interactive and fast response	CBR, fixed small packets, Interactive and fast response							Videophony		
Intolerant Real Time Symmetric	Broadcasting	(i) Inelastic	(i) Inelastic	2-5	<150	50-100	low loss	011010	AF31	Broadcast Video		
		(ii) CBR and VBR	(ii) CBR and VBR							Broadcast Audio		
	Streaming	(i) Elastic	(i) Elastic	2-5	<150	<100	64 k-60 M	low loss	011000	CS3	Streaming media	
(ii) Variable packet size		(ii) Variable packet size	VoD									
Network oriented	Interactive	(i) Inelastic	(i) Inelastic	2-5	<150	<100	64 k-60 M	low loss	100000	CS4		
		(ii) VBR	(ii) VBR									
	Routing and control information	Inelastic-Short messages	Inelastic-Short messages	N/A	1-10 s	N/A	Elastic	Zero Loss	110000	CS6	Routing information	
		OAM	OAM	N/A	50-100	N/A	Elastic	Zero Loss	010000	CS2	OAM	
Operation and Management signaling	Signaling Service Class	Signaling Service Class	N/A	50-100	N/A	Elastic	Zero Loss	101000	CS5	VOIP Signaling		

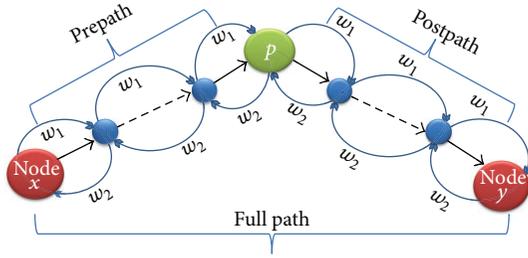


FIGURE 6: Finding MCSP from source  $x$  to destination  $y$ .

when there is an incoming request with QoS requirements to find the path that satisfies QoS constraints from these precomputed paths. Actually, proposed hybrid form can provide enough information to support efficient admission control, as well as less on-line computation overhead and high success rate.

**7.1. Related Notations and Problem Analysis.** For QoS assurance, we have to study all the subpaths between every pair of vertices in the graph instead of working with a single source to obtain more visibility and accurate path computation. However, this can be solved by repeating single source algorithm once for each vertex in the graph, but it requires more computations and incorporates more complexity.

Undirected graphs can be transformed into directed graphs, by replacing the undirected link with two directed links each assigned one weight. Using two generic nodes, labeled as nodes  $x$  and  $y$ , in a network of  $N$  nodes. Notations are as follows. Let  $d_{xy}^p$  denote the length of the shortest path from vertex  $x$  to vertex  $y$ , where only the first  $p$  vertices are allowed to be intermediate vertices. If no such path exists, then let  $d_{xy}^p = \infty$ . From this definition of  $d_{xy}^p$  it follows that  $d_{xy}^0$  denotes the length of the shortest path from  $x$  to  $y$  that uses no intermediate vertices (i.e., directly connected). So,  $d_{xy}^0 = \infty$  if the nodes are not directly connected; otherwise,  $d_{xy}^0$  has a finite value and  $d_{xx}^0 = 0$  for all vertices  $x$ . We have to rely on intermediate nodes and, accordingly, consider a node labeled  $p$  as intermediate between node  $x$  and node  $y$ .

As presented in Figure 6, for finding multi-dimensional shortest path from node  $x$  to  $y$  assume that we know: (i) a shortest path from vertex  $x$  to vertex  $y$  that allows  $p$  vertices as intermediate vertices according to weight ( $w_1$ ) denoted as  $d_{xy}^{p,w_1}$ , and (ii) a shortest path from vertex  $y$  to vertex  $x$  that allows  $p$  vertices as intermediate vertices according to weight ( $w_2$ ) denoted as  $d_{yx}^{p,w_2}$ . The two terms in the minimum operator (2) identify that either node  $p$  is on the shortest path from nodes  $x$  to  $y$  or not:

$$d_{xy}^p = \min \{d_{xp}^{p-1} + d_{py}^{p-1}, d_{xy}^{p-1}\}, \tag{2}$$

$$d_{yx}^p = \min \{d_{yp}^{p-1} + d_{px}^{p-1}, d_{yx}^{p-1}\}.$$

Furthermore,  $d_{xy}^m$  and  $d_{yx}^m$  represent the length of the shortest path from  $x$  to  $y$  in the last iteration according to weights ( $w_1$ ) and ( $w_2$ ), respectively. Ultimately, we wish to determine  $D^m$ , the matrix of the shortest path lengths  $d_{xy}^m$ . The shortest path

algorithm starts with  $D^0$  and calculates  $D^1$  from  $D^0$  and then  $D^2$  from  $D^1$ . This process is repeated until  $D^m$  (the shortest path matrix) is calculated from  $D^{m-1}$  using formula (2).

Formula (2) computes only one path between the nodes. It is beneficial to know the second or third shortest paths between two nodes. In order to compute  $k$ -shortest path, we have to execute a sequence of the arithmetic operations, namely, addition and minimization presented in [79]. Path lengths are now represented by a  $k$ -dimensional vector,  $d_i \in R^k$ .

Let  $a = [a_1, a_2, \dots, a_k]$  and  $b = [b_1, b_2, \dots, b_k]$  be members of  $R^k$ . Generalized minimization denoted by  $+$  and generalized addition denoted by  $\times$  are defined as follows:

$$a + b = \min_k \{a_i, b_i \mid i = 1, 2, 3, \dots, k\},$$

$$a \times b = \min_k \{a_i + b_i \mid i = 1, 2, 3, \dots, k\}. \tag{3}$$

To retrieve multiple paths, utilizing (3) and (2) can be replaced by the following equation:

$$d_{xp}^p = d_{xp}^p \times d_{xy}^p + d_{xy}^{p-1}. \tag{4}$$

The problem is to obtain path(s) from a source  $s$  to destination  $d$  on  $G$  that satisfies multiple QoS constraints  $L_i$ , where  $i = 1, 2, \dots, m$ . QoS routing problem can have different definitions in the literature as follows.

**Definition 1.** In multiconstrained feasible path (MCFP) problem, find a feasible path ( $P$ ) from  $s$  to  $d$  such that

$$w_i(P) = \sum_{(x,y) \in P} w_i(x, y) \leq L_i. \tag{5}$$

**Definition 2.** In multiconstrained optimal path (MC (O) P) problem, find all feasible paths from  $s$  to  $d$  satisfying (2) and in addition minimize some length function  $l(p)$  such that  $l(p) \leq l(p^\blacksquare)$ , for all feasible paths  $p^\blacksquare$  between  $s$  and  $d$  that satisfy MCFP.

**Definition 3.** In multiple-constrained shortest path (MCSP) problem, find optimal constrained shortest path (CSP) from  $s$  to  $d$  that obeys constraints and has the smallest hop count.

**Definition 4.** In  $K$ -multiple constrained shortest path (KMCSPP) problem, find  $K$  feasible constrained shortest path from  $s$  to  $d$  subject to multiple constraints, and list them in order of increasing length, where  $K$  is the number of paths.

Definitions 2 and 4 need path length  $l(P)$  to be able to compare paths. The first choice was a linear path length; Jaffe proposed to use the next definition [32]:

$$l(P) = \sum_{i=1}^m d_i w_i(P) \quad \text{where } d_i > 0. \tag{6}$$

The main advantage of linear path length algorithms is that by replacing each link vector via (6) by a single parameter, Dijkstra's algorithm can be deployed and so it is easy to

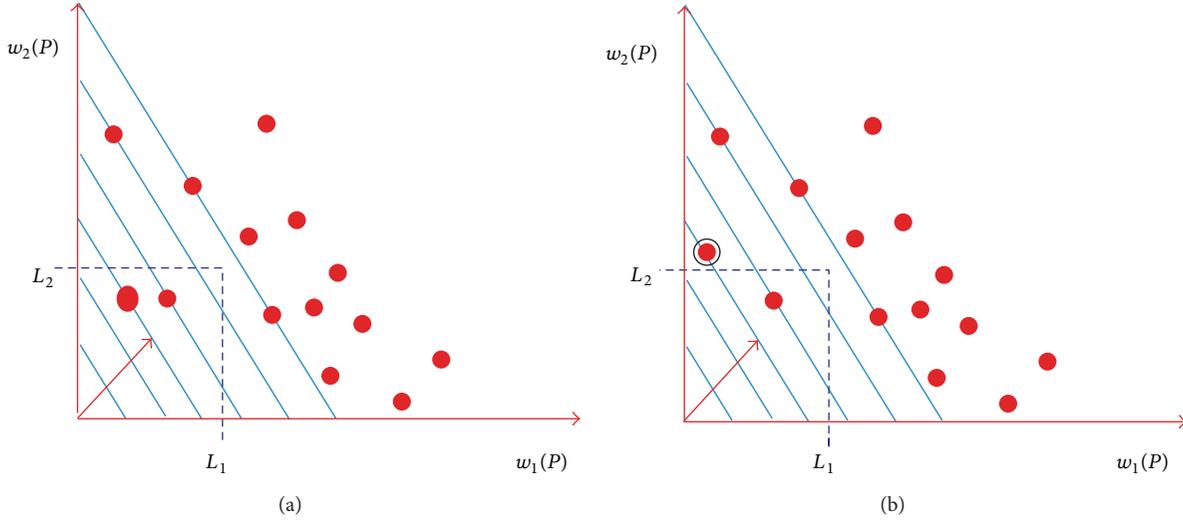


FIGURE 7: Using a linear path length, searching for a solution starts from origin until it hits a point: (a) algorithm succeeds to obtain a solution. (b) The algorithm fails as it finds solution outside constrained area.

develop a polynomial-time algorithm that minimizes  $l(P)$ . Unfortunately, their major drawback, as Figure 7 depicts, is that (i) the shortest path returned by Dijkstra’s algorithm is the first solution intersected by a set of parallel lines which may be infeasible (i.e., path outside the feasible region) and (ii) the area scanned outside the constrained area is considered large. Thus, Dijkstra’s algorithm will not always work satisfactorily.

To overcome drawbacks of relying on linear path length, Van Mieghem and Kuipers noticed that the area scanned outside the constraint area can be further reduced if the straight equi-length lines are replaced by curved equi-length lines that more closely approach the boundary of the constraint area [39]. Therefore, they recommended the deployment of non-linear representation of path length. Normalized nonlinear cost function for any path from the source to the destination is given as follows:

$$l(p) = \left(\frac{w_1(p)}{l_1}\right)^q + \left(\frac{w_2(p)}{l_2}\right)^q + \dots + \left(\frac{w_K(p)}{l_m}\right)^q, \tag{7}$$

where  $q \geq 1$ .

As Figure 8 depicts, feasible region can be scanned precisely. As  $q$  increases, the likelihood of finding a feasible path also increases. Therefore, to increase the probability of finding a feasible path, set  $q$  to  $\infty$  and use the following cost function for a path [34]:

$$l(p) = \max \left\{ \frac{w_1(p)}{l_1}, \frac{w_2(p)}{l_2}, \dots, \frac{w_K(p)}{l_m} \right\}. \tag{8}$$

The length function (7) considers the value of the most critical constraint of a path regarding the end-to-end requirements. Nonlinear length algorithms are likely to outperform linear length algorithms.

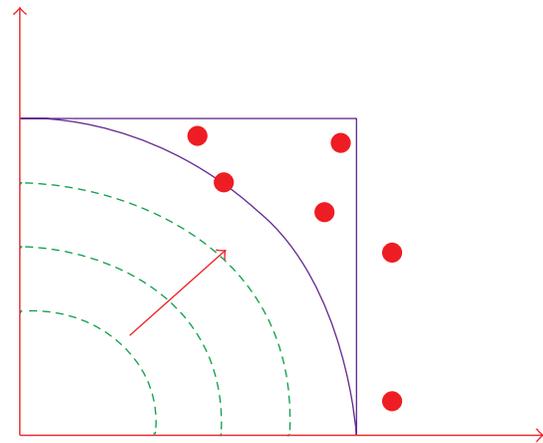


FIGURE 8: Using a nonlinear path length.

Thus, the shortest path from vertex  $x$  to vertex  $y$  according to weight  $(w_1)$  and weight  $(w_2)$  can be given as follows:

$$\delta(x, y) = \max \left( \frac{d_{xy}^m}{l_a}, \frac{d_{yx}^m}{l_b} \right). \tag{9}$$

7.2. Metacode. Metacode of QOPRA starts with initialization. The module INITIALIZE initializes the necessary parameters for the main algorithm (Algorithm 1). Determine the matrix  $D^0$  whose  $xy$ th elements equal the length of the shortest arc according to  $w_1$  from vertex  $x$  to vertex  $y$ , if any. The  $yx$ th element equals the length of the shortest arc from vertex  $x$  to vertex  $y$  according to  $w_2$ . If no such arc exists, let  $d_{xy}^0 = \infty$ . Let  $d_{xx}^0 = 0$ . The actual arcs that comprise each shortest path are also recorded in the  $\pi$  matrix from which we obtain tentative along the shortest paths. The INITIALIZE module sets  $\pi$  matrix as follows: if there is no direct link between

<p><b>Input:</b>  <math>i, j, \pi^m, PathVector</math></p> <p><b>Output:</b>  <math>i, p, \pi^m, PathVector</math></p> <p><b>Steps:</b>  (1) <math>p = \pi_{xy}^M</math>  (2) <b>if</b> (<math>p = x</math>)  (3) Return <math>PathVector</math>  (4) <b>Else</b>  (5) <math>PathVector.add(k)</math>  (6) <i>Return PrintPath</i>(<math>x, p, \pi^M, PathVector</math>)</p>
--

ALGORITHM 3: PrintPath.

nodes  $x$  and  $y$ , it sets  $\pi^0(x, y)$  to Nill. If there is direct link from nodes  $x$  to  $y$ , it sets  $\pi^0(x, y)$  to  $x$  and sets  $\pi^0(x, x)$  to  $x$ .

QOPRA successively determines the elements of  $D^1$  from the elements of  $D^0$  and the elements of  $D^2$  from  $D^1$  until obtaining  $D^M$  from  $D^{M-1}$  using the recursive formula (5). As each element  $D_{xy}^p$  is determined, record the corresponding path through the computation of matrix  $\pi^p$ . We need only to record one vertex for  $P_{xy}^p$ . If  $P_{xy}^p$  is known for all vertices  $x$  and  $y$ , then all the vertices along the shortest path from  $x$  to  $y$  can be found as follows: set  $P_{xy}^p$  equal to  $x$  for all  $y$ . Do this for all vertices  $x$ . Then, as the algorithm is performed, whenever the minimum on the left side of (9) is the first term, set  $P_{xy}^p$  equal to  $P_{py}^p$ . Otherwise, leave  $P_{xy}^p$  unchanged.

Upon termination of the algorithm, the  $xy$ th element of matrix  $D^M$  represents the length of the shortest path from vertex  $x$  to vertex  $y$  according to  $w_1$  and the  $yx$ th element of matrix  $D^M$  represents the length of the shortest path from vertex  $x$  to vertex  $y$  according to  $w_2$ . Also the  $\pi^M$  matrix contains the next-to-last vertex in that path.

The module OQRA (Algorithm 2) calculates the shortest path using nonlinear path length (4) from a source node  $x$  to all nodes using matrix  $D^M$  evaluated from the module QOPRA. The shortest path from vertex  $x$  to vertex  $y$  according to weight ( $w_1$ ) and weight ( $w_2$ ) can be obtained using (6).

OQRA gets the path that satisfies the length function (6) according to constraints ( $l_a, l_b$ ) and prints the corresponding path in a vector called Path[] and its hop count (HC) by calling the PrintPath module. If there is more than one path with the same length, it retrieves the one with the least hop count by calling the PrintPath module which returns the paths and the hop count HC1 corresponding to Path 1 and HC2 corresponding to Path 1.

OQRA procedure obtains the shortest path  $\delta_{(x,y)}$  and the nodes on that path by calling the procedure OQRA and  $\pi^M$  matrix. Obtaining the nodes on the shortest path can be achieved by calling the module PrintPath (Algorithm 3).

We often wish to compute not only the shortest path length, but also the vertices on the shortest path as well. The module PrintPath returns the nodes on the best path by using matrix  $\pi^M$  evaluated from the module QOPRA. It gets the

path by making a recursion calling to itself until the full path is obtained. It also returns the hop count of that path. It starts with an empty vector called PathVector and checks the  $\pi^M$  matrix for the path from  $x$  to  $y$  as follows: it sets  $p = \pi_{xy}^M$  and adds  $k$  to the PathVector and makes a recursive calling to itself until it reaches the source  $x$  in the  $\pi^M$  matrix.

## 8. Proposed QMPF Model

To implement proportional differentiation, the majority of related work proposed priority-based scheduling (PS) algorithms. PS enforces proportional delay differentiation by dynamically adjusting the priority of a given class as a function of the waiting time experienced by packets from that class [80, 81]. Alternatively, rate-based schedulers provide proportional differentiation by dynamically changing the service rates allocated to classes [82]. Here, we provide a slightly different approach. Instead of relying on actual waiting time experienced by each packet, we deploy the proportional model in the differentiation of average class queuing delay to define class priority and investigate appropriate packet scheduling mechanisms (Algorithm 4).

Average class queuing delay is determined by the arrival rate of packets, the amount of traffic removed from the queue, and other service classes (queues). We start by deriving the average queuing delay for each class. The derivation draws inspiration from [80–84]. In Figure 9, the concepts of arrival curve, input curve, and output curve for class  $i$  traffic are depicted.

**8.1. Average Queuing Delay per Class.** Consider a discrete, event-driven time model, where events are traffic arrivals with the following notations:

- (i)  $t(n)$ : the time of the  $n$ th event,
- (ii)  $\Delta t(n)$ : the time elapsed between the  $n$ th and the  $(n + 1)$ th events,
- (iii)  $\lambda_i(n)$ : the class  $i$  arrivals at  $n$ th event,
- (iv)  $\chi_i(n)$ : the quantum allocated to class  $i$  at the time of the  $n$ th event. The quantum is the amount of bytes that class  $i$  could transmit if it is selected for transmission; otherwise, it is set to zero.

The arrival curve for class  $i$  at the  $n$ th event,  $A_i(n)$ , is the total traffic that has arrived to the transmission queue of class  $i$  since the beginning of the current busy period; that is,

$$A_i(n) = \sum_{k=0}^n \lambda_i(k). \quad (10)$$

Assuming that queuing model is lossless and that there are large enough buffers for packets that need to be queued and hence no traffic is dropped, then the input curve,  $R_i^{\text{in}}$ , is the traffic that has been entered into the transmission queue at the  $n$ th event that equals  $A_i(n)$ :

$$R_i^{\text{in}} = A_i(n). \quad (11)$$

**Input:**  
Traffic Units

**Output:**  
Path,  $\chi$

**Steps**

- (1) Receive incoming packets.
- (2) Classify incoming packets and enqueue them in appropriate class buffer using DSCP and 5-tuple information.
- (3) Compute queuing delay of each class using  $D_i(n) = t(n) - t(\sup\{k < n : R_i^{\text{in}} \leq R_i^{\text{out}}\})$
- (4) Compute average queuing delay of each class using  $\bar{d}_i(n) = (D_i(n) + B_i(n))/\chi_i(n)$
- (5) Among all classes, select class  $j$  with  $j = \max_i \bar{d}_i(n)\delta_i$ .
- (6) For each path available for flow  $f$  in class  $j$  calculate each path delay using:  
 $D_p(\chi) = \sum_{a \in P} D_a + (1 - \omega) \sum_{a \in P} (\varphi(l_a, c_a)/\mu_a \cdot \chi) + \omega(q_p/\mu_p)$
- (7) Assign flow  $f$  to the lowest delay path.
- (8) If end-to-end delay observed is less than the required delay, and if the next higher delay path can sustain the delay requirement of this flow, then it jumps to the higher delay path.
- (9) Send flow to lowest delay path whenever it observes a violation of the delay requirement.

ALGORITHM 4: QMPE.

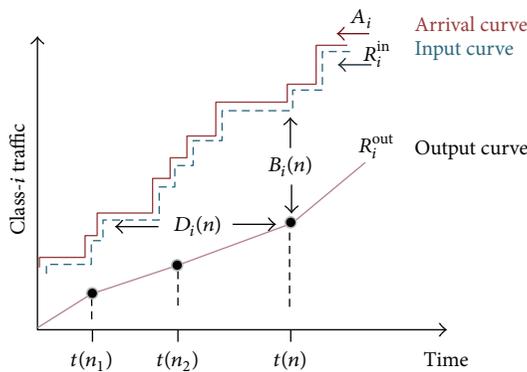


FIGURE 9: Concepts of arrival curve, input curve, and output curve for class  $i$  traffic.

The output curve is the traffic that has been transmitted since the beginning of the current busy period; that is,

$$R_i^{\text{out}} = \sum_{k=0}^n \chi_i(k) \Delta t(n). \tag{12}$$

For event  $n$ , the vertical distance between the input and output curves denotes the class  $i$  backlog  $B_i(n)$  and the horizontal distance denotes class  $i$  delay  $D_i(n)$ . For the  $n$ th event, we have

$$B_i(n) = R_i^{\text{in}} - R_i^{\text{out}}, \tag{13}$$

$$D_i(n) = t(n) - t(\sup\{k < n : R_i^{\text{in}} \leq R_i^{\text{out}}\}).$$

The minimum average delay  $\bar{d}_i(n)$  for all the packets that have already arrived to class  $i$  can be calculated as

$$\bar{d}_i(n) = \frac{D_i(n) + B_i(n)}{\chi_i(n)}. \tag{14}$$

At time  $n$ , the scheduler calculates for all classes (for transmission), the minimum possible normalized average delay, then selects the class with the maximum normalized average delay. That is, the order in which classes are selected for service is determined by the C-PDR-DWRR scheduler according to the following priority function:

$$j = \max_i \bar{d}_i(n) \delta_i. \tag{15}$$

So, high priority queues are serviced first such that the average delay experienced by packets in a delay class is inversely proportional to the delay weight of the class.

Within a selected class, queues are serviced in DWRR fashion. The main objective is to provide accurate control over the amount of data (i.e., quantum of service ( $\chi$ )) sent to the path selected by path selection component. The quantum of service ( $\chi$ ) is proportional to path bandwidth and output buffer size. The quantum of service ( $\chi$ ) is calculated such that maximum end-to-end delay can be minimized. The deficit counter (DC) specifies the total number of bytes that the queue is permitted to transmit. The DC of a queue is incremented by a quantum ( $\chi$ ) each time the queue is visited by the scheduler.

8.2. End-to-End Delay and Service Quantum Computation.

Nodes (vertices) will be labeled with the generic label  $v$  ( $v = 1, 2, \dots, V$ ), links (edges) with label  $a$  ( $a = 1, 2, \dots, A$ ), and flows within a service class with label  $f$  ( $f = 1, 2, \dots, F$ ). The capacity of link  $a$  will be denoted by  $c_a$ . Each flow  $f$  is characterized by the flow volume denoted by  $h_f$ . For flow  $f$  the total number of assigned paths is denoted by  $P_f$  and they are labeled with  $p$  from the first path to the total number of paths; that is  $p = 1, 2, \dots, P_f$ ; this sequence is called the list of candidate paths.

To tie it to generic flow  $f$ , we write the list of paths as  $Pf = (Pf1, Pf2, \dots, PfP_f)$ . Flow volumes are realized by means of flows assigned to paths on their routing lists. The

flow realizing flow  $f$  on path  $p$  is denoted by  $\chi_{fp}$  ( $p = 1, 2, \dots, Pd$ ). Suppose that we denote the vector of flows assigned to flow  $f$  with  $\chi_f = (x_{f1}, x_{f2}, \dots, x_{fPf})$  for path indices  $p = 1, 2, \dots, Pf$ ; then we arrive at

$$\chi_{f1} + \chi_{f2} + \dots + \chi_{fPf} = h_f. \tag{16}$$

In summation notation, we can write this as

$$\sum_p \chi_{fp} = h_f \quad f = 1, 2, 3, \dots, F. \tag{17}$$

In general, the vector of all flows (path-flow variables) will be called flow allocation vector or simply flow vector, which can be written as

$$\begin{aligned} \chi &= (\chi_1, \chi_2, \dots, \chi_f) \\ &= (x_{11}, x_{12}, \dots, x_{1P1}, x_{21}, x_{22}, \dots, \\ &\quad x_{2P2}, \dots, x_{f1}, x_{f2}, \dots, x_{fPf}) \\ &= (x_{fp} : d = 1, 2, \dots, f; p = 1, 2, \dots, Pf). \end{aligned} \tag{18}$$

Assume firstly that flow arrival follows the Poisson process and that the flow size is exponentially distributed. The system can be thought of as the famous  $M/M/1$  queueing system in which arrival rate  $\lambda$  and the service rate  $\mu$  are the same. If the average size of a flow assigned to a link is denoted by  $\chi$  bits, then the average service rate and arrival rate of the link are

$$\begin{aligned} \mu_a &= \frac{c_a}{\chi}, \\ \lambda_a &= \frac{l_a}{\chi}. \end{aligned} \tag{19}$$

Minimizing average path delay being weighted by its corresponding traffic of all paths between source and destination pairs is essential. Path delay is the time it takes a packet to travel across the network from one end to the other end.

Efficient QoS LDM must provide well-tailored load balancing and preserve packet ordering. Well load balancing can be achieved via assigning load on each path properly with respect to path bandwidth and buffer size. The packet ordering is likely to increase in a network with a large degree of parallelism. Assigning packets to different paths which have the same delay leads to increasing the probability of packet ordering preserving.

Thus, minimizing average path delay is crucial and involves minimizing link delay for each link that belongs to the path. The link delay is composed of two components, namely, propagation delay and queuing delay. Propagation delay  $D_a$  is a fixed value, whereas, queuing delay ( $Q_a$ ) varies according to the input traffic rate ( $\lambda$ ), the bandwidth capacity of the path ( $\mu_p$ ), and the traffic splitting ratio ( $\chi$ ). Queuing delay can be decreased using load balancing. Thus, we can write

$$\begin{aligned} D_p &= \sum_{a \in p} D_l + Q_p, \\ Q_p &= \sum_{a \in p} \frac{l_a}{c_a - l_a}. \end{aligned} \tag{20}$$

At the link level, we should minimize average queuing link delay. The average queuing link delay can be evaluated using the following function ( $F$ ):

$$F = \frac{l_a}{c_a - l_a}, \quad \text{for } 0 \leq \frac{l_a}{c_a} < 1. \tag{21}$$

Function  $F$  is a nonlinear convex function which is discontinuous at  $l_a = c_a$ . The good news is that it is possible to substitute convex mathematical programming problems with their piecewise linear approximations problems. Function  $F$  can be transformed into linear by using piecewise linear approximation presented by Fortz et al. [53–55]. Fortz and Throup proposed a six-segment piecewise linear cost function which is useful in tuning IGP metric that is where the routing cost for each arc is an increasing convex function of its utilization. Fortz and Throup function is given by

$$\dot{\varphi} = \begin{cases} 1 & \text{for } 0 \leq \frac{l_a}{c_a} < \frac{1}{3} \\ 3 & \text{for } \frac{1}{3} \leq \frac{l_a}{c_a} < \frac{2}{3} \\ 10 & \text{for } \frac{2}{3} \leq \frac{l_a}{c_a} < \frac{9}{10} \\ 70 & \text{for } \frac{9}{10} \leq \frac{l_a}{c_a} < 1 \\ 500 & \text{for } 1 \leq \frac{l_a}{c_a} < \frac{11}{10} \\ 5000 & \text{for } \frac{11}{10} \leq \frac{l_a}{c_a} < \infty. \end{cases} \tag{22}$$

The Fortz and Throup function is a piecewise linear envelope of the load latency function, scaled by  $c$ . Thus, we can say that  $c_a \cdot F \approx \dot{\varphi}$ . For the objective of minimizing weighted average queuing path delay,

$$\text{Minimize } F(x) = \sum_{a \in P} \frac{\varphi(l_a, c_a)}{c_a}, \tag{23}$$

where

$$\begin{aligned} l_a &= \lambda_a \cdot \chi, \\ C_a &= \mu_a \cdot \chi. \end{aligned} \tag{24}$$

Thus, we can write the end-to-end path delay considering  $M/M/1$  queueing system that can be formulated as follows:

$$D_p(\chi) = \sum_{a \in p} D_a + \sum_{a \in p} \frac{\varphi(l_a, c_a)}{c_a}. \tag{25}$$

Formula (25) is designed for Poisson traffic and is thus likely not practical for a real network under different traffic conditions. As in [67], with the assumption that input traffic is a combination of Poisson traffic and unknown traffic, a third term is added to formula (25). The third term evaluates the

waiting time of the current packet at an input queue. Formula (25) becomes

$$D_p(\chi) = \sum_{a \in P} D_a + (1 - \omega) \sum_{a \in P} \frac{\varphi(l_a, c_a)}{\mu_a \cdot \chi} + \omega \frac{q_p}{\mu_p}, \quad (26)$$

where  $\omega$  is weight factor that controls the weight between theoretical queuing delay and instantaneous queuing delay.  $q_p$  is the current queuing size of the buffer of each path  $p$ . The optimization problem can be formulated as follows:

$$\text{Minimize } \max_{p_i \in P} D_p(\chi) \quad (27)$$

subject to

$$\begin{aligned} \sum_p \chi_{fp} &= h_f \quad f = 1, 2, 3, \dots, F \\ \sum_p u_{fp} &= K_f \quad f = 1, 2, 3, \dots, F \\ x_{fp} &\leq u_{fp} h_f \quad f = 1, 2, 3, \dots, F \end{aligned} \quad (28)$$

constants:

$$\delta_{efp} \begin{cases} =1 & \text{if link } a \text{ belongs to path } p \text{ realizing flow } f \\ =0 & \text{Otherwise} \end{cases}$$

$h_f$ : volume of flow  $f$

$K_f$ : predetermined number of paths for flow  $f$

variables:

$\chi_{fp}$ : quantum of flow allocated on path  $p$

$u_{fp}$ : binary variable corresponding to the flow variable  $f_{dp}$ .

## 9. Performance Evaluation

The aim of experiments presented in this section is to demonstrate the effectiveness of proposed DQARE architecture and prove that it is capable of overcoming OSPFxQoS limitations. During the experiments, a number of simulations were conducted using MATLAB and NS2 [85]. Firstly, OSPFxQoS and DQARE behaviors are evaluated. Secondly, the proposed routing algorithm and multipath forwarding model have to be evaluated with their counterparts in the literature.

**9.1. OSPFxQoS and DQARE Architectures Evaluation.** OSPF and DQARE behavior has been investigated for different scenarios built under network simulation tool (NS2) environment [85]. NS2 is a discrete event driven simulator which means that it starts packet sending at the designated time and stops also at a determined time. In this experiment, the OSPFxQoS environment has been firstly implemented. Then, we modify OSPFxQoS operation with our proposed priority treatment, QoS routing, and forwarding mechanisms. We created network topology with Gt/itm tool that exists in NS2 simulator. We have taken networks of 10, 25, 50, 100, 150, 200, and 250 nodes and simulation time = 300 sec in our scenario files. OSPF costs of the links are assigned randomly according

TABLE 3: Simulation parameters.

Parameter	Value
Number of nodes	10–250
Routing protocol	OSPFxQoS
Simulation time	300 Sec
Packet size	50–100 byte
Traffic flow	TCP–UDP
Session arrival rate	0.02 ms
Traffic sources	CBR

to the guideline given in [5]. Also, each link delay is created randomly.

We generate cross traffic in all scenarios to account for the network traffic flowing through nodes. This cross traffic is generated as follows: source and destination nodes are randomly chosen. Then each source and destination pair exchange traffic, which follows a Poisson distribution. The elastic traffic and real-time traffic are created and delivered via this test network. As shown in Table 3, different parameters are settled; for example, TCP and UDP traffic are considered and, in all simulations, constant bit rate (CBR) is applied with intervals 0.02 ms.

**9.1.1. Performance Metrics.** For performance comparison between DQARE and OSPFxQoS, we choose five key performance metrics, namely, the average end-to-end delay, packet delivery ratio, throughput, and control overhead.

**Average End-to-End Delay of Data Packets (AD).** Average time taken by a data packet to arrive at the destination include delays due to route acquisition, reservations, buffering and processing at intermediate nodes, and retransmission delays at the MAC layer. AD can be expressed mathematically as

$$AD = \sum_{i=1}^N \frac{T_i^r - T_i^s}{N}, \quad (29)$$

where  $T_i^r$  is receiving time of packet  $i$ ,  $T_i^s$  is sending time of packet  $i$ , and  $N$  is the number of connections.

**Throughput (T).** The average of successful message delivery over a communication channel. It can be expressed mathematically as

$$T = \frac{N_D * S b}{TS \ s}, \quad (30)$$

where  $N_D$  is the number of delivered packets,  $S$  is packet size, and  $TS$  is the simulation time.

**Packet Delivery Ratio (PDR).** It is the ratio of number of data packets successfully received by hosts to the total number of data packets sent:

$$PDR = \frac{\sum \text{Number of packets received}}{\sum \text{Number of packets sent}}. \quad (31)$$

**Control overhead** is the ratio of total number of routing control packets sent to describe the changes in the dynamic

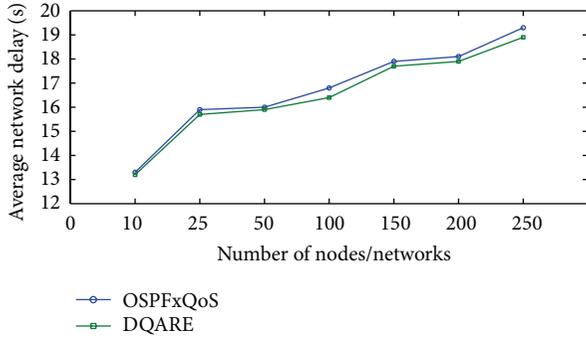


FIGURE 10: Average delay/network versus number of nodes.

topology to the total number of data packets delivered successfully:

$$\text{Control overhead} = \frac{\sum \text{Routing control Packets}}{\sum \text{Data Packets received}}. \quad (32)$$

### 9.1.2. Results

(1) *Varying Number of Nodes.* In the first experiment, we measure the performance of OSPFxQoS and DQARE by varying the number of nodes as 10, 25, 50, 100, 150, 200, and 250. Figure 10 depicts the performance comparison of end-to-end average delay in the network. DQARE almost outperforms OSPFxQoS. This is because OSPFxQoS used RVSP, while DQARE used priority treatment, classification, and scheduling. Actually, improvement in average end-to-end delay in DQARE results from the MPF model which makes use of multiple paths.

The difference of throughput between OSPF and DQARE environments, depicted in Figure 11, is quite obvious. OSPFxQoS gives single shortest paths based upon precomputation routing scheme. Retrieved paths may have common segments that become bottlenecked. When congestion occurs, OSPFxQoS cannot shift traffic to better alternative paths to mitigate congestion. On the other hand, DQARE architecture is supplied with QMPF model and path/congestion manger. Multiple paths exploitation is crucial for circumventing congestions scenarios. Also, if congestion occurs, path/congestion manager transfers quickly traffic flow from congested paths to another better path without severe loss of traffic.

Figure 12 shows the packet delivery ratio of OSPFxQoS and DQARE with different network topologies. We can find that as the number of nodes/network increases packet delivery fraction in case of DQARE outperforms OSPFxQoS. This is because OSPFxQoS lacks self-adaptation mechanism. Routers disseminate information only when topology changes. OSPFxQoS is unable to readjust forwarding paths in order to lessen the impact of failures. Also, OSPFxQoS is unable to load-balance traffic to optimize the performance of the network. On the other hand, DQARE uses an adaptive route mechanism implemented via path/congestion manager and QoS load balancing module. Thus, DQARE significantly enhances the usage of network capacity.

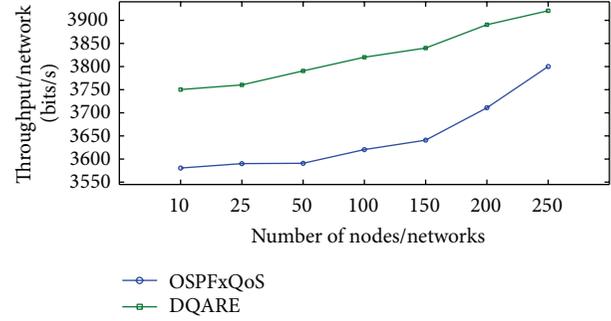


FIGURE 11: Average throughput/network versus number of nodes.

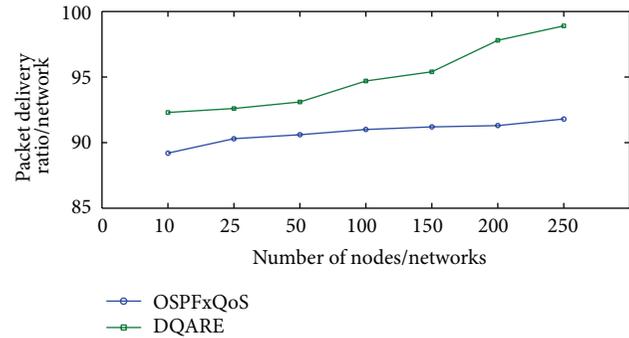


FIGURE 12: Packet delivery ratio/network versus number of nodes.

Figure 13 shows the performance comparison of routing overhead per network, and under this condition DQARE continues to outperform OSPFxQoS. Routing overhead in OSPFxQoS grows rapidly with the changes in the network topology; this is because OSPFxQoS deployed a precomputation routing algorithm routing that suffers from nonoptimal routing during congestion. OSPFxQoS needs to quickly initiate a new route discovery process when a link fails and therefore needs to consume a large amount of routing overhead. Our proposed routing algorithm retrieves multiple paths from a source to destinations and thus has the ability to use these multiple paths, so the overhead is smaller than OSPFxQoS.

(2) *Varying Packet Arrival Rate.* Figures 14, 15, 16, and 17 depict the behavior of both OSPFxQoS and DQARE for a network topology of 100 nodes. At each node the packet arrival rates are changed. Packet arrival rates are 50, 100, 150, 200, and 250 packet/sec. Figure 14 depicts that DQARE outperforms OSPFxQoS in providing lower delay with the increasing arrival rate. The average throughput and delivery ratio gap between DQARE and OSPFxQoS increase along with arrival rate. Figure 17 depicts the control overhead of DQARE architecture considered to be acceptable with the increasing in arrival rate.

9.2. *Routing Algorithm Evaluation.* Our first simulation is to show the intensity of the proposed routing algorithm by comparing the results with heuristics and exact algorithms in

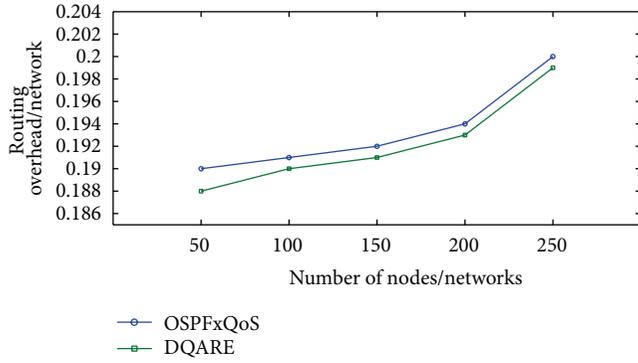


FIGURE 13: Control overhead/network versus number of nodes.

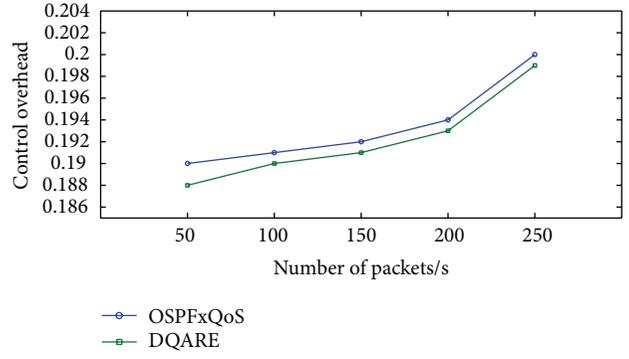


FIGURE 17: Control overhead versus arrival rate.

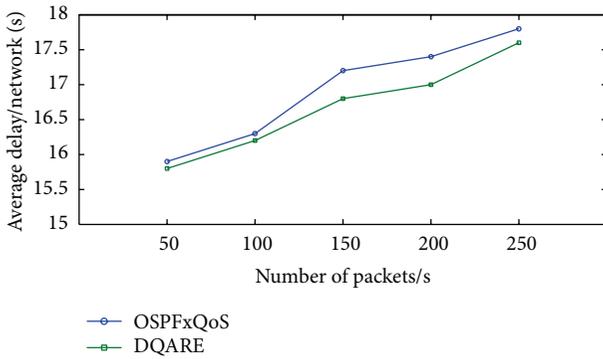


FIGURE 14: Average delay versus arrival rate.

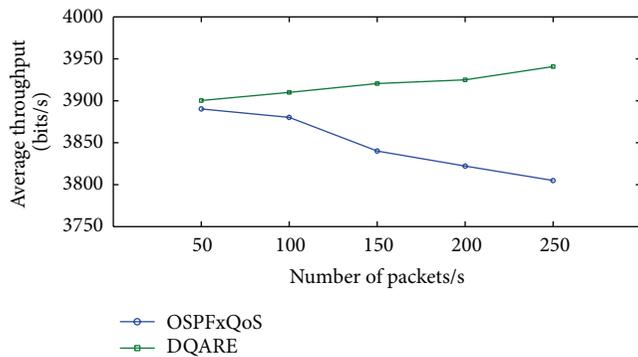


FIGURE 15: Throughput versus arrival rate.

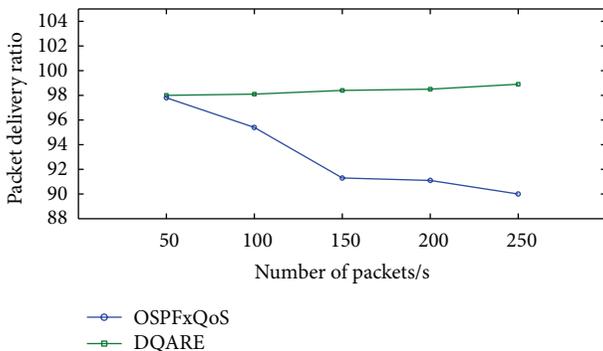


FIGURE 16: Packet delivery ratio versus arrival rate.

the literature. The proposed QoS routing algorithm is compared with RMCOP, HMCOP, SMACRA, HCA, and MPLMR algorithms coded in MATLAB 2012 and implemented on an Intel Core i3, 2.5 GHz CPU with 3 GB RAM running on Windows 7 professional.

9.2.1. *Network Topology.* Network topologies used for simulations are randomly generated based on Waxman’s model [86] with 50, 100, 200, 300, and 500 nodes. Waxman graph is considered counterpart to realistic telecommunication networks. In that sense, the location of nodes is randomly generated within the area of the graph. The probability of the existence of a link between nodes  $x$  and  $y$  is related to some function of the distance between these nodes. Formally, in Waxman graphs the probability ( $p_{xy}$ ) that two nodes  $x$  and  $y$  are connected equals  $f(\vec{r}_x - \vec{r}_y)$ , where  $\vec{r}_x$  and  $\vec{r}_y$  represent the position of node  $x$  and node  $y$ , respectively. So the farther the distance between two nodes, the smaller the need for a direct link between them. The probability function of Waxman’s model is as follows:

$$p_{xy} = \alpha \cdot \exp \left[ \frac{-\delta(x, y)}{\beta L} \right], \quad (33)$$

where  $\alpha$  represents the maximum link probability,  $\beta$  represents the parameter to control the link length,  $L$  is the maximum distance between two nodes in the graph, and  $\delta(x, y)$  is the distance between  $x$  and  $y$ . In experiments, we set  $\alpha = 0.8$  and  $\beta = 0.9$ .

9.2.2. *Performance Metrics.* We contrast the performance of various path selection algorithms using complexity and success rate (SR). Low complexity is the main goal of multi-constrained QoS routing schemes. A significant performance measure for the complexity of routing algorithms is time complexity in terms of execution time (ET). SR is the fraction of connection requests for which a feasible path is found. During all simulations, the success rate and execution time were stored.

9.2.3. *Simulation Model and Performance Measures.* After generating graphs, we associate two randomly generated additive weights with each link  $(i, j)$ . These weights are

TABLE 4: Ranges and correlations of link weights.

No correlation	Negative correlation
$w_1(i, j) \sim \text{unifrom}[1, 100]$	$w_1(i, j) \sim \text{unifrom}[1, 50]$
$w_2(i, j) \sim \text{unifrom}[1, 200]$	$w_2(i, j) \sim \text{unifrom}[100, 200]$
	$w_1(i, j) \sim \text{unifrom}[50, 100]$
	$w_2(i, j) \sim \text{unifrom}[1, 100]$

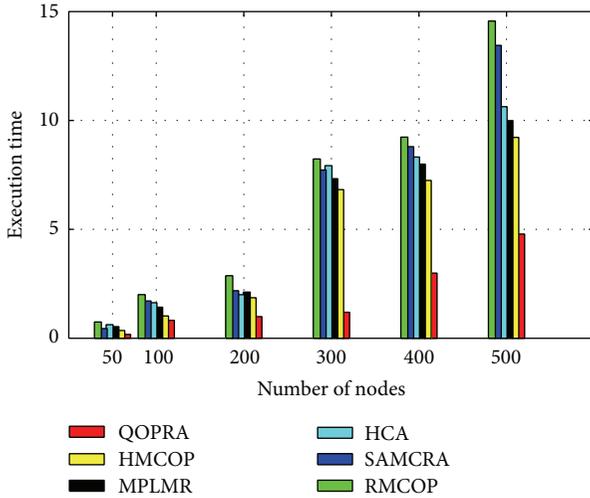


FIGURE 18: ET using uncorrelated link weights.

selected from uniform distribution sets. These weights are assigned, as depicted in Table 4, according to two types of correlation between them. No correlation assumes that both weights are independently selected from one set. While negative correlation assumes that one of the weights is selected from a set with small mean, the other is selected from another set with large mean.

In each run, source and destination nodes and QoS requirements of a request are randomly generated. The results reported in the subsequent sections are averaged over several runs. In each run, 10 random graphs are generated. For each random graph, ten independent link weights are generated using different random seeds. Each graph is subjected to 10 requests with different QoS constraints. There are about 1000 to 3000 connection requests that are generated for graphs with 50, 100, 200, 300, 400, and 500 nodes, respectively.

9.2.4. Results. Using extensive simulations on the Waxman random graph with uncorrelated link weights, Figures 18 and 19 show that under the same level of computational complexity proposed algorithms (QOPRA) outperforms its contenders in its success rate and computational complexity.

If the primary cost of HMCOP is not available it finds only a feasible path and postpath which may misguide the selection of prepath. The performance of HMCOP in finding feasible paths can be improved by using the  $k$ -shortest path algorithm and by eliminating dominated paths. SAMCRA worst case complexity grows exponentially and it may be subject to some error decision rate. The absolute complexity of HCA is greater than SAMCRA complexity.

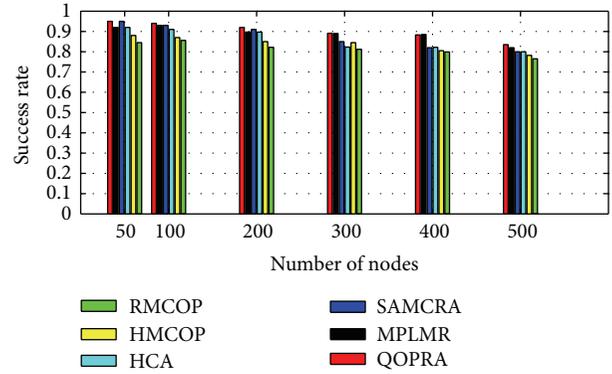


FIGURE 19: SR using uncorrelated link weights.

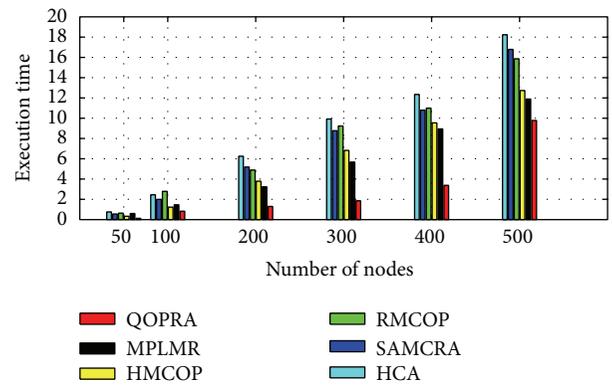


FIGURE 20: ET using negative correlated link weights.

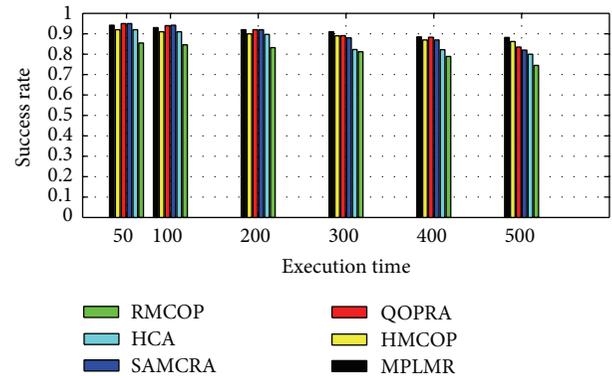


FIGURE 21: SR using negative correlated link weights.

HCA stops at the first feasible path it finds, so it has low execution time. RMCOP algorithm suffers from high complexity. MPLMR time complexity is comparable to HMCOP. However, MPLMR has a higher success rate than HMCOP.

As Figures 20 and 21 depict, using negatively correlated link weights, QOPRA still outperforms other algorithms. Negatively correlated link weights result in more paths in the network for which  $w_1(P) \gg w_2(P)$  and vice versa. This situation degrades the performance of MPLMR and HLA. Also, in such case algorithms such as SAMCRA incur a large

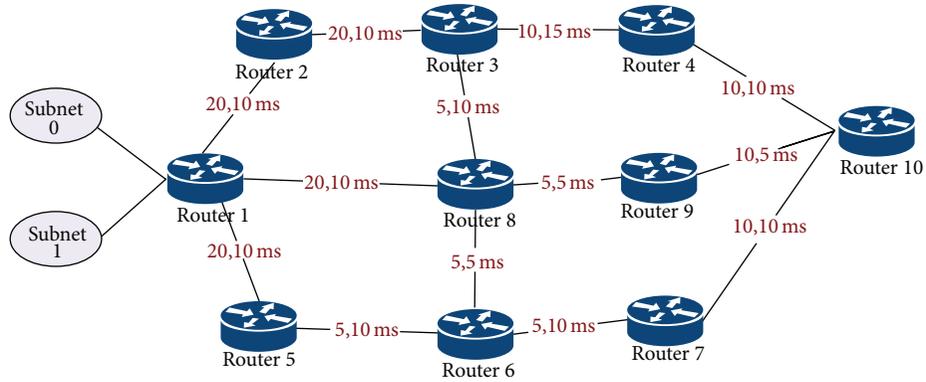


FIGURE 22: Simulated network topology.

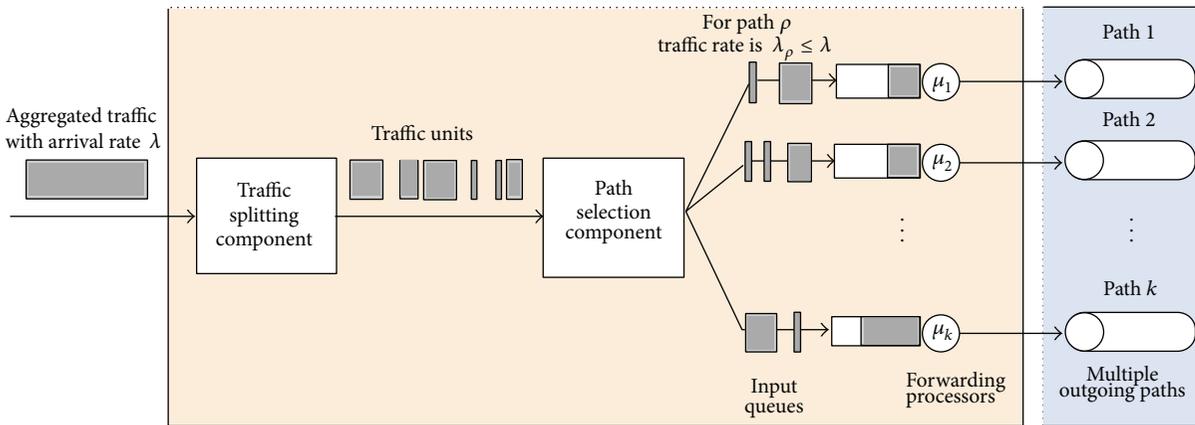


FIGURE 23: Data-plane extensions for MPF.

execution time with the increase in node number. However, SAMCRA achieves a higher success rate because it used path dominance and look-ahead techniques.

9.3. *MPF Algorithm Evaluation.* In this section, extensive network simulations are conducted to evaluate the performance of multipath forwarding mechanisms. We analyze the performance of proposed QMPF. FLARE, EDCLD, and CBM models are used for comparisons.

9.3.1. *Performance Metrics.* Simulation-based verifications are presented in terms of (i) end-to-end delay which can be defined as the sum of propagation delay and queuing delay as defined in (26), (ii) Jitter, variation of end-to-end packet delay, and (iii) total packet delay which is the sum of end-to-end delay and packet reordering recovery delay.

9.3.2. *Simulation Method.* Our simulations are based on NS2 [85]. Figure 22 shows the network topology adopted for the simulations. Each link is assigned with bandwidth and fixed propagation delay. The buffer size of router is set to 220

packets. TCP traffic and UDP traffic are generated from two subnets, 0 and 1, destined for node 10. Each subnet represented 50 traffic-generating hosts.

Each router is supplied with multiple paths to node 10. MPF mechanism is conducted under the environment shown in Figure 23. The input traffic to each node from 1 to 10 will be split into available paths. Load condition varies from low to high. The mean service time is inversely proportional to the bandwidth capacity ( $1/\mu$ ). The parameter  $\lambda$  is proportional to the total bandwidth of the paths. The mean packet arrival rate is chosen such that the ratio of the mean offered load to the mean service rate  $\lambda/\mu$  varies from 0.1 to 0.9 with a step size of 0.1. Parameter  $\omega$  in (23) is chosen to be 0.5.

9.3.3. *Simulation Results.* Figure 24 compares the mean of end-to-end delay achieved by various LDMs. As the ratio of input rate to output rate (i.e.,  $\lambda/\mu$ ) increased, the mean value of total packet delay rises as well. For low to medium load QMPF achieves the least end-to-end delay. When the traffic load becomes heavier, the performance of QMPF degrades due to classification and scheduling overhead. Figure 24 also shows that CBM and EDCLD achieve near results. However,

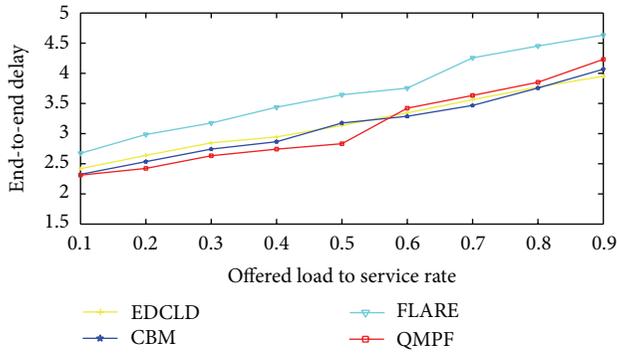


FIGURE 24: End-to-end packet delay.

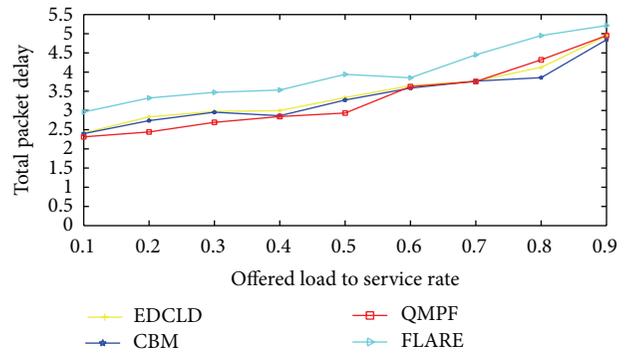


FIGURE 26: Total packet delay.

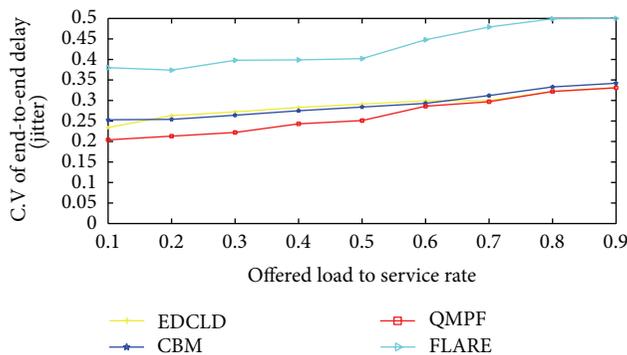


FIGURE 25: Coefficient variation of end-to-end packet delay.

CBM has a smaller end-to-end delay. FLARE achieves the largest end-to-end delay due to load imbalance, especially at the high packet arrival rate.

Packet delay variation is depicted in Figure 25. The relationship between coefficient variation (CV) of end-to-end packet delay and the ratio of offered load to service rate is constructed. A large CV indicates a high risk of packet reordering. In light load QMPF achieves the least delay variation. As the ratio of input rate to output rate increases, CBM and EDCLD outperform QMPF.

The total packet delay is an important indicator for QoS-oriented application. QMPF, CBM, and E-DCLD aim to decrease end-to-end delay and packet reordering delay and can thus efficiently reduce the total packet delay. Figure 26 indicates that when the ratio ( $\lambda/\mu$ ) is larger than 0.6, the total packet delay counted by QMPF is slightly larger than E-DCLD and CBM.

## 10. Conclusion

QoS routing plays an important role in QoS provisioning. This paper introduced a generic distributed QoS adaptive routing engine (DQARE) architecture based on OSPF<sub>x</sub>QoS. DQARE architecture is furnished with three relevant traffic control schemes that shape the design of a QoS solution, namely, service differentiation, QoS routing, and QoS intradomain traffic engineering (TE).

Accordingly, this paper provided a general configuration guideline for service differentiation. Also, this paper introduced a QoS routing algorithm (QOPRA), based on dynamic programming technique. QOPRA attempts to obtain the optimal multiple paths in terms of two additive metrics. This objective is proved in the proposed work with minimal complexity and low error decision rate. This paper also proposed a new effective QoS *load distribution* model (QMPF). QMPF aimed to efficiently utilize multiple available paths and minimize the difference among end-to-end delays, jitter, and packet reordering. NS2-based simulations proved DQARE superiority over OSPF<sub>x</sub>QoS.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

- [1] Z. L. Sun, "IP networking and future evolution," in *Network Performance Engineering*, pp. 951–978, Springer, Berlin, Germany, 2011.
- [2] D. Qiu, "On the QoS of IPTV and its effects on home networks," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, Article ID 253495, 5 pages, 2010.
- [3] X. Xiao, *Technical, Commercial and Regulatory Challenges of QoS: An Internet Service Model Perspective*, Morgan Kaufmann, San Francisco, Calif, USA, 2008.
- [4] J. Baraković, H. Bajrić, M. Kos, S. Baraković, and A. Husić, "Prioritizing signaling information transmission in next generation networks," *Journal of Computer Networks and Communications*, vol. 2011, Article ID 470264, 10 pages, 2011.
- [5] Cisco System, "Routing basics," in *Internetworking Technology*, chapter 5, 4th edition, 2003.
- [6] J. Moy, "OSPF version 2," Internet Request for Comments RFC 1247, 1991.
- [7] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, "A framework for QoS—based routing in the internet," IETF RFC 2386, 1998.
- [8] G. Apostolopoulos, D. Williams, S. Kamat, R. Guerin, A. Orda, and T. Przygienda, "QoS routing mechanisms and OSPF extensions," RFC 2676, 1999.

- [9] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation protocol (RSVP)—version 1 functional specification," IETF RFC 2205, 1997.
- [10] W. C. Hardy, *QoS Measurement and Evaluation of Telecommunications Quality of Service*, John Wiley & Sons, Chichester, UK, 2001.
- [11] R. Wójcik and A. Jajszczyk, "Flow oriented approaches to QoS assurance," *ACM Computing Surveys*, vol. 44, no. 1, article 5, 2012.
- [12] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: an overview," Request for Comments (Informational) RFC 1633, Internet Engineering Task Force, June 1994.
- [13] B. Nandy, N. Seddigh, A. Chapman, and J. H. Salim, "A connectionless approach to providing QoS in IP networks," in *Proceedings of the 8th IFIP Conference on High Performance Networking (HPN '98)*, 1998.
- [14] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF RFC 2475, 1998.
- [15] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers," Request for Comments 2474, Internet Engineering Task Force, 1998.
- [16] S. Giordano, S. Salsano, S. van den Berghe, G. Ventre, and D. Giannakopoulos, "Advances QoS provisioning in IP networks: the European premium IP projects," *IEEE Communications Magazine*, vol. 41, no. 1, pp. 30–36, 2003.
- [17] J. Li and C. Mao, "Providing flow-based proportional differentiated services in class-based DiffServ routers," *IEE Proceedings: Communications*, vol. 151, no. 1, pp. 82–88, 2004.
- [18] M. Montanez, "Deploying QoS in the enterprise," *Packet-Cisco System Users Magazine*, vol. 14, no. 4, pp. 30–34, 2002.
- [19] K. Nguyen and B. Jaumard, "Routing engine architecture for next generation routers: evolutionary trends," *International Journal of Network Protocols and Algorithms*, vol. 1, no. 1, pp. 62–85, 2009.
- [20] H. J. Chao, "Next generation routers," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1518–1558, 2002.
- [21] I. Marsic, *Computer Networks Performance and Quality of Service*, 2013, [http://www.ece.rutgers.edu/~marsic/books/CN/book-CN\\_marsic.pdf](http://www.ece.rutgers.edu/~marsic/books/CN/book-CN_marsic.pdf).
- [22] S. Asthana, C. Delph, H. V. Jagadish, and P. Krzyzanowski, "Towards a gigabit IP router," *Journal of High Speed Networks*, vol. 1, no. 4, pp. 281–288, 1992.
- [23] C. Partridge, P. P. Carvey, E. Burgess et al., "A 50-Gb/s IP router," *IEEE/ACM Transactions on Networking*, vol. 6, no. 3, pp. 237–248, 1998.
- [24] P. Almquist, "Type of Service in the Internet Protocol Suite," 1992.
- [25] J. Babiarz and F. Baker, "Configuration Guidelines for DiffServ Service Classes," RFC 4594, 2006.
- [26] D. Shin, E. K. P. Chong, and H. J. Siegel, "Multi-postpath-based lookahead multiconstraint QoS routing," *Journal of the Franklin Institute*, vol. 349, no. 3, pp. 1106–1124, 2012.
- [27] G. Liu and K. G. Ramakrishnan, "A\*prune: an algorithm for finding K shortest paths subject to multiple constraints," in *Proceedings of the IEEE INFOCOM*, Anchorage, Alaska, April 2001.
- [28] C. Casetti, R. lo Cigno, M. Mellia, and M. Munafò, "A new class of QoS routing strategies based on network graph reduction," in *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '02)*, vol. 2, pp. 715–722, June 2002.
- [29] Z. Yu, F. Ma, J. Liu, B. Hu, and Z. Zhang, "An efficient approximate algorithm for disjoint QoS routing," *Mathematical Problems in Engineering*, vol. 2013, Article ID 489149, 9 pages, 2013.
- [30] Z. Wang and J. Crowcroft, "Quality-of-service routing for supporting multimedia applications," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 7, pp. 1228–1234, 1996.
- [31] F. A. Kuipers and P. van Mieghem, "The impact of correlated link weights on QoS routing," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, pp. 1425–1434, San Francisco, Calif, USA, April 2003.
- [32] J. M. Jaffe, "Algorithms for finding paths with multiple constraints," *Networks*, vol. 14, no. 1, pp. 95–116, 1984.
- [33] H. de Neve and P. van Mieghem, "TAMCRA: a tunable accuracy multiple constraints routing algorithm," *Computer Communications*, vol. 23, no. 7, pp. 667–679, 2000.
- [34] T. Korkmaz and M. M. Krunz, "Routing multimedia traffic with QoS guarantees," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 429–443, 2003.
- [35] A. Bellabas, S. Lahoud, and M. Molnár, "Performance evaluation of efficient solutions for the QoS unicast routing," *Journal of Networks*, vol. 7, no. 1, pp. 73–80, 2012.
- [36] G. Xue, A. Sen, and R. Banka, "Routing with many additive QoS constraints," in *Proceedings of the IEEE International Conference on Communications (ICC '03)*, pp. 223–227, Anchorage, Alaska, USA, May 2003.
- [37] G. Xue and S. K. Makki, "Multiconstrained QoS routing: a norm approach," *IEEE Transactions on Computers*, vol. 56, no. 6, pp. 859–863, 2007.
- [38] G. Xue, W. Zhang, J. Tang, and K. Thulasiraman, "Polynomial time approximation algorithms for multi-constrained QoS routing," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 656–669, 2008.
- [39] P. Van Mieghem and F. A. Kuipers, "Concepts of exact QoS routing algorithms," *IEEE/ACM Transactions on Networking*, vol. 12, no. 5, pp. 851–864, 2004.
- [40] Y. Li, J. Harms, and R. Holte, "Fast exact multiconstraint shortest path algorithms," in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 123–130, June 2007.
- [41] L. Sun, L. Wang, and R. Wang, "Ant colony algorithm for solving QoS routing problem," *Wuhan University Journal of Natural Sciences*, vol. 9, no. 4, pp. 449–453, 2004.
- [42] Y. H. S. Wan, Y. Hao, and Y. Yang, "Approach for multiple constraints based QoS routing problem of network," in *Proceedings of the 9th International Conference on Hybrid Intelligent Systems (HIS '09)*, vol. 2, pp. 66–69, IEEE, Shenyang, China, August 2009.
- [43] J. Y. Yen, "Finding the K shortest loopless paths in a network," *Management Science*, vol. 17, pp. 712–716, 1970/7119.
- [44] E. L. Lawler, "A procedure for computing the K best solutions to discrete optimization problems and its application to the shortest path problem," *Management Science*, vol. 18, pp. 401–405, 1972.
- [45] T. Sanguankotchakorn, S. Maneepong, and N. Sugino, "A relaxing multi-constraint routing algorithm by considering QoS metrics priority for wired network," in *Proceedings of the 5th*

- International Conference on Ubiquitous and Future Networks (ICUFN '13)*, pp. 738–743, Da Nang, Vietnam, July 2013.
- [46] R. Bellman, “On a routing problem,” *Quarterly of Applied Mathematics*, vol. 16, no. 1, pp. 87–90, 1958.
- [47] G. Xue, A. Sen, W. Zhang, J. Tang, and K. Thulasiraman, “Finding a path subject to many additive QoS constraints,” *IEEE/ACM Transactions on Networking*, vol. 15, no. 1, pp. 201–211, 2007.
- [48] X. Jin, “Routing for multi-constrained path problem with imprecise additive link state information,” in *Proceedings of the International Conference on Computer Application and System Modeling (ICCASM '10)*, vol. 3, pp. 472–475, Taiyuan, China, October 2010.
- [49] J. Huang, X. Huang, and Y. Ma, “Routing with multiple quality-of-services constraints: an approximation perspective,” *Journal of Network and Computer Applications*, vol. 35, no. 1, pp. 469–479, 2012.
- [50] R. Hou, K. Luib, K. Leung, and F. Baker, “Performance analysis of quantization-based approximation algorithms for precomputing the supported QoS,” *Journal of Network and Computer Applications*, vol. 40, pp. 244–254, 2014.
- [51] N. Wang, K. H. Ho, G. Pavlou, and M. Howarth, “An overview of routing optimization for internet traffic engineering,” *IEEE Communications Surveys and Tutorials*, vol. 10, no. 1, pp. 33–56, 2008.
- [52] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, “The end-to-end effects of Internet path selection,” in *Proceedings of the ACM SIGCOMM*, August 1999.
- [53] B. Fortz and M. Thorup, “Internet traffic engineering by optimizing OSPF weights,” in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, pp. 519–528, Tel Aviv, Israel, March 2000.
- [54] B. Fortz and M. Thorup, “Optimizing OSPF/IS-IS weights in a changing world,” *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 4, pp. 756–767, 2002.
- [55] B. Fortz, J. Rexford, and M. Thorup, “Traffic engineering with traditional IP routing protocols,” *IEEE Communications Magazine*, vol. 40, no. 10, pp. 118–124, 2002.
- [56] A. K. Parekh and R. G. Gallager, “Generalized processor sharing approach to flow control in integrated services networks: the single-node case,” *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 344–357, 1993.
- [57] M. Lengyel, J. Sztrik, and C. S. Kim, “Simulation of differentiated services in network simulator,” in *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae, Sectio Computatorica*, 2003.
- [58] K. C. Leung and V. O. K. Li, “Generalized load sharing for packet-switching networks I: theory and packet-based algorithm,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 7, pp. 694–702, 2006.
- [59] D. Thaler and C. Hopps, “Multipath issues in unicast and multicast next-hop selection,” RFC 2991, 2000.
- [60] C. Hopps, “Analysis of an equal-cost multi-path algorithm,” RFC 2992, 2000.
- [61] A. Zinin, *Cisco IP Routing, Packet Forwarding and Intradomain Routing Protocols*, Addison-Wesley, Reading, Mass, USA, 2002.
- [62] T. W. Chim and K. L. Yeung, “Traffic distribution over equal-cost-multi-paths,” in *Proceedings of the IEEE International Conference on Communications*, vol. 2, pp. 1207–1211, June 2004.
- [63] S. Kandula, D. Katabi, S. Sinha, and A. Berger, “Dynamic load balancing without packet reordering,” *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 2, pp. 53–62, 2007.
- [64] M. Tian, J. Lan, X. Zhu, and J. Huang, “A routing optimization algorithm of equal-cost-multi-paths based on link criticality,” in *Proceedings of the IEEE International Conference on Advanced Computer Control (ICACC '10)*, pp. 203–207, March 2010.
- [65] Y. Wang and Z. Wang, “Explicit routing algorithms for internet traffic engineering,” in *Proceedings of the 8th International Conference on Computer Communications and Networks (ICCN '99)*, pp. 582–588, IEEE, Boston, Mass, USA, October 1999.
- [66] R. Banner and A. Orda, “Multipath routing algorithms for congestion minimization,” *IEEE/ACM Transactions on Networking*, vol. 15, no. 2, pp. 413–424, 2007.
- [67] S. Prabhavat, H. Nishiyama, N. Ansari, and N. Kato, “Effective delay-controlled load distribution over multipath networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 10, pp. 1730–1741, 2011.
- [68] H. Adishesu, G. Parulkar, and G. Varghese, “A reliable and scalable striping protocol,” *ACM SIGCOMM Computer Communication Review*, vol. 26, no. 4, pp. 131–141, 1996.
- [69] M. Li, H. Nishiyama, N. Kato, K. Mizutani, O. Akashi, and A. Takahara, “On the fast-convergence of delay-based load balancing over multipaths for dynamic traffic environments,” in *Proceedings of the International Conference on Wireless Communications and Signal Processing (WCSP '13)*, Hangzhou, China, October 2013.
- [70] R. Martin, M. Menth, and M. Hemmkeppeler, “Accuracy and dynamics of multi-stage load balancing for multipath Internet routing,” in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 6311–6318, June 2007.
- [71] W. Shi, M. H. MacGregor, and P. Gburzynski, “Load balancing for parallel forwarding,” *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 790–801, 2005.
- [72] G. Rétvári, F. Németh, R. Chaparadza, and R. Szabó, “OSPF for implementing self-adaptive routing in autonomic networks: a case study,” in *Modelling Autonomic Communications Environments: Proceedings of the Fourth IEEE International Workshop, MACE 2009, Venice, Italy, October 26-27, 2009*, vol. 5844 of *Lecture Notes in Computer Science*, pp. 72–85, Springer, Berlin, Germany, 2009.
- [73] C. Dovrolis, D. Stiliadis, and P. Ramanathan, “Proportional differentiated services: delay differentiation and packet scheduling,” *IEEE/ACM Transactions on Networking*, vol. 10, no. 1, pp. 12–26, 2002.
- [74] M. Shreedhar and G. Varghese, “Efficient fair queueing using deficit round robin,” *ACM SIGCOMM Computer Communication Review*, vol. 25, no. 4, pp. 231–242, 1995.
- [75] Y. Chen, T. Farley, and N. Ye, “QoS requirements of network applications on the internet,” *Information Knowledge Systems Management*, vol. 4, pp. 55–76, 2004.
- [76] J. Postel, “Internet Protocol,” RFC 791 (Standard), Internet Engineering Task Force, September 1981, updated by RFC 1349, <http://www.ietf.org/rfc/rfc791.txt>.
- [77] T. H. Cormen, C. E. Leiserson, and R. Rivest, *An Introduction to Algorithms*, MIT Press, Boston, Mass, USA, 2009.
- [78] R. W. Floyd, “Algorithm 97: shortest path,” *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.
- [79] K. A. Rink, E. Y. Rodin, and V. Sundarapandian, “A simplification of the double-sweep algorithm to solve the k-shortest path problem,” *Applied Mathematics Letters*, vol. 13, no. 8, pp. 77–85, 2000.
- [80] H. Saito, C. Lukovszki, and I. Moldován, “Local optimal proportional differentiated services scheduler for relative differentiated

- services,” in *Proceedings of the 9th International Conference on Computer Communications and Networks (ICCCN '00)*, pp. 554–550, Las Vegas, Nev, USA, 2000.
- [81] T. Nandagopal, N. Venkitaraman, R. Sivakumar, and V. Barghavan, “Delay differentiation and adaptation in core stateless networks,” in *Proceedings of the 9th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, pp. 421–430, Tel-Aviv, Israel, April 2000.
- [82] Y. Moret and S. Fdida, “A proportional queue control mechanism to provide differentiated services,” in *Proceedings of the 13th International Symposium On Computer and Information Sciences (ISCIS '98)*, pp. 17–24, Belek, Turkey, October 1998.
- [83] N. Christin, J. Liebeherr, and T. Abdelzaher, “Enhancing class-based service architectures with adaptive rate allocation and dropping mechanisms,” *IEEE/ACM Transactions on Networking*, vol. 15, no. 3, pp. 669–682, 2007.
- [84] R. L. Cruz, “A calculus for network delay. I. Network elements in isolation,” *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 114–131, 1991.
- [85] Network Simulator NS2, <http://www.isi.edu/nsnam/ns/>.
- [86] B. M. Waxman, “Routing of multipoint connections,” *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.

## Research Article

# Precoding Method Interference Management for Quasi-EVD Channel

Wei Duan,<sup>1</sup> Wei Song,<sup>2</sup> Sang Seob Song,<sup>1</sup> and Moon Ho Lee<sup>1</sup>

<sup>1</sup> Division of Electronic and Information Engineering, Chonbuk National University, Chonju 561-756, Republic of Korea

<sup>2</sup> College of Information Technology, Eastern Liaoning University, Dandong, 118003, China

Correspondence should be addressed to Moon Ho Lee; moonho@jbnu.ac.kr

Received 12 March 2014; Revised 5 July 2014; Accepted 10 July 2014; Published 28 August 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Wei Duan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Cholesky decomposition-block diagonalization (CD-BD) interference alignment (IA) for a multiuser multiple input multiple output (MU-MIMO) relay system is proposed, which designs precoders for the multiple access channel (MAC) by employing the singular value decomposition (SVD) as well as the mean square error (MSE) detector for the broadcast Hermitian channel (BHC) taken advantage of in our design. Also, in our proposed CD-BD IA algorithm, the relaying function is made use to restructure the quasideigenvalue decomposition (quasi-EVD) equivalent channel. This approach used for the design of BD precoding matrix can significantly reduce the computational complexity and proposed algorithm can address several optimization criteria, which is achieved by designing the precoding matrices in two steps. In the first step, we use Cholesky decomposition to maximize the sum-of-rate (SR) with the minimum mean square error (MMSE) detection. In the next step, we optimize the system BER performance with the overlap of the row spaces spanned by the effective channel matrices of different users. By iterating the closed form of the solution, we are able not only to maximize the achievable sum-of-rate (ASR), but also to minimize the BER performance at a high signal-to-noise ratio (SNR) region.

## 1. Introduction

Recently, wireless relay networks which are capable of improving the power efficiency, as well as the network coverage, have been studied with a lot of interest because relaying transmission is a promising technique which can be applied to extend the coverage or increase the system capacity. The capacity achieved by a point-to-point MIMO network has been shown to increase linearly with the minimum number of transceiver's antennas [1, 2]. Therefore, by employing multiple antennas at the transmitter or the receiver, the system can significantly improve the transmission reliability.

If multiple antennas are applied at both the transmitter and receiver sides, the channel capacity can be enhanced linearly with the minimum number of transmit and receive antennas [3].

Relay precoder designs for such a system have been reported in [4–6]. The problem of designing optimal beamforming vectors for multicasting is hard in general, mainly due to its nonconvex nature. In [4], the authors propose

a transceive precoding scheme at the relay node by using zero-forcing (ZF) and MMSE criteria with certain antenna configurations. The information theoretic capacity of the multiantenna multicasting channel is studied in [5] with a particular focus on the scaling of the capacity and achievable rates as the number of antennas and users approaches infinity. In [6], the authors develop one algorithm to compute the globally optimal beamforming matrix at the relay node and characterize the system capacity region.

Most of the works mentioned above assume the availability of perfect channel state information (CSI) at the relay node [7, 8]. In practice, the CSI available at the relay node is usually imperfect due to different factors such as estimation error, quantization, and feedback delay. Interference alignment (IA) is proposed to achieve the maximum degree of freedom (DOF) for the  $K$ -user interference channels [9]. It designs the signals transmitted by all users with perfect CSI in such a way that the interfering signals at each receiver fall into a reduced-dimensional subspace. In order to implement IA scheme in the slow fading environment, multiple channels can be

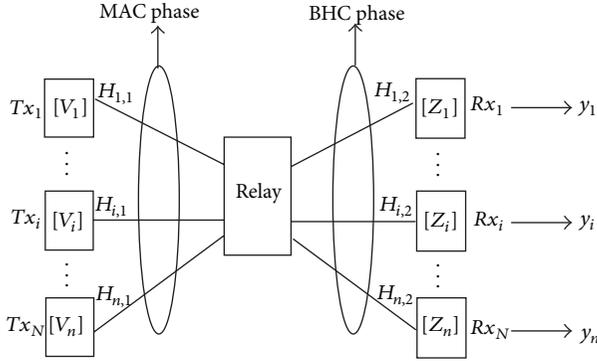


FIGURE 1:  $K$ -pairs single relay-aided interference alignment system.

used for multiple carriers or multiple antennas [10]. Since these resources are limited, IA scheme with time extension is still efficient to support multiple users. In the past decades, researches on information theory have been exploring the capacity regions of Gaussian interference channels [11, 12]. In the  $K$ -user interference channel, it is proved that the IA scheme can provide the following capacity for each user:

$$C_{\text{IA}} = \frac{K}{2} \log(\text{SNR}) + o(\log(\text{SNR})). \quad (1)$$

Thus, in high-SNR regime, the capacity scales linearly with the number of users.

In this paper, we consider the problem of jointly designing the precoders and the relay transformation matrix for a one-way relay MIMO relay system, where all nodes have multiple antennas. Our goal is to use BHC and BD precoding design to decouple MU-MIMO channel into a set of  $K$  parallel independent SU-MIMO channels and CD-BD algorithm to reduce the computational complexity. In particular, the leakage interference is minimized in order to achieve interference alignment. By iterating the closed-form solution and precoding design, we reach the maximum sum-of-rate capacity and better performance in BER as shown in simulations.

The organization of the paper is as follows: Section 2 describes a general system model for the  $K$ -pairs one-way relay system, the definition of quasi-EVD, and global CSI. In Section 3, we propose an iterative CD-BD algorithm and optimal precoder design. In Sections 4 and 5, we discuss the ASR, DOE, and computational complexity for efficient channel model. The simulation results are presented to show the good performance of the proposed algorithm for the  $K$ -pairs relay-aided system in Section 6, and Section 7 concludes the paper.

*Notation.* For matrix  $A$ ,  $\text{tr}(A)$ ,  $\text{rank}(A)$ ,  $|A|$ ,  $A^T$ ,  $A^H$ , and  $A^{-1}$  denote the trace, rank, determinate, transpose, conjugate transpose, and inverse of  $A$ , respectively.  $\mathbb{C}^{x \times y}$  and  $\mathbb{R}^{x \times y}$  denote the space of  $x \times y$  matrices with complex and real entries.  $E(\cdot)$  stands for the expectation and  $D(A) = \text{diag}(a_1, \dots, a_n)$  are the diagonal matrix whose elements on the diagonal are  $a_1, \dots, a_n$ .

## 2. System Model

In this section, we propose the one-way relay system, whose key idea to structure the quasi-EVD channel is using the relay function to cancel the unitary matrices of multiple access channel (MAC) and broadcast hermitian channel (BHC).

*2.1. Protocol Description.* Consider  $K$ -pairs interference single relay-aided system that proceeds in two phases, which are multiple access channels (MAC) and broadcast hermitian channel (BHC) as shown in Figure 1, where transmitter  $Tx_i$  and receiver  $Rx_i$  are equipped with  $M$  antennas, and the relay node has  $NK$  antennas. The channel coefficients  $H_{i,1} \in \mathbb{C}^{N \times M}$  and  $H_{i,2} \in \mathbb{C}^{M \times N}$  define links from the source  $i$  to relay and relay to the destination  $i$ , where  $i = 1, 2, \dots, K$  and  $M \leq N$  (decodable condition). The received signal at relay in the MAC phase is given by

$$r_i = H_{i,1}s_i + \sum_{j \neq i}^K H_{j,1}s_j + n_{i,1}, \quad (2)$$

where  $n_{i,1} \sim \text{CN}(0, \sigma_{i,1}^2 I_N)$  represents the additive white Gaussian noise (AWGN) vector with zero mean and variance  $\sigma_{i,1}^2$ . The transmitted signal form  $Tx_i$  to relay is obtained by the precoding matrix  $V_i \in \mathbb{C}^{M \times M}$ ; that is,  $s_i = V_i x_i$  for  $i = 1, 2, \dots, K$ , where  $x_i = [a_1 \cdots a_i \cdots a_m]^T$  is the transmitted signals form user  $i$  and  $a_i$  is date stream. The proposed precoder  $V_i$  can be obtained in two steps as follows:  $V_i = V_i^a V_i^b$ , which will be further discussed in Section 3. The term  $s_i \in \mathbb{C}^{M \times 1}$  is subject to a power constraint,  $\text{tr}\{E(s_i s_i^H)\} \leq P_i$  with  $E(x_i x_i^H) \leq (P_i/M)I_M$ , where  $P_i$  is the transmit power at  $Tx_i$ .

In the BHC phase, relay sends  $s_r \in \mathbb{C}^{N \times 1}$  which is combined with the linear precoding matrix  $W_i \in \mathbb{C}^{N \times N}$ , to  $Rx_i$  as follows:

$$s_r = W_i r_i, \quad (3)$$

where the relay precoding matrix  $W_i$  is subset of relay filter  $W$ . We assume that the maximum transmission power at relay node is  $P_r$ ; that is,

$$\text{tr} \left\{ W \left( \sum_{i=1}^K H_{i,1} V_{i,1} V_{i,1}^H H_{i,1}^H + \sigma_{i,1}^2 I_N \right) W^H \right\} \leq P_r, \quad (4)$$

where we have used the assumption that the source signals and the relay noise are independent with each other. Then, the relay broadcasts  $s_r$  to the destination nodes and the received signals at  $Rx_i$  can be written as

$$y_i = H_{i,2} s_r + n_{i,2}, \quad (5)$$

where  $n_{i,2}$  denotes the additive noise vector at  $Rx_i$  with  $n_{i,1} \sim \text{CN}(0, \sigma_{i,2}^2 I_M)$ . Due to the received signal given by (5), the destination can detect the message by the MMSE criterion or

$$\varepsilon_i = \arg \min E \left\{ \|Z_i^H y_i - x_i\|^2 \right\}, \quad (6)$$

where  $Z_i$  is an  $M \times M$  linear decode matrix at  $Rx_i$ .

2.2. *Quasi-EVD and Global CSIT.* We assume that the global channel state information (CSI) and the designed precoding matrices are perfectly known at all the nodes; thus, the channel coefficient can be denoted as SVD decomposition or Hermitian of SVD. In our proposed system, the channel matrices may be defined as follows:

- (a) MAC phase:  $H_{i,1} = U_{i,1}^H \Sigma_{i,1} \Lambda_{i,1}$ ,
- (b) BHC phase:  $H_{i,2} = \Lambda_{i,2}^H \Sigma_{i,2}^H U_{i,2}$ ,

where  $(U_{i,1}, U_{i,2}) \in \mathbb{C}^{N \times N}$  and  $(\Lambda_{i,1}, \Lambda_{i,2}) \in \mathbb{C}^{M \times M}$  are unitary matrices.  $\Sigma_{i,1} = [\text{diag}(\lambda_{1,1}, \dots, \lambda_{m,1}) \ 0_{(N-M) \times M}]^T \in \mathbb{C}^{N \times M}$  and  $\Sigma_{i,2} = [\text{diag}(\lambda_{1,2}, \dots, \lambda_{m,2}) \ 0_{(N-M) \times M}]^T \in \mathbb{C}^{N \times M}$  are eigen value matrices, where  $\lambda_{i,1}$  is the element of eigenvalues.

In addition, we propose the channel gain matrix which has its singular value matrix in its middle as well as its eigen matrix and unitary matrix in its right or left side appropriately, which results in the new diagonal matrix. This kind of structure is called quasi-EVD. Firstly, we show a result which is helpful to define the quasi-EVD equivalent channel as follows:

$$\begin{aligned} \Sigma_{i,2}^H \cdot \Sigma_{i,1} &= \text{diag}(\lambda_{1,2}^* \cdot \lambda_{1,1}, \dots, \lambda_{m,2}^* \cdot \lambda_{m,1}) \\ &= \text{diag}(\lambda_{1,2,1}, \dots, \lambda_{1,2,m}) \\ &= \Sigma_{i,i}^2, \end{aligned} \quad (7)$$

where  $\lambda_{a,b,i} = \lambda_{b,i}^* \cdot \lambda_{a,i}$ .

First, we proceed by reviewing the feasibility conditions of interference alignment and cancellation. Next, we turn to structure of the quasi-EVD diagonal channel and the problem of the optimization of the precoders and MSE detectors.

### 3. Optimal Filters Design and CD-BD Algorithm

3.1. *Interference Alignment and Cancellation.* As shown in [15], the IA scheme is a linear precoding technique to align interference in reduced dimensional signal subspace at each receiver. The feasibility conditions for MIMO interference channel (IC) consist of the one interference-free constraint and a signal space rank constraint. The perfect IA requirements for all  $k$  are

$$U_j^H H_j V_j = 0, \quad \forall j \neq n, \quad (8a)$$

$$\text{rank}(U_i^H H_i V_i) = d_i, \quad \forall i \in \{1, 2, \dots, K\}. \quad (8b)$$

An efficient distributed algorithm to find matrices  $U_j$  and  $V_j$  are derived in [16] by using the channel reciprocity. The condition (8a) guarantees that all the interfering signals at destination  $j \in K$  are aligned in a subspace of  $N_k - d_i$  dimensions and can be zero-forced by  $U_j$ . Condition (8b) guarantees that destination  $Rx_i$  is able to decode all  $d_i$  intended data streams successfully. If conditions (8a) and (8b) are satisfied, then the effective channel is free from interference; the structure is feasible for the given DOF  $d_i$ .

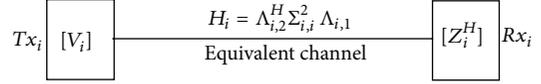


FIGURE 2: Equivalent quasi-EVD channel for relay-aided system.

3.2. *Effective Equivalent Diagonal Channel.* Due to the SVD of channel in Section 2, the equivalent channel for the total system can be described as

$$\begin{aligned} H_i &= H_{i,2} W_i H_{i,1} \\ &= \Lambda_{i,2}^H \Sigma_{i,2}^H U_{i,2} W_i U_{i,1}^H \Sigma_{i,1} \Lambda_{i,1}, \end{aligned} \quad (9)$$

where  $W_i \in \mathbb{C}^{N \times N}$  is the relay precoding matrix. To eliminate the quasi-EVD channel, we adopt the relay precoding matrix defined as

$$W_i = U_{i,2}^H \times U_{i,1}. \quad (10)$$

If  $V_i$  has full rank,  $U_i^H$  are also with full rank. It implies that both pseudoinverses of  $V_i$  and  $U_i^H$  exist. In order to get the optimal leakage interference, the relay filter should satisfy the constraint

$$W_i^H W_i = I_N. \quad (11)$$

By substituting (10) into (11), the above-mentioned equation can be written as

$$\begin{aligned} W_i^H W_i &= (U_{i,2}^H \times U_{i,1})^H (U_{i,2}^H \times U_{i,1}) \\ &= I_N. \end{aligned} \quad (12)$$

Obviously, the relay function  $W_i$  results in optimal leakage interference condition. In order to achieve the optimal leakage interference, it should satisfy the constraint as follows:

$$\min(W_i^H A_i W_i) = 0, \quad (13)$$

where  $A_i = Z_i^H P_r Z_i$ ,  $P_r$  is the relay power constraint shown in (4). Therefore, when interference alignment is feasible, the objective function in (13) can be minimized. By using relay function  $W_i$  and (7), we may structure a quasi-EVD channel as

$$\begin{aligned} H_i &= \Lambda_{i,2}^H \Sigma_{i,2}^H \Sigma_{i,1} \Lambda_{i,1} \\ &= \Lambda_{i,2}^H \Sigma_{i,i}^2 \Lambda_{i,1}. \end{aligned} \quad (14)$$

Subsequently, this efficient channel for the pair of user  $i$  in total system can be shown in Figure 2.

Therefore,  $\text{span}(Z_i^H H_i V_i)$  constitutes the useful signal space in which it is expected to observe all symbols transmitted by user  $i$ , while  $\text{span}(Z_j^H H_j V_j)_{j \neq i}$  is the space where all interference is observed. In addition, to make the leakage interference zero, the relaying function can be inserted at the relay.

The total interference leakage at the destination is given by [17]

$$\Omega_{i,2} = \text{tr} \{ Z_i^H P_r Z_i \}, \quad (15)$$

where  $P_r$  is the power constraint shown in (4). Based on equivalent channel, (15) can be rewritten as

$$\begin{aligned} \Omega_{i,1} &= \text{tr} \{ V_i^H \bar{P}_i V_i \}, \\ \bar{P}_i &= \sum_{i=1}^K \text{tr} \left\{ \frac{1}{d_i} V_i^H H_i^H Z_i Z_i^H H_i V_i \right\}. \end{aligned} \quad (16)$$

For the perfect interference alignment, the leakage interference should be zero, which means that  $\Omega_{i,1} = \Omega_{i,2} = 0$ . This equation is equivalent to the zero-forcing at  $Rx_i$  which is elegantly employed to achieve a good performance in the proposed scheme. The channel state information is perfectly known at every node; the optimization problem in (15) can be written as

$$\begin{aligned} \min_{V_i, Z_i^H} \quad & E \left\{ \left\| Z_i^H y_i - x_i \right\|^2 \right\} \\ \text{s.t.} \quad & \text{tr} \left\{ W \left( \sum_{i=1}^K H_{i,1} V_i V_i^H H_{i,1}^H + \sigma_{i,1}^2 I_N \right) W^H \right\} \leq P_r, \end{aligned} \quad (17)$$

where  $P_r$  is the transmit power at relay. It shows that the optimization problem contains only  $V_i$  and  $Z_i$ ; we will further discuss details in next section.

**3.3. Global Optimal Precoder and Detector Design.** The proposed optimal precoder design involves two steps, that is, MMSE detector design at destination and optimal precoding design at transmitter. It contains two phases as follows.

**3.3.1. MMSE Detector Design.** For the above-mentioned parameters, the sum of leakage interference can be reshaped as

$$\sum_{i=1}^K \Omega_r = \sum_{i=1}^K \text{tr} \{ Z_i^H P_r Z_i \}, \quad (18)$$

and it may be given as follows by denoting that  $Q_i = H_{i,2} W_i H_{i,1} V_i$ :

$$Z_i^{\text{opt}} = Q_i^H \left( Q_i Q_i^H + \sigma_{i,1}^2 \Sigma_{i,2}^2 + \sigma_{i,2}^2 I_M \right)^{-1}, \quad i = 1, \dots, K \quad (19)$$

which is the optimal MMSE decoder design proved in Appendix A. Therefore, the minimum  $\Omega_i$  is equivalent to sum of  $d_i$  least dominant eigenvalues of  $P_i$ .

**3.3.2. Optimal Precoding Design and Iterative Algorithm.** Based on MMSE detector  $Z_i^{\text{opt}}$ , precoding matrices at source nodes should be collaboratively designed. To simply discuss the optimization problem, we assume that the noises are with same variance; that is,  $\sigma_{i,1} = \sigma_{i,2} = \sigma_i$ . By using optimal MSE detector design shown in (19), the MSE matrix of the

signal waveform estimation at receiver can be denoted as  $\varepsilon_i = [(\tilde{x}_i - x_i)(\tilde{x}_i - x_i)^H]$  or

$$\begin{aligned} \min_{V_i} \quad & \varepsilon_i = \text{tr} \left\{ \left[ I_M + \frac{1}{\sigma_i^2} Q_i \Psi_i^{-1} Q_i^H \right]^{-1} \right\} \\ \text{s.t.} \quad & \text{tr} \left\{ W \left( \sum_{i=1}^K H_{i,1} V_i V_i^H H_{i,1}^H + \sigma_{i,1}^2 I_N \right) W^H \right\} \leq P_r, \end{aligned} \quad (20)$$

where  $\Psi_i = I_M + H_{i,2} H_{i,2}^H$ .

**Lemma 1.** *The optimal precoding matrices  $V_i^b$  design is a convex optimization in high-SNR region. For proof see Appendix B.*

By applying the MMSE inversion to the combined channel matrix, we have

$$\begin{aligned} H_{\text{mse}}^\dagger &= H^H (H H^H + \alpha I)^{-1} \\ &= [H_{1,\text{mse}}, H_{2,\text{mse}}, \dots, H_{K,\text{mse}}], \end{aligned} \quad (21)$$

where  $H$  is the combined equivalent channel matrix; that is,  $H = [H_1^T, H_2^T, \dots, H_K^T]^T \in \mathbb{C}^{KM \times KM}$  and  $\alpha$  is the regularization factor. Considering a high-SNR case, it can be shown that  $\alpha$  approaches zero and we have  $H H_{\text{mse}}^\dagger \approx I_{KM}$ . This means the off diagonal block matrices of  $H H_{\text{mse}}^\dagger$  converge to zero with high SNR. In addition, we exclude the  $i$ th pair user's channel matrices and define  $\bar{H}_{i,1}$  and  $\bar{H}_{i,2}$  as

$$\begin{aligned} \bar{H}_{i,1} &= [H_{1,1}^T, \dots, H_{i-1,1}^T, H_{i+1,1}^T, \dots, H_{M,1}^T]^T \\ &\in \mathbb{C}^{(K-1)N \times K(M-1)}, \\ \bar{H}_{i,2} &= [H_{1,2}^T, \dots, H_{i-1,2}^T, H_{i+1,2}^T, \dots, H_{M,2}^T]^T \\ &\in \mathbb{C}^{K(M-1) \times (K-1)N}. \end{aligned} \quad (22)$$

Thus, the equivalent excluded channel may be denoted as

$$\bar{H}_i = \bar{H}_{i,1} \bar{W}_i \bar{H}_{i,2} \in \mathbb{C}^{K(M-1) \times K(M-1)}. \quad (23)$$

Obviously, the matrix  $H_{i,\text{mse}}$  is approximately in the null space of  $\bar{H}_i$  which can be expressed as

$$\bar{H}_i H_{i,\text{mse}} \approx 0. \quad (24)$$

Considering the SVD of  $H_{i,\text{mse}} = U_{i,\text{mse}} \Sigma_{i,\text{mse}} \Lambda_{i,\text{mse}}$ , we have

$$\bar{H}_i H_{i,\text{mse}} = \bar{H}_i U_{i,\text{mse}} \Sigma_{i,\text{mse}} \Lambda_{i,\text{mse}} \approx 0, \quad (25)$$

where  $U_{i,\text{mse}}$  and  $\Lambda_{i,\text{mse}}$  are unitary matrices and  $\Sigma_{i,\text{mse}}$  is eigen value matrix. Since  $U_{i,\text{mse}}$  and  $\Lambda_{i,\text{mse}}$  are invertible, we have

$$\bar{H}_i \Sigma_{i,\text{mse}} \approx 0. \quad (26)$$

Thus,  $\Sigma_{i,\text{mse}}$  satisfies the BD constraint to balance the interference and the noise term. Therefore, the first step precoding design is completed with result  $V_i^a = \Sigma_{i,\text{mse}}$ . On the

(1) Given the channel  $H_{i,1} = [H_{1,1}, H_{2,1}, \dots, H_{K,1}]$  and  $H_{i,2} = [H_{1,2}, H_{2,2}, \dots, H_{K,2}]$  for  $K$ -pair users as in (2), may be decomposed as:

$$\begin{aligned} H_{i,1} &= U_{i,1}^H \Sigma_{i,1} \Lambda_{i,1} \\ H_{i,2} &= \Lambda_{i,2} \Sigma_{i,2}^H U_{i,2} \end{aligned}$$

(2) Fix the relay function  $W_i = U_{i,2}^H \times U_{i,1}$  shown in (10).

(3) Begin iteration.

(4) Applying the MMSE channel inversion:

$$Z_i^{\text{opt}} = Q_i^H (Q_i Q_i^H + \sigma_{i,1}^2 \Sigma_{i,2}^2 + \sigma_{i,2}^2 I_M)^{-1}$$

(5) Compute the Cholesky factorization:

$$L_i^H L_i = I_M + H_{i,2} H_{i,2}^H$$

(6) Compute the precoding matrix:

$$V_i^b = \Lambda_{i,1}^H B_i^{1/2} L_i$$

(7) Compute the MSE matrix of the signal waveform estimation:

$$\epsilon_i = \sigma_i^2 \text{tr} \left[ \Lambda_{i,2} \Sigma_{i,2}^2 \Lambda_{i,2} B_i \right]^{-1}$$

(8) Compute the leakage interference:

$$\begin{aligned} \Omega_{i,1} &= \text{tr} \{ V_i^H \tilde{P}_i V_i \} \\ \tilde{P}_i &= \sum_{i=1}^K \text{tr} \{ (1/d_i) V_i^H H_i^H Z_i Z_i^H H_i V_i \} \end{aligned}$$

(10) Stop iteration until convergence.

ALGORITHM 1: Cholesky decomposition-block diagonalization (CD-BD) algorithm.

other hand, the interference generated to the other users is determined by  $\bar{H}_i V_i^a$ . Thus, the final precoder for user  $i$  may be obtained as

$$V_i = V_i^a V_i^b = \Sigma_{i,\text{mse}} \Lambda_{i,1}^H B_i^{1/2} L_i. \quad (27)$$

After the precoding process, the MU-MIMO channel is decoupled into a set of  $K$  parallel independent SU-MIMO channels by the BD precoding. In order to decode the desired signals at the corresponding receivers, the following constraints should be satisfied [9]:

$$\text{span}(H_{m,n} V_m) = \text{span}(H_{j,n} V_j), \quad \forall m \neq n \neq j, \quad (28)$$

where the precoder  $V_m$  is subject to the signal space. We can optimize the precoder matrix tailored to individual rate. Consequently, the total leakage interference is

$$\sum_{i=1}^K \Omega_{i,1} = \sum_{i=1}^K \text{tr} \{ V_i^H \tilde{P}_i V_i \}. \quad (29)$$

As the variance of noises  $\sigma_{i,1}$  and  $\sigma_{i,2}$  is small enough in the wireless systems, the convexity can be ensured by substituting (10) and (27) into (29). While it is hard to derive a closed-form solution for (29), it can be efficiently solved using the optimal package provided in [18]. Therefore, the minimum  $\Omega_i$  is equal to the sum of the  $d_i$  least dominant eigenvalues of  $\tilde{P}_i$ ; therefore, the optimal precoder and decoder design are completed.

The proposed relay-aided interference alignment algorithm is given in Algorithm 1. By employing the minimization technique, it can iteratively update the coding vectors at transmitters, the zero-forcing vectors at receivers, and relaying function at relay to minimize the total leakage interference.

#### 4. Performance Analysis

In this section, we carry out an analysis of the performance of proposed system. We consider a performance analysis in terms of BER, achievable sum of rate (ASR).

For the RBD precoding [13], the residual interference  $\bar{H}_i V_i^{a(\text{RBD})}$  is not zero between the users which is the solution in high-SNR region shown as follows:

$$\left( \bar{H}_i V_i^{a(\text{RBD})} \right) \left( \bar{H}_i V_i^{a(\text{RBD})} \right)^H \approx I_M. \quad (30)$$

By comparing (26) and (30), we can see that the impact of our proposed precoding would be smaller than that of the conventional RBD precoding algorithm.

Assuming that there exist intersections between desired signal channel and interference signal channel, the following equation will be satisfied:

$$\begin{bmatrix} I_M & -H_{i,1} & 0 \\ I_M & 0 & -H_{j,1} \end{bmatrix} \begin{bmatrix} x_i \\ V_i \\ V_j \end{bmatrix} = 0, \quad (31)$$

where  $x_i$  is the transmitted signals from user  $i$ . After spanned interference signals into one dimension, we can full cancel them [19]. Therefore, the observations at the relay in (2) can yield

$$r_i = H_{i,1} V_i x_i + n_{i,1}, \quad (32)$$

where  $H_{i,1} V_i$  denote column vector of total effective MAC channel matrix with size  $M \times M$ . Consequently, after the relay filter  $W$ , the effective propagation of total system is structured and the observations of user  $i$  for MMSE precoding under the high-SNR scenario can be obtained as

$$y_i = s_i + \sqrt{\eta_1} n_{i,1} + \sqrt{\eta_2} n_{i,2}. \quad (33)$$

Consequently, the factor that  $H_{i,1}V_i^a = U_{i,1}^a \Sigma_{i,1}^a \Lambda_{i,1}^a$  with rank  $\aleph$  and  $H_{i,i}V_i^a = U_{i,i}^a \Sigma_{i,i}^a \Lambda_{i,i}^a$  with rank  $\Gamma$ , it is simple that the normalization factors  $\eta_\varphi$  and  $\eta_\tau$  can be written as

$$\begin{aligned}\eta_\varphi &= \|(H_{i,1}V_i^a)^{-1} s_i\|_F^2 = \text{tr} \left( (\Sigma_{i,1}^a)^{-2} s_i s_i^H \right) \\ &= \sum_{\varphi=1}^{\aleph} \frac{P_\varphi^2}{(\lambda_\varphi^a)^2}, \\ \eta_\tau &= \|(H_{i,i}V_i^a)^{-1} s_i\|_F^2 = \text{tr} \left( (\Sigma_{i,i}^a)^{-2} s_i s_i^H \right) \\ &= \sum_{\tau=1}^{\Gamma} \frac{P_\tau^2}{(\lambda_\tau^a)^2},\end{aligned}\quad (34)$$

where the quantity  $\lambda_\varphi^a$ ,  $\lambda_\tau^a$ ,  $P_\varphi^2$ , and  $P_\tau^2$  are the  $\varphi$ th singular value of  $\Sigma_{i,1}^a$ ,  $\tau$ th singular value of  $\Sigma_{i,i}^a$ , energy of  $\varphi$ th, and  $\tau$ th stream of  $s_i$ , respectively. From (34), the received SNR for  $l$ th date of user  $i$  is obtained as

$$\text{SNR}_l = \frac{P_l}{\sigma_n^2 \left( \sum_{\varphi=1}^{\aleph} P_\varphi^2 (\lambda_\varphi^a)^{-2} + \sum_{\tau=1}^{\Gamma} P_\tau^2 (\lambda_\tau^a)^{-2} \right)}. \quad (35)$$

Then, the SR upper bound for  $i$ th user can be calculated as

$$C_i \leq \sum_{l=1}^{\max(\aleph, \Gamma)} \log \left( 1 + \frac{P_l}{\sigma_n^2 \sum_{\varphi, \tau=1}^{\max(\aleph, \Gamma)} (\eta_\varphi + \eta_\tau)} \right). \quad (36)$$

It shows that  $C_i$  contains only normalization factors  $\eta_\varphi$  and  $\eta_\tau$ . The maximum value of  $C_i$  is achieved only and only if  $P_1^2/(\lambda_1^a)^2 = \dots = P_\varphi^2/(\lambda_\varphi^a)^2 = \dots = P_\tau^2/(\lambda_\tau^a)^2$ ; thus, the ASR for total system at high-SNR region can be expressed as

$$C \leq \sum_{i=1}^K \sum_{l=1}^{\max(\aleph, \Gamma)} \log \left( 1 + \frac{P_l}{2\sigma_n^2 \max(\aleph, \Gamma)} \right). \quad (37)$$

Therefore, the total achievable DOF for this network can be represented as the sum of DOF for each link [20]. Consider

$$\begin{aligned}d_{\text{total}} &= \lim_{\text{SNR} \rightarrow \infty} \sum_{i=1}^K d_{i,j} \\ &= \lim_{\text{SNR} \rightarrow \infty} \sum_{i=1}^K \frac{C}{\log(\text{SNR})},\end{aligned}\quad (38)$$

where  $d_{i,j}$  denotes the DoF for the transmission from user  $i$  to user  $j$ .

## 5. Computational Complexity Analysis

In this section, we will compare the computational complexity of proposed scheme and prior works. We use the total number of floating point operations (FLOPs) to measure the computational complexity. According to [21], the required FLOPs of each matrix operation are described as follows:

- (i) multiplication of  $m \times n$  and  $n \times p$  complex matrices:  $8mnp - 2mp$ ;
- (ii) multiplication of  $m \times n$  and  $n \times m$  complex matrices:  $4nm \times (m + 1)$ ;
- (iii) SVD of and  $m \times n$  ( $m \leq n$ ) complex matrix where only  $\Sigma$  is obtained:  $32(mm^2 - n^3/3)$ ;
- (iv) SVD of and  $m \times n$  ( $m \leq n$ ) complex matrix where only  $\Sigma$  and  $\Lambda$  are obtained:  $32(nm^2 + 2m^3)$ ;
- (v) SVD of and  $m \times n$  ( $m \leq n$ ) complex matrix, where only  $U$ ,  $\Sigma$ , and  $\Lambda$  are obtained:  $8(4n^2m + 8nm^2 + 9m^3)$ ;
- (vi) inversion of an  $m \times m$  real matrix using Gauss-Jordan elimination:  $2m^3 - 2m^2 + m$ ;
- (vii) Cholesky factorization of an  $m \times m$  complex matrix:  $8m^3/3$ .

For the conventional RBD method [13], the authors consider a multiuser MIMO downlink precoding system with a base station communicating with  $K$ -users simultaneously. For the nonregenerative MIMO relay systems [14], the authors consider a 3-node MIMO relay, where multiple antennas are equipped at the source  $S$ , the relay  $R$ , and the destination  $D$ . We compare the required FLOPs of each precoding algorithm for proposed method, conventional RBD, and nonregenerative MIMO relay system in Tables 1, 2, and 3, respectively, where we assume that  $N_T = N_R$  and  $\bar{N}_i = N_T - N_i$ .

For instance, the  $(2, 2, 2) \times 6$  case denote a system with user  $K = 3$ , each user with  $N_i = 2$  antennas, and total transmit antennas is  $N_T = 6$ . The required FLOPs of the proposed method, conventional RBD, and the nonregenerative MIMO relay system are counted as 34638, 40824, and 45306, respectively. From the results, we can see that the reduction in the number of FLOPs and the proposed method precoding are 15.15% and 23.55% as compared to the conventional RBD and the nonregenerative MIMO relay systems. Thus, the proposed algorithms exhibit lower complexity than the conventional RBD and the nonregenerative MIMO relay system approaches, and the complexity advantage grows as  $N_i$ ,  $N_T$ , and  $K$  increase.

## 6. Simulation Results

In this section, we show the performance of the proposed scheme in terms of the computation complexity, achievable sum-of-rate (ASR), and BER performance with some simulation results.

Using Tables 1, 2, and 3, we give the calculated results of FLOPs of the alternative methods in Figures 3 and 4. In the first comparison shown in Figure 3, we consider the case that  $N_T = K \times N_i$ . We set  $N_i = 2$  and express the computation cost as a function of  $K$ .

In Figure 4, we fix user  $K = 4$  and  $N_T = K \times N_i$  while the computation cost as a function of  $N_i$ . For conventional RBD method, the orthogonal complementary vector  $V_{k,0}$  with dimension  $\bar{N}_i \times N_T$  is obtained; it requires  $K$  times SVD operations and if we only want to compute  $V_{k,0}$ , the

TABLE 1: Computational complexity of proposed Algorithm 1.

Steps	Operations	FLOPs	Case (2, 2, 2) × 6
1 (a)	$U_{i,1}^H \Sigma_{i,1} \Lambda_{i,1}$	$8K (4N_T^2 N_i + 8N_T N_i^2 + 9N_i^3)$	13248
1 (b)	$\Lambda_{i,2}^H \Sigma_{i,2}^H U_{i,2}$	$8K (4N_T^2 N_i + 8N_T N_i^2 + 9N_i^3)$	13248
2	$H_{i,2} W H_{i,1}$	$K [8N_i N_T^2 - 2N_i N_T + 4N_i N_T \times (N_i + 1)]$	2088
3	$L_i^H L_i$	$2K [N_i + 2N_T N_i \times (N_i + 1) + 4N_i^3/3]$	508
4	$H_{\text{mse}}^\dagger$	$4N_R^3/3 + 12N_R^2 N_T - 2N_R^2 - 2N_T N_R$	2736
5	$H_{i,i} V_i^a V_i^b$	$8K [4N_T N_i^2 - 4N_i^3/3 + N_i^2 (N_i + 1)]$	2336
6	$(Q_i Q_i^H + \sigma_i^2 \Psi_i)^{-1}$	$K [4N_R N_i \times (N_i + 1) + 3N_i + 2N_i^3 - 2N_i^2]$	474
Total			34638

TABLE 2: Computational complexity of conventional RBD [13].

Steps	Operations	FLOPs	Case (2, 2, 2) × 6
1	$U_i \Sigma_i^a \Lambda_i^{aH}$	$32K(N_T \bar{N}_i^2 + 2\bar{N}_i^3)$	21504
2	$(\Sigma_i^{aT} \Sigma_i^a + \rho^2 I_T)^{-1/2}$	$K(18N_T N_i^2 + \bar{N}_i)$	336
3	$V_i^a D_i^a, (D_i^a \leftarrow 2)$	$8KN_T^3$	5184
4	$H_i P_i^a$	$K(8N_T N_i^2 - 2N_i^2)$	552
5	$U_i^b \Sigma_i^b V_i^{bH}$	$64K((9/8)N_i^3 + N_T N_i^2 + (1/2)N_T^2 N_i)$	13248
Total			40824

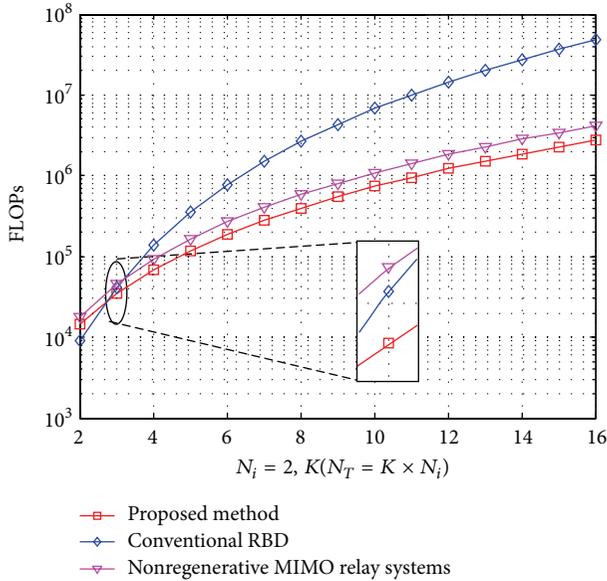


FIGURE 3: The complexity comparisons for required FLOPs versus the number of the users  $K$ .

computational is not efficient. In Step 5, after we got efficiency channel  $H_{\text{eff}} = H_i P_i^a$ , the second SVD operation should be carried out with dimension  $R_{\text{eff}} \times N_T$ , where  $R_{\text{eff}}$  is the rank of  $H_{\text{eff}}$ .

For nonregenerative MIMO relay system method, to simply discuss computational complexity, only the indirect link part algorithm is shown. In Steps 1 and 2, two SVD

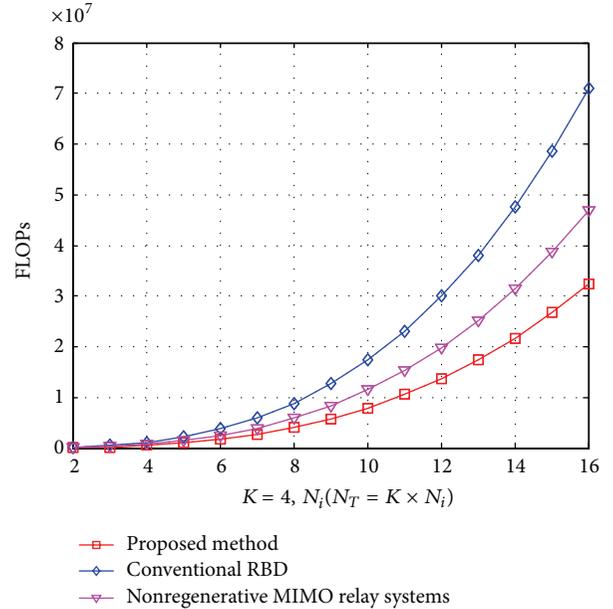


FIGURE 4: The complexity comparisons for required FLOPs versus the number of the receive antennas  $N_i$  for each user.

operations are required for the channels from the source to relay and relay to the destination Two variances  $H_i^H H_i$  and  $H_j^H H_j$  are needed to structure  $A$  as shown in Step 5. Finally, SVD  $A$  and diagonalize  $G$ .

For the proposed algorithm, the second precoding matrix  $V_i^b$  is structured by using Cholesky decomposition instead of

TABLE 3: Computational complexity of nonregenerative MIMO relay system [14].

Steps	Operations	FLOPs	Case (2, 2, 2) × 6
1	$U_i^a \Sigma_i^a \Lambda_i^{aH}$	$8K(4N_T^2 N_i + 8N_T N_i^2 + 9N_i^3)$	13248
2	$U_j^a \Sigma_j^a \Lambda_j^{aH}$	$8K(4N_T^2 N_i + 8N_T N_i^2 + 9N_i^3)$	13248
3	$H_i^H H_i$	$4KN_i N_T(N_i + 1)$	432
4	$H_j^H H_j$	$4KN_i N_T(N_i + 1)$	432
5	$H_i^H [\sigma_1^2 \sigma_2^{-2} (H_j F)^H H_j F + I]^{-1} H_i$	$2K(N_i^3 + 8N_i N_T^2 + 4N_i^2 N_T + 2N_i N_T - N_i^2 + N_i)$	4212
6	$V_A \Lambda_A V_A^H$	$8K(4N_T^2 N_i + 8N_T N_i^2 + 9N_i^3 + (N_i/2))$	13272
7	$\text{diag}(\bar{G})$	$K[4N_i N_T(N_i + 1) + 2N_i^3 - 2N_i^2 + N_i]$	462
Total			45306

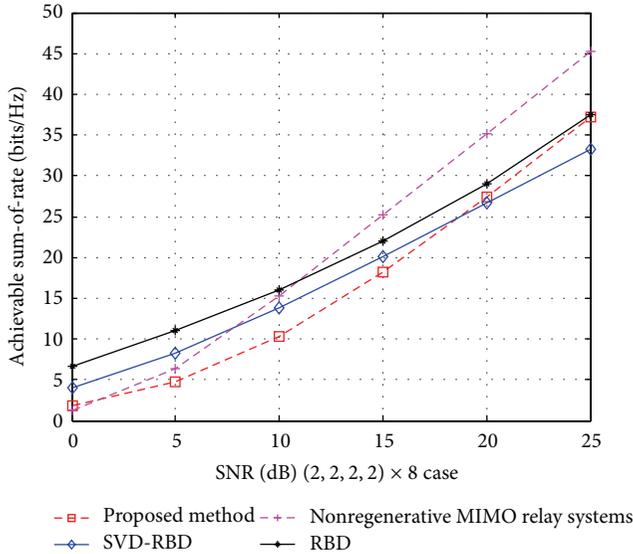


FIGURE 5: The achieved sum-of-rate of SVD-BD, RBD, nonregenerative MIMO relay systems, and proposed method for (2, 2, 2) × 8 case.

SVD operation and the first precoding matrix  $V_i^a$  is calculated by SVD of  $H_{i,\text{mse}}^\dagger$ , but only eigenvalue matrices are obtained. Obviously, the proposed method shows a clear advantage in comparisons.

In Figures 5 and 6, we compare the sum-of-rate of various MU-MIMO schemes under full CSI known at each node. The total capacity is obtained by using [22]

$$C_{\text{sum}} = \log \left( \det \left( I + \sigma_n^{-2} H P P^H H^H \right) \right), \quad (39)$$

and the ASR of proposed method is computed using (35), (36), and (37). Figures 5 and 6 illustrate the sum-of-rate as a function of SNR for (2, 2, 2) × 8 and (2, 2) × 4 cases, respectively.

In Figures 5 and 6, the nonregenerative MIMO relay systems show a better sum-of-rate than others at high SNRs, because its capacity includes direct links from source to the

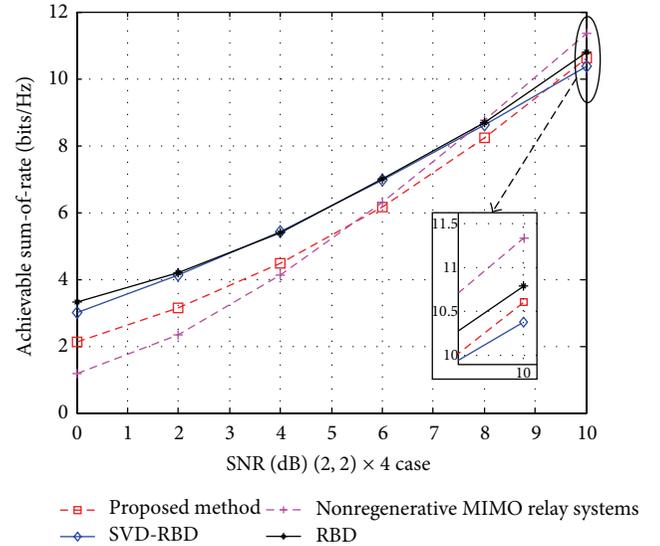


FIGURE 6: The achieved sum-of-rate of SVD-BD, RBD, nonregenerative MIMO relay systems, and proposed method for (2, 2) × 4 case.

destinations and indirect links via relay. The RBD precoding with SVD provides higher ASR than BD at whole SNRs. It is clear that the ASR of our proposed precoding algorithm is lower than the BR at low SNRs, but at high-SNR regime, it is higher than SVD-RBD and almost same as RBD.

In Figure 7, we compare the BER performance of BD-water filling, RBD, SVD-RBD, and proposed method, where QPSK modulation is applied. The proposed algorithm achieved better performance than existing precoding algorithms. As shown in Figure 7, the global optimal scheme in Section 3.3 is evaluated, the reason is that the precoding matrix  $V_i^a$  restricts the interference between the users close to zero while the other precoding algorithm is  $I_M$ . The performances significantly improve with increase of SNR.

## 7. Conclusion

In this paper, motivated by the structure of the quasi-EVD based channel in the relay-aided system, we have

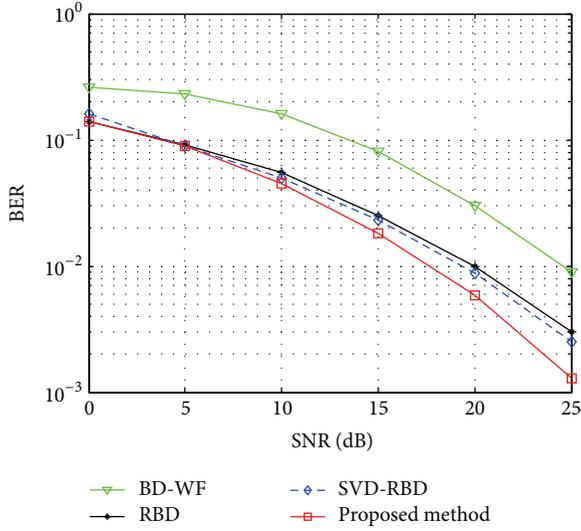


FIGURE 7: BER performance with QPSK.

demonstrated a novel iterative algorithm. Our goal is to achieve the maximum sum-of-rate and the minimum leakage interference. To minimize leakage interference, we use interference alignment to minimize the overlap of the row spaces spanned by the effective channels of different users. The design of the precoding matrix presented in this paper is general, which also can target minimum BER and reduce the computational complexity. In the first step, we use the Cholesky and the singular value decomposition to design the second part of precoder and solve the optimization problem for the total system with MMSE detector. In the next step, we apply the MMSE inversion to the equivalent channel to minimize the BER, which completes the first part of the precoder design. According to the precoding processes, the MU-MIMO channel is decoupled into a set of parallel independent SU-MIMO channels. Simulation results show that the proposed algorithm outperforms the existing techniques.

## Appendices

### A. The proof of Optimal MSE

The MSE at receiver can be further expressed as

$$\begin{aligned}
 \varepsilon_i &= \arg \min E \left\{ \left\| Z_i^H y_i - x_i \right\|^2 \right\} \\
 &= \text{tr} \left\{ \left( Z_i^H (H_{i,2} s_r + n_{i,2}) - x_i \right) \right. \\
 &\quad \left. \times \left( Z_i^H (H_{i,2} s_r + n_{i,2}) - x_i \right)^H \right\} \\
 &= \text{tr} \left( Z_i^H H_{i,2} W_i H_{i,1} P P^H H_{i,1}^H W_i^H H_{i,2}^H Z_i - Z_i^H H_{i,2} W_i H_{i,1} P \right. \\
 &\quad \left. - P H_{i,1}^H W_i H_{i,2}^H Z_i + \sigma_{i,1}^2 Z_i^H H_{i,2} W_i W_i^H H_{i,2}^H Z_i \right. \\
 &\quad \left. + \sigma_{i,2}^2 Z_i Z_i^H + I_M \right), \tag{A.1}
 \end{aligned}$$

where we have assumed that the signals and noise are independent with each other. Based on (10), the derivation of optimal MSE detection matrix  $Z_i^{\text{opt}}$  is equivalent to solving the following equation:

$$\frac{\partial \bar{x}_i}{\partial Z_i} = 2Z_i^H Q Q^H + 2\sigma_{i,2}^2 Z_i^H + 2\sigma_{i,1}^2 H_{i,2} H_{i,2}^H Z_i^H - 2Q^H = 0, \tag{A.2}$$

where  $Q_i = H_{i,2} W_i H_{i,1} V_i$ . To evaluate the efforts of the result,  $\text{tr}(H_{i,2} H_{i,2}^H)$  can be further developed as follows by applying singular value decomposition (SVD) shown in Section 2.2 on BCH channel:

$$\begin{aligned}
 \text{tr} \left( H_{i,2} H_{i,2}^H \right) &= \text{tr} \left( \Lambda_{i,2}^H \Sigma_{i,2}^H U_{i,2} U_{i,2}^H \Sigma_{i,2} \Lambda_{i,2} \right) \\
 &= \sum_{i=1}^M |\lambda_{i,2}|^2, \tag{A.3}
 \end{aligned}$$

where  $\lambda_{i,2}$  is the eigenvalues of  $H_{i,2}$ . Then, the closed-form expression of  $Z_i^{\text{opt}}$  can be obtained, which can be expressed as

$$Z_i^{\text{opt}} = Q_i^H \left( Q_i Q_i^H + \sigma_{i,1}^2 \Sigma_{i,2}^2 + \sigma_{i,2}^2 I_M \right)^{-1}; \tag{A.4}$$

this completes the proof.

### B. The proof of Lemma 1

*Proof.* In the high-SNR region, the objective function  $\varepsilon_i$  can be expressed approximately as

$$\varepsilon_i = \text{tr} \left[ \frac{1}{\sigma_i^2} Q_i \Psi_i^{-1} Q_i^H \right]^{-1}. \tag{B.1}$$

Since the matrix  $\Psi_i$  in the above function is Hermitian and positive definite, we can decompose this matrix using Cholesky factorization as

$$\Psi_i = I_M + H_{i,2} H_{i,2}^H = L_i^H L_i, \tag{B.2}$$

where  $L_i$  is an  $M \times M$  upper triangular matrix. Thus, the MSE  $\varepsilon_i$  can be rewritten as

$$\varepsilon_i = \text{tr} \left[ \frac{1}{\sigma_i^2} Q_i L_i^{-1} (L_i^H)^{-1} Q_i^H \right]^{-1}. \tag{B.3}$$

Using equivalent channel  $H_i = H_{i,2} W_i H_{i,1} = \Lambda_{i,2}^H \Sigma_{i,2}^2 \Lambda_{i,1}$ ,  $Q_i$  can be denoted as  $\Lambda_{i,2}^H \Sigma_{i,2}^2 \Lambda_{i,1} V_i$ , replace  $Q_i$  into (B.1), we can rewrite (B.1) as

$$\varepsilon_i = \text{tr} \left[ \frac{1}{\sigma_i^2} \Lambda_{i,2}^H \Sigma_{i,2}^2 \Lambda_{i,1} V_i L_i^{-1} (L_i^H)^{-1} V_i^H \Lambda_{i,1}^H \Sigma_{i,2}^2 \Lambda_{i,2} \right]^{-1}. \tag{B.4}$$

When MSE of the signal waveform estimation is adopted as the optimal problem in (20) which is solved in [23], the precoding matrices at source can be designed as

$$V_i^b = \Lambda_{i,1}^H B_i^{1/2} L_i, \tag{B.5}$$

where  $B_i \in \mathbb{C}^{M \times M}$  is a diagonal matrix as power constraint,  $A = V_i V_i^H = \Lambda_{i,1}^H B_i^{1/2} L_i L_i^H B_i^{1/2} \Lambda_{i,1} = B_i$ . Replacing the precoding matrix  $V_i$  into  $\varepsilon_i$ , the optimization problem is obtained as

$$\begin{aligned} \varepsilon_i &= \sigma_i^2 \text{tr} \left[ \Lambda_{i,2}^H \Sigma_{i,2}^2 \Lambda_{i,2} B_i \right]^{-1} \\ &= \sigma_i^2 \sum_{i=1}^M \sum_{i=1}^M \frac{|\lambda_{1,2,i}|^{-2}}{b_i}, \end{aligned} \quad (\text{B.6})$$

where  $\lambda_{1,2,i}$  is structured shown in (7); that is,  $\lambda_{1,2,i} = \lambda_{1,i} \cdot \lambda_{2,i}$  and  $b_i$  is the diagonal elements of matrices  $B_i$ . Similar to Lemma 2 in [24],  $\varepsilon_i$  is convex if and only if  $\varepsilon_i = h(A(\lambda_{1,2,1}, \dots, \lambda_{1,2,M}))$  is convex and nonincreasing with  $A$  and  $A = g(V_i)$  is a concave function of  $V_i$ . The Hessian matrices of  $V_i$  is  $\nabla_{V_i} V_i^H A = 0$  which is seminegative definite; it holds that  $g(V_i)$  is a concave function of  $V_i$ . Thus, Lemma 1 has been proven.  $\square$

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The work presented in this paper was supported in part by MEST 2012-002521 NRF, BK 21+ Korea, and Double Innovation Plan Eastern Liaoning University.

## References

- [1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.
- [2] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.
- [3] W. Rhee and J. M. Cioffi, "On the capacity of multiuser wireless channels with multiple antennas," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2580–2595, 2003.
- [4] T. Unger and A. Klein, "Duplex schemes in multiple antenna two-hop relaying," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 128592, 2008.
- [5] N. Jindal and Z. Luo, "Capacity limits of multiple antenna multicast," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '06)*, pp. 1841–1845, Seattle, Wash, USA, July 2006.
- [6] R. Zhang, Y.-C. Liang, C. C. Chai, and S. Cui, "Optimal beamforming for two-way multi-antenna relay channel with analogue network coding," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 5, pp. 699–712, 2009.
- [7] O. Munoz, J. Vidal, and A. Agustin, "Linear transceiver design in nonregenerative relays with channel state information," *IEEE Transactions on Signal Processing*, vol. 55, pp. 2593–2604, 2007.
- [8] Z. Fang, Y. Hua, and J. C. Koshy, "Joint source and relay optimization for a non-regenerative mimo relay," in *Proceedings of the 4th IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings (SAM '06)*, pp. 239–243, Waltham, Mass, USA, July 2006.
- [9] C. M. Yetis, T. Gou, S. A. Jafar, and A. H. Kayran, "On feasibility of interference alignment in MIMO interference networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4771–4782, 2010.
- [10] M. Chen and A. Yener, "Multiuser two-way relaying for interference limited systems," in *Proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 3883–3887, May 2008.
- [11] G. Kramer, "Outer bounds on the capacity of Gaussian interference channels," *IEEE Transactions on Information Theory*, vol. 50, no. 3, pp. 581–586, 2004.
- [12] A. S. Motahari and A. K. Khandani, "Capacity bounds for the Gaussian interference channel," Tech. Rep. UW-ECE 2007-26, Library Archives Canada, 2007.
- [13] H. Wang, L. Li, L. Song, and X. Gao, "A linear precoding scheme for downlink multiuser mimo precoding systems," *IEEE Communications Letters*, vol. 15, no. 6, pp. 653–655, 2011.
- [14] R. Mo and Y. H. Chew, "Precoder design for non-regenerative MIMO relay systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5041–5049, 2009.
- [15] C. M. Yetis, T. Gou, S. A. Jafar, and A. H. Kayran, "Feasibility conditions for interference alignment," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '09)*, pp. 1–6, November 2009.
- [16] N. Lee and J.-B. Lim, "A novel signaling for communication on MIMO y channel: signal space alignment for network coding," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '09)*, pp. 2892–2896, July 2009.
- [17] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3309–3322, 2011.
- [18] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming," July 2010, <http://cvxr.com/cvx>.
- [19] S. Gollakota, S. David Perli, and D. Katabi, "Interference alignment and cancellation," in *Proceedings of the ACM SIGCOMM*, ACM, Barcelona, Spain, 2009.
- [20] V. R. Cadambe and S. A. Jafar, "Degrees of freedom of wireless networks with relays, feedback, cooperation, and full duplex operation," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2334–2344, 2009.
- [21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.
- [22] S. Vishwanath, N. Jindal, and A. Goldsmith, "On the capacity of multiple input multiple output broadcast channels," in *Proceedings of the International Conference on Communications (ICC '02)*, pp. 1444–1450, May 2002.
- [23] Y. Rong, "Optimal joint source and relay beamforming for MIMO relays with direct link," *IEEE Communications Letters*, vol. 14, no. 5, pp. 390–392, 2010.
- [24] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H.-H. Chen, "Denoise-and-forward network coding for two-way relay MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 775–788, 2014.

## Research Article

# Spontaneous Ad Hoc Mobile Cloud Computing Network

Raquel Lacuesta,<sup>1</sup> Jaime Lloret,<sup>2</sup> Sandra Sendra,<sup>2</sup> and Lourdes Peñalver<sup>2</sup>

<sup>1</sup> Universidad San Jorge, A-23, km 299, Villanueva de Gállego, 50830 Zaragoza, Spain

<sup>2</sup> Universidad Politécnica de Valencia, Camino Vera S/N, 46022 Valencia, Spain

Correspondence should be addressed to Jaime Lloret; [jlloret@dcom.upv.es](mailto:jlloret@dcom.upv.es)

Received 28 April 2014; Accepted 10 July 2014; Published 17 August 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Raquel Lacuesta et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cloud computing helps users and companies to share computing resources instead of having local servers or personal devices to handle the applications. Smart devices are becoming one of the main information processing devices. Their computing features are reaching levels that let them create a mobile cloud computing network. But sometimes they are not able to create it and collaborate actively in the cloud because it is difficult for them to build easily a spontaneous network and configure its parameters. For this reason, in this paper, we are going to present the design and deployment of a spontaneous ad hoc mobile cloud computing network. In order to perform it, we have developed a trusted algorithm that is able to manage the activity of the nodes when they join and leave the network. The paper shows the network procedures and classes that have been designed. Our simulation results using Castalia show that our proposal presents a good efficiency and network performance even by using high number of nodes.

## 1. Introduction

A mobile ad hoc network (MANET) is a self-configuring network of mobile devices connected by wireless links. Each device in a MANET is free to move independently in any direction and will therefore frequently change its links to other devices. Each device must forward traffic unrelated to its own use and therefore could act as a router.

A spontaneous ad hoc network is a type of ad hoc network that is formed during a certain period of time, with no dependence on a central server and without the intervention of an expert user [1]. This network is made of several independent nodes which are in the same place at the same time in order to communicate with each other. Nodes are free to join and leave the network at will [2]. Spontaneous networking happens when neighboring nodes discover each other within a short period of time; however, discovery velocity is paid in terms of energy consumption [3]. Spontaneous networks are conceptually in a higher level of abstraction than ad hoc ones; they are basically those who seek to imitate human relationships in order to work together in groups, running on the already existing technology. Their objective is the integration of services and devices in an environment which allows the provision to the user of an

instant service with minimum manual intervention, ensuring important aspects, such as the multimedia quality [4] or network lifetime [5]. The concept of spontaneous networks was introduced in depth by Feeney et al. in [6].

The biggest problem in these networks is the security issue [7, 8]. The use of a certificate authority (CA) server is not a good idea because of the lack of a robust infrastructure and the distance. The device could be very far from the CA, so their connection could be a big issue. There is a need of a two-phase protocol to allow the exchange of an introductory description of each device, that is, a handshake of the devices within the same area. In order to achieve this, a protocol where each device has to exchange an identity card and will have a neighbour card list has been proposed. Thus, each node takes over the role of the CA. The model uses the trust between nodes as a key base of the proposed protocol. Castalia simulator has been used in order to validate our proposed spontaneous network.

The remainder of this paper is structured as follows. Section 2 shows some previous works about spontaneous ad hoc networks and mobile cloud computing. The proposed spontaneous ad hoc network model for mobile computing is explained in Section 3. It details the analytical model and the cryptographic and the trust system for protecting

the mobile cloud network. Section 4 explains the designed node algorithms, the network procedures, and the designed classes. The network performance is shown in Section 5. The deployment is explained in section. Finally, Section 7 shows the conclusion and future work.

## 2. Related Work

In mobile computing, there are some inherent problems such as resource scarcity, frequent disconnections, and mobility that make exploiting its full potential difficult. In [9], authors propose to address these problems by executing mobile applications on resource providers external to the mobile device. They provide an extensive survey of mobile cloud computing research highlighting the motivation for mobile cloud computing as the dominant model for mobile applications in the future. On mobile cloud computing, devices can act as clients or resource providers. Some requirements such as adaptability, scalability, availability, and self-awareness need to be met in a cloud. They present a taxonomy of the issues found in this area and the approaches in which these issues have been tackled. They focused their study on operational level, end user level, service and application level, and security and context-awareness level. They remark that although many of the reviewed frameworks mention the need for security and trust, very few of them have actually implemented it; they have left the implementation for future directions.

In this kind of network, it is very important to determine the application in which the network will be dedicated to. As a function of this, factors such as network size, the type of devices, the software applications, and the shared services will be defined. Another important issue is the routing protocol used to communicate all nodes. The routing protocols used in spontaneous ad hoc and sensor wireless networks could be the same as regular ones, so we should consider the same constraints such as transmission power, energy resources, bandwidth usage, delay, hop count, and QoS, among others [10]. All of these factors will be affected by issues such as link stability or level of mobility in nodes [11], which, somehow, will depend on the environment where the network is deployed.

Researchers claim that spontaneous networking presents the need of improving the wide-scale applications fully exploiting its potential and that this is due to the intrinsic complexity of spontaneous network management, unsuitable to be directly handled by application developers. Following this approach, Bellavista et al. [12] proposed a middleware called RAMP for managing the autonomic and cross-layer application of spontaneous networks. The RAMP prototype can be considered as a useful tool for the community of researchers in the field of generation of spontaneous networks. RAMP enables the dynamic sharing of all resources available via multiple, heterogeneous, intermittent, infrastructure-based, and ad hoc links, which are orchestrated in a lightweight way to compose the multihop paths needed to share applications at runtime. RAMP performance is evaluated considering aspects such as delay requirements

in real-time multimedia streaming. Finally, the simulation results show that the feasibility of this proposal achieves good results in a wide range of practical situations with different file sizes and path lengths.

Christensen [13] examines the architectural considerations of creating next generation mobile applications using smart mobile devices, context enablement using sensors on the device and cloud computing and RESTful web services. Mobile applications are enhanced with REST based cloud computing technologies to create applications on the smart mobile device with offload processing. To best leverage this, they consider the capabilities and constraints of these architectures.

Mani et al. presented in [14] a platform called SCOPE that implements an architecture to provide a P2P and spontaneous solution for social networking in local areas. SCOPE follows the hierarchical P2P model because in a network there are nodes with higher computing capability which can form an overlay and provide the distributed data management system for the P2P social network, meanwhile, client nodes connect to supernodes and rely on them for sharing their contents or accessing to the shared information. SCOPE is based on IEEE 802.11 ad hoc mode and needs no infrastructure. SCOPE is developed to work on mobile devices spontaneously without any dedicated network resources. The proposal provides the distributed database and lookup services deployed on distributed hash table (DHT) technology where it defines the rules for information management and creates our social networking overlay. As authors conclude, this proposal is able to provide session based communication services and provides a rich menu of social networking services from simple link/text sharing to P2P IP telephony.

There are lots of applications where spontaneous networks can be very useful. Regarding the environmental monitoring, Liu et al. presented an adaptive and efficient peer-to-peer search (AEPS) approach for distributed service discovery for dependable service integration on service-oriented architecture [15]. The proposal is able to efficiently discover desirable services for decision making of disaster monitoring and relief by interacting with connected nodes with incomplete information. AEPS builds a social network for each sensor node which contributes to an effective service discovery. AEPS is evaluated for rescue capability provision where the results demonstrated that distributed nodes can self-organize in a peer-to-peer way and discover the required service and information without any central administration. In this case, the creation of a spontaneous network is used to assist an emergency rescue team to make the correct real-time decisions in very changing environments.

In [16], the authors studied how the underused computing resources within an enterprise may be harnessed to improve their utilization and create an elastic computing infrastructure. They propose to use an ad hoc cloud model that allows complex cloud-style applications to exploit untapped resources on nondedicated hardware. They have outlined a case for ad hoc cloud computing, a set of resulting research challenges, and they propose an architecture. No protocol is designed and developed in this paper.

Taking into account that mobile devices are resource-constrained and some applications demand more resources than they can afford, in [17], Huerta-Canepa and Lee propose to create a virtual cloud computing platform using mobile phones. They present the preliminary design of a framework to create ad hoc cloud computing providers. The framework creates a cloud among the devices in the vicinity, allowing them to execute jobs between the devices. However, the work presented is preliminary, and no protocol has neither been designed nor detailed.

Another solution is the presented by Divya in [18]. He proposes a conceptual architecture where a mobile application platform shares a service among multiple users. A proof-of-concept prototype is developed using Android. The server platform shares Android OS among multiple users to obtain high performance on virtual image-based virtualization for mobile applications.

In this sense, several authors of this paper have previously deployed a spontaneous ad hoc network for multiple purposes, but never for cloud computing. In [19], a secure protocol for spontaneous wireless ad hoc networks was proposed. It is based on the behavior of human relationships. It uses a hybrid symmetric/asymmetric scheme and the trust between users in order to exchange the initial data and to exchange the secret keys that will be used to encrypt the data. In this paper, authors explained the procedures for the node's self-configuration and for providing DNS service. In [20], two flexible secure spontaneous wireless ad hoc network protocols for wireless mesh clients that are based on the computational costs are proposed. Proposals are based on a trust network, where the session key allows node confidentiality. They have been implemented over the DSR routing protocol. The developed protocols provide node authenticity and intermediate node authenticity when packets are transmitted. Integrity checking, random checking, and verification distribution are also considered in the protocol. In [21], we propose a secure spontaneous ad hoc network, based on direct peer-to-peer interaction, to grant a quick, easy, and secure access to the users when they surf the web. The protocol allows the users to collaborate during a period of time to accomplish a collaborative task. The proposal is also compared with other caching techniques published in the related literature. The proposed solution presents a distributed model where the interaction required between devices is minimal. In [22], we proposed a secure spontaneous network to create communities. Each community has an identity that acts as a unity on a world based on internet connection. Trust chains are established among users. Chains of confidence allow the establishment of groups or communities to access the services as well as for spreading group information.

As we have seen, no previous spontaneous network has been focused on mobile cloud computing.

### 3. Network Model and Description

In this section, the proposed spontaneous ad hoc network for mobile cloud computing and its model is described.

Our network model meets the following requirements.

- (1) Devices can move freely in the given area. Even out of each other's range.
- (2) Every node is also a router. It has a limited communication range towards other nodes.
- (3) The different identities are given by IP addresses. Each address is obtained dynamically following our previous proposal [23].
- (4) There is no central administration.
- (5) Devices can come from everywhere and join and leave at will.
- (6) Resources for cloud computing can be provided by any node if it has enough capacity to do it.

During the start-up of a node, it broadcasts messages in order to find neighbors. In this kind of networks, it is very important to select those nodes which offer better performance to the whole network [24]. A node will accept another node as its neighbor as a function of the amount of messages it has received to this node. When a new node has defined all its neighbors, it sends its identity card to all its neighbors. If the neighbor sends back a message to inform that it has received the identity card and the content of this message is correct (by checking the hash of the message), then the new node trusts these neighbors. When a node trusts a second node, it can send messages directly to the second node, unless the second node does not trust the first node; in this case, the communication is not allowed. The system does not follow the commutative and associative properties; that is, although a first node can trust a second node, the second node may not trust the first node. It also happens with three nodes in a chain. If a node wants to send a message towards a nontrusted node, then it has to do it through a trusted node.

The system follows the next steps.

- (1) Broadcast messages searching neighbor nodes.
- (2) Send its identity card to the neighbors.
- (3) Acknowledge/not acknowledge the reception of the messages from its neighbors.
- (4) Set the neighbor node as a trusted or nontrusted node.

Our protocol is based on the use of two information structures: (1) an IDC (identity card) and (2) a certificate. On one hand, the IDC is composed by two parts. The first one is the public part which is formed by a logical identity (LID). It is unique for each user and allows nodes to identify it. LID includes information such as name, photograph, and other user identification. Public part also contains information about the public key of the user ( $K_i$ ) and the information signature. On the other hand, private part is composed by the private key ( $k_i$ ) and this information is not accessible by other devices.

A certificate of a user consists of a validated identification card, signed by the user that gives its validity, for example, user "j". Thus, the certificate of user "i", validated and signed by the user "j", corresponds to " $C_{ij}$ ". The user will introduce its

LID only the first time that the user uses the system because the security information is generated only the first time the user joins the network. Security data is stored persistently in the device for its future use.

In order to explain the procedure, the following case is detailed. When there are three persons and two of them know certain data, then there is only one way that lets the third person know this data: one of the two persons must trust the third person. This simple example explains basically the concept of trusted network and how data are exchanged between nodes.

If the ad hoc network covers a large area, then the cloud computing services can be obtained by using ad hoc routing. In this proposal, the ad hoc routing is performed only between trusted nodes, so there are two important facts: (1) nodes should not be trusted without the proper authentication of the node and user and (2) confidentiality, integrity, availability, and access control with authentication, all of them must be based on encryption mechanisms that must be offered without central administration.

If we want to create a spontaneous ad hoc network for mobile cloud computing, we need trust establishment, key management, and membership control. Network availability and routing security must also be added. Techniques that enable the creation of ad hoc networks based on the spontaneity of human interactions (people who are near each other can communicate, exchange things, and ask people to relay information to others) should be added. In our model, each node will send its public key towards its neighbors. When the node obtains a public key, it is considered valid only if it is sure that it belongs to the owner. Not valid means that it is not sure that the key belongs to the owner. If the node trusts the key, it signs the key with its private key and considers the node as a trusted neighbor. Then, a trust network is created. When a new device joins the network and it does not have a pair of keys, it must generate them to perform authentication and to communicate with other nodes. When a node leaves the network, the network maintains the data for a period of time in case it wants to come back later. But it has to authenticate again. A node does not have to obtain the public key from every other node; in other words, one node does not have to broadcast its authentication information to all other nodes in the network. Nodes can obtain this information through the “network of trust.” Now, we provide a simple example where the network is formed by three nodes. Node 1 and node 2 know and trust each other. Then, node 2 trusts a third node, node 3. If the second node receives a public key from the third node and signs it with its private key, we consider that the owner of this key is “trusted.” Later, if the first node wants to obtain that key, it can be obtained from the second node, and since the first node trusts the second one, it “validates” this new key by signing it with its private key. If the third node is not trustworthy, any key signed by the third node will not be considered a trusted key. Furthermore, the first node will never sign the third node key, although it might forward it to other nodes in the spontaneous ad hoc mobile cloud computing network.

**3.1. Analytical Considerations.** In this section, we analyze our proposal analytically. Our purpose is to model the behavior of the spontaneous ad hoc network when there are nodes joining and leaving during its existence. On one hand, we will take care of the authentication and trust between nodes (by using trust links and trusted communication graphs), and, on the other hand, we model the network behavior (in terms of number of nodes in the network because of leavings and joining of new nodes) by using the conditional probability density function. Let  $N(t)$  be a set of users having a meeting with wireless devices, with  $|N(t)| = n$  being the maximum number of users during a certain period of time  $t$ . These users are located in a certain bounded region  $R$ . The trusted communication graph is the directed graph  $G(t) = (N(t), E(t))$  such that each pair of users  $(u, v) \in E(t)$  only if the user's device  $v$  is within  $u$ 's transmitting range at the current transmit power level at time  $t$ . The graph contains all possible wireless links between the nodes in the network. Given any two nodes  $u, w \in N$ , a path connecting  $u$  and  $w$  in  $G$  is a sequence of nodes  $\{u = u_0, u_1, \dots, u_{k-1}, u_k = v\}$  such that for any  $i = 0, \dots, k-1, (u_i, u_{i+1}) \in E$ . The length of the path is the number of edges in the path. Moreover,  $u$  has a pair of values  $(T, V)$  for each node  $w$  which gives the trust ( $T$ ) and validity ( $V$ ) values for each user. The trust and validity can only have two values:  $T = \{0, 1\}$  and  $V = \{0, 1\}$ .

Users can join and leave the spontaneous network at will, so a range assignment RA is said to be connected at time  $t$ , if the resulting communication graph at time  $t$  is strongly connected; that is, if for any pair of nodes  $u$  and  $v$ , there exists at least one trusted connection from  $u$  to  $v$ . In other words, the trusted directed wireless link  $(u, v)$  exists if and only if nodes  $u$  and  $v$  are at distance of at most RA( $u$ ) at time  $t$  and their *trusted* parameter value is equal to 1. In this case,  $v$  is said to be a 1-hop neighbor, or *neighbor* for short, of node  $u$ . The trust nodes set of node  $u$ , denoted as TNS( $u$ ), is defined as it is shown in the following expression:

$$\text{TNS}(u) = \{z \in N : (z, u) \in E, T = 1\}. \quad (1)$$

A trusted wireless link is said to be bidirectional, or symmetric, at time  $t$  if  $(u, z) \in E(t)$ ,  $(z, u) \in E(t)$ ,  $u$  trusts  $z$ , and  $z$  trusts  $u$ . The trusted communication graph generated can be considered as undirected, since  $(u, z) \in E(t) \Leftrightarrow (z, u) \in E(t)$ .

Let us suppose that a user  $u$  has authenticated the user  $w$  and  $u$  has  $w$ 's public key, then  $u$  sends a message encrypted with the session key to  $w$ . In this case, we say that there is a trusted directed graph from  $u$  to  $w$ . For any trusted directed graph  $H \in G$ , if two users  $u$  and  $w$  are in  $H$ , and there is a trusted directed path from  $u$  to  $w$  in  $H$ , then we say that  $w$  is reachable from  $u$  in  $H$  and we denote this by  $(u \leftrightarrow w)_H$ ; thus,  $w$  is also reachable from  $u$  in  $G$ , and we denote this by  $(u \leftrightarrow w)_G$ .

In order to use a public key distribution system for user authentication and session key sharing, each user maintains a local repository of public key certificates and their *trust* values. When the user  $u$  wants to use the resources shared by user  $w$ , first  $w$  must trust  $u$ , so they must merge their subgraphs and try to find a trusted directed path from  $u$  to  $w$ .

Thus, we can apply a subgraph theory based similar to work performed by Capkun et al. in [25]. But in that case they merged subgraphs to authenticate a public key in trusted authorities or certificate repositories, not trusted paths, while we merge subgraphs to validate trusted paths.

We assume that each user has the same subgraph selection algorithm  $A$  to build its subgraph. We denote  $S_A(H, u)$  by the algorithm  $A$  executed in  $H \in G$  by the user  $u$ . When we merge the subgraph  $S_A(H, u)$  of user  $u$  with  $S_A(H, v)$  of user  $v$ , we obtain  $S_A(H, u, v)$ . When the trusted communication graph is undirected,  $S_A(H, u, v) = S_A(H, v, u)$ .

The performance of the subgraph selection algorithm, denoted by  $P_A(H)$ , is defined as the ratio of the number of user pairs  $(u, w)$ , where there is a trusted directed path from  $u$  to  $w$  in the merge subgraph of  $u$  and  $w$  to the number of user pairs  $(u, w)$ , where there is a directed path from  $u$  to  $w$  in the trust graph. The following expression shows  $P_A(H)$

$$P_A(H) = \frac{\text{Card} \{(u, v) \in NxN : (u \leftrightarrow v)_{S_A(G,u,v)}\}}{\text{Card} \{(u, v) \in NxN : (u \leftrightarrow v)_G\}}, \quad (2)$$

where *card* denotes the cardinality of a set. The performance of  $A$  can be increased by selecting larger subgraphs, that is, using more information about the trust graph, but the devices of the users will need more memory to store their subgraphs. Moreover, the devices will need high amount of knowledge to execute it.

In order to model the network behavior when the users join the network during the meeting time, we have used the diffusion approximation. In the spontaneous network, there will be users that join and leave the ad hoc network at will. Let  $t_i$  be the arrival time of the user  $i$  to the network, and let  $t'_i$  be the departure time of the user  $i$ . That is,  $0 \leq t_i < t'_i \leq T$ , where  $T$  is the network lifetime. Let  $A(t)$  and  $D(t)$  represent the cumulative number of arrivals and departures, respectively, up to time  $t$ . The number of users in the spontaneous network at time  $t$ ,  $N(t)$ , is given by the following expression:

$$N(t) = A(t) - D(t). \quad (3)$$

Let the consecutive interarrival time  $a_i = t_i - t_{i-1}$  and the consecutive interdeparture time  $d_i = t'_i - t'_{i-1}$  be both independent and identically distributed with the (mean, variance) given by  $(1/\mu_a, \sigma_a^2)$  and  $(1/\mu_d, \sigma_d^2)$ , respectively. Let their squared coefficients of variation be  $C_a^2 = \sigma_a^2 \cdot \mu_a$  and  $C_d^2 = \sigma_d^2 \cdot \mu_d$ , respectively. We define the sum of a set of consecutive interarrival times as  $T_k = \sum_{i=1}^k a_i$ . We assume that they are independent and identically distributed random variables; hence, according to the central limit theorem, the standardized random variable,  $T_k^*$ , shown in expression (4) tends to a standard normal distribution with  $k \rightarrow \infty$ , as it is shown in the following expression:

$$T_k^* = \frac{T_k - k \cdot \mu_a}{\sigma_a \sqrt{k}}, \quad (4)$$

$$N(t) = \lim_{k \rightarrow \infty} P \left[ \frac{T_k - k \cdot \mu_a}{\sigma_a \sqrt{k}} \leq n \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^n e^{-(t^2/2)} dt. \quad (5)$$

If  $k$  is large enough, there will be many arriving users between  $t$  and  $t + k$  and may be approximated by the normal distribution with mean  $\mu_a t$  and variance  $\sigma_a^2 \mu_a^3 t$ . Similarly, the number of leaving users during that time will be approximately normally distributed with mean  $\mu_d t$  and variance  $\sigma_d^2 \mu_d^3 t$ . Consequently, the changes of  $N(t)$  within the interval  $[t, t+k]$ , then  $N(t+k) - N(t)$ , should be approximately normally distributed with the mean as is shown in the following expression:

$$\beta = (\mu_a - \mu_d) t. \quad (6)$$

And the variance is given by the following expression:

$$\alpha = (\sigma_a^2 \mu_a^3 + \sigma_d^2 \mu_d^3) t. \quad (7)$$

The diffusion approximation replaces  $N(t)$  by a continuous diffusion process (also known as Wiener-Levy process)  $x(t)$ , normally distributed with the mean  $\beta \cdot dt$  and variance  $\alpha \cdot dt$ . Given the initial value  $x_0 = 0$ , the unrestricted process  $x(t)$  would have the conditional probability density function at time  $t$  given by the following expression:

$$P(x, t) = \frac{1}{\sqrt{2\pi\alpha t}} e^{-((x-\beta t)^2/2\alpha t)}, \quad (8)$$

which satisfies Kolmogorov diffusion equation (also known as Fokker-Planck equation) given in the following expression:

$$\frac{\partial f(x, t)}{\partial t} = -\beta \frac{\partial f(x, t)}{\partial x} + \frac{\alpha}{2} \frac{\partial^2 f(x, t)}{\partial x^2}. \quad (9)$$

Deriving expression (8) in expression (9) and treating  $x = 0$  as a reflecting barrier for all  $t > 0$ , we obtain the following expression:

$$\lim_{x \rightarrow \infty} \left[ -\beta P(x, t) + \frac{\alpha}{2} \frac{\partial P(x, t)}{\partial x} \right] = 0. \quad (10)$$

Now, we can estimate the solution when  $t \rightarrow \infty$  and  $\mu_a < \mu_d$ . Expression (11) shows the equilibrium distribution of the conditional probability density function [26]

$$P(x) = \frac{2|\beta|}{\alpha} e^{-(2|\beta|x/\alpha)}, \quad (11)$$

where  $\alpha$  and  $\beta$  are defined in (6) and (7), which are related to the probability of nodes' interarrival and interdeparture times.

**3.2. Cryptographic System.** The cryptographic algorithm election has been taken bearing in mind their strong security and simple key management features. Symmetric algorithms and summary functions have lower computational cost than public key cryptography, but public key cryptography has stronger security and it could be feasible in devices with low computation capacity.

In our proposal, we use a summary function with symmetric and asymmetric algorithms with the purpose of taking their benefits. The security management is based

TABLE I: Trust and validity values.

Parameter	Level	When does it happen?
Trust	0	(i) There has been a greeting process between users, but the one in the network does not trust him (ii) The user has reduced the level of confidence to the other one
	1	(i) It has been a greeting process between user and the one in the network that trusts him (ii) The user has increased the level of confidence to the other one
Validity	0	(i) It is not obtained from the greeting process with that user (ii) The validity has not been obtained through a trusted node
	1	(i) It has been obtained directly from the greeting process with that user (ii) It has been obtained through a trusted node

on the public key infrastructure and the symmetric key encryption scheme. Asymmetric key encryption scheme is mainly used in the distribution of session key and in the user authentication process. It lets us also generate a distributed certification authority. The symmetric key is used as a session key to cipher the confidential messages between trust nodes, because it has less energy requirements [27–29]. Asymmetric key encryption scheme is used to authenticate the users. The hybrid symmetric/asymmetric scheme is introduced to exchange the initial data and to exchange the secret keys that will be used to encrypt the data. The hash function lets us improve the data integrity. Now, we are going to discuss which algorithms are the best for our purpose.

We have used advanced encryption standard (AES) algorithm for the symmetric encryption scheme [30]. It presents a high security level because its design structure removes subkey symmetry. It is also resistant to lineal and differential cryptanalysis. AES is actually considered as one of the most secured ones. Moreover, the execution times and the energy consumption in the cryptography processes are adequate for low power devices.

The asymmetric encryption scheme should overload the devices as less as possible. On one hand, elliptic curve cryptosystem (ECC) is presented as a high performance scheme that is recommended by many researchers [31]. On the other hand Rivest, Shamir and Adleman cryptographic algorithm (RSA) is very secure and it has been checked and recommended by many scientists [32]. ECC needs fewer bits than RSA (163 bits in ECC versus 1024 bits in RSA) in order to obtain the same security level, so it is able to achieve high security level with low size keys without consuming too much system resources, thus needing less bandwidth. ECC is usually adequate for small devices with few memory resources and low computing resources (such as cellular phones and smart cards). In order to have flexibility in our protocol and because both cryptographic algorithms have good performance, we have included both (RSA and ECC) in our protocol. The election of one of them will be taken in the network formation. Both algorithms will be shown later in our performance study.

We have selected secure hash algorithm (SHA-1) for the summary function [33]. SHA-1 is commonly used because of its equilibrium between its speed and its security. This performance is also maintained in low computing devices. This feature does not happen in other functions, because they mainly depend on the processor. Its execution time and its

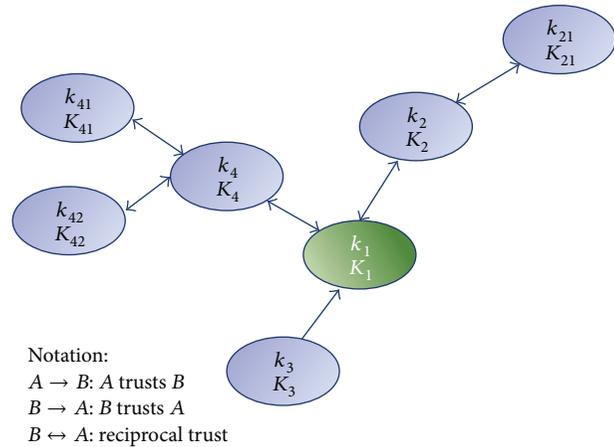


FIGURE 1: Example of trust nodes.

energy consumption are not so high when they are compared with other functions.

**3.3. Trusted Network.** The proposed model is based on the creation and management of a trusted network. A node will trust other nodes through personal view and criteria. That is, the trust is based on the relationship of the users rather than on a central certification authority. The user of the device will identify the other users and will be in charge of establishing a trust value (0 or 1) associated with each one of them. The parameters used for configuring this trust network are *trust* and *validity*.

Trust refers to the person who owns the key and its value will be established by the relationship between the user that grants it and the user that is granted. It should be granted to reliable persons when their IDCs are exchanged. The *trust* can always be changed manually by the user later. *Validity* indicates that a certificate belongs to that person/device. Table I shows the *trust* and *validity* values.

**3.4. Certification Authority.** The certification authority of a node could be any node in the group of nodes that this node trusts. This system lets us build a distributed certification authority between trust nodes. When a node wants to communicate with other nodes and see if it is a valid node, it can request the certificate of that node to its trust nodes. After obtaining this certificate, it will be able to sign this node

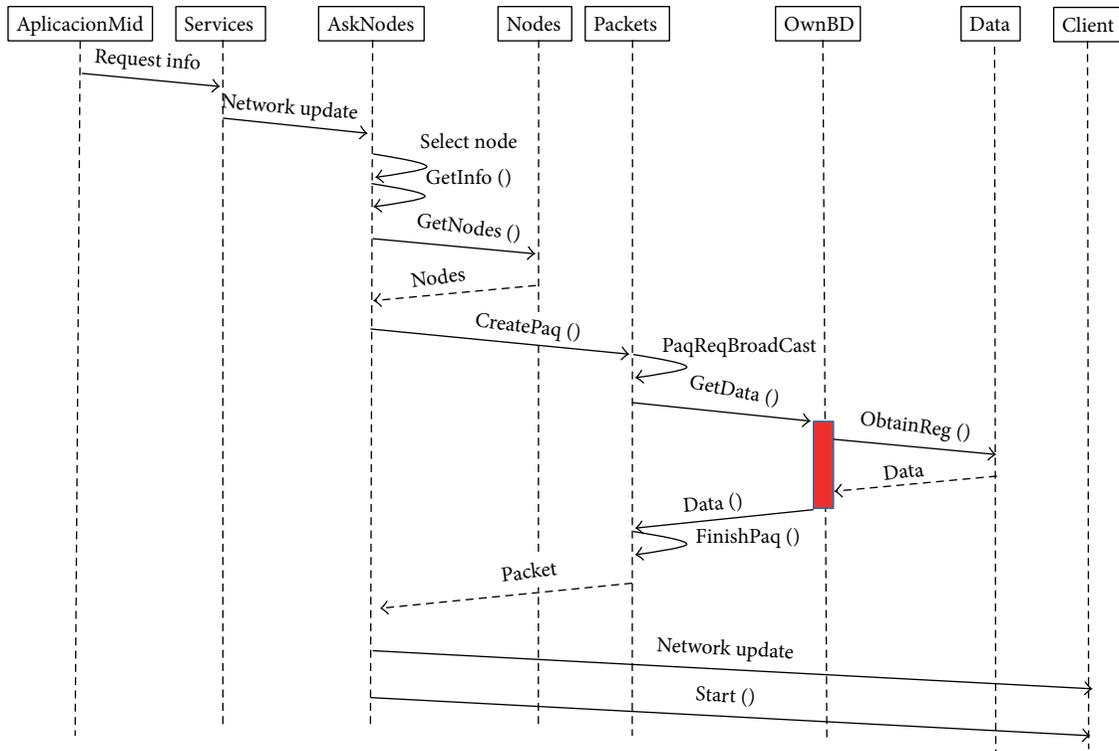


FIGURE 2: Procedure to request an update from all network nodes.

as a valid node. All nodes can be both, client, requesting information or authentication to other nodes, and server, serving requests for information or authentication from other nodes. Figure 1 shows an example. Each  $n$  node has its public key ( $K_n$ ) and its private key ( $k_n$ ). Nodes 2 and 4 are trust nodes of node 1, but not of node 3. Thus, nodes 2 and 4 could act as a certification authority of node 1.

#### 4. System Design

In this section, we explain the designed algorithms for the nodes, the network procedures, and the classes designed for Castalia in order to simulate it.

**4.1. Network Procedures.** After defining the network model and the security features that our proposed spontaneous ad hoc network should present, it is important to specify how a node should work and the set of actions it should perform to ensure the correct operation of the whole network. This subsection explains the operation of the network and the main processes included.

In order to design the flow chart diagrams, we have used the Unified Modeling Language (UML) [34]. UML is an industry standard modeling language with a rich graphical notation and comprehensive set of diagrams and elements that can be used to model object oriented systems.

**4.1.1. Procedure to Request the Update to all Network Nodes.** A user requests a data updating from all nodes. Firstly, the information about network nodes is obtained. Secondly, data

updating packet is prepared and the request is sent. Figure 2 shows the procedure and the primitives and services offered and served by the nodes.

**4.1.2. Procedure to Process a Request.** When a request is received, the information requested is checked and a reply is sent to the source node that sent the request. Then, the request is forwarded to the rest of network nodes. After receiving the data, request form a new client can be attended. The process to attend each request is shown in Figure 3. As the diagram shows, each task is validated by the replying of packets with data information, node information, or control packets. After the confirmation, a new packet is sent in order to complete the communication between nodes.

#### 4.2. Node Algorithms

**4.2.1. Packet Control.** When a packet is received by a node, it applies a packet control process to the received data. Figure 4 shows it. Data are checked to know whether it is correct and it has not been modified during the transmission or not. The source IP, packet number, and retry number are some of the checked parameters. This check is performed just for security reasons, despite of its analysis at lower levels such as the underlying wireless communication technology (Bluetooth, Wi-Fi, etc.).

When Bluetooth is used, the Bluetooth frames are used in the authentication process. In this case, the packet digest (hash) is not ciphered with the session key because the receiver node does not have this key yet. The packet includes

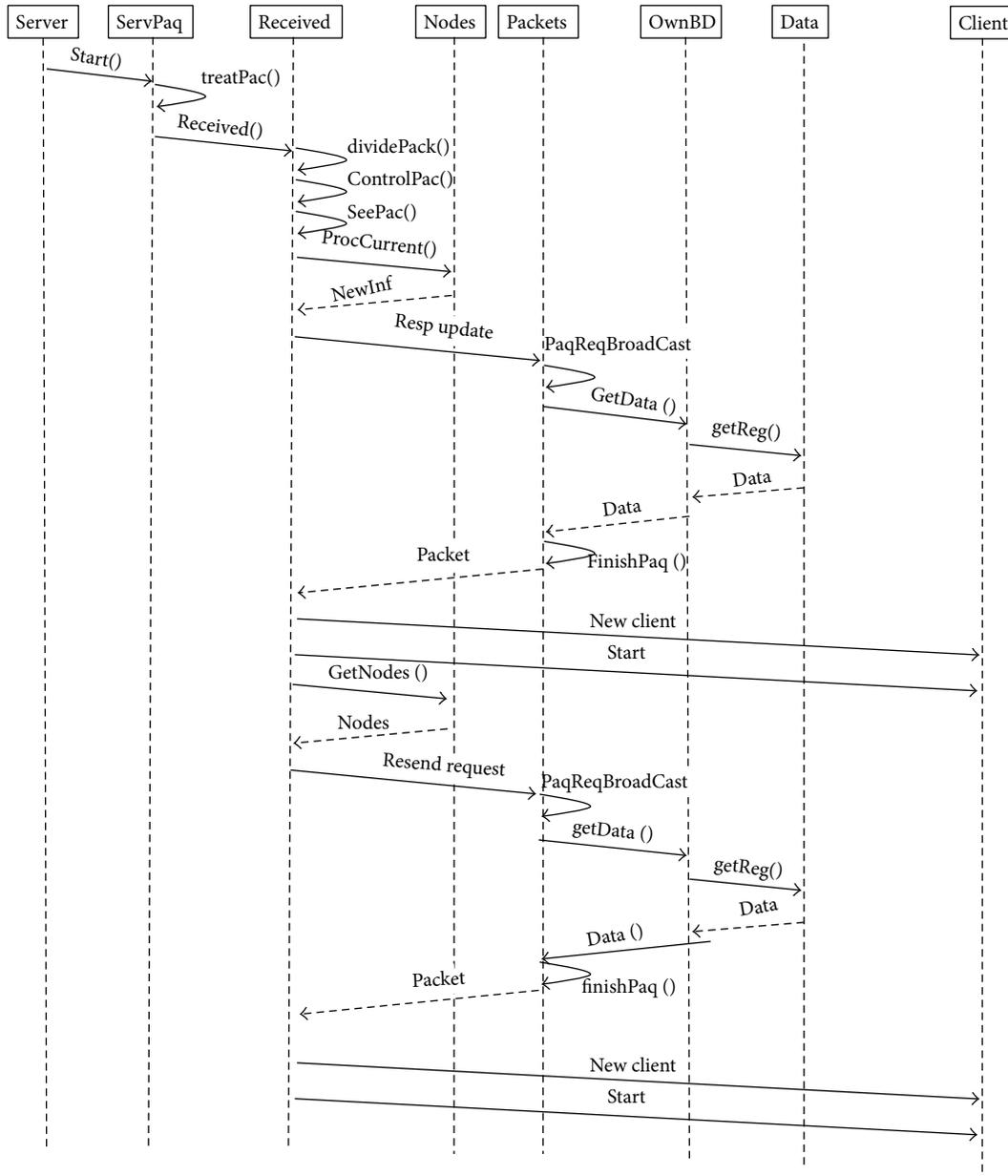


FIGURE 3: Procedure to process a request.

the sender’s node certificate ciphered with the sender’s private key. Sender’s public key is also included in the packet. Sender’s node certificate is deciphered with its public key.

When Wi-Fi is used, the packet digest is ciphered with the session key. When the frame is received, data are firstly deciphered and then the frame is checked.

In both cases, if the hash comparison is wrong, the system shows a message, informing that the frame is wrong. If the results of the comparisons are valid, packets are processed and this process ends.

4.2.2. *Modification of Keys.* When a user decided to modify its asymmetric keys, he is notified of the risks. This notification is shown as a text message in an emergent window. If the user decides to modify it, user’s keys are regenerated

and the certificate is modified. As Figure 5 shows, after these changes, new data is stored. Finally, the system allows user to stop the process of modification keys before they are changed.

4.2.3. *Main Menu.* For the development of the software application, we have designed a main menu that includes a submenu the services offered in the mobile cloud computing network, a submenu that allows the user to exchange data, and a submenu to see its own data. Figure 6 shows all the possibilities that the main menu offers.

4.2.4. *Request of Information.* In our system, there are two types of information request:

- (i) request for one node: there is a request for one node about specific information;

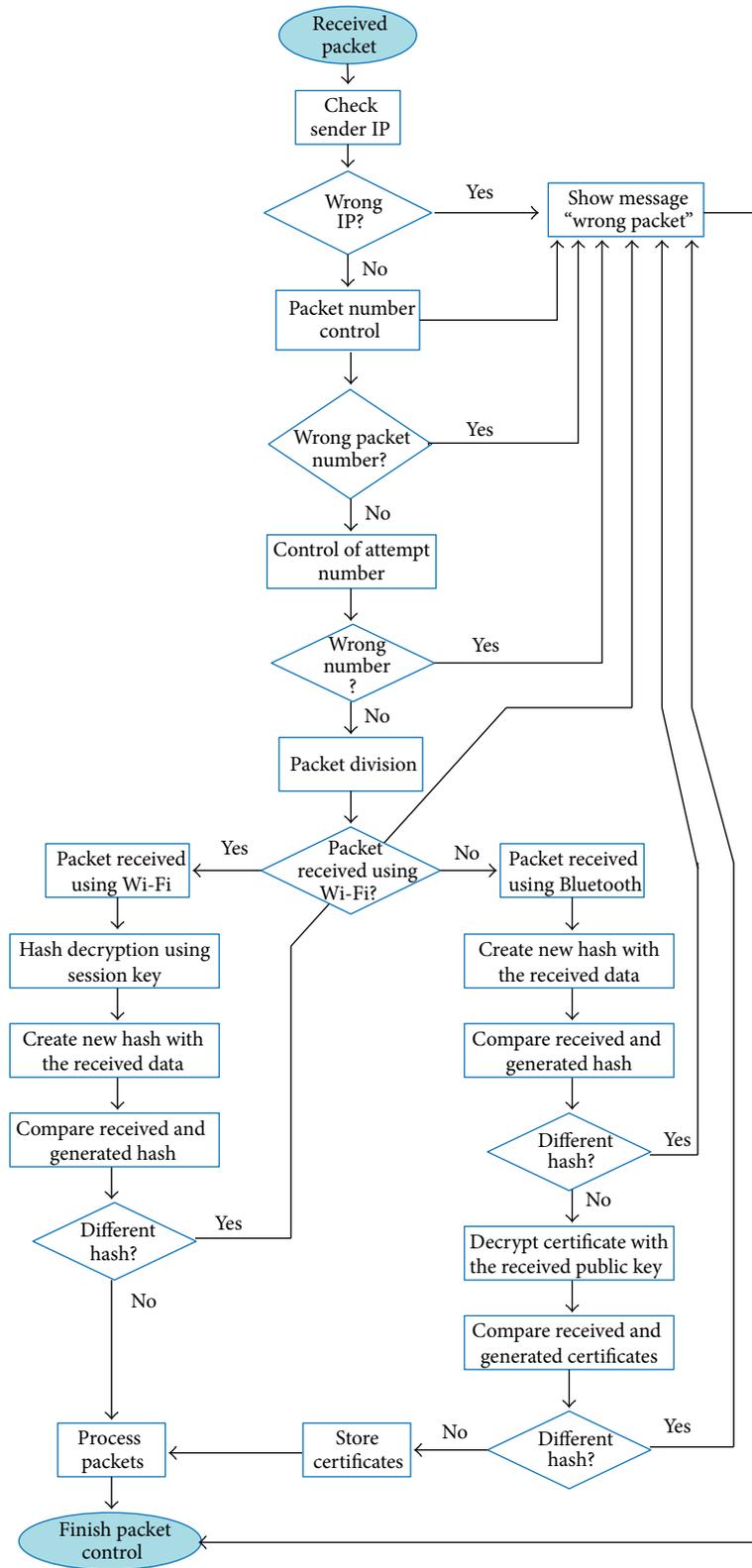


FIGURE 4: Packet control.

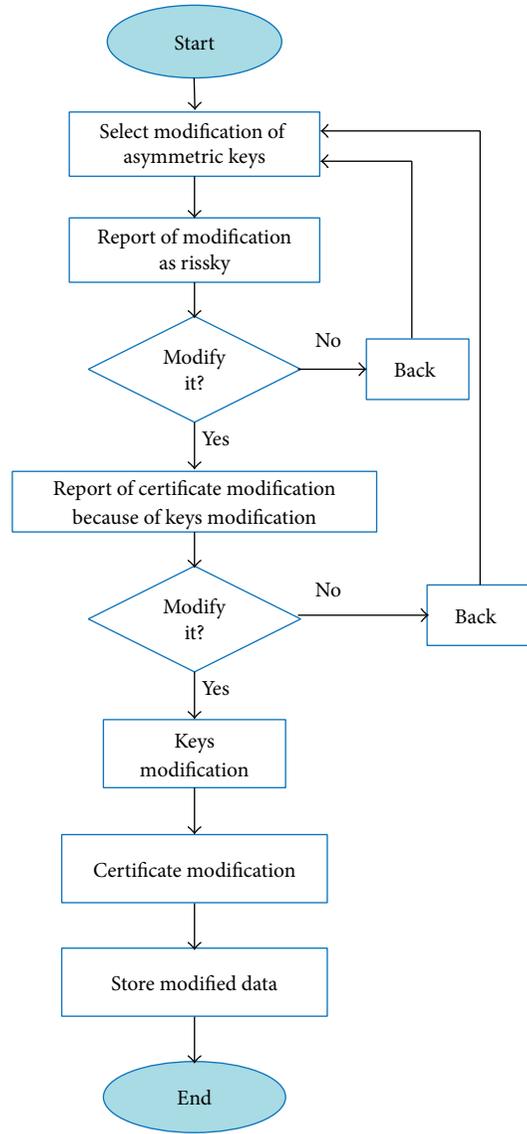


FIGURE 5: Modification of keys.

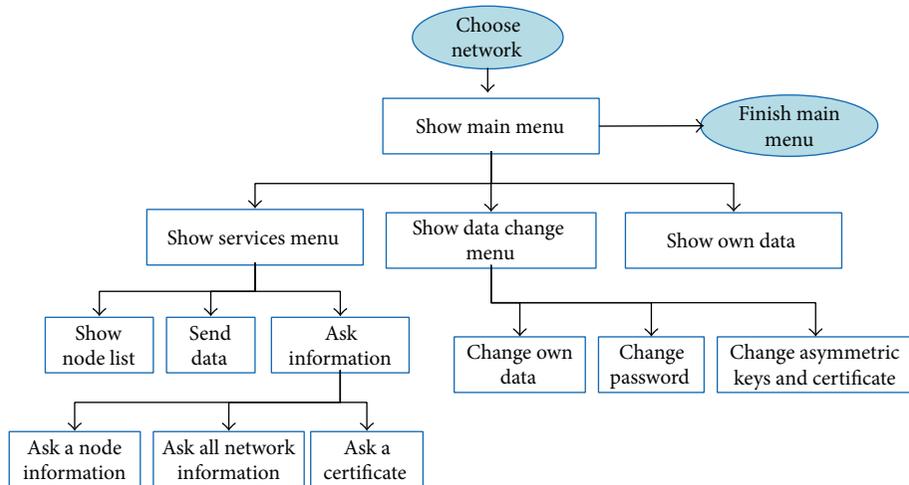


FIGURE 6: Main menu.

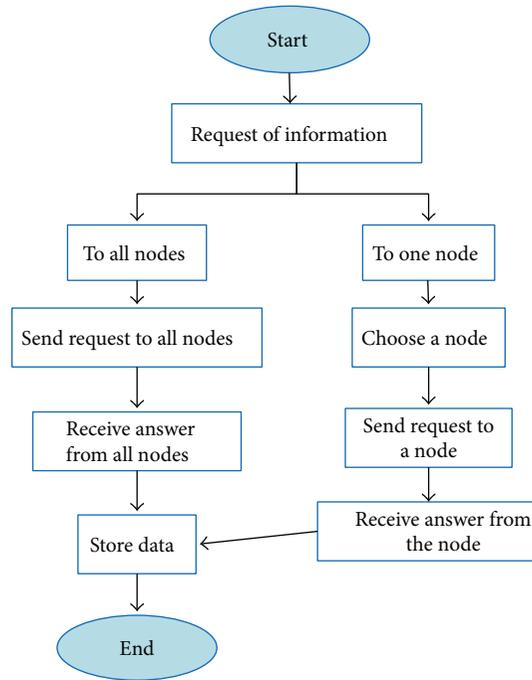


FIGURE 7: Request of information.

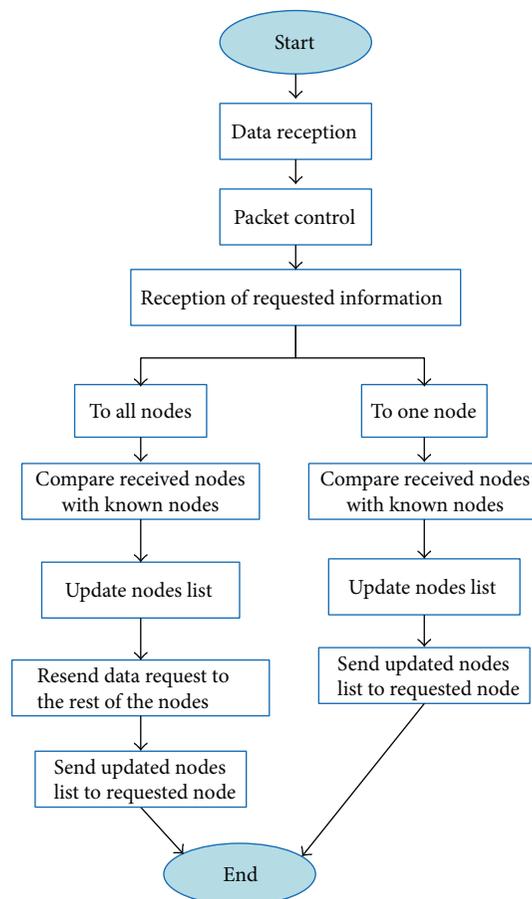


FIGURE 8: Procedure when information is requested.

```

class trustednet_Ident
{
    private:
        int id;
        std::string computeHash() const;
    public:
        in tip;
        trustednet_Key pair;
        time_t timestamp;
        CSHA1 sha1;
        trustednet_Ident(void);
        trustednet_Ident(int_id);
        int getID() const;
        writeIdent(std::ostream & os) const;
        bool test() const;
};

```

ALGORITHM 1: Program code of the class trustednet ident.

```

class trustednet_Node
{
    private:
        double* coordinates;
        trustednet_Neighbor** neighbors;
    public:
        int nNeighbors;
        trustednet_Ident identity;
        trustednet_Node();
        trustednet_Neighbor* getNeighborById(int_id);
        trustednet_Neighbor* getNeighbor(int_id);
        void addNeighbor(trustednet_Node* node);
        void setProperty(Trustednet_Node* node);
        void writeNode(std::ostream & os) const;
        double getXCOR();
        double getYCOR();
};

```

ALGORITHM 2: Program code of the class trustednet node.

- (ii) request for all nodes: there is a request for all network nodes requesting for the available information in the network (such as shared resources).

As Figure 7 shows, the information can be requested to one node or to all nodes. After receiving the information requested, these data are stored.

**4.2.5. Reply to a Request of Information.** Figure 8 shows the process of how to reply to a request of information. When the node receives a request, the reply depends on the type of the received request. If the request is about all network data, the node will reply with the updated data and will forward the request to the rest of the nodes. If the request is just to one node, the receiver node replies with the data request by the source node.

**4.3. Classes Design.** In this subsection, we describe the main designed classes for the proper operation of the spontaneous ad hoc mobile computing network.

**4.3.1. Trustednet Ident Class.** The *trustednet ident class* (see Algorithm 1) creates an identity card for the node. It contains most of the information about a node and the encryption algorithms that are going to be used by the node. The default constructor generates a timestamp, public key, and a private key.

After the neighbor discovery, the different nodes have to send messages towards their neighbors. These messages contain the identity card of the node. When the neighbor receives this card, it checks if there is nothing changed in the card. This can be checked by calculating the hash of the card. If this new calculated hash is the same as the hash which is included in the message, then nothing is changed.

**4.3.2. Trustednet Node Class.** It is shown in Algorithm 2. Objects of this class type are nodes. The program/class which uses this class can declare and initialize the node objects. These objects contain a group of neighbors and an identity and the location coordinates.

```

class trustednet_Key
{
    private:
        int privateKey;
    public:
        int publicKey;
        trustednet_Key(void):privateKey(rand()%1000), publicKey(rand()%1000){};
        void writeKey(std::ostream %os) const;
};

```

ALGORITHM 3: Program code of the class trustednet key.

```

Typedef map<int,trustednet_Node * > Hash;
Class trustednet_Graph
{
    private:
        Hash hash;
        Double getDistance(trustednet_Node* source,
                           trustednet_Node* destination) const;
    public:
        trustednet_Graph();
        trustednet_Node* getNode(int_id);
        vector<int> dijkstra(trustednet_Node *source) const;
        void writeGraph(std::ostream & os) const;
        void addNode(trustednet_Node *node);
};

```

ALGORITHM 4: Program code of the class trustednet graph.

The *double\* coordinates* variable stands for the location coordinates of the node. The *Neighbor\*\* neighbors* variable is an array of pointers that contains the neighbors of the node. *nNeighbors* contains the number of neighbors and *identity* contains the ID card. The method *test()* generates a new hash out of the data fields and compares it to the SHA-1 hash.

**4.3.3. Trustednet Key Class.** Trustednet key class has two keys the private and the public one. They are generated when the default constructor is called. The generated keys can have a value up to 999. This is a very basic class. But, it is specially designed for integrating different key generation algorithms. Algorithm 3 shows the code of this class.

**4.3.4. Trustednet Graph Class.** Objects of this class type are graphs. It is shown in Algorithm 4. The class which uses this class can declare and initialize graph objects. The graph contains nodes. These nodes are connected by neighbors. The main purpose of this class is to estimate which neighbor it has to send the message to (till the destination receives the message). It performs the routing protocol tasks. The *hash* variable contains a pointer to the network nodes. There is no need to delete a node that does not have neighbors or trusted neighbors because it is a pointer. The *dijkstra* method calculates the path to the nodes it can reach.

Some designed helper methods that are not included before are the following ones. The method *onReceiveMessage* takes a look at the type of the received message. If this is a public key sending or returning a message, then we check if there was no data loss or change in the identity card. If it is a public key message, then the node sends back a public key return message. When the first node receives a public key return message and there is no data loss or changes, it sets the neighbor as trusted. When there is a broadcasting message, the node calls to the *updateNeighborTable* method. The method *send2NetworkDataPacket* is two times declared in the source with different parameters. That happens because the messages for broadcasting are different from those for sending public keys. Mainly they are doing the same. They are setting/getting the data of a trustednet DataPacket message. Afterwards, the node sends the message to the destination node updates *neighborTable* by using the method *updateNeighborTable*. These messages are used to discover the neighbor nodes. The table holds an entry for every node in the network if it receives a message from that node. When all the messages are broadcasted, it checks the *neighborTable*. It will only accept a node as a neighbor if the amount of received messages is above a certain threshold. For instance, we can declare the threshold for every node on 95%, or, in case of 3 nodes in the network, we can declare an independent threshold for each node. So, in this case, node 0 will accept a neighbor if it receives 25% of the broadcast messages.

## 5. Performance Simulation

Castalia 2 is a wireless sensor network simulator based on the OMNeT++ 3 platform [35]. It can be used by developers and researchers who want to test their algorithms and protocols with a realistic node behavior and wireless channel radio model. It is very important to perform the most accurate channel characterization in order to reproduce the system operation in real environments [36]. It can also be used to evaluate different platform characteristics for specific applications. Because it is highly tunable and it can simulate a wide range of platforms. The main features of Castalia are advanced channel/radio model based on empirically measured data, detailed state transition for the radio, highly flexible physical process model, sensing device noise, bias and power consumption, node clock drift and CPU power consumption, and resource monitoring, allowing the design of medium access control protocols with a large number of parameters to tune. Castalia lets us easily implement and import our designed algorithms and protocols while making use of the features provided by the simulator. The modularity, reliability, and speed of Castalia are partly enabled by OMNeT++, which is an excellent framework for building event-driven simulators.

**5.1. Simulation Parameters.** Castalia simulator uses objects from OMNeT++. In Castalia, each model has 3 different methods: initialize(), handle message(), and finish() methods. They are called in this sequence for every single node. In the next subparagraphs, we will talk about different Castalia modules and how to configure them. Each parameter of these modules has a different meaning. All these parameters are initialized in different files. The channel model used is the log shadowing wireless channel model which gives the power loss in dB, given the distance of two nodes  $d$  and a few parameters. Based on the power loss and the transmission power of a transmitter, we can calculate the power of the signal received at a receiver. By knowing the noise or interference at this receiver, we can calculate the signal to noise ratio or signal to interference ratio, SNR or SIR. Castalia allows us to dynamically calculate the interference from different transmitting nodes and thus dynamically calculate the SNR's or SIR's and the resulting packet reception probabilities.

The radio module tries to capture many features of a real generic low power radio, which is used in wireless sensor network platforms. The following parameters like noise, Bandwidth, modulationType, and encodingType affect the probability of reception. Another parameter is noise floor, which depends on temperature and bandwidth. The receiverSensitivity gives the sensitivity of the receiver. Other parameters are rxPower, listenPower, and sleepPower or transmission parameters like txPowerLevels and txPowerConsumptionPerLevel.

There is a separate module for the medium access control. There are different interesting parameters. The dutyCycle parameter is the fraction of the time that a node listens to the channel. The listenInterval is the time the node stays on listening. Knowing the duty cycle, we can then define the amount of time the node sleeps. If a node is sleeping,

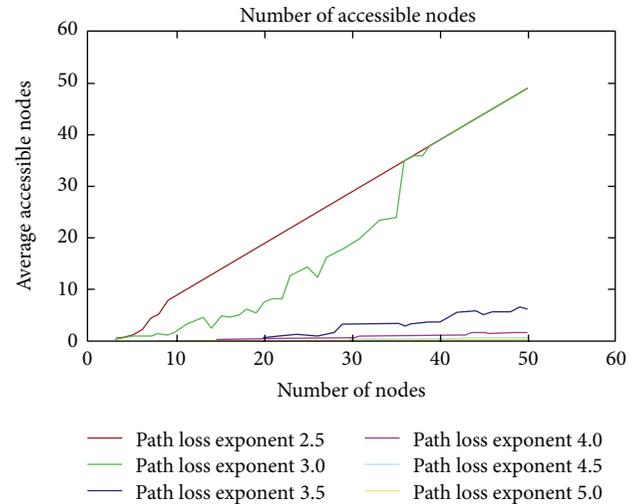


FIGURE 9: Path loss exponent.

then the BeaconIntervalFraction lets us wake up this node. BackOffType let us put the transmission back-off for some time if the channel is not clear and puts the radio to sleep mode.

Castalia lets us work in two models, linear model and nonlinear model. In a linear model, there are 3 different states: (1) a node can be impossible to reach because of a not trusted state, (2) a node can be directly reached if the destination is the node itself or a direct neighbor of the node, (3) and a node has to bypass a message towards another node if the destination can be reached through a neighbor. The nonlinear model uses the Castalia built-in generator. There are three deployments: uniform random deployment, grid deployment, and randomized grid deployment (grid + noise).

**5.2. Performance Results.** Because we want the most accurate results with large number of nodes, we ran the model several times and we provide the most important results in the wireless channel and the MAC layer parameters.

**5.2.1. Wireless Channel.** This section discusses what happens with our result of our model when we adjust the most important parameters of the wireless channel.

We use path loss exponent of a transmitter to calculate the power of the received signal. Figure 9 shows the number of the average accessible nodes as a function of number of nodes using different path loss exponent. This simulation is tested in a 100 m by 100 m area. As we know, if the path loss exponent decreases, the power of the signal received at the receiver increases. Taking into account this fact, we can see that the lowest path loss exponent presents the highest average number of accessible nodes. However, the highest path loss exponent shows very small average number of accessible nodes.

The PL  $d_0$  is the known path loss at a reference distance  $d_0$ . It let us know the initial signal power for each node, also called equivalent isotropically radiated power (EIRP) or, alternatively, effective isotropically radiated power. For

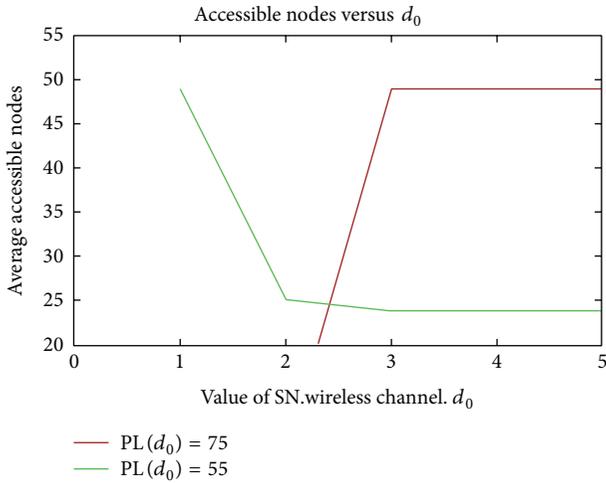


FIGURE 10: Reference distance.

the following test, we used  $PL d_0 = 55$  dBm and 75 dBm (which simulates nodes with different transmission power or different antenna gains). The test increases the reference distance with the same known path loss. We used 50 nodes in an area of 100 m by 100 m. Our results are shown in Figure 10.

5.2.2. *Medium Access Control.* The parameter BackOffType has a close connection with carrier sensing. It means that before a node starts to transmit a message and before it starts to transmit potential beacons, it checks with the radio to see if the channel is clear. If the channel is not clear, the node has to back off for some time. We have different backOffTypes (Figure 11). Using value 1, the back-off time is constant and it is defined by the BackoffBaseValue parameter. Using value 2, the back-off time has a multiplying value, for example,  $1 * a$ ,  $2 * a$ ,  $3 * a$ ,  $4 * a$ , ... We back off for (BackoffBaseValue) \* (times). Using a value that is equal to 3, the back-off time is an exponential value (e.g., 2, 4, 8, 16, 32, ...). As we can see in Figure 11, the exponential value is the better than the multiplying value and the constant value. We used the additive interference model to simulate this behavior.

In the following test, we increase the BackOffBaseValue parameter and look at the difference between the backOff types. For all these tests, we used the additive interference model to simulate this behavior. Figure 12 shows different BackOffBaseValues (from 0 to 256) for BackOffType 1. Meanwhile, there are not too much difference between BackOffBaseValue 32 and BackOffBaseValue 64, and there is a clear difference between BackOffBaseValue 128 and BackOffBaseValue 256, although we can find some peaks in BackOffBaseValue 128 which have higher values than BackOffBaseValue 256.

Figure 13 shows different BackOffBaseValues (from 0 to 256) for BackOffType 2. We can see that except BackOffBaseValue 0, the rest tend to have similar average accessible nodes when the number of nodes increases. The values obtained in this case for the average accessible nodes are higher than those for BackOffType 1.

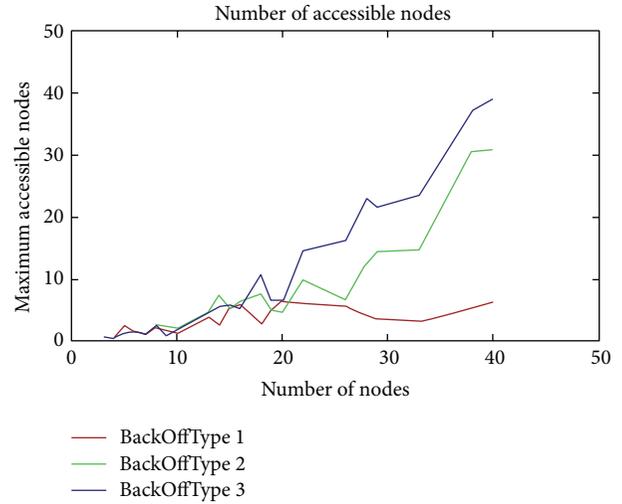


FIGURE 11: Maximum accessible nodes as a function of the number of nodes.

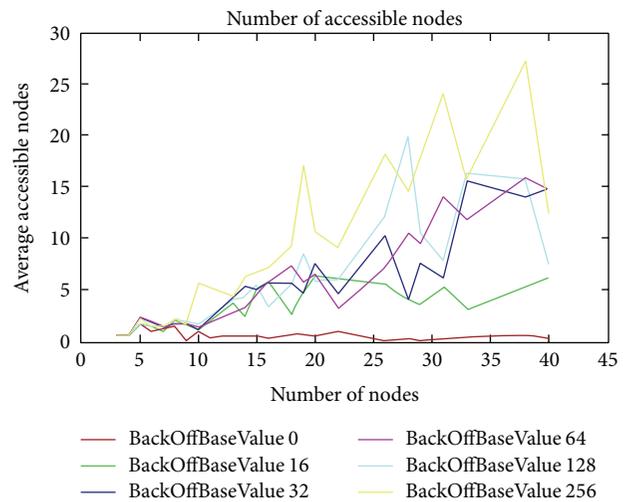


FIGURE 12: BackOffType = 1.

Figure 14 shows the obtained graphs for different BackOffBaseValues (from 0 to 256) for BackOffType 3. We obtain similar average accessible node values than for BackOffType 2, but in this case the tend of the graphs of BackOffBaseValues are parallel when the number of nodes increases (except for BackOffBaseValue 0).

We have observed that if the BackOffBaseValue gets very high, there is almost no difference between the number of neighbors. But when the BackOffBaseValue is low, for instance, between 0 and 16, then there is a big difference between the amount of neighbors.

Castalia uses a random time offset when a node decides to transmit something instead of transmitting it immediately (because it helps to avoid collisions). Now, we made a test of the randomTxOffset parameter. We used again the additive interference model. We observed that the randomTxOffset

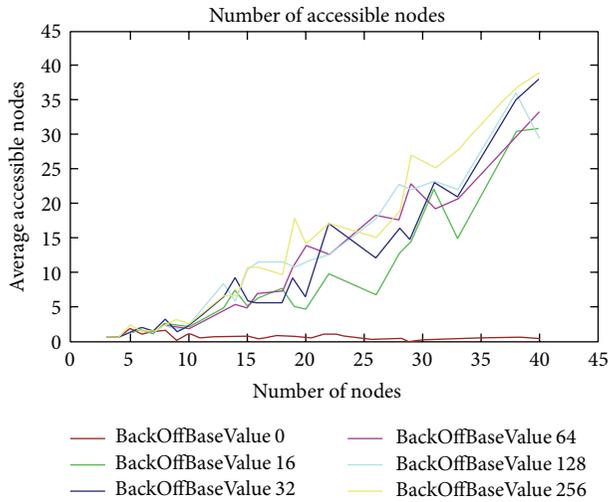


FIGURE 13: BackOffType = 2.

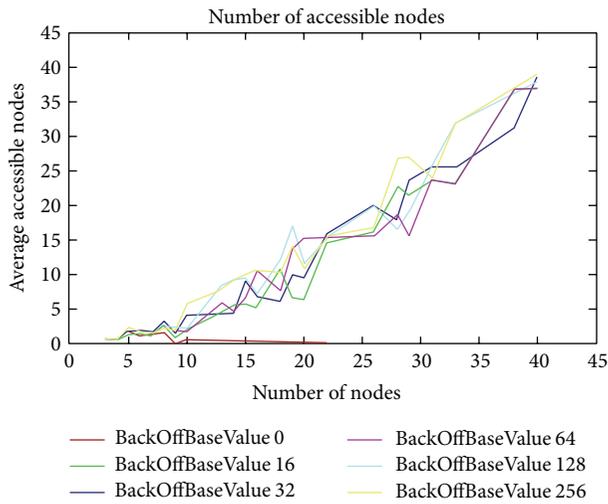


FIGURE 14: BackOffType = 3.

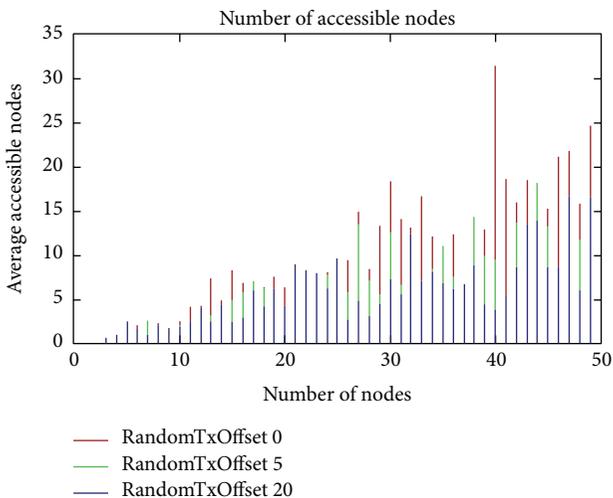


FIGURE 15: RandomTxOffset.

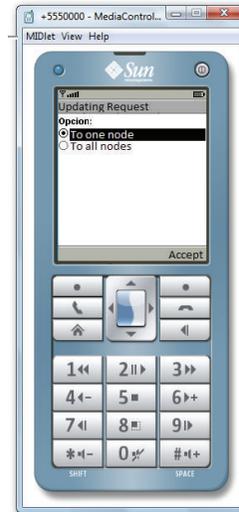


FIGURE 16: Request menu.



FIGURE 17: Image in the mobile when there is a request to one node.

parameter scores better with a value of 0 than in higher randomTxOffset. Figure 15 shows the results of our test.

### 6. Deployment

A prototype to simulate the creation of a virtual cloud computing platform using a spontaneous network has been developed. When the network is created, users can update the network information, sharing resources and services, by asking other nodes. They can ask all network nodes or just one specific node. The receiver node reply to the request with the information requested or it can decide not to share and deny the request.

Figures 16 and 17 show the designed windows to request some network information. A user can decide to request the data only for a specific network node or for all network

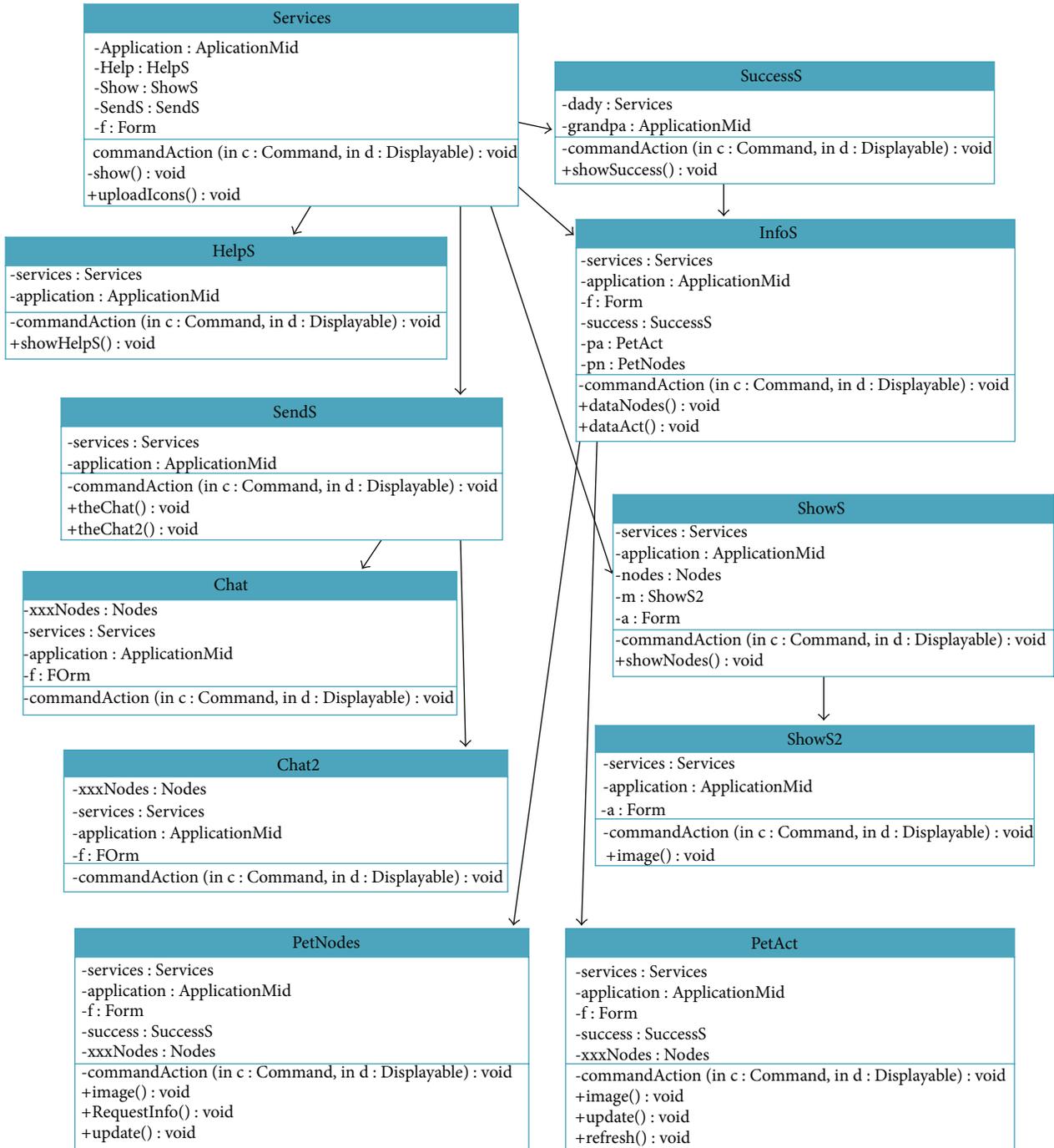


FIGURE 18: Request diagram class.

nodes. If the user selects to request the data for one node, the user must choose the node to request the data. When a node receives the request, it can decide to reply or not to the request.

Figure 18 shows the diagram class of the request process. HelpS class helps the user with the use of the services menu. SuccessS class informs the user if the process has been processed successfully or if there has been a failure. SendS class offers the user the option to choose between

sending the request to one node or to all network nodes. Chat Class lets the user send messages to a node. Chat2 Class lets the user send messages to all network nodes. ShowS Class shows the user the list of the trusted network nodes. ShowS2 class shows the detailed data of one node. InfoS class offers the user a menu to choose a request. PetNodes Class manages the information request about all network nodes. PetAct class manages the information request of one network node.

## 7. Conclusion

Mobile cloud computing networks allow mobile users to share computing resources and applications. In this paper, we proposed a trusted algorithm for creating spontaneous ad hoc mobile cloud computing network. We have developed and tested some algorithms that allow managing the nodes that join and leave the spontaneous ad hoc network. In order to guarantee the network security and the reliability of the communications and transmitted data, we have also developed a trusted algorithm. This algorithm is based on the advanced encryption standard (AES) algorithm and it has implemented a symmetric encryption scheme with simple key management features. We have also deployed the communication protocol procedures and the designed classes. Finally, using Castalia simulator in the OMNeT++ 3 platform, we have implemented a prototype to simulate the creation of a mobile cloud computing system using a spontaneous ad hoc network.

From our results, we can see that, in some cases, as the number of nodes in network increases, the network performance is slightly reduced. However, there are combinations of parameters that maintain the performance level in very promising values.

As future work, we would like to include secure processes based on trust mechanisms and analyze the delay of the secure procedures (proposed in our system) versus the procedures without security systems. Moreover, we will compare the simulation results with real values. In future works, we will test our system in unsecure public cloud environments [37].

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

- [1] S. Preu and C. H. Cap, "Overview of spontaneous networking—evolving concepts and technologies," in *Proceeding of the Workshop on Future Services for Networked Devices (FuSeNetD '99)*, Heidelberg, Germany, 1999.
- [2] O. Bello, A. Bagula, and H. A. Chan, "Automatic network service discovery and selection in virtualization-based future internet," *Network Protocols and Algorithms*, vol. 4, no. 2, pp. 5–29, 2012.
- [3] S. Gallo, L. Galluccio, G. Morabito, and S. Palazzo, "Rapid and energy efficient neighbor discovery for spontaneous networks," in *Proceedings of the 7th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '04)*, pp. 8–11, Venice, Italy, October 2004.
- [4] J. J. P. C. Rodrigues, L. Zhou, L. D. P. Mendes, K. Lin, and J. Lloret, "Distributed media-aware flow scheduling in cloud computing environment," *Computer Communications*, vol. 35, no. 15, pp. 1819–1827, 2012.
- [5] A. H. Mohsin, K. A. Bakar, A. Adekiigbe, and K. Z. Ghafoor, "A survey of energy-aware routing protocols in mobile Ad-hoc networks: trends and challenges," *Network Protocols and Algorithms*, vol. 4, no. 2, pp. 82–107, 2012.
- [6] L. M. Feeney, B. Ahlgren, and A. Westerlund, "Spontaneous networking: an application-oriented approach to ad hoc networking," *IEEE Communications Magazine*, vol. 39, no. 6, pp. 176–181, 2001.
- [7] K. Z. Ghafoor, K. A. Bakar, M. A. Mohammed, and J. Lloret, "Vehicular cloud computing: trends and challenges," in *Mobile Networks and Cloud Computing Convergence for Progressive Services and Applications*, pp. 262–274, IGI Global, 2013.
- [8] H. Modares, J. Lloret, A. Moravejosharieh, and R. Salleh, "Security in mobile cloud computing," in *Mobile Networks and Cloud Computing Convergence for Progressive Services and Applications*, pp. 79–91, IGI Global, 2013.
- [9] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: a survey," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 84–106, 2013.
- [10] J. Lloret, L. Shu, R. Lacuesta, and M. Chen, "User-oriented and service-oriented spontaneous ad hoc and sensor wireless networks," *Ad-Hoc and Sensor Wireless Networks*, vol. 14, no. 1-2, pp. 1–8, 2012.
- [11] A. K. Gupta, H. Sadawarti, and A. K. Verma, "Performance analysis of AODV, DSR & TORA routing protocols," *IACSIT International Journal of Engineering and Technology*, vol. 2, no. 2, pp. 226–231, 2010.
- [12] P. Bellavista, A. Corradi, and C. Giannelli, "The real ad-hoc multi-hop peer-to-peer (RAMP) middleware: an easy-to-use support for spontaneous networking," in *Proceedings of the 15th IEEE Symposium on Computers and Communications (ISCC '10)*, pp. 463–470, Riccione, Italy, June 2010.
- [13] J. H. Christensen, "Using RESTful web-services and cloud computing to create next generation mobile applications," in *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, pp. 627–633, ACM, October 2009.
- [14] M. Mani, A. Nguyen, and N. Crespi, "SCOPE: a prototype for spontaneous P2P social networking," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops '10)*, pp. 220–225, Mannheim, Germany, March-April 2010.
- [15] L. Liu, J. Xu, N. Antonopoulos, J. Li, and K. Wu, "Adaptive service discovery on service-oriented and spontaneous sensor systems," *Ad-Hoc and Sensor Wireless Networks*, vol. 14, no. 1-2, pp. 107–132, 2012.
- [16] G. Kirby, A. Dearle, A. Macdonald, and A. Fernandes, "An approach to ad hoc cloud computing," <http://arxiv.org/abs/1002.4738>.
- [17] G. Huerta-Canepa and D. Lee, "A virtual cloud computing provider for mobile devices," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing and Services: Social Networks and Beyond (MCS '10)*, June 2010.
- [18] V. L. Divya, "Mobile application with cloud computing," *International Journal of Scientific and Research Publications*, vol. 2, no. 4, 2012.
- [19] R. Lacuesta, J. Lloret, M. Garcia, and L. Peñalver, "A secure protocol for spontaneous wireless Ad Hoc networks creation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 4, pp. 629–641, 2013.
- [20] R. Lacuesta, J. Lloret, M. Garcia, and L. Peñalver, "Two secure and energy-saving spontaneous ad-hoc protocol for wireless mesh client networks," *Journal of Network and Computer Applications*, vol. 34, no. 2, pp. 492–505, 2011.
- [21] J. Lloret, R. Lacuesta, M. Garcia, and L. Pealver, "A spontaneous ad hoc network to share www access," *Eurasip Journal on*

- Wireless Communications and Networking*, vol. 2010, Article ID 232083, 2010.
- [22] R. Lacuesta, G. Palacios-Navarro, C. Cetina, L. Peñalver, and J. Lloret, "Internet of things: where to be is to trust," *Eurasip Journal on Wireless Communications and Networking*, vol. 2012, article 203, 2012.
- [23] R. Lacuesta and L. Peñalver, "Automatic configuration of ad-hoc networks: establishing unique IP link-local addresses," in *Proceedings of the International Conference on Emerging Security Information, Systems and Technologies (SECURWARE '07)*, Valencia, Spain, October 2007.
- [24] F. Schatz, S. Koschnicke, N. Paulsen, and M. Schimmler, "Master/Slave assignment optimization for high performance computing in an EC2 cloud using MPI," *Network Protocols and Algorithms*, vol. 4, no. 1, pp. 22–33, 2012.
- [25] S. Capkun, L. Buttyán, and J.-P. Hubaux, "Self-organized public-key management for mobile ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 2, no. 1, pp. 52–64, 2003.
- [26] T. Czachorski and F. Pekergin, "Diffusion approximation as a modeling tool in congestion control and performance evaluation," in *Proceedings of the 2nd International Working Conference Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs '04)*, West Yorkshire, UK, July 2004.
- [27] A. S. Wandert, N. Gura, H. Eberle, V. Gupta, and S. C. Shantz, "Energy analysis of public-key cryptography for wireless sensor networks," in *Proceedings of the 3rd IEEE International Conference on Pervasive Computing and Communications (PerCom '05)*, pp. 324–328, Kauai Island, Hawaii, USA, March 2005.
- [28] N. R. Potlapally, S. Ravi, A. Raghunathan, and N. K. Jha, "Analyzing the energy consumption of security protocols," in *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED '03)*, Seoul, Republic of Korea, August 2003.
- [29] J. Goodman and A. Chandrakasan, "An energy efficient reconfigurable public-key cryptography processor architecture," in *Cryptographic Hardware and Embedded Systems—CHES 2000*, vol. 1965 of *Lecture Notes in Computer Science*, pp. 175–190, Springer, 2000.
- [30] S. Landau, "Communications security for the twenty-first century: the advanced encryption standard," *Notices of the American Mathematical Society*, vol. 47, no. 4, pp. 450–459, 2000.
- [31] A. Kumar, A. Aggarwal, and C. Charu, "Performance analysis of MANET using elliptic curve cryptosystem," in *Proceedings of the 14th International Conference on Advanced Communication Technology (ICACT '12)*, pp. 201–206, February 2012.
- [32] R. Mayrhofer, F. Ortner, A. Ferscha, and M. Hechinger, "Securing passive objects in mobile ad-hoc peer-to-peer networks," *Electronic Notes in Theoretical Computer Science*, vol. 85, no. 3, pp. 105–121, 2003.
- [33] FIPS 180-1-Secure Hash Standard, "SHA-1. National Institute of Standards and Technology," <http://www.itl.nist.gov/fipspubs/fip180-1.htm>.
- [34] J. Rumbaugh, I. Jacobson, and G. Booch, *Unified Modeling Language Reference Manual*, Addison-Wesley Professional, Boston, Mass, USA, 2nd edition.
- [35] Castalia Website, <http://castalia.research.nicta.com.au/index.php/en/>.
- [36] L. D. P. Mendes, J. J. P. C. Rodrigues, J. Lloret, and S. Sendra, "Cross-layer dynamic admission control for cloud-based multimedia sensor networks," *IEEE Systems Journal*, vol. 8, no. 1, pp. 235–246, 2013.
- [37] R. Dutta and B. Annappa, "Protection of data in unsecured public cloud environment with open, vulnerable networks using threshold-based secret sharing," *Network Protocols and Algorithms*, vol. 6, no. 1, pp. 58–75, 2014.

## Research Article

# Indoor Positioning in Wireless Local Area Networks with Online Path-Loss Parameter Estimation

Luigi Bruno,<sup>1</sup> Paolo Addresso,<sup>2</sup> and Rocco Restaino<sup>2</sup>

<sup>1</sup> German Aerospace Center (DLR), Institute of Communications and Navigation, P.O. Box 1116, 82230 Oberpfaffenhofen, Germany

<sup>2</sup> DIEM, University of Salerno, Via Giovanni Paolo II No. 132, 84084 Fisciano, Italy

Correspondence should be addressed to Paolo Addresso; [paddresso@unisa.it](mailto:paddresso@unisa.it)

Received 8 March 2014; Accepted 21 June 2014; Published 4 August 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 Luigi Bruno et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Location based services are gathering an even wider interest also in indoor environments and urban canyons, where satellite systems like GPS are no longer accurate. A much addressed solution for estimating the user position exploits the received signal strengths (RSS) in wireless local area networks (WLANs), which are very common nowadays. However, the performances of RSS based location systems are still unsatisfactory for many applications, due to the difficult modeling of the propagation channel, whose features are affected by severe changes. In this paper we propose a localization algorithm which takes into account the nonstationarity of the working conditions by estimating and tracking the key parameters of RSS propagation. It is based on a Sequential Monte Carlo realization of the optimal Bayesian estimation scheme, whose functioning is improved by exploiting the Rao-Blackwellization rationale. Two key statistical models for RSS characterization are deeply analyzed, by presenting effective implementations of the proposed scheme and by assessing the positioning accuracy by extensive computer experiments. Many different working conditions are analyzed by simulated data and corroborated through the validation in a real world scenario.

## 1. Introduction

Indoor positioning has been drawing remarkable interest since it is pivotal in location based services (LBS), such as visitors monitoring for security issues, automated navigation to points of interest, and customized advertising for pedestrians in malls [1, 2]. The need for local, low cost, and reliable technologies arises from the inaccuracy of satellite based navigation system indoor [3]. Wireless communication technologies, like wireless local area networks (WLANs), represent a valid alternative for their pervasive presence; moreover, the use of received signal strengths (RSSs) obtained from the beacon signals does not affect privacy issues because it does not require exchange of sensitive information.

The complexity of indoor environments has a deep impact on radio propagation, since reflection and diffraction of the radio waves on surfaces and edges make the field propagation highly random. Furthermore, since WLANs usually operate at frequencies between 2 GHz and 5 GHz, interaction with small objects causes time-variant scattering, causing diffraction and multipath contributions, which generate

slow or fast fading effects, respectively [4, 5]. A further technological problem, which affects the performances of positioning algorithms, is intercalibration: different receivers have different antenna gains, thus requiring a calibration procedure that is specific for each employed device [6, 7]. The criticality of this step has attracted a relevant number of contributions of the recent devoted scientific literature [6–12].

The harshness of the indoor propagation channel modeling has endorsed the development of positioning techniques based on scene analysis (or fingerprinting), which use an empirical representation of the field emitted by the transmitting access points (APs), constituting the radio map (RM) of the environment. To this aim, an offline stage is usually performed for measuring RSS at a number of known positions (an independent localization system is required in this phase). During localization, RSS measurements collected at the unknown position are compared to the RM, allowing inferring the user location through a deterministic or probabilistic rule [13]. RADAR is the most famous fingerprinting algorithm which simply adopts RSS mean values of the RM and is shown to achieve positioning accuracy down to 2-3

meters in office buildings [14]. Although these results are very appreciable, the construction of the RM makes the algorithm hardly scalable with the size of the building and, above all, variability of radio propagation should be accounted in order to make the algorithm robust. In [6] the RM is periodically corrected under the arbitrary assumption that the change is uniform across the area. A more flexible system is proposed in [15] which makes use of model trees to adapt the RM online by using RSS measurements at some reference points and without assuming explicit transformation functions. More recently, [11, 16] propose the use of projections techniques to extract features from the RM, which can be more easily updated during the online stage. Focusing on the related problem of intercalibration in [9] develops a solution for addressing the incoherence of the RM with the current operating conditions based on a transformation function, whose training online causes a transient in the algorithm performance (1-2 minutes in the proposed real scenarios).

Although the cited techniques can alleviate the variability issue at the cost of a moderate increase of complexity, the need for an on-site training of the RM still represents the principal drawback of fingerprinting approaches. The development of methods exploiting a theoretical propagation model constitutes the unique possibility of avoiding this demanding step. In this case the key phase is represented by an accurate statistical characterization of the RSS. The *path-loss model*, based on Friis formula, is a very addressed representation for radio propagation and consists in an additive model (in decibel) composed of a deterministic part, accounting for the mean intensity and a zero-mean random term.

The first factor is completely specified by two parameters: the transmitted power, which depends also on the antenna gains, and the path-loss exponent, which describes the decay of signal intensity with distance [17]. The sensitivity of positioning algorithms to errors on path-loss parameters and different solutions for the setting of this crucial quantities has been explored in several papers, for example, in [18], where an empirical study based on RSS measurements in the IEEE 802.11.b network is proposed. Some authors focus on the sole path-loss exponent, with the aim of optimizing least squares position estimation methods [8, 19] or of mitigating the impact of its uncertainty in the spring-relaxation algorithm [20].

The second term of the path-loss model characterizes the random nature of the RSS. Accordingly, it is commonly used to describe the principal corruption effects due to the indoor propagation channel disturbances and in particular the fading effects due to diffraction and reflection phenomena. A widespread model consists in employing a Gaussian distribution for describing the additive random contribution to the RSS in dB. Actually this model, whose success is especially due to its mathematical tractability, is particularly suited for describing the received signal intensity in the presence of slow fading, which corresponds to a Lognormal distribution of the RSS in linear measure units [4]. On the other side, the Gaussian hypothesis is often unrealistic, as it happens, for example, when fast fading effects are present [4]. According to this observation, some authors have dropped out the parametric functional description of the statistical

model, resorting to an approach based on a demanding kernel-based density estimation method [21].

In this paper we develop a sequential Bayesian localization algorithm, aimed at reducing the effect of the inaccurate propagation model knowledge, which commonly affects the indoor positioning problem. The Bayesian scheme constitutes the recursive implementation of the maximum a posteriori probability approach [22] for the estimation of the whole mobile user trajectory. Exploiting the correlation between successive positions has been proven useful also for fingerprinting approaches [23], but the Bayesian scheme represents the most used framework for encompassing this information [24]. The objective of this work is to improve the applicability of this approach by incorporating an estimation phase, which is able to adapt the algorithm to different working scenarios, without requiring a preliminary training phase. This last goal distinguishes the method described in this paper from similar contributions that aim at jointly estimating the user position and the path-loss parameters [25].

More in detail, the proposed algorithm allows to keep on tracking parameters online, by simultaneously estimating the user's trajectory and the path-loss parameters for all APs. The method is based on a particle filter implementation [26], whose suitability for indoor localization was already shown in [27]. In a previous study, we have developed and tested a simple joint Bayesian algorithm, in which the unknown parameters were added to the state space and sampled from a fictitious Gaussian process [28]. In this paper we develop a more advanced localization algorithm based on the Rao-Blackwellized Particle Filter [29]. In this paper we only deal with one parameter of the transmitted power, this way accounting for time-varying obstacles and intercalibration, while the path-loss exponent is approximated by the free space value. This fits several empirical studies, which evidence the appropriateness of affine transformations for modeling the intensities differences between the various devices, and, more specifically, the similarity of the experienced power decay coefficients [7, 30]. Indeed, the latter parameter is mainly influenced by the propagation characteristics of the specific environment, thus resulting essentially independent of the user equipment.

A second main contribution of this paper concerns the extension of the proposed Bayesian algorithm to non-Gaussian statistical model. More specifically, a general approximate approach for implementing the Rao-Blackwellization scheme is presented. This generalization is applied to the crucial case of fast fading, for which the statistical model based on the Rice (or Nakagami-n) distribution is employed [4].

The paper is composed of the following. In Section 2 we detail the state space dynamic system employed for describing the user motion and the observed signal, with particular focus on the statistical characterization of the RSS. In Section 3 the Bayesian approach to the simultaneous estimation of state and parameters is presented, while the computer simulations, performed to analyze the performance of the proposed scheme for adaptive indoor positioning, are shown in Section 4. In Section 5 the results are validated on a real scenario (an indoor parking lot). Final remarks and

further lines of research arising from this study are reported in Section 6.

## 2. State and Observation Models

The algorithm proposed in this paper estimates the location of a mobile user based on the RSS measurements. More specifically, we use a Bayesian sequential approach, which tracks the user during walk by using several scans of RSS. A crucial step of Bayesian approaches is the choice of suitable statistical models for both mobile user kinematics and received signals, which are required to yield an accurate description within an affordable mathematical framework.

In this work we ignore the vertical coordinate of the user position, which is thus encoded in the two-dimensional vector  $\theta \in \mathcal{R}^2$ . The movement is described according to a discrete linear nearly constant velocity model (NCVM), sampled at the time instants  $k\tau$  [31]

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + \mathbf{v}_k, \quad k = 0, 1, 2, \dots, \quad (1)$$

in which the state  $\mathbf{x}_k$  is the 4-dimensional vector composed of the user's position and velocity

$$\mathbf{x}_k = \left[ \theta_k^T, \dot{\theta}_k^T \right]^T, \quad (2)$$

where the superscript  $T$  indicates the transposition operator and  $\mathbf{v}_k$  are the samples of a zero-mean white process, henceforth supposed Gaussian. In (1) the  $4 \times 4$  matrix  $F$  is defined like

$$F = \begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix} \otimes \mathbf{I}_2, \quad (3)$$

having for simplicity introduced the identity matrix  $\mathbf{I}_2$  of order 2 and the Kronecker product  $\otimes$ . The covariance matrix  $Q$  of the noise  $\mathbf{v}_k$  is

$$Q = E[\mathbf{v}_k \mathbf{v}_k^T] = \sigma_v^2 \begin{pmatrix} \frac{1}{3}\tau^3 & \frac{1}{2}\tau^2 \\ \frac{1}{2}\tau^2 & \tau \end{pmatrix} \otimes \mathbf{I}_2, \quad (4)$$

where  $\sigma_v^2$  is the noise variance and multiplies all entries. In other terms, the velocity changes over a sampling period  $\tau$  are of the order of

$$\sqrt{Q_{22}} = \sigma_v \sqrt{\tau}. \quad (5)$$

In particular, the key assumption of the NCVM is that the expected velocity variations are much smaller than the actual velocity. Finally, at  $k = 0$ , we assign a known prior distribution  $p_{\mathbf{x}}(\mathbf{x}_0)$  to the state.

The mobile user device collects signals transmitted by  $N_{AP}$  APs, which are deployed in the environment in known positions  $\theta_j^{AP} \in \mathcal{A}$ ,  $j = 1, \dots, N_{AP}$ . Several statistical models are available in the technical literature for describing the amplitudes  $r_{j,k}$  of the radio signal emitted by the  $j$ th AP and received by the user at instant  $k$  [32]. We selected two credited

models that are able to describe the most common signal degradations. In the case of slow fading, the conditional probability density function (pdf) of  $r_{j,k}$  is well described by a Lognormal distribution [4]

$$r_{j,k} \sim p_L(r) = \frac{1}{\sqrt{2\pi}\sigma_{j,k}r} \exp\left(-\frac{(\ln r - \mu_{j,k})^2}{2\sigma_{j,k}^2}\right), \quad (6)$$

where  $\mu_{j,k}$  and  $\sigma_{j,k}$  are the pdf parameters, dependent on the distance between user and AP; instead, fast fading is better fitted by the Rice (or Nakagami- $n$ ) distribution [4]

$$r_{j,k} \sim p_R(r) = \frac{2(1+K_f)r}{\Omega_{j,k}} \exp\left(-K_f - \frac{(K_f+1)r^2}{\Omega_{j,k}}\right) \cdot I_0\left(2r\sqrt{\frac{K_f(K_f+1)}{\Omega_{j,k}}}\right), \quad r \geq 0, \quad (7)$$

whose parameters are  $K_f \geq 0$  and  $\Omega_{j,k} = E[r^2]$  and  $I_0(\cdot)$  is the zero-th order modified Bessel function of the first type. In detail,  $K_f$  is related to the signal-to-noise ratio of the received signal and is reported to assume values in the range  $K_f \in [0, 20]$  [33].

In both cases, by expressing the amplitudes in dBm,

$$y_{j,k} = 20 \log_{10} r_{j,k}, \quad (8)$$

the noise becomes an additive component. In the slow fading case the measurements in dBm follow a Gaussian pdf, with mean and variance

$$E[y_{j,k}] = \kappa \mu_{j,k}, \quad (9)$$

$$\text{VAR}[y_{j,k}] = \kappa^2 \sigma_{j,k}^2, \quad (10)$$

where  $\kappa = 20/\ln(10)$ .

In the case of fast fading, the conditional pdf of the RSS in dBm is

$$\begin{aligned} y_{j,k} \sim p_R(y) &= \frac{2(1+K_f)}{\kappa} \\ &\times \exp\left(\frac{2(y - \Omega_{dB,j,k})}{\kappa} - K_f\right) \\ &\quad - (K_f + 1) \exp\left(\frac{2(y - \Omega_{dB,j,k})}{\kappa}\right) \\ &\cdot I_0\left(2\sqrt{K_f(K_f+1)} \exp\left(\frac{y - \Omega_{dB,j,k}}{\kappa}\right)\right), \end{aligned} \quad (11)$$

where  $\Omega_{dB,j,k} = (\kappa/2) \ln(\Omega_{j,k})$  is a shift parameter and affects only the expectation

$$E[y_{j,k}] = \Omega_{dB,j,k} - e(K_f), \quad (12)$$

$$\text{VAR}[y_{j,k}] = v(K_f). \quad (13)$$

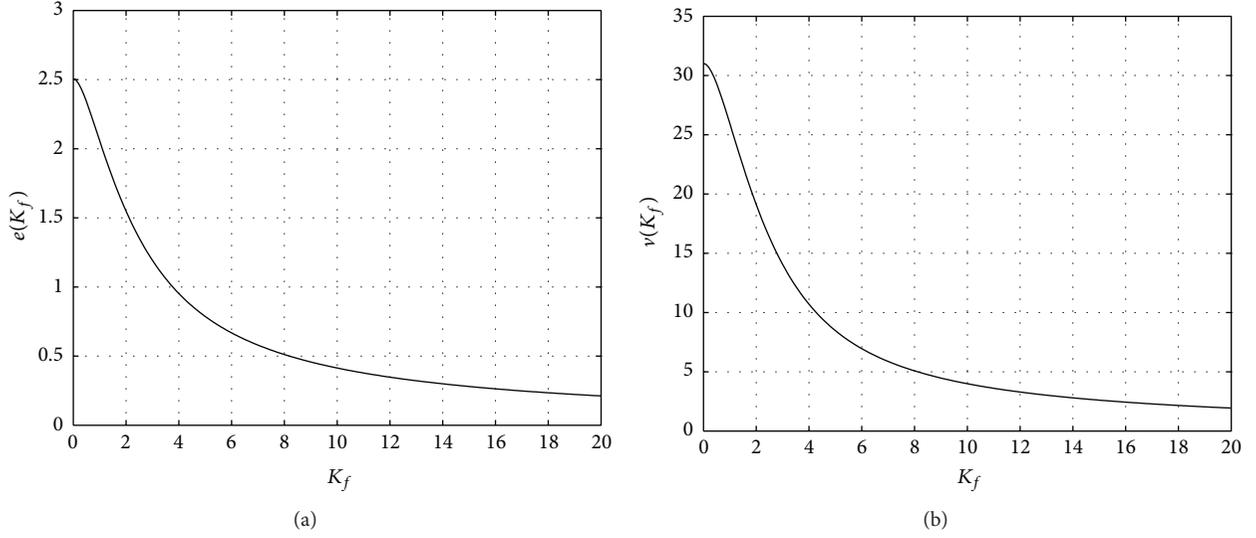


FIGURE 1: Plot of the functions  $e(K_f)$  (a) and  $v(K_f)$  (b) versus  $K_f$ .

The functions  $e(K_f)$  and  $v(K_f)$  can be numerically evaluated and they are depicted in Figure 1 in a typical range for  $K_f$ .

These considerations motivate the use of the additive model for the observations, since for both slow and fast fading cases the RSS measurement can be written as

$$y_{j,k} = E[y_{j,k}] + n_{j,k}, \quad (14)$$

where  $E[y_{j,k}]$  is given by either (9) or (12) and  $n_{j,k}$  are the zero-mean observation noises, which are supposed to be independent among the APs. In the slow fading case  $n_{j,k}$  is a zero-mean Gaussian variable with variance given by (10), while in the presence of fast fading  $n_{j,k}$  is distributed like in (11), but with zero mean, as obtained by setting  $\Omega_{dB,j,k} = e(K_f)$ . The expected RSS value  $E[y_{j,k}]$  models the average attenuation experienced by the strength of the signal emitted by  $j$ th AP at a given distance  $d_j = \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_j^{\text{AP}}\|$ . Its value in dBm is commonly described through the path-loss model [17]

$$E[y_{j,k}] \cong h_j - 20\alpha_j \log_{10} \left( \frac{d_j}{d_0} \right), \quad (15)$$

where  $d_0$  is a reference distance and  $h_j$  and  $\alpha_j$  are static parameters denoting the RSS value at distance  $d_0$  and the path-loss decay exponent, respectively.

By defining the observation vector  $\mathbf{y}_k = [y_{1,k}, \dots, y_{1,N_{\text{AP}}}]^T$ , the observation noise vector  $\mathbf{n}_k = [n_{1,k}, \dots, n_{1,N_{\text{AP}}}]^T$  and the nonlinear functions  $\mathbf{g}(\mathbf{x}_k) = -20\alpha_j \log_{10}(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_j^{\text{AP}}\|/d_0)$ , (15) can be put in a vector form

$$\mathbf{y}_k = \mathbf{g}(\mathbf{x}_k) + \mathbf{h} + \mathbf{n}_k, \quad (16)$$

which evidences the linear dependence of observations  $\mathbf{y}_k$  on the parameters  $\mathbf{h}$  and the nonlinear dependence of  $\mathbf{y}_k$  on the state  $\mathbf{x}_k$ .

### 3. Online Sequential Bayesian Estimation of State and Parameters

We assume an incomplete knowledge of the path-loss model (15). More in detail, the parameters  $h_j$ ,  $j = 1, \dots, N_{\text{AP}}$ , in (15) are unknown, while the decay exponents  $\alpha_j = \alpha$  are set to a fixed value. This formalization fits realistic situations in which the AP's transmitted powers or, more frequently, the sensitivity of the receiving antennas is unavailable.

The localization of the mobile user is here recast within the sequential Bayesian framework, which aims at estimating its whole trajectory by means of the observations acquired at successive instants. Besides motion and observation models, we provide a fictitious probabilistic model to the parameter vector  $\mathbf{h}$ , based on the identity transition matrix with the addition of noise [34]

$$\mathbf{h}_{k+1} = \mathbf{h}_k + \mathbf{r}_k, \quad (17)$$

in which  $\mathbf{r}_k$  is assumed to be a Gaussian white noise with zero mean and a suitable covariance matrix  $R_k^r$ .

Summing up, the dynamics of the faced localization problem can be resumed by the dynamic system

$$\mathbf{x}_{k+1} = F\mathbf{x}_k + \mathbf{v}_k, \quad (18)$$

$$\mathbf{h}_{k+1} = \mathbf{h}_k + \mathbf{r}_k, \quad (19)$$

$$\mathbf{y}_k = \mathbf{g}(\mathbf{x}_k) + \mathbf{h}_k + \mathbf{n}_k, \quad (20)$$

with priors  $p_0(\mathbf{x}_k)$  and  $p_0(\mathbf{h})$  at time 0.

As estimator of the mobile user trajectory we use the maximum a posteriori probability (MAP) estimate, given the available observations. In other terms, it consists in maximizing, at each instant  $k$ , the posterior pdf of the user's trajectory  $\mathbf{x}_{0:k} = [\mathbf{x}_0, \dots, \mathbf{x}_k]$ , given the RSS measurements  $\mathbf{y}_{1:k} = [\mathbf{y}_1, \dots, \mathbf{y}_k]$ , namely, in finding

$$\hat{\mathbf{x}}_{0:k} = \arg \max_{\mathbf{x}_{0:k}} p(\mathbf{x}_{0:k} | \mathbf{y}_{1:k}). \quad (21)$$

The calculation of the state posterior pdf at  $k$ , given the observed data

$$p(\mathbf{x}_{0:k} | \mathbf{y}_{1:k}), \quad (22)$$

can be obtained through the recursive factorization:

$$p(\mathbf{x}_{0:k} | \mathbf{y}_{1:k}) \propto p(\mathbf{y}_k | \mathbf{x}_{0:k}, \mathbf{y}_{1:k-1}) \cdot p(\mathbf{x}_k | \mathbf{x}_{0:k-1}, \mathbf{y}_{1:k-1}) \cdot p(\mathbf{x}_{0:k-1} | \mathbf{y}_{1:k-1}), \quad (23)$$

which is a straightforward consequence of the Bayes theorem. The term

$$p(\mathbf{x}_k | \mathbf{x}_{0:k-1}, \mathbf{y}_{1:k-1}) = p(\mathbf{x}_k | \mathbf{x}_{k-1}) \quad (24)$$

is completely specified by the model used to derive (18), which drops the dependence of  $\mathbf{x}_k$  on  $\mathbf{x}_{0:k-2}$  and  $\mathbf{y}_{1:k-1}$ , given  $\mathbf{x}_{k-1}$ ; the last right term of (23) is the posterior pdf at instant  $k - 1$ .

On the other side, the evaluation of the first term of (23), corresponding to the RSS likelihood function, requires the marginalization over  $\mathbf{h}$ :

$$\begin{aligned} p(\mathbf{y}_k | \mathbf{x}_{0:k}, \mathbf{y}_{1:k-1}) &= \int p(\mathbf{y}_k | \mathbf{x}_k, \mathbf{h}_k) p(\mathbf{h}_k | \mathbf{x}_{0:k}, \mathbf{y}_{1:k-1}) d\mathbf{h}_k \\ &= \int p(\mathbf{y}_k | \mathbf{x}_k, \mathbf{h}_k) p(\mathbf{h} | \mathbf{x}_{0:k-1}, \mathbf{y}_{1:k-1}) d\mathbf{h}_k. \end{aligned} \quad (25)$$

In the last line we dropped the dependence of the parameter pdf on the current state since the corresponding measurement is missing. Evaluation of the integral (25) constitutes the key point of the adopted Bayesian procedure and highlights the dependence of the MAP user trajectory on the parameters distribution.

**3.1. Existing Approaches.** Dual and joint estimation algorithms constitute the most diffuse approaches to Bayesian estimation in the presence of unknown parameters. The first one consists in running two interacting concurrent algorithms, one devoted to the state estimation and another devoted to the parameters [35]. Instead, joint estimation is performed by constructing a single augmented state vector including both the kinematic quantities, namely, the position and the velocity of the mobile user, and the unknown parameters [36].

Classical Bayesian approaches for state and parameters estimation rely upon the use of Kalman filters (KFs), which are optimal for linear dynamical systems corrupted by Gaussian noise. Extended KFs (EKFs), achieved after the linearization of the equations, are a suitable solution also in the presence of nonlinear models, for both dual and joint estimation methods [34].

More accurate implementations of Bayesian algorithm for general dynamic equations are constituted by Monte Carlo schemes, which are commonly referred to as *particle filters* [26]. In this approach an empirical approximation of the posterior pdf, consisting in a summation of delta measures centered at a finite set of support points (or particle),

is employed to simplify the computation of the Bayesian procedure recursions. Application of particle filtering to the joint estimation of position and propagation parameters of a mobile user connected to a WLAN has been tested in a previous contribution by the authors [28]. In this paper the sequential importance sampling with resampling (SIR) scheme [37] has been employed, underlying the capabilities of the methods, but evidencing, at the same time, its drawbacks. The most critical issue is surely related to the augmentation of the state space dimensionality; this is due to the addition of the parameters to the vector of estimating quantities. This implies the exponential growth of the particles number, required to preserve an adequate particle density within the state space.

**3.2. Rao-Blackwellized Particle Filter.** In this paper we attain the solution of the Bayesian problem through a different approach, which exploits the Rao-Blackwell Theorem to reduce the state estimation error by means of the parameter marginalization [29]. More in detail, the *Rao-Blackwellized Particle Filter (RBPF)* consists in applying the Monte Carlo approximations only for the state estimation and in deriving the parameter pdf through analytical procedures, instead. This is done to avoid including the parameters in the state space, which, therefore, keeps a constant dimensionality. Accordingly, the main hypothesis required for its utilization consists in the availability of a deterministic algorithm to recursively compute the parameter conditional pdf. A noticeable case is represented by parameters evolving, given the state, according to a *conditionally linear Gaussian (CLG)* system.

According to the Monte Carlo rationale, the user's state posterior pdf at  $k$  is written as

$$p(\mathbf{x}_k | \mathbf{y}_{1:k-1}) = \sum_{i=1}^N w_i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \quad (26)$$

in which  $\delta(\mathbf{x}_k - \mathbf{x}_k^i)$  denotes the delta measure centered at the support point  $\mathbf{x}_k^i$  and  $w_i$  is the corresponding weight. In the SIR scheme adopted in this work, the  $i$ th particle  $\mathbf{x}_k^i$  is obtained by sampling the state space according to the predictive pdf:

$$\mathbf{x}_k^i \sim p(\mathbf{x}_k | \mathbf{x}_{0:k-1}^i, \mathbf{y}_{1:k-1}) = p(\mathbf{x}_k | \mathbf{x}_{k-1}^i), \quad (27)$$

which is replaced, at the initial time  $k = 0$ , by the prior

$$\mathbf{x}_0^i \sim p(\mathbf{x}_0). \quad (28)$$

For  $k > 1$ , the particle weights are obtained in a recursive way by following the factorization illustrated in (23), namely, as

$$\begin{aligned} w_k^i &= w_{k-1}^i \cdot p(\mathbf{y}_k | \mathbf{x}_{0:k}^i, \mathbf{y}_{1:k-1}), \\ &= w_{k-1}^i \cdot \int p(\mathbf{y}_k | \mathbf{x}_k^i, \mathbf{h}) \\ &\quad \times p(\mathbf{h} | \mathbf{x}_{0:k-1}^i, \mathbf{y}_{1:k-1}) d\mathbf{h}, \end{aligned} \quad (29) \quad (30)$$

whereas for  $k = 0$  the initial weights are uniformly set to  $w_0^i = N^{-1}$ ,  $\forall i = 1, \dots, N$ . Therefore, in order to completely specify the RBPF algorithm, we need to compute the parameters density function, conditioned on the state trajectory sample,  $\mathbf{x}_{0:k}^i$ , and on data  $\mathbf{y}_{1:k}$ :

$$p(\mathbf{h} \mid \mathbf{x}_{0:k}^i, \mathbf{y}_{1:k}). \quad (31)$$

In the following sections we detail two approaches for calculating  $p(\mathbf{h} \mid \mathbf{x}_{0:k}^i, \mathbf{y}_{1:k})$ , with reference to the observation models presented in Section 2. The former represents an efficient and exact implementation, which fits the CLG model of parameters, as it is the case of Lognormally distributed noise; the latter is a very general method that constitutes an approximated solution exploitable for all nonlinear non-Gaussian (NLNG) models.

**3.3. Lognormal Fading: Continuous Model for the Parameter.** If the RSS likelihood function is assumed to be Lognormal or, equivalently, data in dBm follow a Gaussian distribution, the parameter pdf (31) can be computed by means of the Kalman filter (KF) [34]. In particular, by starting from a Gaussian prior also for the parameter vector, the integrand function in (25) is always the product of two Gaussian distributions. The result is a Gaussian density, except for a normalization constant  $c$ ,  $0 < c < 1$ ,

$$p(\mathbf{y}_k \mid \mathbf{h}, \mathbf{x}_k^i) p(\mathbf{h} \mid \mathbf{x}_{0:k-1}^i, \mathbf{y}_{1:k-1}) = cs(\mathbf{h}), \quad (32)$$

where the Gaussian pdf has been denoted by  $s(\cdot)$  and its mean and variance can be easily obtained. Indeed, one can use the following result: if  $s_i(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}_i, \Sigma_i)$ ,  $i = 1, 2$ , the function  $s(\mathbf{x}) = s_1(\mathbf{x}) \cdot s_2(\mathbf{x})$  is proportional to a multivariate Gaussian pdf with mean and covariance matrix  $\mathbf{m} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mathbf{m}_1 + \Sigma_2^{-1}\mathbf{m}_2)$ ,  $\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$ . Note that the first factor on the left part of (32) is a normal pdf w.r.t.  $\mathbf{y}_k$ : here we further exploit the Gaussianity of  $\mathbf{h}$  which can be easily derived by solving 13 for  $\mathbf{h}$ , yielding

$$\mathbf{h} = \mathbf{g}(\mathbf{x}_k) + \mathbf{y}_k + \mathbf{n}_k. \quad (33)$$

By using (32) in the integral (30) defining the particle weight, we find

$$w_k^i \propto w_{k-1}^i \cdot c, \quad (34)$$

with  $c$  being on turn the ratio

$$c = \frac{p(\mathbf{y}_k \mid \mathbf{h}, \mathbf{x}_k^i) p(\mathbf{h} \mid \mathbf{x}_{0:k-1}^i, \mathbf{y}_{1:k-1})}{s(\mathbf{h})}, \quad (35)$$

which can be calculated at an arbitrary value of the variable  $\mathbf{h}$ , for example, at its expected value. This algorithm will be referred to as RBPF-KF in the following.

**3.4. General Case: Discrete Model for the Parameter.** If the RSS distribution is not Gaussian, as in the fast fading case, we need another method to evaluate the integral of (25). Unfortunately numerical techniques often represent a bottleneck

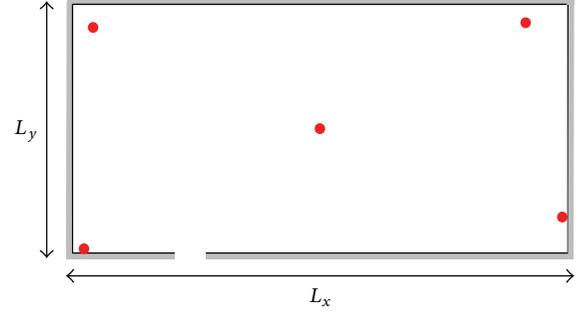


FIGURE 2: Testbed adopted in the simulations; in the figure  $L_x = 40$  m and  $L_y = 20$  m and the APs are in the positions denoted by red circles.

from a computational point of view and, therefore, we use a grid-based approach, that is computationally suitable even for nonlinear and non-Gaussian (NLNG) models. In detail, we decompose the range of variation of the  $j$ th component  $h_{j,k}$  of the vector parameter  $\mathbf{h}_k$  into a finite number  $N_c$  of disjoint cells  $\{H_c\}_{c=1,\dots,N_c}$  and quantize the values within each cell to its mean value, say  $\bar{h}_c$ . In different words, the parameter vector is approximated by a discrete random process  $\tilde{\mathbf{h}}_k$ , whose independent components  $\tilde{h}_{j,k}$  assume values in the set  $\{\bar{p}_c\}_{c=1,\dots,N_c}$  and admit a probability mass function given by

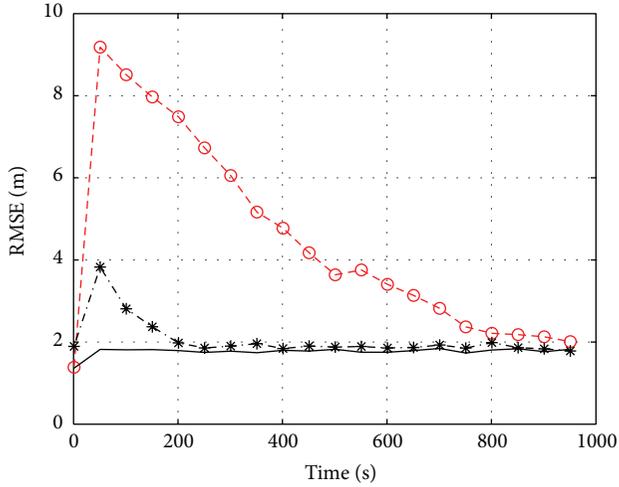
$$\Pr\{\tilde{h}_{j,k} = \bar{h}_c\} = \Pr\{h_{j,k} \in H_c\} = \int_{H_c} p(h_{j,k}) dh_{j,k}. \quad (36)$$

By resorting again to the factorization of the posterior pdf reported in (23), we address the RBPF algorithm, but we compute differently the terms concerning the parameters. In particular, in such hypotheses, the parameter distribution corresponding to the  $i$ th particle is given by the pmf defined in (36). Recursive computation of the above distribution can be performed by means of the *Approximated Grid-Based (AGB)* algorithm presented in [37], which is the counterpart of the KF in a discrete state space. We denote this algorithm by RBPF-AGB.

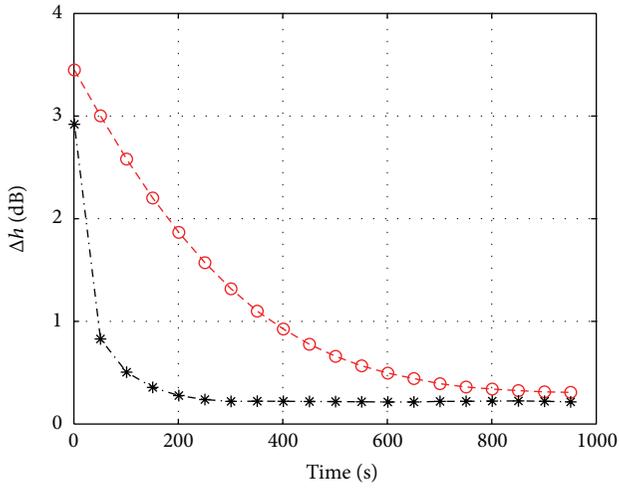
## 4. Computer Experiments

Several simulations were designed in order to analyze the performance of our proposals. We have chosen to separately evaluate the effects of fast and slow fading to avoid combined effects which would be difficult to discriminate. The synthetic testbed, represented in Figure 2, is composed of a  $40 \times 20$  m open area where 5 APs, denoted by red circles, periodically emit beacon signals.

A user walks according to the model described in Section 2 with  $\sigma_v = 0.1$  m/s<sup>2</sup> and  $\tau = 1$  s. The initial state  $\mathbf{x}_0$  is drawn from a multivariate Gaussian (MG) prior distribution with diagonal covariance matrix, whose nonzero terms are set to 1 for the positions ( $\sigma_{\theta_1}^2$  and  $\sigma_{\theta_2}^2$ ) and 0.1 for the velocities ( $\sigma_{v_1}^2$  and  $\sigma_{v_2}^2$ ). The mean RSS is given by the path-loss model described by (15), in which the free space value  $\alpha = 2$  is assumed for all the APs. In particular we draw the starting



(a) Positioning RMSE



(b) Error of the parameter

FIGURE 3: Slow fading effect evaluated by means of computer experiments based on the testbed of Figure 2. JSIR and RBPf-KF are applied with  $\sigma_h^2 = 9$ ,  $\sigma_y^2 = 5$ , and  $N_p = 1000$  particles; we show (a) the positioning RMSE and (b) estimation error  $\Delta h_k$  of the parameters averaged over all APs.

value of  $\mathbf{h}$  from a MG distribution with known mean  $\mathbf{h}_0$  and diagonal covariance matrix with elements  $\sigma_h^2$ ; in some simulation settings a stepwise variation of some component of  $\mathbf{h}$  is also impressed. Finally, all results are averaged over a series of independent experiments and are presented in terms of a numerical evaluation of the positioning RMSE.

4.1. Slow Fading (RBPf-KF). To test the slow fading effects, measurements in dB are drawn according to a multivariate distribution whose components are independent Gaussian

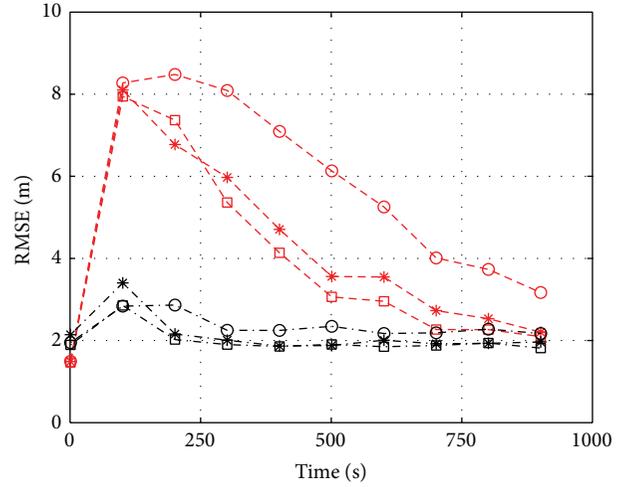


FIGURE 4: Positioning RMSE (both JSIR and RBPf-KF) related to computer experiments concerning slow fading in the setup of Figure 2 with variable number of particles in the range  $N_p = 200 \div 1000$ ; here,  $\sigma_h^2 = 9$ ,  $\sigma_y^2 = 5$ .

random processes with means given by (15) and a common fixed variance  $\sigma_h$ .

Thus, we employ the RBPf-KF and compare its performances with the JSIR approach presented in [28]. The first test is carried out by setting  $\sigma_h^2 = 9$  and, as a reference, we also draw the corresponding performance obtained by the clairvoyant SIR algorithm that is fed up with the true values of the reference power  $\mathbf{h}$ . All algorithms are applied with 1000 particles and their RMSEs, calculated only on the user position, are plotted against time in Figure 3(a). The initial RMSE value is related to the covariance matrix of the state prior; namely,

$$\sqrt{E \left[ \|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_t\|^2 \right]} = \sqrt{\sigma_{\theta_1}^2 + \sigma_{\theta_2}^2} = \sqrt{2}; \quad (37)$$

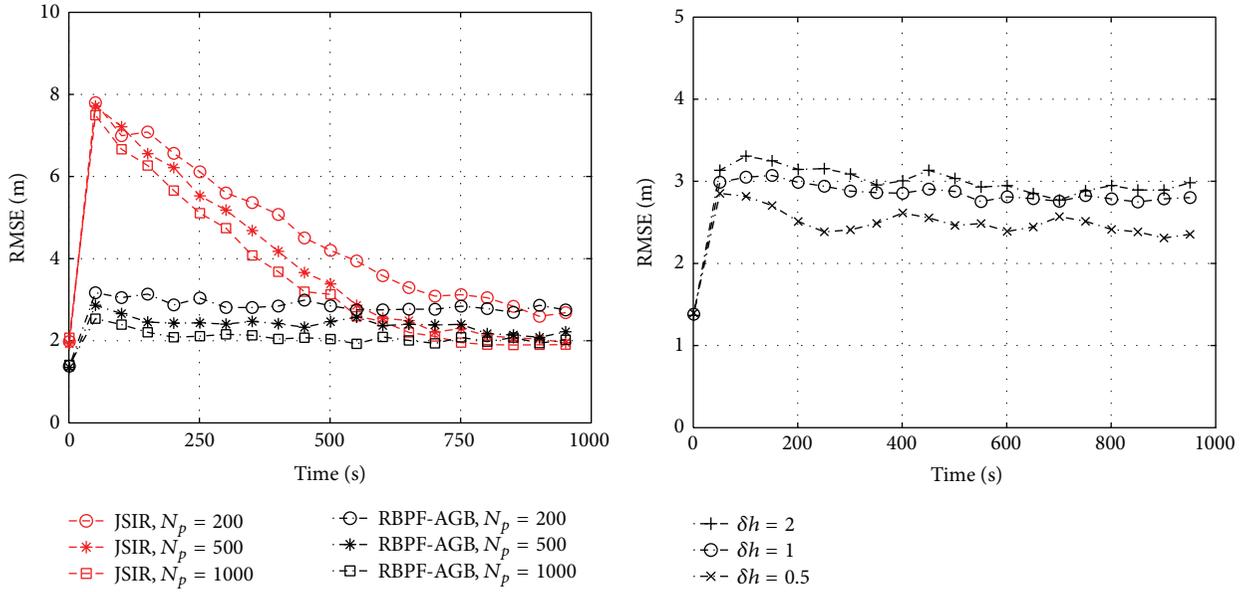
then, both adaptive algorithms are characterized by a transient during which the parameters are estimated; after this phase they attain the same performance shown by the clairvoyant algorithm. The differences between JSIR and RBPf-KF lay in the amplitude of the RMSE overshoot and in the speed of convergence; in both cases relevant benefits are achieved by RBPf. This is a direct consequence of the algorithms adaptivity: as it is shown in Figure 3(b), the error, averaged over all APs, of the estimated reference power  $\hat{\mathbf{h}}$ ,

$$\Delta h_k = \frac{1}{N_{AP}} \sum_{j=1}^{N_{AP}} |h_{j,k} - \hat{h}_{j,k}|, \quad (38)$$

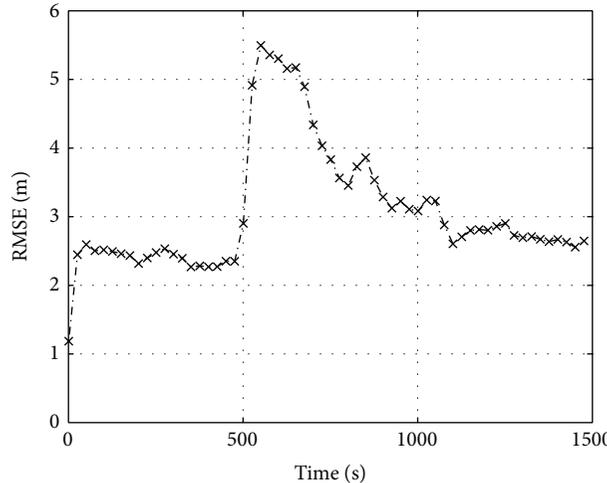
is rapidly torn down in the RBPf case to a steady state value.

Let us dig deeper into the algorithms evaluation. In Figure 4 we show the results of our algorithms applied in





(a) Number of particles (b) Stepsize of the parameter



(c) Step variation

FIGURE 8: Computer experiments concerning fast fading in the testbed of Figure 2; we report the positioning RMSE plotted against time obtained by (a) both JSIR and RBPf-AGB with different sets of particles in the range  $N_p = 200 \div 1000$ , (b) RBPf-AGB with step size in the range  $\delta h = 0.5 \div 2$ , and (c) RBPf-AGB in the presence of a  $-5$  dB step variation on one AP's reference power at the time instant  $k = 500$ ; if not otherwise specified, we use  $\sigma_h^2 = 9$ ,  $\sigma_y^2 = 5$ ,  $N_p = 200$ , and  $\delta h = 1$  dBm.

TABLE 1: Steady state values of the positioning RMSE computed by means of RBPf algorithm (for both KF and AGB implementations) for different values of the measurement variance in the range  $\sigma_y^2 = 1 \div 15$ . Here  $\sigma_h^2 = 9$ ,  $N_p = 500$ , and  $\delta h = 1$  dBm.

	RBPf-KF	RBPf-AGB
$\sigma_y^2 = 3$	1.6961 m	2.2424 m
$\sigma_y^2 = 5$	1.9600 m	2.8125 m
$\sigma_y^2 = 10$	2.4623 m	2.8657 m
$\sigma_y^2 = 15$	2.7777 m	2.9516 m

Thus, we test the RBPf-AGB algorithm, using again the JSIR algorithm as a yardstick. Figure 7 highlights a comparison between our proposals and the clairvoyant SIR algorithm, applied to the testbed of Figure 2 with  $\sigma_h^2 = 9$ ,  $\sigma_y^2 = 5$ , and  $N_p = 1000$ . The RBPf-AGB effectiveness is clearly shown, thanks to a very sharp convergence with respect to JSIR, although the steady state value is slightly higher than that of JSIR. This is due to the discrete set of parameter values assumed in RBPf-AGB, whose choice is key for the algorithm performance. We prefer a uniform sampling of  $\mathbf{h}$  in a suitable set, to account for sudden changes during the estimation. The step size, say  $\delta h$ , can be tuned by considering the full

mismatch case: the maximum difference between the true value of the parameter and the closest discretized value is  $\delta h/2$  and must be lower than the expected error  $\Delta h$ . Since we have found out in the computer experiments that usually  $\Delta h \approx 0.5$  dBm, then we choose

$$\delta h = 1 \text{ dBm} \tag{39}$$

as a suitable balance between algorithm complexity and performance.

The results of the analysis relative to the number of particles ( $N_p = 200 \div 1000$ ), step size ( $\delta h = 0.5 \div 2$ ), and downside variation of one AP's reference power are shown in Figure 8, subplots (a), (b), and (c), respectively. In detail, Figure 8(b) confirms that there is room for improvement by setting a lower  $\delta h$ . The results about variations of  $\sigma_h^2$  and  $\sigma_y^2$  do not present relevant differences compared to the slow fading case. We only report the RMSE steady state values against the measurement variance  $\sigma_y^2$  in the last column of Table 1.

### 5. Real Data Experiments

We assess our algorithms on the testbed already presented in [28, 38] and shown in Figure 9. It is a  $45 \times 40$  m indoor parking lot, one floor below the ground level, in which a 802.11 (WiFi) network with 5 APs 3COM 7760 operates. Thick walls, columns made of concrete, and car dispositions which change rapidly make this environment very challenging for indoor localization. That is why in [38] the RADAR algorithm, in its finer weighted version, is shown to exhibit poor performance (sample RMSEs are not lower than 7 meters). In that case the training set, computed on the base of about 30 measurements per 50 positions distributed all over the parking lot, was filled soon before the online stage. In our methods, instead, we estimated in the training stage the decay exponent  $\alpha$ , the noise variance  $\sigma_y^2$ , and the RSS model. Analysis of measured data reveals an inhomogeneous scenario. As an example, in Figure 10, we show the RSS measured in the target area. The path-loss model defined by (15) is roughly observed with evident fluctuations dependent on the environment configuration. We use for all APs the values  $\alpha = 3$  for the decay exponent and  $\sigma_y^2 = 20$  for the noise variance. We have also observed that the Lognormal model for RSS is dominant and thus we simplify the algorithms by neglecting the fast fading contribution.

The results, presented in Figure 11 for both JSIR and RBPF-KF with different numbers of particles, refer to a 10-minute dataset, acquired along the path shown in Figure 9. They are given in terms of localization RMSE and the ground truth is provided by a set of places known with high accuracy. We can see that both algorithms are convergent to consistent values of the localization RMSE: specifically, RBPF-KF takes less than a minute to achieve errors lower than 6 meters if it is run with at least 250 particles and errors lower than 7 meters if only 100 particles are employed. JSIR is slower, but with 1000 particles its RMSE converges to 6 meters in about 5 minutes.

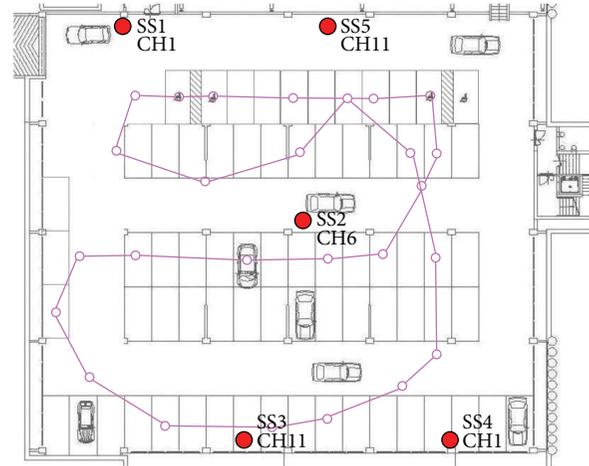
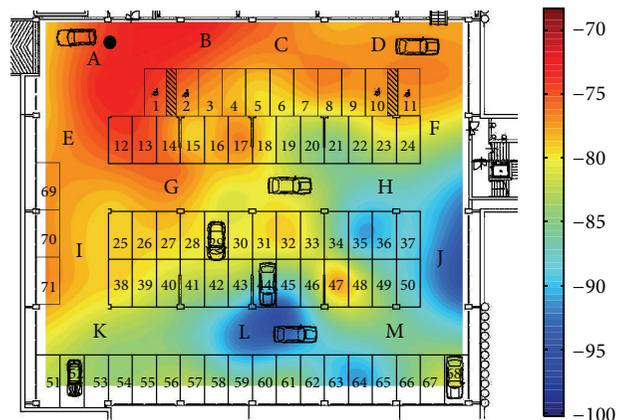
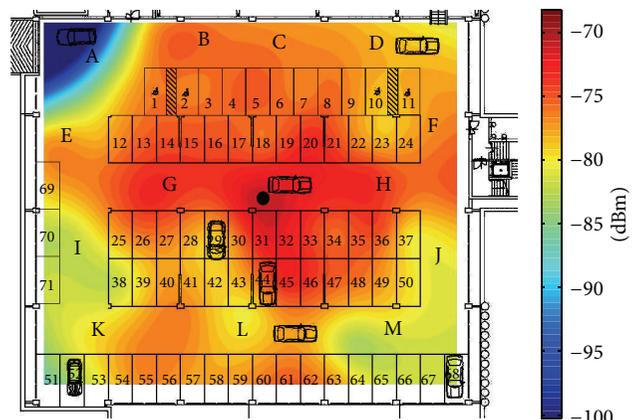


FIGURE 9: Real world experiments: experimental setting used for data collection. The APs are marked by a red circle and indicated with SS*i*,  $i = 1, \dots, 5$ . Also the channel of the 802.11 band used by each AP is indicated. Note that we use only 3 channels in order to avoid interferences, being the distances between SS3 and SS5 and between SS1 and SS4 greater than the APs range in reception. The line in magenta represents the path used in the test and is run several times.



(a) SS1



(b) SS2

FIGURE 10: Mean RSS measured in the parking lot related to AP SS1 (a) and AP SS2 (b).

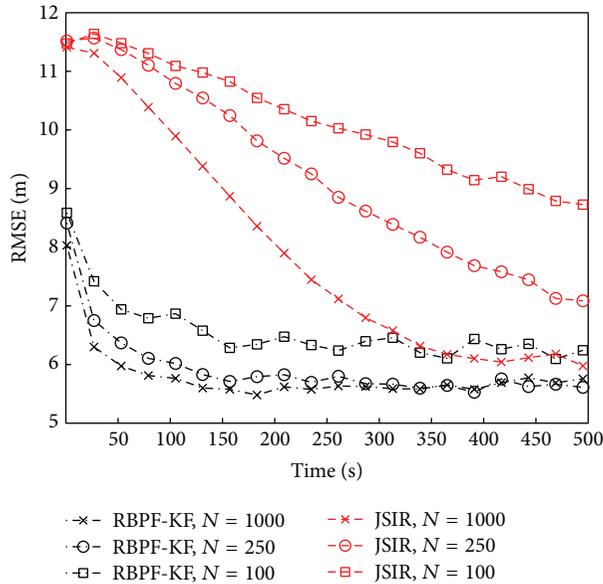


FIGURE 11: RMSE versus time is represented for RBPf-KF (dashed lines) and JSIR (continuous lines) algorithms applied to real data, with reference to the scenario in Figure 9 for different numbers of particles ( $N_p = 50 \div 1000$ ). We recall that RADAR obtains an RMSE of 7 meters in the same conditions.

## 6. Conclusions

Indoor localization employing not perfectly known signals is a challenge still far from a complete solution. We processed RSS measurements, freely available in infrastructured WLANs, by means of an adaptive Bayesian framework which is able to deal with unpredictable effects such as intercalibration and fading. At this aim we referred to simple but very addressed models for signal propagation, whose calibration was carried out online by avoiding time-consuming training stages. Extensive computer experiments and real world data collected in a harsh environment showed the effectiveness of our approaches, evidencing the remarkable convergence properties of the RBPf implementation. A natural continuation of the current work consists in including further propagation parameters within the estimating quantities.

Other future lines of research concern a deeper analysis of the propagation models, aimed at improving the localization accuracy. On the other hand, the development of other tracking techniques that are able to follow other kinds of changes (such as the noise variance) is of paramount importance. A final interesting working case, which will be addressed in the next future, includes the lack of a perfect knowledge of the APs' positions.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

The authors would like to thank both the management and the staff of the Coritel Lab, which have kindly provided the experimental datasets used in this paper.

## References

- [1] S. A. Ahson and M. M. Ilyas, *Location-Based Services Handbook: Applications, Technologies, and Security*, CRC Press, New York, NY, USA, 2009.
- [2] W. Kolodziej and J. Hjelm, *Local Positioning Systems: LBS Applications and Services*, CRC Press, New York, NY, USA, 2006.
- [3] P. Misra and P. Enge, *Global Positioning System, Signals, Measurements, and Performance*, Ganga-Jamuna Press, 2006.
- [4] J. Parsons, *The Mobile Radio Propagation Channel*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2000.
- [5] N. Yarkony and N. Blaunstein, "Prediction of propagation characteristics in indoor radio communication environments," in *Proceedings of the 2nd European Conference on Antennas and Propagation (EuCAP '07)*, pp. 1–9, IET, 2007.
- [6] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom '04)*, pp. 70–84, Philadelphia, Pa, USA, October 2004.
- [7] J.-G. Park, D. Curtis, S. Teller, and J. Ledlie, "Implications of device diversity for organic localization," in *Proceedings of IEEE INFOCOM*, pp. 3182–3190, Shanghai, China, 2011.
- [8] X. Li, "RSS-based location estimation with unknown pathloss model," *IEEE Transactions on Wireless Communications*, vol. 5, no. 12, pp. 3626–3633, 2006.
- [9] A. Tsui, Y. Chuang, and H. Chu, "Unsupervised learning for solving rss hardware variance problem in wifi localization," in *Mobile Network Applications*, pp. 677–691, Springer, New York, NY, USA, 2009.
- [10] G. Lui, T. Gallagher, B. Li, A. G. Dempster, and C. Rizos, "Differences in RSSI readings made by different Wi-Fi chipsets: a limitation of WLAN localization," in *Proceeding of the International Conference on Localization and GNSS (ICL-GNSS '11)*, pp. 53–57, Tampere, Finland, June 2011.
- [11] S. H. Fang, C. H. Wang, S. M. Chiou, and P. Lin, "Calibration-free approaches for robust Wi-Fi positioning against device diversity: a performance comparison," in *Proceedings of the IEEE 75th Vehicular Technology Conference (VTC Spring '12)*, pp. 1–5, June 2012.
- [12] A. K. M. Mahtab Hossain, Y. Jin, W. Soh, and H. N. Van, "SSD: a robust RF location fingerprint addressing mobile devices' heterogeneity," *IEEE Transactions on Mobile Computing*, vol. 12, no. 1, pp. 65–77, 2013.
- [13] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man and Cybernetics C: Applications and Reviews*, vol. 37, no. 6, pp. 1067–1080, 2007.
- [14] P. Bahl and V. Padmanabhan, "RADAR: an in-building rf-based user location and tracking system," in *Proceedings of 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '00)*, pp. 775–784, March 2000.

- [15] J. Yin, Q. Yang, and L. M. Ni, "Learning adaptive temporal radio maps for signal-strength-based location estimation," *IEEE Transactions on Mobile Computing*, vol. 7, no. 7, pp. 869–883, 2008.
- [16] S. Fang and C. Wang, "A dynamic hybrid projection approach for improved Wi-Fi location fingerprinting," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 3, pp. 1037–1044, 2011.
- [17] T. S. Rappaport, *Wireless Communications: Principles and Practice*, IEEE Press, Piscataway, NJ, USA, 1st edition, 1996.
- [18] A. Bose and H. F. Chuan, "A practical path loss model for indoor WiFi positioning enhancement," in *Proceedings of the 6th International Conference on Information, Communications and Signal Processing (ICICS '07)*, pp. 1–5, December 2007.
- [19] S. Mazuelas, A. Bahillo, R. M. Lorenzo et al., "Robust indoor positioning provided by real-time rssi values in unmodified WLAN networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 821–831, 2009.
- [20] Q. Zhang, C. H. Foh, B.-C. Seet, and A. C. M. Fong, "Variable elasticity spring-relaxation: improving the accuracy of localization for WSNs with unknown path loss exponent," *Personal and Ubiquitous Computing*, vol. 16, no. 7, pp. 929–941, 2012.
- [21] J. Prieto, S. Mazuelas, A. Bahillo, P. Fernandez, R. M. Lorenzo, and E. J. Abril, "Adaptive data fusion for wireless localization in harsh environments," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1585–1596, 2012.
- [22] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley & Sons, 2001.
- [23] C. Wu, Z. Yang, Y. Liu, and W. Xi, "WILL: wireless indoor localization without site survey," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 4, pp. 839–848, 2013.
- [24] D. Madigan, E. Elnahrawy, R. P. Martin, W. Ju, P. Krishnan, and A. S. Krishnakumar, "Bayesian indoor positioning systems," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '05)*, vol. 2, pp. 1217–1227, Miami, Fla, USA, March 2005.
- [25] H. Nurminen, J. Talvitie, S. Ali-Loytty et al., "Statistical path loss parameter estimation and positioning using RSS measurements in indoor wireless networks," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '12)*, pp. 1–9, Sydney, Australia, November 2012.
- [26] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-gaussian Bayesian state estimation," *IEE Proceedings F*, vol. 140, no. 2, pp. 107–113, 1993.
- [27] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Bordello, "Bayesian filtering for location estimation," *IEEE Pervasive Computing*, vol. 2, no. 3, pp. 24–33, 2003.
- [28] P. Addesso, L. Bruno, and R. Restaino, "Adaptive localization techniques in WiFi environments," in *Proceeding of the IEEE 5th International Symposium on Wireless Pervasive Computing 2010 (ISWPC '10)*, pp. 289–294, Modena, Italy, May 2010.
- [29] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 176–183, Morgan Kaufmann, 2000.
- [30] Y. Kim, H. Shin, Y. Chon, and H. Cha, "Smartphone-based Wi-Fi tracking system exploiting the RSS peak to overcome the RSS variance problem," *Pervasive and Mobile Computing*, vol. 9, no. 3, pp. 406–420, 2013.
- [31] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*, John Wiley & Sons, New York, NY, USA, 2001.
- [32] H. Hashemi, "Indoor radio propagation channel," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 943–968, 1993.
- [33] L. J. Greenstein, D. G. Michelson, and V. Erceg, "Moment-method estimation of the Ricean  $K$ -factor," *IEEE Communications Letters*, vol. 3, no. 6, pp. 175–176, 1999.
- [34] S. Haykin, *Kalman Filtering and Neural Networks*, John Wiley & Sons, New York, NY, USA, 2001.
- [35] A. T. Nelson, *Nonlinear estimation and modeling of noisy time series by dual kalman filtering methods [Ph.D. thesis]*, 2000, AAI9984703.
- [36] H. Cox, "On the estimation of state variables and parameters for noisy dynamic systems," *IEEE Transactions on Automatic Control*, vol. 9, pp. 5–12, 1964.
- [37] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.
- [38] M. D. Mauro, G. D. Corte, A. L. Robustelli, P. Addesso, and M. Longo, "A WLAN-based location system for indoor parking areas," in *Proceedings of the 17th International Conference on Software, Telecommunications and Computer Networks (SoftCom '09)*, pp. 186–190, Hvar island, Croatia, September 2009.

## Research Article

# Describing the Access Network by means of Router Buffer Modelling: A New Methodology

**Luis Sequeira, Julián Fernández-Navajas, Jose Saldana,  
José Ramón Gállego, and María Canales**

*Communications Technology Group (GTC), Aragón Institute of Engineering Research (I3A), Department of IEC, EINA, University of Zaragoza, Ada Byron Building, 50018 Zaragoza, Spain*

Correspondence should be addressed to Luis Sequeira; [sequeira@unizar.es](mailto:sequeira@unizar.es)

Received 7 March 2014; Accepted 25 May 2014; Published 24 July 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Luis Sequeira et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The behaviour of the routers' buffer may affect the quality of service (QoS) of network services under certain conditions, since it may modify some traffic characteristics, as delay or jitter, and may also drop packets. As a consequence, the characterization of the buffer is interesting, especially when multimedia flows are transmitted and even more if they transport information with real-time requirements. This work presents a new methodology with the aim of determining the technical and functional characteristics of real buffers (i.e., behaviour, size, limits, and input and output rate) of a network path. It permits the characterization of intermediate buffers of different devices in a network path across the Internet.

## 1. Introduction

Traditionally, the available bandwidth, delay, and jitter between two end-to-end devices have been used as a parameter that can give a rough idea of the expected quality of service (QoS). But nowadays, we know that QoS is also affected by the behaviour of the intermediate router buffer, which is mainly determined by its size and its management policies. So, the buffer may cause different packet loss behaviour and may also modify some QoS parameters.

Many multimedia applications and services (e.g., video-conferencing, video streaming, peer-to-peer, and VoIP services) take advantage of various available bandwidth estimations techniques and tools (ABETT) in order to improve some QoS parameters. But all these techniques have one thing in common: they are focused on links estimations at the core network where buffer behaviour and its parameters are not the principal priorities.

In general, buffers are used as a traffic regulation mechanism in network devices. Current core routers make extensive use of AQM (active queue management) disciplines which are able to maintain a shorter queue length than drop-tail queues; this fights against bufferbloat and reduces latency. But these techniques (e.g., RED and SRED) require careful tuning of their parameters in order to provide good performance [1].

There exist QoS scheduling algorithms as weighted fair queuing (WFQ) which is a data packet scheduling technique allowing different scheduling priorities to statistically multiplexed data flows.

However, some network points may become critical bottlenecks, mainly in access networks, because these networks' capabilities are lower than the ones available in the backbone, being the main cause of packet loss of the discarding of packets in router queues. Mid- and low-end routers, which do not implement advanced traffic management mechanisms, are usually used in access networks. In this scenario, SME (small and medium enterprises) environments may be principally affected because of their modest infrastructure. So the design characteristics of router buffers and the implemented scheduling policies are of primary importance in order to ensure the correct delivery of the traffic of different applications and services, so it will be useful to include buffer parameters in the link capacity estimation.

On the other hand, it is true that the performance of TCP (transmission control protocol) has been extensively studied and a big number of variants (SACK, New Reno, Vegas, etc.) have been deployed in order to improve it. Nevertheless, many multimedia applications and real-time

services transport their information under UDP (user datagram protocol). So, the applications have to describe certain network behaviour for optimizing traffic.

Hence, the characterization of the technical and functional parameters of router buffer in SME environments becomes critical when planning a network or trying to provide certain levels of QoS. As a consequence, if the size and the behaviour of the buffer are known, some techniques can be used so as to improve link utilization, for example, multiplexing a number of small packets into a big one, fragmentation, smoothing traffic, and so forth.

A new methodology is presented in this paper in order to describe the access network by means of router buffer modelling (e.g., behaviour, size, limits, and input and output rate) by the use of four simple steps which will be detailed in Section 3.3. In particular we will estimate bandwidth, size, and behaviour because this gives us more useful link information, than only an estimation with ABETT techniques. In addition, this methodology has been deployed in order to solve problems of resource consumption for processing data and inaccuracies when obtaining input rate estimations and buffer size as well.

The paper is organized as follows: Section 2 presents the related work, and Section 3 describes the test methodology. The next section covers the experimental results, and the paper ends with the conclusions.

## 2. Related Work

**2.1. Bandwidth Estimations.** There exist several estimation techniques for obtaining available bandwidth. A performance evaluation of Pathload, Pathchirp, Spruce, IGI, and Abing in a flexible test bed was presented in [2]. The results demonstrate that ABETTs are far from being ready to be applied in all applications and scenarios. The evaluation includes scenarios in which the packet loss rate and the propagation delays of the links have been varied. Also, the amount of cross-traffic, the capacity of the links, and the cross-traffic packet size were tested for different values.

In [3], the authors define probabilistic available bandwidth as the largest input rate at which we can send a traffic flow along a path while achieving, with specified probability, an output rate that is almost as large as the input rate. The method is expressed directly in terms of the measurable output rate and includes adjustable parameters that allow the user to adapt to different application requirements. It was proposed as a new definition for available bandwidth and a novel framework that addresses some issues: provide a confidence interval, be suited to the task of multipath estimation in large-scale networks, and provide enough flexibility in terms of accuracy, overhead, latency, and reliability to adapt to the requirements of various applications.

Moreover, in [4] a novel tool for end-to-end ABETT PRM (probe gap models) called quality monitoring and estimation (QMoEs) was proposed, as a key component embedded in the RUBENS (rethinking the usage of broadband access for experience-optimized networks and services) architecture. The results indicate that QMoEs present a good performance in terms of accuracy and estimation time which validates the ABETT within the RUBENS framework.

**2.2. Buffer Issues.** Buffers are used to reduce packet loss by absorbing transient bursts of traffic when routers cannot forward them at that moment. They are instrumental in keeping output links fully utilised during congestion times.

For many years, researchers accepted the so-called *rule of thumb* (or bandwidth delay product, BDP) to obtain the amount of buffering needed at a router's output interface. This rule was proposed in [5] and it is given by  $B = C \times RTT$ , where  $B$  is the buffer size,  $RTT$  is the average round-trip time, and  $C$  the capacity of the router's network interface. In [6] a reduced buffer size was proposed by dividing the BDP by the square root of the number of TCP flows,  $N$ ,  $B = C \times RTT / \sqrt{N}$ . This model was called *small buffer*. In [7] the use of even smaller buffers, called *tiny buffers*, was suggested considering a size of some tens of packets.

In [8], the authors presented a simple algorithm to manage a drop-tail queue, which adapts its size based on traffic conditions in order to obtain a minimum size and providing high level of utilization. The results show that adaptive drop-tail achieves significantly smaller queues than current approaches at the expense of 1-2% of the link utilization.

The buffer can be measured in different ways: maximum number of packets, amount of bytes, or even queueing time limit [9, 10]. For example, in [11] the routers of two manufacturers were compared, and it was observed that one measures the buffer in packets, whereas the other one does it in milliseconds.

Moreover, the buffer must play an important role when planning a network because it can influence the packet loss of different services and applications. The reason for this is that there is a relationship between router buffer size and link utilization, since an excessive amount of memory would generate a significant latency increment when the buffer is full. On the other hand, a very small amount of memory in the buffer will increase packet loss in congestion time. As a consequence, the knowledge of the buffer behaviour is an interesting parameter which can be considered when trying to improve link utilization.

**2.3. The Influence of the Buffer on Real-Time Services.** The influence of the router buffer on the subjective quality experienced by users of a interactive service (i.e., an online game) with very tight real-time requirements was studied in [4], showing the mutual relationship between the size and policies of the buffer and the obtained subjective quality, which mainly depends on delay and jitter in this case.

The study has been conducted, showing that *tiny buffers* are more adequate in order to maintain game quality in acceptable levels. Since, if it is too big it may add delay and jitter which are not acceptable for gamers.

A popular online, multiplayer, game server was studied in [12]; the paper shows the results of a 500-million packet trace in which the traffic behaviour describes a loaded game server and can be attributed to the fact that current game designs target the saturation of the narrowest, last-mile link. Online games try to provide relatively uniform experiences between all players, maximizing the interactivity of the game, so they fix their usage requirements in such a way as to saturate the network link of their lowest speed players.

The authors also comment that facing the stringent demands on interactivity, routers must be designed with enough capacity to manage such bursts without delay. But current routers are designed for bulk data transfers with larger packets, so a significant deployment of online game servers will have the potential for overwhelming the current networking equipments.

Several studies have characterized P2P video streaming applications and have measured their impact in the communication networks traffic. So, in [13] a traffic characterization of major P2P-TV application was presented and they concluded that the traffic consists in a mixture of small packets (signalling packets) and large packets (video packets). In addition, the generation of high rates of small packets may penalize the video packets and consequently the peer's behaviour within a P2P structure may not be as expected. The application relies on UDP traffic and it is revealed that this application faces a high overhead; about 60% of packets correspond to signalling, and the other 40% correspond to video data packets.

In [14], the results show that the presence of a bursty application (video surveillance) causes packet loss for all the coexisting applications, even for those generating constant bit rate traffic (VoIP). In addition, packet loss decreases when buffer size is increased, because big buffers can absorb the burst produced by the traffic mix. As expected, packet loss increases when link utilization grows in the case of 40-packet buffer (Figure 1).

The tests were deployed in a scenario in which two IP camera flows, one videoconferencing session, and two VoIP calls are used as test traffic in two different tests: in the first one, the Internet access link was set to an average link utilization of 70% and different values of the buffer size were tested. In the second tests, the buffer size of the Internet access router was fixed at 40 packets and different values of the access bandwidth were used, so consequently different levels of link utilization ranged from 50% to 90%.

### 3. Methodology for the Characterization of Internet Paths

The network path (Figure 2) is uncertain for most applications and services which only measure the available bandwidth, in order to limit the generated traffic and rarely to attune it (e.g., trough smoothing it). Thus, we propose a method for applications to discover and get advantage of knowing network characteristics by means of finding the buffer behaviour models which may be useful to correctly attune the traffic.

**3.1. Network Path Model.** Traditionally a network path can be characterized by bandwidth, packet loss, and delay. The premises of this work are that most of the network characteristics can be explained by buffer models. We recommend a characterization by including buffer parameters (size and input and output rate) and buffer behaviour; see Figure 3.

In this work, the buffer model only considers FIFO queues since they are the most common in real access network devices as it was observed in [15]. The same work

also reported that the behaviour of these buffers can be characterized as follows: once the buffer gets completely full, no more packets or bytes are accepted until a certain amount of memory is available; see Figure 4. Thus, an *upper limit* and a *lower limit* can be defined: when the *upper limit* is reached, no more packets or bytes are accepted until the size of the buffer corresponds to the *lower limit*. There are some cases in which the difference between *lower limit* and *upper limit* is as small as one packet; however, this difference sometimes reaches tens of packets. Generally speaking, a buffer drops packets in bursts. The number of packets per burst depends on the relationship between the input and output rates. When these rates are very similar, the number of packets in the burst can be just one packet.

**3.2. Test Procedure.** The scheme of the tests is shown in Figure 2. There is a "system under test" (SUT from now), which may be either a single device or an entire network. The test is based on sending a burst of UDP packets, all with the same length, from the source to the destination machine, so as to produce a buffer overflow in the SUT. All the transmitted packets are identified by a sequence number included in the payload.

**3.3. Methodology.** The methodology is based on the premise that the output rate can be obtained from traffic capture at the destination device. This output rate depends on the technology used in each case (Ethernet, Wi-Fi, Cablemodem, and others). Different buffers can be detected by means of a packet loss pattern analysis when more than one bottleneck is in the path. In most cases, one buffer is the main cause of packet loss in a network path and sometimes two buffers can get into overflow at the same time in an access network; for this reason, we will use as an example the case of two concatenated buffers to present this methodology (Figure 5).

The output rate can be easily determined by the destination trace because we know all arrived packets and the time of each one; see Figure 6. Also, the input rate can be estimated counting the dropped packets in the destination trace by means of the sequence number included in the payload.

A packet loss map is useful not only to determine the amount of losses but also to observe the packet loss patterns which may define each buffer model. It is a simple packet loss scatter (see Figure 7). We should note that the information on this map is equivalent to the buffer occupancy under the same conditions, also a simple way to see the packet loss patterns (hop sequence) which are difficult to deduce from Figure 6.

The methodology consists of four simple steps which will be described as follows:

- Methodology {
    - ⇒ Analyze packet loss patterns
    - ⇒ Determine rates
    - ⇒ Infer locations
    - ⇒ Estimate buffer size.
- (1)

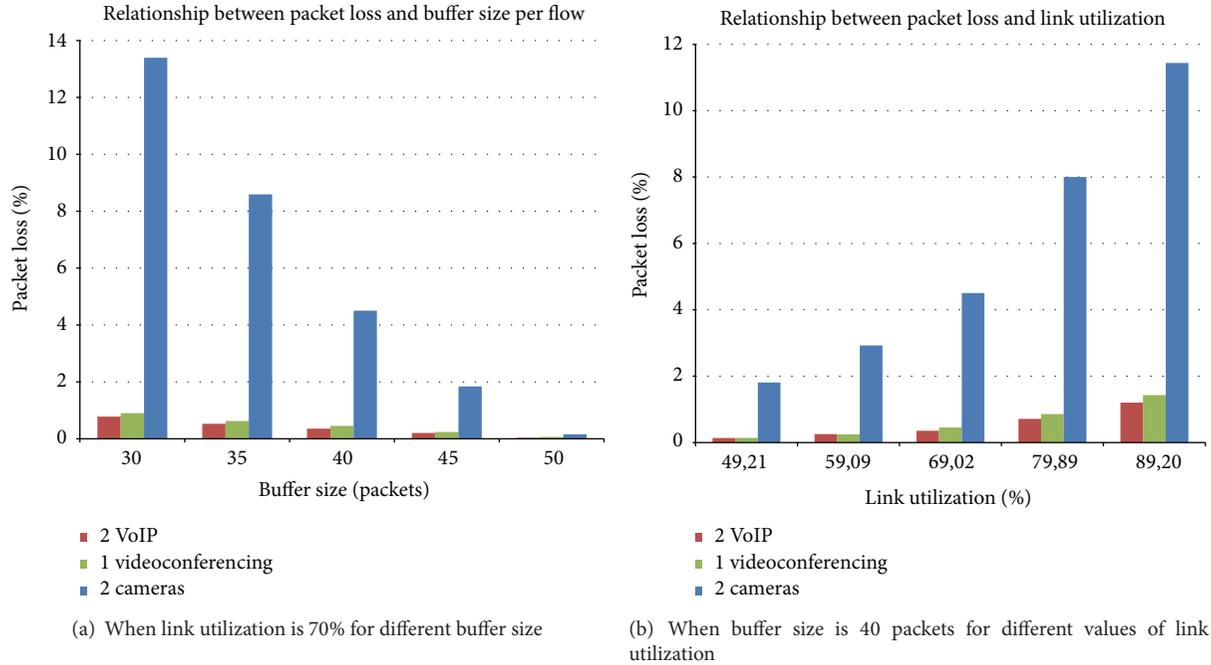


FIGURE 1: Packet loss.

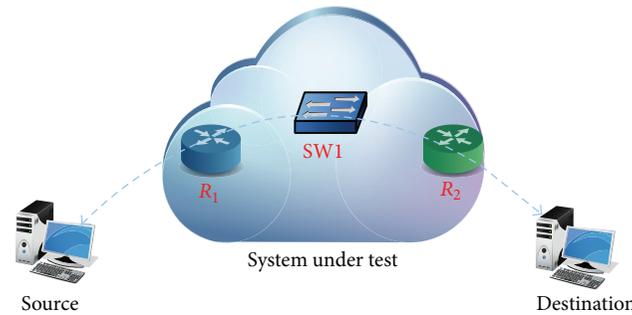


FIGURE 2: Network path and topology used for tests.

3.3.1. *Analyze Packet Loss Patterns.* The analysis consists of determining the number of congested buffers and the packet loss rate of each one. As an example, in Figure 7 we can see two different patterns which correspond to two buffers. The packet loss for the first is 60 packets per burst, while the second only drops 1 packet per burst. Also, we can find the number of dropped packets in a given period of time  $T_{rx}$  for *Buffer 1* ( $N_1$ ) and *Buffer 2* ( $N_2$ ).

3.3.2. *Determine Rates.* The output rate ( $R_3$  in Figure 6) can be calculated with (2), from the destination trace because we know all the arrived packets,  $N$ , and the arrival time of each one; see Figure 6. The input rate ( $R_1$  in Figure 6) is obtained with the arrived packets and all the dropped ones (3). We must estimate the rates in the period in which we observed all the packet loss patterns  $T_{rx}$  (see Figures 6 and 7) since it has been observed that it is the most stable period in terms of output rate and packet loss. Consider

$$R_3 = \frac{P_L}{T_{R_x}} \times N, \tag{2}$$

$$R_1 = \frac{P_L}{T_{R_x}} \times (N + N_1 + N_2), \tag{3}$$

where  $P_L$  is the *packet length*.

If we obtain several packet loss patterns (see Figure 6), we can also calculate intermediate rates. When there are two buffers in the path a general equation (4) can be defined in which two values can be found. The intermediate rate calculation ( $R_2$  in Figure 6) will generate two different possible values. In order to obtain the correct value, a new test with a new generated rate between these two value rates is needed. If the results show (by observing packet loss maps) that both buffers are overflowed then the correct value is the lowest one; otherwise it is the highest. Consider

$$R_2 = \begin{cases} \frac{P_L}{T_{R_x}} \times (N + N_1) \\ \frac{P_L}{T_{R_x}} \times (N + N_2). \end{cases} \tag{4}$$

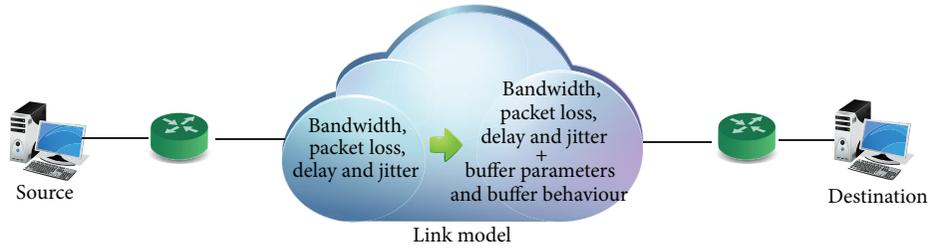


FIGURE 3: Link model parameters.

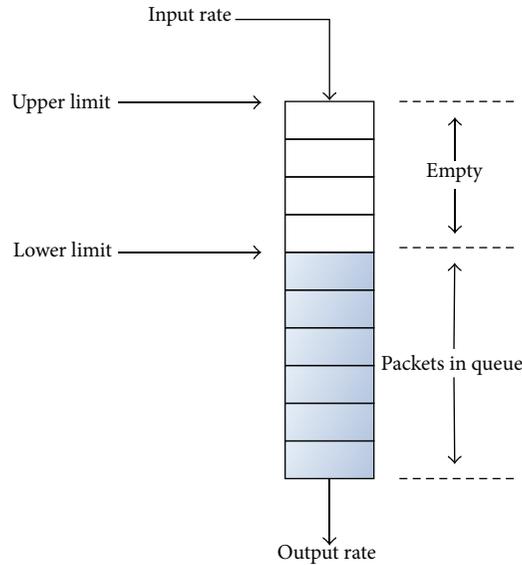


FIGURE 4: A particular buffer behaviour.

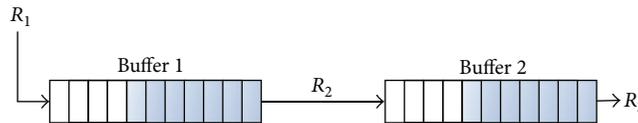


FIGURE 5: Two concatenated buffers.

3.3.3. *Infer Locations.* The packet loss rate is defined by the relationship between input and output rates of any buffer. So, when the packet loss rate of each buffer is known, we can compare the rate with the relationship of the input and output rates ( $R_1$ ,  $R_2$ , and  $R_3$ ) to obtain the buffer position; see Figure 6.

3.3.4. *Estimate Buffer Size.* When we know the buffer input and output rates, then buffer size can be estimated if we find the latency of a packet in the buffer when it is full. In this case, we use the last received packet before the first packet loss, as it is shown in Figure 8, because this is the packet which completely fills the buffer. In addition the sequence number (SN) of this packet  $n$  can give us the number of packets sent in certain time  $T$ . The  $n$ th packet gets into the  $Buffer_m$  in a time:  $T_m = n * P_L/R_m$ , and the output time is  $T_{m+1} = n * P_L/R_{m+1}$ . The packet latency is  $T_{m+1} - T_m$ , so the buffer size, in packets, of the  $m$  buffer can be estimated using (5), which only

depends on the rates relationship and the  $n$  arrived packets before the packet loss in each buffer.

The number of arrived packets to a certain buffer, before the first packet loss, depends on which buffer drops packets first and the physical location of the buffer (see Figure 6) described above. From Figure 6 we know that *Buffer 1* is physically before *Buffer 2*, so the dropped packets by *Buffer 1* will never arrive to *Buffer 2*. From Figure 7 we know that *Buffer 1* is the first one in dropping packets, so we can deduce that  $n_1 = SN_1$  and  $n_2 = SN_2 - N_1$ . The same analysis can be applied when *Buffer 2* is before *Buffer 1*. In this case the dropped packets by *Buffer 2* successfully pass through *Buffer 1*, so  $n_1 = SN_1$  and  $n_2 = SN_2$ . Consider

$$\begin{aligned}
 size_{Buffer_m} &= (T_{m+1} - T_m) \times \frac{R_{m+1}}{P_L} \\
 &= n_m \times \left( 1 - \frac{R_{m+1}}{R_m} \right).
 \end{aligned}
 \tag{5}$$

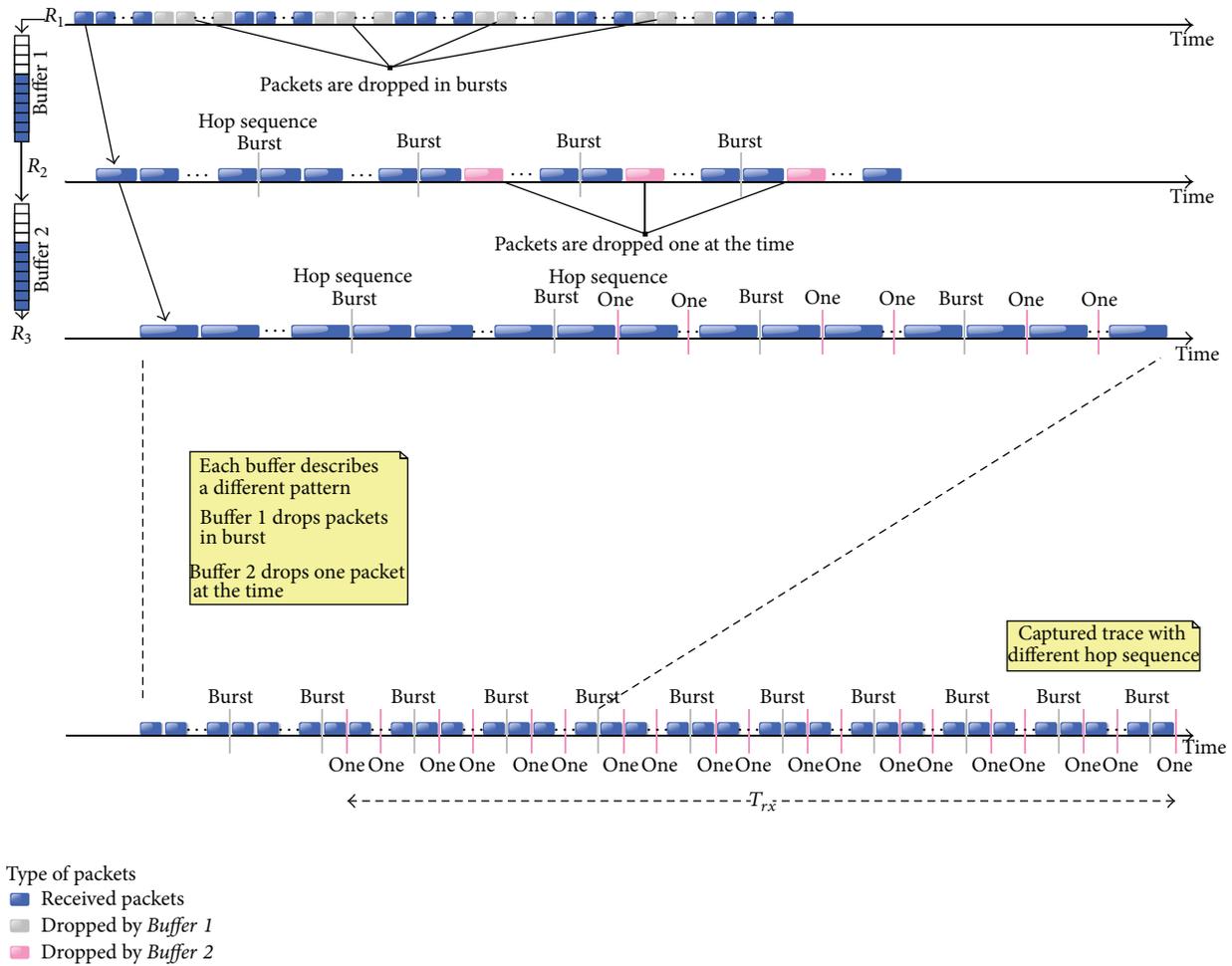


FIGURE 6: Dropped packets for two concatenated buffers.

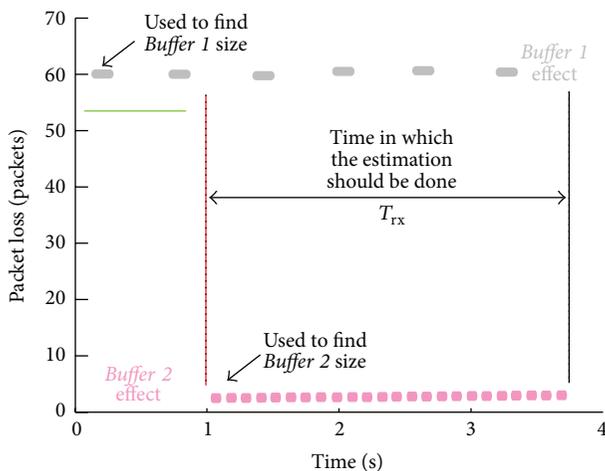


FIGURE 7: Packet loss map for two concatenated buffers.

With the aim of determining if the buffer is measured in number of bytes or packets, a new test should be done; in this case, the new test burst of UDP packets should use a different packet length. Now, we can calculate buffer size for all tests

and compare the results: if the buffer size is the same for all tests, the buffer is measured in number of packets if not in bytes.

### 4. Experimental Results

Real tests have been deployed in a testbed and results are analyzed according to the procedures cited above. We have implemented a controlled network environment in order not only to reproduce the scenario in Figure 5 but also to study two different devices: a switch (3COM) and a Wi-Fi access point (Linksys WAP54G). The topology is shown in Figure 9.

The wireless link capacity is set to 11Mbps. Packets of different sizes (200, 400, 1000, 1500 bytes) are used to determine if the buffer is measured in number of packets or in bytes. The presented results are the most significant.

#### 4.1. Characterizing the Path

4.1.1. Analyze Packet Loss Patterns. In order to obtain a reliable packet-loss map that permits the pattern analysis and be the least intrusive as possible, we generated a traffic of

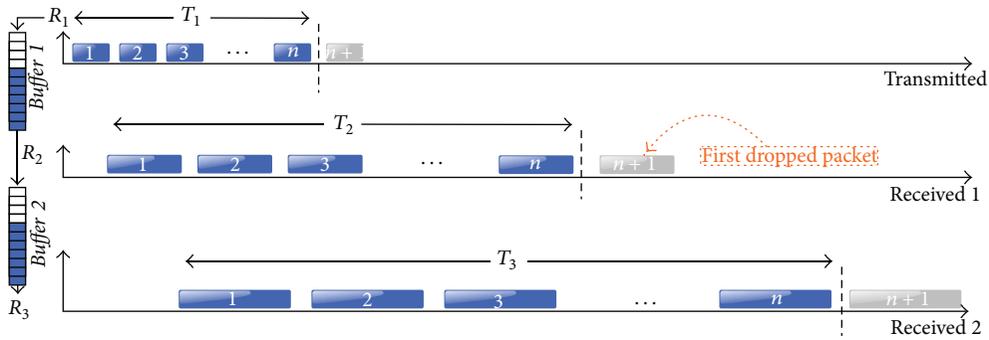


FIGURE 8: Estimating buffer size, from the last received packet before first packet loss.

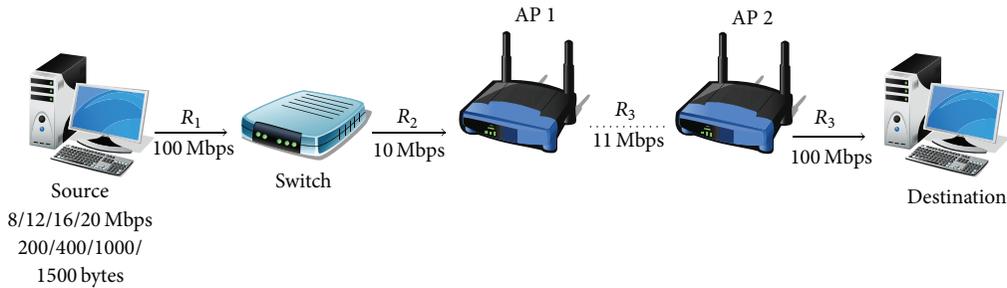


FIGURE 9: Topology used for estimating the buffer size in wired and wireless network.

20 Mbps with a packet length of 1500 bytes, which is considerably bigger than link capacity and it is not so intrusive as the interface maximum output rate. Figure 10(a) shows two different packet loss patterns which can be determined by observing the groups of packets around the same packet loss value. The first pattern is roughly 45 packets and the second one 205. The second pattern is very constant while the first presents some dispersion.

**4.1.2. Determine Rates.** We obtain output and input rates using (2) and (3) in the stable period; however output rate presents more variations due to Wi-Fi behaviour. The results are  $R_3 = 6.5$  Mbps (average) and  $R_1 = 20$  Mbps. Figure 10(a) shows two patterns, so there is an intermediate rate  $R_2$ . From (4) we obtain two possible values:  $R_{2a} = 18$  Mbps and  $R_{2b} = 10$  Mbps. In order to choose the correct value a new test is deployed with a rate of 16 Mbps (a value between  $R_{2a}$  and  $R_{2b}$ ). The results can be seen in Figure 10(b), which shows that two packet loss patterns appear. This means that both buffers get into overflow (the first pattern still maintains its value in 45 packets, but the second one has decreased to 160); thus the intermediate rate is 10 Mbps.

**4.1.3. Infer Locations.** If we consider the buffer order as shown in Figure 6 and the obtained results ( $R_1 = 20$  Mbps,  $R_2 = 10$  Mbps, and  $R_3 = 6.5$  Mbps), *Buffer 1* must lose half the arrived packets, but *Buffer 2* only 35%. We compare these results with each packet loss pattern and we infer that *Buffer 1* drops bursts of 205 packets and *Buffer 2* drops bursts of 45. Furthermore, *Buffer 2* may correspond to the Wi-Fi access point buffer because it has more variations in the number of packets per burst.

**4.1.4. Estimate Buffer Size.** From the destination capture we can observe that the first burst of dropped packets corresponds to *Buffer 2* despite having a lowest filling rate than *Buffer 1*. This is due to the time to completely fill *Buffer 2* is smaller than *Buffer 1*, so  $size_{Buffer_2} < size_{Buffer_1}$ .

We obtained the buffer size for *Buffer 1* (switch) and *Buffer 2* (access point) using (5) with  $n_1 = 240$  and  $n_2 = 143$ . The maximum buffer sizes are roughly 120 packets for the switch and 50 packets (average) for the access point, which correspond to the technical characteristics provided by the manufacturer. The test was repeated with different packet sizes obtaining the same results for buffer sizes, so we conclude that both buffers are measured in number of packets (Figure 11). This methodology can be used for a deep analysis of a single device, the most restrictive, for example, observing the packet size effect.

## 5. Conclusion

This paper has presented a methodology which is useful in order to describe the technical and functional characteristics of commercial buffers on a network path. This characterization is important, taking into account that the buffer may modify the traffic characteristics.

Tests using commercial devices have been deployed in a controlled laboratory scenario, including wired and wireless devices. Accurate results of the buffer size and other parameters have been obtained. We have demonstrated that buffers may be analyzed independently of other devices. As a future line, more than two buffers will be studied by packet loss pattern analysis.

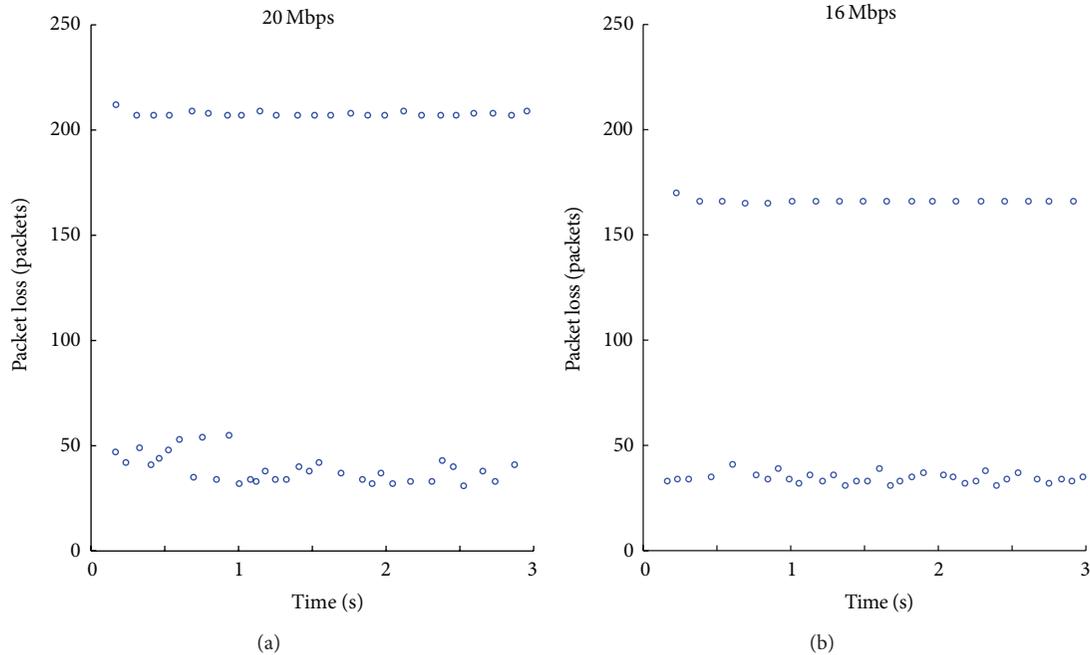


FIGURE 10: Packet loss patterns in the switch and the access point for different bandwidth amounts when packet size is 1500 bytes.

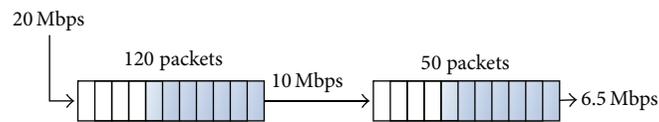


FIGURE 11: Estimated parameters for two concatenated buffers.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work has been partially financed by the European Social Fund in collaboration with the Government of Aragón, Spanish Government, through Grant TEC2011-23037 from the Ministerio de Ciencia e Innovación (MICINN) and PECIO Project, University of Zaragoza, and Fundación Carolina.

## References

- [1] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993.
- [2] C. D. Guerrero and M. A. Labrador, "On the applicability of available bandwidth estimation techniques and tools," *Computer Communications*, vol. 33, no. 1, pp. 11–22, 2010.
- [3] F. Thouin, M. Coates, and M. Rabbat, "Large scale probabilistic available bandwidth estimation," *Computer Networks*, vol. 55, no. 9, pp. 2065–2078, 2011.
- [4] J. Saldana, J. Fernández-Navajas, J. Ruiz-Mas, E. Viruete Navarro, and L. Casadesus, "The effect of router buffer size on subjective gaming quality estimators based on delay and jitter," in *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC '12)*, pp. 482–486, Las Vegas, Nev, USA, January 2012.
- [5] C. Villamizar and C. Song, "High performance TCP in ANSNET," *SIGCOMM Computer Communication Review*, vol. 24, no. 5, pp. 45–60, 1994.
- [6] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '04)*, R. Yavatkar, E. W. Zegura, and J. Rexford, Eds., pp. 281–292, ACM, 2004.
- [7] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Part III: routers with very small buffers," *Computer Communication Review*, vol. 35, no. 3, pp. 83–90, 2005.
- [8] R. Stanojević, R. N. Shorten, and C. M. Kellett, "Adaptive tuning of drop-tail buffers for reducing queueing delays," *IEEE Communications Letters*, vol. 10, no. 7, pp. 570–572, 2006.
- [9] A. Vishwanath, V. Sivaraman, and G. N. Rouskas, "Considerations for sizing buffers in optical packet switched networks," in *Proceedings of the 28th Conference on Computer Communications (IEEE INFOCOM '09)*, pp. 1323–1331, Rio de Janeiro, Brazil, April 2009.
- [10] A. Dhamdhere and C. Dovrolis, "Open issues in router buffer sizing," *Computer Communication Review*, vol. 36, no. 1, pp. 87–92, 2006.
- [11] J. Sommers, P. Barford, A. G. Greenberg, and W. Willinger, "An SLA perspective on the router buffer sizing problem,"

*SIGMETRICS Performance Evaluation Review*, vol. 35, no. 4, pp. 40–51, 2008.

- [12] W. C. Feng, F. Chang, W. C. Feng, and J. Walpole, “Provisioning online games: a traffic analysis of a busy counter-strike server,” in *Proceedings of the Internet Measurement Workshop*, pp. 151–156, ACM, 2002.
- [13] B. Fallica, Y. Lu, F. Kuipers, R. Kooij, and P. Van Mieghem, “On the quality of experience of SopCast,” in *Proceedings of the 2nd International Conference on Next Generation Mobile Applications, Services, and Technologies (NGMAST '08)*, pp. 501–506, September 2008.
- [14] L. Sequeira, J. Fernández-Navajas, L. Casadesus, J. Saldana, I. Quintana, and J. Ruiz-Mas, “The influence of the buffer size in packet loss for competing multimedia and bursty traffic,” in *Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '13)*, pp. 134–141, July 2013.
- [15] L. Sequeira, J. Fernández-Navajas, J. Saldana, and L. Casadesus, “Empirically characterizing the buffer behaviour of real devices,” in *Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '12)*, pp. 1–6, July 2012.

## Research Article

# Protection of HEVC Video Delivery in Vehicular Networks with RaptorQ Codes

**Pablo Piñol, Miguel Martínez-Rach, Otoniel López, and Manuel Pérez Malumbres**

*Physics and Computer Architecture Department, Miguel Hernández University, Avenida de la Universidad, s/n, 03202 Elche, Spain*

Correspondence should be addressed to Pablo Piñol; pablop@umh.es

Received 11 April 2014; Accepted 25 June 2014; Published 17 July 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Pablo Piñol et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With future vehicles equipped with processing capability, storage, and communications, vehicular networks will become a reality. A vast number of applications will arise that will make use of this connectivity. Some of them will be based on video streaming. In this paper we focus on HEVC video coding standard streaming in vehicular networks and how it deals with packet losses with the aid of RaptorQ, a Forward Error Correction scheme. As vehicular networks are packet loss prone networks, protection mechanisms are necessary if we want to guarantee a minimum level of quality of experience to the final user. We have run simulations to evaluate which configurations fit better in this type of scenarios.

## 1. Introduction

Our lives have experienced a radical change since cell phones are no longer just telephones and have become smartphones, with good processing capabilities, large store capacity, and above all, great connectivity. This connectivity allows us to have all kinds of information available at every moment. And we are not only information consumers but active producers as well. At the day when vehicular networks will be integrated by a high percentage of our vehicles and infrastructures in our cities and roads, a vast number of applications of all types (some of them nowadays unthinkable) will arise. Many of them will be oriented to safety and others to entertainment, economizing, and so forth. Within vehicular networks applications, video streaming can be very useful. But, on the one hand, vehicular networks are inhospitable environments where packet losses appear, and on the other hand, video transmission is heavily resource demanding. The combination of these two characteristics makes video streaming over vehicular networks a hard to manage task.

The aim of this work is to evaluate the protection of video encoded with the emerging standard High Efficiency Video Coding (HEVC) by using RaptorQ codes and how they both behave in vehicular scenarios. For this task we will vary a series of parameters in both HEVC and RaptorQ and will present the performance of those configurations.

There are lots of works that evaluate HEVC efficiency (like [1, 2]) although most of them do not take into consideration lossy environments. Some works evaluate HEVC performance under packet loss conditions as the authors do in [3]. In that work the authors have developed a complete framework for testing HEVC under different packet loss rates, bandwidth restrictions, and network delay. As HEVC decoder is not robust against packet losses (it crashes if packets are missing) their framework decodes the complete bitstream and then overrides the areas corresponding to the missing packets. Several works propose mechanisms for protecting video streaming over vehicular networks although quality evaluation refers to percentage of lost packets [4]. In [5] the authors do a thorough research of protection of content delivery in vehicular environments searching for the best packet size in order to maximize throughput. They use different FEC techniques to protect data and offer results in terms of packet arrival ratio and file transfer time.

Our work differs in some aspects from the cited research works. The main difference is that our work combines several features of the cited works into one research. And also it includes features that are not included in previous works. In our work we have used real maps to design our vehicular scenarios. We have not used synthetic losses but the real losses obtained by simulations of vehicles moving through the scenario. We have used HEVC reference software with some

modifications in order to make it robust against packet losses (avoiding crashes when packets are lost). So we have decoded the bitstream with missing packets directly with the reference software. We have used RaptorQ codes in order to protect the video stream and we have tuned them to test which configurations get better results. And not only statistics about percentage of recovered packets are provided, but also the quality of the real reconstructed video sequence is presented (in terms of PSNR versus the original sequence).

The rest of the paper is structured as follows. In Section 2 the three main components of the scenarios are presented: vehicular networks, HEVC, and RaptorQ codes. In Section 3 the framework where the tests have been done is explained. Section 4 explains the results obtained for the different configurations that we have tested. At last, in Section 5, several conclusions are drawn.

## 2. Components

In this section we describe the three integral parts of the scenarios of our research: vehicular networks, HEVC, and RaptorQ codes.

*2.1. Vehicular Networks.* In future, vehicles will be equipped with lots of sensors (nowadays they already are) which will be able to collect internal and external measurements. They will also be provided with a certain computing capability which will allow them to process information, and they will also carry communications equipment. These three elements will make Intelligent Transport Systems (ITS) possible, where vehicles will have the ability to communicate with each other and with infrastructure in an intelligent way. This capacity of vehicles to communicate with each other and with a fixed network will bring both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) networks. These networks will be used by applications regarding areas like people safety, fuel consumption savings, reduction of CO<sub>2</sub> emissions, infotainment, and so forth. Video streaming will be used by diverse types of applications like digital entertainment, Video On Demand (VOD), tourist information, contextual advertising, traffic flow density, and other regarding safety issues like emergency video call and so forth. Vehicular networks will have many challenges due to their nature. The combination of wireless communications with the relative high speed of nodes, variability of routes, and obstacles disturbing wireless signal (buildings, other vehicles, etc.) will make them packet loss prone networks. In packet loss prone networks, video streaming applications, which produce big deals of data and have high bandwidth requirements, may decrease the network performance and may have to deal with high rates of packet loss. These big deals of data need to be efficiently compressed to diminish the bandwidth needed for streaming. So we will need the assistance of proficient video codecs.

*2.2. HEVC.* On January 2013, High Efficiency Video Coding [6] was agreed upon as the new video coding standard. By 2010, ITU-T Video Coding Experts Group (VCEG)

and ISO/IEC Moving Pictures Experts Group (MPEG) joined their research efforts and constituted the Joint Collaborative Team on Video Coding (JCT-VC) in order to develop a new video coding standard that would improve the previous one, H.264/AVC (Advanced Video Coding) [7], and could keep pace with the growing resolutions and frame rates of new video contents. The new standard follows the same hybrid compression scheme as its predecessor but includes a good number of refinements and new features that nearly doubles its coding efficiency. Here we will explain one of the features of HEVC that we will use in our tests: slices. For a deeper insight into the standard, some of the members of JCT-VC provide a complete overview in [8].

Slices are components of an encoded video stream that were introduced in the previous video coding standard, H.264/AVC. They are coded fragments of a frame that can be independently decoded. This makes them especially useful in packet loss prone scenarios. If we encode each frame of a video sequence using only one slice and that slice is bigger than the network MTU, then we will have to divide the slice in several fragments which will travel in several network packets. If one of these packets gets lost, then the rest of the fragments of the slice will become completely useless, because a slice cannot be decoded if it is not complete. In this way the loss of a single packet implies the effective loss of the whole frame. But if we divide each frame into several slices and each slice is smaller than the network MTU, then we will be able to send each slice in one packet. The loss of a single packet would not imply the loss of the whole frame but only the loss of a slice because the rest of the packets could be independently decoded.

But there is a drawback in dividing a frame into several slices. As slices are independently decodable, prediction is not allowed farther away of the slice limits. For instance, spatial prediction using areas of other slices is not allowed. The same happens to motion vectors prediction that must remain inside slice limits. This causes a decrease in coding efficiency. For every slice we will have a slice header which will also introduce some overhead. And if we divide a frame into too many slices, then the size of each slice may be much smaller in comparison with other headers used for streaming (e.g., RTP) and this would introduce a considerable overhead. In this work we will study how HEVC behaves in video streaming over vehicular networks for different number of slices per frame with the aid of RaptorQ codes to deal with packet loss.

*2.3. RaptorQ Codes.* Raptor codes, invented by Shokrollahi [9, 10], are a type of Fountain codes and are based in Luby Transform (LT) codes [11]. Raptor codes are a Forward Error Correction (FEC) technology which implements application layer protection against network packet losses. RaptorQ codes are a new family of codes that provide superior flexibility, support for larger source block sizes, and better coding efficiency than Raptor codes. They have several features that make them an interesting technology. One of them is that they can encode (protect) and decode (restore) data with linear time. They can also add variable levels of protection to better suit the protection to the network characteristics

(e.g., packet loss ratio, maximum bandwidth, etc.). They have very good recovery properties, because they can completely recover the original data if they receive approximately the same amount of data than the original one, regardless of whether the received packets are original packets or repair packets. RaptorQ codes are very efficient and have small memory and processing requirements, so they can be used in a wide variety of devices (from smartphones to big servers). Raptor and RaptorQ codes have been standardized by IETF [12, 13] and are used in 3GPP Multimedia Broadcast Multicast Services (MBMS) for file delivery and streaming.

This is how RaptorQ operates. The RaptorQ encoder receives a data stream (source packets) during a specified protection period. These packets are put together in memory to form a source block. A 4-byte FEC trailer is added to each received packet. This trailer identifies the packet and the protection period to which it belongs. These FEC-protected packets are sent through the network. When the protection period finishes, the source block in memory is FEC-encoded into repair symbols which are placed into repair packets and sent through the network. At the receiver side, the RaptorQ decoder receives protected source packets and repair packets. Some of these packets may get lost or corrupted and the RaptorQ decoder will try to recover lost packets out of the group of source and repair packets correctly received.

Latency (and, in some cases, memory consumption) is the drawback of RaptorQ protection scheme. As the protection window increases, the delay grows. Live events or real-time applications like video conference will have to use short periods for the protection window to keep latency into reasonable limits. Other types of streaming applications like IPTV may tolerate wider protection windows (while keeping latency inside a reasonably interaction response time). And some other applications like Video On Demand can be much more flexible in enlarging the protection period which will provide more bandwidth efficiency.

### 3. Framework

In our tests, several tools and simulators have been used. For the construction of the vehicular scenario we have used the OpenStreetMap project [14]. The OpenStreetMap project is a public domain geographic data base, built upon contributions of volunteers of all around the world. It is a project like Wikipedia but with geographic data, where you can find varied information like streets (including the number of lanes and their direction), parks, squares, schools, bus stops, singular buildings, drugstores, rivers, and so forth. From this page we have downloaded a real map of the city of Kiev. This map (XML data) has to be converted in order to be handled by SUMO (simulation of urban mobility) [15]. SUMO is a traffic simulator which is well known by the scientific community. It is able to run traffic simulations including vehicles, traffic lights, crossroads, priorities, and so forth and allows defining characteristics like the size, acceleration, deceleration and maximum speed of vehicles, and gathering different statistics like fuel consumption and CO2 emissions. One of the tools which is included in SUMO

is TraCI (Traffic Control Interface). By using this interface, a bidirectional communication is possible between SUMO and other applications. In our tests we will connect SUMO (which will run the simulation of vehicles mobility) with OMNeT++ [16] where the vehicular network will be tested.

OMNeT++ is not a simulator itself but a framework for the development of simulators. It is a public domain software and a lot of projects (which implement the actual simulators) are available for it. Two of these projects have been put together to provide a vehicular network simulator, MiXiM [17] and Veins [18]. MiXiM is a project which implements wireless communications both fixed and mobile. Veins adds some protocols which have been standardized for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) wireless communications, like IEEE 802.11p [19] and the IEEE 1609 family of standards [20].

Inside this vehicular network simulator we have developed an application for driving the experiments. It basically allows the injection of a video stream and the insertion of background traffic to produce congestion in the network to simulate adverse conditions. It also gathers some statistics for subsequent analysis.

For the encoding and decoding of HEVC video we have used the HEVC reference software [21]. That software has been modified by us in order to generate RTP packets inside the bitstream. In addition we have modified the HEVC decoder to make it resilient to packet losses. We have used two different encoding modes, specified by the JCT-VC. On the one hand we have used All Intra (AI) mode, in which every frame of the video sequence is encoded as an I frame. An I frame is encoded without using other frames as a reference, and therefore, it is independent of the rest of the frames. This is called intramode. An I frame exploits the spatial redundancy that exists inside the frame. On the other hand, we have used Low-Delay P (LP) mode. In this mode the first frame is encoded as an I frame and the rest of the frames are P (predictive) frames. P frames use other frames previously encoded (and decoded) as a reference to exploit temporal redundancy. They use motion estimation/compensation. This is called inter mode. Inter mode is more efficient than intramode; thus, better compression rates are obtained by LP encoding mode than by AI encoding mode. In LP mode, P frames depend on other frames used as a reference. This makes this mode more vulnerable to packet losses. In AI mode the loss of a slice will only affect one frame, but in LP mode, the loss of a slice will affect the frame which it belongs to and also the frames that use this frame as a reference. The frames that use these affected frames will also be damaged and so on. To stop this drift and try to alleviate packet losses (those that could not be recovered with RaptorQ codes) we have used the intra-refresh mechanism in LP encoded streams. This mechanism consists basically in inserting an I frame every 32 frames. If all the slices that belong to this frame arrive to their destination (or equivalently, RaptorQ codes are able to recover them) the sequence is refreshed and the propagation of errors ends.

For the protection of the bitstream with RaptorQ codes we have used the Qualcomm (R) RaptorQ (TM) Evaluation Kit [22]. RaptorQ software has different options to better tune

it and suit it to fit your needs or preferences. For instance, you can specify the amount of protection to add to the bitstream (i.e., 10%, 20%, etc.). The bigger the amount of protection you add, the higher the probability of successful recovery of lost packets will be, but also the higher the bandwidth required for the transmission will become. You can also adjust the length of the temporal window which will divide the bitstream in fractions in order to generate FEC packets. You can also set the symbol size and the repair packet size. An adequate symbol size can make the encoding and decoding of FEC data more computationally efficient and also reduce the amount of memory required for these computations. The final aim of this work is to evaluate how HEVC and RaptorQ codes behave and how they can collaborate to protect video transmissions in vehicular networks and determine which are the most suitable configurations of both in each situation. This is a necessary step to propose adaptive mechanisms that can optimize resources and provide error resilience techniques that assure a good quality of experience in video streaming via vehicular networks.

For the performance of the tests we have followed these steps. Once the scenario has been implemented, we have encoded a raw video sequence at different number of slices per frame producing several HEVC bitstreams (in RTP format). Each one of these bitstreams has been FEC encoded with several configurations using RaptorQ. Protected sequences have been used to run simulations. Each simulation produces a file with several statistics (including the packet loss ratio) and a file with the received packets. This file is FEC decoded by means of RaptorQ in order to try to recover lost packets and to produce an HEVC file (in RTP format). This RTP file is decoded by HEVC decoder and finally the reconstructed raw video sequence is compared to the original one in order to calculate PSNR and evaluate final video quality.

In next section we will give some extra details about tests and we will analyze the most relevant results obtained.

## 4. Experiments and Results

The study case is based on the scenario shown in Figure 1. It is an area of 2000 m × 2000 m of the city of Kiev. In it we can find a long avenue that crosses that area from north to south. Along the avenue, 3 road side units (RSUs) have been positioned, named A, B, and C in the figure. These RSUs will transmit the video sequence simultaneously (in a synchronized way). The coverage radius of the wireless devices is 500 m. RSUs A and B have a small area where their signals overlap. And RSUs B and C have a small shaded area where neither of the two can reach. Therefore we have three different types of areas regarding transmission: areas where a vehicle can receive the data from only one RSU, one area where the vehicle receives the signal from two RSUs (A and B), and one area where signal is momentarily lost (between B and C coverage areas). A total of 50 vehicles have been inserted into the scenario, driving in different routes. Those vehicles send a beacon every second through the control channel (following IEEE 1609.4 multichannel operations).

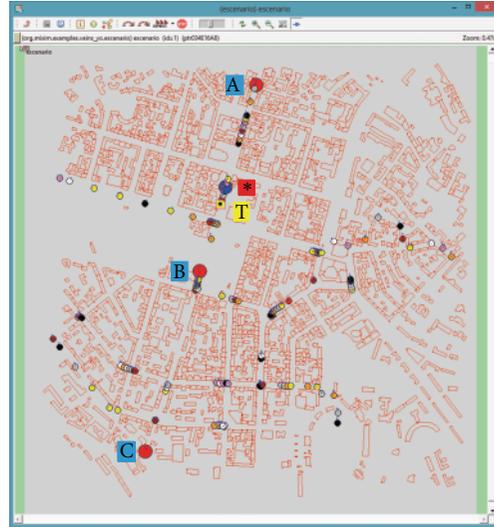


FIGURE 1: Vehicular network scenario in OMNeT++ (red circles A, B, C = RSUs//blue circle \* = video client//yellow square T = background traffic source//small circles = other vehicles//red rectangles = buildings).

We also have a vehicle driving near the video client that can act as a background traffic source (labeled as T in Figure 1), sending packets through the wireless network at different packets per second (pps) rates. The video client (marked in the figure as \*) will experience isolated packet losses (mainly due to background traffic) and bursty packet losses (around the limits of RSUs coverage). RSUs send periodically through the control channel advertisements of the video service that they offer, indicating the service channel used for the video stream. The video client receives that invitation and commutes to the specified service channel in order to receive the video stream.

**4.1. HEVC Evaluation.** Now we present the evaluation of the video sequence behavior when it is encoded at a different number of slices per frame in both encoding modes used (AI and LP).

The sequence chosen for the tests is RaceHorses, one of the test sequences used by JCT-VC for HEVC evaluation in common test conditions [23]. It has a resolution of 832x480 pixels and a frame rate of 30 frames per second. We have encoded it at 1, 2, 4, 8, 13, and 26 slices per frame (slc/frm). For the encoding process we have used a value of 37 for the quantization parameter (QP). For that value we obtain a mean PSNR value of 32.12 dB for AI mode (1 slc/frm) and a value of 30.19 dB for LP mode (1 slc/frm). Video quality is higher in AI mode but bitrate in LP mode is much lower. As predictions cannot cross slice boundaries, when we split each frame into several slices we are reducing coding efficiency because we cannot use information of nearby areas if they do not belong to the slice. An example of this penalization is that slices cannot use intraprediction between slices (in AI mode) and slices cannot use predictions for motion vectors (in LP mode). Figure 2 shows the percentage of increment

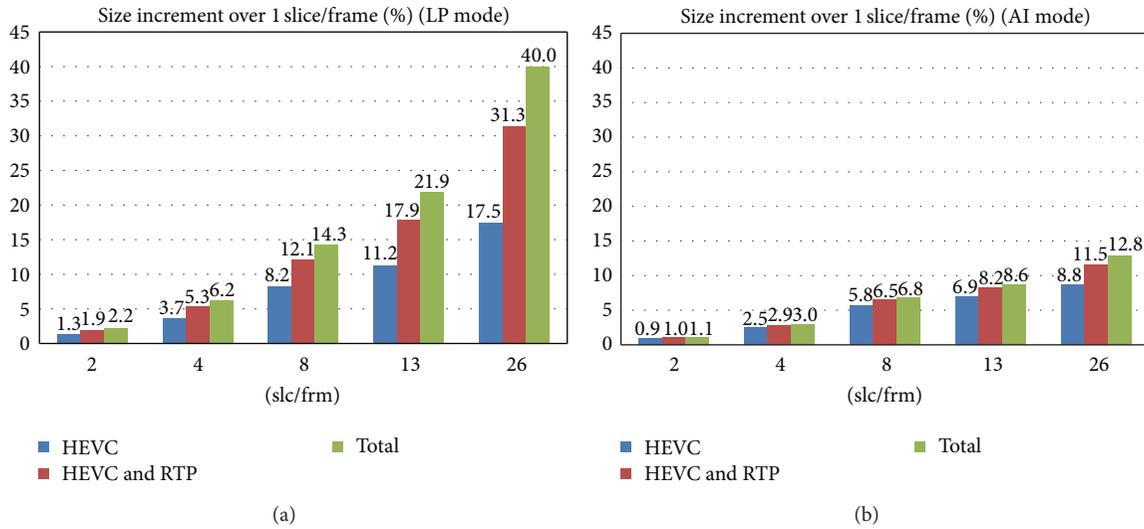


FIGURE 2: Percentage of bitrate increase (without FEC protection) for different number of slices per frame. (a) LP mode. (b) AI mode. (HEVC raw bitstream//HEVC + RTP header//HEVC + RTP header + fragmentation header.)

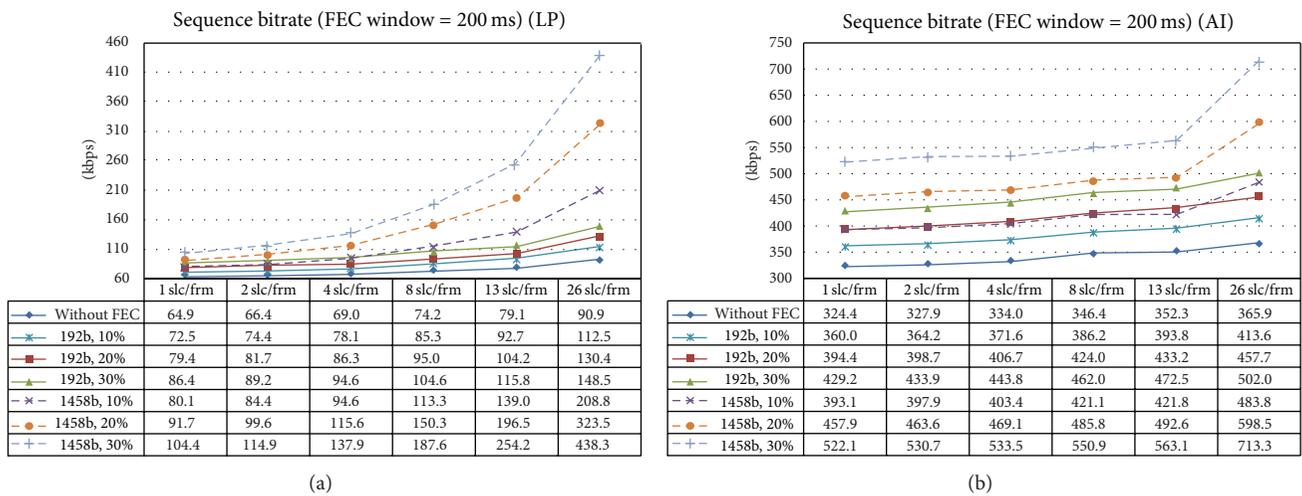


FIGURE 3: Bitrate (kbps) without FEC protection and with FEC protection for different symbol sizes, different protection windows, and different number of slices per frame (including RTP and fragmentation headers). (a) LP mode. (b) AI mode.

of the encoded sequence size at different number of slices per frame compared to 1 slc/frm. It shows the percentage of increment of HEVC raw bitstream and also the percentage of increment after adding the RTP headers to each slice. If the size of one slice (including its RTP header) is greater than the network MTU, then it will be divided into some fragments. If one of the fragments of a slice gets lost, then the whole slice will be discarded because it will be undecodable. So the rest of the fragments of the slice are automatically discarded. We identify every fragment of a slice with a header in order to know if a slice has received all its fragments. In Figure 2, data labeled as "TOTAL" shows the bitrate increment with respect to 1 slc/frm when both RTP and fragmentation headers are included. As slices generated by LP mode are much smaller than slices generated by AI mode (because of LP mode coding efficiency), the overhead introduced by RTP and fragmentation headers is much greater for LP mode, and consequently

the same happens to the total percentage of increment in the bitstream. For example, encoding at 13 slc/frm increases the bitstream around a 20%, and encoding at 26 slc/frm increases it around a 40% for LP mode. But for the same number of slices per frame, in AI mode the increments are around 8.6% and 12.8%, respectively.

In Figure 3, if you look at the curve labeled "without FEC," you can see the bitrate value (not the percentage of increment) of the sequence for LP and AI modes. In LP mode the bitrates range from 64.9 kbps (1 slc/frm) to 90.9 kbps (26 slc/frm). In AI mode the bitrates range from 324.4 kbps (1 slc/frm) to 365.9 kbps (26 slc/frm). As we stated before in this section, the bitrate for mode LP is much lower than for AI mode; specifically, at 1 slice per frame the bitrate of LP mode is 5 times lower than AI's one.

A factor that can be of great relevance when protecting data is the proportion of network packets (fragments) with

TABLE 1: Mean proportion of fragments (network packets) for every RTP packet (slice).

Fragments/RTP	1 sl	2 sl	4 sl	8 sl	13 sl	26 sl
LP mode	1.87	1.37	1.21	1.02	1.00	1.00
AI mode	7.90	4.29	2.38	1.51	1.07	1.00

TABLE 2: Packet rate (packets per second) for LP mode without FEC protection and with FEC protection at 30% of redundancy for different symbol sizes, different protection windows, and different number of slices per frame.

Packets/sec.	1 sl	2 sl	4 sl	8 sl	13 sl	26 sl
Without FEC	56.1	82.0	134.6	244.9	391.3	780.0
192 b, 133 ms	75.4	102.7	156.2	270.0	420.7	824.6
192 b, 200 ms	73.8	100.5	155.0	268.7	419.6	822.7
192 b, 250 ms	73.1	100.1	154.6	268.4	419.0	821.5
192 b, 333 ms	72.9	99.8	154.2	267.7	418.4	820.7
192 b, 500 ms	72.1	99.2	153.3	266.5	417.0	821.5
192 b, 1000 ms	71.2	97.9	152.0	265.2	415.5	820.9
1458 b, 133 ms	83.2	116.3	183.0	320.4	509.1	1016.7
1458 b, 200 ms	82.9	114.9	181.3	321.7	509.8	1015.1
1458 b, 250 ms	81.8	114.5	180.8	321.6	509.1	1010.0
1458 b, 333 ms	81.8	114.0	180.7	320.0	507.9	1008.9
1458 b, 500 ms	80.5	113.7	179.1	317.8	504.3	1003.9
1458 b, 1000 ms	79.5	111.6	176.2	313.2	498.1	1013.9

respect to RTP packets (each RTP packet includes one slice). As stated before, if we have many fragments for one RTP packet, then the probability of losing this RTP packet increases, because the real loss of only one of the fragments will render the RTP packet useless. And this will result in the effective loss of the rest of the fragments of that RTP packet. But if we have one fragment per RTP packet, then the loss of one fragment will not affect the rest of packets. When the proportion tends to 1 every lost packet which can be recovered by RaptorQ is contributing to improve the final quality of the reconstructed video. When the proportion moves far away from 1 the recovery of packets by RaptorQ does not always directly turn into an improvement on video quality.

The mean proportion of fragments per RTP packet is shown in Table 1. Data show that when using AI mode and few slices per frame the proportion is far from 1, and this indicates that packet recovery will not be as productive as for LP mode. Data of Table 1 will be different if we encode other video sequences with larger (or smaller) resolutions or with different values for QP parameter or with different values for intra-refresh period. The conclusion is that before protecting video streams with RaptorQ codes, we should take into consideration this proportion in order to better decide the suitable number of slices per frame that will take a greater advantage of FEC protection.

**4.2. RaptorQ Codes Setup.** For the evaluation of RaptorQ and determining which configurations are most suitable, we have protected each of the generated bitstreams with different setups for RaptorQ. 6 different sizes for the temporal window have been used (133, 200, 250, 333, 500, and 1000

milliseconds), 3 different levels of protection have been assigned (10%, 20% and 30%), and 2 different combinations of symbol size and repair packet size have been used (a small symbol size of 192 bytes together with a repair packet size of 7 and a big symbol size of 1458 bytes together with a repair packet size of 1).

First of all we have analyzed the overhead generated by the protection added with RaptorQ codes. In addition to the global bitrate increment we have measured the increment in the packet rate (which is caused by the addition of repair packets). In previous works we found out that in vehicular networks packet losses are more influenced by packet rate than by packet size, so if we do not keep the extra packets per second low, then the solution to our problem (adding extra repair packets) can become the problem of the network.

Table 2 shows the number of packets per second in the transmission of the LP encoded sequence without any FEC protection and also with FEC protection with a redundancy of 30% for different symbol sizes and different temporal windows. When we use a big symbol size, more packets per second are generated than when we use a small symbol size. This gets worse as the number of slices increases. It can also be seen that varying the temporal window does not change significantly the packet rate.

Table 3 shows the packet rate without FEC protection and by using FEC protection with a symbol size of 192 bytes and a temporal window of 200 ms. If we compare the number of packets per second for LP and AI modes without using FEC we can see the correlation with Table 1. When we encode the video sequence at 26 slc/frm, we have a proportion of 1.00 fragment per RTP packet for both LP and AI modes. This means that every single RTP packet is smaller than the MTU

TABLE 3: Packet rate (packets per second) without FEC protection and with FEC protection for a symbol size of 192 bytes and a protection window of 200 ms for different coding modes, different percentages of redundancy, and different number of slices per frame.

Packets/sec.	1 sl	2 sl	4 sl	8 sl	13 sl	26 sl
(LP) w/o FEC	56.1	82.0	134.6	244.9	391.3	780.0
(LP) 10%	63.6	89.9	143.3	254.9	402.0	795.4
(LP) 20%	68.5	95.0	148.9	261.9	410.8	808.9
(LP) 30%	73.8	100.5	155.0	268.7	419.6	822.7
(AI) w/o FEC	237.0	257.3	285.3	362.8	417.6	780.0
(AI) 10%	265.0	285.6	314.8	392.9	449.3	815.0
(AI) 20%	290.2	311.0	340.6	421.1	478.2	847.6
(AI) 30%	316.0	336.8	367.7	449.4	507.4	880.7

and the packet rate can be directly calculated by multiplying the frame rate (30 fps) by the number of slices per frame (26 slc/frm); this is 780.0 pps. At the opposite side of Table 3, if we look at AI mode without using FEC encoded at 1 slc/frm, we can see that the packet rate (237.0 pps) is very far from the multiplication of the frame rate (30 fps) by the number of slices per frame (1 slc/frm). At 1 slc/frm AI mode and LP mode produce very different packet rates. Observing FEC-protected data it can be seen that when fewer slices per frame are used the proportion of extra packets per second is greater.

Figure 3 shows the bitrate (kbps) of the encoded sequences in modes LP and AI without FEC protection and with FEC protection for three different levels of redundancy and two symbol sizes. Selecting 192 bytes as the symbol size leads to much lower bitrates. This is more emphasized for LP coding mode.

4.3. *OMNeT++ Tests.* In this section we will present the tests performed in the simulations. For the experiments we have used the framework previously depicted, using SUMO, OMNeT++, MiXiM, and Veins.

In simulations we connect SUMO and OMNeT++ via TraCI. SUMO tells OMNeT++ the position of every vehicle at every instant and OMNeT++ (using MiXiM and Veins) runs the simulation of the vehicular network. Video sequences which have been previously protected with RaptorQ codes are transmitted from the 3 RSUs to the video receiver. At the end of the simulation a file is generated with some statistics, including the percentage of packets lost. Another file is also generated including the packets received by the vehicle. This file is processed using RaptorQ decoder in order to generate a file with RTP packets, trying to recover the lost packets. The output of this process is passed to the HEVC decoder which will try to restore the video sequence. After this process we compute the PSNR value for the reconstructed sequence.

We have run simulations for the following combinations: both modes of encoding (AI and LP), different number of slices per frame (1, 4, 8, and 13 slc/frm), and three protection levels (10%, 20%, 30%). For all these combinations we have run tests injecting background traffic at four different rates (0, 30, 240, and 390 pps).

In areas of full coverage, packet losses are due to background traffic. Here we encounter isolated losses. On the contrary, in areas near the limits of the RSUs coverage,

TABLE 4: Total percentage of network packet loss, percentage of RTP packet loss after recovery, and difference in PSNR of the reconstructed video, for a background traffic of 390 pps and a level of protection of 30%, for areas with good signal coverage. LP encoding mode.

slc/frm	Measurement	192 b	1458 b
1 sl	TOTAL loss (%)	11.28	10.89
1 sl	RTP loss (%)	1.28	0.18
1 sl	PSNR diff (dB)	0.99	0.18
4 sl	TOTAL loss (%)	13.55	14.36
4 sl	RTP loss (%)	1.84	0.62
4 sl	PSNR diff (dB)	2.85	0.75
8 sl	TOTAL loss (%)	14.59	15.05
8 sl	RTP loss (%)	1.47	0.09
8 sl	PSNR diff (dB)	1.90	0.11
13 sl	TOTAL loss (%)	13.97	13.25
13 sl	RTP loss (%)	0.29	0.00
13 sl	PSNR diff (dB)	0.52	0.00

packet losses are bursty because the signal is completely lost for some period. For isolated losses RaptorQ codes do a good job in recovering lost packets. Tables 4 and 5 show the total percentage of network packet loss, the percentage of RTP packet loss after recovery, and the difference in PSNR of the reconstructed video for LP and AI modes, respectively (for areas of good coverage). As it can be seen, RaptorQ can recover a high percentage of network packets and the result is that only a small percentage of RTP packets are lost. This is not true for AI mode and few slices per frame (1 slc/frm and 2 slc/frm), but this is the expected behavior taking into account that, as we stated before, in these configurations the proportion of fragments per RTP packet is high. We can observe from those tables that AI mode is inherently more error resistant than LP mode. This is also the expected behavior because in LP mode reference pictures with incorrect data propagate errors and in AI mode errors do not propagate. Some techniques like unequal error protection methods could be useful to introduce different levels of protection regarding the importance of the video packets (I or P frames) in the final quality of the reconstructed sequence. Regarding areas near the limits of RSUs coverage, we can state that RaptorQ is not able to deal with bursty losses

TABLE 5: Total percentage of network packet loss, percentage of RTP packet loss after recovery and difference in PSNR of the reconstructed video, for a background traffic of 390 pps and a level of protection of 30%, for areas with good signal coverage. AI encoding mode.

slc/frm	Measurement	192 b	1458 b
1 sl	TOTAL loss (%)	19.07	19.10
1 sl	RTP loss (%)	14.81	16.94
1 sl	PSNR diff (dB)	2.20	2.52
4 sl	TOTAL loss (%)	18.34	18.25
4 sl	RTP loss (%)	6.76	4.17
4 sl	PSNR diff (dB)	1.32	0.81
8 sl	TOTAL loss (%)	16.65	16.13
8 sl	RTP loss (%)	2.56	0.93
8 sl	PSNR diff (dB)	0.71	0.23
13 sl	TOTAL loss (%)	15.06	14.72
13 sl	RTP loss (%)	0.46	0.21
13 sl	PSNR diff (dB)	0.15	0.06

because within the protection period very few packets are received and there is not enough data to carry out a recovery. This problem could be addressed with the introduction of techniques like interleaving, where, with the trade-off of the introduction of some delay, bursty losses can be easily converted to isolated losses (where RaptorQ can get good percentages of recovery). We can see that a symbol size of 1458 bytes provides better recovery results but if we take into consideration data from Figure 3, then the overhead introduced is not bearable. At last, from the very specific conditions of our tests we can state that the optimum number of slices per frame is 8, which produces the best trade-off between the introduced overhead (both packets per second and total bitrate) and the percentage of recovery and the final video quality.

## 5. Conclusions

In this work we have analyzed the protection of video delivery in vehicular networks. The video encoder selected for the tests is the new emerging standard HEVC. To protect the video stream we have used RaptorQ codes. We have first analyzed the behavior and performance of HEVC for two coding modes (AI and LP) and for different number of slices per frame. Then we have protected the encoded sequences by means of RaptorQ with several configurations and observed the effects of this selection. We have seen that varying HEVC and RaptorQ parameters leads to very different situations, regarding total bitrate and packet rate. At last we have run simulations in a vehicular environment to measure the ability of RaptorQ in protecting video packets in this type of scenarios. The reduction in the number of lost packets because of RaptorQ codes recovery properties has been presented as well as the quality of the decoded video reconstructions. As a general conclusion we can state that there are a lot of parameters that can be fine-tuned to adapt the video protection to the requirements of the specific network conditions

(bandwidth, packet loss ratio, isolated/bursty losses, etc.) and to the requirements of the specific application (encoding mode, level of quality, resolution of video sequence, etc.). So there is not a general formula which will fit into all situations to provide the best level of protection. On the contrary, a previous evaluation of the real conditions and user preferences is mandatory.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research was supported by the Spanish Ministry of Education and Science under Grant TIN2011-27543-C03-03, the Spanish Ministry of Science and Innovation under Grant TIN2011-26254, and Generalitat Valenciana under Grant ACOMP/2013/003.

## References

- [1] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards-including high efficiency video coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, 2012.
- [2] R. Garcia and H. Kalva, "Subjective evaluation of HEVC in mobile devices," in *Multimedia Content and Mobile Devices*, vol. 8667 of *Proceedings of SPIE*, February 2013.
- [3] J. Nightingale, Q. Wang, and C. Grecos, "HEVStream: a framework for streaming and evaluation of high efficiency video coding (HEVC) content in loss-prone networks," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 404–412, 2012.
- [4] J. Park, U. Lee, and M. Gerla, "Vehicular communications: Emergency video streams and network coding," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 57–68, 2010.
- [5] C. T. Calafate, G. Fortino, S. Fritsch, J. Monteiro, J.-C. Cano, and P. Manzoni, "An efficient and robust content delivery solution for IEEE 802.11p vehicular environments," *Journal of Network and Computer Applications*, vol. 35, no. 2, pp. 753–762, 2012.
- [6] B. Bross, W.-J. Han, J.-R. Ohm et al., "High Efficiency Video Coding (HEVC) text specification draft 10," Tech. Rep. JCTVC-L1003, Joint Collaborative Team on Video Coding (JCT-VC), Geneva, Switzerland, 2013.
- [7] ITU-T and ISO/IEC JTC 1, "Advanced Video Coding for Generic Audiovisual Services," ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC) version 16, 2012.
- [8] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
- [9] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, 2006.
- [10] M. A. Shokrollahi and M. Luby, "Raptor codes," *Foundations and Trends in Communications and Information Theory*, vol. 6, no. 3-4, pp. 213–322, 2009.

- [11] M. Luby, "LT codes," in *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science*, pp. 271–280, November 2002.
- [12] M. Luby, A. Shokrollahi, M. Watson, and T. Stockhammer, "Raptor forward error correction scheme for object delivery," in *IETF RMT Working Group, Work in Progress*, RFC 5053, 2007.
- [13] M. Luby, A. Shokrollahi, M. Watson, T. Stockhammer, and L. Minder, "RaptorQ Forward Error Correction Scheme for Object Delivery," IETF RMT Working Group, Work in Progress, RFC 6330, 2011.
- [14] OpenStreetMap, <http://www.openstreetmap.org/>.
- [15] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO—simulation of urban mobility," *International Journal on Advances in Systems and Measurements*, vol. 5, no. 3-4, pp. 128–138, 2012.
- [16] OMNeT++, <http://www.omnetpp.org/>.
- [17] MiXiM, <http://mixim.sourceforge.net/>.
- [18] C. Sommer, R. German, and F. Dressler, "Bidirectionally coupled network and road traffic simulation for improved IVC analysis," *IEEE Transactions on Mobile Computing*, vol. 10, no. 1, pp. 3–15, 2011.
- [19] "IEEE Standard for Information technology—Local and metropolitan area networks—Specific requirements—Part: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments," IEEE Std 802.11p-2010, pp. 1–51, 2010.
- [20] "IEEE standard for Wireless Access in Vehicular Environments (WAVE)—multi-channel operation," Tech. Rep. IEEE Std 1609.4-2010, 2011.
- [21] H M Reference Software vers. 9.0, <https://hevc.hhi.fraunhofer.de/svn/svn.HEVCSoftware/tags/HM-9.0/>.
- [22] Qualcomm (R) RaptorQ (TM) Evaluation Kit, <http://www.qualcomm.com/raptor-evaluation-kit>.
- [23] F. Bossen, "Common test conditions and software reference configurations," Tech. Rep. JCT VC-L1100, Joint Collaborative Team on Video Coding (JCT-VC), Geneva, Switzerland, 2013.

## Research Article

# A Secure 3-Way Routing Protocols for Intermittently Connected Mobile Ad Hoc Networks

Ramesh Sekaran<sup>1</sup> and Ganesh Kumar Parasuraman<sup>2</sup>

<sup>1</sup> Anna University Regional Centre, Madurai, Tamil Nadu 625007, India

<sup>2</sup> K.L.N. College of Engineering, Sivagangai, Tamil Nadu 630611, India

Correspondence should be addressed to Ramesh Sekaran; itz\_ramesh87@yahoo.com

Received 10 March 2014; Accepted 21 June 2014; Published 17 July 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 R. Sekaran and G. K. Parasuraman. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The mobile ad hoc network may be partially connected or it may be disconnected in nature and these forms of networks are termed intermittently connected mobile ad hoc network (ICMANET). The routing in such disconnected network is commonly an arduous task. Many routing protocols have been proposed for routing in ICMANET since decades. The routing techniques in existence for ICMANET are, namely, flooding, epidemic, probabilistic, copy case, spray and wait, and so forth. These techniques achieve an effective routing with minimum latency, higher delivery ratio, lesser overhead, and so forth. Though these techniques generate effective results, in this paper, we propose novel routing algorithms grounded on agent and cryptographic techniques, namely, location dissemination service (LoDiS) routing with agent AES, A-LoDiS with agent AES routing, and B-LoDiS with agent AES routing, ensuring optimal results with respect to various network routing parameters. The algorithm along with efficient routing ensures higher degree of security. The security level is cited testing with respect to possibility of malicious nodes into the network. This paper also aids, with the comparative results of proposed algorithms, for secure routing in ICMANET.

## 1. Introduction

The era of network started with the traditional wired networks, that is, lasting from many decades, that communicate through physical medium. It leads to wireless networks where communication does not enfold the physical medium. Another form of network evolved where nodes are intended to move within the network named mobile networks. Subsequent to which is the wireless sensor networks (WSN) where the communication channels operate through the inbuilt wireless sensor devices within the nodes. A new form of network raised as a challenge for routing called ad hoc network with a special feature of dynamically changing topology. Mobile ad hoc networks (MANET) [1–4] run through into existence where the topology keeps changing frequently in addition to the mobile nature of nodes. At present, the intermittently connected mobile ad hoc networks (ICMANET) [5] explored a new era where the connectivity between nodes never occurs with high density of nodes resulting in spasmodic environment.

The routing in all forms of networks is possible through traditional routing schemes like distance vector routing, link state routing (LSR), open shortest path first (OSPF), opportunistic adaptive routing, distance source routing (DSR) [6], ad hoc on demand distance vector (AODV) [7], and destination-sequenced distance vector (DSDV) [8]. But these schemes are not applicable for ICMANET.

In general ICMANET is a delay tolerant network (DTN) [9, 10] capable of holding larger delays. It is designed to operate effectively over extreme distances such as those encountered in space communications or an interplanetary scale. The sparse or dense nature of intermittent network is mainly due to high mobility of the nodes. The nodes stir within fractions of time and hence their topology changes in a dynamic way. Typical examples of intermittent network are wild life tracking, habitat monitoring sensor networks, military networks, nomadic community networks, vehicular networks, and so forth. Due to typical distorted nature of the network, routing becomes an onerous task.

Many routing algorithms are proposed for efficient routing in the intermittent network, namely, flooding [5], epidemic [11], direction based routing, adaptive routing, utility based routing, probabilistic routing, copy case routing, spray and wait [12] routing, and so forth. These algorithms furnish a proficient channel for data transmission. But they do not clear out a way for efficient secure routing.

The secure communication in MANET adheres to enormous technical challenges due to unique characteristics of MANET. In addition to that it opens for eaves dropping due to intermediate communication and also holds many security threats. Hence security in MANET is adopted using intrusion detection system (IDS) or by creating a trust [13] based environment. The IDS shows extended varieties like behavior based IDS [14], knowledge based IDS [14], distributed IDS [15], real-time IDS [15], multilayer integrated anomaly detection system [4], clustering approach [16], mobile based detection system [17], cooperative approach, [18] and so forth. These known systems enhance the secure communication in MANET.

The security protocols designed for MANET are not suited for ICMANET which is mainly due to the disconnected nature of the network. In the trust based system all IDS must comprise a central server to ensure a node's authentication. The impossibility of setting central server in ICMANET does not adopt these techniques. Hence a new mode of security protocol is to be designed that does not demand for a central system.

The target of secure routing is to prevent malicious attacks by the intruders. This also prevents unwanted attacks or threats of data in the network. The secure routing should be provided in an efficient manner such that they should not degrade the normal performance of data routing in the networks, that is, increase in delay, higher overhead, and maximum storage capacity; invariant bandwidth should not occur. In this paper, we put forth optimum secure algorithms, namely, LoDiS routing with agent AES (LA), A-LoDiS with agent AES (A-LA) routing, and B-LoDiS with agent AES (B-LA) routing that aims at providing a high range of security with optimum result than the existing protocols. The normal routing with ant and bee operates at milliseconds whereas on agent setup the routing operates at nanoseconds and hence the delay in performing the authentication process will be maintained at an ordinal form. This ensures no degradation in the performance of routing. In this way, an efficient secure communication is proposed here. This paper also frames that B-LA provides the better result in contrast to LA and A-LA. The degree of security is measured with introducing malicious nodes into network and ordeal with the three proposed algorithms.

The data transmission in each proposed scheme undergoes diverse algorithms. LA, A-LA, and B-LA employ location aware routing for delay tolerant networks (LAROD) routing, ant colony optimization (ACO) technique, and bee colony optimization (BCO) technique, respectively. All the three algorithms ensure security by means of agent setup and cryptographic technique. All nodes in this network have agent set within them. The agent performs three tasks in prior to data transmission, namely, node analyzing, data

aggregation, and data broadcasting. The data aggregator is a simple database that holds all the data regarding the respective node. Each node has unique id, passcode, origin, grid card representation, and specific pattern. All these information along with the mobility model are stored in the data aggregator of each node. The node analyzer tests whether a node is malicious node based on these agent parameters mentioned above. Once a node is assured for a trusted node the data broadcaster broadcasts the data packet to the nearest possible node. The agent setup imposes secure communication, to propose secure data transmission; the advanced encryption standard (AES) algorithm is used. This algorithm is mainly chosen for its efficiency in handling timing and power attacks. In depth comparison is made between these three algorithms to show the best of the optimum results delivered by them.

The paper is prearranged as the following sections. Section 2 describes the work related to ICMANET. Section 3 portrays the mechanism of secure routing in LA, A-LA, and B-LA. The simulation results are depicted in the Section 4.

## 2. Related Work

The Intermittently connected network is a new form of emerging network where routing data packets is seemed to be monotonous task. Many research works have proved the possibility of routing in ICMANET. This section provides an overview of routing techniques applicable in the intermittent network. It also conveys the general concept of ACO and BCO algorithms.

*2.1. Routing in ICMANET.* The traditional routing scheme that forms a basis for the routing schemes in ICMANET is the flooding based routing. In this, one node sends packet to all other nodes in the network. Each node acts as both a transmitter and a receiver. Each node tries to forward every message to every one of its neighbors [19]. The result in every message eventually is delivered to all reachable parts of the network.

The Epidemic routing oeuvres on the basis of the traditional flooding based routing protocol, which states that periodic pair-wise connectivity is necessitate for message delivery [2]. The protocol banks on immediate dissemination of messages across the network. Routing occurs based on the node mobility of carriers that are within distinctive position of the network.

The beaconless routing protocol [8] is grounded on the hypothesis where there never exists an intervallic diffusion of beacons into the network. Routing primarily makes a choice of forwarding node in a dispersed modus amidst its neighbors, without any form of erudition about their location or prevalence.

The context aware routing (CAR) [7] algorithm paves the forethought of asynchronous communication in ICMANET. The algorithm endows a basement of organizing the messages in the network. It addresses that the nodes are able to exploit the context information to make local decisions which imparts the good delivery ratios and latencies with less overhead. CAR is pain staked as a general framework

to predict and evaluate context information for superior delivery of messages.

The Brownian gossip [6] is an amalgamation of gossip and the random node mobility which provides a scalable geographical routing. In this routing, each node forwards the query related to other nodes information with certain values of probability. Gossiping is a resourceful approach for information dissemination and is done with a probability, namely, P gossip. The probability value makes certain that the query can reach the secondary nodes in the network with highest probability.

The mobility profile based routing [1] addresses, a hub-level routing method and two versions of user-level routing methods [3]. The routing involves a SOLAR-HUB (sociological orbit aware location approximation and routing) which manipulates the user profiles that aids in hub-level routing.

The direction based geographic routing (DIG) [20] algorithm is grounded on geographic location of packets that are routed in an average approximate ideal path towards destination. The algorithm postulates that when two nodes encounter each other, the nodes exchange the knowledge of their current location, moving direction, and packets. The packets are forwarded to nodes whose distance and moving direction are closest to destination.

The single copy case routing [21]: from its nomenclature it postulates that only a single copy of message packet is carried to destination. The routing scheme includes direct transmission, randomized routing, utility based routing, seek and focus, and oracle based routing.

The multiple copy case [22] scheme deals with the mechanism of spraying a few copies of message and then routing each copy in isolated manner to the destination. The algorithm that holds multiple copy case routing are spray and wait and spray and focus.

The semi-probabilistic routing (SPR) [12] algorithm considers that the network is partitioned into tiny portions that have a stable topology. The protocol upholds the information about host mobility and connectivity changes for more accurate message forwarding.

The contention based routing postulates that the efficiency of routing can be achieved only by taking into account the contention and dead end [23]. The spray select and focus provides a better performance considering the contention and dead ends.

The spray and hop [24] is a routing protocol that holds two phases, namely, spray phase that sprays few copies of message into the network. Hop phase which occurs after the spraying phase, a node that was not able to find the destination, switches to the hop phase.

The spray and wait [11] is a scheme that sprays into the network a fewer number of message copies and waits until one of these nodes that holds the copies reaches the destination. It is simple to implement and can be optimized to achieve the depicted performance.

The LAROD-LoDiS [5] routing is a geographical routing that uses a beaconless routing protocol and a store forward carry technique. It also uses a database to communicate among them to achieve routing. It is done by Gossiping

protocol. It provides constant overhead and higher delivery ratio.

*2.2. Optimization Techniques.* The algorithms adopted in this paper use certain optimization techniques. To enhance the performance of data transfer, optimization techniques, namely, ACO and BCO, are used.

ACO is a form of swarm intelligence, a relative approach to problem solving. The swarm intelligence takes a token of inspiration from the social behaviour of insects and of the other animals. In concord to this ACO [16] takes inspiration from the foraging behaviour of ant species. The general behaviour of ant involves depositing pheromone, a chemical substance used by ant to find a path in search of its food. The pheromone deposition acts as an indicator of the way to other members of the colony.

The ants in general use a stigmergy mode of communication. This communication holds two main attributes, namely, (i) an indirect nonsymbolic form of communication where insects exchange information by modifying their environment and (ii) local information which is assessed by those insects that visits the immediate neighborhood.

The general ACO technique involves initialization, traversing, pheromone deposition, and updating the pheromone.

BCO [25] is an optimization technique under the swarm intelligence, a part of artificial intelligence which is based on the actions of individuals in various decentralized systems. The decentralized system is composed of individual systems that are capable to communicate, cooperate, collaborate, and exchange information among them. BCO [26] is a "bottom-up" [27] approach. Artificial bees are created in BCO that acts as artificial agents inspired by the general behavior of natural bees aiding in the solution for optimization problems.

BCO [28] is inspired by the natural behavior of bees and is a population-based algorithm. The basic idea behind BCO is to create a group of artificial bees. The artificial bees represent agents, each generating a new solution. The process is to generate an optimal solution. The BCO algorithm consists of two phases namely the forward and backward phases respectively. The forward phase is a search phase during which artificial bees undergoes a predefined number of moves, constructing the solution and hence yielding a new solution. The new solution obtained is the partial solution. The artificial bees then start the backward phase, where they share information about their solution with each other. The information sharing is estimated by the objective value function.

The security in network plays an imperative role in preventing threats or data theft by the intruders. The secure measures are led by series of checkpoints to ensure safe data transfer. The privation for security is essential in network communication.

*2.3. Agent Technology.* The agent is a program module that functions incessantly in a meticulous environ [29]. It is proficient in carrying out activities in a supple and intellectual comportment, that is responsive to changes in the surrounding environ. The agent is not a complete program but is

an interface [30] responsible for performing the preassigned chores. Agent is autonomous which takes actions grounded on its innate knowledge and its precedent experiences [31].

On setting agent at each node, security [32] is achieved by incorporating certain agent parameters at each node. The agent parameters [33] are described as follows.

- (i) *Node ID*: a unique identifier for each node in the network.
- (ii) *Passcode*: a common password for the nodes in the network.
- (iii) *Mobility model*: the mobility model of the network topology.
- (iv) *Origin of placement*: the initial placement of each node in the network.
- (v) *Grid card*: an  $n \times n$  matrix in which each grid contains a particular data in it. Each node contains a unique grid.
- (vi) *Pattern formation*: the network is associated with certain geometric silhouettes and using these, each node encompasses a unique pattern.

These agent parameters settle on the security issue coupled in the ICMANET.

Agent is set in each node and it includes three components [33, 34], namely:

- (i) Data Aggregator.
- (ii) Node Analyzer.
- (iii) Data Broadcaster.

*Data Aggregator*. The data aggregator is similar to a database that holds an aggregate of information about all the nodes within the network. It includes detailed information of each and every node. The aggregator holds the agent parameters of every node. In plain, it is just the collector of information.

*Node Analyzer*. The node analyzer analyses whether the node that accept the information is a family node, that is, node that belong to the topology. The analysis is based on the agent parameters that are hoarded in the data aggregator. It selects any one of the parameter in a random manner to conclude that the node is a family node. If a malicious node is sensed, the node analyzer broadcasts the presence of intruder.

*Data Broadcaster*. In the data broadcaster, once after a node is determined to be a family node, it allows the sender or any relay node to transfer the message packet to the secondary relay node. It acts as a gateway that provides access for communication amidst the encountered nodes.

From Figure 1 the architecture and working of agent set at each node. When a secondary node receives a message packet from a primary node, the node analyzer tests whether the node belongs to the home network or not. The node analyzer selects one of the parameters randomly and checks for the authorized node from the data aggregator.

The data aggregator is a database, if a match is found within the data aggregator; it is preceded towards the data

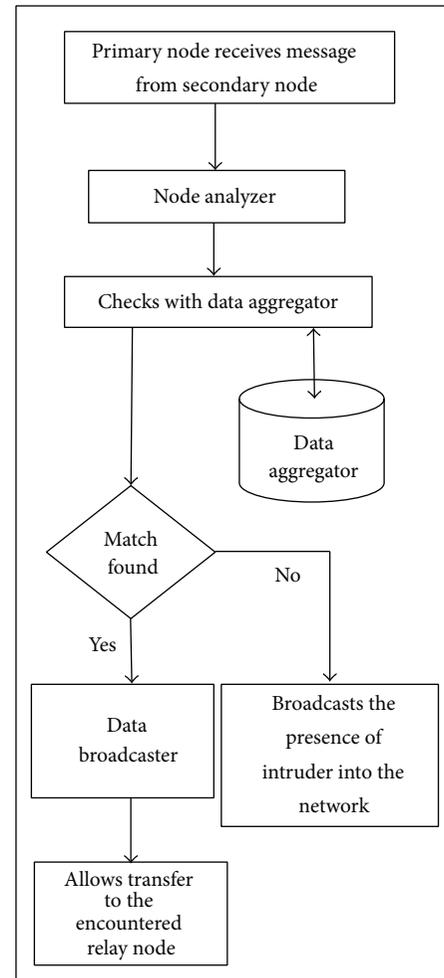


FIGURE 1: Working of agent.

broadcaster. If the node is not valid, node analyzer broadcasts the presence of the intruder within the network. The data broadcaster allows the node to transfer the message packet to the encountered node.

*2.4. AES Algorithm*. The AES algorithm is chosen mainly due to its reliable characteristics of security, cost and code compactness and its design and implementation simplicity. As AES [35] accepts data block sizes of 128, 192, and 256 and a key size of 128 bits, which can be variably expanded, it can accommodate a wide spectrum of security strengths for various application needs. Multiple encryptions use a plural number of keys, since it has been avoided in AES, a reduction on number of cryptographic keys for an application to manage is reduced and hence the design of security protocols and systems are simplified. The AES algorithm is chosen mainly for the following reasons.

- (i) Effectiveness in high speed applications.
- (ii) Simplicity—code compactness.
- (iii) Flexible—varies with the size of input key.
- (iv) Prevent timing and power attacks.

- (v) Used in restricted space environments.
- (vi) Cost effective.

### 3. Routing Mechanisms

In this section a secure communication with the aid of LAROD-LoDiS, A-LoDiS, and B-LoDiS routing is described. An agent is set at each node. During routing, when a sender node A wishes to transmit a data to a destination node X, it initially sends it within its boundary. Node A sends the data only if it confirms that node R1 is a trusted node within the network, that is, it is a node belonging to that particular network. The confirmation on trusted node is done using the agent technology. The agent present at each node generates a test towards R1. The node analyzer of agent at A selects one of the test parameters of agent and passes it to R1. If R1 replies with the correct reply, it is assured to be the trusted node and A passes data towards it. The data actually resides in an encrypted form and sent to R1. R1 just delivers it to another node either relay node  $Rx$  ( $x = 2, 3 \dots n$ ) or destination node X. The encryption and decryption is done by AES algorithm.

**3.1. LAROD Routing with Agent AES.** In general the selections of the relay nodes are done with the help of LoDiS that uses the gossiping technique by which each node can determine the location about its immediate nodes. Hence the destination can be reached by LAROD that uses the store-carry-forward [5] and beaconless technique [5] in aid with LoDiS. Thus, a secure communication is enchanted. The pseudocode for secured LAROD-LoDiS is depicted in Algorithm 1.

**3.2. A-Lodis with Agent AES.** The selections of the relay nodes in A-LA are done with the help of Ant routing scheme using pheromones. Each node in the network acts as an artificial ant. Initially the ants (here ant refers to the node in the network) will be in the sleep mode and the pheromone value (PH) is set to 0. As a data packet is generated at a node A, it searches the relay node in a random manner. The efficiency of the relay node, that is, capability of the relay node to deliver data packet towards destination, is determined using the gossiping method. With the successful delivery of data packet by the relay node, the pH value is incremented each time. For every half minute, when a relay node is inactive, the PH value is decremented. Hence based on PH value the relay nodes are selected and are used for transmission.

LoDiS of A-LA aids in routing by means of the gossiping technique by which each node can determine the location about its immediate nodes. Hence the destination can be reached by Ant routing in aid with LoDiS technique. The pseudocode for secured A-LA is depicted in Algorithm 2 and the Ant routing for A-LA is shown in Algorithm 3.

**3.3. B-Lodis with Agent AES.** The selections of the relay nodes in B-LA are done with the help of bee routing scheme using objective value (OV). Each node in the network acts as an artificial bee. Initially the bees (here ant refers to the node in the network) will be in the sleep mode and the OV is set

TABLE 1: Basic simulation parameters.

Parameters	One simulator
Area	2000 × 2000 m
Mobility model	Random Waypoint
Node density	50 nodes
Node speed	1.5 m/s
Radio range	250 m
Packet life time	600 s

to 0. As a data packet is generated at a node A, it searches the relay node in a random manner. The search of relay node is a step of forward phase. The efficiency of the relay node that is capability of the relay node to deliver data packet towards destination is determined using the OV estimated during the backward phase and the gossiping method. The gossiping is mainly used to have knowledge about the positions of node in the network to transfer data through it. With the successful delivery of data packet by the relay node, the OV value is incremented each time and the efficiency of the path is shared during the backward phase.

LoDiS of B-LA aids in routing by means of the gossiping technique by which each node can determine the location about its immediate nodes. Hence the destination can be reached by Bee routing in aid with LoDiS technique. The pseudocode for secured B-LA is depicted in Algorithm 4 and the B-LA is shown in Algorithm 5.

## 4. Simulation Results

This section describes the simulation results of the proposed algorithms LA, A-LA and B-LA. The performance of these algorithms is described and they are compared with each other to highlight the best of three algorithms. B-LA exerts optimum performance. It outstands LA and A-LA. This variation is depicted in this section evidently. The comparison is made with respect to various network parameters and also it is made in contrast to the influence of the malicious nodes. Section 4.1 clearly shows the scenario setup for evaluation. Section 4.2 expresses the various network parameters under which the evaluation is made. The influence of number of nodes with varying parameters is portrayed in Section 4.3. The Section 4.4 insists the performance with respect to the varying transmission range. The Section 4.5 shows the influence of malicious nodes and the behaviour entrusted by LA, A-LA and B-LA.

**4.1. Scenario Setup.** The parameters set are the basic one simulator [36–38] environ parameters and are given in Table 1. The One Simulation [39], in this paper uses the random waypoint mobility model. The nodes move in an area of 2000 × 2000 m with a speed limit within bounds 0.5 to 1.5 m/s. The radio range is set to 250 m. The efficiency of any routing protocol is determined by the node density that is the total number of nodes within the set network.

The packets are generally generated with the initial setup of the simulation and holds through the overall simulation

```

At Source node,
  Choose the destination node using Location Services
  Perform the authentication mechanism with the node encountered
  Encrypt the data pkt using AES algorithm and broadcast it
  Initialize a timer value for rebroadcasting the data pkt
At Destination node,
  Decipher the received pkt by AES decryption mechanism
  If pkt is received for the first time
    Deliver the pkt
  Transmit ack pkt
At Relay nodes,
  Check whether sender is trusted node
  //done by means of agent test parameters
  Update Location Services with data packet Location Information
  Receive the pkt
  If the pkt is ack
    Resend it to the sender
  Else if the node is within the forwarding area
    If the node does not have copy of pkt
      Initialize timer value for rebroadcasting
    //if current node is ahead of destined node
  If node has copy of pkt
    Remove the pkt.
At ack pkt reception
  Check whether the sender is an authenticated one
  Update location service with ack pkt location information
  If the node has a copy of the pkt
    Remove pkt
When the pkt's rebroadcasting timer expires
  If the pkt's TTL has expired
    Remove pkt
  Else
    Update location information in pkt with location server data
    Broadcast data pkt
    Set up timer for rebroadcasting the pkt
At a set interval broadcast location data
  Select location data: vector with elements (node, location, timestamp)
  Broadcast the data
When a LoDiS broadcast is received
  For each received more recent location data
  Update the entry in the LoDiS server
When location data is received from the routing protocol
  If the supplied information is more recent
  Update the entry in the LoDiS server

```

ALGORITHM 1: Secured LA pseudocode.

time. The time to live (TTL) or the packet life time is set as 600 s initially that are varied lately on consideration to the performance criterion. When evaluating, the simulation is run for 3000 s.

*4.2. Parametric Measures.* This section provides an insight of the various network parameters that are used in the evaluation of LA, A-LA, and B-LA.

Three main parameters are used for evaluation, namely, overhead, delivery latency, and delivery probability. To show the effect of maximum secured routing among LA, A-LA, and B-LA, it is measured with number of malicious nodes

isolated from the network as well as number of packets routed through malicious nodes. These two are evaluated with respect to mobility and number of nodes in the network.

Overhead is one of the main constraints that are to be contemplated for efficient routing. Overhead is any combination of excess or indirect computation time, memory, bandwidth, or other resources that are required to attain a particular goal.

In general the latency is defined as the amount of time it takes for a packet to travel from source to destination. The delivery latency is the time interval taken for the source or any relay node to reach the destination. In general due to the sparse nature of the ICMANET, that is, due to its highly

```

At Source node,
  Choose the destination node using Ant routing
  Perform the authentication mechanism with the node encountered
  Encrypt the data pkt using AES algorithm and broadcast it
  Initialize a timer value for rebroadcasting the data pkt
At Destination node,
  Decipher the received pkt by AES decryption mechanism
  If pkt is received for the first time
    Deliver the pkt
  Transmit ack pkt
At Relay nodes,
  Check whether sender is trusted node
  //done by means of agent test parameters
  Update information at encountered node
  //done by means of gossiping technique

```

ALGORITHM 2: Pseudocode of A-LA routing.

```

Ant in sleep mode
  PH = 0;
Generate packet at node A
Random mode()
//node move in random manner in search of another node to deliver the data packet
do
  Relay Ant()
  //chose the relay node by using gossiping technique
  Trail()
  PH += 1;
Until destined node is reached
if Relay Ant is inactive
  for every t == 30 sec
    PH --;
if PH == 0
  node is ahead of transmission

```

ALGORITHM 3: Pseudocode for Ant routing in A-LA.

```

At Source node,
  Choose the destination node using B-LA
  Perform the authentication mechanism with the node encountered
  Encrypt the data pkt using AES algorithm and broadcast it
  Initialize a timer value for rebroadcasting the data pkt
At Destination node,
  Decipher the received pkt by AES decryption mechanism
  If pkt is received for the first time
    Deliver the pkt
  Transmit ack pkt
At Relay nodes,
  Check whether sender is trusted node
  //done by means of agent test parameters
  Update information at encountered node
  //done by means of gossiping technique

```

ALGORITHM 4: Pseudocode of secure B-LA.

```

Initialize
  OV = 0;
Repeat
  For all nodes
    Bee_Route();
Until route found
Bee_Route()
  If no route found
    Fwd_phase()//search for relay nodes
  End if
  If route found
    Bwd_phase()//estimate the path using OV
    OV += 1;
  End if
  For OV = max
    Deliver data packet
End Bee_Route()
    
```

ALGORITHM 5: Pseudocode for B-LA.

mobile nature, the time period to deliver the data packets is desirably high.

Third metric is the probability to deliver the data to the desired location at specified speed. Under these parametric considerations the protocol performance is evaluated.

4.3. *Influence on Number of Nodes.* The variations in number of nodes indicate a great impact on the performance. The node density is varied from the initial setup of 50 nodes to 250 nodes. As this gets increased, each routing protocol shows the influence made by the number of nodes on overhead, delivery latency, and delivery probability. The result variations are depicted pictorially in the following graphs.

The variation on overhead ratio of LA, A-LA, and B-LA are portrayed in Figure 2. The Figure 2 clearly shows that B-LA incurs a minimum overhead that is acceptable. LA and A-LA exerts slightly higher ratio of overhead than B-LA. Even though three algorithms provide optimum that is acceptable range of overhead B-LA seems to be minimum. The main reason is that B-LA uses the objective value to estimate the efficiency of the route selected. Also since it uses the backward propagation to evaluate the route it delivers minimum overhead than the other two. Figure 2 says B-LA shows 44.62% (approx) of overhead whereas LA and A-LA show a slight higher value of 57.036% (approx) and 52.87% (approx).

In general the delay in ICMANET is large and the routing protocol should tolerate the delays. The LA, A-LA have trivial higher delivery delay when compared to B-LA. The Algorithm 4 clearly portrays the delay in LA and A-LA is higher when compared to that of B-LA. It concludes that B-LA exerts minimum delay in delivering data packets. From Figure 3 it is understood that B-LA exerts 34% (approx) of delay in delivering data packets. LA and A-LA incur delivery delay at a rate of 45% (approx) and 48.05% (approx). Hence B-LA results in better performance than LA and A-LA.

The delivery rate in B-LA is higher when compared to LA and A-LA. The Figure 4 shows the variations on the three

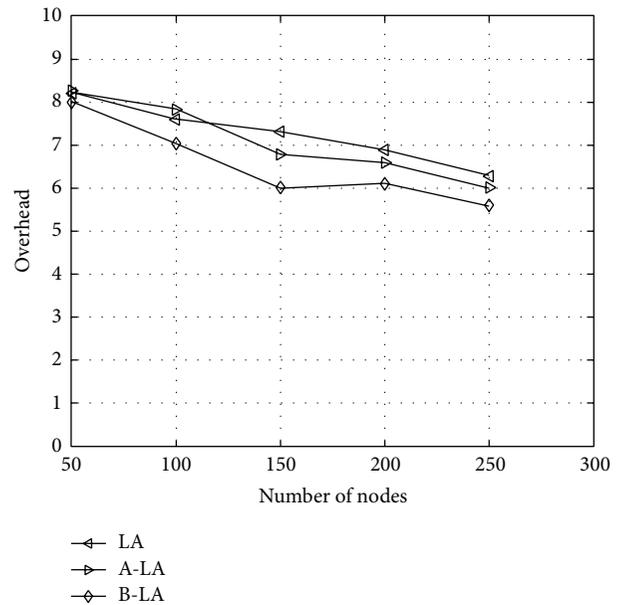


FIGURE 2: Overhead with respect to number of nodes.

protocols and it evidently depicts that B-LA incurs 96.8% (approx) delivery of data packets towards the destined region. In contrast A-LA and LA deliver data packets at an average rate of 95% (approx) and 90% (approx), respectively. As a whole B-LA delivers data at better rate compared to LA and A-LA.

4.4. *Influence on Transmission Range.* Transmission range has a greater impact in the behaviour of routing protocols. The three protocols LA, A-LA and B-LA are evaluated varying the transmission range from 50 to 250. The transmission range illustrates the coverage region of particular node to transmit the data packet. The metrics latency, delivery ratio and overhead are measured and compared for evaluating

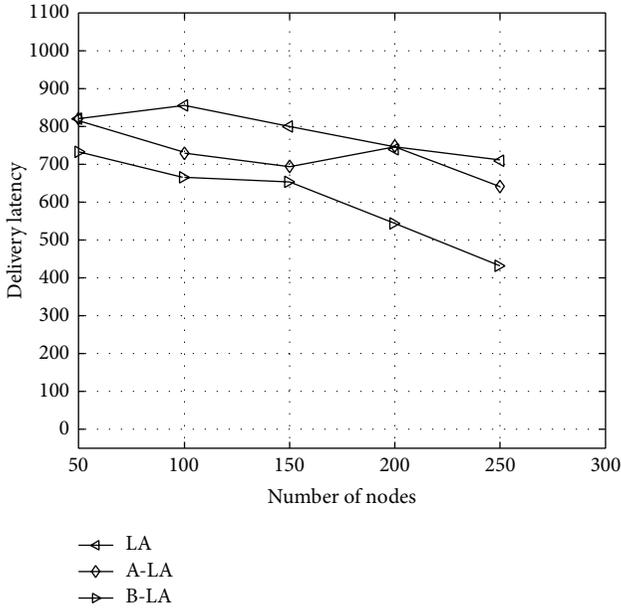


FIGURE 3: Delivery latency with respect to number of nodes.

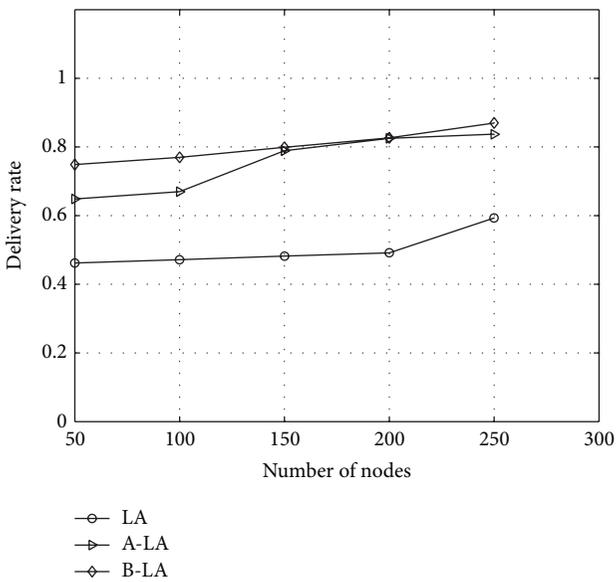


FIGURE 4: Delivery rate with respect to number of nodes.

the higher performance among three protocols LA, A-LA, and B-LA. On varying the transmission range, the protocols show the behaviour as depicted in Figure 5. Among the three proposed protocols, B-LA wield minimum overhead compared to LA and A-LA. They vary periodically in the ratios described in subsequent. LA and A-LA have an overhead of 36.9% (approx) and 34.005% (approx). B-LA has overhead ratio at an average rate of 29.66% (approx). Hence B-LA shows minimum overhead when compared with LA and A-LA.

In general the delay in ICMANET is large and the routing protocol should tolerate the delays. The LA and A-LA have

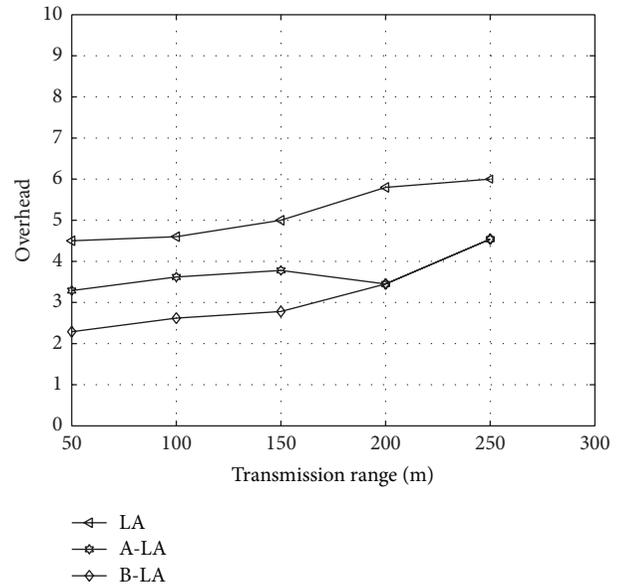


FIGURE 5: Overhead with respect to transmission range.

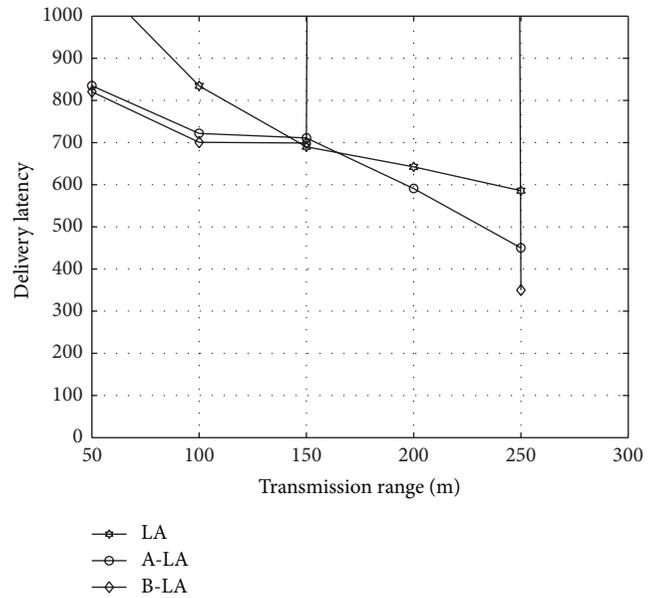


FIGURE 6: Delivery latency with respect to transmission range.

higher delivery delay when compared to B-LA. Since in B-LA the backward propagation ensures estimating the capacity of the route chosen to deliver the data accurately to the destined node in timely manner, B-LA has better results in contrast to LA and A-LA. The Figure 6 clearly portrays the delay in LA and A-LA is higher when compared to that of B-LA. LA has a delay of 41.48% (approx), A-LA has 37.15% (approx) whereas B-LA has 31.808% (approx) of delay in an average rate. It concludes that B-LA exerts minimum delay in delivering data packets.

The delivery rate in B-LA is higher when compared to LA and A-LA. The Figure 7 shows the variations on the three

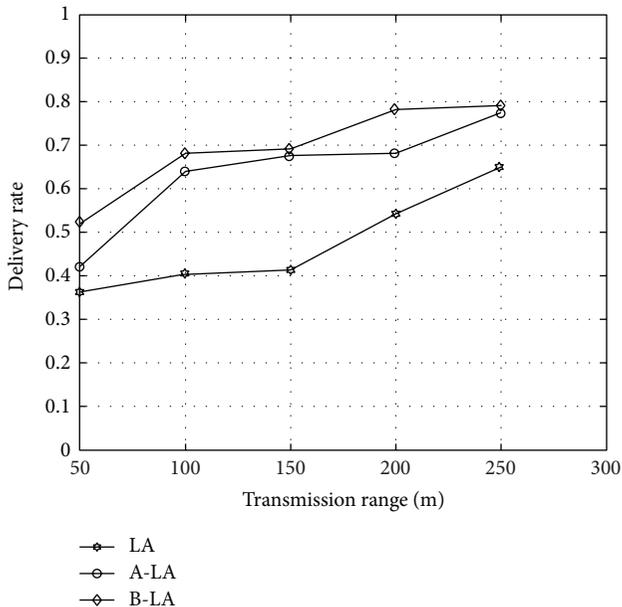


FIGURE 7: Delivery rate with respect to transmission range.

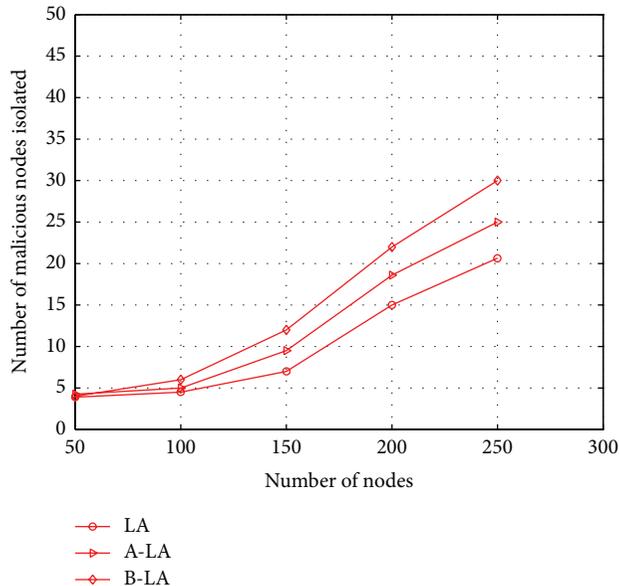


FIGURE 9: Number of malicious nodes isolated with respect to number of nodes.

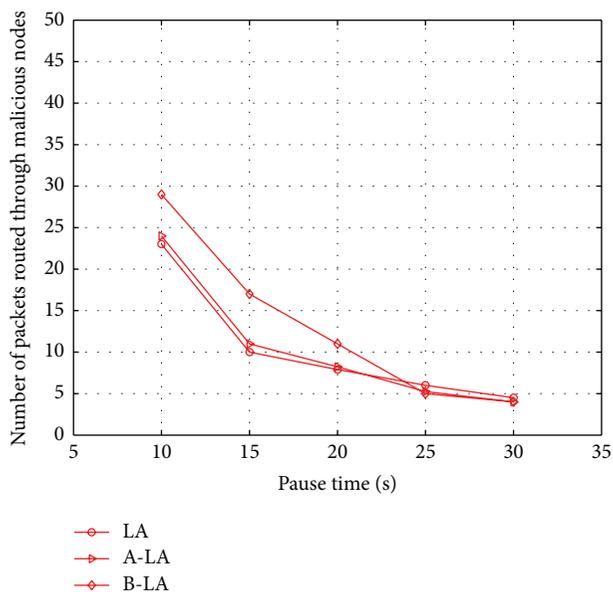


FIGURE 8: Number of malicious nodes isolated with respect to mobility.

protocols and it evidently depicts that B-LA incurs 97.09% (approx) delivery of data packets towards the destined region. In contrast A-LA and LA deliver data packets at an average rate of 94.814% (approx) and 93.73% (approx), respectively. As a whole B-LA delivers data at better rate compared to LA and A-LA.

4.5. Influence of Malicious Nodes. The secure routing through LA, A-LA, and B-LA is evaluated with number of malicious

nodes isolated from network and the number of packets routed through such malicious nodes. The isolated malicious nodes form network depicts the potency of secure routing scheme with detection of intruders in to the network. The Figures 8 and 9 portray the possibility of the proposed protocols to detect the malicious node in the network at higher rate. It is mainly due to the fact that these protocols use the authentication terminologies. When a node fails to meet the authentication terminologies, it is suspect to be a malicious node in the network. Here comparison is made to evaluate the best among LA, A-LA, and B-LA in secure transmission. In these protocols as the authentication scheme uses sharing process, when the number of nodes is increased in the network, more malicious nodes are detected. Figures 10 and 11 say the ratio of the protocols in detecting the presence of malicious nodes in the network with respect to mobility and number of nodes. B-LA detects at a rate of 78% (approx), A-LA detects at a rate of 76% (approx) whereas LA detects at 66% (approx) with respect to mobility. B-LA detects at a rate of 75% (approx), A-LA detects at a rate of 72% (approx) whereas LA detects at 69.66% (approx) with respect to number of nodes. This result proves that B-LA detects the malicious nodes in the network in an efficient way compared to LA and A-LA.

Considering the fact of number of packets routed through the malicious node, the probability of routing across malicious node is found to be less in B-LA compared to LA and A-LA. It is clearly depicted in Figures 10 and 11, respectively. It clearly says the ratio of the protocols in routing packets through malicious node with respect to mobility and number of nodes. B-LA detects at a rate of 38.716% (approx), A-LA detects at a rate of 33.22% (approx) whereas LA detects at 28.55% (approx) with respect to mobility. B-LA detects at a rate of 36.47% (approx), A-LA detects at a rate of

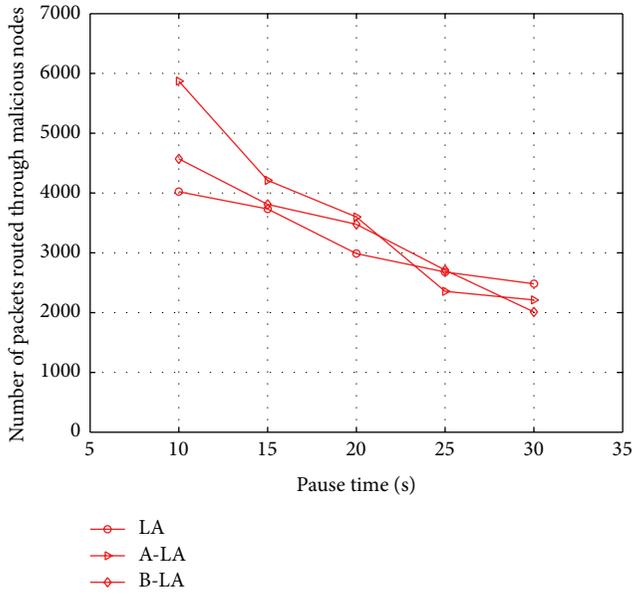


FIGURE 10: Number of packets routed through malicious nodes with respect to mobility.

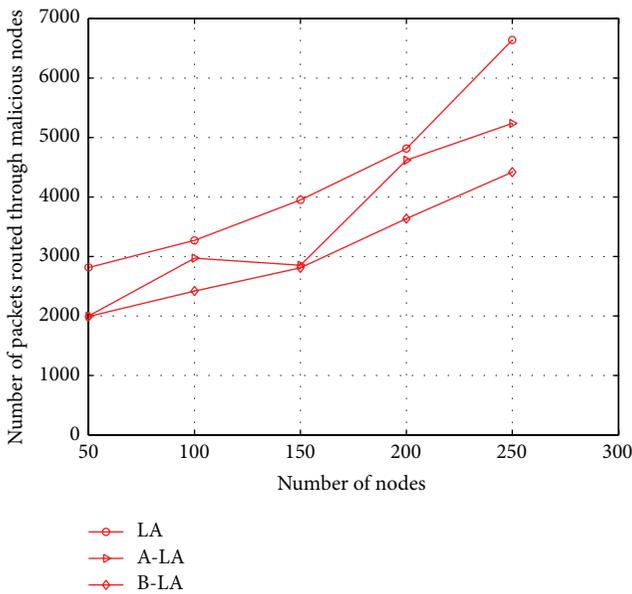


FIGURE 11: Number of packets routed through malicious nodes with respect to number of nodes.

30.56% (approx) whereas LA detects at 28.88% (approx) with respect to number of nodes. This result proves that B-LA routes lesser packets across network through the malicious nodes compared to LA and A-LA. The lesser transmission of packets through malicious nodes is mainly due to the efficient detection of malicious nodes by these proposed algorithms.

Thus, these metrics clearly show that B-LA is an efficient protocol for secure routing in ICMANET among the proposed three algorithms, namely, LA, A-LA, and B-LA.

## 5. Conclusion

In this paper, we have demonstrated the efficient secure routing by means of the 3-way routing protocols LA, A-LA, and B-LA in ICMANET. The proposed routing protocols provide a higher degree of security in the intermittently connected mobile networks. The routing performance metrics are not degraded with the implementation of the security mechanism. This paper also provides a comparative analysis of the performance provided by the three algorithms and clearly predicts that B-LA paves a better way for secure transmission in ICMANET. B-LA outstands LA and A-LA with better delivery ratio, minimum overhead, and latency. It also has higher probability in detecting the presence of malicious nodes and transferring minimum data through these malicious nodes. B-LA has average overhead of 37.14% (approx) in contrast to LA and A-LA with respect to number of nodes and transmission range. It has a maximum delivery ratio of 96.945% (approx) with respect to number of nodes and transmission range. Considering number of nodes and transmission range, B-LA has a minimum delay of 32.904% (approx). It also detects proposition of malicious nodes in network and routing through it at an average rate of 76.5% (approx) and 37.593% (approx), respectively. These theoretical and analytical results prove the efficiency of LA, A-LA, and B-LA aiding in secure transmission and also pictures that B-LA provides better means of secure data transfer for ICMANET. This paper proves the efficiency of secure routing in ICMANET. As an extension to the work, multiagent will be set at each node aiding at faster and secure data transfer between nodes than setting a single agent.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

- [1] J. Ghosh, H. Q. Ngo, and C. Qiao, "Mobility profile based routing within intermittently connected mobile ad hoc networks (ICMAN)," in *Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC '06)*, pp. 551-556, July 2006.
- [2] A. Vahdat and D. Becker, "Epidemic routing for partially connected Ad hoc networks," Tech. Rep. CS-2000-06, Duke University, Durham, NC, USA, 2000.
- [3] J. Ghosh, C. Westphal, H. Ngo, and C. Qiao, "Bridging intermittently connected mobile Ad Hoc networks (ICMAN) with sociological orbits".
- [4] S. Bose, S. Bharathimurugan, and A. Kannan, "Multi-layer integrated anomaly intrusion detection system for mobile adhoc networks," in *Proceedings of the International Conference on Signal Processing, Communications and Networking (ICSCN '07)*, pp. 360-365, February 2007.
- [5] E. Kuiper and S. Nadjm-Tehrani, "Geographical routing with location service in intermittently connected MANETs," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 592-604, 2011.

- [6] R. R. Choudhury, "Brownian gossip: exploiting node mobility to diffuse information in ad hoc networks," in *Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 1–5, December 2005.
- [7] M. Musolesi, S. Hailes, and C. Mascolo, "Adaptive routing for intermittently connected mobile ad hoc networks," in *Proceedings of the 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM '05)*, pp. 183–189, 2005.
- [8] M. Heissenbüttel, T. Braun, T. Bernoulli, and M. Wälchi, "BLR: Beaconless Routing Algorithm for Mobile Ad Hoc Networks," *Computer Communications*, vol. 27, no. 11, pp. 1076–1088, 2004.
- [9] Z. Zhang, "Routing in intermittently connected mobile Ad Hoc networks and delay tolerant networks: overview and challenges," *IEEE Communications Surveys and Tutorials*, vol. 8, no. 4, pp. 24–37, 2006.
- [10] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proceedings of the ACM Conference on Computer Communications (SIGCOMM '04)*, pp. 145–157, September 2004.
- [11] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *Proceedings of the ACM SIGCOMM Workshop on Delay-tolerant Networking (WDTN '05)*, pp. 252–259, August 2005.
- [12] K. E. Shi, "Semi-probabilistic routing in intermittently connected mobile ad hoc networks," *Journal of Information Science and Engineering*, vol. 26, no. 5, pp. 1677–1693, 2010.
- [13] F. Yin, X. Feng, Y. Han, L. He, and H. Wang, "An improved intrusion detection method in mobile AdHoc network," in *Proceedings of the 8th IEEE International Symposium on Dependable, Autonomic and Secure Computing (DASC '09)*, pp. 527–532, Chengdu, China, December 2009.
- [14] L. P. Rajeswari, R. A. X. Annie, and A. Kannan, "Enhanced intrusion detection techniques for mobile ad hoc networks," in *Proceeding of the IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES '07)*, pp. 1008–1013, Tamil Nadu, India, December 2007.
- [15] I. Stamouli, P. G. Argyroudis, and H. Tewari, "Real-time intrusion detection for ad hoc networks," in *Proceedings of the 6th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM '05)*, pp. 374–380, 2005.
- [16] K. Samad, E. Ahmed, W. Mahmood, K. Sharif, and A. A. Chaudhry, "Efficient clustering approach for intrusion detection in ad hoc networks," in *Proceedings of the Student Conference on Engineering Sciences and Technology (SCONEST '05)*, pp. 1–6, Karachi, Pakistan, August 2005.
- [17] A. F. Farhan, D. Zulkhairi, and M. T. Hatim, "Mobile agent intrusion detection system for mobile ad hoc networks: a non-overlapping zone approach," in *Proceedings of the 4th IEEE/IFIP International Conference in Central Asia on Internet (ICI '08)*, September 2008.
- [18] H. Otrok, M. Debbabi, C. Assi, and P. Bhattacharya, "A cooperative approach for analyzing intrusions in mobile ad hoc networks," in *Proceedings of the 27th International Conference on Distributed Computing Systems Workshops (ICDCSW '07)*, June 2007.
- [19] D. Çokuslu and K. Erciyeş, "A flooding based routing algorithm for mobile ad hoc networks," in *Proceedings of the IEEE 16th Signal Processing, Communication and Applications Conference (SIU '8)*, pp. 1–5, April 2008.
- [20] Z. Li and H. Shen, "A direction based geographic routing scheme for intermittently connected mobile networks," in *Proceedings of the 5th International Conference on Embedded and Ubiquitous Computing (EUC '08)*, pp. 359–365, December 2008.
- [21] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the single-copy case," *IEEE/ACM Transactions on Networking*, vol. 16, no. 1, pp. 63–76, 2008.
- [22] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the multiple-copy case," *IEEE/ACM Transactions on Networking*, vol. 16, no. 1, pp. 77–90, 2008.
- [23] E. J. JebraJothi, V. Kavitha, and T. Kavitha, "Contention based routing in mobile ad hoc networks with multiple copies," *International Journal of Engineering and Technology*, vol. 2, no. 2, pp. 93–96, 2010.
- [24] W. K. Lai, W. K. Chung, J. B. Tsai, and C. S. Shieh, "Spray and hop: efficient utility-mobility routing for intermittently connected mobile networks," in *Proceeding of the 4th International Conference on Communications and Networking in China (CHINACOM '09)*, pp. 1–5, Xian, China, August 2009.
- [25] L.-P. Wong, M. Y. H. Low, and C. S. Chong, "Bee colony optimization with local search for traveling salesman problem," in *6th IEEE International Conference on Industrial Informatics (INDIN '08)*, pp. 1019–1025, Daejeon, Republic of Korea, July 2008.
- [26] L. P. Wong, M. Y. Hean Low, and C. S. Chong, "An efficient bee colony optimization algorithm for traveling salesman problem using frequency-based pruning," in *Proceedings of the 7th IEEE International Conference on Industrial Informatics (INDIN '09)*, pp. 775–782, IEEE, Wales, UK, June 2009.
- [27] M. H. Saffari and M. J. Mahjoob, "Bee colony algorithm for real-time optimal path planning of mobile robots," in *Proceedings of the 5th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW '09)*, pp. 1–4, IEEE, September 2009.
- [28] L. Wong, M. Y. H. Low, and C. S. Chong, "A bee colony optimization algorithm for traveling salesman problem," in *Proceedings of the 2nd Asia International Conference on Modelling and Simulation (AMS '08)*, pp. 818–823, May 2008.
- [29] I. Stamouli, *Real-time intrusion detection for ad-hoc networks [M.S. dissertation]*, University of Dublin, 2003.
- [30] I. M. Hegazy, T. Al-Arif, Z. T. Fayed, and H. M. Faheem, "A multi-agent based system for intrusion detection," *IEEE Potentials*, vol. 22, no. 4, pp. 28–31, 2003.
- [31] K. Ioannis, T. Dimitriou, and F. C. Freiling, "Towards intrusion detection in wireless sensor networks," in *Proceedings of the 13th European Wireless Conference*, Paris, France, April 1997.
- [32] O. Kachirski, R. Guba, D. Schwartz, S. Stoecklin, and E. Yilmaz, "Casebased agents for packet-level intrusion detection in Ad-hoc networks," in *Proceedings of the 17th International Symposium on Computer and Information Sciences*, pp. 315–320, CRC Press, October 2002.
- [33] S. Ramesh, R. Indira, R. Praveen, and P. Ganesh Kumar, "Agent technology for secure routing in intermittently connected MANETs," in *Proceedings of the National Conference on Recent Advances in Computer Vision and Information Technology (NCVIT '13)*, 2013.
- [34] L. Besson and P. Leleu, "A distributed intrusion detection system for ad-hoc wireless sensor networks: the AWISSENET distributed intrusion detection system," in *Proceedings of the 16th International Conference on Systems, Signals and Image Processing (IWSSIP '09)*, IEEE, June 2009.

- [35] W. Stallings, *Cryptography and Network Security, Principles and Practices*, Pearson Education, Upper Saddle River, NJ, USA, 4th edition, 2005.
- [36] A. Keranen, J. Ott, and T. Karkkainen, "The one simulator for DTN protocol evaluation," in *Proceedings of the 2nd International Conference on Simulation Tools and Techniques (ICST '10)*, ACM, May 2010.
- [37] A. Keranen, T. Karkkainen, and J. Ott, "Simulating mobility and DTNs with the ONE," *Journal of Communication*, vol. 5, no. 2, pp. 92–105, 2010.
- [38] A. Keranen, "Opportunistic network environment simulator," Special Assignment Report, Department of Communications and Networking, Helsinki University of Technology, 2008.
- [39] TKK/COMNET, "Project page of the ONE simulator," 2009, <http://www.netlab.tkk.fi/tutkimus/dtn/theone/>.

## Research Article

# Design and Performance Evaluation of a Distributed OFDMA-Based MAC Protocol for MANETs

Jaesung Park,<sup>1</sup> Jiyoung Chung,<sup>2</sup> Hyungyu Lee,<sup>2</sup> and Jung-Ryun Lee<sup>2</sup>

<sup>1</sup> Department of Information Security, The University of Suwon, Suwon 445-743, Republic of Korea

<sup>2</sup> School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 156-756, Republic of Korea

Correspondence should be addressed to Jung-Ryun Lee; jrlee@cau.ac.kr

Received 8 April 2014; Accepted 21 June 2014; Published 16 July 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Jaesung Park et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, we propose a distributed MAC protocol for OFDMA-based wireless mobile ad hoc multihop networks, in which the resource reservation and data transmission procedures are operated in a distributed manner. A frame format is designed considering the characteristics of OFDMA that each node can transmit or receive data to or from multiple nodes simultaneously. Under this frame structure, we propose a distributed resource management method including network state estimation and resource reservation processes. We categorize five types of logical errors according to their root causes and show that two of the logical errors are inevitable while three of them are avoided under the proposed distributed MAC protocol. In addition, we provide a systematic method to determine the advertisement period of each node by presenting a clear relation between the accuracy of estimated network states and the signaling overhead. We evaluate the performance of the proposed protocol in respect of the reservation success rate and the success rate of data transmission. Since our method focuses on avoiding logical errors, it could be easily placed on top of the other resource allocation methods focusing on the physical layer issues of the resource management problem and interworked with them.

## 1. Introduction

The orthogonal frequency-division multiple access (OFDMA) has received attention as a promising air interface for next-generation wireless systems by providing high data rates while supporting good coverage and mobility [1]. OFDMA can simultaneously satisfy the communication requirements of multiple mobile stations by allocating one or more subcarriers to each mobile station at the same time unit (OFDMA symbol) [2]. Since OFDMA provides the flexibility in radio resource management, OFDMA has been adopted in many infrastructure based wireless networks such as IEEE 802.16 [3] and 3GPP LTE-A [4]. On the other hand, the wireless mobile ad hoc multihop communication paradigm is also receiving attention as a solution, not only for extending the coverage of wireless communications, but also for enhancing the quality of communication services in shadow areas [5, 6]. However, unlike an infrastructure-based network, an ad hoc network is self-organized by participating nodes without any regulations of centralized control entities such as base stations (BSs)

or access points (APs). Hence, nodes in an ad hoc network should contend for communication resources, which may result in collisions in resource allocations. The main reason of the collision is that multiple nodes try to use the same resources at the same time. Since OFDMA allows multiple simultaneous communications by allocating a portion of resources to different nodes at the same time, combining the two promising technologies (OFDMA and wireless mobile ad hoc networks) is expected to provide enhanced communication opportunities.

Relay network has been designed by standard bodies to exploit the OFDMA in mobile ad hoc networks [7, 8]. The purpose of these relay networks is to extend the coverage of a BS and to increase the overall system throughput. In this network, a relay node plays the roles of both a BS and a mobile station (MS). However, a distinguishing feature of an ad hoc network is that each node could be not only a data source and destination but also a data forwarder to assist other nodes. Since a relay network evolves from infrastructure-based networks, a relay method is designed to operate with

backward compatibility towards existing systems. Thus, the frame format of the legacy cellular systems, which is strictly divided into an uplink (UL) part and a downlink (DL) part in the time or frequency domain, does not reflect the unique features of an ad hoc network [9, 10]. Furthermore, the relay standards are mainly designed for two-hop communications between an MS and a BS through a relay node and the radio resources are still controlled by a BS. Therefore, it is difficult for a relay network to support more than 2-hop communication and it is not well-suited for an ad hoc network.

In OFDMA, the smallest resource allocation unit is defined by both time and frequency which will be called a protocol data unit (PDU) bin hereafter. In such networks, since every node can be a transmitter and a receiver, there could be multiple resource contentions for a set of PDU bins among different nodes while other nodes are exchanging data. However, since mobile ad hoc networks are required to be self-organized and operate without a centralized coordinator, collisions may occur during resource reservation and data transmission processes. Therefore, the utilization of radio resources would deteriorate unless the resource contention process is orchestrated in a distributed manner by taking into account the characteristics of such networks. So, to increase the utilization of radio resources by fully exploiting the flexibility provided by OFDMA in a mobile ad hoc multihop network while providing enhanced communication experiences to mobile users, an efficient MAC protocol operating in a distributed manner is required.

There are a few proposals on resource management for an OFDMA-based MAC protocol in mobile ad hoc networks. The method proposed in [11] mainly focuses on allocating PDU bins to maximize system throughput without considering medium access control. A signal strength based MAC protocol is proposed to reduce the cochannel interference and signaling overheads [12]. Each node selects a PDU bin to send data according to the interference level in the corresponding receiver. In [13], a resource allocation and conflict correction algorithm for an ad hoc network in a disaster area is proposed by considering two-hop interferences. Under the assumption that the interference range is twice as large as the communication range, the authors propose a resource allocation method to maximize the spatial reuse of resources. The authors in [14] take a cross-layer approach to design a resource management method in OFDMA-based ad hoc networks. They integrate a MAC layer and a routing layer to maximize the network throughput. They allocate resources based on the received signal strength at a physical layer to avoid interference.

Since these proposals focus on the physical layer issues of the resource allocation problem, they could either minimize the cochannel interference or maximize the system throughput by making a node to select a PDU bin based on the signal measurement at a physical layer. However, since it is very difficult (even if not impossible) for all nodes in an ad hoc network to have global information on network states in real-time, multiple nodes could select the same PDU bin while the other nodes are sending and receiving data via this PDU bin. Therefore, the corresponding data transfer would

result in a collision even if each node reserves a PDU bin successfully. In [15], the logical errors are identified in the name of a multichannel hidden terminal problem. However, they extended the IEEE 802.11 MAC to operate in a multichannel environment without considering the important characteristics of OFDMA (i.e., support of simultaneous data transmission or reception to or from multiple nodes). In [16, 17], optimization approaches are taken to solve the resource management problem. In [16], an optimization problem is defined to maximize the throughput of mesh routers in an OFDMA-based mesh backhaul network. They proposed three heuristic methods to solve the optimization problem. However, they formulated the problem in a restricted ad hoc network where each node plays only one of the roles of a source, destination, and a relay. In [17], a convex optimization problem is solved by an interior point method to obtain optimal data routes, subchannel schedules, and power allocations to maximize a weighted sum rate of data communicated over the network. The proposals taking an optimization approach deal with the maximization of long term average utility of a network without considering the packet-level dynamics. Therefore, they could be used to explain the average behavior of data transmissions across the network and be useful in network planning. However, further elaboration is needed when they are adopted to design an online resource allocation method.

In this paper, we propose a distributed OFDMA-based MAC protocol for wireless mobile ad hoc multihop networks focusing on the resource management strategy to avoid logical errors that could happen in a resource management process. Therefore, proposed MAC protocol could operate on top of the other resource allocation methods that mainly focused on the physical layer issues of the resource management problem. To the best of our knowledge, this is the first distributed OFDMA-based MAC protocol for mobile ad hoc multihop network in the sense that the proposed MAC protocol fully exploits the characteristics of OFDMA-based MAC (simultaneous transmission or reception to or from multiple users using just one transceiver) and operates in a fully distributed manner without limitation on maximum hop count. The contributions in this paper are summarized as follows.

- (i) We design a frame format reflecting the characteristics of a mobile ad hoc multihop network in which every node could be a transmitter as well as a receiver. In contrast to relay networks, a frame is not divided into a UL part and a DL part because radio resources might be wasted if the division cannot adapt to asynchronous traffic loads. Instead, we divide a frame into a data part and a control part. In the data part, radio resources are divided into protocol data unit (PDU) bins, each of which could be used for sending or receiving data according to the roles of nodes. The PDU bins are managed in a distributed fashion to support the features of an ad hoc network. All the nodes play equal roles in managing the PDU bins with the information in the control part of a frame.

- (ii) Since the proposed distributed MAC protocol uses the information of 1-hop neighbors not to compromise the autonomous nature of an ad hoc network, there exist some logical errors. We identify the scenarios of the logical errors and divide them into five categories according to their root causes. We present a detailed radio resource management method to avoid the logical errors and it is shown that three types of the logical errors are avoided while two types of them are inevitable under the proposed MAC protocol.
- (iii) The information of 1-hop neighbors is periodically broadcasted on a contention basis. We analyze the effect of the advertisement period on the delay from when a node broadcasts a message to when all of its neighbors successfully receive this message and determine the optimal advertisement period minimizing this delay.

The rest of the paper is organized as follows. In Section 2, we present the frame format and the distributed radio resource management method is detailed in Section 3. After we evaluate the performance of the proposed method through simulation studies in Section 4, we conclude the paper with possible future research directions in Section 5.

## 2. Frame Structure

In this section, we propose a MAC frame structure for a wireless mobile ad hoc multihop network reflecting the necessary distributed resource management procedures. Figure 1 shows the structure of a MAC frame. A frame is divided into a data subframe and a control subframe with time. The data subframe is composed of  $N$  PDU bins. A PDU bin represents the minimum radio resource unit for data transmission and reception, which is made of  $n_s$  subcarriers and  $n_t$  OFDM symbols.

The control subframe is used for the nodes in a network to exchange control information so as to reserve a PDU bin in a distributed manner. According to the type of control information, the control subframe is further divided into a network management unit (NMU) zone, an acknowledgement (ACK) zone, and a read-to-send and clear-to-send (RCTS) zone. There are a number of channels in each zone separated by subcarriers. The number of channels in a zone is independent of those of the other zones, while the size and the function of the channels in the same zone are the same. The NMU zone is composed of  $M$  channels and is used for a node to announce its presence in an area. Each node periodically contends for a channel in the NMU zone to broadcast its presence to its 1-hop neighbors. By inspecting the NMU zone, a node could estimate its 1-hop neighbors in a distributed way. Each channel in the ACK zone corresponds to each PDU bin. Thus, the number of channels in the ACK zone ( $A$ ) is the same as the number of PDU bins ( $N$ ). A channel in the ACK zone is used to control the data transmission and reception through the corresponding PDU bin. The usage of the ACK zone depends on the type of service request from an upper layer. If an upper layer requests a reliable data transfer service between adjacent nodes, a receiver sends

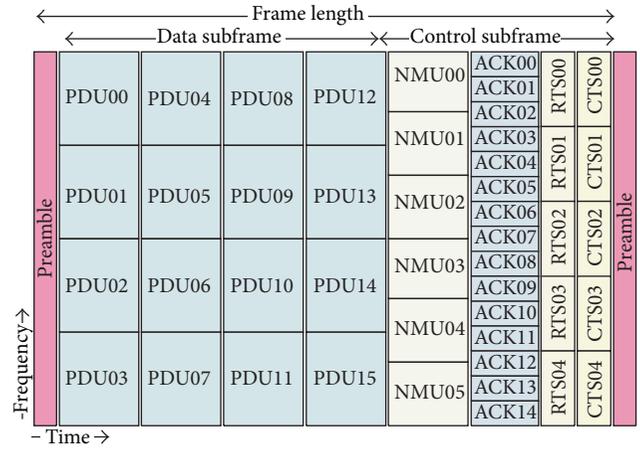


FIGURE 1: Example of frame structure ( $N = 16$  PDU bins,  $M = 6$  NMUs,  $A = 16$  ACKs, and  $B = 5$  RTS-CTS pairs).

an acknowledgement to a sender through the ACK channel when it receives data through the corresponding PDU bin. In contrast, if an upper layer requests a time-sensitive service to serve error-tolerant applications such as VoIP or video streaming, a channel in the ACK zone is used to control the data flow between a sender and a receiver. For example, a receiver may use the channel to feed back a message for flow control or to report the quality parameters of data reception such as the average delay and the data error rate [18–20].

When a node needs to send data, it uses the RCTS zone to reserve a PDU bin. The RCTS zone is further divided into an RTS region and a CTS region with time. Since these channels are used to convey small control frames for resource reservation, there could be  $B$  such pair of channels at the same time. The rationale behind having multiple simultaneous RTS-CTS pairs is to support the characteristics of mobile ad hoc multihop networks, where multiple reservations could be made at the same time. An RTS channel is coupled with a CTS channel. A pair of nodes reserves a PDU bin by exchanging an RTS frame and a CTS frame through a pair of RTS and CTS channels. For example, if a node trying to reserve a PDU bin sends an RTS frame to a receiver through the  $k$ th RTS channel, the receiver must respond by sending a CTS frame via the  $k$ th CTS channel.

A node could use multiple PDU bins for data transmission and reception if the PDU bins are separated in time. However, since we assume that a node has a single radio interface, a node cannot transmit while it is receiving and vice versa. Accordingly, a node cannot use PDU bins with the same OFDM symbols for both data transmission and reception at the same time. For example, a node cannot receive anything through any of PDU bin 04, PDU bin 06, and PDU bin 07, while it is transmitting data using PDU bin 05.

## 3. Operational Procedures

In this section, we propose a distributed radio resource management method using the MAC frame structure introduced

in Section 2. The resource management method is composed of a network state estimation process and a resource reservation process. The network state estimation process operates in a management plane whenever a node receives a frame. The process is used for a node to estimate a set of 1-hop neighbor nodes and a set of PDU bins being used by them. The resource reservation process operates in a control plane to reserve a PDU bin when a node has data to send. The notations we use hereafter are summarized in Notations section.

**3.1. Network State Estimation.** Each node advertises its presence to its 1-hop neighbors using a channel in the NMU zone. Since  $M$  channels in the NMU zone are shared by all the nodes, a collision may occur if more than two nodes in the transmission range of each other select the same NMU channel of the same frame at the same time. In addition to the randomness in selecting an NMU channel in a frame, we introduce additional randomness to reduce the collision probability. As shown in Figure 2, during every time period  $T$ , a node randomly selects a frame among the frames within  $T$  and chooses an NMU channel in the selected frame at random. Whenever a node receives a frame, it keeps updating  $NN_X$  by analyzing the NMU zone of the frame.

A node might manage  $U_X$  by examining the data sub-frame whenever it receives a frame. However, an exposed node problem may occur if a node estimates  $U_X$  by analyzing the data subframe. Since each ACK channel is coupled with its corresponding PDU bin and data frames collide at a receiving node, we take an approach that involves a node managing  $U_X$  by analyzing the ACK zone of each frame it receives.

A node  $X$  successfully receives an NMU message sent by  $Y$  in  $AN_X$  only if the other neighboring nodes in  $AN_X$  do not use the same NMU channel in the same frame selected by  $Y$  and  $X$  is in the receiving mode. Therefore, the probability that  $X$  successfully receives an NMU message sent by  $Y$  is derived as follows. The probability that  $X$  is in the receiving mode when  $Y$  is sending an NMU message is  $P_1 = (T-1)/T$  and the probability that none of the neighbors of  $X$  except  $Y$  sends an NMU message at the same time when  $Y$  broadcasts its NMU message also becomes  $P_1$ . In addition, the probability that  $Z$  is in  $AN_X$  ( $Z \neq Y$ ) and  $Y$  uses different NMU channels even if  $Z$  and  $Y$  simultaneously send their NMU messages becomes  $P_2 = 1/T \times (M-1)/M$ . Therefore, the probability that  $X$  successfully receives the NMU message sent by  $Y$  is given as

$$p_{NMU} = P_1(P_1 + P_2)^{|AN_X|-1} = \frac{T-1}{T} \left( \frac{T-1}{T} + \frac{1}{T} \left( \frac{M-1}{M} \right) \right)^{|AN_X|-1} \quad (1)$$

Since the frame length is  $T_F$ , the average delay that  $X$  receives an NMU message from  $Y$  becomes

$$D_o = T_F \cdot T \sum_{k=0}^{\infty} (k+1) (1 - p_{NMU})^k p_{NMU} = \frac{T_F \cdot T}{p_{NMU}} \text{ (ms)} \quad (2)$$

From (1) and (2), we can derive the optimal period  $T^*$  that minimizes  $D_o$  by solving  $dD_o/dT = 0$  as

$$T^* = \frac{2 - |AN_X|(a-1) + \sqrt{(|AN_X|(a-1))^2 + 4a}}{2}, \quad (3)$$

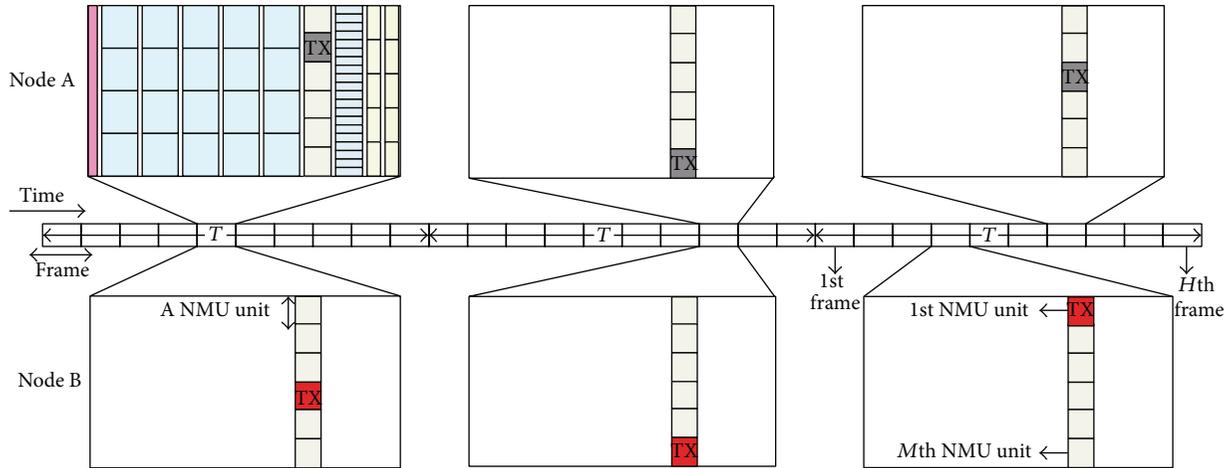
where  $a = (M-1)/M$ . Therefore, if the number of NMU channels and the number of neighboring nodes are given, each node can determine the optimal period  $T^*$ .

To construct an exact  $NN_X$ , a node  $X$  must receive all the NMU messages from all of its neighboring nodes. Since a node could move around, it becomes important to know the average delay until  $X$  makes up an exact  $NN_X$ . The probability that  $X$  receives an NMU message from  $Y$  in  $AN_X$  at least once before the  $k$ th period is given as  $p_a = 1 - (1 - p_{NMU})^k$ . Thus, the probability that  $X$  receives all the NMU messages from all the nodes in  $AN_X$  before the  $k$ th period is given as  $P_b(k) = p_a^{|AN_X|}$ . Therefore, the average delay that a node receives NMU messages from all of its neighbors becomes

$$D_a = \sum_{k=1}^{\infty} \frac{kdP_b(k)}{dk} \quad (4)$$

**3.2. Distributed Resource Reservation Process.** A node with data to send starts a resource reservation process by selecting a PDU bin to reserve. Among the  $N - |U_X|$  available PDU bins, a node  $X$  randomly chooses a PDU bin  $i$  (i.e.,  $i \notin U_X$ ), where  $|U_X|$  denotes the cardinality of the set  $U_X$ . Then, the node  $X$  selects an RTS channel  $j$  at random among the unused RTS channels and sends  $R_i^j$  to a receiver node  $Y$ . When  $Y$  successfully receives  $R_i^j$  from  $X$  and no possible errors are detected, it answers with  $C_i^j$ , notifying  $X$  of successful reservation of the PDU bin  $i$ . If  $X$  successfully receives  $C_i^j$  from  $Y$ , the PDU bin  $i$  is reserved between  $X$  and  $Y$ . After the reservation,  $X$  sends data to  $Y$  through the PDU bin  $i$ . When  $Y$  receives data without an error,  $Y$  sends an acknowledgement using the ACK channel  $i$  that corresponds to the PDU bin  $i$  if an upper layer requests a reliable data transfer service. If an error is reported, the sender  $X$  immediately retransmits the data frame using the same PDU bin used before. After successful data transmission between the sender and the receiver, the reserved PDU bin is returned. However, if the size of data is larger than the size of the PDU bin, the PDU bin is preempted. In other words, the sender sends data consecutively through the reserved PDU bin without additional exchange of RTS/CTS frames.

However, since each node uses only the local information estimated from the control subframes to reserve a PDU bin, the uncertainty in the estimated network state might fail a resource reservation attempt between two nodes. The failures can be classified into two categories. The first category of the failures represents the cases where a control message sent from a node collides with the other control messages sent from the other nodes. We will call this type of failure as a physical error. In contrast, the second category of the failures occurs when a reserved PDU bin leads to the failure of data transfer, even if the PDU bin is reserved in advance between



Case (a) Nodes A and B broadcast at the same frame but using different NMU units. Neighbor nodes can receive both NMU messages but nodes A and B cannot receive the NMU message of each other.  
 Case (b) Nodes A and B broadcast at the same frame using the same NMU unit. Neighbor nodes cannot receive both NMU messages.  
 Case (c) Nodes A and B broadcast at different frames. Neighbor nodes receive both NMU messages. Nodes A and B receive the NMU messages of each other.

FIGURE 2: During time period  $T$ , each node selects a frame randomly among the frames within  $T$  and chooses an NMU channel in the frame at random to broadcast its presence to its 1-hop neighbors.

two nodes through successful RTS/CTS frame exchange. We will call this type of failure a logical error. When a physical error occurs, our protocol operates as follows.

(i) *RTS Frame Collision.* If  $Z \in NN_Y$  sends an RTS frame using the RTS channel  $j$  when  $X$  sends an RTS frame to  $Y$  through the same RTS channel  $j$ , the two RTS frames collide with each other at  $Y$ . Since a node cannot receive a frame while it is sending,  $X$  cannot detect the collision, even if  $Z \in NN_X$ . However, when the RTS frame from  $X$  collides,  $Y$  cannot respond with a CTS frame. If  $X$  does not receive a CTS frame from  $Y$  at the following CTS zone, it considers that the previous reservation request has failed. When  $X$  detects the RTS collision,  $X$  retransmits an RTS frame with the probability  $p_{rt} = 1/(NN_X + 1)$  so as to avoid successive collisions.

(ii) *CTS Frame Collision.* If  $Z \in NN_X$  sends a CTS frame through the CTS channel  $j$  when  $Y$  sends a CTS frame to  $X$  through the same CTS channel  $j$ , the two CTS frames collide with each other at  $X$ . Since  $X$  is in the listening mode,  $X$  could detect the collision of the CTS frames. After detecting the collision,  $X$  restarts the reservation process with probability  $p_{rt}$ .

The logical errors occur because multiple node pairs could begin their resource reservation processes at the same time, while the other nodes are sending and receiving data. Since a logical error means that data transmission with a reserved PDU bin inevitably fails, network resources are wasted once a logical error takes place. However, since the uncertainties in estimating network states cannot be eliminated completely, we try to avoid logical errors in the resource reservation process by checking the RTS region of the frame

once a node receives an RTS frame. The rationale behind this operation is that a node requesting a reservation does not know the resource usage situations of 1-hop neighbors of its receiver and data collides only at a receiver node. In Table 1, we classify the types of logical errors according to the situations in which they occur and describe how our protocol operates when it detects them. When a node detects a logical error, it begins another reservation process without random backoff to expedite the process. In the following, we explain the operational procedures for detecting and treating each type of logical error in detail.

(iii) *Logical Error Type 1 (LET1).* When a node  $X$  sends an RTS frame to  $Y$ , it selects a PDU bin  $i$  randomly among those that are not in  $U_X$ . Therefore, a logical error type 1 occurs if  $i$  is in  $U_Y$  and a node in  $NN_Y$  that reserved the PDU bin  $i$  keeps using the PDU bin  $i$  when  $X$  sends data to  $Y$  through the same PDU bin  $i$ . Since  $Y$  cannot know how long the PDU bin  $i$  will be occupied, we take a conservative approach to avoid this type of logical error. When  $Y$  receives a request to reserve a PDU bin  $i$  from  $X$ ,  $Y$  checks whether or not PDU bin  $i$  is in  $U_Y$ . If  $i \in U_Y$ , the data transmission from  $X$  through PDU bin  $i$  might collide with the other data transmission by a node in  $NN_Y$  at  $Y$ , even if  $X$  and  $Y$  reserve PDU bin  $i$  successfully. Thus,  $Y$  informs  $X$  of an LET1 by sending a CTS frame to  $X$  if the PDU bin  $i$  is in  $U_Y$ . When  $X$  receives an LET1 from  $Y$ ,  $X$  starts the resource reservation process again by selecting another available PDU bin  $j$  randomly.

(iv) *Logical Error Type 2 (LET2).* When multiple adjacent node pairs try to reserve the same PDU bin using different RTS channels at the same time, a logical error type 2 may occur. We illustrate an example scenario in Figure 3. Since

TABLE 1: The types of logical errors ( $X$ : a node requesting a reservation of a PDU bin and  $Y$ : a node receiving a resource reservation request from  $X$ ).

	Cause	Detection method	Actions on detection
LET1	When $X$ selects a PDU bin $i$ for reservation, it could not know $U_Y$ .	Inspecting $U_Y$ .	$Y$ sends a CTS frame with error type LET1. $X$ restarts the reservation process.
LET2	Multiple node pairs attempt to reserve the same PDU bin using different RTS channels at the same time.	$Y$ checks the RTS region of the frame from which it receives a resource reservation request.	$Y$ sends a CTS frame with error type LET2. $X$ restarts the reservation process.
LET3	Multiple nodes send RTS frames simultaneously to the same node $Y$ through different RTS channels to reserve the same PDU bin.	$Y$ checks the RTS region of the frame from which it receives resource reservation requests.	$Y$ randomly selects a winner and assigns another PDU bin to a loser. If $X$ is a loser, it uses the assigned PDU bin if it is available. If not, it restarts the reservation process.

$A$  and  $C$  are two-hop neighbors to each other, they could select the same PDU bin 1 when they begin their resource reservation process. Figure 3(a) shows a situation where a node  $A$  is sending  $R_1^2$  to a node  $D$ , while a node  $C$  is sending  $R_1^1$  to a node  $B$  to reserve the same PDU bin 1 at the same time. The RTS frame sent from  $A$  to  $D$  also reaches  $B$ . However, since the RTS channel used for  $A$  to send the RTS frame to  $D$  is different from the one that  $C$  used to send its RTS frame to  $B$ , the two RTS frames do not collide at  $B$ . If PDU bin  $i$  is neither in  $U_B$  nor in  $U_D$ ,  $D$  and  $B$  send  $C_1^2$  and  $C_1^1$  to  $A$  and  $C$ , respectively, to inform  $A$  and  $C$  of the successful resource reservation at time  $t + 1$ , respectively (Figure 3(b)). However, the data sent from  $A$  to  $D$  also reaches  $B$ , which makes the data transmission from  $C$  to  $D$  fail, because the two data arrive at  $B$  through the same PDU bin at the same time (Figure 3(c)). To detect this type of logical error, if a node receives a request to reserve a PDU bin from one of its neighbors, it checks not only  $U_X$ , but also  $A_X$ , by investigating the RTS region of the frame from which it receives a resource reservation request. If a node detects an LET2, it sends a CTS frame to the node requesting a resource reservation with an error code LET2. When a node receives a CTS frame with LET2, it begins another resource reservation process. For example, in Figure 3(b), since the PDU bin that  $A$  ( $A \in NN_B \wedge A \neq C$ ) attempts to reserve with  $D$  is in  $A_B$ ,  $B$  perceives an LET2. Then,  $B$  informs  $C$  of LET2 by sending a CTS frame with an error code LET2.  $C$  restarts the resource reservation process by selecting another PDU bin  $j$  randomly so that  $j \notin U_C \wedge j \neq i$ .

(v) *Logical Error Type 3 (LET3)*. When multiple nodes are sending their RTS frames to the same node through different RTS channels at the same time, a logical error type 3 occurs. In Figure 4, we show an example where an LET3 takes place. Since nodes  $A$ ,  $B$ , and  $C$  are located such that  $A$  and  $C$  are members of  $NN_B$ ,  $A \notin NN_C$ , and  $C$  is not in  $NN_A$ ,  $A$  does not know the resource usage situation around  $C$ , nor does  $C$  know  $U_A$ . If  $A$  sends  $R_1^1$  to  $B$  while  $C$  is requesting  $B$  to reserve the same PDU bin 1 by sending  $R_1^2$ , the two RTS frames do not collide at  $B$ . Since  $B$  checks the RTS region of the received frame,  $B$  can detect that both  $A$  and  $C$  are trying to reserve a PDU bin 1. In this case,  $B$  could send RTS frames with an error code LET3 to fail

both of the reservation requests. However, to reduce the overhead incurred by repeated resource reservation attempts,  $B$  randomly selects a winner. If  $A$  is selected as a winner,  $B$  sends  $C_1^1$  to  $A$  to inform that the PDU bin 1 has successfully been reserved. On the other hand,  $B$  may send  $C_1^2$  to  $C$  with an error code LET3 for  $C$  to restart the reservation process. However, to further reduce the overhead of repeated reservation attempts, in our protocol,  $B$  sends  $C_1^2$  to  $C$  with a new PDU bin 3 that is neither in  $U_B$  nor in the RTS region of a frame received at time  $t$  (Figure 4(b)). If  $C$  receives a CTS frame containing a PDU bin that is not the same as the one it requested,  $C$  detects an LET3. Then,  $C$  uses the assigned PDU bin 3 if it is not in  $U_C$ . Otherwise, it restarts the resource reservation process with another randomly selected PDU bin.

*3.3. Inevitable Resource Reservation Errors.* In our protocol, when a node  $X$  receives a resource reservation request, it investigates  $U_X$  and  $A_X$  to detect a logical error. Each node manages the local information ( $U_X$  and  $A_X$ ) by checking the control subframe of every frame that it receives. However, since each node competes for the NUM zone and the RCTS zone in a control subframe, estimation errors are involved in  $U_X$  and  $A_X$ . In addition, a node cannot know the resource usage situation of its 2-hop neighbors. Therefore, there might be logical errors that could not be detected using only the information of 1-hop neighbors. According to the causes of errors, these logical errors can be further divided into two types. The first one is called a logical error type 4 (LET4) and is attributed to the fact that the  $U_X$  and the  $A_X$  might not exactly reflect the actual resource usage states of 1-hop neighbors all the time. The second one is denoted by a logical error type 5 (LET5) and takes place because a node does not know the states of its 2-hop networks.

In Figure 5, we illustrate a situation where an LET4 occurs. At time  $t$ ,  $A$  sends  $R_1^1$  to  $B$ ,  $C$  sends  $R_1^2$  to  $D$ ,  $E$  sends  $R_2^2$  to  $G$ , and  $H$  sends  $R_1^2$  to  $F$  to start the resource reservation processes simultaneously (Figure 5(a)). After receiving an RTS frame from  $A$ ,  $B$  first checks  $U_B$  to detect an LET1. If the PDU bin 1 requested by  $A$  is not in  $U_B$ ,  $B$  checks  $A_B$  to detect LET2 and LET3. At time  $t$ , the RTS frames sent by  $C$  and  $E$  also reach  $B$ , even if they are destined to 2-hop neighbors of  $B$ . However, since the two frames are sent through the same

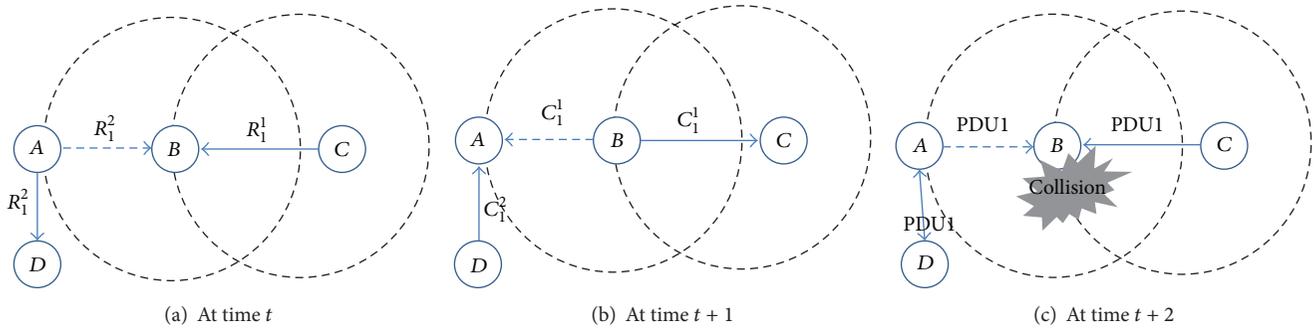


FIGURE 3: An example of operational procedure for a node to detect and treat an LET2. (In the following figures, we use frame time units. A circle represents the transmission range of a node located at the center of the circle. The solid arrows represent the intended direction of a frame and the dotted arrows represent the propagation of a frame to the nodes that are not the destinations of the frame. PDU $x$  denotes data sent through PDU bin  $x$ .)

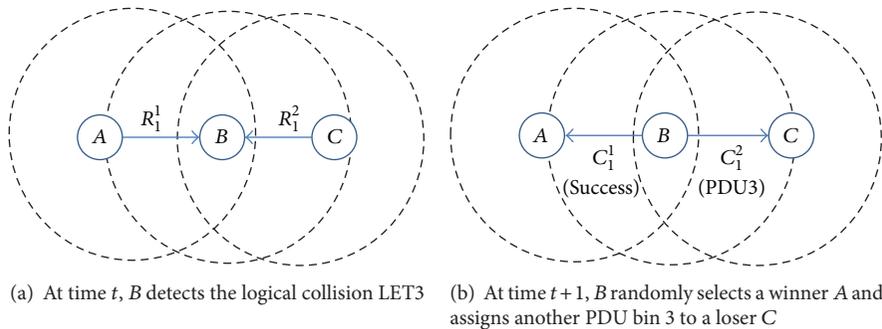


FIGURE 4: An example of operational procedure to detect and treat an LET3. (When  $B$  detects an LET3,  $B$  randomly selects a winner and informs the winner of the successful reservation.  $B$  sends a  $C_1^2$  to  $C$  with another PDU bin 3 that is not in  $U_B$ .)

RTS channel 2, they collide at  $B$ . Therefore,  $B$  cannot know that its neighbor  $C$  is attempting to reserve the same PDU bin 1 that  $A$  requested. In other words,  $B$  makes a mistake that the PDU bin 1 is not in  $A_B$  and sends a CTS frame  $C_1^1$  to  $A$  to confirm the reservation. While the CTS frame  $C_1^1$  sent from  $B$  is arriving at  $A$ , nodes  $D$ ,  $F$ , and  $G$  also send CTS frames to their corresponding nodes at the same time  $t + 1$ . In this case, the CTS frames sent by  $B$  and  $F$  collide at  $C$  because they use the same CTS channel. However, since the CTS frames are not destined to  $C$ , all the reservation attempts succeed at time  $t + 1$  (Figure 5(b)). Consequently, at time  $t + 2$ , nodes  $A$ ,  $C$ ,  $E$ , and  $H$  send their data using the reserved PDU bins. Since nodes  $C$  and  $E$  are in  $NN_B$ , not only the data sent by  $A$  but also the data sent by  $C$  and  $E$  also arrive at  $B$ . Accordingly, the data sent from  $A$  to  $B$  collides with the data sent from  $C$  to  $D$  because  $A$  and  $C$  reserved the same PDU bin 1 (Figure 5(c)). As a result, the data transmission from  $A$  to  $B$  fails, even if they reserved a PDU bin successfully.

In our protocol, when a node  $X$  detects an LET3, it randomly selects a winner and notifies the winner that a resource reservation request has succeeded. The node also sends a CTS frame to the loser with another PDU bin that is different from the one requested by the loser. Instead of failing all the nodes, causing an LET3, our design choice might increase the success rate of a PDU bin reservation

and decrease the singling overhead in a resource reservation process. However, such an operation might bring about another type of logical error called an LET5, because a node cannot know the resource usage states of its 2-hop neighbors.

Figure 6 shows an example scenario of an LET5. At time  $t$ , both  $A$  and  $C$  are sending RTS frames to  $C$  to reserve a PDU bin 1 while  $E$  sends  $R_3^3$  to  $D$  (Figure 6(a)). Since  $A$  is using an RTS channel 1 while  $C$  is sending an RTS frame through an RTS channel 2, the two RTS frames do not collide at  $B$  but cause an LET3. At time  $t + 1$ ,  $D$  confirms the resource reservation request by sending  $C_3^3$  to  $E$ . In contrast,  $B$  detects an LET3, because  $B$  perceives that more than two nodes are asking it to reserve the same PDU bin at the same time. If  $A$  is randomly selected as a winner,  $B$  sends  $C_1^1$  to notify  $A$  of the successful reservation of the PDU bin 1. In addition,  $B$  randomly selects a PDU bin 3 that is neither in  $U_B$  nor in  $A_B$  and sends  $C_3^2$  to  $C$  to expedite the resource reservation process of  $C$  (Figure 6(b)). When  $C$  receives  $C_3^2$ , it detects an LET3 and checks whether or not the assigned PDU bin 3 is in  $U_C$ . Since  $C$  manages  $U_C$  by analyzing the ACK zone of a frame, the PDU bin 3 is not in  $U_C$ , even though  $D$  is in  $NN_C$  and reserves the PDU bin 3 at time  $t + 1$ . As a result,  $C$  sends data to  $B$  using the PDU bin 3 at time  $t + 2$ . However, since the data also reaches  $D$ , the data transmission from  $E$  to  $D$  fails, even though they use the reserved PDU bin (Figure 6(c)).

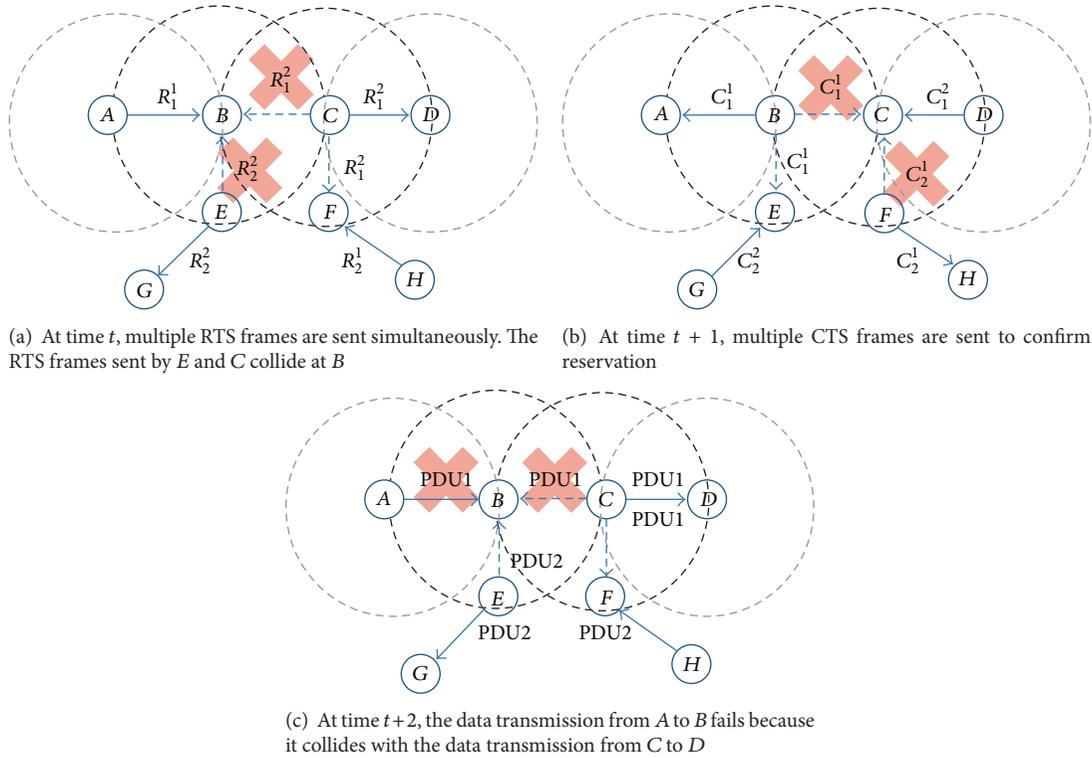


FIGURE 5: An LET4 takes place because local information cannot reflect the network state completely.

#### 4. Performance Evaluation and Discussion

In this section, we evaluate the performance of our MAC protocol by analyzing simulation results in a variety of operational environments. Here, it is noted that the performance of the proposed MAC protocol is not compared to those of previous OFDMA-based MAC protocols since there are no protocols that conform to the basic operational principles of our MAC protocol (simultaneous transmission or reception to or from multiple users using just one transceiver, fully distributed scheduling, and no limitation on maximum possible hop count).

We evaluate the performance of the proposed resource management method with two performance metrics. The first metric is the reservation success rate ( $Cr$ ) defined as the ratio of the number of CTS frames successfully received to the number of RTS frames sent including retransmissions. The second metric is the success rate of data transmission ( $Dr$ ), which is defined as the proportion of the number of successful data receptions to the number of data transmissions including retransmissions. Regarding the MAC frame format for the simulation studies, we use the frame structure presented in Figure 1. There are 16 PDU bins in the data subframe. Accordingly, the number of channels in the ACK zone is set to 16. The number of NMU channels is configured to be 6 and there are 5 pairs of the RTS/CTS channels. Assuming a unit disk model, we set the same transmission radius of 2 km for all the nodes. We uniformly deploy nodes in a 16 km  $\times$  16 km area. The performance of our resource reservation protocol is influenced not only by the amount of traffic generated per

node but also by the density of nodes. In this simulation topology, we use  $\rho$  to denote the number of nodes in a 4 km  $\times$  4 km region (the density of nodes) and we vary  $\rho$  from 2 to 12.

In a multihop network, the amount of data that a node transmits is the sum of a volume of data generated and an amount of data forwarded for its neighbors. The latter will vary according to the routing protocol used to determine the next hop of a flow, even if the operational environment of a network is the same. However, the focus of this section is to evaluate the performance of our resource reservation protocol regardless of other protocols. Thus, to exclude the influence of a routing protocol, we set up simulation scenarios in which a node transmits only the data it generated without forwarding data received from its neighbors. Data in a node is generated as follows. The data generation rate of a node follows a Poisson distribution with mean  $\lambda$  and the size of data follows an exponential distribution with mean  $\mu$ . By varying  $\lambda$  and  $\mu$ , we control the amount of data produced by a node. We further assume that one PDU bin is reserved for data transmission between a sender and a receiver. Once a reserved PDU bin is used to transmit data, it is returned, and a node should contend for a PDU bin again to transmit more data. However, if the size of data is larger than the size of a PDU bin, the data is segmented to fit into the PDU bin. Once a PDU bin is reserved for the first segment, the rest of the segments are consecutively transmitted through the PDU bin without additional resource reservation. In addition, if an error occurs while transmitting data, a node retransmits the data immediately through the same PDU bin reserved

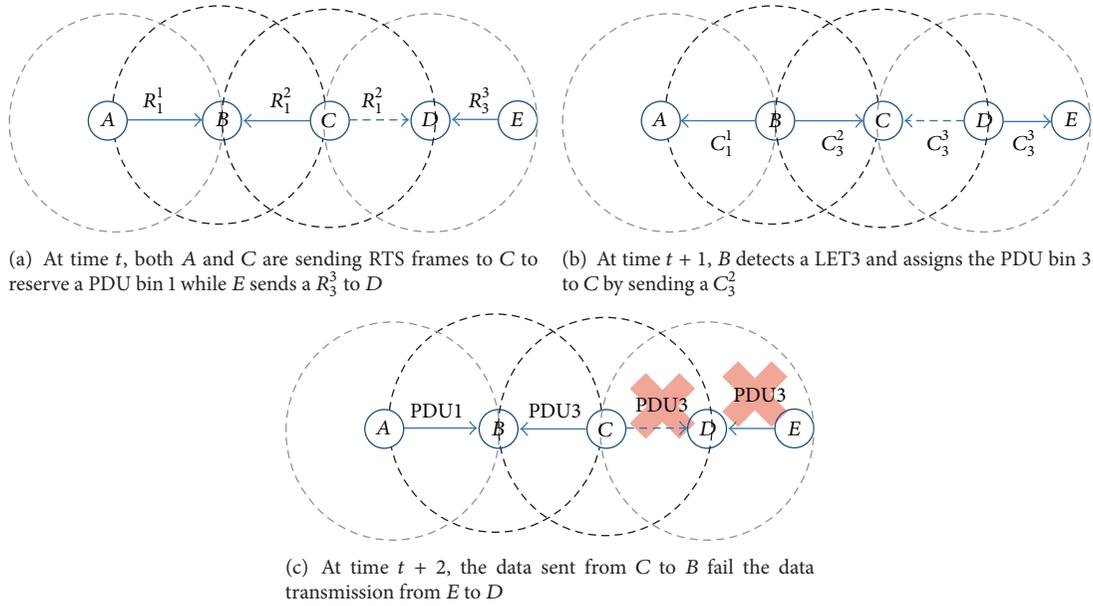


FIGURE 6: An example scenario of an LET5.

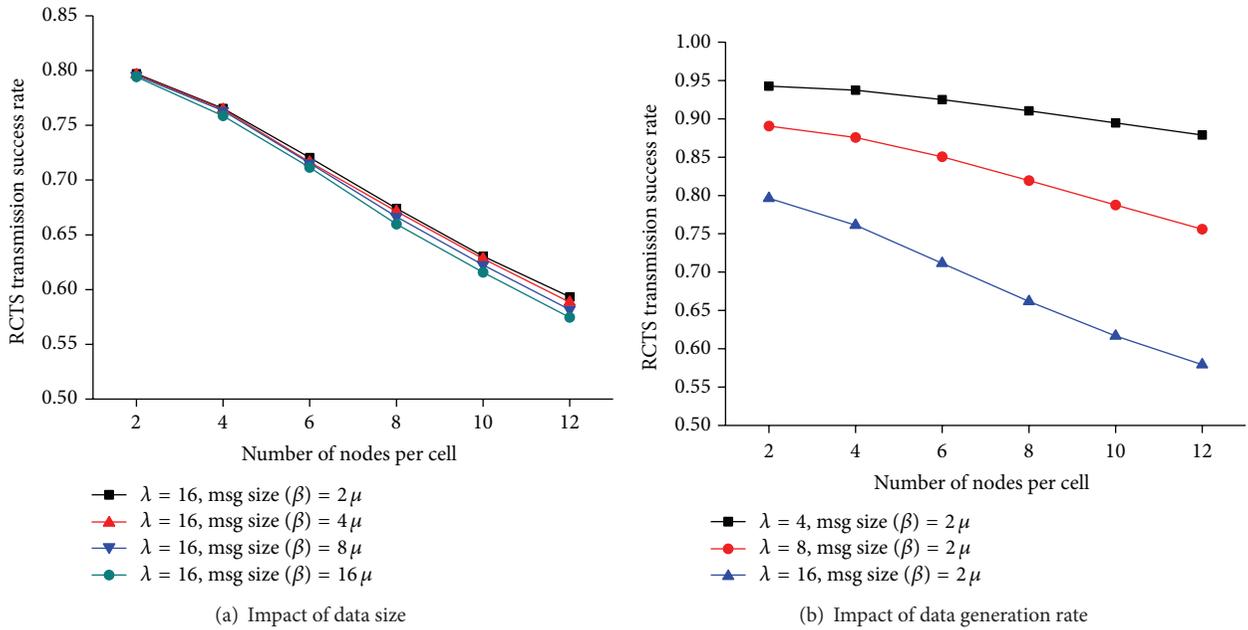


FIGURE 7: Resource reservation success rate at the MAC layer.

before without going through a resource reservation process again. We limit the number of retransmissions to 2. Thus, if data transmission fails after retransmitting twice, the data is discarded.

**4.1. MAC Level Performance.** In this section, we evaluate the performance of our distributed resource reservation protocol at the MAC layer in a static environment, where nodes do not move around and frames are not lost in a wireless link. This ideal operational environment is configured to evaluate

the pure performance of our resource reservation protocol. In this environment, we measure the  $C_r$  and the  $D_r$  after all the nodes identify their 1-hop neighbors exactly to exclude the effect of the network state estimation process. In addition, in this operational environment, data transmissions fail only when collision occurs in a resource reservation stage or a data transmission stage.

Figure 7 shows the reservation success rate for different node densities, data generation rates, and data sizes. A resource reservation process fails when a physical error takes place. A physical error occurs when an RTS frame or a CTS

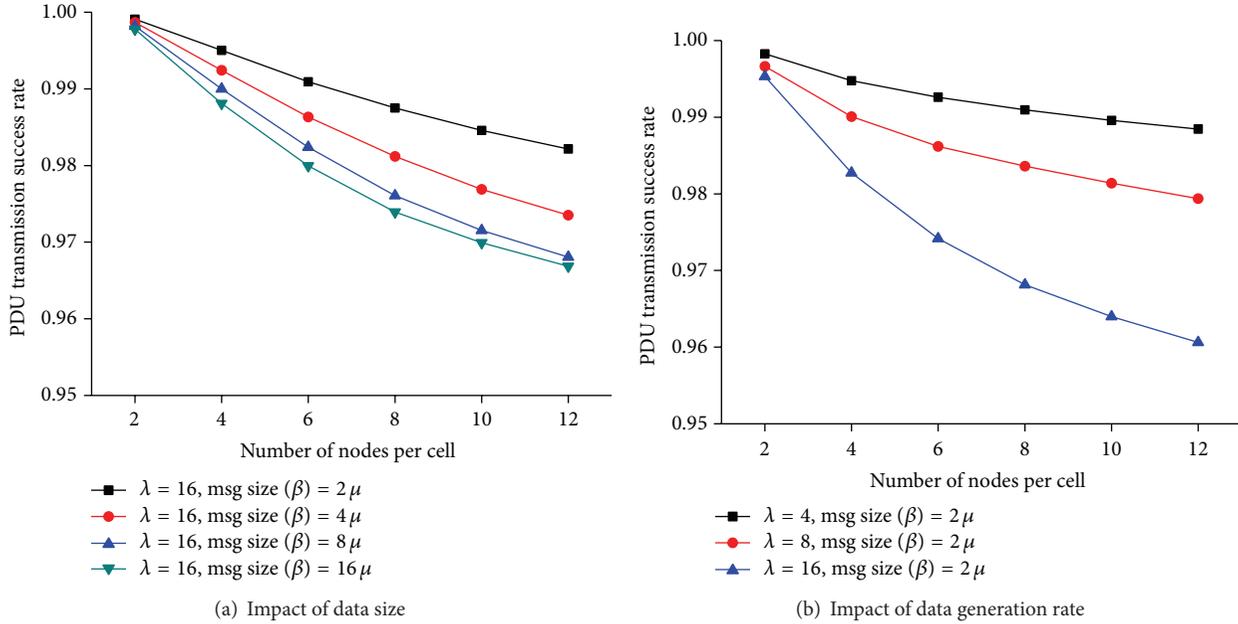


FIGURE 8: Data transmission success rate.

frame collides. A node  $Y$  cannot successfully receive an RTS frame  $R_i^j$  sent from  $X$  if more than two nodes in  $NN_Y$  send  $R_k^j$ 's while  $Y$  is receiving  $R_i^j$ . Similarly, a node  $X$  cannot decode a CTS frame  $C_i^j$  sent from  $Y$  if more than two nodes in  $NN_X$  send  $C_k^j$ 's while  $X$  is receiving  $R_i^j$ .

As the number of nodes in a network increases, it becomes more likely that more than two neighboring nodes are sending RTS frames or CTS frames using the same RCTS channel at the same time. Therefore,  $Cr$  decreases with the density of nodes. In our protocol, once a node reserves a PDU bin, the node uses the PDU bin consecutively if the size of data is larger than that of a PDU bin. Therefore, a node holds the reserved PDU bin longer as the size of data increases. Given the same operational environment, the average number of reserved PDU bins increases with the data size. If a node with data to send senses that a PDU bin is not available (i.e.,  $N = |U_X|$ ), the node defers data transmission until it becomes  $N > |U_X|$ . As the number of nodes waiting to start a reservation process becomes larger, the number of nodes beginning to send RTS frames simultaneously increases, which leads to the relative decrease in  $Cr$  with  $\mu$  (Figure 7(a)). On the other hand, the number of data a node has to transmit increases with  $\lambda$ . Since a node has to reserve a PDU bin before it transmits data, the number of nodes attempting to reserve PDU bins at the same time increases with  $\lambda$ . Consequently, the probability that more than two neighboring nodes simultaneously choose the same RCTS channel increases with  $\lambda$ , which results in a decrease in  $Cr$  with  $\lambda$  (Figure 7(b)).

Figure 8 shows the influence of  $\rho$  on the success rate of data transmission. There are logical errors (LET4 and LET5) that cannot be detected through our distributed resource reservation protocol using only the 1-hop information of a

node. When these undetectable logical errors occur, data transmission fails, even if a PDU bin is reserved successfully. Both the number of available PDU bins and the number of free RTS channels become smaller as the node density increases. Accordingly, it becomes more probable that neighboring nodes are attempting to reserve the same PDU bins using the same RTS channels at the same time. Consequently, since the LET4 and the LET5 take place more often with  $\rho$ ,  $Dr$  decreases with the density of nodes, as shown in Figure 8.

Figures 8(a) and 8(b) show the impact of  $\mu$  and  $\lambda$  on  $Dr$ . As the size of data increases, a node holds the reserved PDU bin longer. Thus, the level of contention for a PDU bin increases with the size of data. Similarly, the number of data a node needs to send per second grows as  $\lambda$  increases. This makes the contention levels for an RTS channel and a PDU bin increase. Accordingly, the probabilities that LET4 and LET5 take place become higher because it is more likely that nodes are reserving the same PDU bin at the same time. As a consequence,  $Dr$  decreases with the data size and the data generation rate. However, since  $Dr$  is more than 95% for all the simulation environments, the gain in our design choice when a node detects an LET3 outweighs the loss because of the LET5.

**4.2. System Level Performance.** In this section, we evaluate the performance of the proposed resource management method at the system level by considering not only the bit errors in a wireless link, but also the mobility of a node. In this work, we use a simplified packet error rate table shown in Table 2 to exclude the effects of physical layer issues of the resource management problem and focus on verifying the ability of our MAC protocol. When a receiver detects an error in a message, we assume that the receiver immediately discards the message without further processing.

TABLE 2: The characteristics of a wireless link in terms of delivering a message between a sender and a receiver. (Each column except the first one represents the probability that a message corresponding to its first row is received at a receiver without an error.)

Distance from a sender (km)	PDU	NMU	ACK	RTS	CTS
0.0–0.5	1	1	1	1	1
0.5–1.0	0.96	0.98	0.99	0.99	0.99
1.0–1.5	0.89	0.95	0.97	0.97	0.97
1.5–2.0	0.67	0.83	0.92	0.92	0.92
Otherwise	0	0	0	0	0

We modify the random waypoint model [21] to represent the mobility pattern of a node. At the beginning of a simulation, a node uniformly selects the speed in  $[0, 72 \text{ km/h}]$  and the direction in  $[0, 2\pi]$ . We assume that nodes do not change the initial speed and direction until the end of the simulation. If a node moves beyond the simulation topology, the node enters the network at a symmetrical point to the position where it moves out with respect to the center of the topology. We measure  $Cr$  and  $Dr$  by varying  $\lambda$ ,  $\mu$ ,  $T$ , and  $\rho$ .

The level of resource contention grows as the node density increases. Therefore, the contention success rate deteriorates with  $\rho$  (Figure 9). However, compared with that in the ideal simulation environment,  $Cr$  becomes lower in this simulation environment under the same  $\lambda$ ,  $\mu$ ,  $T$ , and  $\rho$ . This is attributed to the frame loss and the mobility of a node that prevents the successful RTS/CTS frame exchange besides the physical errors. A node cannot reserve a PDU bin if an RTS frame or a CTS frame is lost in a wireless link. Furthermore, if one of the nodes performing a resource reservation process moves out of the transmission range of the other node, it cannot complete the RTS frame and the CTS frame exchange. In the range of the parameters, the data generation rate dominates  $Cr$  because it has the greatest effect on the frequency at which a node begins a resource reservation process.

Figure 10 shows the success rate of data transmission. In addition to the loss in data transmission, the logical errors that could not be detected by local information also result in data transmission failure. Since a node manages  $U_X$  and  $A_X$  by overhearing messages not destined to it,  $U_X$  and  $A_X$  become more inaccurate as the number of lost frames increases. Therefore, the probability that a node reserves a PDU bin that is being used by its 1-hop or 2-hop neighbors grows. Consequently, the increase in the number of the LET4 and the LET5 reduces  $Dr$ . Moreover, since nodes are moving around, data transmission also fails if a node moves out of the transmission range of its corresponding node after successfully reserving a PDU bin. The larger the data size becomes, the longer the reserved PDU bin is used. If the number of available PDU bins becomes small, the LET5 is more likely to occur. Therefore,  $Dr$  deteriorates (Figure 10(a)) as the data size increases. However, since we limit the maximum number of data retransmissions to two, a PDU bin is returned if a data transmission does not succeed

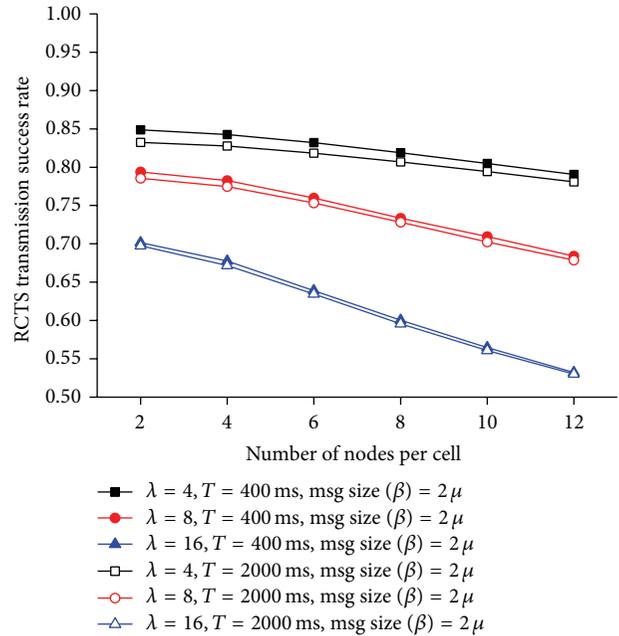
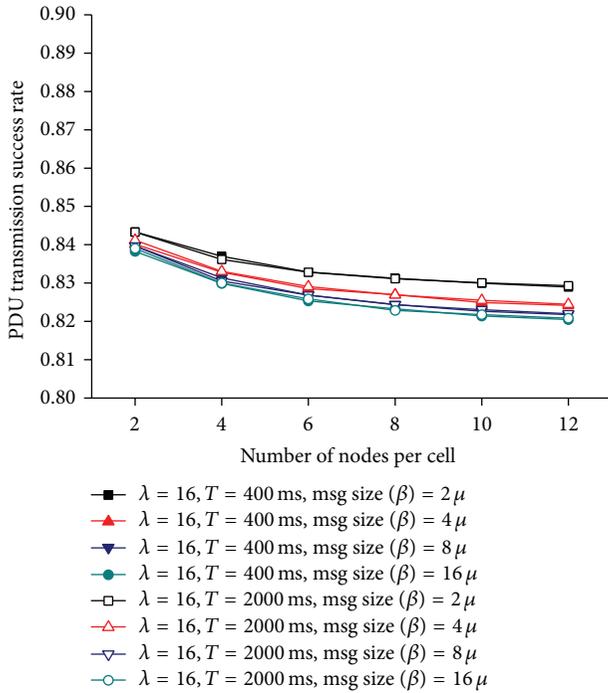


FIGURE 9: Resource reservation success rate in a mobile ad hoc environment ( $\lambda = 16$ ).

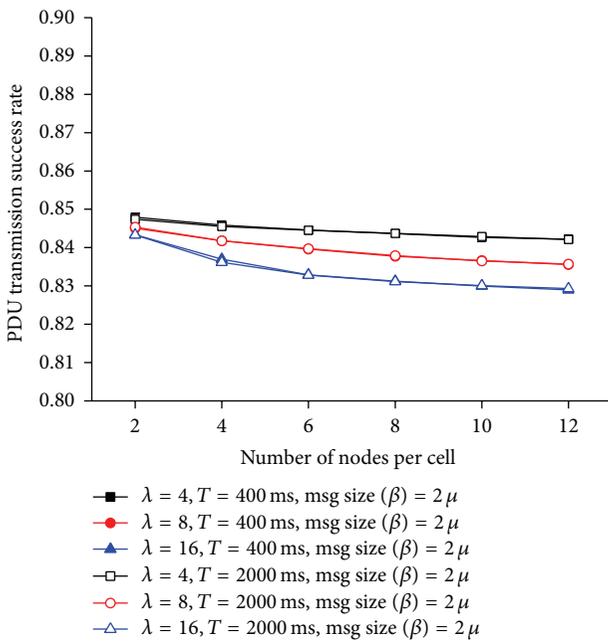
for two consecutive trials. Therefore, the decrease in  $Dr$  does not show significant difference with  $\mu$ .

The parameter  $T$  controls the frequency at which a node announces its presence to a network. Thus, the time for a node to construct an exact  $NN_X$  decreases as  $T$  becomes smaller. If  $NN_X$  is equal to  $AN_X$ , the  $NN_X$  can be considered to be exact. However, since a node maintains  $U_X$  not by  $NN_X$  but by the information in the ACK zone and uses  $NN_X$  to calculate the backoff probability when a collision occurs in a resource reservation process, the impact of  $T$  on the  $Cr$  and the  $Dr$  was marginal. On the other hand, the timeliness of an exact  $NN_X$  affects the performance of a routing protocol. Therefore, in the following section, we analyze the impact of  $T$  on the system performance in terms of the probability that a node successfully receives a NUM message from one of its neighbors. We also derive an optimal  $T$  that minimizes the average delay until a node successfully receives NUM messages from one of its neighbors.

4.3. *Performance of Network Estimation Procedure.* The simulation environment for the performance evaluation of the network estimation procedure is as follows. All the nodes are configured to have the same transmission radius of 1 km and the same broadcast period  $T$ . We uniformly deploy nodes in a  $5 \text{ km} \times 5 \text{ km}$  region. We denote the number of nodes per cell (i.e., the number of nodes in a  $1 \text{ km} \times 1 \text{ km}$  region) by  $n$ . To include the hidden nodes in the contention for the NMU channels, we select a node  $X$  randomly among the nodes located in the center cell of the topology (Figure 11). For the selected node  $X$ ,  $p_{NMU}$  is calculated as the ratio of the number of nodes that successfully received the NMU message from  $X$  to  $|AN_X|$  after  $X$  sends an NMU message. To evaluate the



(a) Impact of data size



(b) Impact of data generation rate

FIGURE 10: Data transmission success rate in a mobile ad hoc environment.

effect of  $T$  on the performance of our protocol, we exclude the dynamics of a radio propagation environment by assuming that frames are not lost in a wireless link.

As we can see in Figure 12, the  $p_{\text{NMU}}$  values obtained by the mathematical analysis are in accord with those from simulations for various node densities in a cell and  $T_s$ . Given  $T$  and  $M$ , it is likely that nodes choose the same

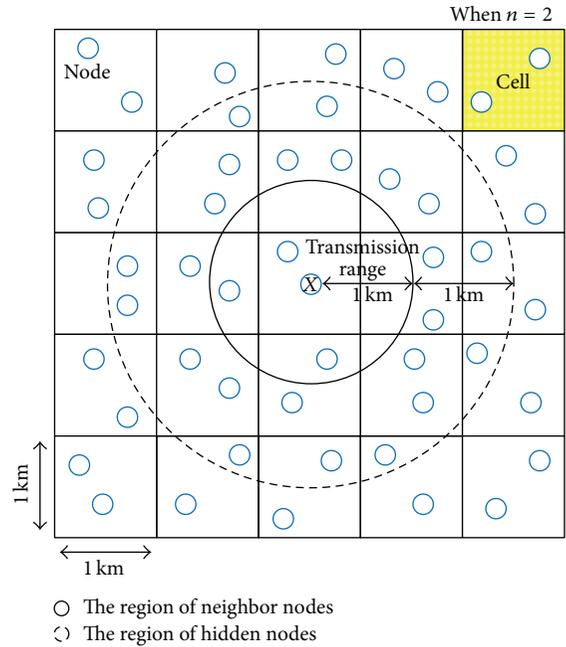


FIGURE 11: A simulation topology to evaluate the impact of  $T$  ( $n = 2$ ).

NMU channel at the same time as  $n$  increases. Therefore,  $p_{\text{NMU}}$  becomes smaller with the number of nodes per cell (Figure 12(a)). If a node density is given, the contention level for the NMU channels decreases as the number of NMU channels increases. Therefore,  $p_{\text{NMU}}$  becomes higher with  $M$ . Figure 12(b) shows the impact of  $T$  on  $p_{\text{NMU}}$  for a few  $n$  when  $M = 6$ . As  $T$  becomes longer, it is more unlikely that more than two nodes select the same frame for advertising their presence. Thus, the probability of successfully receiving an NMU message increases with  $T$ .

Figure 13(a) shows the average delay for a node to successfully receive an NMU message from one of its neighbors as  $T$  varies from 100 ms to 1000 ms when  $M = 6$ . The simulation results are in accord with the analytical results in (2). Since the contention level for an NMU channel decreases as the number of neighboring nodes becomes smaller,  $D_o$  becomes longer with  $n$ . For a given node density, we can see that there is an optimal  $T$  that minimizes  $D_o$ . As can be seen in Figure 13(b), the changing pattern of  $D_a$  is the same as that of  $D_o$ , and  $D_a$  from simulations also coincides with those from the analysis.

## 5. Conclusions and Future Works

We have proposed a distributed OFDMA-based MAC protocol for mobile ad hoc multihop networks. A MAC frame format that supports resource reservation before data transmission was presented. The proposed MAC frame format is divided into data and control subframes. The data subframe is composed of multiple PDU bins, each of which contains data to be transferred. The control subframe is composed of NMU, ACK, and RCTS zones. The roles of each zone are explained

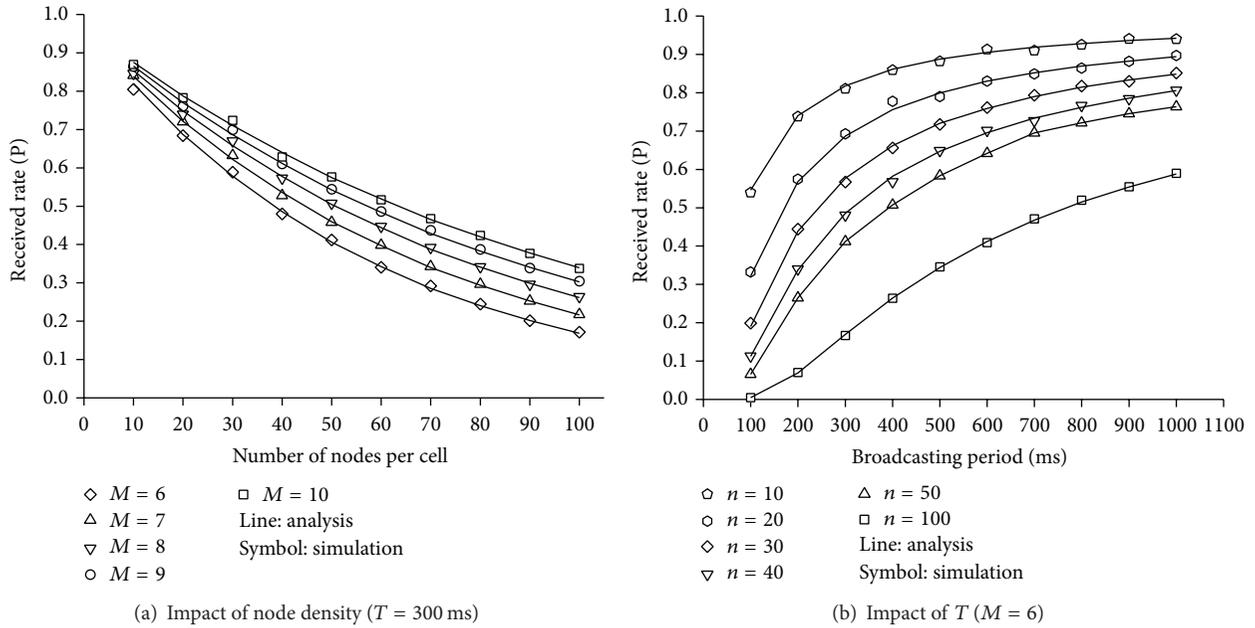


FIGURE 12: Probability of successfully receiving an NMU message.

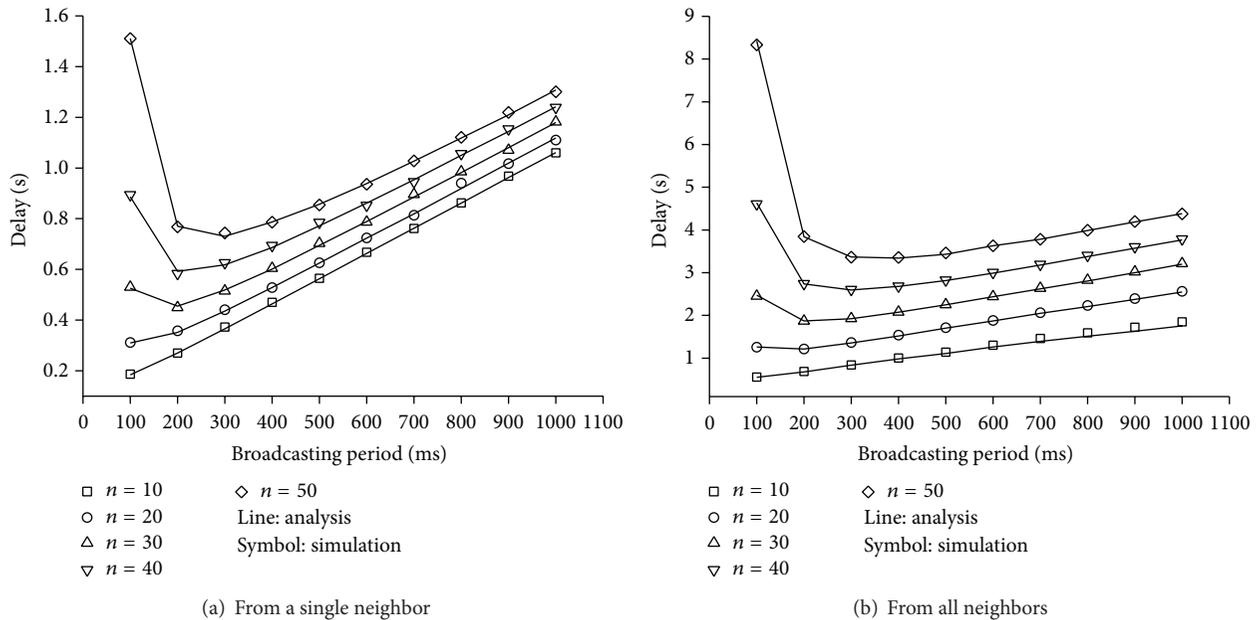


FIGURE 13: Average delay for a node to successfully receive the NMU message from its neighbors.

in detail. Also, we classify logical error scenarios into 5 categories, from LET1 to LET5, some of which are unique features of the MAC protocol using OFDMA. By extensive simulation studies and analysis, we have evaluated the performance of the proposed MAC protocol and verified the effects of the distributed resource reservation procedure and inevitable logical error on system performance. Specifically, the RCTS/PDU transmission success rates were evaluated as functions of message size and traffic generation ratio in the MAC-level simulation results. Also, we have evaluated the

system level simulation results of RCTS/PDU transmission efficiency, considering both bit errors in a wireless link and the mobility of a node. Finally, we analyzed and evaluated the effect of the NMU advertisement period on the delay before a node receives NMU messages from all of its neighbors.

In our work, a receiving node does not send ACK message when time-sensitive service is required by the upper layer. Instead, the receiving node can control the data transmission of the sending node by the feedback of data quality information such as the success rate of data transmission and data

transmission delay. This transmission control methodology should consider the quality of service to be guaranteed in upper layers and optimized usage of system resources which will be the focus of further work.

## Notations

- $NN_X$ : An estimated set of 1-hop neighbor nodes of a node  $X$
- $AN_X$ : An actual set of 1-hop neighbor nodes of a node  $X$
- $U_X$ : An estimated set of PDU bins being used by 1-hop neighbors of a node  $X$  including those that  $X$  is using
- $R_i^j$ : An RTS frame sent through an RTS channel  $j$  to request to reserve a PDU bin  $i$
- $C_i^j$ : A CTS frame sent through a CTS channel  $j$  to inform about the reservation state of a PDU bin  $i$ , which is requested by a node, starting a reservation process
- $A_X$ : A set of PDU bins that 1-hop neighbors of  $X$  attempt to reserve when  $X$  receives an RTS frame from one of its neighbor nodes  $Y$ .  $A_X$  is inferred from the RTS region and it excludes the PDU bin that  $Y$  requests
- $T_F$ : A frame length.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research was supported by the Chung-Ang University Research Scholarship Grants in 2012. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A2A2A01014170).

## References

- [1] A. Maeder and N. Zein, "OFDMA in the field: current and future challenges," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 5, pp. 71–76, 2010.
- [2] M. Sternad, T. Svensson, T. Ottosson, A. Ahlen, A. Svensson, and A. Brunstrom, "Towards systems beyond 3G based on adaptive OFDMA transmission," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2432–2455, 2007.
- [3] IEEE Std 802.16e-2005, IEEE Standard for Local and Metropolitan Area Network, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Band and Corrigendum 1, 2006.
- [4] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-advanced: next-generation wireless broadband technology," *IEEE Wireless Communications*, vol. 17, no. 3, pp. 10–22, 2010.
- [5] T. Braun, A. Kassler, M. Kihl, V. Rakocevic, V. Siris, and G. Heijenk, "Multihop wireless networks," *Lecture Notes in Electrical Engineering*, vol. 31, pp. 201–265, 2009.
- [6] M. Conti and S. Giordano, "Multihop ad hoc networking: the theory," *IEEE Communications Magazine*, vol. 45, no. 4, pp. 78–86, 2007.
- [7] 3GPP TR 36.814 V1.2.1, Further Advancements for EUTRA: Physical Layer Aspects, Technical Specification Group Radio Access Network Rel. 9, June 2009.
- [8] IEEE P802.16j/D9, Draft Amendment to IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Multihop Relay Specification, 2009.
- [9] M. Salem, A. Adinoyi, M. Rahman et al., "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 3, pp. 422–438, 2010.
- [10] Y. Yang, H. Hu, J. Xu, and G. Mao, "Relay technologies for WiMAX and LTE-advanced mobile systems," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 100–105, 2009.
- [11] G. Kulkarni and M. Srivastava, "Subcarrier and bit allocation strategies for OFDMA based wireless ad hoc networks," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '02)*, pp. 92–96, November 2002.
- [12] H. Xiong and E. Bodanese, "A signal strength based medium access control for OFDMA based wireless ad hoc networks," in *Proceedings of the 18th International Conference on Telecommunications (ICT '11)*, pp. 439–443, May 2011.
- [13] S. Pomportes, A. Busson, J. Tomasik, and V. Vèque, "Resource allocation in ad hoc networks with two-hop interference resolution," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference (GLOBECOM '11)*, December 2011.
- [14] K. N. Quang, V. D. Nguyen, T. D. Nguyen, T. H. Nguyen, and G. Gelle, "MAC and routing integration performance improvements in OFDMA-based multi-hop and Ad-hoc Networks," in *Proceedings of the International Conference on Computing, Management and Telecommunications (ComManTel '13)*, pp. 5–10, January 2013.
- [15] J. So and N. Vaidya, "Multi-channel mac for ad hoc networks: handling multi-channel hidden terminals using a single transceiver," in *Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'04)*, pp. 222–233, 2004.
- [16] K. Karakayali, J. H. Kang, M. Kodialam, and K. Balachandran, "Cross-layer optimization for OFDMA-based wireless mesh backhaul networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'07)*, pp. 276–281, Kowloon, China, March 2007.
- [17] R. Rashtchi, R. H. Gohary, and H. Yanikomeroglu, "Joint routing, scheduling and power allocation in OFDMA wireless ad hoc networks," in *IEEE International Conference on Communications (ICC '12)*, pp. 5483–5487, June 2012.
- [18] X. Yu, D. B. Hoang, and D. Feng, "A novel QoS feedback control for supporting compressed video," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '01)*, vol. 4, pp. 2484–2488, San Antonio, Tex, USA, November 2001.
- [19] U. Tos and T. Ayav, "Adaptive RTP rate control method," in *Proceedings of the 35th Annual IEEE International Computer Software and Applications Conference Workshops (COMPSACW '11)*, pp. 7–12, July 2011.

- [20] H. Gharavi, "Control based mobile ad-hoc networks for video communications," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 383–391, 2006.
- [21] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 2, no. 3, pp. 257–269, 2003.

## Research Article

# A New Graph Drawing Scheme for Social Network

**Eric Ke Wang<sup>1</sup> and Futai Zou<sup>2</sup>**

<sup>1</sup> Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China

<sup>2</sup> School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 518055, China

Correspondence should be addressed to Eric Ke Wang; wk\_hit@hitsz.edu.cn

Received 11 April 2014; Accepted 21 June 2014; Published 16 July 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 E. K. Wang and F. Zou. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of social networks, people have started to use social network tools to record their life and work more and more frequently. How to analyze social networks to explore potential characteristics and trend of social events has been a hot research topic. In order to analyze it effectively, a kind of techniques called information visualization is employed to extract the potential information from the large scale of social network data and present the information briefly as visualized graphs. In the process of information visualization, graph drawing is a crucial part. In this paper, we study the graph layout algorithms and propose a new graph drawing scheme combining multilevel and single-level drawing approaches, including the graph division method based on communities and refining approach based on partitioning strategy. Besides, we compare the effectiveness of our scheme and FM<sup>3</sup> in experiments. The experiment results show that our scheme can achieve a clearer diagram and effectively extract the community structure of the social network to be applied to drawing schemes.

## 1. Introduction

Graph drawing is a combination technique of information science and mathematics, which is employed in multiple research areas such as social network analysis. Since social networks are commonly very complex in large amount of data about features and relationships, it is difficult for people to understand the huge data. Fortunately, graphs help analytics in visualization and rationalization. Graph drawing is, given a set of nodes and sets (edge sets) of their relationships, to calculate the position of each node and plot the edges as curves. In other words, it is a transforming way from abstract data such as text and digits to static or dynamic visualized results in order to let people easily understand the principle and inner meaning of huge amount of complex data. It helps people make judgmental and analytic decision from the macro view. But, although graph drawing for social networks has been studied for several years, there are still many problems to be solved.

Currently, in most schemes of graph drawing, one social network is regarded as one kind of community structure

to draw the graph instead of multiple communities in one social network. However, a social network commonly has possible features of various communities. Thus, it leads to an appearance that many graph drawing algorithms can perform well in some data sets with certain features but perform badly in more complex data. Therefore, how to detect various community structures in social networks and adapt drawing to current structure are important research problems to be solved.

Besides, in many data sets of social networks, there are various semantic information fusions and exchanges among members; however, in current drawing approaches, the impact of visualization for the semantic information is not considered; only topology model or structure features are employed. Thus, it may result in many graphs being unreadable and readers hardly fully extract the information of members or communities they care about from the drawn graphs. Therefore, we need a new drawing approach which combines topology and semantic information to make the drawn graphs readable and reasonable.

## 2. Related Works

Currently, there are mainly several categories of graph drawing approaches such as node-link [1–4], space filling [5, 6], matrix [7–9], and mix [10, 11]. Node-link is relatively simple, considering nodes as vertices, only calculating their positions and representing edge as curve or fold line; space filling is a reduction of multidimensional problems, for example, reducing 3-dimensional problems into 2-dimensional problems. A nested curve such as Hilbert m-Peano curve is recursively refined to represent the data. Matrix approach represents a diagram as a connected matrix,  $(i, j)$  is represented as the edge from node  $i$  to node  $j$ , and the attributes of the edge are encoded in visual features such as color, form, or size.

For node-link approach, there are two main drawing algorithms (single-level drawing algorithm and multilevel drawing algorithm).

In single-level drawing approaches, there are several typical types such as tree based [12], radical [13], and force directed [6]. Among them, force directed is widely used for drawing. The idea of force-directed way is proposed by Eades [14]. It is that, mapping the relationships into physics mechanics models, nodes are replaced by small solid balls with some certain radius; the edges are replaced by springs. In initialization, the coordinates of each small ball are generated randomly. And then, to use the elastic force of springs on the balls to move the positions of the balls until the energy of whole system is minimal at last is what we call optimal state. Lately, the spring algorithm is updated in several schemes [2] and [15–17]; the major difference among them is the way to compute elastic force.

Multilevel scheme is mainly used to improve the effectiveness of layout and shorten drawing time. The main idea of multilevel scheme is to recursively apply the coarsening of diagrams. The coarsening of diagrams is the abstraction representation of fine-grained diagrams in multilevel and it can be drawn much faster. In other words, it can be applied for larger data sets to enhance the visualization effectiveness and reduce the running cost at the same time. FM<sup>3</sup> (fast multipole multilevel method) [15] is a classical multilevel algorithm applying for most of the graphs. In FM<sup>3</sup>, the diagram is segmented to several child diagrams called “solar systems”; each “solar system” is compressed to be a node and repeats the process until forming a hierarchical diagram. The method had a better effectiveness than former approaches [18]. Walshaw [19] proposed a kind of evaluation method to coarsen by maximum matching. Maximum matching is a greedy algorithm to contain the largest possible number of edges. ACE algorithm [20] decides the number of partitions by solving Laplacian matrix. The feature vectors are computed by constructing a hierarchical coarsening matrix and recursively evaluating the feature vector of each level to achieve the vector of the original diagram. Archambault et al. [21] proposed a multilevel approach based on topological feature. In the approach, the interested topological feature is firstly detected and the child diagram with the topological feature is replaced with a node in the coarsening level. And then recursively execute detection for the features and compression process. In

the process, for the topological feature of each child diagram, a proper drawing approach is selected.

In this paper, we propose a new drawing scheme combining multilevel drawing and single-level drawing approaches, including the graph partition method based on communities and layout refining approach based on partitioning strategy. Graph partition based on communities is employed in the stage of graph division of multilevel drawing. Single force-directed algorithm is used for the setting of initial coordinates of layout refining process; the layout refining process based on partitions is used for the iteration of initial coordinates and optimizing process to achieve the best layout effectiveness. Besides, we compare the effectiveness of our scheme with FM<sup>3</sup> in experiments. The experiment results show that our scheme can achieve a clearer diagram and effectively extract the community structure of the social network to be applied to drawing algorithm.

The objective of our graph drawing scheme is to quickly present readable graphs to the users which can also precisely reflect the data principle. It mainly includes three goals:

- (1) recognizing the communities accurately,
- (2) adaptive layout,
- (3) reasonable use the layout space to reflect the strong and weak relations among vertices.

We propose a new adaptive scheme to achieve the above goals.

## 3. Assumption and Notations

In this paper, we mainly research on undirected graphs. Undirected graph can be represented by  $G = \{V, E\}$ , where  $V$  represents the set of vertices and  $E$  represents the set of edges. In our work, we target on connected undirected graph. The definitions of notations are as follows:

$|V|$  is number of vertices;

$|E|$  is number of edges;

$r(v)$  is neighbor set of node  $v$ ;

$e(i, j)$  is the edge between nodes  $i$  and  $j$ ;

$\text{Pos}(x_i)$  are the coordinators for distributing nodes;

$\|x_i - x_j\|$  is the distance between two nodes in the graph;

$d(i, j)$  is the distance between any two nodes.

In our scheme, we adopt small-world network theory. As we all know, researchers have studied small-world network theory for a long time, but most of the researches focus on exploring the principle and topology of small-world networks. For example, a common job of social networks analysis is recognition of the modes and relations among the connected nodes which represent some social implications such as social status. Actually, many networks to be visualized have some features such as community characteristics. Those features can be easily recognized by people straightforwardly if they are shown in a graph. However, most of researchers

focus on data and topological analysis of social networks, while, in the area of information visualization, small-world network theory is not fully employed. Therefore, we propose a graph drawing scheme based on small-world network theory. We separate a network to some small hierarchical communities which are highly connected with each other inside each community. And it is much more convenient for users to observe the relations and groups among members and understand the relations structures in the graph.

### 4. The Graph Drawing Scheme

It mainly involves two steps: (1) communities partition and (2) adaptive refinement.

4.1. *Communities Partition.* In community partition, we adopt a filtering approach to separate a graph into a hierarchy of subnetworks by finding out the weakest edges as the separation starting edges. The procedure is as shown in Figure 1.

The process is to calculate the edge strength to find out the weakest edges in the network and then delete the weakest edges so that it can be separated into subnetworks with stronger connections inside each subnetwork.

The procedure of community partition can be divided into 3 steps.

(1) *Filter Out Weak Edges.* The edge strength represents its contribution to the clustering coefficient. If an edge connects two uninteracted groups of neighbors, then strength of the edge is considered zero. Thus, the edge is weak and filtered out.

We can set up a threshold value  $\tau$ , and once strengths of the edges are lower than  $\tau$ , they would be filtered out. Thus, the original graph can be divided into some subnetworks. Based on our observation, we found that the threshold value  $\tau$  is related to the maximum of edges strengths instead of empirical value. Then we propose an approach to identify the threshold value. Find out the maximum and minimum of the strengths of all edges, and then  $\tau = \min_{ES} + (\max_{ES} - \min_{ES}) \times \text{ratio}$ . When ratio is close to the biggest strength such as 0.95, it can guarantee the accuracy.

Given an edge  $e(u, v)$ , the edge strength can be calculated as follows (as shown in Figure 2).

- (1) Separate the neighbors of  $u$  and  $v$  into three subsets which have no interaction with each other.
- (2)  $M(u)$  represents the set of all  $u$ 's neighbors which are not adjacent to  $v$ .
- (3) Similarly,  $M(v)$  represents the set of all  $v$ 's neighbors which are not adjacent to  $u$ .
- (4)  $W(u, v)$  represents the set of common neighbors of  $u$  and  $v$ .  $r(A, B)$  represent the number of edges between set  $A$  and set  $B$ .
- (5)  $S(A, B) = r(A, B) / (|A| * |B|)$  represents the ratio of the real exist edges and all possible edges between  $A$  and  $B$ .

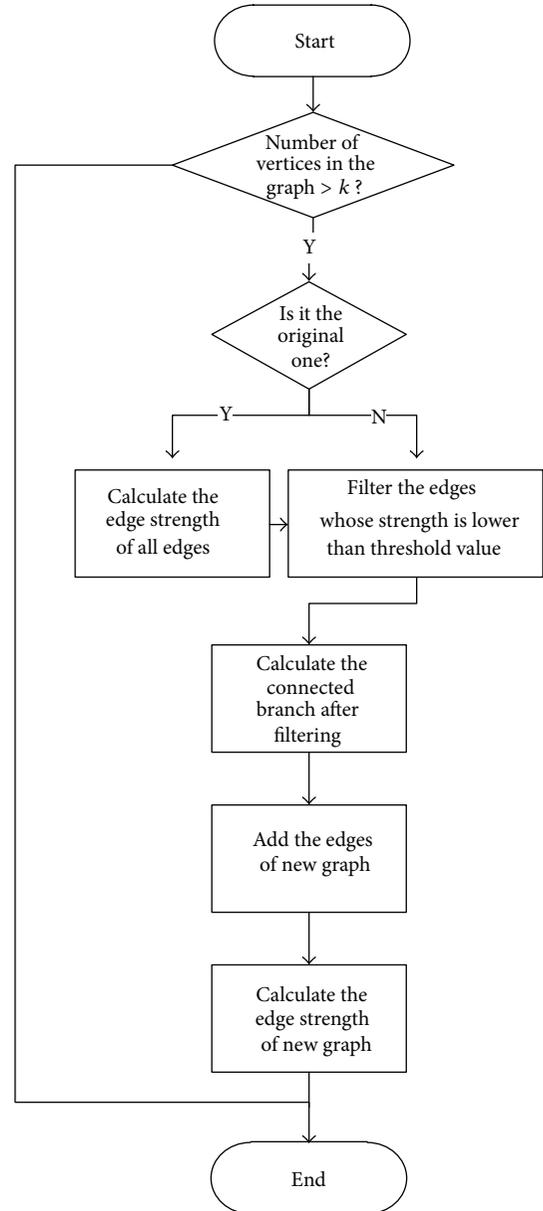


FIGURE 1: Flow of layout compression.

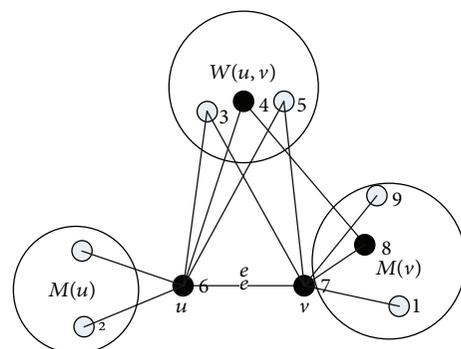


FIGURE 2:  $e(u, v)$  and neighbors.

- (6) Any edge between  $M(u)$  and  $M(v)$  or  $W(u, v)$  is a part of a 4-edge closed loop which must have an edge of  $e(u, v)$ .
- (7) We define that  $|W(u, v)|/(|M(u)| + |W(u, v)| + |M(v)|)$  is the ratios of 3-edge closed loop including  $e(u, v)$ . Then the strength of  $e(u, v)$  can be calculated by the following equations:

$$ES(u, v) = s(M(u), W(u, v)) + s(W(u, v), M(v)) + s(W(u, v), M(u)) + s(M(u), M(v)) + \frac{|W(u, v)|}{|M(u)| + |W(u, v)| + |M(v)|} \tag{1}$$

(2) *Hierarchical Decomposition.* Recursively apply filtering out weak edges until the final graph has enough small size to maintain the communities characteristics.

(3) *Selection of Threshold Value  $\tau$  Decides the Form Way of Clusters.* The technique of graph drawing provides a good view way for the data sets. When graph becomes very big, subgraphs would be presented by dense areas.

However, this kind of hierarchical clustering way has a disadvantage which is that deciding the edges strength of higher level graph is difficult after each time of hierarchical decomposition. For example, there are three subgraphs  $H_i, H_j,$  and  $H_k$  which are the same size. In the original graph, there were 10 edges between  $H_i$  and  $H_j,$  while there are 2 edges between  $H_j$  and  $H_k.$  Apparently, the relation between  $H_i$  and  $H_j$  is much closer than that between  $H_j$  and  $H_k.$  But, after the hierarchical decomposition, in the high level graph, the number of edges between  $H_i$  and  $H_j$  becomes 1, so does the number of edges between  $H_j$  and  $H_k.$  It completely loses the weight of relations. It is not reasonable. Therefore, we propose an approach to solve the problem.

Suppose the original graph is  $G_0,$  after a time of hierarchical clustering, a higher graph  $G_1$  is generated, and then  $G_2$  and  $G_3$  are generated sequentially. Suppose  $e(u, v)$  is one edge of  $G_0$  and its strength is smaller than  $\tau;$  that means that in the clustering process it would be deleted. After it is deleted,  $u$  and  $v$  are located in two different clusters  $C_u$  and  $C_v;$  we use  $H_u$  and  $H_v$  to represent two clusters. Then, in the higher level graph, there would be an edge  $e(H_u, H_v)$  which connects  $H_u$  with  $H_v.$  Then, in  $G_1,$  the new edge strength is calculated as the following formula:

$$ES(H_u, H_v) = \frac{1}{n_{C_u, C_v}} \sum_{e(u', v')} ES(u', v') \quad u' \in C_u, v' \in C_v. \tag{2}$$

At the same time,  $n_{C_u, C_v}$  represents the number of edges between  $C_u$  and  $C_v$  in  $G_0.$  That means the edges strength in  $G_1$  is decided by  $G_0.$

4.2. *Adaptive Refinement.* In graph drawing of social networks, analyzers are often interested in the topology of

relationships and consider drawing strategy according to physical model, instead of the semantic information. Thus, it leads to much important relationship information which would be covered or lost. Therefore, we propose a new refinement approach to fully use layout space to make the visualization result intuitionistic based on topology and semantic characteristics. It mainly achieves three objectives: (1) make the communities partition as clear as possible by regionalization that is making drawing of the different communities in different regions, and the size of regions is expected to reflect the size of communities; (2) make full use of layout region which can minimize the intersection part of each of subregion; (3) decide the distance between two vertices which is expected to reflect their relation strength, and the distance between two communities is also expected to reflect two communities relation strength. To achieve the three goals, the approaches are as follows.

Suppose there are constructed hierarchical graphs list  $List G = \{G_0, G_1, \dots, G_i\}, i = 0, 1, 2, \dots, k,$  for graphs  $G_i$  and  $G_{i+1},$  where  $G_{i+1}$  is compressed result from  $G_i.$  Suppose  $G_i$  has  $N_{G_i}$  nodes,  $N_{C_u}$  is the size of the community  $C_u$  nodes set,  $u$  is one vertex of  $G_{i+1}$  which is compressed result from  $C_u$  of  $G_i,$   $Area_u$  is the distributed area of  $u$  in the process of layout for  $G_{i+1},$  and  $Rect(u)$  is the rectangle area of  $u$  in the limited area of  $G_{i+1}.$  The area of each vertex of the layout for  $G_{i+1}$  is calculated as follows.

- (1) Calculate the area of  $u:$

$$Area(u) = \frac{N_{C_u}}{N_{G_i}} \times Area_{G_{i+1}}, \tag{3}$$

- (2) Calculate the rectangle area of  $u$  partitioned in the layout of  $G_i,$  and the width and height of  $Rect(u)$  should satisfy the equations:

$$Rect(u) = Width(u) \times Height(u) = Area(u),$$

$$\frac{Width(u)}{Height(u)} = \frac{Width(G_0)}{Height(G_0)}. \tag{4}$$

Since there maybe overlaps between rectangles, we need to reduce the overlaps.

- (3) Locate the subrectangle and minimize the overlap of rectangles. In other words, reducing the overlap area represents more usage of space. After the step (2), get the rectangle areas of all the vertices of  $G_{i+1};$  we need to distribute them to appropriate positions to make the overlap minimum. For human view, each subrectangle can be laid flat on the original rectangle. The overlap areas of subrectangles are calculated as follows (shown in Figure 3).

Suppose  $u$  and  $v$  of  $G_{i+1}$  represent two subrectangles, their areas are marked as  $Rect(u)$  and  $Rect(v),$  their width and height are  $width_u$  and  $width_v$  and  $height_u$  and  $height_v,$  and the coordinates of their center points are  $P(u) = (x_u, y_u)$  and  $P(v) = (x_v, y_v).$  Consider  $P(u)_x = x_u$  and  $P(v)_x = x_v.$  To identify whether two rectangles have intersection area, we

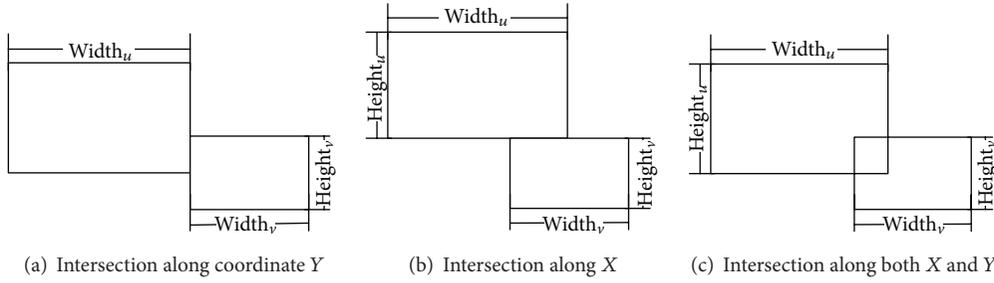


FIGURE 3: Status of intersection of rectangles.

need to check whether the distance between two center points is bigger than half of the sum of their width or height along the coordinates  $X$  or  $Y$ . The following equations should be satisfied:

$$|P(u)_x - P(v)_x| \leq \frac{1}{2} (\text{width}_u + \text{width}_v), \quad (5)$$

$$|P(u)_y - P(v)_y| \leq \frac{1}{2} (\text{height}_u + \text{height}_v).$$

If the above two equations are satisfied, two rectangles would have intersection part.

After identifying that they have intersection part, we can calculate the overlap area as the following formula:

$$\begin{aligned} \text{Area}(\text{Rect}(u), \text{Rect}(v)) \\ = |P(u)_x - P(v)_x| \times |P(u)_y - P(v)_y|. \end{aligned} \quad (6)$$

After computing the overlap areas of all subrectangles, we need to maintain a matrix 0-1 to record whether they are intersected with each other. Then we need to check them one by one; if there are  $N_{G_{i+1}}$  vertices in the graph  $G_{i+1}$ , we need to maintain a 0-1 matrix whose size is  $N_{G_{i+1}} \times N_{G_{i+1}}$ , marked as  $\text{ShadowM}$ . Then all overlap areas can be calculated as the following formula:

$$\sum_{i=1}^{N_{G_{i+1}}} \sum_{j=1}^{N_{G_{i+1}}} \text{ShadowM}(i, j) \times \text{Area}(\text{Rect}(u), \text{Rect}(v)). \quad (7)$$

Then, according to the objectives, we need to make the overlap areas as small as possible.

In order to achieve the third goal, we define the standard distance between two entities: suppose the edge strength of two entities is  $\text{ES}(e)$ , and then their standard distance is calculated by the following formula:

$$d_{u,v} = \frac{1}{\text{ES}(e_{uv})}. \quad (8)$$

Suppose  $P(u) = (x_u, y_u)$  and  $P(v) = (x_v, y_v)$  are the actual coordinates of vertices  $u$  and  $v$ , and then the real distance  $d'_{u,v} = 1/\sqrt{|x_u - x_v|^2 + |y_u - y_v|^2}$ ; we use  $\Delta d_{u,v} = d_{u,v} - d'_{u,v}$  to represent the deviation of real distance and the edge strength. So  $\Delta d_{u,v}$  is expected to be as small as possible to make sure the real distance can reflect the edge strength.

Considering the three above goals, the objective function is as follows:

$$\begin{aligned} F = \sum_{i=1}^{N_{G_{i+1}}} \sum_{j=1}^{N_{G_{i+1}}} (\text{ShadowM}(i, j) \times \text{Area}(\text{Rect}(u), \text{Rect}(v)) \\ + \Delta d_{u,v}). \end{aligned} \quad (9)$$

## 5. Main Algorithms

In our scheme, we propose three main algorithms:

- (1) community partition algorithm,
- (2) hierarchical compression algorithm,
- (3) optimization algorithm based on blocks.

The first one is used to compress the first layer; the second algorithm is for generating hierarchical compression map which is refined in the third algorithm.

**5.1. Community Partition Algorithm.** Community partition algorithm is the first step of our scheme. The procedure is as follows.

- (1) Calculate the set of neighbors of all edges, including the single neighbor-sets of the two nodes of one edge which have no intersection part, 3-edge circle neighbor-sets which can form a circle including 3 edges between the neighbor and the two nodes, and 4-edge circle neighbor-sets which can form a circle including 4 edges between the neighbor and the two nodes. These neighbor-sets are the base of calculating the strength of edges in the next step.
- (2) Calculate the strengths of the above edges.
- (3) Compare and find out the maximum and minimum of edge and calculate the filter threshold value.
- (4) Delete all the edges whose strengths are lower than threshold value and update the diagram.
- (5) Recalculate the connected components of updated diagram and return the value.
- (6) Compress each connected component into a packed node as a new vertex of the updated graph.

```

Input: Graph  $g$ 
Output: New graph after Community partition

compressGraph(Graph  $g$ )
for each  $e$  in  $g$ , do
   $u\_e.node1$ ;  $v\_e.node2$ ;
  // Compute the single neighbours, 3-circles common neighbours and 4-circles common neighbours of  $u$  and  $v$ .
  computesES( $e$ );
   $maxESmax(Set(ES))$ ;  $minESmin(Set(ES))$ ;
  thresholdcomputeThresh( $maxES$ ,  $MinES$ );
  filterLowEdges( $g$ , threshold);
  computeComponent( $g$ );
  for each component in  $g.components$ 
    newNodecompSubG( $c$ ); newG.add(newNode);
    newG.addEdges();

```

ALGORITHM 1: Community partition algorithm.

```

Input: Graph  $oriG$ ;
Output: Hierarchical graphs after compression

hierGraphs.push( $oriG$ );
while( $hierGraphs.topG.nodes.size > tN$ )
  newGcompressGraph( $hierGraphs.topG$ );
  hierGraphs.push(newG);
returnhierGraphs;

```

ALGORITHM 2: Hierarchical compression algorithm.

```

Input: hierarchical compressed graphs-hierGraphs
Output: the layout result

while( $hierGraphs$  not null)
  springlayout( $hierGraphs.topG$ );
  layoutdistrArea( $hierGraph.topG.nodes$ );
  optimize(layout);
  hierGraphs.pop();

```

ALGORITHM 3: Optimization algorithm based on blocks.

And we add edges for new vertices if they have relationship in their original diagrams.

The pseudocode is shown in Algorithm 1.

5.2. *Hierarchical Compressions.* Hierarchical compression is the second step; its pseudocode is shown in Algorithm 2.

The algorithm is built on algorithm 1. It firstly compresses the original diagrams and puts them into a stack, and then judge whether the top element of queue satisfies the compression condition; if yes, take it out, compress the front element, and push it into the queue again.

5.3. *Optimization Algorithm Based on Blocks.* The main procedure of single-level partition algorithm includes three steps:

- (1) adopt spring algorithm to locate the coordinates of single-level diagram,
- (2) set partitions and put each vertex into the partitions,
- (3) execute iteratively gradient descent to achieve the minimum of intersection of the partitions.

The algorithm is a loop procedure which processes the diagram of each layer. In the first step, it adopts spring to locate the initial positions of single-level diagram; secondly, it traverses all the vertices; because each vertex corresponds to the new child diagram connected component from the top

of the stack, the size of each component is different from the others. Then, according to the size of component, it allocates subareas which are in proportion to the width and height of original area, and set the center coordinate as the initial coordinate of the vertex from the first algorithm. In the third step, based on iteratively gradient descent approach for the intersected parts of subareas, the minimum value can be achieved and the final coordinates of all areas are created. Its pseudocode is shown in Algorithm 3.

## 6. Evaluation

In community partition, we compare our scheme with a popular partition approach [3] which is based on empirical value. We adopt 10 data sets of social networks with communities' structure. The configuration of experiment is operating system, WIN7, CPU, Intel(R) Core(TM)4 Quad CPU2.33 GHZ; Memory, 4 G, and area of layout is 1500 \* 850.

In multilevel drawing stage, we compare our scheme with fast multilevel algorithm FM<sup>3</sup>. The data sets are from Newman classical data sets which include two groups of artificial social networks graphs with communities, three real data sets "subScience," "football," and "polbooks." The artificial graphs include a social network with 128 nodes and 1009 edges and a scientists working network with 379 nodes and 914 edges. Football graph represents the networks

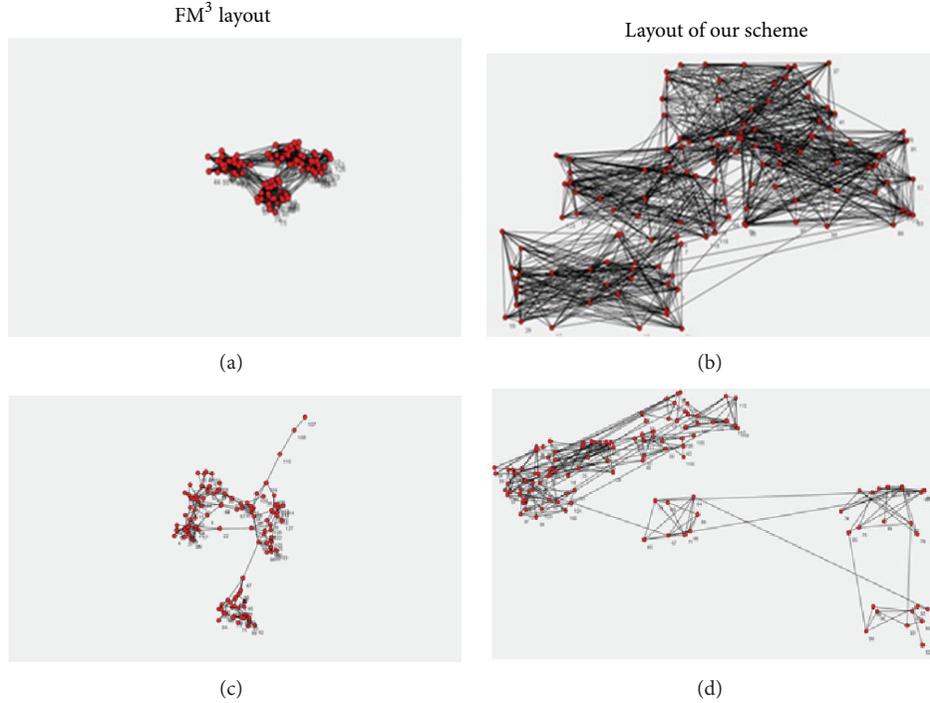


FIGURE 4: Comparison of community discovery for our scheme and FM<sup>3</sup>.

of competitions which involve the football teams in a competition season, which include 114 nodes (teams) and 615 edges (competitions); polbooks is the data set of the sales of American political books in <http://www.amazon.com/> where edge represents that two books are bought together in one order.

**6.1. Evaluation Metrics.** We adopt cluster quality value MQ to evaluate the accuracy of community partition. MQ is the average value of the density of edges inside a community. After partition for a graph, if MQ is bigger, it represents that the partition result is closer to real community result. MQ is calculated as follows:

$$MQ(C, G) = \frac{1}{p} \sum_{i=1}^n S(C_i, C_i) - \frac{1}{p(p-1)/2} \sum_{i < j} S(C_i, C_j). \tag{10}$$

In several common algorithms, the selection of MQ threshold value is decided by empirical value based on statistics approach. Given a value of MQ as “c,” we can identify the probability of partition effect higher than c.

In our scheme, the threshold value is computed by  $\tau = \min_{ES} + (\max_{ES} - \min_{ES}) \times \text{ratio}$ , which is described in Section 4.1.

**6.2. Community Partition Comparison.** According to 10 social networks data sets, we compare our scheme with the threshold approach in cluster quality value; the result is as shown in Table 1.

TABLE 1: Threshold value and MQ on several data sets.

Graph	Threshold value of our scheme	MQ	Empirical value	MQ based on empirical value
128 my	2.34	-2042	1.75	-52
128	2.82	-536	1.95	-206
Football	2.5	-1261	1.84	-241
Polbooks	3.045	-894	2.27	-734
SubScience	3.03	-1882	2.22	-867

From Table 1, the MQ value based on empirical value is less than MQ of our scheme. We can find that the average value of MQ based on empirical value is smaller than that of MQ based on threshold value selection approach of our scheme.

**6.3. Comparison of Layout Effects.** Figure 4 shows the drawing effectiveness comparison of FM<sup>3</sup> algorithm and our algorithm on the artificial social networks data sets. The first data set has 128 nodes and 1009 edges and four obvious communities; the second data set has 5 obvious communities.

From Figure 4, our scheme and FM<sup>3</sup> can both recognize four communities in the graph; the difference is that FM<sup>3</sup> often presents the whole layout in a very small region; the communities partition is clear but the degree of overlapping inside the communities is high; the main reason is that the configuration of parameters needs a lot of experiments and validations to adjust the empirical values for various kinds of layout regions. Therefore, if operators are not familiar

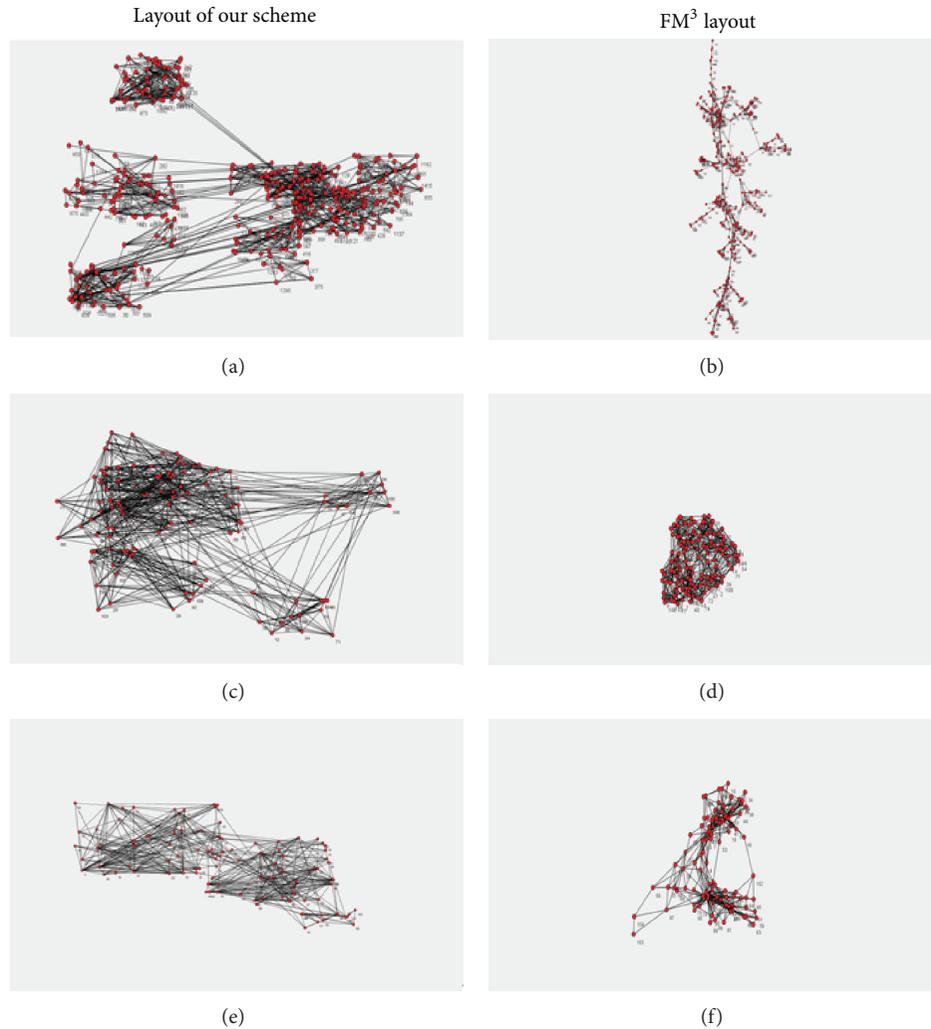


FIGURE 5: Comparison of layout effectiveness between our scheme and  $FM^3$ .

with the algorithms or do not have enough experience, they need much longer time to adjust experiments to achieve an ideal drawing result. However, in our scheme, not only four communities are recognized correctly, but also the layout space is better for visualization. Besides, it only needs one parameter to be configured (the percentage of users' expectation on final layout space). Thus, it is easy for users to understand and operate.

Figure 5 shows the comparison of the drawing results between our scheme and  $FM^3$  on three groups of real data sets with some certain community structures. The first data set is a researchers collaboration relationship graph, "subScience," with 379 vertices and 914 edges; the second data set is a competition schedule graph of a football club, "football," with 114 vertices and 616 edges; the third is the sales situation of books about American Politics on <http://www.amazon.com/>, "polbooks," with 105 vertices and 882 edges.

For the first data set, our scheme can recognize the main community's structures and make full use of layout space; for the second, our scheme can present the distribution of

competitions of teams; for the 3rd data set, our scheme can partition 3 categories of buyers and make good use of layout space. In general, our scheme has better layout effects than  $FM^3$ .

## 7. Conclusion

We studied graph drawing schemes and propose a new scheme for social networks which improves the graph drawing effectiveness. Besides, we compare the effectiveness of our scheme with  $FM^3$  in experiment. The experiment result shows that our scheme can effectively extract the community structure of the social network to apply into drawing algorithm and achieve a clearer diagram.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research was supported by National Natural Science Foundation of China (no. 61100192) and Research Fund for the Doctoral Program of Higher Education of China (no. 20112302120074) and was partially supported by Shenzhen Strategic Emerging Industry Development Foundation (no. JCYJ20120613151032592 and no. ZDSY20120613125016389), National Key Technology R&D Program of MOST China under Grant no. 2012BAK17B08 and National Commonweal Technology R&D Program of AQSIQ China under Grant no. 201310087. The authors thank the reviewers for their comments.

## References

- [1] Y. F. Hu, "Visualizing graphs with node and edge labels," *ACM Computing Research Repository*, 2009.
- [2] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software—Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
- [3] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon, "Multiscale visualization of small world networks," in *Proceedings of the 9th annual IEEE Conference on Information Visualization (InfoVis '03)*, pp. 75–81, Washington, DC, USA, October 2003.
- [4] S. Takahashi and S. Miyashita, "A constraint-based approach for visualization and animation," *Constraints*, vol. 3, no. 1, pp. 61–86, 1998.
- [5] J. Heer and D. Danahoy, "Vizster: visualizing online social networks," in *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '05)*, pp. 32–39, Minneapolis, Minn, USA, 2005.
- [6] D. Archambault, T. Munzner, and D. Auber, "Smashing peacocks further: drawing quasi-trees from biconnected components," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 813–820, 2006.
- [7] N. Elmqvist, T.-N. Do, H. Goodell, N. Henry, and J. Fekete, "ZAME: interactive large-scale graph visualization," in *Proceedings of IEEE Pacific Visualization Symposium (PacificVIS '08)*, pp. 215–222, IEEE, Kyoto, Japan, March 2008.
- [8] N. Henry and J. Fekete, "MatrixExplorer: a dual-representation system to explore social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 677–684, 2006.
- [9] N. Henry and J. D. Fekete, "MatLink: enhanced matrix visualization for analyzing social networks," in *Human-Computer Interaction—INTERACT 2007*, Lecture Notes in Computer Science, pp. 288–302, Springer, Berlin, Germany, 2007.
- [10] N. Henry, J. Fekete, and M. J. McGuffin, "NodeTriX: a hybrid visualization of social networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1302–1309, 2007.
- [11] J. Abello and F. Van Ham, "Matrix zoom: a visual interface to semi-external graphs," in *Proceedings of the IEEE Symposium on Information Visualization (INFO VIS '04)*, pp. 183–190, October 2004.
- [12] A. Frick and A. Ludwig, "A fast adaptive layout algorithm for undirected graphs," in *Lecture Notes in Computer Science*, vol. 894 of *Lecture Notes in Computer Science*, pp. 388–403, Springer, Berlin, Germany, 1995.
- [13] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon, "Multiscale visualization of small world networks," in *Proceedings of the 9th Annual IEEE Symposium on Information Visualization (InfoVis '03)*, pp. 75–81, October 2003.
- [14] T. Dwyer and K. Marriott, "Integrating edge routing into force-directed layout," in *Graph Drawing*, vol. 4372 of *Lecture Notes in Computer Science*, pp. 8–19, Springer, Berlin, Germany, 2007.
- [15] S. Hachul and M. Junger, "Drawing large graphs with a potential field-based multi-level algorithm," in *Graph Drawing*, vol. 3383 of *Lecture Notes in Computer Science*, pp. 285–295, Springer, Berlin, Germany, 2004.
- [16] A. Frick and A. Ludwig, "A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration)," in *Graph Drawing*, vol. 894 of *Lecture Notes in Computer Science*, pp. 388–403, Springer, Berlin, Germany, 1995.
- [17] R. Davidson and D. Harel, "Drawing graphs nicely using simulated annealing," *ACM Transactions on Graphics*, vol. 15, no. 4, pp. 301–331, 1996.
- [18] S. Hachul and M. Jünger, "An experimental comparison of fast algorithms for drawing general large graphs," in *Graph Drawing*, vol. 3843 of *Lecture Notes in Computer Science*, pp. 235–250, Springer, Berlin, Germany, 2006.
- [19] C. Walshaw, "A multi-level algorithm for force-directed graph drawing," in *Graph Drawing*, Lecture Notes in Computer Science, pp. 253–285, Springer, Berlin, Germany, 2003.
- [20] Y. Koren, L. Carmel, and D. Harel, "ACE: a fast multiscale eigenvectors computation for drawing huge graphs," in *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '02)*, pp. 137–144, 2003.
- [21] D. Archambault, T. Munzner, and D. Auber, "TopoLayout: multilevel graph layout by topological features," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 2, pp. 305–316, 2007.

## Research Article

# Video Texture Synthesis Based on Flow-Like Stylization Painting

**Qian Wenhua, Xu Dan, Yue Kun, and Guan Zheng**

*Department of Computer Science and Engineering, School of Information Science and Engineering,  
Yunnan University, Kunming 650091, China*

Correspondence should be addressed to Qian Wenhua; [qwhua003@sina.com](mailto:qwhua003@sina.com)

Received 19 March 2014; Accepted 7 June 2014; Published 15 July 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Qian Wenhua et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The paper presents an NP-video rendering system based on natural phenomena. It provides a simple nonphotorealistic video synthesis system in which user can obtain a flow-like stylization painting and infinite video scene. Firstly, based on anisotropic Kuwahara filtering in conjunction with line integral convolution, the phenomena video scene can be rendered to flow-like stylization painting. Secondly, the methods of frame division, patches synthesis, will be used to synthesize infinite playing video. According to selection examples from different natural video texture, our system can generate stylized of flow-like and infinite video scenes. The visual discontinuities between neighbor frames are decreased, and we also preserve feature and details of frames. This rendering system is easy and simple to implement.

## 1. Introduction

Most natural phenomena can use image or video to demonstrate directly. However, both natural and artificial phenomena cannot adequately be captured by a single static photo or image. On the other hand, though video is considered as the best media to show natural scene, it also has some shortcomings, such as if video scene should be stored on a computer or other storage devices, we always apply finite duration video clip. Therefore, video needs a beginning, a processing, and the ending frame. Further, though video captures the time-varying behavior of the phenomenon, it lacks the “timeless” quality of the photograph or image. In addition, it needs too much storage space.

In order to solve the above problems, Schödl and Essa proposed a new type of medium called “video texture” [1]. Video texture has qualities between photograph and video. This medium provides a continuous, infinitely varying stream of video images, and it can show an arbitrary length of video. The natural short video can be captured to synthesize infinite scene based on the technique of video texture synthesis. Then video texture works well for motions not only repetitive but also quasi-repetitive, such as swaying trees, waterfalls, and flickering flames. So, video texture is the best media to represent natural phenomena of our nature. Because of

its wide application in the entertainment and industry and so forth, computer-based phenomena video texture becomes hot topic.

Generally, video texture may be regarded as expansion of image not only in the time domain but also in natural phenomena simulation. Schödl and Essa also employed video texture technique to animation. They used L2 distance to measure similarity between neighbor frames and maintained playing continuity among video frames. The proposed method has realized highly third dimension scenery animation successfully. Based on image cutting and synthesizing technique, Kuwahara improved video texture synthesis method. To ensure continuity, the input frames were divided into some parts, and then they would be synthesized to video texture [2]. Bhat et al. put forward an editing nature phenomena method to generate the final video texture. Their method first analyzed the motion of texture particles and segmented frames into patches, then they reorganized the patches to reduce error dither [3]. Their method can merge an arbitrary length video based on natural scene. Agarwala et al. described an automatic method for taking the output of a single panning video camera and creating a panoramic video texture [4]. Their method applied a dynamic programming step through hierarchical min-cut optimization process to improve synthesis speed.

Though there are so many video synthesis methods to synthesize natural video phenomena, the synthesis results are traditionally striven photorealism scenes. That is to say photorealism effects can not produce artistic sensation. Sometimes, people want to enjoy artistic video with cartoon, oil, and water painting effects. In addition, photorealistic images often include more details than necessary to communicate intended information. It is known that nonphotorealistic rendering (NPR) is the subject of intense debate for representative methods of artistic style creation. Artistic expression can often convey a specific mood which is difficult to imbue in a photorealistic scene. NPR also can focus viewer's attention on the important information while downplaying extraneous or unimportant features. Additionally, a NPR look is often more engaging than the traditional, photorealism computer graphic rendering [5]. So, if video texture can be rendered to some nonphotorealistic effects, we can obtain artistic video with infinite playing scene.

Some NPR methods address the computer generated abstract stylization artistic effects, and this process fit perceptual task performance and evaluation. Decaudin applied modeling tools to generate abstract effect from natural scene [6], and his method kept details very well and shadow stayed coherent. Based on mean-shift, DeCarlo and Santella extended three-dimensional video volumes to automatically generate abstract artistic [7]. Particularly in the presence of occlusions and camera movement, contour tracking required substantial user correction of the segmentation results.

This paper presents an improved method to synthesize phenomena video with flow-like abstract painting. The new NP-video model can convert natural phenomena to a flow-like stylization video fast, and we also can achieve the infinite video texture easily. The amalgamations of the NPR and video texture synthesis can make the user enjoy artistic phenomena work, looking on the original phenomena in different artistic style, immersing vision and spirit into the artistic kingdom, and avoid some critical problems in photorealism systems.

## 2. NP-Video Phenomena Model System

We will introduce the basic architecture of our NP-video phenomena model in this section. As shown in Figure 1, our system proceeds in two steps. (1) Flow-like effects rendering: the natural phenomena video should be rendered to NPR artistic scene. (2) Video texture synthesis: the phenomena with flow-like effect will be synthesized to infinite playing video. We define the final play sequence of video texture to eliminate visual discontinuities. In addition, frames dividing and recombination method is used to arrive coherent playing frames. Our architecture supports a common NPR technique merging to video texture, and the final NP-video can give user flexibility and enjoyment.

## 3. Flow-Like Artistic Style Simulating

Many arts create flow-like artistic work, and people always enjoy this work. Van Gogh and Munch emphasized flow-like structures and directional features in their paintings.

The paintings are so famous because they are harmonic, interesting, and pleasant. Directional coherence and flow structure in this artistic work can help salience region features and boundaries. It also helps to evoke viewer's imagination. So, before synthesizing video texture, we transfer the input phenomena to flow-like art first, and our method provides a versatile NPR flow-like rendering interface.

As same as the abstract effects, flow-like stylization artistic work can generate using edge-preserving filter method. Bilateral filter [8] and mean-shift method [9] are very famous examples of edge-preserving filter. Kang et al. improved the filter shapes, which is based on the vector field derived from the salient image features [10]. However, because edge-preserving filter cannot obtain good result with weak vector, this method may lose some details such as weak edge and structure. Kuwahara filter provides an overall stereoscopic painting visual and maintains a uniform level of abstraction across the image [11]. Unfortunately, Kuwahara filter can result in clustering artifacts during filter process. Papari et al. introduced a new weighting filter window, and the shape of the filter window can be changed based on local feature directions [12]. Kyprianidis also generated an abstract effect along the local feature directions [13], but the stylized surface of their methods cannot preserve both spatial and temporal coherence. Like bilateral filter and Kuwahara filter, Kang and Lee used Shock filter and mean curvature flow to generate abstraction effect, but the boundaries are not salient and simplified [14]. Recently, Kim et al. put forward bristle maps to generate aggregation, abstraction, and stylization of spatiotemporal data [15]. Isenberg explored visual map technique for generating of stylized renderings of 2D map data [16]. His method can render different stylization artistic work.

Though there are many techniques that can be used to obtain flow-like artistic work, these methods should be improved to adopt phenomena video. After capturing phenomena video, we will resolve this video and generate flow-like stylization painting of every frame. Our method begins with the anisotropic Kuwahara filtering. Firstly, eigenvalues and eigenvectors can be used to calculate structure tensor. Secondly, local orientation and anisotropy structure are used to guide Kuwahara filter process. Then the flow-like effects can be generated using improved line integral convolution.

*3.1. Local Orientation Calculate.* Brox et al. obtained anisotropy through calculating eigenvalues and eigenvectors of structure tensor [17]. We use Brox's method to calculate structure tensor first. Let  $G_{\sigma,x}$  and  $G_{\sigma,y}$  be the spatial derivatives in  $x$ - and  $y$ -direction and let  $f$  be the input frame with standard deviation  $\sigma$ ; structure tensor can be calculated as

$$G_{\sigma,y} = \frac{1}{2\pi\sigma^2} e^{-(y/\sqrt{2}\sigma)^2}, \quad G_{\sigma,x} = \frac{1}{2\pi\sigma^2} e^{-(x/\sqrt{2}\sigma)^2}. \quad (1)$$

The pixel's gradient vector is  $\mathbf{G} = (G_x, G_y)^T$ . Let  $*$  denote convolution operation; the partial derivatives of  $f$  can be calculated:

$$fx = G_{\sigma,x} * f, \quad fy = G_{\sigma,y} * f. \quad (2)$$

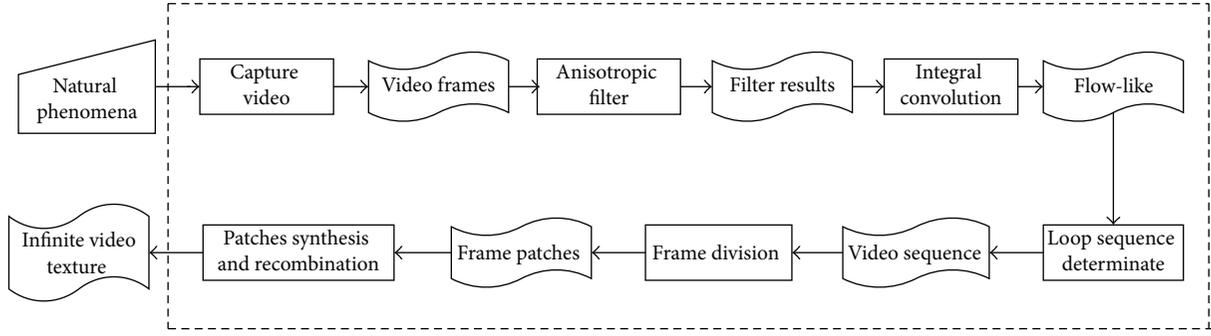


FIGURE 1: NP-video phenomena of our system.

The structure tensor of each video frame  $\mathbf{f}$  can be defined as

$$\mathbf{T}_{x,y} = \mathbf{G} \times \mathbf{G}^T = \begin{bmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}. \quad (3)$$

Let  $d_1$  and  $d_2$  be the minimum and the maximum eigenvalue of  $\mathbf{T}$ , which represents gray scale variation degree of vector  $\mathbf{v}_1$  and  $\mathbf{v}_2$

$$g_{\sigma,\gamma,\theta} = \frac{d_1 + d_2}{2\pi\sigma^2} \exp \left[ -\frac{x^2 + r^2 y^2}{2\theta^2 / (d_1 + d_2)} \right], \quad (4)$$

where parameter  $\sigma$  denotes filter kernel of Gauss function variance,  $r$  denotes filter radius which is calculated using  $d_1/d_2$ . If  $d_1$  and  $d_2$  are close, this filter field tends to be circular. The parameter  $\theta$  represents edge direction of pixel  $(x, y)$ , which reflects curve direction of eigenvalue. Then anisotropic structure tensor can be expressed as  $\mathbf{T}'$ :

$$\mathbf{T}' = g_{\sigma,\gamma,\theta} \times \mathbf{T} = \begin{bmatrix} g_{\sigma,\gamma,\theta} \times \mathbf{T}_{11} & g_{\sigma,\gamma,\theta} \times \mathbf{T}_{12} \\ g_{\sigma,\gamma,\theta} \times \mathbf{T}_{21} & g_{\sigma,\gamma,\theta} \times \mathbf{T}_{22} \end{bmatrix}. \quad (5)$$

Because the local vector field maybe discontinuous, we use Gaussian filter technique to smooth this vector field, and eigenvalues of structure tensor are given:

$$Z_{1,2} = \frac{T'_{11} + T'_{22} \pm \sqrt{(T'_{11} - T'_{22})^2 + 4T'_{12}T'_{21}}}{2}. \quad (6)$$

Then, the eigenvector in local curve direction is calculated as

$$t = \begin{pmatrix} Z_1 - T'_{11} \\ -T'_{12} \end{pmatrix}. \quad (7)$$

The local orientation is  $\varphi = \arg t$ , and we calculate anisotropy based on Yang's method [18]:

$$A = \frac{Z_1 - Z_2}{Z_1 + Z_2}, \quad (8)$$

where parameter  $A$  ranges between 0 and 1, which denotes isotropic and anisotropic regions, respectively.

**3.2. Anisotropic Filter.** Because pixel  $(i, j)$  of filter result is determined by the spatial distance from the center pixel  $(x, y)$ , as well as its relative intensity difference, the filtering output  $Q$  at pixel  $(i, j)$  is calculated:

$$Q_{i,j}(x, y) = \frac{\sum_{x,y \in \Omega} a_{i,j}(x, y) m_{i,j}(x, y)}{\sum_{x,y \in \Omega} a_{i,j}(x, y)}, \quad (9)$$

where parameter  $\Omega$  denotes filter spatial space.  $G_{i,j}$  is the local intensity or color similarity with weighted  $w_{i,j}$  and  $a_{i,j}$  denotes the squared standard deviations,  $m$  and  $a$  can be defined:

$$m_{i,j}(x, y) = \frac{1}{k} \sum_{x,y \in \Omega} f_{i,j}(x, y) w_{i,j} G_{i,j}(x, y),$$

$$a_{i,j}(x, y) = \frac{1}{k} \sqrt{\sum_{x,y \in \Omega} f_{i,j}^2(x, y) w_{i,j} G_{i,j}(x, y) - m_{i,j}^2}, \quad (10)$$

$$k = \sum_{x,y \in \Omega} w_{i,j} G_{i,j}(x, y).$$

Let  $\varphi$  be the local orientation,  $A$  be anisotropy, and  $R\varphi$  be matrix defining a rotation, the filter spatial space  $\Omega$  is

$$\Omega_{i,j} = \{(x, y) \in R^2 : \text{abs}(SR_{-\varphi}(x, y)) \leq h\}, \quad (11)$$

where parameter  $h = 2\sigma_r$ . Parameter  $S$  is coefficient to adjust anisotropy  $A$ . So,  $SR_{-\varphi}$  is a linear coordinate transform which maps  $\Omega$  to a disc with radius  $h$ . This anisotropic filter ensures that if local region has lower standard deviation, the more weight value is given during filter process. Figure 2 shows our filter results. Figure 2(a) is input image, and Figures 2(b) and 2(d) are our results with different  $h$ . If  $h$  is smaller, more details will be retained. Figures 2(e) and 2(h) are filter results with other methods. Figure 3 also shows another result with blowing water. Figures 3(a) and 3(d) are different frame of natural texture, and Figures 3(e) and 3(h) are filter results. From these results, we can find that our method can maintain more details and features, such as edge, color, and local structure. Our method also eliminates artifacts of shape-simplifying and Shock filter method.



FIGURE 2: Comparison of different filter.

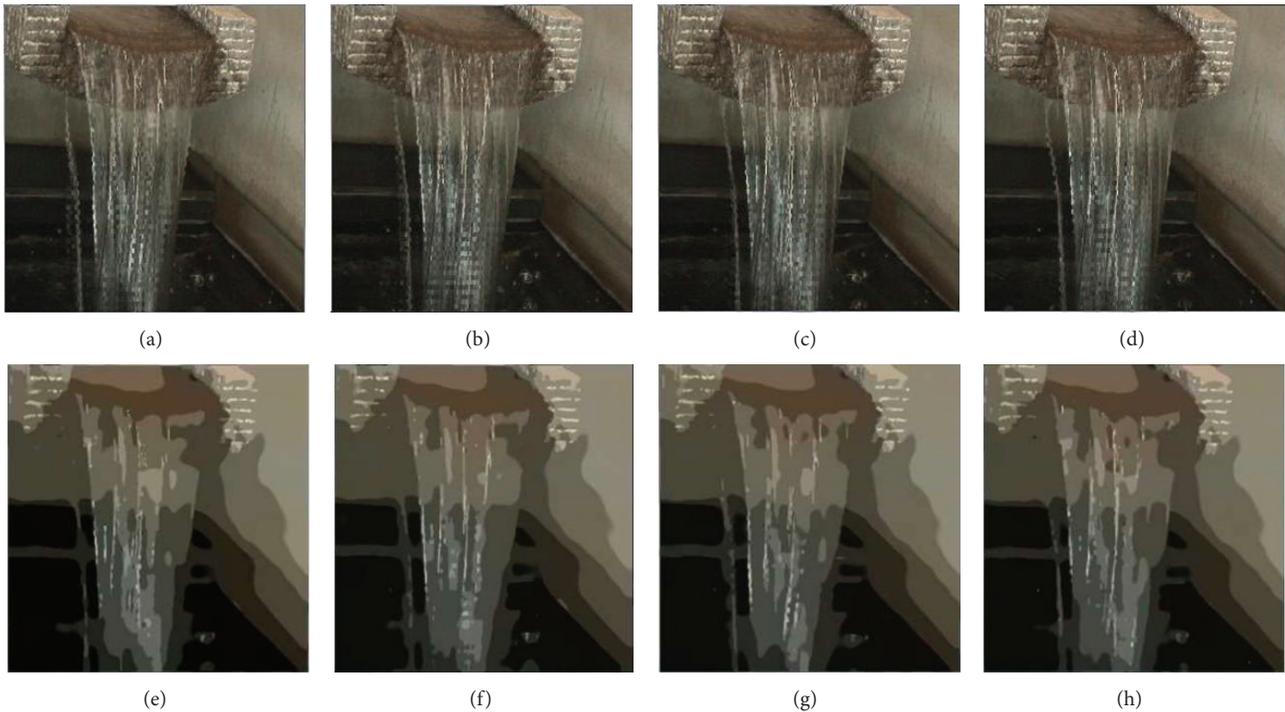


FIGURE 3: Video texture filter of blowing water.

**3.3. Integral Convolution.** We benefit from our efficient stylization painting not only in anisotropic filtering, but also with flow-like effect. Papari and Petkov applied integral convolution technique to produce nonphotorealistic rendering glass

patterns texture effect [19]. We refer their integral convolution method to generate final artistic work. During convolution process, our method calculates a local stream line which moves out from the positive and negative directions. Let  $I_{out}$

be flow-like output image, and let  $\sigma(s)$  be a stream line with length of  $L$ . Parameter  $p_0$  denotes each pixel in  $\sigma(s)$ . Let  $s_0$  be current pixel; we can obtain  $p_0 = \sigma(s_0)$ . Parameter  $k(s)$  denotes convolution kernel, and we utilize Hanning window function in this paper.  $I(\sigma(s))$  denotes all the pixels in this line, the convolution implement is defined as follows [20]:

$$I_{\text{out}}(p_0) = \int_{s_0-L/2}^{s_0+L/2} k(s-s_0) I(\sigma(s)) ds, \quad (12)$$

where parameter  $L$  denotes convolution length. If  $L$  is small, an insufficient amount of filtering flow occurs. However, too large  $L$  will result in close convolution value and regular output. Because variance can well reflect local statistic information, it can be used to calculate local convolution length automatically:

$$L_{(x,y)} = L_{\text{max}} - \frac{L_{\text{max}} - L_{\text{min}}}{\sigma_{\text{max}} - \sigma_{\text{min}}} (\sigma_{(x,y)} - \sigma_{\text{min}}), \quad (13)$$

where  $L_{\text{max}}$  and  $L_{\text{min}}$  are the largest and the smallest integral length which can be assigned. Parameter  $\sigma$  is variance value of local area. Based on anisotropic filtering and integral convolution, input video can be rendered to flow-like effects. Next we will synthesize nonphotorealistic video texture to infinite video scene.

#### 4. Phenomena Video Texture Synthesis

After phenomena video is rendered to flow-like effects, we will synthesize video texture clip to infinite video scene through video texture synthesis method. There are two significant challenges for video texture synthesis. (1) Because infinite video scene is generated through playing period video sequence repeat, the end frame in one sequence has large difference with the beginning frame in the next loop sequence. So, there are visual breaks between different loop arrays. (2) During video texture playing, visual discontinuities exist in different frames in one sequence. When frames dispose to nonphotorealistic artistic work, the discontinuities will be more prominent. So, we will determinate loop sequence of input video frames first. Then, methods of frame division and recombination, patches synthesis will be applied to eliminate visual discontinuities.

**4.1. Loop Sequence Determinate.** Many natural phenomena present playing cycle in video texture. If we play video cycle of natural phenomena repeat, people can enjoy infinite scene, and visual discontinuities will be reduced. So, we should find the loop cycle of a certain video texture. We use Schödl's method of L2 distance to measure the similarity among different video frames [1].

Because human's vision is sensitive to the change of luminance in frames, during the process of L2 distance calculation, we transfer the color space of frame from RGB to YCbCr. So L2 distance can be calculated only in luminance

channel Y. If  $N_i, N_j$  denote two different frames in one video texture, similarity can be acquired as

$$E(N_i, N_j) = \sum_{p \in n_i, p' \in n_j} [I_i(p) - I_j(p')]^2, \quad (14)$$

where  $I_i, I_j$  are luminance information of  $N_i$  and  $N_j$ . Parameters  $p, p'$  represent pixel positions of  $N_i$  and  $N_j$ . So, if  $E(N_i, N_j)$  is smaller than the threshold  $k$ , frames  $N_i$  and  $N_j$  can be regarded as the same loop sequence. That is to say,  $N_j$  and  $N_i$  are neighbor frames. If  $n$  denotes frame number of input natural video texture and  $m$  denotes each frame in  $n$ ; threshold  $k$  can be calculated:

$$k = \frac{1}{n-1} \sum_{m=1}^{n-1} E_{m,m+1}. \quad (15)$$

Based on similarity, video texture can be divided into different sequences. Because L2 distance of neighbor frames in one sequence is small, we consider using different sequences to combine final infinite video scene. If L2 distance satisfies formula (16), then there only exists one sequence, and this sequence of input video texture has  $n$  frames:

$$E_{1,2} < E_{1,3} < \dots < E_{1,m} < E_{1,m+1} > \dots > E_{1,n} < E_{1,n+1}. \quad (16)$$

Figure 4 shows one natural video texture of stream water. There are fifteen frames  $I_1$  to  $I_{15}$  in this video, and we can obtain similarity in Table 1 using formula (14).

From Table 1, L2 distance satisfied  $E_{1,2} < E_{1,3} < \dots < E_{1,7} > E_{1,8} > \dots > E_{1,14} < E_{1,15}$ . Then frames  $I_1$  to  $I_{14}$  belong to the one sequence, and fourteen is the cycle length of this video. If we want to obtain infinite video texture, frame  $I_1$  should be played when frame  $I_{14}$  is ended, and these fourteen frames should play repeatedly.

**4.2. Frame Division and Recombination.** Though L2 distance is smaller than threshold in one sequence, when playing video sequence repeatedly, there also exists visual discontinuity between neighbor frames, such as ending and beginning frames. Therefore, we utilize Bhat's method to eliminate these discontinuities. Most natural phenomena video texture is similar particle moving, such as waterfall, fountain, flame, and stream water. Figure 5(a) shows a particle moving procedure along a flow line, and we can find that a patch texture of one frame in Figure 5(b) will shift to neighbor frame in next playing moment.

One frame in video texture can be divided into some texture patches, such as Figure 5(c). If the sequence of input video is four, then each frame should be divided into four texture parts, and each patch will shift along flow line like Figure 5(a). The patch will move to different position in neighbor frames. For example, when  $T = 3$ , the patch moves to the third part in the fourth frame, whose beginning is the first part in the second frame. If frames are divided into three parts, Figure 5(d) shows the result after frames recombined together, and the same color stands for the same texture patch. We can find that the same patch is in different frames when this video scene is playing.

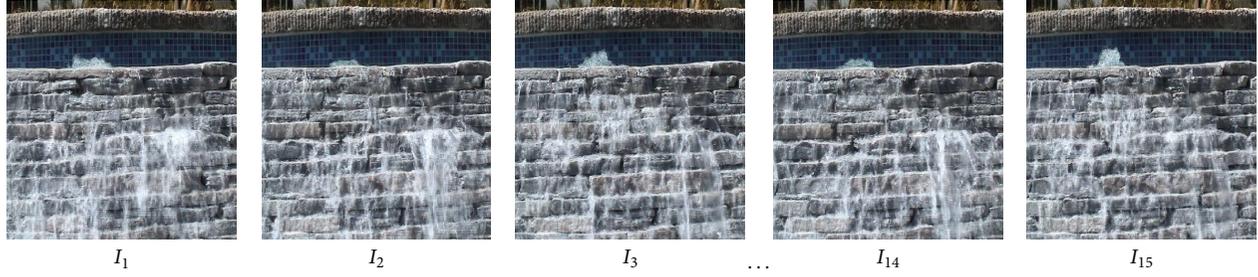


FIGURE 4: Natural input video texture.

TABLE 1: L2 distance between the first frame and others.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4776	6206	7069	7124	7125	7454	7453	7114	6952	6117	5822	5549	5049	5305

TABLE 2: Original frame loop sequence.

										→ T
1-1	2-1	3-1	4-1	5-1	1-1	2-1	3-1	4-1	5-1	
1-2	2-2	3-2	4-2	5-2	1-2	2-2	3-2	4-2	5-2	
1-3	2-3	3-3	4-3	5-3	1-3	2-3	3-3	4-3	5-3	

TABLE 3: Frame division and recombination.

									→ T
1-1	2-1	3-1	1-1	2-1	3-1	1-1	2-1	3-1	
1-2	2-2	3-2	4-2	2-2	3-2	4-2	2-2	3-2	
1-3	2-3	3-3	4-3	5-3	3-3	4-3	5-3	3-3	

Table 2 shows original playing sequence without frame dividing and recombination. The different column expresses different playing video frames. When this video texture is playing repeatedly, there is a visual discontinuity between the end frame of front cycle and the beginning frame of next cycle. If sequence length of this input video texture is three, then we divide each frame into three texture parts. After using frames recombination method, we obtain new group frames like Table 3. From Table 3, we can find that the discontinuity is scattered to different frames and forms a ladder-shaped discontinuity. Because this technique disperses visual dither to the whole playing sequence, visual break between different loop arrays will be reduced.

The method of frames division and recombination is suitable to natural phenomena moving downwards, such as rivers, waterfalls, and fallen leaves. If input video texture moves upwards, such as fountains or flames, the frames should be divided into patches and recombination retroflex like Table 3. However, if a natural phenomenon is complicated and its sequence is more than ten, it is impossible to divide every frame into ten patches. Then, if video texture's sequence is more than appointed threshold, each frame of this video can be divided using this threshold.

**4.3. Patches Synthesis Method.** Though visual discontinuity can be removed applying frame division and recombination, some distinct boundaries will be produced between different patches in one frame. Texture synthesis method can be used to eliminate these boundaries.

Texture synthesis method can be used to synthesize texture sample to an unlimited size image. Efros and Freeman proposed a texture quilting method to synthesize surprisingly

good results for a wide range of textures [21]. Efros and Freeman chose some blocks from the sample texture to fill in the final texture image. The blocks satisfy the overlap constraints within some error tolerance. As shown in Figure 6, there have been some overlap regions between neighbor blocks.

To eliminate discontinuity between overlap regions, based on the error cost path in overlap region, Efros found the best seam line to synthesize neighbor blocks. Hence, when frames of input video texture are divided into patches, we apply Efros' method to recombine these patches. Let  $E_{(i,j)}$  be error cost function and  $M$  width of overlap region.  $E_{color}$  denotes distance error of color,  $E_{geometry}$  denotes structure error, and the minimum error boundary can be calculated as

$$E(i, j) = \min \sum_{k=1}^M (E_{color}(i, j)^2 + E_{geometry}(i, j)). \quad (17)$$

Figure 7 shows the slide window when we calculate  $E(i, j)$ . Let  $E_{pre}$  be error cost of previous slider window and  $e_{old}$  error cost of region A;  $e_{new}$  is error cost of region B. Equation (18) is used to calculate the error cost between *New* and *Old* windows, and this method can optimize synthesis process between different patches:

$$E(N_1, N_2) = E_{pre}(N'_1, N'_2) - e_{old}(N'_1, N'_2) + e_{new}(N_1, N_2). \quad (18)$$

## 5. Experiment Results

To verify the feasibility and effectiveness of our proposed methods in this paper, we implemented the relevant algorithms and carried out our corresponding experiments.

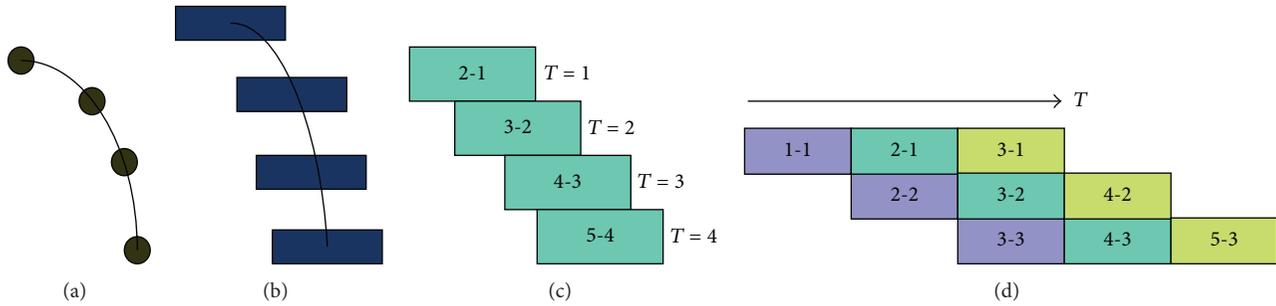


FIGURE 5: Natural phenomena moving like particle.

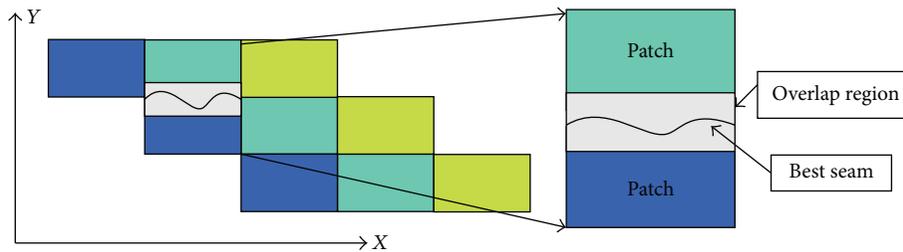


FIGURE 6: Overlap regions between patches.

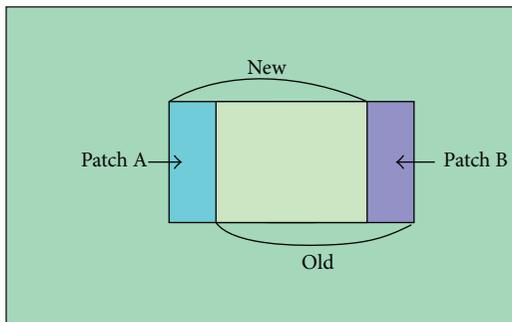


FIGURE 7: The slide of matching window.

Figures 7 and 8 show the experimental results of video texture about blowing water and fountain scene. Our system works on Microsoft Visual C and the OpenGL texture mapping routines.

Figure 8 shows blowing water video scene, and this input video has seventeen frames. Figure 9 shows fountain video scene, and this input video has fifty-eight frames. Figures 8(a) and 8(d) and Figures 9(a) and 9(d) are input frames from natural texture samples. The size of input frames is  $256 * 256$ . Based on L2 distance calculation, we choose eight frames as loop sequence from blowing water and choose eighteen frames as loop sequence from fountain video scene. Figures 8(e) and 8(h) and Figures 9(e) and 9(h) show final frames with flow-like stylization effects based on our nonphotorealistic rendering and video texture synthesis methods. We can obtain the infinite video scenes playing loop sequence repeat, and this infinite video scenes has flow-like artistic effect.

We also can conclude that the experimental results are realistic and they all have some esthetics.

### 6. Conclusions

In this paper, we have presented an NP-video texture synthesis system based on natural phenomena. Based on a simple and effective anisotropic filtering and integral convolution, flow-like stylization video effects can be generated. In addition, through loop sequence determination, frame division, and recombination, our algorithm removes visual discontinuities between different frames during video texture synthesis process. The experiment results indicate that our algorithm is easy to synthesize infinite video scene with nonphotorealistic artistic work.

An obvious limitation of our method is that the synthesis speed should be improved. Some strategies should be used to protect highly important area in the frames. We may further extend the GPU implementation of our algorithm to process video in real time.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgments

This research was funded by the Grants (nos. 60663010, 61163019, and 61271361) from the Research Natural Science Foundation of China, (no. 2010CD024) from the Research Foundation of Yunnan Province, (no. 2012Z047) from

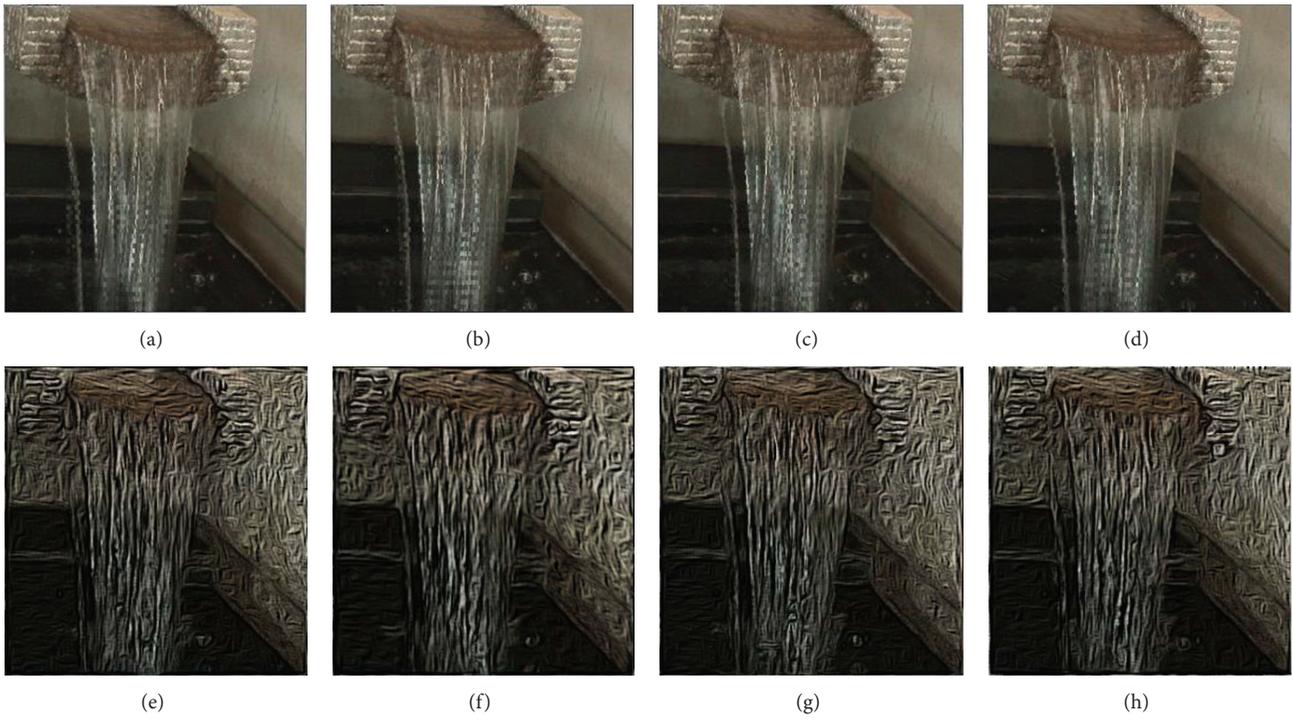


FIGURE 8: Video texture of blowing water.



FIGURE 9: Video texture of fountain.

the Research Foundation of the Educational Department of Yunnan Province, and (no. 20125301120008) from the Research Foundation of New Teacher Fund for Doctor Station, the Ministry of Education.

## References

- [1] A. Schödl and I. Essa, "Controlled animation of videosprites," in *Proceedings of ACM SIGGRAPH*, pp. 121–127, San Antonio, Tex, USA, 2002.
- [2] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," *ACM Transactions on Graphics (TOG)—Proceedings of ACM SIGGRAPH 2003*, vol. 22, no. 3, pp. 277–286.
- [3] D. H. Bhat, S. M. Seitz, K. J. Hodgins et al., "Flow based video synthesis and editing," *Proceedings of ACM SIGGRAPH*, vol. 23, no. 3, pp. 360–363, 2004.
- [4] A. Agarwala, K. C. Zheng, C. Pal et al., "Panoramic video textures," in *Proceedings of the ACM SIGGRAPH (SIGGRAPH '05)*, pp. 821–827, August 2005.
- [5] A. W. Klein, W. Li, M. M. Kazhdan, W. T. Corrêa, A. Finkelstein, and T. A. Funkhouser, "Non-photorealistic virtual environments," in *Proceedings of the 27th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*, pp. 527–534, July 2000.
- [6] D. Philippe, *Cartoon-looking rendering of 3D-scenes*, [Ph.D. thesis], Université de Technologie de Compeigne, Inria, France, 1996.
- [7] D. DeCarlo and A. Santella, "Stylization and abstraction of photographs," in *Proceedings of the ACM Transactions on Graphics (ACM SIGGRAPH '02)*, pp. 769–776, July 2002.
- [8] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the 6th International Conference on Computer Vision (ICCV '98)*, pp. 839–846, Bombay, India, January 1998.
- [9] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
- [10] H. Kang, S. Lee, and C. K. Chui, "Flow-based image abstraction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 1, pp. 62–76, 2009.
- [11] M. Kuwahara, K. Hachimura, S. Eiho, and M. Kinoshita, *Digital Processing of Biomedical Images*, Plenum Press, 1976.
- [12] G. Papari, N. Petkov, and P. Campisi, "Artistic edge and corner enhancing smoothing," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2449–2462, 2007.
- [13] J. E. Kyprianidis, H. Kang, and J. Döllner, "Image and video abstraction by anisotropic Kuwahara filtering," *Computer Graphics Forum*, vol. 28, no. 7, pp. 1955–1963, 2009.
- [14] H. Kang and S. Lee, "Shape-simplifying image abstraction," *Computer Graphics Forum*, vol. 27, no. 7, pp. 1773–1780, 2008.
- [15] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. S. Ebert, and T. Isenberg, "Bristle maps: a multivariate abstraction technique for geovisualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1438–1454, 2013.
- [16] T. Isenberg, "Visual abstraction and stylisation of maps," *The Cartographic Journal*, vol. 50, no. 1, pp. 8–18, 2013.
- [17] T. Brox, J. Weickert, B. Burgeth, and P. Mrázek, "Nonlinear structure tensors," *Image and Vision Computing*, vol. 24, no. 1, pp. 41–55, 2006.
- [18] G. Z. Yang, P. Burger, D. N. Firmin, and S. R. Underwood, "Structure adaptive anisotropic image filtering," *Image and Vision Computing*, vol. 14, no. 2, pp. 135–145, 1996.
- [19] G. Papari and N. Petkov, "Continuous glass patterns for painterly rendering," *IEEE Transactions on Image Processing*, vol. 18, no. 3, pp. 652–664, 2009.
- [20] B. Cabral and L. C. Leedom, "Imaging vector fields using line integral convolution," in *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '93)*, pp. 263–270, August 1993.
- [21] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the Computer Graphics Annual Conference (SIGGRAPH '01)*, pp. 341–346, Los Angeles, Calif, USA, August 2001.

## Research Article

# Reliable Multihop Broadcast Protocol with a Low-Overhead Link Quality Assessment for ITS Based on VANETs in Highway Scenarios

Alejandro Galaviz-Mosqueda,<sup>1</sup> Salvador Villarreal-Reyes,<sup>2</sup> Hiram Galeana-Zapién,<sup>1</sup> Javier Rubio-Loyola,<sup>1</sup> and David H. Covarrubias-Rosales<sup>2</sup>

<sup>1</sup> Information Technology Laboratory, Center for Research and Advanced Studies (Cinvestav), Science & Technology Park TECNOTAM, Km. 5.5 Cd. Victoria-Soto La Marina Highway, 87130 Ciudad Victoria, TAMPS, Mexico

<sup>2</sup> Center for Scientific Research and Higher Education at Ensenada (CICESE), 22860 Ensenada, BC, Mexico

Correspondence should be addressed to Salvador Villarreal-Reyes; [svillar@cicese.mx](mailto:svillar@cicese.mx)

Received 8 March 2014; Accepted 21 June 2014; Published 15 July 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 Alejandro Galaviz-Mosqueda et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vehicular ad hoc networks (VANETs) have been identified as a key technology to enable intelligent transport systems (ITS), which are aimed to radically improve the safety, comfort, and greenness of the vehicles in the road. However, in order to fully exploit VANETs potential, several issues must be addressed. Because of the high dynamic of VANETs and the impairments in the wireless channel, one key issue arising when working with VANETs is the multihop dissemination of broadcast packets for safety and infotainment applications. In this paper a reliable low-overhead multihop broadcast (RLMB) protocol is proposed to address the well-known broadcast storm problem. The proposed RLMB takes advantage of the hello messages exchanged between the vehicles and it processes such information to intelligently select a relay set and reduce the redundant broadcast. Additionally, to reduce the hello messages rate dependency, RLMB uses a point-to-zone link evaluation approach. RLMB performance is compared with one of the leading multihop broadcast protocols existing to date. Performance metrics show that our RLMB solution outperforms the leading protocol in terms of important metrics such as packet dissemination ratio, overhead, and delay.

## 1. Introduction

Intelligent transport systems (ITSs) aim to integrate information and communication technologies with transportation systems to make transport more efficient, green, safe, and seamless. The ITSs rely on wireless technologies to achieve both vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications. In V2I communications, vehicles communicate with a fixed infrastructure via the wireless media. On the other hand, in the V2V approach vehicles are equipped with wireless communications solutions to directly communicate with vehicles nearby without the need for any infrastructure. The vehicles with V2V capabilities form an ad hoc network, commonly referred as vehicular ad hoc network (VANET). Therefore, in V2V each vehicle must be able to

send, receive, and relay safety or infotainment information throughout the VANET. When compared to V2I networks, VANETs provide ubiquitous information sharing and their use results in lower implementation costs, as they work without fixed access network nodes.

The research, development, and standardization communities have identified V2V communications as a key technology to radically improve road safety conditions. It is argued that V2V communications can potentially address above 79% of precrash scenarios involving unimpaired drivers [1]. Regarding nonsafety related applications, it is expected that V2V communications will allow a rapid development and deployment of infotainment applications such as multimedia streaming [2, 3], Internet access [4], and additional infotainment services (e.g., taxi service [5]).

Before VANET technology can fulfill all its expected potential, several difficulties must be addressed. Particularly, the design of methods to effectively disseminate messages through multiple hops is of paramount importance to successfully deploy both safety and nonsafety applications for ITSs [6–8]. For instance, road safety applications attempt to increase the awareness range of drivers by transmitting messages from vehicles internal sensors (e.g., speed) and about surrounding conditions (e.g., crash scenario). These messages are broadcasted to all vehicles located within a specific geographic region or zone of relevance (ZOR). For example, a ZOR can be defined by the lanes of vehicles travelling towards the crash site. From a network perspective, for this scenario the use of broadcast packets to disseminate the messages is more suitable than the use of unicast packets, as the message is of general interest. Additionally, the message must be delivered as soon as possible to all vehicles within the ZOR, such that preventive measures can be timely implemented. In this kind of pre/postcrash warning scenarios, a multihop broadcast (MB) protocol is needed to reach all vehicles within the ZOR at the lowest delivery time.

The design of MB protocols is a challenging task because of spatial-temporal changes in the wireless channel (e.g., fading), different mobility patterns followed by the vehicles, the density of vehicles, and infrastructure availability. In turn, these conditions are closely related with the specific deployment scenario of VANETs [9], the urban and highway scenarios being the most relevant. It is important to note that the constraints imposed by these scenarios are different. Therefore, a MB protocol designed for urban scenarios might not be able to cope with the higher speeds of vehicles in highway scenarios. Furthermore, ITSs implementation in urban scenarios is more likely to rely on V2I solutions, which makes the protocol design more tractable. On the other hand, V2V solutions are the most attractive technology for ITSs deployments in highway scenarios from the cost-efficiency point of view [10]. Nevertheless, the design of MB protocols for these deployments becomes a challenge because of the constraints imposed by highway scenarios. It is noteworthy that highways account for a significant amount of the road infrastructure deployed throughout several countries. For example, highways represent about 75% of the total statute miles in the US [11]. Hence, the design of MB protocols for highway scenarios is an important issue that must be addressed as recently research has pointed out [11–13].

Different wireless technologies have been considered to enable V2V communications (e.g., RFID, IEEE 802.11b, IEEE 802.15.4, Bluetooth, etc.). However, nowadays the most prominent option is the IEEE 802.11p standard [14, 15]. The IEEE 802.11p PHY layer is based on the IEEE 802.11a standard. Similarly, its MAC layer uses carrier sense multiple access with collision avoidance (CSMA/CA). The well-known enhanced distributed channel access (EDCA) mechanism defined in the IEEE 802.11e standard is included in IEEE 802.11p to provide four different access priorities: background, best effort, video, and voice, [14]. However, the IEEE 802.11p standard leaves open the design of efficient broadcast protocols and the solution of issues like the broadcast storm problem (BSP) [7].

The simplest protocol for MB is basic flooding, where each node that receives a packet for the first time retransmits it with no further restrictions. When using CSMA/CA the dissemination of packets by flooding can introduce an important number of redundant broadcasts. This is because of the shared wireless medium nature in CSMA/CA and the lack of any protection mechanism for the broadcast packets. This results in the BSP, where an increase in the medium access delay and in the number of collisions is observed. The BSP is prevalent in networks with high node densities, like those found in vehicular scenarios. The BSP has a negative impact on the arrival time of packets and it can even lead to a significant packet loss.

In order to solve the BSP, beaconless protocols for VANETs aiming at reducing the number of redundant broadcasts have been proposed in the literature [7, 16–18]. Unfortunately, these protocols are not efficient while trying to provide a good trade-off between overhead and reliability [19]. Beacon-assisted protocols that use the neighbors' information to reduce the redundant broadcast have been proposed as well for VANETs scenarios [6, 13, 19, 20]. Although beacon-assisted protocols have shown better performance than beaconless protocols, the accuracy of the information used to make a rebroadcast decision is highly dependent on the frequency of the beacon messages. Even though the overhead/reliability trade-off is better addressed in beacon-assisted protocols, the overhead introduced by the beacon messages can significantly affect the protocol performance.

This paper introduces a new reliable low-overhead multihop broadcast (RLMB) protocol with a cooperative link quality assessment for VANETs in highway scenarios. The proposed RLMB solution provides a good trade-off between overhead and reliability. This is addressed through a beacon-assisted approach along with an implicit acknowledgement mechanism and a position prediction algorithm.

The rest of the paper is structured as follows. Section 2 discusses the approaches taken by previous studies, while the details of the proposed RLMB protocol are presented in Section 3. The performance of the proposed solution is presented in Section 4, and the concluding remarks are presented in Section 5.

## 2. Related Work

As previously mentioned, multihop broadcast of packets in VANETs was initially addressed with simple flooding, resulting in the BSP. In order to solve the BSP problem, several broadcast storm mitigation protocols for vehicular scenarios have been proposed [12, 16, 19, 20]. In order to discuss different broadcast protocols previously proposed in the literature, the MB protocols will be classified into two main groups within this paper: beaconless (BL) and beacon-assisted (BA) protocols (see Figure 1). A brief explanation of these two groups is provided hereafter.

Basically, BL protocols only use information contained in the disseminated message to decide whether to retransmit it. On the other hand, BA protocols take advantage of the beacon messages that each node in the network transmits periodically. In addition, BA protocols can be further classified as

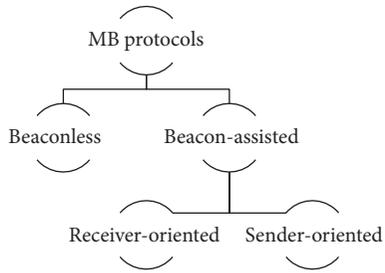


FIGURE 1: A classification of multihop broadcast protocols for VANETs.

sender-oriented or receiver-oriented. In the latter ones the rebroadcast decision is made at each node when the message is received. Contrastingly, in BA sender-oriented protocols, the next set of rebroadcast nodes is chosen a priori at the previous transmitter node. In both BL and BA protocols, relay decisions can be made based on operational parameters like received power, distance, density of neighbors, timer, or some combination of the former parameters.

There are several BL protocols reported in the literature [7, 16, 17]. In these protocols, each node determines whether to retransmit a message based only on the information contained in the disseminated message. In BL protocols the redundant broadcasts cannot be entirely eliminated, especially under the high variability of scenario conditions present in VANETs. Thus, in BL protocols the trade-off between overhead and reliability cannot be properly addressed.

In BA receiver-oriented (BARO) protocols the exchanged beacons are used for detection of different scenario conditions (e.g., vehicles density). If a current scenario condition reaches a predefined state (e.g., number of transmissions heard, number of neighbors found), then the message is disseminated with a BL approach. Otherwise, a strategy of store-carry-forward is applied (e.g., [12, 20, 21]). Hence as in BL protocols, the trade-off between overhead and reliability cannot be properly addressed in BARO protocols. Therefore, BL and BARO protocols are not suitable for applications with requirements such as high reliability and low overhead.

In the context of BA sender-oriented protocols, the set of relay nodes is formed a priori in the transmitter. The set is chosen based on the stored information gathered through the exchange of hello messages between the vehicles. Thus, given that each node has neighborhood information, the BA sender-oriented approach can potentially reduce the redundant broadcasts in a more efficient way than BL or BARO protocols.

In [22], the enhanced multipoint relay (EMPR) is proposed. This protocol considers the mobility of nodes and an additional area of coverage to select the set of relay nodes. In [23], BA sender-oriented protocol called BPAB is introduced. BPAB performs a repetitive 2-partition method to divide the area inside transmission range. Then, a vehicle to retransmit the message is chosen in the furthest segments. Nevertheless, EMPR and BPAB do not consider the fading nature of the wireless channel when selecting the set of relay nodes. Because of the time-varying channel conditions in V2V communications, the link quality between vehicles

could be significantly degraded. Thus, high levels of packet losses and/or delays can occur, as the fading nature of the wireless channel was not considered in the protocol design. Additionally, because of the multipath components, messages beyond the vehicles nominal radio range (NRR) can be occasionally received. As such, these nodes could be considered when selecting the next relay, which would lead to wrong decisions with the consequent waste of resources.

The work presented in [13] introduces a BA sender-oriented protocol, whose aim is to group its neighbor vehicles in clusters formed through the periodic exchange of hello messages. Then, a message is disseminated cluster-to-cluster through the formed transient clustering infrastructure. The proposed cluster formation algorithm and the next relay selection criterion only consider the vehicles mobility. Thus neither of them considered the impairments of the wireless channel in its design, assuming ideal channel conditions. In highway scenarios such assumption can turn into packet losses and/or delays, because it may be difficult to form clusters or an excessive number of clusters may be formed (consisting of a single node), depending on the particular channel conditions at any given moment. Additionally, selecting the next relay only based on vehicles mobility, without observing channel conditions, may also lead to significant packet losses in the presence of a highway V2V channel.

BR-NB and FUZZBR are two sender-oriented MB protocols proposed in [6, 19], respectively. In these two works neighbors are ranked using a fuzzy inference system based on three parameters, namely, vehicle mobility, intervehicle distance, and link quality. Then, the node with the highest rank among its neighbors in the message propagation direction is selected as the next relay. The link quality between two neighbors is estimated in [6] using the hello reception ratio (HRR). A fixed frequency of hello messages is assumed when computing the link quality. Thus for low HRR situations the BR-NB protocol could lead to either transmissions of hello messages with unnecessary high frequency in dense scenarios or to information losses in more dynamic scenarios. Furthermore, the HRR described in [6] is updated for each 10 seconds interval. Thus the reliability of the protocol is likely to be low in high mobility scenarios. For instance, in a highway scenario, vehicles traveling in opposite directions with a typical highway speed (e.g., 32 m/s) can easily go out of range from each other within the assumed 10 seconds period. Additionally, BR-NB needs the 2-hop neighborhood information to estimate both the intervehicular distance and the mobility of vehicles. However, using 2-hop neighborhood information in scenarios with high number of neighbors could cause an exponential growth of the overhead generated by hello packets.

The FUZZBR protocol [19] uses position information of neighbors contained in the hello messages to estimate both intervehicle distances and mobility of neighbors. Because of the predefined mobility patterns in VANETS, using position information is a suitable feature for highway scenarios, as most of the times the same set of vehicles could be used to forward information [9]. Additionally, the FUZZBR protocol described in [19] computes the link quality between two nodes based on the received signal strength indicator

(RSSI). Furthermore, the calculation of the vehicle mobility, intervehicle distance, and link quality parameters in FUZZBR is highly dependent on the frequency of hello messages. A consequence of this dependence is that the trade-off between overhead and relevance of the information cannot be entirely addressed. Despite this weakness, FUZZBR is able to adapt to different scenario conditions and overcome several of the problems shown by the MB protocols previously mentioned.

Most of the BA sender-oriented protocols described earlier consider a point-to-point approach for the evaluation of its neighbors. That is, a node establishes the “relay node” suitability of each neighbor using the hello messages broadcasted by the neighbor itself. Hence, in this approach the accuracy of the stored information and the quality of the relay node selection are directly proportional to the hello messages rate (HMR). In conclusion, this approach makes the protocol performance highly dependent on the HMR.

Including the wireless channel impairments in the relay node selection is a challenging and critical issue, as acknowledged by FUZZBR and BR-NB protocols. This issue is studied as well in [24], where the most forward within adjusted radius (MFWAR) mechanism is proposed to adjust the NRR of vehicles in highway scenarios (however note that MFWAR was designed as a discovery service for a unicast routing protocol and cannot be directly compared with a broadcast protocol). Thus, when the point-to-point approach is used for link quality computations in highway scenarios, the needed HMR should be enough to address the time varying nature of the channel. Setting the HMR too high can lead to a larger number of collisions. In contrast, setting a relatively low HMR can lead to not having enough information to perform a proper decision. Hence, with a point-to-point approach BA sender-oriented protocols may achieve a good performance in terms of the existing trade-off between redundant broadcasts and reliability. However, a relevant issue is that the existing trade-off between overhead and the relevance of the information was not addressed.

Based on the previous discussion, it can be stated that a significant issue that must be addressed when designing a new MB protocol for VANETs in highway scenarios is the design of a reliable algorithm for selecting the set of relay nodes while reducing the dependence of the MB protocols from the HMR. As explained in Section 3, the protocol introduced in this paper addresses these design constraints.

### 3. Proposed RLMB Protocol

This paper introduces a new reliable low-overhead multihop broadcast (RLMB) protocol. The development of RLMB considered the inherent constraints of the wireless channels and the high dynamics of vehicles in VANETs for highway scenarios.

RLMB is a BA sender-oriented protocol, thus a subset of relay nodes is selected a priori among the neighbors of the current relay. Then, the selected subset of relay nodes is attached to the message header before the current relay retransmits the message. When receiving a message each node will retransmit it if its own ID is in the header; otherwise the message is dropped. Additionally, each node periodically

broadcasts its transmission power, geographic position, and speed in the hello messages.

When selecting the relay set, RLMB proposes the use of a dynamic factor ( $\beta_{rd}$ ) to adjust the nominal radio range (NRR) according to the vehicles relative direction (rd). Then, the farthest neighbor in each message direction within the adjusted radio range is included in the relay set. The main objective of the  $\beta_{rd}$  factor is to adjust the NRR to the larger distance where a stable communication can be achieved, this considering the current conditions of the deployment scenario. For this purpose, RLMB uses the signal strength attenuation of the received *hello* messages in each rd and the receiver-transmitter distance for  $\beta_{rd}$  computation.

It is expected that while the transmitter-receiver distance increases the signal strength attenuation increases as well. However, depending on the scenario channel conditions (e.g., shadowing, fading or interference), a larger or smaller signal attenuation than the expected may occur. Thus, an adaptive  $\beta_{rd}$  computation is needed in dynamic scenarios. As such, RLMB uses a smart and adaptive fuzzy inference system for the radio range adjustment.

RLMB also uses a position prediction (PP) algorithm. The goal of this algorithm is to fill the gap between the last values stored in the neighborhood table and the most up-to-date values. Additionally, an implicit acknowledgment (iACK) mechanism is coupled to RLMB in order increase its reliability. The iACK mechanism is used when the current relay does not overhear the retransmitted message from one or more of its neighbors in the relay set.

The different mechanisms of the RLMB protocol are explained in the following subsections.

**3.1.  $\beta_{rd}$  Radio Adjustment Calculation.** The coverage range of each radio transceiver depends on radio channel impairments, which in turn are affected by specific characteristics such as density of nodes, weather, and buildings. Therefore, it can be inferred that considering the transmitter NRR coverage for the selection process may not be the best option, since the channel impairments may be the cause of significant packet losses near the border of the NRR. As such, the RLMB protocol proposes the use of the dynamic scaling factor  $\beta_{rd}$ . This factor is used as an estimation of the influence of the conditions of the radio channel and it is calculated based on the received hello packets from neighbors.

Unlike similar BA sender-oriented proposals, in RLMB the assessment is not made for each node. Instead, the link evaluation is made for a particular zone in a cooperative way. As it will be shown in Section 4, this approach allows RLMB making a proper adjustment of the NRR while maintaining a low overhead. Furthermore, according to [25], the relative direction between vehicles has a great impact in the propagation of the signal. Hence, an adjustment factor  $\beta_{rd}$  of this kind needs to consider the relative direction of vehicles. Thus, RLMB calculates the  $\beta_m$ ,  $\beta_c$ , and  $\beta_a$  adjustment factors, which consider the three possible relative directions taken by two vehicles, as illustrated in Figure 2.

In RLMB the NRR adjustment is performed considering the received power and distance between neighbors. These two aspects are introduced in the RLMB protocol by means of

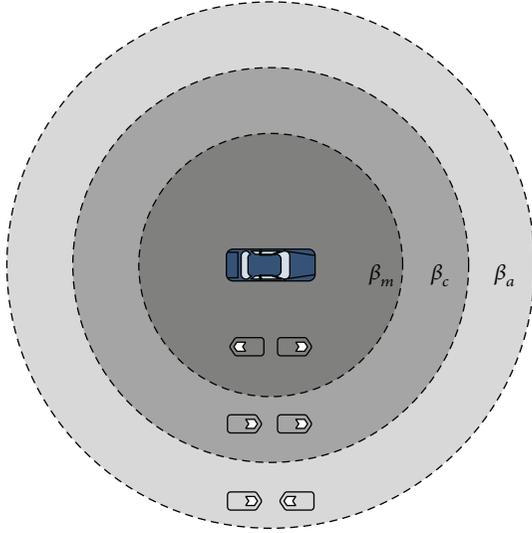


FIGURE 2: Use of dynamic  $\beta_m, \beta_c$ , and  $\beta_a$  factors to determine the maximum allowed transmission range according to the relative direction of vehicles.

two factors, namely, the  $\gamma$  and  $\delta$  factors, which are described next.

The  $\gamma$  factor is defined as

$$\gamma = 1 - \frac{PL_r}{PL_{max}}, \quad (1)$$

where  $PL_{max}$  is the maximum allowed linear path loss and  $PL_r$  is the current linear path loss. This factor is a measure of how far is the received power from the reception threshold, that is, the minimum acceptable power to detect and decode a packet. A received power close to the reception threshold indicates a poor link quality and  $\gamma$  will tend to 0. Conversely, a value far from the reception threshold indicates a better link quality and  $\gamma$  will tend to 1.

Note that because of the wireless channel phenomena (e.g., propagation loss and shadowing) a poor link quality is expected for distant nodes. Similarly, a good link quality is expected for the nearby nodes. Thus, if only the  $\gamma$  factor is taken into account for the radio adjustment, the  $\beta_{rd}$  factor would only oscillate between low and high values. In order to address this issue, calculating the  $\gamma$  factor considering only the distant nodes seems to be a suitable option at first glance. However, ignoring hello messages from the middle nodes can reduce the efficiency of the protocol because of the reduced availability of information.

To improve the radio adjustment accuracy while maintaining a low overhead, the transmitter-receiver distance is considered in the calculation of the  $\beta_{rd}$  factor by means of the  $\delta$  factor:

$$\delta = \begin{cases} \frac{D_{(x,y)}}{R}, & \text{if } D_{(x,y)} \leq R \\ 1, & \text{otherwise,} \end{cases} \quad (2)$$

where  $D_{(x,y)}$  is the distance between node  $x$  and node  $y$  and  $R$  is the NRR. This factor is a measure of the distance between neighbors and it is proportional to the NRR.

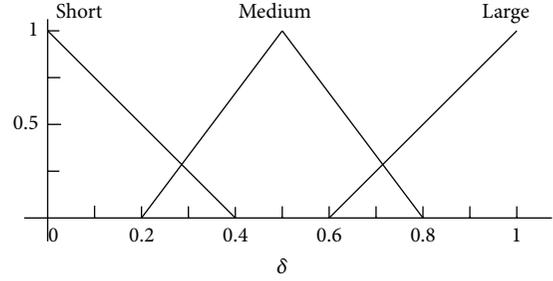


FIGURE 3: Fuzzy membership function for input  $\delta$  factor.

The value of  $\beta_{rd}$  must vary according to the channel variations dictated by propagation conditions, so that it can effectively represent the dynamics of vehicular networks. Furthermore, the information available to calculate  $\beta_{rd}$  may be imprecise because of different factors like a noisy GPS measure or a wrong power measurement at the PHY layer when receiving a hello message. Hence, using a closed expression to calculate the radio adjustment factor  $\beta_{rd}$  could restrict the protocol accuracy to very specific scenario/channel conditions.

In order to provide RLMB with an adaptable decision making mechanism capable of achieving a satisfactory performance, a smart and adaptive solution for the NRR adjustment based on fuzzy logic [26] is implemented in RLMB. Fuzzy logic has been used in the context of VANETs [19, 26] to make intelligent and adaptive decisions. However, to the best of the authors' knowledge, the fuzzy logic based algorithm presented in this paper is the first to consider the inherent constraints of highway scenarios to dynamically adjust the NRR with a low dependence of the HMR.

### 3.1.1. Design of the Decision Making System Using Fuzzy Logic.

When receiving a hello message at PHY layer, each node calculates the  $\gamma$  factor with expression (1). This factor is included in a header of the packet before sending it towards the network layer. In the network layer the  $\delta$  factor is calculated from the received information in the hello packet. Then, both  $\gamma$  and  $\delta$  are used as inputs of the proposed fuzzy inference system (FIS) in order to calculate the dynamic adjustment factor  $\beta_{rd}$ .

The proposed decision-making algorithm contains two phases.

*Phase I.* The initial set up phase when the FIS is formed from two sets of data:

- (a) the construction of the fuzzy memberships based on the individual linguistic values of distances, losses, and adjustments as shown in Figures 3, 4, and 5;
- (b) the if/then rules shown in Table 1. These rules specify the actual combination of these values to properly map the fuzzy values of the distance and losses, to the radio range adjustment fuzzy values.

*Phase II.* The decision-making algorithm (see Figure 6), which is invoked at each hello message arrival and it is fed

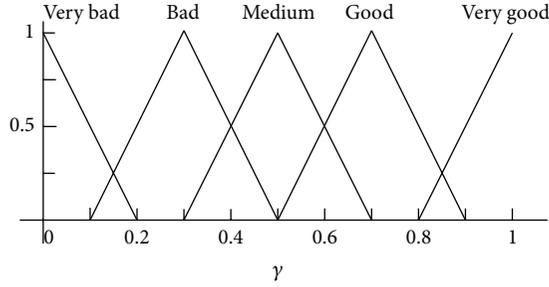


FIGURE 4: Fuzzy membership function for input  $\gamma$  factor.

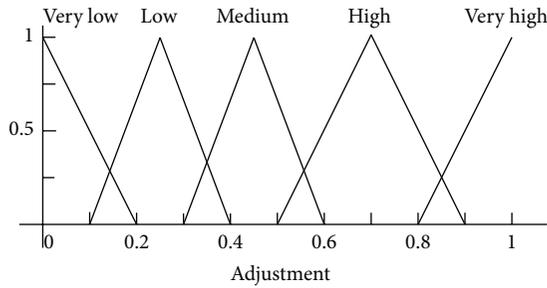


FIGURE 5: Fuzzy membership function for output variable.

with the values of distance and losses calculated from the hello message. The decision-making algorithm executes the following main steps:

- (a) mapping the values of the  $\gamma$  and  $\delta$  factors to a fuzzy value of  $\beta_{rd}$  through the FIS;
- (b) using a defuzzification method in order to map from fuzzy value to a crisp value.

To cope with highly transient values at the output of the FIS due to shorter than expected path losses, a moving average is used to smooth such undesired behaviors. The moving average is defined by

$$\beta_{rdn} = \beta_{rdi} + \epsilon * (\beta_{rdcurrent} - \beta_{rdi}), \quad (3)$$

where  $\beta_{rdn}$  is the new radio range adjustment factor in the rd direction;  $\beta_{rdcurrent}$  is the value of the current  $\beta$  in the rd direction; and  $\beta_{rdi}$  is the inferred value of radio range adjustment in the rd direction. The value of  $\epsilon$  defines the weight of the accumulated evaluation. Because of the high dynamic nature of the VANETs, a higher weight for the current value is desired. Thus, in this paper a value of 0.4 is chosen for  $\epsilon$  in expression (3).

**3.2. Position Prediction Algorithm.** When a vehicle needs to send or relay a packet, the position prediction (PP) algorithm must fill the gap between the last values stored in the neighborhood table and the most updated values. Thus, when the set of relay nodes has to be selected, the PP algorithm is invoked before performing the selection. The PP algorithm implemented in RLMB works as follows.

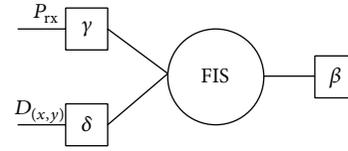


FIGURE 6: The decision-making system designed to update the adjustment factors.

- (1) The position of all neighbors is updated in the neighborhood table by means of

$$P_e = P_c + \left( \hat{v}_i * dt + \frac{acc_i}{2} * dt^2 \right), \quad (4)$$

where  $\hat{v}_i$  is the last stored velocity vector of the neighbor  $i$ ;  $acc_i$  is the last stored acceleration vector of the neighbor  $i$ ; and  $dt$  is the information dwell time of the neighbor  $i$  in the one-hop neighbors table.

- (2) The number of entries in the neighborhood table is updated. If the estimated distance for one particular neighbor is larger than the NRR, then this particular neighbor is deleted from the one-hop neighborhood table.
- (3) The updated neighborhood table is passed to the relay set selection algorithm.

**3.3. Relay Set Selection Method.** Two classes of nodes are considered in the RLMB protocol: source nodes and relay nodes. The source nodes are vehicles with data to send (e.g., a warning message), while relay nodes are vehicles within the ZOR that must relay the original packet. If a node (source or relay) wants to disseminate a packet, before broadcasting the packet, a set of relay nodes is selected as follows.

- (1) Using the respective  $\beta$  factor, a threshold for the maximum allowed distance is set for each relative direction rd with expression

$$Th_{rd} = \beta_{rd} * R, \quad (5)$$

where  $Th_{rd}$  is the threshold for the relative direction rd;  $\beta_{rd}$  is the adjustment radio range factor for the rd direction; and  $R$  is the NRR.

- (2) Neighbors whose distance is below the corresponding threshold,  $Th_{rd}$ , are grouped in three different sets based on their relative direction to the current relay. Specifically, vehicles traveling in opposite directions and moving away belong to group  $V_m$ ; vehicles traveling in opposite directions and approaching belong to group  $V_a$ ; and vehicles traveling in the same direction belong to group  $V_c$ .
- (3) Finally the set of relay nodes is defined considering the following.

- (i) If the current node is a relay node, the next relay node is the farthest node among the vehicles in the  $V_m$ ,  $V_a$ , and  $V_c$  groups in the message propagation direction.

TABLE 1: Knowledge structure based on fuzzy rules.

If $\delta$	&	If $\gamma$	Then Adjustment
Short		Very bad	Very low
Short		Bad	Very low
Short		Medium	Medium
Short		Good	High
Short		Very good	High
Medium		Very bad	Low
Medium		Bad	Low
Medium		Medium	Medium
Medium		Good	High
Medium		Very good	Very high
Large		Very bad	Low
Large		Bad	Medium
Large		Medium	High
Large		Good	Very high
Large		Very good	Very high

- (ii) If the current node is the source node or if it is located at an intersection, a set of relay nodes must be selected. This set is integrated by the farthest node among the vehicles in the  $V_m$ ,  $V_a$ , and  $V_c$  groups in each direction of the ZOR.

**3.4. Implicit Acknowledgment Mechanism.** As previously mentioned, the aim of the RLMB selection mechanism is to reduce the drop of broadcast packets by considering the constraints imposed by the radio channel and the dynamics of vehicles in highway scenarios. Nevertheless, sometimes the packets may not reach the intended primary relay node because of larger than expected path losses or inaccurate position predictions made by the PP algorithm. Therefore, in order to increase its reliability, the RLMB algorithm implements a basic implicit acknowledgment mechanism. If the current relay does not overhear the sent packet after an iACK time, then the relay retransmits the original packet. Here, the iACK is determined for each node with expression

$$\text{iACK} = \tau + X_{a,b}, \quad (6)$$

where  $\tau$  is a constant value and  $X_{a,b}$  is a uniform distributed random variable between  $a$  and  $b$ .

## 4. Performance Evaluation

In this section, the performance of the proposed RLMB protocol is evaluated. The V2V highway scenario was simulated using the OPNET Modeler simulator [27]. The simulated scenario considered two lanes for cars travelling in one direction and other two lanes for cars travelling in the opposite direction. The length of each lane was set to 3 km and the width to 4 m. In the simulation when a vehicle reaches the end of the road, it is reinserted in the lane with vehicles traveling in the opposite direction. The maximum allowed speed was set to 40 m/s. The radio channel propagation

model introduced in [25] for V2V highway scenarios was considered in the simulation setup. The mobility pattern for each vehicle was generated following the intelligent driver model introduced in [28]. This is a popular model used to generate mobility patterns for highway scenarios (e.g., [29, 30]). Additionally, free flow, medium, high, and jam vehicles densities [31] were also considered to evaluate its effects in the achieved performance of the MB protocols.

The FUZZBR and the proposed RLMB protocols were evaluated using the previously described scenario. The FUZZBR protocol was used for benchmarking purposes as it is the leading BA sender-oriented protocol in the literature. In fact, in [19] several performance metrics such as delay and packet dissemination ratio were provided for FUZZBR, showing that in a well-connected network this protocol outperforms EMPR and two well-known beaconless MB protocols. FUZZBR was implemented in the simulation testbed following the guidelines provided in [19]. In the case of the RLMB protocol, Figure 7 details the flowchart considered for its implementation.

The performance evaluation was realized in terms of the following performance metrics.

(a) *Packet Delivery Ratio (PDR)*. This metric measures the ratio between the number of vehicles that receive a broadcast packet and the number of vehicles in the ZOR times the packets transmitted by the source. We only consider the first received broadcast to determine the PDR.

(b) *Average End-to-End Delay (EED)*. The EED is calculated as the average delay from the source to each vehicle within a particular segment of the highway. For EED calculation, the highway length was divided in equally spaced segments of NRR width called milestones.

(c) *Retransmissions*. This metric is measured in terms of the number of retransmissions per packets made by each protocol during the simulation.

The aforementioned metrics were obtained considering a crash scenario where a source transmits a warning message with a data rate of 10 Kbytes/s from one edge of the ZOR. A specific propagation distance or a specific time can define the ZOR. For this paper, the entire highway was considered as the ZOR and all vehicles within it were considered as intended recipients. At the beginning of every simulation trial, every transmitter was set to wait for 20 s before starting any data transmission. This 20 s period was set to better allow the exchange of hello messages. In order to reach a stable state, a minimum of 100 trials were performed for each vehicle density,  $\lambda$ . The values of each variable used in the simulation are presented in Table 2.

The PHY/MAC layers of the IEEE 802.11p standard were implemented in our simulation, taking as a starting point the IEEE 802.11a project available in OPNET Modeler. The necessary adaptations were performed so that all PHY/MAC settings and parameters correspond to those found in the IEEE 802.11p standard. This approach was previously used in [24, 32] for the evaluation of AODV and DSR routing protocols in VANETs equipped with IEEE 802.11p

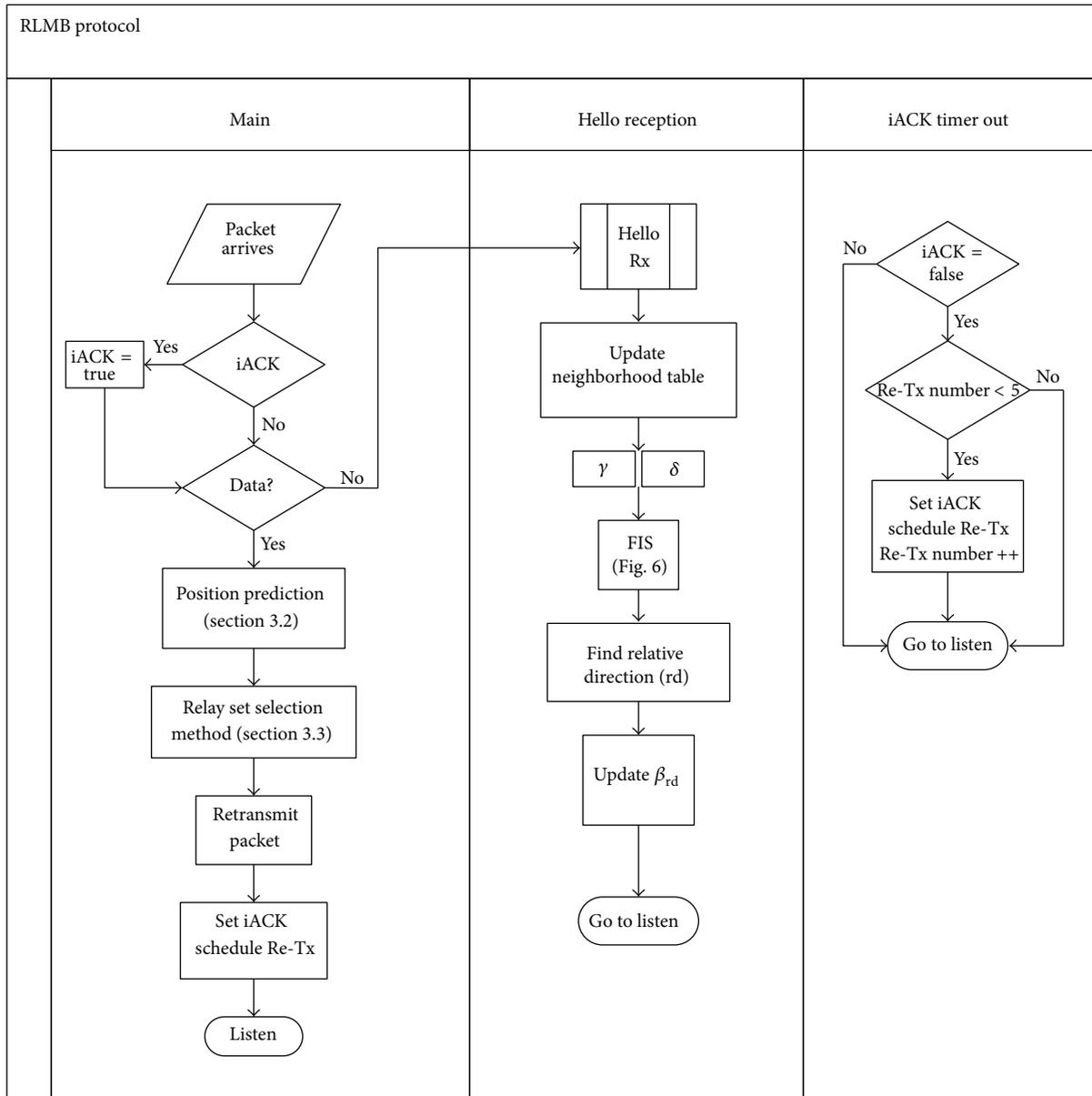


FIGURE 7: Flowchart of the main processes of RLMB broadcast protocol.

transceivers. Furthermore, the IEEE 802.11p DCF parameters corresponding to the best effort traffic over service channels (see [33]) were used to modify the IEEE 802.11a OPNET model. Therefore, the minimum and maximum contention window sizes and the time slot length were adjusted to IEEE 802.11p best effort traffic values. Similarly, the DIFS value was replaced by the corresponding AIFS value. Regarding the PHY layer adaptation, the bandwidth and operating frequency of the IEEE 802.11a OPNET model were adjusted to 10 MHz and 5.880 GHz, respectively, as defined by the IEEE 802.11p standard. For the rest of this paper the modified model will be referred to as adapted IEEE 802.11a/p model.

**4.1. Numerical Results.** Figure 8 presents graphically the PDR that each protocol is able to achieve versus the vehicles

density,  $\lambda$ , in the scenario. For this plot a fixed rate of 1 hello messages per second (hps) was set for FUZZBR, while a rate of 0.25 hps was set for RLMB. As shown in this figure, FUZZBR exhibits a poor performance compared to RLMB, especially under higher vehicles densities. This is caused by the larger overhead introduced by FUZZBR when using a rate of 1 hps. Thus, when evaluating FUZZBR in higher vehicles densities, the probability that the retransmission does not reach the intended relay node is increased because of the increment in the number of collisions. In contrast, RLMB provides a stable PDR regardless of the  $\lambda$  value. In fact, a drop of less than 1% is shown for the RLMB protocol in higher densities. RLMB exhibits this behavior because the zone-based link quality assessment depends on a much lower degree on the hello messages rate (HMR). Thus, a lower

TABLE 2: Values used for each variable in the simulation scenario.

Scenario parameter	
Maximum velocity	40 m/s
Vehicles density, $\lambda$ ,	[33, 66, 100, 133] v/km
Highway length	3 km
Packet size	1 Kbyte
Base frequency	5.880 GHz
Data rate	6 Mbps
Transmission range	250 m
FIS parameters	
Combination method	Min-max
Defuzzification Method	Centroid of area

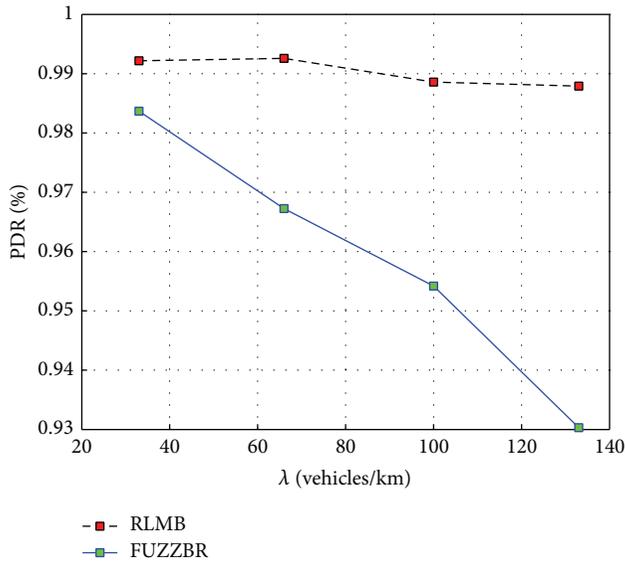


FIGURE 8: Packet delivery ratio (PDR) obtained with RLMB (0.25 hps) and FUZZBR (1 hps) when increasing the vehicles density,  $\lambda$ .

overhead is introduced with the consequent decrement in the number of collisions.

Figure 9 shows the PDR achieved when the rate of hello messages per second was set to 0.25 hps for FUZZBR. Thus, FUZZBR and RLMB had the same HMR for this plot. It can be seen in Figure 9 that using a low HMR in FUZZBR improves its performance for high vehicles densities. In high vehicles densities the dynamic of vehicles is lower. Thus, reducing the HMR along with the low dynamic of vehicles enables a performance improvement for FUZZBR in these cases (even though there are a major number of vehicles sharing the channel). However, the suitability evaluation for the selection of the relay nodes in FUZZBR is closely related to the HMR. Thus, FUZZBR has an important performance drop when the dynamic of vehicles increases, as observed in Figure 9 for the first two values of  $\lambda$ .

Despite the FUZZBR improvement observed in Figure 9 for high vehicles densities, it can be readily seen that RLMB still provided a better PDR than FUZZBR for all densities. Therefore, RLMB exhibited a better PDR when compared to

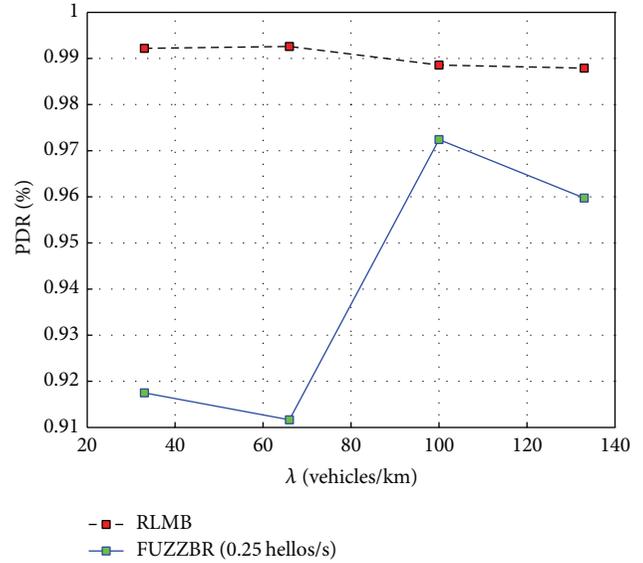


FIGURE 9: Packet delivery ratio (PDR) obtained with RLMB (0.25 hps) and FUZZBR (0.25 hps) when increasing the vehicles density,  $\lambda$ .

the one achieved by the benchmark protocol for both low and high HMRs.

Figure 10 illustrates the effects of increasing the vehicles density on the average number of retransmissions. Both protocols need a higher number of retransmissions when  $\lambda$  increases. The reason for this is because with the increment in the number of vehicles per km, the number of nodes sharing the channel increases. Thus, there is an increment in the probability of collision, which may cause that a packet does not reach the next relay or that the implicit acknowledgment does not reach the previous relay. However, because of the lower HMR, RLMB introduces less overhead. Thus, the retransmissions needed for the different densities are always lower than those needed in FUZZBR. Additionally, note that even when the HMR of FUZZBR is reduced, RLMB provides a better performance. This is the effect of the proposed PP algorithm, which handles the dynamic of vehicles efficiently.

Figure 11 graphically presents the effects of increasing the source-destination distance for different vehicle densities on the EED. Note in this figure that the packets disseminated by RLMB reach the end of the ZOR in a lower time for the different values of  $\lambda$ . This is the result of two aspects: (1) RLMB requires a lower number of retransmissions; and (2) the introduced overhead related to the hello messages is also lower in RLMB than in FUZZBR.

When disseminating the packet along the ZOR, fewer delays are introduced by RLMB compared to FUZZBR. Figure 12 shows a comparison of the EED obtained for RLMB with the EED obtained for FUZZBR with a HMR that matches the one used by RLMB. Under these conditions, FUZZBR achieves better performance for EED when compared to those obtained for FUZZBR with high HMR. However, the evaluation of the vehicles mobility performed by FUZZBR is closely related to the HMR. Thus, a low HMR

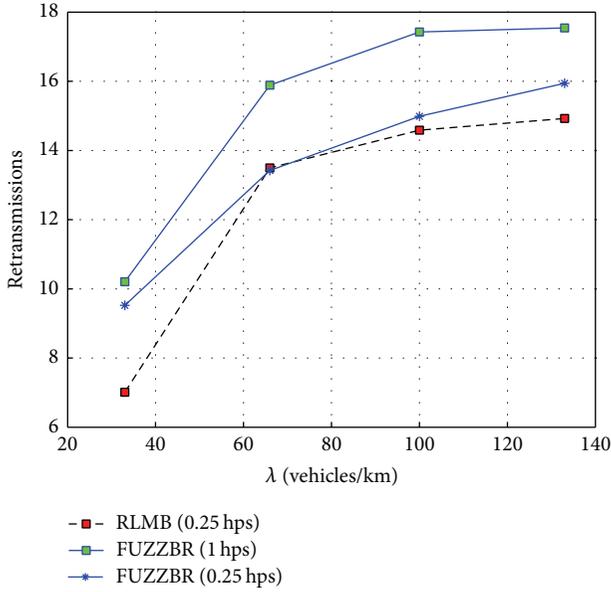


FIGURE 10: Number of retransmissions obtained for RLMB with 0.25 hps compared with the retransmissions obtained for two HMR (0.25 hps, 1 hps) of FUZZBR when increasing the vehicles density,  $\lambda$ .

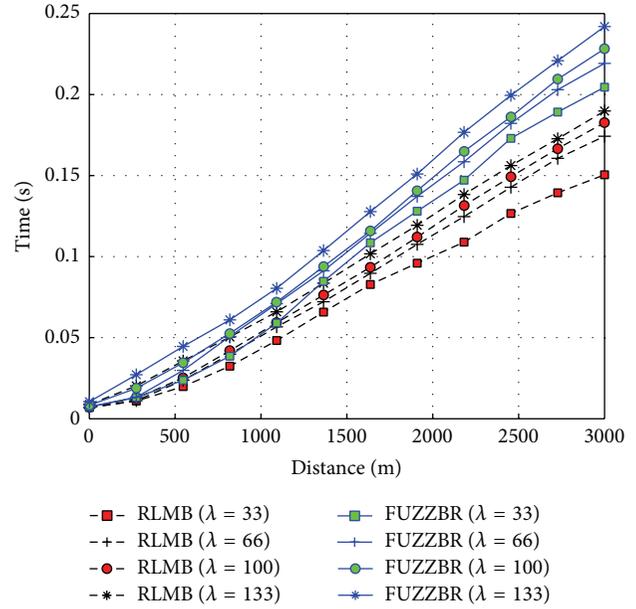


FIGURE 12: EED obtained with RLMB (0.25 hps) and FUZZBR (0.25 hps) when increasing the transmitter-receiver distance for different vehicles densities,  $\lambda$ .

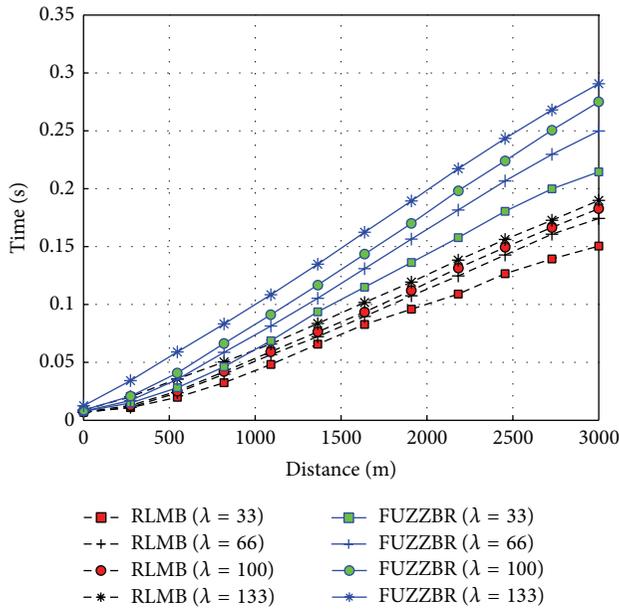


FIGURE 11: EED obtained with RLMB (0.25 hps) and FUZZBR (1 hps) when increasing the transmitter-receiver distance for different vehicles densities,  $\lambda$ .

causes a higher number of errors in the next relay selection method of FUZZBR. Hence, a higher number of retransmissions are needed in comparison with RLMB. Therefore, RLMB outperforms FUZZBR in both scenarios as it can be seen in Figure 12.

### 5. Conclusions and Future Work

In this paper a new low-overhead and reliable multihop broadcast protocol for VANETs called RLMB is introduced. The RLMB protocol proposes a point-to-zone computation of the link quality instead to a point-to-point assessment. The design of RLMB was aimed to address the two trade-offs arising in BA multihop broadcast protocols deployed in highway scenarios: the trade-off between reliability and redundant broadcast and the trade-off between reliability and the number of hello messages needed. As such, the performance of RLMB was compared with the performance shown by the FUZZBR protocol, which is one of the few broadcast protocols that consider both the dynamic of vehicles and the radio channel in the selection of the set of relay nodes. These contribution findings indicate that for low vehicle densities in highway scenarios, RLMB provides a higher PDR with a lower number of retransmissions compared to the results obtained when using the leading protocol FUZZBR. This is achieved by means of the PP algorithm altogether with the proposed radio adjustment of RLMB. Furthermore, the average EED is lower as well in RLMB than in FUZZBR for this case. In the case of high vehicles density scenarios in highways (which is likely to be the most relevant crash scenario), the proposed RLMB protocol also outperforms FUZZBR. The reason for this is because the point-to-zone link quality assessment reduces the dependence of RLMB on the HMR. Therefore, because of the reduced HMR and the lower number of retransmissions needed, RLMB introduces less overhead than FUZZBR. This feature of RLMB also helps in reducing the number of collisions. Additionally, the PP algorithm and the adaptive  $\beta_{rd}$  factor help to perform better relay set selections in highway scenarios. This can be drawn

from the fact that a lower number of retransmissions are generated when using RLMB compared to FUZZBR.

With the results presented in this paper, it has been shown that, with a point-to-zone link quality assessment, the dependence of the broadcast protocol from the HMR is reduced. Therefore, an improved performance for different vehicle densities can be achieved. In that sense, it can be stated that in a well-connected vehicular network RLMB efficiently handles the reliability/overhead trade-off. Furthermore, RLMB outperforms the leading BA sender-oriented protocol for multi-hop broadcast in connected highway scenarios.

In this paper a fixed HMR is used for RLMB. Thus, future work includes developing an adaptive HMR mechanism based on the current network conditions in order to make RLMB suitable for urban scenarios. Additionally, the implemented FIS was designed based on the overall performance in different network scenarios. The RLMB protocol performance can be improved if the FIS can be adapted to specific scenario conditions. Therefore, future work will include optimizing the FIS for different deployment scenarios such as a disconnected VANET.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work was funded by the Council for Science and Technology of Mexico (CONACyT) under the Project FORDE-CyT 2011-01-174509. A special acknowledgment is given to the OPNET University Program for their support.

## References

- [1] G. Najm Wassim, J. Koopmann, J. D. Smith, and J. Brewer, "Frequency of Target Crashes for IntelliDrive Safety Systems," U.S. Department of Transportation, DOT HS 811 381, 2010.
- [2] C.-J. Huang, Y.-W. Wang, H.-M. Chen et al., "An adaptive multimedia streaming dissemination system for vehicular networks," *Applied Soft Computing*, vol. 13, no. 12, pp. 4508–4518, 2013.
- [3] C. Rezende, A. Mammeri, A. Boukerche, and A. A. F. Loureiro, "A receiver-based video dissemination solution for vehicular networks with content transmissions decoupled from relay node selection," *Ad Hoc Networks*, vol. 17, pp. 1–17, 2014.
- [4] J. Lloret, A. Canovas, A. Catalá, and M. Garcia, "Group-based protocol and mobility model for VANETs to offer internet access," *Journal of Network and Computer Applications*, vol. 36, no. 3, pp. 1027–1038, 2013.
- [5] T. W. Chim, S. M. Yiu, L. C. K. Hui, and V. O. K. Li, "VANET-based secure taxi service," *Ad Hoc Networks*, vol. 11, no. 8, pp. 2381–2390, 2013.
- [6] C. Wu, S. Ohzahata, and T. Kato, "Practical solution for broadcasting in VANETs using neighbor information," *IEICE-Transactions on Communications*, vol. E96-B, no. 11, pp. 2856–2864, 2013.
- [7] N. Wisitpongphan, O. K. Tonguz, J. S. Parikh, P. Mudalige, F. Bai, and V. Sadekar, "Broadcast storm mitigation techniques in vehicular ad hoc networks," *IEEE Wireless Communications*, vol. 14, no. 6, pp. 84–94, 2007.
- [8] R. S. Schwartz, A. E. Ohazulike, C. Sommer, H. Scholten, F. Dressler, and P. Havinga, "Fair and adaptive data dissemination for Traffic Information Systems," in *Proceedings of the IEEE Vehicular Networking Conference (VNC '12)*, pp. 1–8, November 2012.
- [9] A. Fonseca and T. Vazão, "Applicability of position-based routing for VANET in highways and urban environment," *Journal of Network and Computer Applications*, vol. 36, no. 3, pp. 961–973, 2013.
- [10] R. Bauza and J. Gozalvez, "Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications," *Journal of Network and Computer Applications*, vol. 36, no. 5, pp. 1295–1307, 2013.
- [11] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis, "Highway vehicular delay tolerant networks: information propagation speed properties," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1743–1756, 2012.
- [12] R. S. Schwartz, H. Scholten, and P. Havinga, "A scalable data dissemination protocol for both highway and urban vehicular environments," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, article 257, 2013.
- [13] Y.-C. Chu and N.-F. Huang, "An efficient traffic information forwarding solution for vehicle safety communications on highways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 631–643, 2012.
- [14] "IEEE Standard for information technology—local and metropolitan area networks—specific requirements—part 11: wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications amendment 6: wireless access in vehicular environments," IEEE Std 80211p-2010 Amend. IEEE Std 80211-2007 Amend. IEEE Std 80211k-2008 IEEE Std 80211r-2008 IEEE Std 80211y-2008 IEEE Std 80211n-2009 IEEE Std 80211w-2009, pp. 1–51, July 2010.
- [15] M. Amadeo, C. Campolo, and A. Molinaro, "Enhancing IEEE 802.11p/WAVE to provide infotainment applications in VANETs," *Ad Hoc Networks*, vol. 10, no. 2, pp. 253–269, 2012.
- [16] S. Panichpapiboon and W. Pattara-Atikom, "A review of information dissemination protocols for vehicular ad hoc networks," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 3, pp. 784–798, 2012.
- [17] M. Bakhouya, J. Gaber, and P. Lorenz, "An adaptive approach for information dissemination in Vehicular Ad hoc Networks," *Journal of Network and Computer Applications*, vol. 34, no. 6, pp. 1971–1978, 2011.
- [18] C. Liu and C. Chigan, "RPB-MD: providing robust message dissemination for vehicular ad hoc networks," *Ad Hoc Networks*, vol. 10, no. 3, pp. 497–511, 2012.
- [19] C. Wu, S. Ohzahata, and T. Kato, "VANET broadcast protocol based on fuzzy logic and lightweight retransmission mechanism," *IEICE Transactions on Communications*, vol. E95-B, no. 2, pp. 415–425, 2012.
- [20] O. K. Tonguz, N. Wisitpongphan, and F. Bai, "DV-CAST: a distributed vehicular broadcast protocol for vehicular ad hoc networks," *IEEE Wireless Communications*, vol. 17, no. 2, pp. 47–57, 2010.
- [21] R. S. Schwartz, R. R. R. Barbosa, N. Meratnia, G. Heijenk, and H. Scholten, "A directional data dissemination protocol for vehicular environments," *Computer Communications*, vol. 34, no. 17, pp. 2057–2071, 2011.
- [22] C. Wu, K. Kumekawa, and T. Kato, "A novel multi-hop broadcast protocol for vehicular safety applications," *Journal of Information Processing*, vol. 18, pp. 110–124, 2010.

- [23] J. Sahoo, E. H. K. Wu, P. K. Sahu, and M. Gerla, "BPAB: binary partition assisted emergency broadcast protocol for vehicular ad hoc networks," in *Proceedings of the 18th International Conference on Computer Communications and Networks (ICCCN '09)*, pp. 1–6, San Francisco, Calif, USA, August 2009.
- [24] G. A. Galaviz-Mosqueda, R. Aquino-Santos, S. Villarreal-Reyes, R. Rivera-Rodríguez, L. Villaseñor-González, and A. Edwards, "Reliable freestanding position-based routing in highway scenarios," *Sensors*, vol. 12, no. 11, pp. 14262–14291, 2012.
- [25] J. Karedal, N. Czink, A. Paier, F. Tufvesson, and A. F. Molisch, "Path loss modeling for vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 1, pp. 323–328, 2011.
- [26] K. Z. Ghafoor, K. Abu Bakar, M. van Eenennaam, R. H. Khokhar, and A. J. Gonzalez, "A fuzzy logic approach to beaconing for vehicular ad hoc networks," *Telecommunication Systems*, vol. 52, no. 1, pp. 139–149, 2013.
- [27] "Network Planning and Configuration — Network Simulation (OPNET Modeler Suite) — Riverbed," 2014, [http://www.riverbed.com/products/performance-management-control/network-performancemanagement/network-simulation.html#Modeler\\_University\\_Program](http://www.riverbed.com/products/performance-management-control/network-performancemanagement/network-simulation.html#Modeler_University_Program).
- [28] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, vol. 62, no. 2, pp. 1805–1824, 2000.
- [29] F. Dressler, C. Sommer, D. Eckhoff, and O. K. Tonguz, "Toward realistic simulation of intervehicle communication," *IEEE Vehicular Technology Magazine*, vol. 6, no. 3, pp. 43–51, 2011.
- [30] K. Ibrahim and M. C. Weigle, ASH: Application-aware SWANS with highway mobility.
- [31] H. Wang, D. Ni, Q.-Y. Chen, and J. Li, "Stochastic modeling of the equilibrium speed-density relationship," *Journal of Advanced Transportation*, vol. 47, no. 1, pp. 126–150, 2013.
- [32] M. A. Iqbal, F. Wang, X. Xu, S. M. Eljack, and A. H. Mohammad, "Reactive routing evaluation using modified 802.11a with realistic vehicular mobility," *Annales des Telecommunications*, vol. 66, no. 11-12, pp. 643–656, 2011.
- [33] J. Mišić, G. Badawy, and V. B. Mišić, "Performance characterization for IEEE 802.11p network with single channel devices," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1775–1787, 2011.

## Research Article

# IDMA-Based MAC Protocol for Satellite Networks with Consideration on Channel Quality

Gongliang Liu, Xinrui Fang, and Wenjing Kang

*School of Information and Electrical Engineering, Harbin Institute of Technology, No. 2 West Wenhua Road, Weihai 264209, China*

Correspondence should be addressed to Gongliang Liu; [liugl@hit.edu.cn](mailto:liugl@hit.edu.cn)

Received 8 April 2014; Accepted 21 June 2014; Published 13 July 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Gongliang Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to overcome the shortcomings of existing medium access control (MAC) protocols based on TDMA or CDMA in satellite networks, interleave division multiple access (IDMA) technique is introduced into satellite communication networks. Therefore, a novel wide-band IDMA MAC protocol based on channel quality is proposed in this paper, consisting of a dynamic power allocation algorithm, a rate adaptation algorithm, and a call admission control (CAC) scheme. Firstly, the power allocation algorithm combining the technique of IDMA SINR-evolution and channel quality prediction is developed to guarantee high power efficiency even in terrible channel conditions. Secondly, the effective rate adaptation algorithm, based on accurate channel information per timeslot and by the means of rate degradation, can be realized. What is more, based on channel quality prediction, the CAC scheme, combining the new power allocation algorithm, rate scheduling, and buffering strategies together, is proposed for the emerging IDMA systems, which can support a variety of traffic types, and offering quality of service (QoS) requirements corresponding to different priority levels. Simulation results show that the new wide-band IDMA MAC protocol can make accurate estimation of available resource considering the effect of multiuser detection (MUD) and QoS requirements of multimedia traffic, leading to low outage probability as well as high overall system throughput.

## 1. Introduction

Compared to a narrow-band system, a wide-band system [1] can support much higher traffic rates and provide satisfactory multimedia services. Broadband satellite network is a typical system with limited bandwidth and power. How to effectively utilize the precious communication resources while ensuring the requirements for quality of all kinds of services at the same time is thus an urgent problem to be solved.

Multiple access control (MAC), which provides a mechanism of sharing the satellite resource efficiently, plays a vital role in enhancing the utilization of radio resource. In order to effectively optimize the allocation of resources and the onboard power, a new medium access control (MAC) techniques should be proposed. Besides, the core of the MAC technology is multiple access schemes, which can coordinate users sharing the limited resources to achieve efficient and reliable transmission.

In most of the existing broadband satellite communication systems, it is feasible to use frequency division multiple

access (FDMA) or time division multiple access (TDMA). However, there exist some technical bottlenecks, particularly in the frequency reuse aspect and system capacity.

The system of code-division multiple access (CDMA) is attractive for its outstanding capacity, frequency-spectrum utilization, and reliability, while tremendous computational cost on multiuser detection (MUD) to eliminate multiple access interference (MAI) is needed in this system.

Interleave-division multiple access (IDMA), derived from CDMA [2], provides a new solution for multiple access in satellite networks. As in the latest proposed multiple access scheme, the key thought of IDMA is to use different interleavers to distinguish multiple users, in which way users can transmit their information simultaneously. Consequently, the QoS is satisfied without the utilization of complicated slot management or packet scheduling which are both necessary in TDMA or CDMA systems, leading to the reduction in the complexity of onboard queuing and switching. Furthermore, IDMA adopts the iterative chip-by-chip (CBC) detection scheme to combat both intercell and intracell multiple access

interference (MAI). Compared to CDMA, IDMA solves the problem of MAI at a lower computational complexity, the decrement of which is linear with the number of users [2].

The studies in [3] forecast and modify the ground mobile environment with the time-series model such as glide average, exponential smoothness, and linearity regress. Considering that the satellite channel is different from wireless channels on the ground, channel prediction model of ground system cannot completely be suitable for satellite channel. Meanwhile, due to the long-time delay and nonstationary of satellite channel, the predictable length of most prediction algorithm is limited by the relative time of the input sequence. In this case, prediction data for controlling the rate adaptation and power allocation strategy is out of date, and therefore it is impossible to achieve real-time resource allocation. In [4], channel quality can be obtained directly by physical information such as signal-to-noise ratio (SNR), received signal strength, and bit error rate (BER). However, this method is rather complex and time consuming especially under bad channel condition. In this paper, ARIMI model based on smoothing processing and force feedback is proposed, achieving a relatively accurate prediction with the simple implementation structure.

In this paper, we introduce a new MAC protocol for wide-band IDMA satellite communication systems. Firstly, we develop a new IDMA-based power allocation algorithm combined with the technique of SINR evolution and channel prediction to provide good quality of service to as many mobiles as possible, even for the users under poor channel conditions. In [5], a study on a minimum-power allocation for multimedia traffic focuses on minimizing the received interference in wide-band CDMA networks. However, the optimum power allocation cannot be obtained due to ignoring the efficiency of MUD. A novel power allocation algorithm [6] combining the technique of SINR evolution and load balance can be accurately estimated when channel is assumed perfect, leading to low outage probability. Whereas the transmission environment may change sharply over time, the allocation of resources is inaccurate and unequal in the current. In a word, the new power allocation considered here is jointly based on the SINR evolution and variable channel quality, which is different from that in [6].

By taking the satellite channel prediction and adaptive transmission technique into account, the rate adaptation of multimedia can be achieved dynamically according to the satellite feedback control. And in order to improve system performance and capacity further, a new admission control mechanism based on channel quality and rate adaptation is proposed for IDMA system considering the effect of MUD. Rate adaptation algorithm ensures dynamic resource allocation by considering the physical and MAC layer jointly, which optimizes the system performance. In addition, considering the sufferable delay of handoff and new calls, the strategy of buffer queue for new calls [7] is also introduced to improve the blocking probability. Although most of the previous call admission control schemes combine the rate adaptation strategy and buffer queue strategy together to improve the end-to-end performance, very few of them consider the validity for buffering the new calls due to strict

delay requirement and fairness of different users. What is more, the effect of multiuser detection on the performance of CAC algorithm is also evaluated in this paper.

The rest of this paper is organized as follows. In Section 2, the overall MAC protocol is described. In Section 3, the satellite channel prediction is illustrated. The dynamic power allocation algorithm is derived for wide-band IDMA system in Section 4. Based on the result of channel prediction, the multimedia wide-band IDMA rate scheduling scheme is developed in Section 5, and the effective CAC scheme is derived in Section 6. In Section 7, performance of the new MAC protocol is evaluated through simulation. Finally, the conclusion of this paper is drawn in Section 8.

## 2. New Mac Protocol

*2.1. System Model.* The wide-band IDMA system has the following features [8].

- (i) No complex frame synchronization process: as different users occupy different interleavers in the process of transmission, the IDMA system supports asynchronous transfer mode (ATM) and does not need complex frame synchronization process. Thus, for each user without complex transmission scheduling, their QoS requirements can be satisfied easily.
- (ii) IDMA-CBC MUD and SINR evolution technique: considering the effects of the detector signal-to-noise ratio on MUD efficiency, resource allocation on board can be estimated accurately. Thus, based on the semi-analytical SINR evolution technique, MUD efficiency is considered here as the percentage of the intracell interference cancelled by the multiuser detector, and the efficiency fully reflects the performance advantages of IDMA system.
- (iii) Interleave-division S-ALOHA access mechanism: compared to the traditional S-ALOHA, the interleave-division S-ALOHA (IDSA) access mechanism can improve the efficiency of the random access and shorten the access delay effectively. Considering that different interleavers are used to distinguish the access requests of multiple users in IDMA, we can balance the allocated access interleavers with traffic load interleavers to make a tradeoff between system performance and onboard processing complexity.
- (iv) Variable-rate transmission and variable processing gain: the IDMA system saves the bandwidth which is occupied for spread coding in the conventional CDMA system for channel coding. Thus, the coding gain maximization can be achieved by low bit-rate coding. For the IDMA system, the processing gain is determined by two factors: FEC coding gain and spreading gain. Considering that the greater coding gain can improve the BER performance of IDMA system in the physical layer, the demanding rate can be well adjusted by the FEC coding rate to meet the needs of various users. In that case, this paper uses a variable processing gain IDMA (VSG IDMA) model.

In Figure 1, the schematic shows the MAC protocol of IDMA system. The uplink physical channel is divided into a random access channel (RACH) to send requests from a mobile terminal to the satellite repeater and a traffic channel (TCH) to send feedback of resource allocation from satellite repeater to mobile terminals by different interleavers. Ground station, via the RACH, sends requests to the onboard processor which makes decisions according to the current system load, service level, channel quality, and so on. Meanwhile, onboard processor needs to allocate channel resources (interleaver) and suitable power to the ground station.

Access request needs to include the following information: the coding type, the supportable maximum transmitted power, service classes, QoS requirements in bit error rate, bandwidth, and so forth.

*2.2. Procedures of the MAC Protocol.* The mechanisms of the new wide-band IDMA MAC protocol based on channel quality in wireless multimedia networks are shown in Figure 1. Compared with other multiple access systems, IDMA may provide users with flexible multirate and multiQoS support by controlling coding rate and power. Thus, it integrates the factors of service types, service quality, coding scheme, and permissible power factor in the existing MAC protocols. What is more, an optimized resource allocation proposal integrated with link quality prediction, power control, and call admission control can ensure dynamic resource distribution and change adaptively according to the environmental variation.

As shown in Figure 2, QoS guarantee mechanism includes the following main functions [9].

- (i) Channel quality prediction: the dynamic estimation of satellite channel provides the foundation and prerequisite for efficient resource allocation, particularly for resources such as bandwidth and power. Meanwhile, because of the long-transmission delay for satellites communication, the communication channel between the satellites and the mobile terminal constrains the performance of the whole system. Hence, timely measurement and accurate prediction on satellite channel are rather necessary to ensure effective resource allocation and adjustment.
- (ii) Power allocation algorithm: the performance of power control is the key to the quality of call admission control. The novel power allocation algorithm proposed combining the technique of SINR evolution and channel quality estimation can realize the optimized allocation of resources for each user and adjust the transmitted power adaptively according to the wireless environment. The algorithm not only illustrates the high efficiency of CBC MUD but also provides reliable communication for a growing number of users, especially when suffering the poor channel quality problem.
- (iii) Rate scheduling: in the transmission state, user sends a transmission request in RACH before it transmits a batch of packets. The wireless channel measurement

results can be divided into superior channel quality and inferior channel quality. Users with inferior channel quality perform rating downgrades to maximize the throughput of the system. Rate scheduling strategies guarantee the priority of different services, meeting the requirement of service with maintainable bit rates.

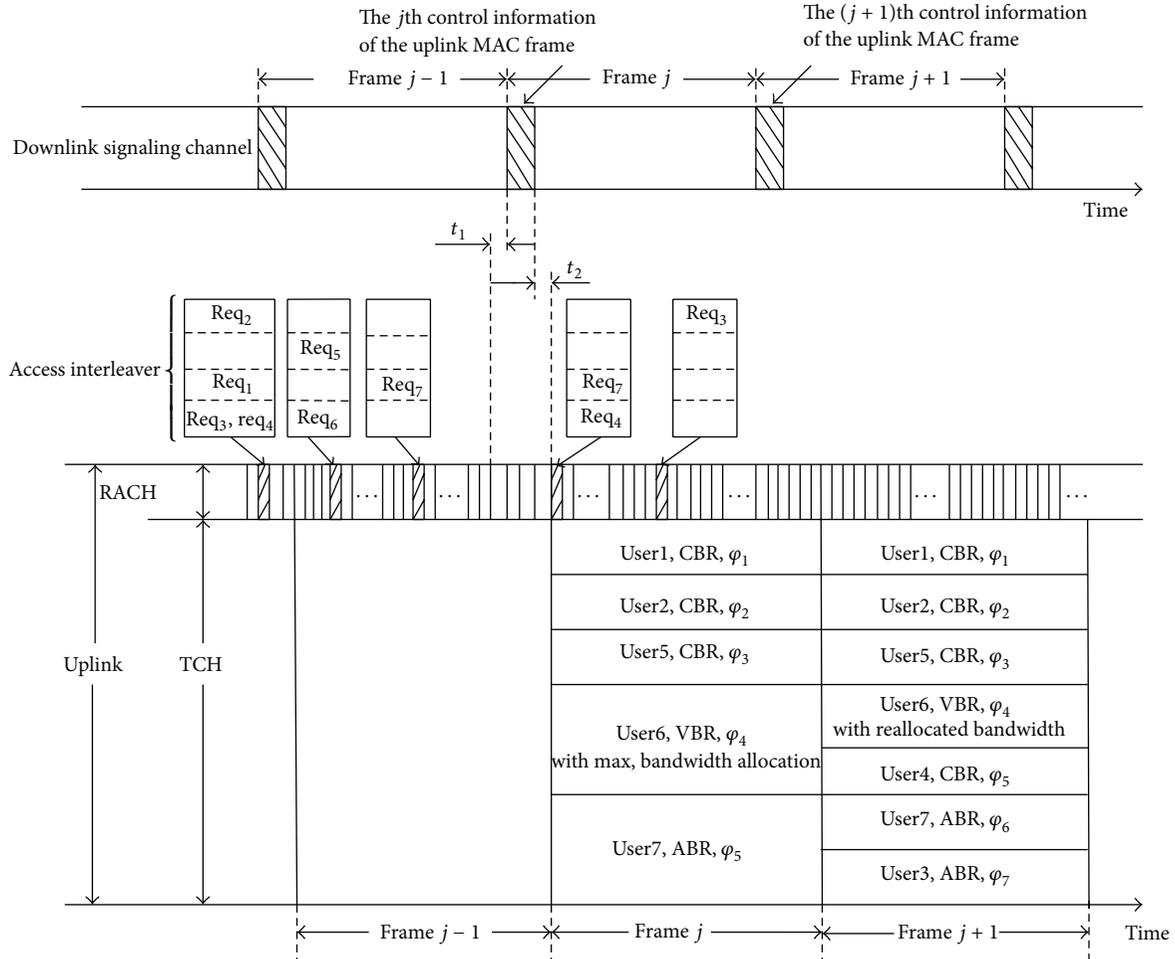
- (iv) Call admission control: when communication is initiated, a mobile terminal sends an admission request in RACH with a randomly selected interleaver. When the satellite receives such a request, a new effective interference-based CAC algorithm is invoked to check if enough resources are available in the system. The result is sent back to the mobile terminal through BCH. If the answer is positive, the request is accepted and the interleaver is reserved for the mobile terminal. And then, the mobile terminal enters the transmission state. Otherwise, considering that handover calls own a higher priority than new calls, call admission policy will change with different types of calls. On one hand, because new calls are insensitive for time delay, the caching strategy of exponential backoff can efficiently decrease blocking probability. On the other hand, as for handover calls, the rate for the accepted user can be demoted further on the premise of fulfilling the user's requirements. Meanwhile, the new CAC can make accurate estimation of available resource considering the effect of MUD, leading to low outage probability as well as low blocking and dropping probability.
- (v) Mapping between physical and MAC layer: a cross-layer resource allocation method for IDMA systems is proposed by considering jointly the physical and MAC layer, which optimizes the system performance. The adaptive resource allocation is not an isolated problem of access control or power allocation problem, but a global optimization one restrained by QoS and channel quality [10].

### 3. Satellite Channel Quality Prediction

With the properties of time-varying, multipath effect, shadow fading, and Doppler shift, satellite communication environment impacts the reliability of digital signal transmission and the effectiveness of onboard resource allocation. Thereby, the communication channel between the satellite and the mobile terminal constrains the performance of the whole system.

In order to effectively predict the change of the channel quality and assess the rationality of the resource allocation scheme, channel model should be established veritably to reflect the practical environment.

*3.1. Model of the Satellite Communication System.* Based on the transmission characteristics of satellite mobile communication channel, we present and simulate three common channel models of wide-band satellite mobile communication. The fading characteristics of the received signal, the level crossing rate (LCR), and average fade duration (AFD)



$t_1$ : propagation delay from the user to the satellite and response time of the satellite  
 $t_2$ : propagation delay from satellite to user and response time of user  
 $Req_k$ : the user- $k$  access request,  $k = 1, 2, 3, \dots$   
 $\varphi_k$ : traffic load interleaver

FIGURE 1: MAC protocol of IDMA.

for C. Loo, Corazza, and Lutz are simulated and analyzed. These important statistical properties drive the statistical regularities about fading speed and duration of the received signal, which provide efficiency and accuracy of theoretical support for the following research. The following provides a detailed example of Corazza model:

$$r(t) = [z(t) + d(t)] s(t) = R(t) s(t), \quad (1)$$

where  $s(t)$  is in line with random process of the lognormal distribution, which is regarded as large scale fading and  $z(t)$  and  $d(t)$  represent light-of-sight (LOS) and multipath component, respectively.

Various parameters of the Corazza model with a least squares fitting method are deduced in the literature [11]. The main parameters studied were the coefficients of mathematical expectation  $\mu$ , variance  $d_0$  of  $\ln s$ , and Rice factor  $K$  varied

with satellite elevation. The parameters of the literature [11] apply to any elevation ranging from 20 to 80.

The cumulative probability curve of the theoretical formula and the measured data were separately fitted at the speed of 60 km/h, in the countryside. With the satellite elevation equal to 20, the result is shown in Figure 3.

At the same time, the second order statistics which mean average level crossing rate as well as average fade duration of each composite fading channel are analyzed and, according to the literature [11], corresponding uniform expressions can be got. Simulation results on the fitting are shown in Figures 4 and 5.

Obviously, the channel influences the signals in two ways: the large scale fluctuation mainly caused by the shadow effect and the small scale fluctuation by multipath effect and Doppler shift. Indeed the occlusion caused by the terminal movement always remains a minor variation or constant

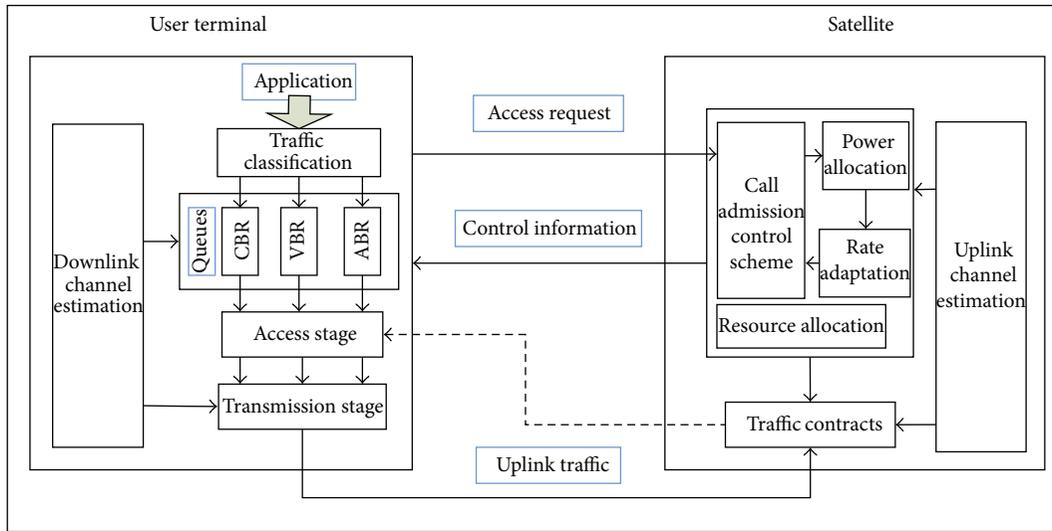


FIGURE 2: The framework of QoS guarantee mechanism.

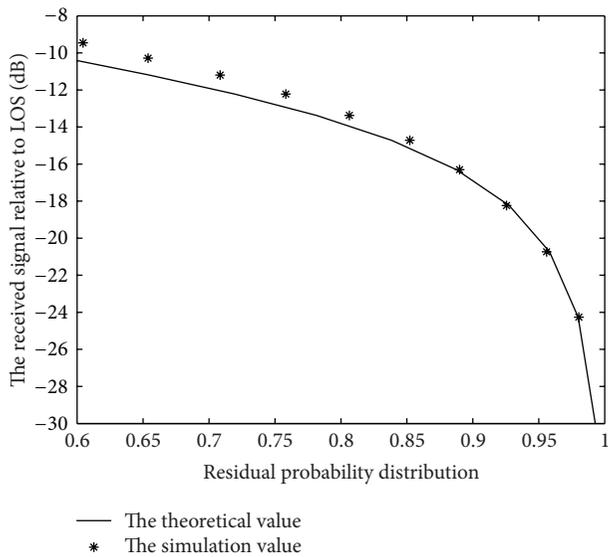


FIGURE 3: The cumulative probability curve of Corazza model.

over time, while rich multipath signal changes sharply with the mobile terminal. Therefore, the change of large scale attenuation rate is generally far below that of the small scale fading.

3.2. ARIMA Model for Satellite Channel Quality Prediction.

The comparisons of signal containing small scale fading and that without the small scale fading under severe environments proved that fluctuation speed of the actual signal is mainly dominated by small scale fading in the literature [12]. Moreover, regardless of the impact of small scale fading, signal wave speed will be greatly reduced. The actual measurement of large scale fading signal and fitting model has been given in the literature [13]. The calculation results show that channel quality forecast is feasible and practical when the

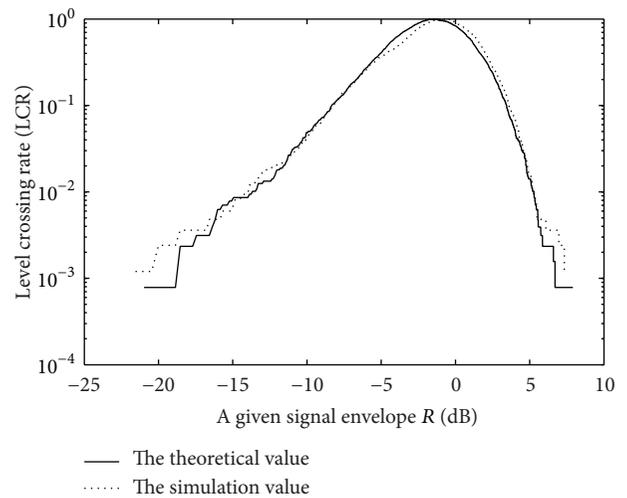


FIGURE 4: The level crossing rate of Corazza model.

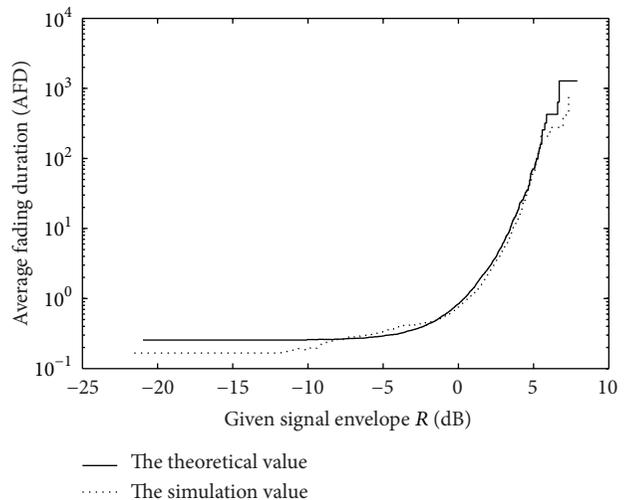


FIGURE 5: The average fades duration of Corazza model.

actual signal is replaced by the signal affected by large scale fading only.

Based on the above analysis, we choose ARIMA prediction channel to forecast the quality of satellite channel when removing the effect of small scale fading.

Supposing that TTI length is 1 ms, the actual satellite data received is a string sequence with 1 ms intervals. The SNR of downlink signal received by all mobile users can be able to directly indicate the channel quality. According to the literature related, the end-to-end delay from the ground station via satellite to the destination end (user), caused by double jump mode, is 540 ms. To reduce the prediction step length, it is feasible to adopt 100 TTI channel status to estimate a channel state, namely, 100 ms statistics for a state. Therefore, the prediction of channel quality by 540 ms requires just six steps of prediction, concluded in 500 ms to 600 ms time period.

As fluctuation values within a certain range were calculated as a single state, relative time of each state will be longer than that of the specific value. Then, reasonable resource allocation can be realized according to the predicted state of channel quality. Figure 6 illustrates the measurement and quantitative data of satellite channel quality and depicts the rapid change of the channel quality.

The prediction data of satellite channel quality with smoothing processing for many times can be seen from Figure 7. As is shown in Figure 7, the signal of eliminating the small scale fading effect is stable.

The autocorrelation for the prediction data of satellite channel quality with smoothing processing is shown in Figure 8. Because the correlations of signals are tending towards stability after operation many times, as can be seen, we can give prediction to the data by using the data correlation. Correlation time between states would be, in some sense, longer than that of the actual data. In other words, ARIMA model would obviously improve the precision of forecasting and increase the reliability of prediction with enhancement in temporal sequence correlation.

As the data window of high order model is fairly large, the volume of historical data which is used to predict trend is increasingly large. Under the condition of the small window, we can develop low order model by introducing data retroactivity to approximate high order model. Figure 9 illustrates the quality of satellite channel while utilizing the step length 90 to predict the step data length 96 shows a certain error, but it is sufficient for handling a pretty good precision.

#### 4. A Dynamic Power Control Algorithm

*4.1. IDMA-CBC MUD and SINR Evolution Technique.* Combining with the specific IDMA-CBC MUD, we consider the capacity analysis further, which effectively resists the internal MAI. Complete structure and procedure of IDMA-CBC MUD are given in the literature [2].

With single path channel being synchronous and with modulation sett as BPSK in this study, the performance of IDMA-CBC detection scheme is mainly reflected by the

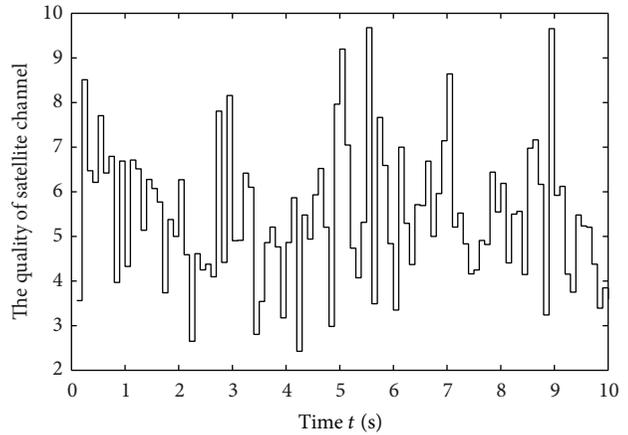


FIGURE 6: The state of channel quality.

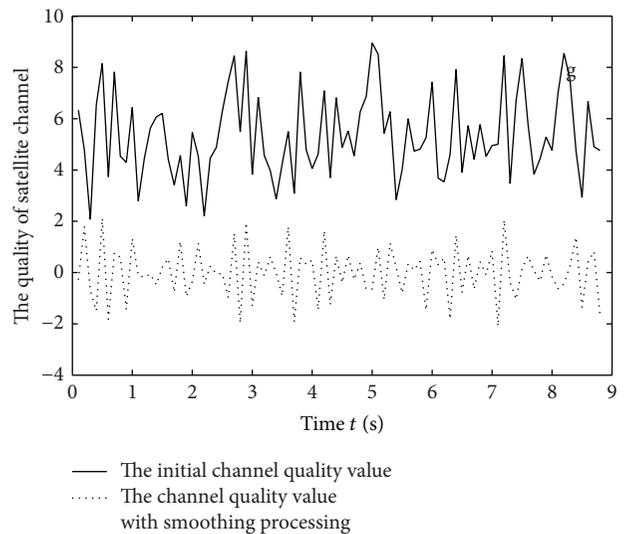


FIGURE 7: The prediction data of satellite channel quality with smoothing processing.

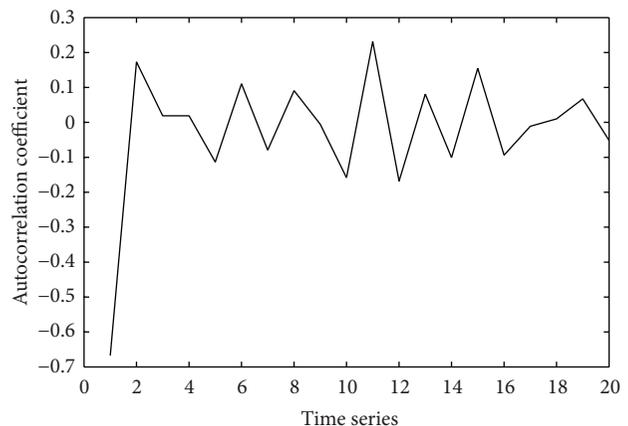


FIGURE 8: Autocorrelation analysis.

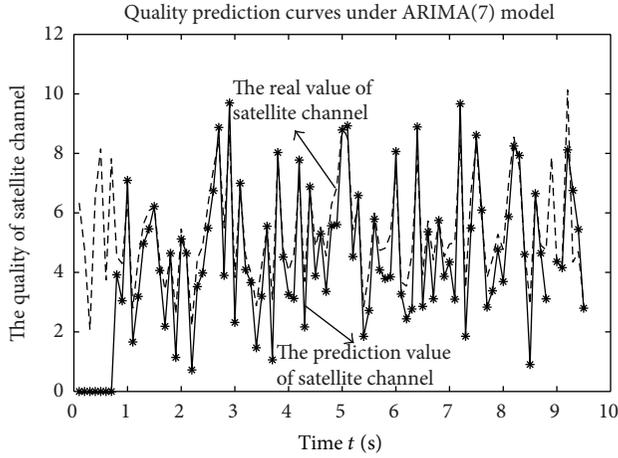


FIGURE 9: Comparison of channel quality sequence and data with smoothing processing.

decrease in the variable variance, for example, the variance of  $\{x_k(j), \forall k, j\}$ . It can be written as

$$V_k = 1 - \tanh^2\left(\frac{Y_{\text{SINR}_k}}{2}\right), \quad k = 1, \dots, K. \quad (2)$$

As shown in (2),  $V_k$ , for example, the variance of an arbitrary chip from user- $k$ , which is the corresponding power interference factor in the iteration, is the function of  $\text{SINR}_k$ . Besides an anti-interference percentage with fixed SINR, the function  $f(\text{SINR})$  is referred as the expectation of the interference power and written as

$$f(\text{SINR}_k) = E(V_k) = 1 - E\left[\tanh^2\left(\frac{Y_{\text{SINR}_k}}{2}\right)\right], \quad (3)$$

$$k = 1, \dots, K.$$

When the iteration reaches the iteration convergence point for user- $k$ , equivalently, the system achieves maximum MAI elimination capacity; thus we define

$$\text{SINR}_k = \frac{P_0}{\sum_{i \neq k} P_i E(V_i) + P_N} \geq \gamma_k, \quad (4)$$

where  $\gamma_k$  represents  $(E_b/I_0)_{\text{req}} \cdot (R_b/B)$ , and the total interference can be expressed as

$$I_{\text{total}} = \sum_{i \neq k} P_i f(\text{SINR}_i) + P_N, \quad (5)$$

where  $f(\text{SINR}_i)$  is negatively correlated with SINR, ranging from 0 to 1, which has been verified in the literature [2].

**4.2. Minimum-Power Allocation.** Based on the variable spreading gain (VSG) of the program, multibit transmission rate can be got by adjusting the variable spreading gain. Therefore, in order to satisfy the QoS requirements of all users, we need to adjust the transmitted power of each user to minimize the interference to other users' further.

Assume that there exist  $N_k$  users for service type  $k$  supported in the multicell IDMA system, in which each user adopts the same spreading code and coding rate to transmit information.  $W$  is the spread-spectrum bandwidth;  $R_i$  is the data rate of the user  $i$  determined only by the spread gain SG;  $P_i$  and  $h_i$  represent the transmitted power and the uplink channel gain of the user  $i$ , respectively.

Considering the home cell,  $(E_b/I_0)_k$  of the special user  $n_k$  can be written as

$$\left(\frac{E_b}{I_0}\right)_k = \frac{h_{n_k} P_{n_k}}{(I_{\text{total}} - h_{n_k} P_{n_k})} \cdot \frac{W}{R_{n_k}} \geq \gamma_k, \quad (6)$$

$$n_k = 1, 2, \dots, N_k.$$

The constraints that the transmitted power and the data rate must fulfill are as follows:

$$0 < P_{n_k} < p_i, \quad R_{n_k} > r_{n_k}, \quad n_k = 1, 2, \dots, N_k, \quad (7)$$

where  $s_i$  and  $r_i$  represent maximum permissible transmitted power and maintain data rate, respectively.

To adjust the transmission power of users, we add the general constraints to (6) and (7):

$$\text{Minimize } \sum_{n_k=1}^{N_k} P_{n_k}, \quad n_k = 1, 2, \dots, N \quad (8)$$

subject to  $0 < P_{n_k} < p_i, \quad R_{n_k} > r_{n_k}$ .

**Assumptions.**

- (1) QoS requirement of each user is equal to the target signal-to-noise ratio.
- (2) Data rate of each user is higher than the maintained data rate.

The optimal power value  $P_{n_k}^*$  obtained in the sense of the above assumptions is expressed as

$$\frac{h_{n_k} P_{n_k}^*}{(I_{\text{total}} - h_{n_k} P_{n_k}^*)} \cdot \frac{W}{R_{n_k}} = \gamma_k, \quad \forall n_k = 1, \dots, N_k. \quad (9)$$

The  $I_{\text{total}}$  in this case can be calculated as

$$I_{\text{total}} = h_{n_k} P_{n_k}^* \left(1 + \frac{W}{\gamma_k R_{n_k}}\right). \quad (10)$$

Considering the multicell IDMA system, the total interference power including the intracell interference from users  $I_{\text{intra}}$ , the received power from adjacent cells  $I_{\text{inter}}$ , and the thermal background noise  $P_N$  can be calculated as

$$I_{\text{total}} = I_{\text{intra}} + I_{\text{inter}} + P_N. \quad (11)$$

According to the interference calculation model, the other-beam interference factor  $f_{\text{other}}$ , which is defined as the ratio of the interference power received from the other beams

$I_{inter}$  to the interference power produced by users in local beam  $I_{intra}$ , can be calculated as

$$f_{other} = \frac{I_{inter}}{I_{intra}}. \quad (12)$$

As average interference is utilized, the other-beam interference factor  $f_{other}$  presented in previous studies [14] can be seen as a constant 0.55. Consequently, with (11) and (12), we can represent the  $I_{total}$  as

$$I_{total} = (1 + f_{other}) \cdot I_{intra} + P_N. \quad (13)$$

Based on the semianalytical SINR evolution technique, MUD efficiency is considered here as the percentage of the intracell interference cancelled by the multiuser detector. Thus, by (5), the intracell interference received by base station considering the effect of CBC MUD is

$$I_{total} = (1 + f_{other}) \cdot \sum_{k=1}^K \sum_{n_k=1}^{N_k} h_{n_k} P_{n_k}^* f(\gamma_k, G_{n_k}) + P_N. \quad (14)$$

Based on (10) and (14), we can derive the following expression:

$$\begin{aligned} & h_{n_k} P_{n_k}^* \left( 1 + \frac{W}{\gamma_k R_{n_k}} \right) \\ &= (1 + f_{other}) \cdot \sum_{k=1}^K \sum_{n_k=1}^{N_k} h_{n_k} P_{n_k}^* f(\gamma_k, G_{n_k}) + P_N \\ &= (1 + f_{other}) \cdot \sum_{k=1}^K \sum_{n_k=1}^{N_k} \frac{1}{1 + (W/\gamma_k R_{n_k})} f(\gamma_k, G_{n_k}) I_{total} \\ & \quad + P_N. \end{aligned} \quad (15)$$

So, we have

$$\begin{aligned} & h_{n_k} P_{n_k}^* \left( 1 + \frac{W}{\gamma_k R_{n_k}} \right) \\ & \quad \times \left( 1 - (1 + f_{other}) \cdot \sum_{k=1}^K \sum_{n_k=1}^{N_k} \frac{1}{1 + (W/\gamma_k R_{n_k})} f(\gamma_k, G_{n_k}) \right) \\ &= P_N. \end{aligned} \quad (16)$$

According to (16), we can evaluate the optimal power value  $P_{n_k}^*$

$$\begin{aligned} & P_{n_k}^* \\ &= (P_N) \left( h_{n_k} \left( 1 + \frac{W}{\gamma_k R_{n_k}} \right) \right) \end{aligned}$$

$$\begin{aligned} & \times \left( 1 - (1 + f_{other}) \right. \\ & \quad \left. \cdot \sum_{k=1}^K \sum_{n_k=1}^{N_k} \frac{1}{1 + (W/\gamma_k R_{n_k})} f(\gamma_k, G_{n_k}) \right)^{-1}. \end{aligned} \quad (17)$$

Since each user has power constraint, it is required that the received optimal power  $P_{n_k}^*$  of user  $n_k$  is less than  $p_{n_k}$ . Thus,

$$\begin{aligned} & \sum_{k=1}^K \sum_{n_k=1}^{N_k} \frac{1}{1 + (W/\gamma_k R_{n_k})} f(\gamma_k, G_{n_k}) \\ & \leq 1 - \frac{P_N}{h_{n_k} (1 + (W/\gamma_k R_{n_k})) (1 + f_{other}) p_{n_k}}, \end{aligned} \quad (18)$$

for all  $\forall n_k = 1, \dots, N_k$  and  $\forall k = 1, \dots, K$ .

Define  $\Delta = (P_N / (h_{n_k} (1 + (W/\gamma_k R_{n_k})) (1 + f_{other}) p_{n_k}))$ ; (18) becomes  $\sum_{k=1}^K \sum_{n_k=1}^{N_k} (1 / (1 + (W/\gamma_k R_{n_k}))) f(\gamma_k, G_{n_k}) \leq 1 - \Delta$ .

## 5. Rate Scheduling Strategy

An effective management of rate scheduling depending on the channel quality and the network load situation assigns rate for online users in the cell, and the main rate adjustment is degradation process. However, rate selection strategy is not specified for multiple rates supported in IDMA system.

In this case, a VSG- (variable spreading gain-) IDMA model [15] is selected for the following reasons.

- (1) VSG model is suitable for the IDMA system. A serious concern in CDMA systems is that the spreading gain is too small to maintain good cross correlation especially for users on condition that data rate is very high. While in IDMA systems, the spreading gain which can realize rate adaptation via the variation of the spreading sequence is irrelevant to distinguish users.
- (2) VSG model is suitable for satellite system. Rate adaptation in IDMA satellite systems can be simply taken as an adaptive spread gain strategy.

By taking the time-varying property of the satellite mobile communication channel and adaptive transmission technique into account, we can adjust the sending rate dynamically, according to the satellite feedback control and the ARIMI model, to improve the adaptive ability further.

The rate adaptation can be created by the following procedures. When an incoming call arrives to an overloaded cell, the degradation procedure is activated by reducing the service rate of online user in the terrible channel condition. The strategy can ensure different priority levels for different class calls to satisfy well the QoS requirements of various services.

Degradation procedure: the existing calls are degraded according to their priority. Calls of the highest priority get

the least consideration for the reduction procedure, whereas calls of low priority can be immediately degraded to a lower bit rate in order to maximize the system capacity. The system assigns the transmission rate, transmitted power, and the only interleaver of the request, to service in the order of voice traffic, video traffic, and background traffic at each frame.

When the resources become overloaded at some frames, background traffic which is pretty relaxed about time delay will firstly reduce transmitted power and the transmission rate of ongoing calls in terrible channel quality. Similarly, if the system is still overloaded, video traffic operates in the same way until degradation factor reaches its maximum value or the system can admit an incoming call. Meanwhile, due to severe restrictions on bit error rate for calls with low priority, the call can be buffered in the system for another access attempt. It not only decreases the blocking probability but also increases the utilization of resource.

In this paper, the basic transmission rate is 15 kb/s and the packets are allowed to transmit at seven rates {15, 30, 60, 120, 240, 480, 960 kb/s} [15]. Correspondingly, the spreading factor  $G$  ranges from 256 to 4 expressed as  $256/2^k$  ( $k = 0, 1, 2, 3, 4, 5, 6$ ). In order to analyze conveniently, we use the parameter  $w_{i,j}$  ( $w_{i,j} \in \{1, 2, \dots, 7\}$ ) to indicate the level of degradation in the transmission rate of call  $j$  in class  $i$  and it also means the rate adaptation will increase the degradation factor of class  $i$  by one in progress. Considering the fairness among users, it is unbearable to frequently downgrade rate for the ongoing calls. Thus, degradation factor of class  $i$  should be lower than allowable degradation limits  $w_i$  of class  $i$ . In other words, degradation factor of class  $i$  cannot be increased once again unless  $w_{i,j} < w_i$ .

The transmission rates for multimedia calls can be adjusted to accommodate more calls while satisfying the minimum signal-to-interference ratio (SIR) and transmission rate requirement. So the rate adaptation and power allocation mechanisms can work jointly to decide whether the user should be accepted and assigned with resources.

## 6. Call Admission Control Scheme Based on Channel Quality

In this part, we focus on the CAC algorithm for IDMA systems with various services. In [16], the distinct capacity bottleneck in uplink and the one in downlink are considered, respectively, while both uplink CAC based on interference and downlink CAC based on the base station transmitted power (UD-CAC) are implemented in [9] at the same time. Although the UD-CAC scheme achieves a nice tradeoff between capacity and stability of the system and maximizes the performance of the system, it does not take the effect of MUD into consideration. In [17], the scheme can make accurate estimation of available resource considering the effect of MUD, leading to low outage probability as well as low blocking and dropping probability. However, most of the existing CAC algorithms ignoring the influence of satellite channel cannot adaptively change with dynamic environment. On this basis, a multiservice call admission control strategy based on channel quality is proposed. The

proposed scheme which is conjunct with rate scheduling and buffering strategy can guarantee high power efficiency and throughput for multimedia traffic even in heavy load conditions, illustrating the high efficiency of CBC MUD. Especially when communication quality of users get worse, system can take into account the quality of service guarantees, interference and channel quality, and so forth to make judgments on whether it is reasonable and feasible to admit new calls and whether to adjust the ongoing calls to increase access opportunity.

The special concerns in designing the scheme are as follows:

- (i) the satellite resource allocation adaptively changes with dynamic environment and can further solve the global optimization problems;
- (ii) the proposed rate adaptation and buffering strategy achieve better performance;
- (iii) the traffic asymmetry and distinct capacity bottlenecks between uplink and downlink will be discussed in detail.

The CBC MUD scheme and SINR evolution technique for fast performance evaluation of IDMA are briefly introduced in Section 4. Here we extend this accurate and effective technique to the estimation of interference level. The proposed CAC scheme working in conjunction with rate scheduling and buffering strategy is explained in Figure 10. The transmission rates for multimedia calls can be adjusted to accommodate more calls according to different traffic priorities while meeting their minimum signal-to-interference ratio (SIR) and QoS requirement, only when congestion occurs. Also, the buffering strategy is introduced to hold the call which cannot be admitted at once to increase access opportunity according to different delay characteristics of the traffic.

**6.1. Estimation of Uplink Interference Level.** With (11) and (12), we can represent  $I_{\text{total}} = (1 + f_{\text{other}}) \cdot I_{\text{intra}} + P_N$ . Further, assume that the required transmitted power of active user  $n_k$  is  $S_{n_k}$ . Then,  $S_{n_k}$  is written as

$$S_{n_k} = h_{n_k} P_{n_k}. \quad (19)$$

According to (6), the received bit energy to interference power spectral density ratio for service type  $k$ ,  $(E_b/I_0)_k$  can be written as

$$\left(\frac{E_b}{I_0}\right)_k = \frac{S_{n_k}}{(I_{\text{total}} - S_{n_k})} \cdot \frac{W}{R_{n_k}} \geq \gamma_k. \quad (20)$$

In order to evaluate the effect of an active user to the system interference, we define the load factor of a single connection as

$$L_{n_k} = \frac{S_{n_k}}{I_{\text{total}}} = \frac{1}{\left(1 + \left(W/\gamma_k R_{n_k}\right)\right)}. \quad (21)$$

Based on the above formulas and the effect of CBC MUD, the total intracell interference power from users in home cell can be written as

$$I_{\text{intra}} = \sum_{k=1}^K \sum_{n_k=1}^{N_k} L_{n_k} \cdot I_{\text{total}} f(\gamma_k, G_{n_k}). \quad (22)$$

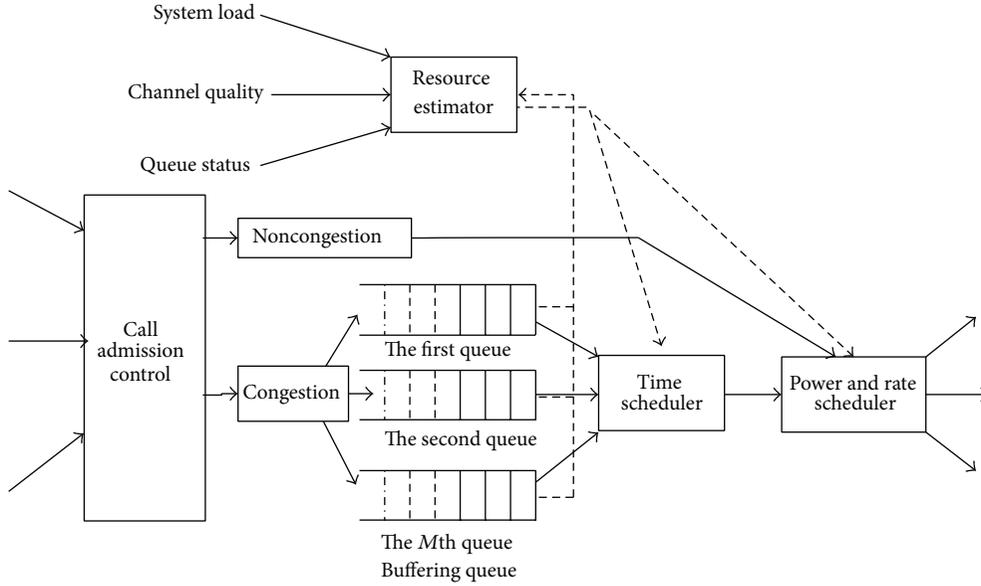


FIGURE 10: The IDMA MAC protocol based on channel quality.

Similarly, we define the fractional load factor in the home cell  $\eta$  as

$$\eta = (1 + f_{\text{other}}) \sum_{k=1}^K \sum_{n_k=1}^{N_k} L_{n_k} f(\gamma_k, G_{n_k}), \quad (23)$$

which is normally used as the home cell load indicator [14]. Based on (11) and (22), the total interference received in the home cell can be written as

$$I_{\text{total}} = \eta \cdot I_{\text{total}} + P_N = \frac{P_N}{1 - \eta}. \quad (24)$$

With the derivative form of (24), the uplink power increase of the total interference level due to a new requiring user- $n_k$  with CBC MUD in IDMA systems can be estimated as follows:

$$\Delta I = \frac{I_{\text{total}} f(\gamma_k, G_{n_k})}{1 - \eta - f(\gamma_k, G_{n_k}) \Delta L} \Delta L. \quad (25)$$

**6.2. Estimation of Downlink Transmitted Power Level.** Similar to the uplink, all users share the common bandwidth and each new connection increases the interference level of other connections, affecting the service quality expressed in terms of a certain  $(E_b/I_0)_i^d$ . For  $N$  users receiving signals simultaneously from a given cell, the received  $(E_b/I_0)_i^d$  can be written as

$$\left(\frac{E_b}{I_0}\right)_i^d = \frac{W}{R_i^d} \cdot \frac{g_{i0} P_i^N}{\sum_{j=1, j \neq i}^N \theta g_{i0} P_j^N + g_{i0} P_p + I_i + P_N} \geq \gamma_i^d, \quad (26)$$

$$P_{\text{total}_N} = \sum_{j=1}^N P_j^N + P_p,$$

where  $P_{\text{total}_N}$  represents the base station transmitted power,  $P_j^N$  ( $j = 1, 2, \dots, N$ ) is the power devoted to the  $j$ th user,

$I_i$  is the intercell interference observed by the  $i$ th user,  $g_{i0}$  is the path loss to the user  $i$ ,  $P_p$  is the power assigned to pilot channel, and  $\theta \in (0, 1]$  is the orthogonality factor in the downlink direction. The minimum transmitted power  $P_i^N$  satisfying the  $i$ th user demands can be expressed as

$$P_i^N = L_i^d \left( \theta \sum_{j=1}^N P_j^N + P_p + \frac{I_i + P_N}{g_{i0}} \right), \quad (27)$$

$$L_i^d = \frac{\gamma_i^d R_i^d}{W + \theta \gamma_i^d R_i^d}.$$

The latter expression is commonly known as the downlink load factor for the  $i$ th user.

Assume that the subscript number of a new call is 0; the total amount of the users in home cell is  $N + 1$  if it is accepted. Now the transmitted power to the  $i$ th user is

$$P_i^{N+1} = L_i^d \left( \theta \sum_{j=0}^N P_j^{N+1} + P_p + \frac{I_i + P_N}{g_{i0}} \right). \quad (28)$$

From the above, the increase in power demand to the  $i$ th user  $\Delta p_i$  is estimated as follows:

$$\Delta p_i = L_i^d \left( \theta \sum_{j=1}^N \Delta P_j + \theta P_0 \right), \quad (29)$$

and transmitted power to the 0th user can be written as

$$P_0 = L_0^d \left( \theta \sum_{j=1}^N P_j^N + P_p + \frac{I_0 + P_N}{g_{00}} \right) \frac{(1 - \theta \sum_{j=1}^N L_j^d)}{(1 - \theta \sum_{j=0}^N L_j^d)}. \quad (30)$$

After accumulating all the  $\Delta p_j$  ( $j = 1, \dots, N$ ), total transmitted power for  $N + 1$  users is

$$P_{\text{total}_N+1} = P_{\text{total}_N} + \sum_{j \neq 0} \Delta p_j + p_0$$

$$= \frac{L_0^d P_{\text{total}_N} (\theta + ((I_0 + P_N) / (g_{00} P_{\text{total}_N})))}{(1 - \theta \sum_{j=0}^N L_j^d)} + P_{\text{total}_N}, \quad (31)$$

where  $\sum_{j=0}^N L_j^d$  is the downlink fractional load factor.

Similar to the uplink,  $f_{\text{other}}^d$  is defined as the ratio of the total base station transmitted power from adjacent cells to intracells, and the power generated by base station in the home cell is set as its typical value which is 0.55 [18]. When the background noise is ignored, (31) can be written as

$$P_{\text{total}_N+1} = \frac{L_0^d P_{\text{total}_N} (\theta + f_{\text{other}}^d)}{(1 - \theta \sum_{j=0}^N L_j^d)} + P_{\text{total}_N}. \quad (32)$$

Considering that the downlink receivers in IDMA systems also benefit from CBC detection [10], the uncanceled percentage of intracell interference is  $f(\text{SINR})$ . Then the orthogonality factor in the downlink direction can be equivalent to  $f(\text{SINR})$ . With the aid of SINR evolution, the downlink load factor for the  $i$ th user and the total transmitted power can be accurately and easily estimated as

$$L_i^d = \frac{\gamma_i^d R_i^d}{W + f(\text{SINR}_i) \gamma_i^d R_i^d}, \quad (33)$$

$$P_{\text{total}_N+1} = \frac{L_0^d P_{\text{total}_N} (f(\text{SINR}_0) + f_{\text{other}}^d)}{(1 - \sum_{j=0}^N f(\text{SINR}_j) L_j^d)} + P_{\text{total}_N}.$$

**6.3. The Proposed Admission Control Algorithm.** According to UD-CAC scheme [9], determination conditions of uplink CAC based on interference and downlink CAC based on the base station transmitted power (UD-CAC) are  $I_{\text{total}_\text{old}} + \Delta I \leq I_{\text{THRESHOLD}}$  and  $P_{\text{total}_\text{old}} + \Delta P_{\text{total}} < P_{\text{threshold}}$ . Based on feedback interval of channel quality, predicted values of channel quality  $\gamma^i$  for the user  $i$  apart from those of voice traffic are compared with the target  $\text{SNR}_i$ . The proposed CAC scheme consists of seven stages, explained in Figure 11. Follow those steps according to the priority of different services.

**Stage 1.** Channel quality detection.

According to the service requirements (such as transmission rate and activating factor) of the call, determine whether or not  $\gamma^i > \gamma_{\text{req}}^i$ . If yes, no further operations are performed; if not, start the rate adaptation.

**Stage 2.** Rate adaptation algorithm.

Plus one for the user's downgrade factor  $w_{i,j}$  and update the current transmission rate and interference factor. Meanwhile, judge whether degradation factor of class  $i$  is lower than allowable degradation limits  $w_i$ . If not, do nothing; else, repeat *Stage 1*.

**Stage 3.** Resource estimation and forecast period.

For uplink, we estimate the total interference  $I_{\text{total}}$  for the current system and incremental interference  $\Delta I$  after accepting new users. And for downlink, the total transmitted power from the base station  $P_{\text{total}_\text{old}}$  and the expected increase in transmitted power  $\Delta P_{\text{total}}$  should be estimated.

When a handoff call arrives, it will be accepted if

$$I_{\text{total}_\text{old}} + \Delta I \leq I_{\text{threshold}},$$

$$P_{\text{total}_\text{old}} + \Delta P_{\text{total}} < P_{\text{threshold}} \quad (34)$$

is satisfied. Else, go to *Stage 4*. But for a new call, determine whether the buffering queue is empty. If not, go to *Stage 5*; else, decide whether to admit the call according to

$$I_{\text{total}_\text{old}} + \Delta I \leq I_{\text{THRESHOLD}},$$

$$P_{\text{total}_\text{old}} + \Delta P_{\text{total}} < P_{\text{threshold}}. \quad (35)$$

If the above equation is workable, the call is accepted. Otherwise, step to *Stage 6*.

**Stage 4.** Degradation procedure for handoff user.

In accordance with class priorities in ascending order, the ongoing calls with poor channel quality perform degradation process. When degradation factors of all ongoing users (excluding voice services) have reached the maximum and (34) still does not valid, then the handoff call will be refused.

**Stage 5.** Implementation of buffering strategy (see the part D).

**Stage 6.** Degradation procedure for new user.

Similarly, according to class priorities in ascending order, the ongoing calls with poor channel quality perform degradation process. But unlike the processes for handoff call, when degradable rate of every class has been exhausted and (35) still does not valid, then put the new call in cache queue to be detected when the backoff window value decreases to 0.

**Stage 7.** When the call is completed, release all resources and update the available capacity in the system at the same time.

**6.4. Buffering Strategy.** When a new call requests the access to the system, test whether the delay queue is empty and when the delay queue is nonempty, enforce buffering strategy. Due to the necessity for satisfying the different requirements of various services for delay time, time-delay counters for multimedia services designed with different threshold values should judge in real time whether the cumulative delay exceeds a threshold. If so, reject the call; if not, generate a random backoff period. The measurements for random delay time are divided discretely in timeslot  $\tau$ .

The initial settings for the maximum and minimum backoff window value are  $\text{BW}_{\text{max}}$  and  $\text{BW}_{\text{min}}$ , respectively. Specific steps are as follows.

**Stage 1.** Depending on the type of new call and the backoff window in the request packet, the backoff window is set as the minimum value  $\text{BW} = \text{BW}_{\text{min}}$ .

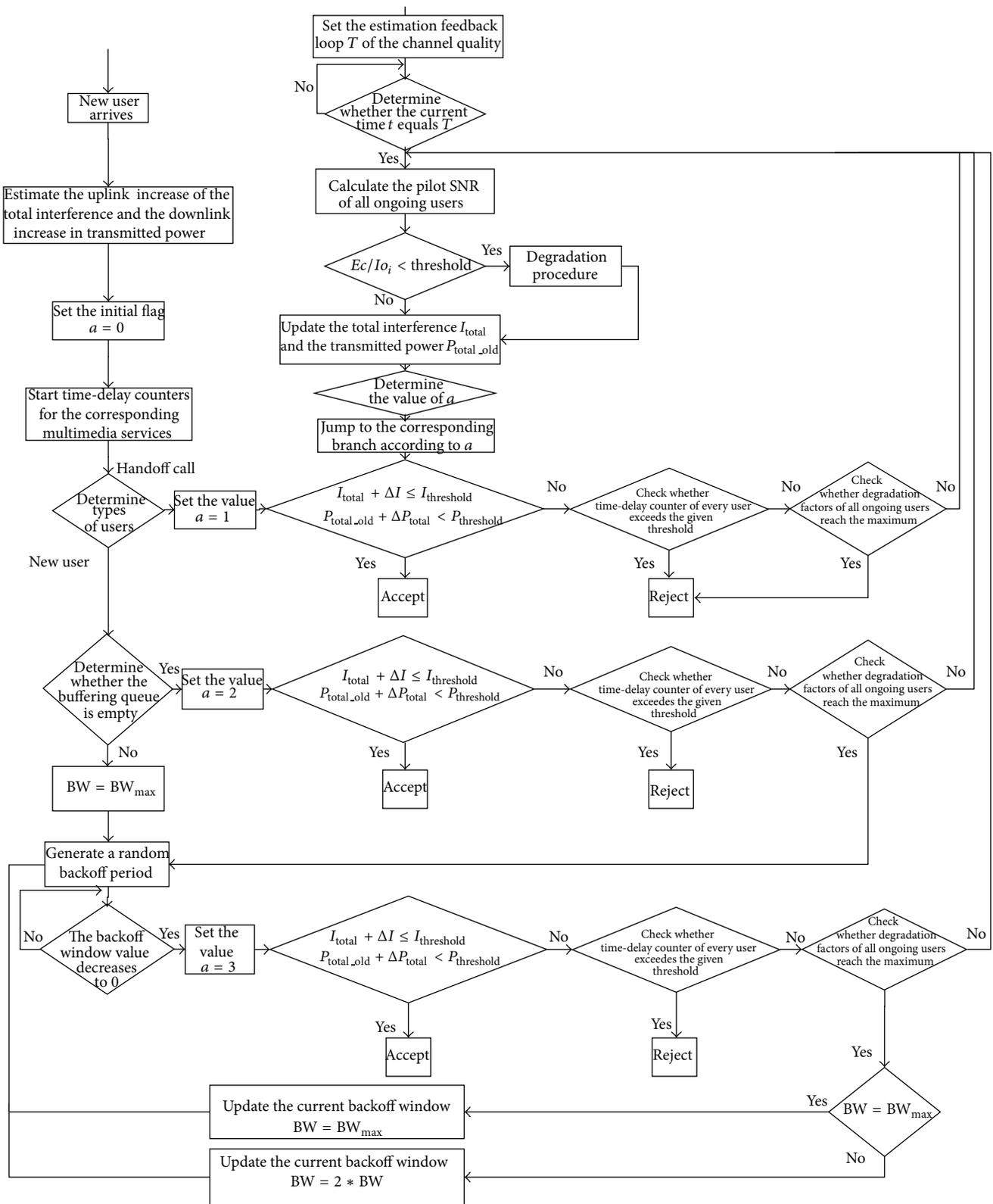


FIGURE 11: Flowchart of the proposed CAC scheme.

TABLE 1: Traffic model.

Link	Conversational class		Streaming class		Interactive class	
	Uplink	Downlink	Uplink	Downlink	Uplink	Downlink
Activity factor	0.6	0.6	0.00285	0.95	0.00285	0.015
Data rate/(kbit/s)		15	{15, 30, 60, 120, 240, 480, 960}		{15, 30, 60, 120, 240, 480, 960}	
$E_b/N_0$ target/dB		7		5		3.7
Portion of arrival						
New call		30%		5%		15%
Handoff		30%		5%		15%
Mean call duration/s		100		1200		1200
Priority		Premium		Assured		Best effort
Max delay for new calls/frame calls		2		7		20
$w_i$		1		3		6

*Stage 2.* Random delay time  $T$  is the product of the timeslot and the random backoff window value  $BW_{\text{now}}$  ranging from 0 to  $BW$ , which is  $T = BW_{\text{now}} \cdot \tau$ .

*Stage 3.* The initial setting for the current backoff window value is  $BW_{\text{now}}$ , and reset the backoff window counter.

*Stage 4.* When the backoff window value gradually decreases to 0, check whether the condition (35) is satisfied. If so, the call is accepted; if not, execute a new rate scheduling immediately and check whether the condition (35) is satisfied. If so, the call is also accepted. Otherwise, step to next stage.

*Stage 5.* Check whether time-delay counter of every user exceeds the given threshold. If satisfied, reject; otherwise, go to next stage.

*Stage 6.* If  $BW_{\text{now}} \geq BW_{\text{max}}$ , update the current backoff window  $BW = BW_{\text{min}}$ ; otherwise, update the current backoff window  $BW = 2 * BW_{\text{min}}$ . Then, continue to *Stage 3*.

## 7. Performance Evaluation

*7.1. Traffic Model and Simulation Parameters.* A 37-cell layout is considered, in which mobiles are distributed uniformly. Suppose the uplink and downlink bandwidth are 3.84 MHz. For each cell there are three classes of traffic, that is, the class of conversational, streaming, and interactive. Conversational class is constant bit rate (CBR) under an ON/OFF activity model. As in [18], streaming class is modeled as a discrete state, continuous-time Markov process. The interactive traffic is approximately modeled as the Pareto process. Further, there are two major types of calls which can arrive at any cell: new calls originated from the local cell and handoff calls coming from adjacent cells corresponding to each traffic. Since it is more reluctant to block a handoff call than a new call, the handoff calls should be given a higher priority. Based on the multimedia traffic requirements, the traffic characteristics and QoS requirements are defined in Table 1.

*7.2. Simulation Results and Performance Evaluation.* The paper analyzes several parameters for ease of comparison: the

blocking probability of new calls, the dropping probability of handover calls, outage probability, throughput, package loss, and average delays (waiting time in the queue to start transmitting).

With the call arrival rate varying, the blocking probability of new calls and the dropping probability of handover calls under two strategies and different classes are shown in Figures 12 and 13, respectively. We could see that the higher the priority level is, the lower the average blocking probability and the dropping probability are. Meanwhile, due to rate adaptation and buffering policy, the proposed MAC strategy based on the channel quality greatly reduces the average blocking probability and the dropping probability. Performance of proposed algorithm is obviously much more excellent than the scheme in [9], which guarantees the priority and fairness between new calls and handoff calls.

In order to verify the ability of the proposed scheme to assure the QoS of all users during their whole service time, the schemes are assessed in terms of outage probability. Figure 14 illustrates the changing of the outage probability with the arriving rate. As is shown, the outage probability is relevant to the priority of various classes and the arriving rate. What is more, the proposed MAC strategy based on the channel quality adopts the rate degradation to handle blocking problems, which reduce the link load factor and the whole interference. At the same time, buffering strategies can balance the traffic load effectively. The proposed scheme cannot only ensure a low average blocking probability as well as a low dropping probability but also ensure an ideal overflow probability of the IDMA system.

Figure 15 shows the resource utilization in the condition of different traffic loads. The throughput is defined as the total bit rate that system can maintain. We can observe that the throughput increases with the increment of network load and starts declining when arriving rate exceeds a point. For the proposed MAC strategy, the throughput is higher than the one in the scheme in [9] when congestion occurs, most users are accepted by means of buffering strategy, thus guaranteeing QoS and ensuring the fairness of each service. In accordance with simulation testing, the proposed MAC strategy can relax the load in the network effectively and improve the whole system quality for the communication to some extent.

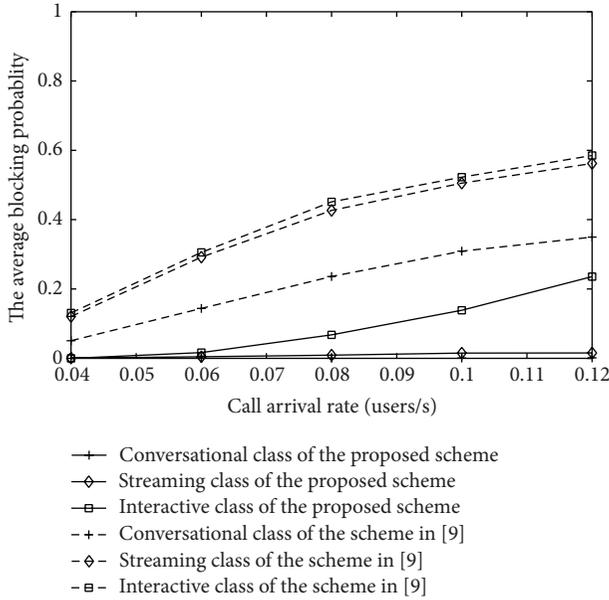


FIGURE 12: The blocking probability of IDMA system.

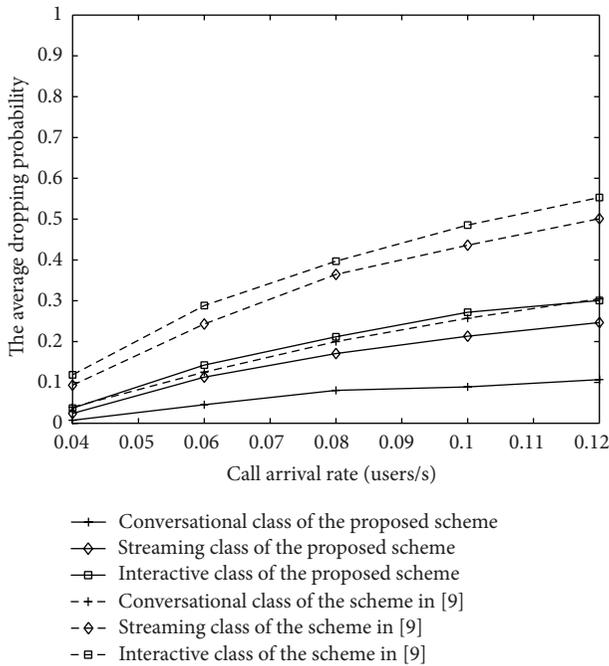


FIGURE 13: The dropping probability of IDMA system.

The package losses of different priorities are compared in Figure 16. When congestion occurs, the call can be buffered in the system for another access attempt. However, when the backoff window length is too large, unpredictable delay leads to the packet loss unavoidably. In addition, rate adaptation, that the existing calls are degraded according to their priority in order to accept more session class will lead to increasing the rate of the losing packet. Therefore, relative to strategy in [9], the proposed MAC strategy based on the channel quality sacrifices parts of the transmission rate to ensure the better performance.

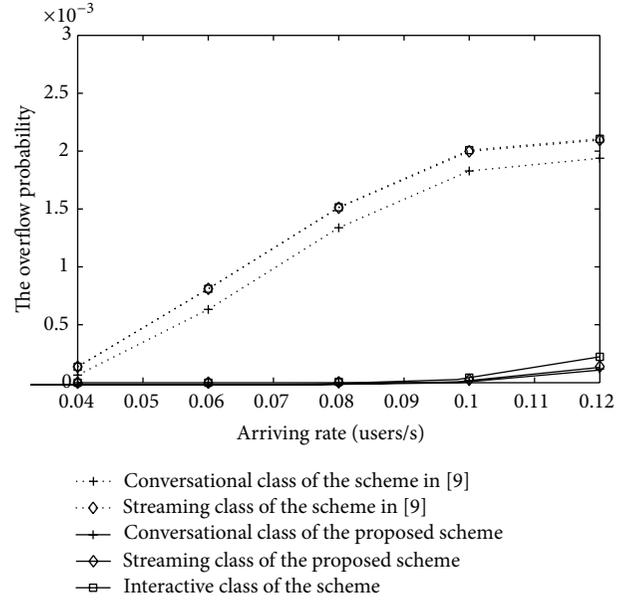


FIGURE 14: The overflow probability of IDMA system.

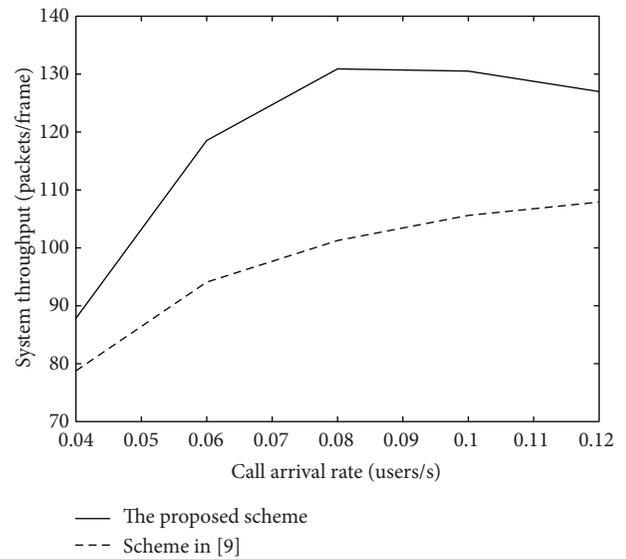


FIGURE 15: System throughput.

The average end-to-end delay is another important factor of system evaluation. The influence of an improved schedule strategy on the average end-to-end delay is simulated in Figure 17. As is shown in Figure 17, the average end-to-end delay of the proposed MAC strategy is greater than that of scheme in [9]. Here we focus on the waiting time including access delay that is the time cost in sending an access request successfully and the queuing time that the messages spend in the buffer. With the proposed MAC, the unaccepted call would cache in a delay queue temporarily and this increases the waiting time compared with the scheme in [9] when considering the same access delay.

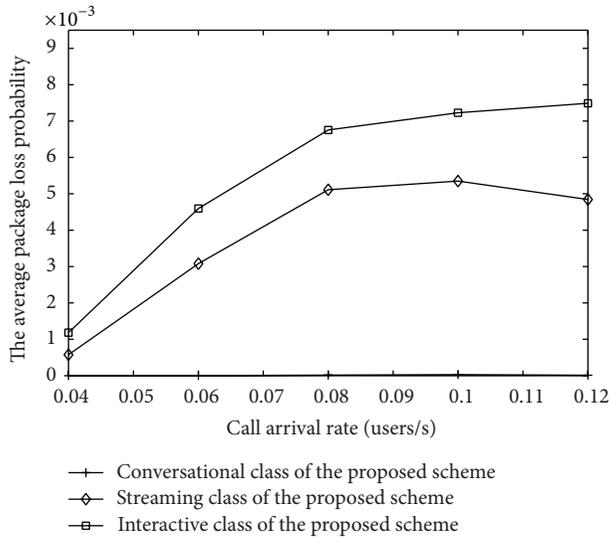


FIGURE 16: The package loss of IDMA system.

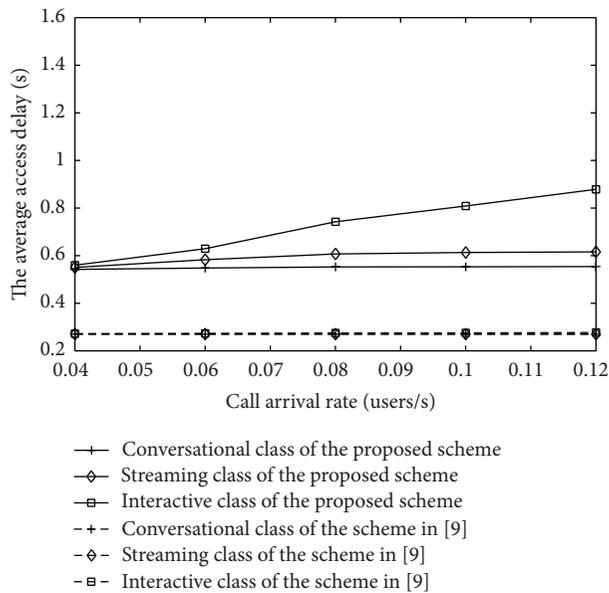


FIGURE 17: The average end-to-end delay of IDMA system.

### 8. Conclusions

In this paper, we derived the minimum-power allocation algorithm for mobile terminals that transmit multimedia traffic in wide-band IDMA system. We also proved that prediction of satellite channel quality is requisite of meeting dynamically changing environments. With effective power control and accurate channel information per timeslot, we proposed a new rate adaptation for the MAC protocol of IDMA wireless system. Furthermore, to enhance the performance of the MAC protocol, we developed a new CAC algorithm based on minimum-power allocation as well as rate scheduling and buffering strategies. The new MAC protocol can be adaptive to guarantee the QoS requirements of all

kinds of services, improving the fairness and the utilization of resources efficiency. In summary, IDMA-based MAC protocol with consideration on channel quality is a promising protocol for satellite networks.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgments

The presented work is sponsored by the National Natural Science Foundation of China (61371100 and 61001093), the Promotive Research Fund for Excellent Young and Middle-Aged Scientist of Shandong Province (BS2012DX001), the Natural Scientific Research Innovation Foundation in Harbin Institute of Technology (HIT.NSRIF.2011114), and the Fundamental Research Funds for the Central Universities (HIT.NSRIF.2013136).

### References

- [1] M. Haardt, A. Klein, R. Koehn et al., "TD-CDMA based UTRA TDD mode," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 8, pp. 1375–1385, 2000.
- [2] L. Ping, "Interleave-division multiple access and chip-by-chip iterative multi-user detection," *IEEE Communications Magazine*, vol. 43, no. 6, pp. S19–S23, 2005.
- [3] P. Sharma and K. Chandra, "Prediction of state transitions in Rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 416–425, 2007.
- [4] D.-D. Luong, J. Grégoire, and Z. Dziong, "Pattern-based channel quality prediction for adaptive coding and modulation in wireless networks," in *Proceedings of the IEEE International Conference on Communications (ICC '10)*, pp. 1–6, May 2010.
- [5] Q. Guo, X. Yuan, and L. Ping, "Multi-user detection techniques for Potential 3GPP long term evolution (LTE) schemes," in *Proceedings of the International Workshop on Multi-Carrier Spread Spectrum*, vol. 6, 2007.
- [6] X. Wang, "An FDD wideband CDMA MAC protocol with minimum-power allocation and GPS-scheduling for wireless wide area multimedia networks," *IEEE Transactions on Mobile Computing*, vol. 4, no. 1, pp. 16–28, 2005.
- [7] Q. Shen, G. Zhu, G. Liu, and W. Zhang, "Joint rate-adaptive and backoff waiting admission control for differentiated services in CDMA cellular network," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '07)*, pp. 957–960, Shanghai, China, September 2007.
- [8] K. Kusume and G. Bauch, "CDMA and IDMA: iterative multiuser detections for near-far asynchronous communications," in *Proceedings of the IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '05)*, vol. 1, pp. 426–431, September 2005.
- [9] X. Yang, G. Feng, and D. S. C. Kheong, "Call admission control for multiservice wireless networks with bandwidth asymmetry between uplink and downlink," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 1, pp. 360–368, 2006.
- [10] D. Adami, R. G. Garroppo, and S. Giordano, "Resource management and QoS architectures in DAMA satellite access

- networks,” in *Proceedings of the IEEE International Conference on Communications*, pp. 2978–2982, May 2002.
- [11] P. Chini, G. Giambene, and S. Kota, “A survey on mobile satellite systems,” *International Journal of Satellite Communications and Networking*, vol. 28, no. 1, pp. 29–57, 2010.
- [12] F. Fontán, M. Castro, and C. Cabado, “Statistical modeling of the LMS channel,” *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 1549–1567, 2001.
- [13] F. Fontan and M. A. Vazquez-Castro, “S-band LMS propagation channel behavior for different environments, degrees of shadowing and elevation angles,” *IEEE Transactions on Broadcasting*, vol. 44, no. 1, pp. 40–76, 1998.
- [14] M. Shuklal, V. K. Srivastava, and S. Tiwari, “Analysis and design of optimum interleaver for iterative receivers in IDMA scheme,” *Wireless Communications and Mobile Computing*, vol. 9, no. 10, pp. 1312–1317, 2009.
- [15] X. Wang, “Wide-band TD-CDMA MAC with minimum-power allocation and rate- and BER-scheduling for wireless multimedia networks,” *IEEE Transactions on Networking*, vol. 12, no. 1, pp. 103–116, 2004.
- [16] H. Holma and J. Laakso, “Uplink admission control and soft capacity with MUD in CDMA,” in *Proceedings of the IEEE VTS 50th Vehicular Technology Conference (VTC '99)*, vol. 1, pp. 431–435, September 1999.
- [17] X. Ge, G. Liu, H. Wang, and N. Zhang, “Call admission control scheme for IDMA-based multi-beam satellite systems in the downlink direction,” in *Proceedings of the 6th International ICST Conference on Communications and Networking in China (CHINACOM '11)*, pp. 609–613, August 2011.
- [18] Q. Huang, S. Chan, K.-T. Ko, L. Ping, and P. Wang, “A QoS architecture for IDMA-based multi-service wireless networks,” in *Proceedings of the IEEE International Conference on Communications (ICC '07)*, pp. 5070–5075, Glasgow, UK, June 2007.

## Research Article

# A Mobile Anchor Assisted Localization Algorithm Based on Regular Hexagon in Wireless Sensor Networks

Guangjie Han,<sup>1,2</sup> Chenyu Zhang,<sup>1</sup> Jaime Lloret,<sup>3</sup> Lei Shu,<sup>2</sup> and Joel J. P. C. Rodrigues<sup>4</sup>

<sup>1</sup> Department of Information & Communication Systems, Hohai University, Changzhou 213022, China

<sup>2</sup> Guangdong Provincial Key Lab of Petrochemical Equipment Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming 525000, China

<sup>3</sup> Integrated Management Coastal Research Institute, Universidad Politecnica de Valencia, 46000 Valencia, Spain

<sup>4</sup> Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal

Correspondence should be addressed to Jaime Lloret; [jlloret@dcom.upv.es](mailto:jlloret@dcom.upv.es)

Received 17 April 2014; Accepted 28 May 2014; Published 13 July 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Guangjie Han et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Localization is one of the key technologies in wireless sensor networks (WSNs), since it provides fundamental support for many location-aware protocols and applications. Constraints of cost and power consumption make it infeasible to equip each sensor node in the network with a global position system (GPS) unit, especially for large-scale WSNs. A promising method to localize unknown nodes is to use several mobile anchors which are equipped with GPS units moving among unknown nodes and periodically broadcasting their current locations to help nearby unknown nodes with localization. This paper proposes a mobile anchor assisted localization algorithm based on regular hexagon (MAALRH) in two-dimensional WSNs, which can cover the whole monitoring area with a boundary compensation method. Unknown nodes calculate their positions by using trilateration. We compare the MAALRH with HILBERT, CIRCLES, and S-CURVES algorithms in terms of localization ratio, localization accuracy, and path length. Simulations show that the MAALRH can achieve high localization ratio and localization accuracy when the communication range is not smaller than the trajectory resolution.

## 1. Introduction

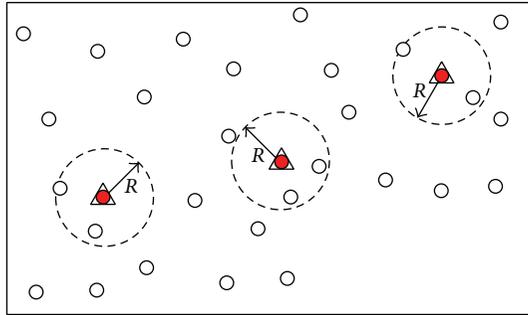
Wireless sensor networks (WSNs) consist of a large number of sensor nodes deployed in a given region of interest (ROI) to fulfill tasks such as area surveillance, biological detection, home care, object tracking, and sending information to sink nodes via multihop communication [1–4].

In WSNs, determining unknown nodes' locations is a critical task since it provides fundamental support for many location-aware protocols and applications, such as location-based routing protocol; the location information is critical for sensor nodes to make optimal routing decisions [5, 6].

The problem of localization is a process of finding location information of the unknown nodes in a given coordinate system [7–9]. Based on the distance measurement technique used, localization algorithms can be classified into range-based localization algorithms and range-free localization

algorithms. The range-based localization means that distances between sensor nodes are estimated by using physical properties of communication signal, that is, received signal strength indicator (RSSI), time of arrival (ToA), time difference of arrival (TDoA), and angle of arrival (AoA) [10]. Range-free localization algorithms do not need the distance or angle information for localization [11].

Constraints of cost and power consumption make it infeasible to equip each sensor node in the network with a GPS unit, especially for large-scale WSNs. A promising method to localize unknown nodes is to use several mobile anchors which are equipped with GPS units moving among unknown nodes and periodically broadcasting their current locations (beacon points) to help nearby unknown nodes with localization [12–15], as shown in Figure 1. This kind of architecture offers significant practical benefits, since the mobile anchor node is not as energy constrained as an



- △ Anchor point
- Mobile anchor
- Unknown node

FIGURE 1: Mobile anchor assisted localization.

unknown node and the localization accuracy also can be improved by carefully designing the mobile anchor's movement trajectory. Moreover, the size of a robot is much larger than the size of a sensor and thus it is much easier to install a GPS unit on it [16].

Generally, mobile anchor assisted localization algorithm involves three stages: (i) mobile anchor traverses the ROI while periodically broadcasting beacon packets which include their current positions; (ii) unknown nodes within the communication ranges of the mobile anchors receive the beacon packets and estimate distances to the anchors by using physical properties of communication signal when needed; and (iii) unknown nodes calculate their positions if they fall inside the overlapping communication ranges of at least three (four) noncollinear (noncoplanar) anchor nodes by using appropriate localization schemes in two-dimensional (2D) (three-dimensional (3D)) WSNs.

In this paper, we propose a mobile anchor assisted localization algorithm based on regular hexagon (MAALRH) with objectives of maximizing localization ratio and localization accuracy. To cover the entire ROI, we present a boundary compensation method (BCM) to ensure that the unknown nodes could fall inside the overlapping communication ranges of at least three noncollinear beacon points.

The rest of this paper is organized as follows. Section 2 gives an overview of mobile anchor assisted localization algorithms. Section 3 describes network model and theoretical background. Section 4 introduces the proposed MAALRH and the comparing algorithms. Simulation results and performance analysis are shown in Section 5. Finally, Section 6 concludes this paper and discusses future research issues.

## 2. Related Work

**2.1. Path Planning Scheme.** A fundamental research issue of mobile anchor assisted localization algorithm is to design path planning scheme that mobile anchor should move along in a given ROI in order to minimize the localization error as well as the time required to localize the whole network.

Path planning schemes can be either static or dynamic. Static path planning scheme designs movement trajectory

before starting execution; mobile anchor follows the predefined trajectory during the localization process. Dynamic path planning scheme designs movement trajectory dynamically or partially according to the observable environments or deployment situations and so forth.

**2.1.1. Static Path Planning Scheme.** Koutsonikolas et al. [17] proposed SCAN, DOUBLE-SCAN, and HILBERT to satisfy network coverage. Movement trajectories of SCAN and DOUBLE-SCAN are composed of a series of straight lines. HILBERT curve divides the 2D area into square cells and connects the centers of those cells using line segments. Compared with SCAN and DOUBLE-SCAN, the HILBERT can provide more noncollinear beacon points for unknown nodes. To reduce the collinearity during localization, Huang and Záruba proposed two path planning schemes, namely, CIRCLES and S-CURVES. CIRCLES consists of a sequence of concentric circles centered within a ROI. S-CURVES is based on SCAN, which progressively scans the monitoring area from left to right taking an "S" curve. Hu et al. [19] proposed a mobile anchor centroid localization (MACL) method. The mobile anchor traverses the monitoring area following a spiral trajectory while periodically broadcasting beacon packets which contain its current position and so forth. Zhang et al. [20] proposed a collaborative localization scheme using a group of mobile anchor nodes (GMAN). A GMAN is composed of three anchor nodes, which form an equilateral triangle and each anchor node locates at one of the three vertexes. Cui et al. [21] introduced five movement trajectories for 3D WSNs. LAYERED-SCAN and LAYERED-CURVE divide the 3D ROI into several layers along one axis and regard each layer as a 2D ROI. Thus, in each layer of LAYERED-SCAN and LAYERED-CURVE, the mobile anchor traverses along one dimension using SCAN and S-CURVES, respectively. TRIPLE-SCAN and TRIPLE-CURVE divide the ROI into several layers along three axes. 3D HILBERT has more turns compared with LAYERED-SCAN and TRIPLE-SCAN to overcome collinearity and coplanarity problems. Cui and Wang [22] proposed a four-mobile-beacon assisted weighted centroid localization method. The four mobile beacons form a regular tetrahedron and traverse the given ROI following the LAYERED-SCAN trajectory which consists of several parallel layers of SCAN.

**2.1.2. Dynamic Path Planning Scheme.** A large amount of dynamic path planning schemes were proposed to consider the real distribution of unknown nodes in the given ROI.

Li et al. [23] regard a WSN as a connected undirected graph. They proposed a Breadth-First (BRF) algorithm and a Backtracking Greedy (BTG) algorithm to transform the path planning issue into seeking spanning trees of the undirected graph and traversing through the graph. Thus, the movement trajectory of the mobile anchor node changes dynamically accordingly to the distribution of unknown nodes. Fu et al. proposed a novel dynamic movement trajectory based on virtual force, which is constructed by interaction force between mobile anchor and unknown nodes [24]. Each unknown node is equipped with an omnidirectional antenna.

The mobile anchor uses directional antennas to receive feedback messages from unknown nodes and calculates the total virtual force on itself. In [25], three mobile anchors form a regular triangle with the length of its communication range and move in a ROI. Unknown nodes that do not know their own positions request the mobile anchor to deliver more beacon messages. The mobile anchor decides its movement trajectory on the basis of the received request messages. In [26], six optional positions are provided to be chosen based on geometry. The mobile anchor finds a new position among the six optional positions. The unknown node with most neighbors has the most chance to be the next position of the mobile beacon.

**2.2. Localization Scheme.** Another research issue of mobile anchor assisted localization algorithm is to design localization scheme by which unknown nodes calculate their positions based on beacon points received from mobile anchors.

Ssu et al. [27] developed a localization mechanism using the geometry conjecture, that is, perpendicular bisector of a chord. If any two chords are obtained, the location of the sensor node can be easily computed based on the conjecture. In [28], instead of using the absolute RSSI values, by contrasting the measured RSSI values from the mobile beacon to a sensor node, perpendicular intersection (PI) utilizes the geometric relation of PI to compute the position of the node. Guerrero et al. [29] intruded an azimuthally defined area localization (ADAL) algorithm which utilizes a beacon with a rotary directional antenna to send message in a determined azimuth periodically, and an unknown node uses the centroid of intersection area of several beacon messages as its position. Arrival and departure overlap (ADO) [30] uses a possible area delimited by two circles with the same radius at different centers. To estimate its position, an unknown node should obtain prearrival position, arrival position, departure position, and postdeparture position of the moving beacon to compute its ADO. To improve the localization accuracy of Ssu's scheme, Lee et al. [31] proposed a method based on geometric constraints utilizing three beacon points, where two are used for obtaining the intersection area and the third is used further to delimit this area. The borderline measurement schemes determine some straight lines that pass through a sensor node and use the intersection point of these lines as its position.

### 3. Network Model and Theoretical Background

**3.1. Network Architecture and Assumptions.** Network architecture of this paper is shown in Figure 2. There are two types of sensor nodes in the network, namely, unknown node and mobile anchor. All the sensor nodes have the same communication range. Unknown nodes are deployed uniformly in the ROI. The mobile anchor travels among unknown nodes following the predefined trajectory while periodically broadcasting its current location to help nearby unknown nodes with localization. Unknown nodes estimate distances to the mobile anchor by using RSSI technique. Once

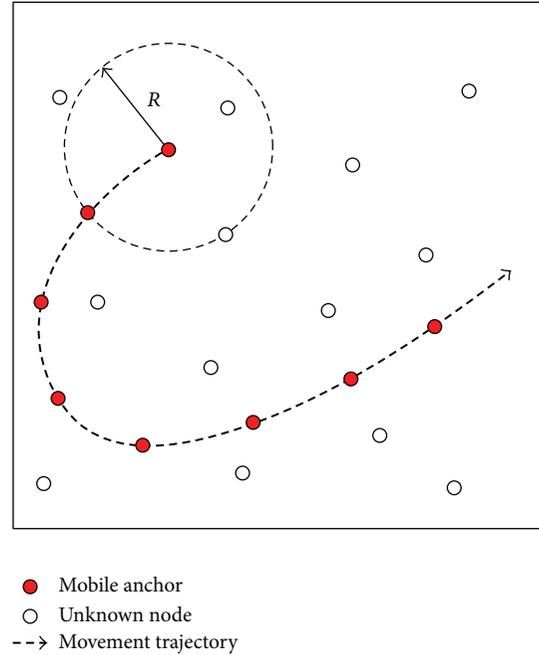


FIGURE 2: Network architecture.

an unknown node receives at least three noncollinear anchor points, it will calculate its position by using trilateration method.

Two assumptions are made.

- (a) The mobile anchor has sufficient energy for moving and broadcasting anchor packets during localization. The speed of the mobile anchor is adjustable and uniform in the process of localization.
- (b) The communication model is perfect spherical radio propagation and there exists measurement errors. The mobile anchor has identical communication range  $r$  at all anchor points. Only the sensors within the communication range are assumed to be able to receive anchor packets sent by the mobile anchor.

**3.2. Theoretical Background.** In a two-dimensional ROI, suppose that the unknown node  $p(x_0, y_0)$  can receive three anchor coordinates  $p_i(x_i, y_i), i = 1, 2, 3$ . Distances between  $p$  and  $p_i$  are  $r_i, i = 1, 2, 3$ . Assume that the measurement error ranges from  $-\varepsilon_i$  to  $\varepsilon_i, \varepsilon_i > 0$ . Thus, we can obtain

$$C_{p_i} = (x, y) \mid (r_i - \varepsilon_i)^2 \leq (x + x_i)^2 + (y + y_i)^2 \leq (r_i + \varepsilon_i)^2, \quad i = 1, 2, 3. \quad (1)$$

Unknown node calculates its coordinates by using of the trilateration. Thus, the localization error is defined as

$$\text{area}(C_{p_i}) = (x, y) \mid x \in \bigcap_{i=1}^3 C_{p_i}, \quad y \in \bigcap_{i=1}^3 C_{p_i}. \quad (2)$$

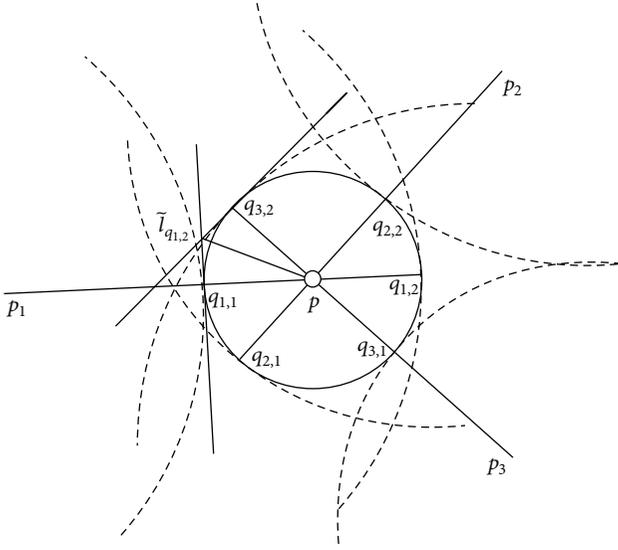


FIGURE 3: Analysis of localization error.

When the measurement error  $\varepsilon$  is relatively small,  $C_{p_i}$  can be linearized and approximated by  $\tilde{C}_{p_i}$ . We proved that the localization error is the smallest when three anchor nodes are placed symmetrically; namely, three anchor nodes form a regular triangle [32]. As shown in Figure 3, let  $l_{p,p_i}$  be the straight line passing through both  $p$  and  $p_i$ . Thus,  $l_{p,p_i}$  will intersect with  $S_p$  at two points  $q_{i,1}$  and  $q_{i,2}$ . For  $j = 1, 2$ , we define the line passing through  $q_{i,j}$  and tangent to  $S_p$  as  $\tilde{l}_{q_{i,j}}$ . Therefore [32],

$$\text{area}(\tilde{C}) = 2\varepsilon^2 \left( \tan \frac{\alpha_{1,2}}{2} + \tan \frac{\alpha_{2,3}}{2} + \tan \frac{\alpha_{3,1}}{2} \right). \quad (3)$$

Note that

$$\alpha_{1,2} + \alpha_{2,3} + \alpha_{3,1} = \pi. \quad (4)$$

Since  $(\tan x)'' = 2 \tan x(1 + \tan x) \geq 0$ , when  $0 \leq x \leq \pi/2$ , we, thus, obtain

$$\begin{aligned} \text{area}(\tilde{C}) &= 6\varepsilon^2 \frac{1}{3} \left( \tan \frac{\alpha_{1,2}}{2} + \tan \frac{\alpha_{2,3}}{2} + \tan \frac{\alpha_{3,1}}{2} \right) \\ &\geq 6\varepsilon^2 \tan \frac{\alpha_{1,2} + \alpha_{2,3} + \alpha_{3,1}}{6} = 6\varepsilon^2 \tan \frac{\pi}{6}. \end{aligned} \quad (5)$$

The equality holds when

$$\alpha_{1,2} = \alpha_{2,3} = \alpha_{3,1} = \frac{\pi}{3}. \quad (6)$$

In other words, the localization error is the smallest when three anchor nodes form a regular triangle.

#### 4. Mobile Anchor Assisted Localization

The problem of path planning for mobile anchor is to design movement trajectory satisfying the following properties: (i) it should pass closely to as many potential node positions

as possible, aiming at localizing as many unknown nodes as possible; (ii) it should provide each unknown node with at least three (four) noncollinear (noncoplanar) anchor points in a 2D (3D) WSN to achieve unique estimation of known node's position; (iii) it should be as short as possible to reduce the energy consumption of mobile anchors and time for localization.

The performances of mobile anchor assisted localization algorithm are influenced by the following factors.

- (a) Communication range: a larger communication range of the mobile anchor covers more unknown nodes. Thus, the unknown nodes have more choices to select appropriate anchor points to calculate their coordinates.
- (b) Movement trajectory: a well designed movement trajectory can eliminate collinearity (coplanarity) problem and make full use of the real-time information, that is, the distribution of the known nodes, environment information, and so forth.
- (c) Broadcast interval: a shorter broadcast interval means that the mobile anchor would broadcast its location more frequently, which may bring about a better localization performance.
- (d) Path length: a longer path length means that the mobile anchor has more opportunities to broadcast its location and pass by more unknown nodes; however, it will consume more energy.

Thus, we should solve the above four problems when designing a mobile anchor assisted localization algorithm.

4.1. MAALRH. The general procedure of MAALRH consists of four steps, as shown in Algorithm 1.

4.1.1. Network Segmentation. We assume that the ROI is a square. We divide the ROI into several subrectangles according to the length of the square. The communication range of mobile anchor nodes can be adjusted according to the length of subrectangles. The distance between two successive segments of the subrectangles is defined as the resolution ( $R$ ). Figure 4 gives an example of network segmentation. The length of the ROI is  $L$ . The square can be divided into  $n$  subrectangles which satisfy  $L = nR$ ,  $n \in N^*$ .

4.1.2. Movement Trajectory. Assume that the ROI is a square with the area of  $L \times L$ ; the vertex coordinates of the ROI are  $(x_{\min}, y_{\min})$ ,  $(x_{\max}, y_{\min})$ ,  $(x_{\min}, y_{\max})$ , and  $(x_{\max}, y_{\max})$ , respectively. The mobile anchor is initially located at the centroid of the ROI. The initial coordinates of the mobile anchor can be calculated by using

$$\begin{aligned} x_0 &= \frac{|x_{\max}| - |x_{\min}|}{2} \\ y_0 &= \frac{|y_{\max}| - |y_{\min}|}{2}. \end{aligned} \quad (7)$$

*Step 1.* The ROI is divided into  $n$  subrectangles which satisfy  $L = nR, n \in N^*$ . The communication range of the mobile anchor is equal to the resolution, that is,  $R = r$ .

*Step 2.* The mobile anchor traverses the ROI following the regular hexagon movement trajectory (the concretely movement trajectory is depicted in Figures 5 and 7) while periodically broadcasting anchor packets  $\{T_{\text{send}}, (x_i, y_i), ID\}$ , where  $T_{\text{send}}$  denotes the sending time,  $(x_i, y_i)$  stands for its current location, and  $ID$  represents the packet ID;

*Step 3.* Unknown nodes receive the anchor packets broadcasted by the mobile anchor and estimate distances to them by using the RSSI technique;

*Step 4.* Each unknown node decides if any of the three noncoplanar anchor coordinates can almost form a regular triangle and if it is within the regular triangle. If so, the unknown node calculates its location by using the trilateration.

ALGORITHM 1: MAALRH algorithm.

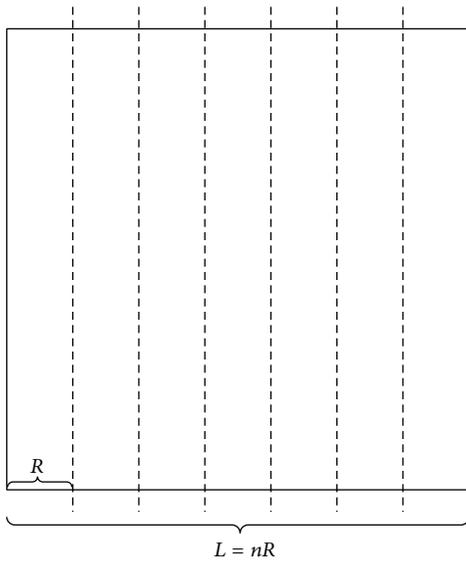
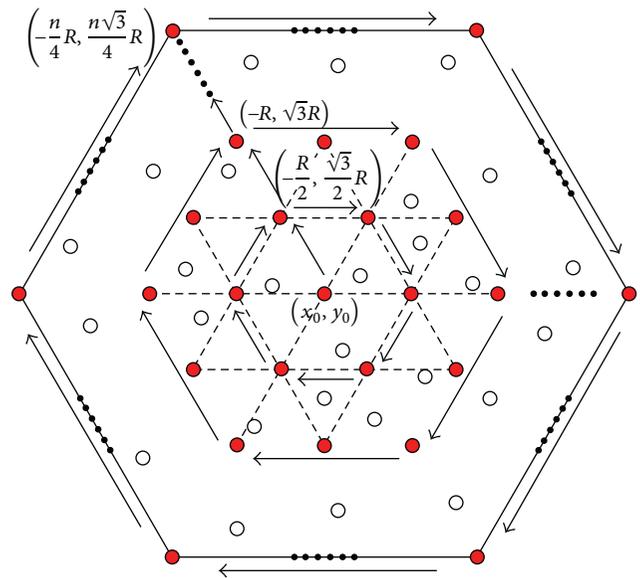


FIGURE 4: An example of network segmentation.

The mobile anchor traverses the entire ROI following the regular hexagon trajectory at the speed of  $v$  and broadcasts its current location  $(x_i, y_i)$  with an interval  $T$  and a communication range  $r$  as depicted in Figure 5.

A rectangular coordinate system is constructed with the origin at  $(x_0, y_0)$ . At first, the mobile anchor moves from  $(x_0, y_0)$  to one of the vertexes of the regular hexagon with the side length of  $R$ ; for instance, the mobile anchor moves from  $(x_0, y_0)$  to  $(-1/2)R, (\sqrt{3}/2)R$ . Then, the mobile anchor moves along the sides of the first regular hexagon with the side length of  $R$ . When the mobile anchor arrives at the point  $(-1/2)R, (\sqrt{3}/2)R$  once again, it moves to  $(-R, \sqrt{3}R)$  and then moves along the sides of the second regular hexagon with the side length of  $2R$ . The side length of regular hexagons increases by  $R$  each time, and the mobile anchor traverses the ROI along the sides of  $n/2$  regular hexagons. The cycle repeats until the mobile anchor arrives at the point  $(-n/4)R, (n\sqrt{3}/4)R$  twice. Thus, the total path length without a boundary compensation method can be calculated by using

$$L'_{\text{MAALRH}} = \frac{3}{4}n^2R + \frac{3}{2}nR = \frac{3}{4}\frac{L^2}{R} + \frac{3}{2}L. \quad (8)$$



- Mobile anchor
- Unknown node
- Movement trajectory

FIGURE 5: Movement trajectory of MAALRH without a boundary compensation method.

Thus, for a given ROI, the total path length depends on the  $R$ . A smaller  $R$  results in a larger path length. Since  $R = vT$ , with the same movement speed, the smaller  $R$  is, the less anchor packets are broadcasted. By this means, the broadcasted anchor points form many regular triangles.

**4.1.3. Boundary Compensation Method.** Since the regular hexagon movement trajectory leaves four corners of ROI uncovered, to improve localization ratio, we present a boundary compensation method to enhance the MAALRH. In BCM, mobile anchor travels in the sensing area which is larger than the ROI, as shown in Figure 6. Unknown nodes are deployed uniformly in the ROI while mobile anchor moves in the sensing area according to the movement trajectory. Assume that the length of the ROI is  $L$ , the length of the sensing area is  $L'$ , the relation of  $L$  and  $L'$  can be expressed as  $L' = L + X, X \in R_+$ , and  $X$  is determined

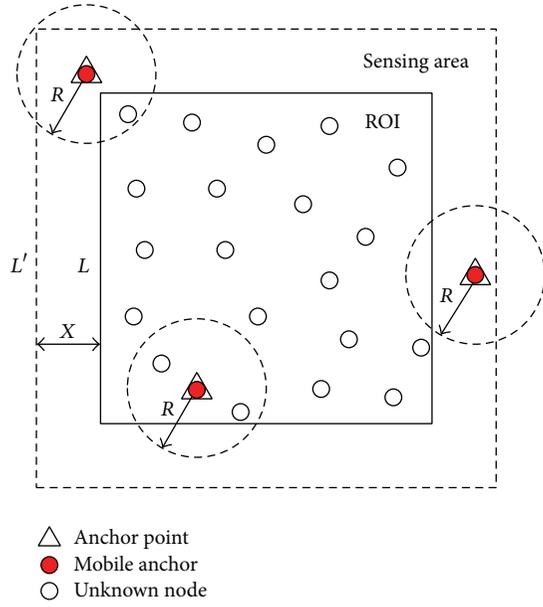


FIGURE 6: The relation of sensing area and ROI.

by the movement trajectory and the communication range of mobile anchor. In other words, by enlarging movement area of mobile anchor, the ROI can be traversed entirely. Therefore, unknown nodes which are at the boundary of the ROI can be localized.

We choose  $X = 2R$  here. Thus, when the mobile anchor arrives at the point  $((-n/4)R, (n\sqrt{3}/4)R)$  once again after it moves along the  $(n/2)$ th regular hexagon with the side length of  $3nR$ , the mobile anchor moves to  $(-(n+2)/4)R, ((n+2)\sqrt{3}/4)R$  and moves along the  $n/2 + 1$ th regular hexagon with the side length of  $(3n+6)R$ . Then, the mobile anchor moves to  $(-(n+4)/4)R, ((n+4)\sqrt{3}/4)R$  and moves along the  $n/2 + 2$ th regular hexagon with the side length of  $(3n+12)R$  to ensure that unknown nodes at the boundary of the ROI could fall inside the overlapping communication ranges of at least three noncollinear anchors, as shown in Figure 7.

Thus, path length of the MAALRH with a boundary compensation method can be calculated by using

$$L_{MAALRH} = \frac{3}{4}n^2R + \frac{15}{2}nR + 18R. \quad (9)$$

**4.1.4. Trilateration.** An example of the trilateration is shown in Figure 8. Suppose that the unknown node  $D(x, y)$  receives three anchor packets from the mobile anchor, namely,  $A(x_a, y_a)$ ,  $B(x_b, y_b)$ , and  $C(x_c, y_c)$ . Distances between  $A, B, C$ , and  $D$  are  $d_a, d_b$ , and  $d_c$ , respectively. Since the unknown node  $D$  is within the regular triangle which is composed of  $A, B$ , and  $C$ , unknown node  $D$  will calculate its location by using

$$\begin{aligned} (x - x_a)^2 + (y - y_a)^2 &= d_a^2, \\ (x - x_b)^2 + (y - y_b)^2 &= d_b^2, \\ (x - x_c)^2 + (y - y_c)^2 &= d_c^2. \end{aligned} \quad (10)$$

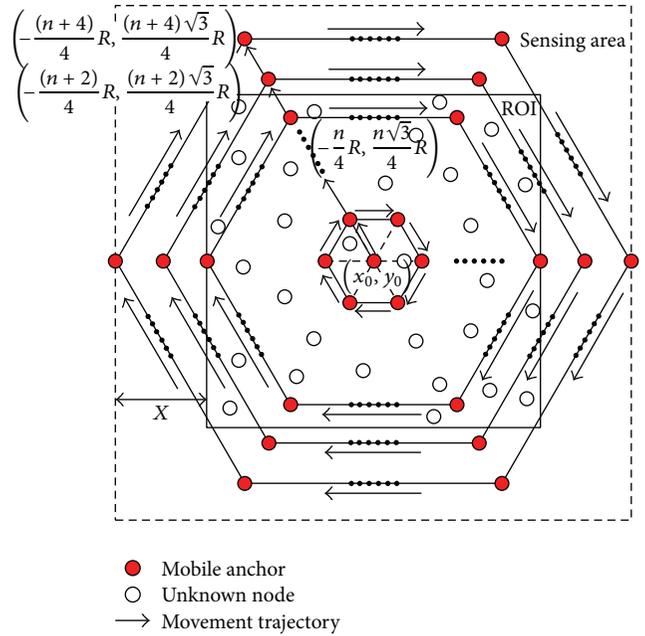


FIGURE 7: Movement trajectory of MAALRH with a boundary compensation method.

Hence,

$$D(x, y)^T = \frac{1}{2}A^{-1}B, \quad (11)$$

where

$$\begin{aligned} A &= \begin{pmatrix} x_a - x_c & y_a - y_c \\ x_b - x_c & y_b - y_c \end{pmatrix}, \\ B &= \begin{pmatrix} x_a^2 - x_c^2 + y_a^2 - y_c^2 + d_c^2 - d_a^2 \\ x_b^2 - x_c^2 + y_b^2 - y_c^2 + d_c^2 - d_b^2 \end{pmatrix}. \end{aligned} \quad (12)$$

**4.2. Comparing Algorithms.** Various path planning schemes have been proposed for single mobile anchor assisted localization. We choose HILBERT, CIRCLES, and S-CURVES to be compared with our proposed MAALRH.

**4.2.1. HILBERT.** HILBERT can reduce the collinearity without significantly increasing the path length compared with SCAN and DOUBLE-SCAN. A  $level-n$  HILBERT curve divides the  $L \times L$  ROI into  $4^n$  square cells and connects the centers of those cells using  $4^n$  line segments [17]. The resolution of the HILBERT curve is defined as the length of each line segment. Thus,  $L, R$ , and  $n$  satisfy  $4^n = (L/R) \times (L/R)$ . Therefore, the path length of HILBERT curve can be calculated by using

$$L_{HILBERT} = n^2 \times R. \quad (13)$$

**4.2.2. CIRCLES.** CIRCLES consists of a sequence of concentric circles centered within the ROI [18]. The resolution is

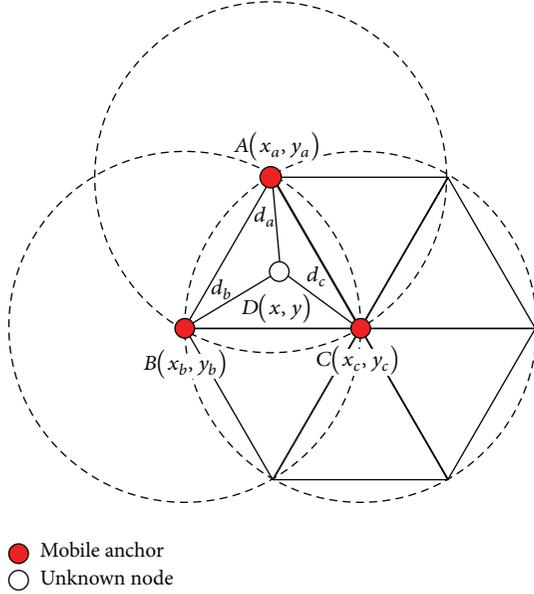


FIGURE 8: An example of the trilateration.

defined as the diameter of the innermost circle. For each outer circle the radius is increased by  $R$  sequentially. CIRCLES can reduce collinearity of the anchor points; all areas within the concentric circles can be localized. However, CIRCLES leaves four corners of the ROI uncovered. The path length of CIRCLES can be calculated by using

$$L_{\text{CIRCLES}} = \frac{n^2 \pi R}{4} + \left(\frac{n}{2} - 1\right) R. \quad (14)$$

**4.2.3. S-CURVES.** S-CURVES is based on the SCAN, which progressively scans the ROI from left to right. However, S-CURVES takes an “S” curve instead of moving in a straight line [18]. For a  $L \times L$  ROI and a resolution of  $R$ , there are  $\lfloor 2(n-1)/3 \rfloor + 1$  curve. Each vertical “S” curve consists of  $n-1$  semicircle with the radius of  $R/2$ . Therefore, the path length of S-CURVES can be calculated by using

$$L_{\text{S-CURVES}} = \frac{(n-1)\pi R}{2} \left( \left\lfloor \frac{2(n-1)}{3} \right\rfloor + 1 \right) + (n-2)R + \frac{\pi R}{2}. \quad (15)$$

## 5. Performance Evaluation

### 5.1. Evaluation Criteria

**Localization Ratio.** Localization ratio is the ratio of the number of localizable unknown nodes to the number of unknown nodes. This metric also indicates the coverage degree of the movement trajectory. Localization ratio is defined as

$$L_{\text{ratio}} = \frac{N_l}{N_o}, \quad (16)$$

TABLE 1: Parameters used in the simulation.

ROI size	480 m × 480 m
Communication range	60–140 m
Resolution	60 m, 80 m, and 120 m
Movement speed	10 m/s
Number of unknown nodes	100–500

where  $N_l$  is the number of localizable unknown nodes and  $N_o$  is the number of unknown nodes.

**Localization Accuracy.** The localization error of unknown node  $i$  is defined as

$$e_i = \frac{\sqrt{(u_i - x_i)^2 + (v_i - y_i)^2 + (w_i - z_i)^2}}{r}, \quad (17)$$

where  $(u_i, v_i, w_i)$  are real coordinates of an unknown node  $i$ ,  $(x_i, y_i, z_i)$  are estimated coordinates of an unknown node  $i$ , and  $r$  is the communication range of sensor nodes.

We evaluate the localization accuracy by using average and standard deviation of localization errors of unknown nodes, which are defined as

$$\mu_e = \frac{1}{N_l} \sum_{i=1}^{N_l} e_i, \quad (18)$$

$$\sigma_e = \sqrt{\frac{1}{N_l} \sum_{i=1}^{N_l} (e_i - \mu_e)^2},$$

where  $N_l$  is the number of localizable unknown nodes in a WSN.

**Path Length.** To save energy consumption and time for localization, the path length of the mobile anchor node should be as short as possible.

**Scalability.** Scalability means that the localization performance is independent of the unknown nodes density.

**5.2. Experiment Parameters.** Our simulations are performed using Matlab. Suppose that the ROI is a square. Table 1 lists parameters used in simulations. Five movement trajectories are compared in this section, HILBERT, CIRCLES, S-CURVES, MAALRH, and MAALRH\_BCM (we name the MAALRH with the boundary compensation method as MAALRH\_BCM). The trilateration is used to calculate coordinates of unknown nodes. To ensure reliability of evaluation results, 50 simulation runs were performed for each set of simulation condition, with a different uniform deployment of unknown nodes on each occasion.

**5.3. Simulations and Analysis.** We evaluate performances of five movement trajectories under three resolutions: 60 m, 80 m, and 120 m in terms of localization ratio, localization accuracy, path length, and scalability.

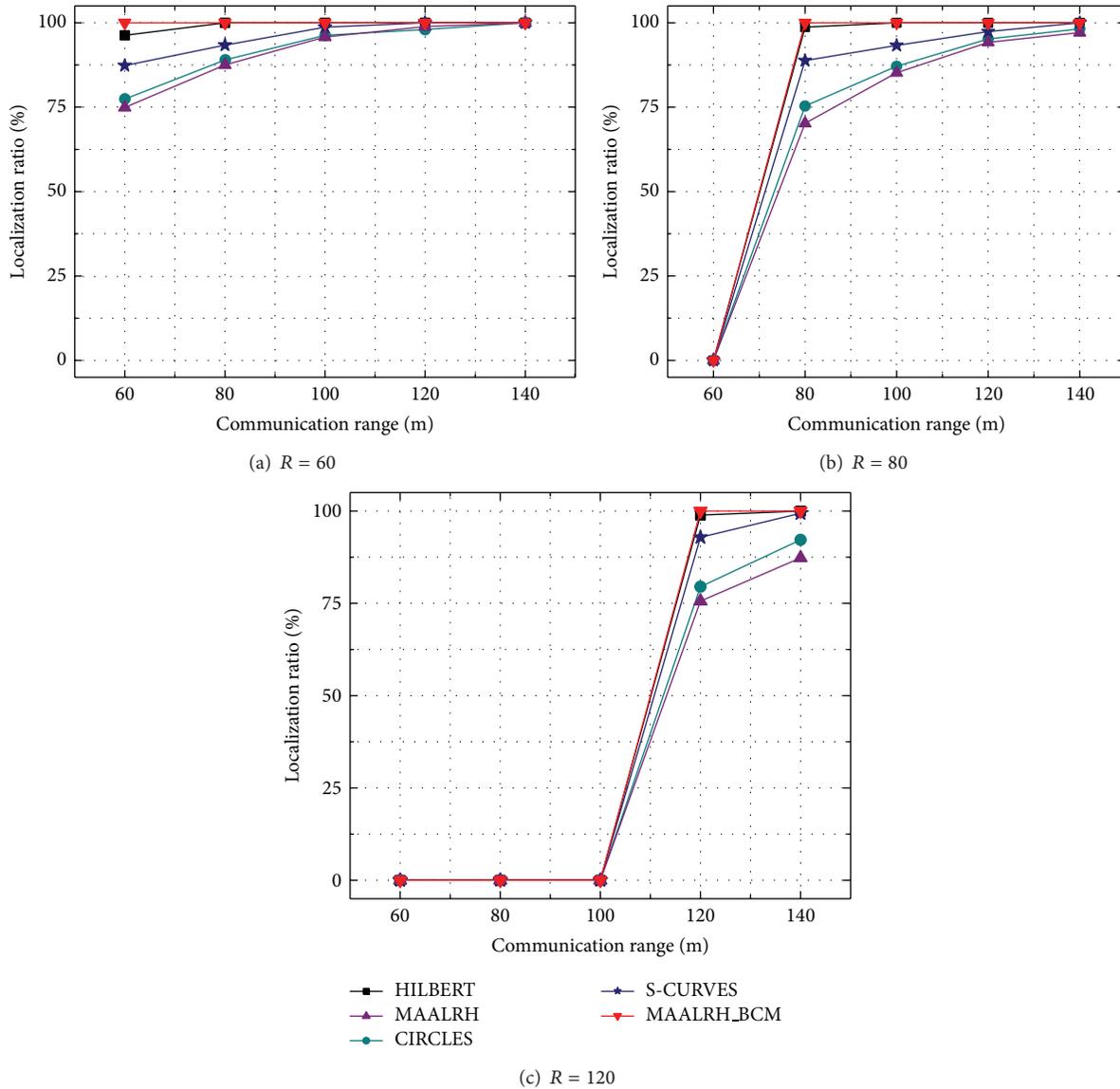


FIGURE 9: Localization ratio with different resolutions.

5.3.1. *Localization Ratio.* Figure 9 depicts the relation between localization ratios and communication ranges under three resolutions. The MAALRH\_BCM outperforms HILBERT, CIRCLES, and S-CURVES and MAALRH in general. Localization ratios of MAALRH\_BCM and HILBERT are similar when the communication range is not smaller than the resolution. Localization ratio of S-CURVES is larger than that of MAALRH and CIRCLES, since MAALRH and CIRCLES leave four corners of ROI uncovered. Localization ratio of MAALRH\_BCM can reach 100% as long as the communication range is not smaller than the resolution, since unknown nodes can receive three noncollinear anchor points which can form a regular triangle to estimate their positions. When the resolution is relatively small, that is,  $R = 60$  m, localization ratios increase rapidly with the increase of the communication range and the localization ratios of the

five algorithms can all reach 100% when the communication range is 140 m. However, with the increase of the resolution, the mobile anchor does not broadcast anchor packets as frequently as in the previous case (i.e.,  $R = 60$ ); only MAALRH\_BCM and HILBERT can reach 100% localization ratio (i.e.,  $R = 80$  m and  $R = 120$  m). That is because with the boundary compensation method MAALRH\_BCM can provide noncollinear anchor points to ensure that the movement trajectory can cover the entire ROI. When the resolution is much larger than the communication range, unknown nodes cannot receive enough anchor points to estimate their coordinates, which results in zero localization ratios. For MAALRH and CIRCLES, the movement trajectories leave four corners of the ROI uncovered and unknown nodes on the boundary of the ROI cannot receive at least three noncollinear anchor packets from the mobile anchor,

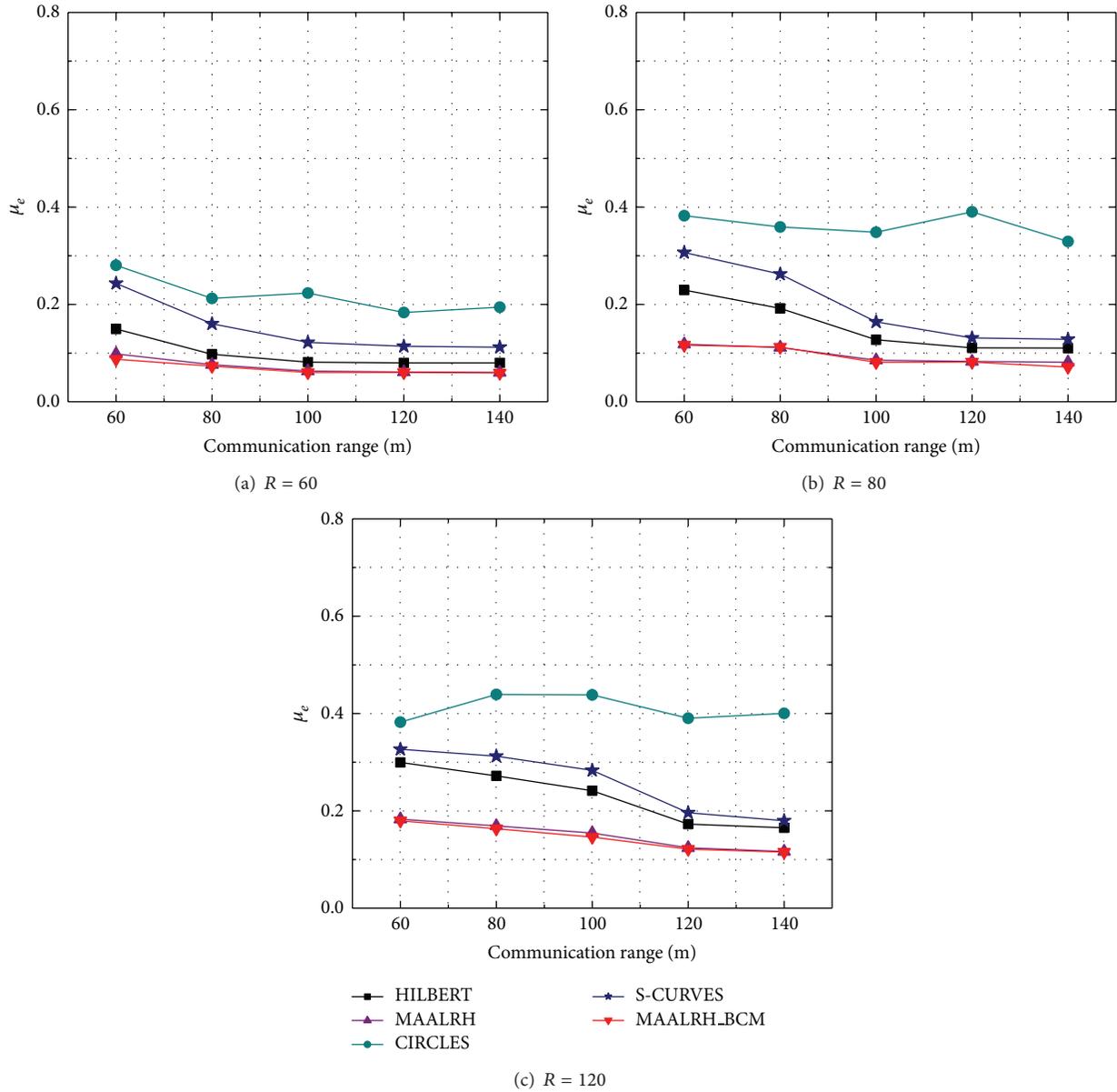


FIGURE 10: Average deviation with different resolutions.

especially when the resolution is large. Thus, MAALRH and CIRCLES perform worse compared with the other three algorithms.

From the simulation, we can conclude that the localization ratio depends on the communication range of the sensor nodes and the resolution of the movement trajectory which determines the amount and the distance interval of the anchor points.

**5.3.2. Localization Accuracy.** Figure 10 presents the variation of  $\mu_e$  under three resolutions. We can observe that with the increase of the resolution,  $\mu_e$  increase correspondingly, because distance between two neighboring anchor points increases with the increase of the resolution and results in a larger measurement error. MAALRH and MAALRH\_BCM

almost have the same average deviation, since both of them use three anchor points which form a regular triangle to estimate unknown nodes' coordinates. Besides, the MAALRH\_BCM has a boundary comprehension method to ensure the localization accuracy of all the unknown nodes within a ROI. HILBERT and S-CURVES have larger average deviations compared with that of MAALRH and MAALRH\_BCM. CIRCLES performs worst among the five movement trajectories. Because in HILBERT, CIRCLES, and S-CURVES, ordinary nodes randomly select three noncollinear anchor points to calculate their positions.  $\mu_e$  of MAALRH, MAALRH\_BCM, HILBERT, and S-CURVES decrease apparently when the communication range is equal to the resolution (with the variation from 60 m to 140 m), because more unknown nodes can receive three noncoplanar

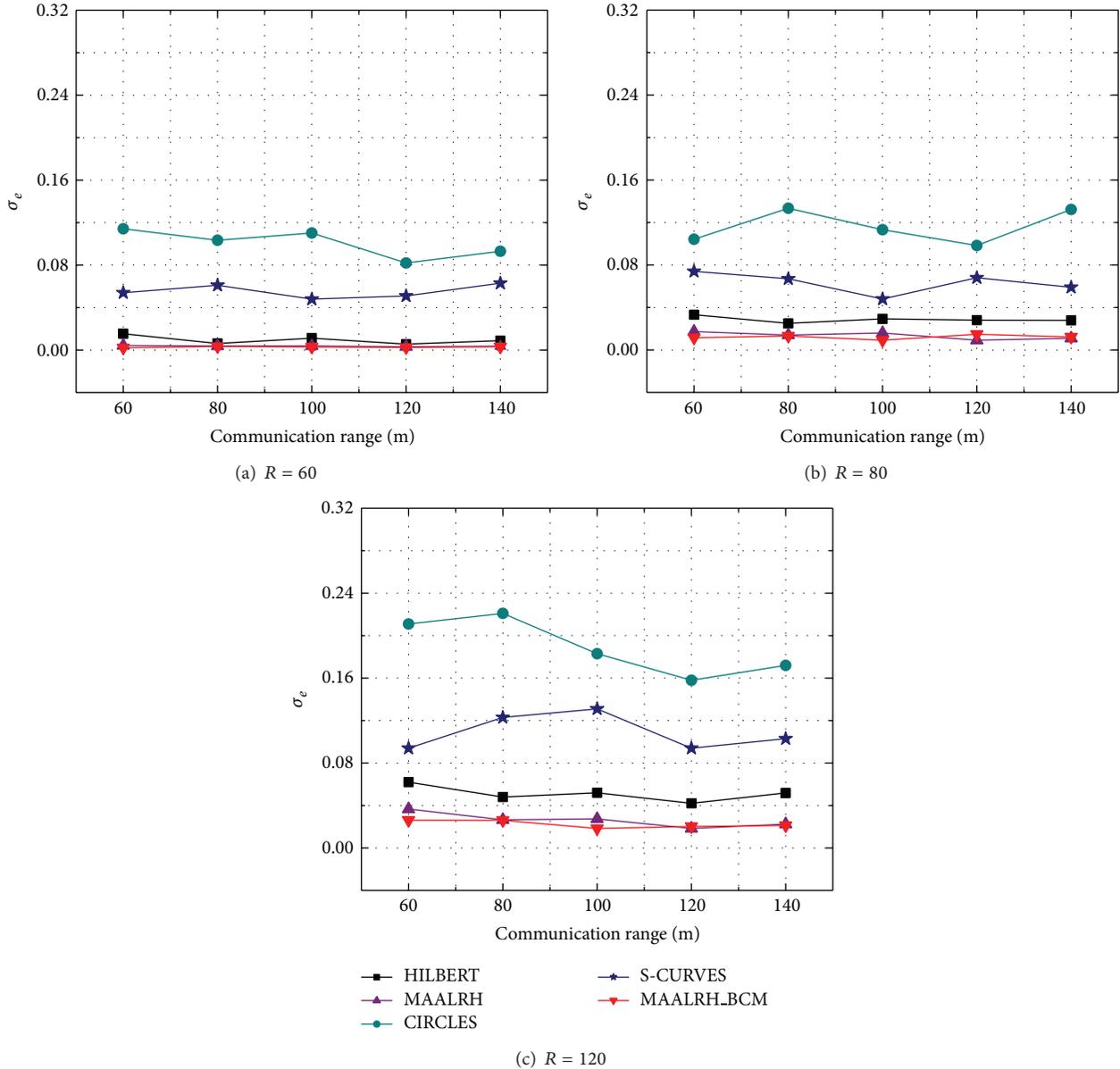


FIGURE 11: Standard deviation with different resolutions.

anchor points. However, the movement trajectory of the CIRCLES leaves four corners of the ROI uncovered, which results in larger average deviations compared with other four algorithms.

Figure 11 shows the standard deviation of the five mobile anchor assisted localization algorithms. MAALRH and MAALRH.BCM have the smallest  $\sigma_e$  compared with HILBERT, S-CURVES, and CIRCLES. From the simulation, we can draw the conclusion that the localization errors of MAALRH and MAALRH.BCM concentrate nearby the  $\mu_e$ , and the distribution of localization errors of HILBERT, S-CURVES, and CIRCLES is more dispersed than that of MAALRH and MAALRH.BCM, especially for CIRCLES.

5.3.3. Path Length. For each of the five movement trajectories, the path length is a function of  $n$  and  $R$ . From Figure 12

we can observe that the CIRCLES has the shortest path length and MAALRH, HILBERT, and S-CURVES have the similar path lengths. The MAALRH.BCM has the longest path length, because the mobile anchor moves in the sensing area which is  $(n + 2)^2 R^2 - n^2 R^2$  larger than the ROI to ensure that unknown nodes which are deployed at the boundary of the ROI could fall inside the overlapping communication ranges of at least three noncollinear anchors. Thus, the path length of the MAALRH.BCM is  $6nR + 18R$  longer than that of the MAALRH. The MAALRH.BCM sacrifices path length to maximize the localization ratio and the localization accuracy.

5.3.4. Scalability. We vary the number of the unknown nodes from 100 to 500 with the step of 100, communication range of 80 m, and resolution of 80 m to test the scalability of

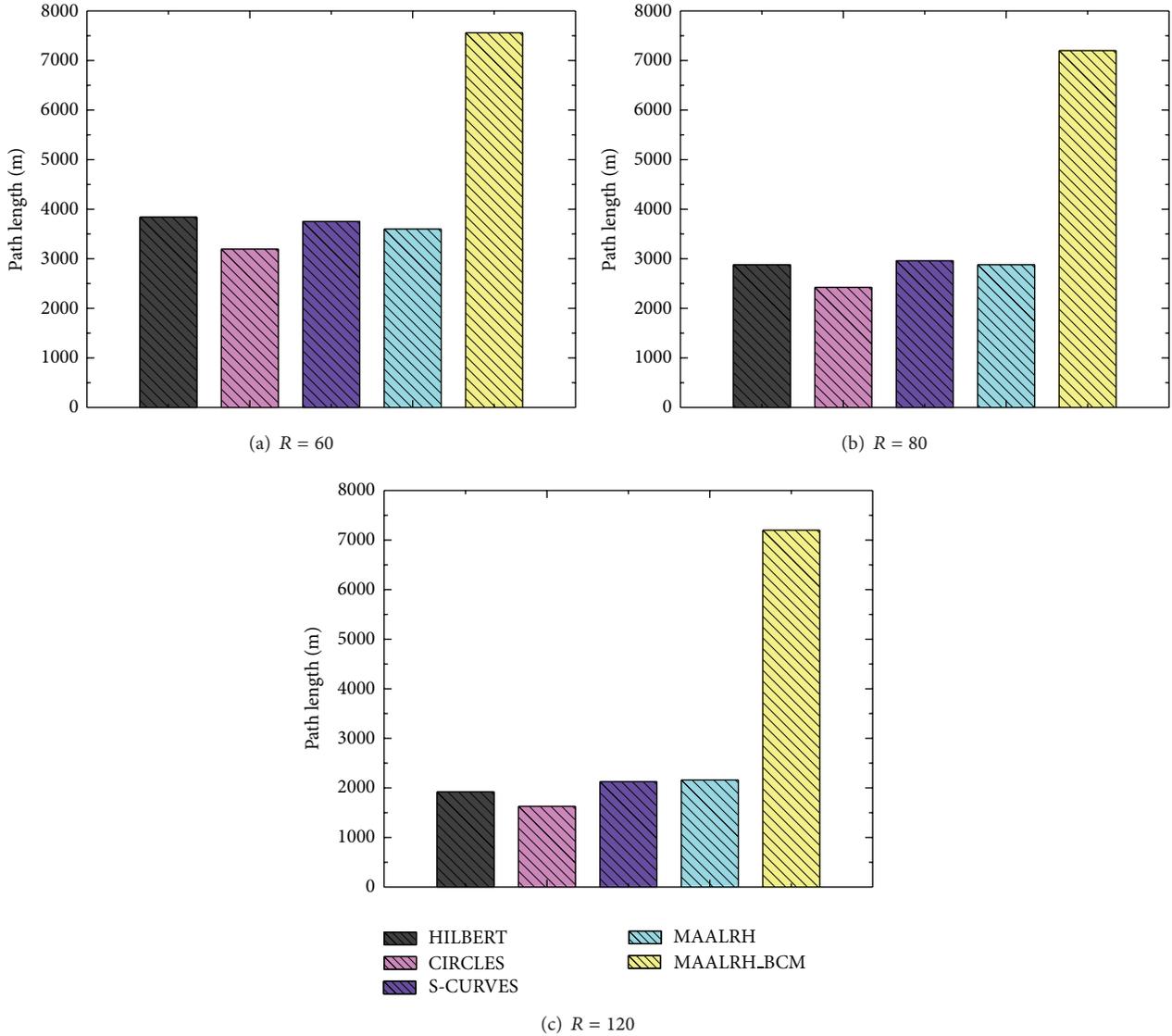


FIGURE 12: Path length with different resolutions.

the MAALRH\_BCM. Table 2 shows the relation between the localization ratio,  $\mu_e$ ,  $\sigma_e$ , and the number of unknown nodes. As depicted in Table 2, the localization ratio remains 100% with the increase of the number of unknown nodes; since the communication range is equal to the resolution, unknown nodes can receive three noncollinear anchor points which can form a regular triangle to estimate their positions.  $\mu_e$  and  $\sigma_e$  change litter under different unknown node densities, where the maximum  $\mu_e$  is 0.0214 m larger than the minimum  $\mu_e$  and the maximum  $\sigma_e$  is 0.0071 m larger than the minimum  $\sigma_e$ . This result proves that the localization ratio and localization accuracy of MAALRH\_BCM do not depend on the unknown node density; they depend on the sensor node's communication range and mobile anchor's trajectory resolution which determine the amount and the distance interval of anchor points.

TABLE 2: Scalability of MAALRH\_BCM.

	100	200	300	400	500
$L_{ratio}$	100%	100%	100%	100%	100%
$\mu_e$	0.1324	0.1538	0.1454	0.1459	0.1397
$\sigma_e$	0.0237	0.0264	0.0193	0.0240	0.0204

## 6. Conclusion

In this paper, we propose a mobile anchor assisted localization algorithm based on regular hexagon in two-dimensional WSNs, which can cover a square ROI entirely with a boundary compensation method. Simulations indicate that compared with HILBERT, CIRCLES, and S-CURVES algorithms, the MAALRH\_BCM can achieve higher localization ratio and

localization accuracy when the communication range is not smaller than the resolution. In summary, a carefully designed movement trajectory can significantly improve localization performances.

The future research issues in the area of mobile anchor assisted localization possibly are as follows.

- (i) In real applications, obstacle-resistant mobile anchor assisted localization algorithms are needed to deal with the obstacles in a given ROI. Movement trajectories of mobile anchors should be designed dynamically or partially according to the observable environment or deployment situations to make full use of the real-time information during localization.
- (ii) Single mobile anchor assisted localization algorithm takes a long time to locate all the unknown nodes in a ROI, especially for a large-scale WSN. Thus, collaborative mobile anchor assisted localization algorithm which uses several mobile anchors should be specifically designed to reduce localization time and improve localization accuracy.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The work is supported by the Natural Science Foundation of Jiangsu Province of China, no. BK20131137; the Applied Basic Research Program of Nantong Science and Technology Bureau, no. BK2013032; and the Guangdong University of Petrochemical Technology's Internal Project, no. 2012RC0106. Jaime Lloret's work has been partially supported by the "Ministerio de Ciencia e Innovacion," through the "Plan Nacional de I+D+i 2008–2011" in the "Subprograma de Proyectos de Investigacion Fundamental," Project TEC2011-27516. Joel J. P. C. Rodrigues's work has been supported by "Instituto de Telecomunicações," Next Generation Networks and Applications Group (NetGNA), Covilhã Delegation, by national funding from the Fundação para a Ciência e a Tecnologia (FCT) through the Pest-OE/EEI/LA0008/2013 Project.

## References

- [1] Y. Liu, Z. Yang, X. Wang, and L. Jian, "Location, localization, and localizability," *Journal of Computer Science and Technology*, vol. 25, no. 2, pp. 274–297, 2010.
- [2] H. Akcan, V. Kriakov, H. Brönnimann, and A. Delis, "Managing cohort movement of mobile sensors via GPS-free and compass-free node localization," *Journal of Parallel and Distributed Computing*, vol. 70, no. 7, pp. 743–757, 2010.
- [3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.
- [4] L. Gui, T. Val, and A. Wei, "A novel two-class localization algorithm in wireless sensor networks," *Network Protocols and Algorithms*, vol. 3, no. 3, pp. 1–16, 2011.
- [5] N. V. Doohan, S. Tokekar, and JitendraPati, "Mobility of sink using hexagon architecture in highly data centric wireless sensor networks," *International Journal of Scientific & Engineering Research*, vol. 3, no. 9, pp. 528–540, 2012.
- [6] S. Vupputuri, K. K. Rachuri, and C. Siva Ram Murthy, "Using mobile data collectors to improve network lifetime of wireless sensor networks with reliability constraints," *Journal of Parallel and Distributed Computing*, vol. 70, no. 7, pp. 767–778, 2010.
- [7] Y. Zeng, J. Cao, J. Hong, S. Zhang, and L. Xie, "Secure localization and location verification in wireless sensor networks: a survey," *Journal of Supercomputing*, vol. 64, no. 3, pp. 685–701, 2013.
- [8] G. Han, H. Xu, T. Q. Duong, J. Jiang, and T. Hara, "Localization algorithms of wireless sensor networks: a survey," *Telecommunication Systems*, vol. 52, no. 4, pp. 2419–2436, 2013.
- [9] A. Al-Fuqaha, "A precise indoor localization approach based on particle filter and dynamic exclusion techniques," *Network Protocols and Algorithms*, vol. 5, no. 2, pp. 50–71, 2013.
- [10] V. K. Chaurasiya, N. Jain, and G. C. Nandi, "A novel distance estimation approach for 3D localization in wireless sensor network using multi dimensional scaling," *Information Fusion*, vol. 15, pp. 5–18, 2014.
- [11] O. Diallo, J. J. P. C. Rodrigues, and M. Sene, "Real-time data management on wireless sensor networks: a survey," *Journal of Network and Computer Applications*, vol. 35, no. 3, pp. 1013–1021, 2012.
- [12] I. Amundson and X. D. Koutsoukos, "A survey on localization for mobile wireless sensor networks," in *Mobile Entity Localization and Tracking in GPS-less Environments*, vol. 5801 of *Lecture Notes in Computer Science*, pp. 235–254, 2009.
- [13] Y. Ding, C. Wang, and L. Xiao, "Using mobile beacons to locate sensors in obstructed environments," *Journal of Parallel and Distributed Computing*, vol. 70, no. 6, pp. 644–656, 2010.
- [14] H. Chenji and R. Stoleru, "Mobile sensor network localization in harsh environments," in *Distributed Computing in Sensor Systems*, vol. 6131 of *Lecture Notes in Computer Science*, pp. 244–257, Springer, Berlin, Germany, 2010.
- [15] A. N. Campos, E. L. Souza, F. G. Nakamura, E. F. Nakamura, and J. J. P. C. Rodrigues, "On the impact of localization and density control algorithms in target tracking applications for wireless sensor networks," *Sensors Journal*, vol. 12, no. 6, pp. 6930–6952, 2012.
- [16] C. Ou and W. He, "Path planning algorithm for mobile anchor-based localization in wireless sensor networks," *IEEE Sensors Journal*, vol. 13, no. 2, pp. 466–475, 2013.
- [17] D. Koutsonikolas, S. M. Das, and Y. C. Hu, "Path planning of mobile landmarks for localization in wireless sensor networks," *Computer Communications*, vol. 30, no. 13, pp. 2577–2592, 2007.
- [18] R. Huang and G. V. Záruba, "Static path planning for mobile beacons to localize sensor networks," in *Proceedings of the 5th Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops '07)*, pp. 323–328, White Plains, NY, USA, March 2007.
- [19] Z. Hu, D. Gu, Z. Song, and H. Li, "Localization in wireless sensor networks using a mobile anchor node," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '08)*, pp. 602–607, Xian, China, July 2008.

- [20] B. Zhang, F. Yu, and Z. Zhang, "Collaborative localization algorithm for wireless sensor networks using mobile anchors," in *Proceeding of the 2nd Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA '09)*, vol. 1, pp. 309–312, Wuhan, China, November 2009.
- [21] H. Cui, Y. Wang, and J. Lv, "Path planning of mobile anchor in three-dimensional wireless sensor networks for localization," *Journal of Information and Computational Science*, vol. 9, no. 8, pp. 2203–2210, 2012.
- [22] H. Cui and Y. Wang, "Four-mobile-beacon assisted localization in three-dimensional wireless sensor networks," *Computers and Electrical Engineering*, vol. 38, no. 3, pp. 652–661, 2012.
- [23] H. Li, J. Wang, X. Li, and H. Ma, "Real-time path planning of mobile anchor node in localization for wireless sensor networks," in *Proceedings of the IEEE International Conference on Information and Automation (ICIA '08)*, pp. 384–389, Changsha, China, June 2008.
- [24] Q. Fu, W. Chen, K. Liu, and X. Wang, "Study on mobile beacon trajectory for node localization in wireless sensor networks," in *Proceedings of the IEEE International Conference on Information and Automation (ICIA '10)*, pp. 1577–1581, Harbin, China, June 2010.
- [25] K. Kim, B. Jung, W. Lee, and D. Du, "Adaptive path planning for randomly deployed wireless sensor networks," *Journal of Information Science and Engineering*, vol. 27, no. 3, pp. 1091–1106, 2011.
- [26] S. Li, D. Lowe, X. Kong, and R. Braun, "Wireless sensor network localization algorithm using dynamic path of mobile beacon," in *Proceedings of the 17th Asia Pacific Conference on Communications (APCC '11)*, pp. 344–349, Sabah, Malaysia, October 2011.
- [27] K. Ssu, C. Ou, and H. C. Jiau, "Localization with mobile anchor points in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1187–1197, 2005.
- [28] Z. Guo, Y. Guo, F. Hong et al., "Perpendicular intersection: locating wireless sensors with mobile beacon," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 7, pp. 3501–3509, 2010.
- [29] E. Guerrero, H. G. Xiong, Q. Gao, G. Cova, R. Ricardo, and J. Estévez, "ADAL: a distributed range-free localization algorithm based on a mobile beacon for wireless sensor networks," in *Proceedings of the International Conference on Ultra Modern Telecommunications and Workshops*, pp. 1–7, Saint Petersburg, Russia, October 2009.
- [30] B. Xiao, H. Chen, and S. Zhou, "Distributed localization using a moving beacon in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 5, pp. 587–600, 2008.
- [31] S. Lee, E. Kim, C. Kim, and K. Kim, "Localization with a mobile beacon based on geometric constraints in wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 5801–5805, 2009.
- [32] G. Han, D. Choi, and W. Lim, "Reference node placement and selection algorithm based on trilateration for indoor sensor networks," *Wireless Communications and Mobile Computing*, vol. 9, no. 8, pp. 1017–1027, 2009.

## Research Article

# Vehicle Density Based Forwarding Protocol for Safety Message Broadcast in VANET

Jiawei Huang, Yi Huang, and Jianxin Wang

*School of Information Science and Engineering, Central South University, Changsha 410083, China*

Correspondence should be addressed to Jiawei Huang; [jiawei Huang@csu.edu.cn](mailto:jiawei Huang@csu.edu.cn)

Received 7 March 2014; Accepted 14 May 2014; Published 10 July 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Jiawei Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In vehicular ad hoc networks (VANETs), the medium access control (MAC) protocol is of great importance to provide time-critical safety applications. Contemporary multihop broadcast protocols in VANETs usually choose the farthest node in broadcast range as the forwarder to reduce the number of forwarding hops. However, in this paper, we demonstrate that the farthest forwarder may experience large contention delay in case of high vehicle density. We propose an IEEE 802.11-based multihop broadcast protocol VDF to address the issue of emergency message dissemination. To achieve the tradeoff between contention delay and forwarding hops, VDF adaptably chooses the forwarder according to the vehicle density. Simulation results show that, due to its ability to decrease the transmission collisions, the proposed protocol can provide significantly lower broadcast delay.

## 1. Introduction

During the last decade, the intelligent transportation systems (ITSs) have used advanced wireless communication technologies to enhance current road transportation systems. Most of ITS applications utilize vehicular ad hoc networks (VANETs) to provide the communications between vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) [1]. As a kind of self-organizing wireless networks, VANET chooses the suitable candidate vehicle as the forwarder to accomplish the multihop data delivery.

The accident detection and avoidance by disseminating safety messages are considered as one of the most important services of VANETs. When accident happens, the safety message should be broadcasted through the following vehicles (namely, a platoon) covering a specific area of few kilometers [2–4]. Lowering the dissemination delay between the time of an accident event and the time at which all vehicles of the following platoon receive the emergency message, the probability of chain collisions can be reduced. Since the wireless transmission range is limited (about 250 m), the vehicles have to relay the safety messages by multihop broadcasting. In the case of accident, there exist two issues. The first is how to accomplish the multihop broadcasting in

very tight message delivery time, typically few hundreds of milliseconds [5]. The second one is that the messages should be delivered to all vehicles with very high delivery reliability.

Due to the high mobility and restricted mobility patterns of vehicles, it is challenging work to design emergency message dissemination scheme with low delay and high reliability. Because the wireless channel is shared by all vehicles, the flooding broadcasting will lead to transmission contention and collision among neighboring vehicles, which degrades reliability and efficiency [6–8]. Many researchers have widely investigated the problem of disseminating safety messages in IEEE 802.11p protocol, which is MAC-layer standard in VANET [9]. Most of the proposed multihop broadcast protocols use the geographical information to avoid the broadcast storm problem. In these protocols, each vehicle is aware of its own position by the GPS devices. To reduce the number of forwarding hops in multihop broadcast, the farthest node in broadcast range is chosen as the forwarder. Consequently, the end-to-end broadcast delay is decreased.

In this paper, we use mathematical analysis and simulation results to demonstrate that, when the vehicle density is large, the farthest forwarder may experience large number of collisions, which bring about the high contention delay. Thus,

we propose an efficient vehicle density based forwarding (VDF) protocol for safety message dissemination in multihop VANETs. The protocol adaptably chooses the forwarder according to the vehicle density to obtain the good tradeoff between contention delay and forwarding hops.

The remainder of this paper is organized as follows. In Section 2, we present a brief overview of the related broadcast protocols. To investigate the problem of current protocols in detail, the mathematical analysis and simulation results are given in Section 3. The proposed protocol is described in Section 4. Section 5 compares and analyzes the protocol performances by the NS2 simulations. Finally, we draw conclusions in Section 6.

## 2. Related Work

In wireless networks, how to achieve high efficient multihop broadcast or coverage is a challenging task [10, 11]. In the multihop broadcasting, when a source vehicle broadcasts safety message, some of the vehicles within the vicinity of the source will become the next forwarding vehicles and perform relaying by rebroadcasting the message further. In the naive pure flooding scheme [12], each vehicle rebroadcasts the packet. It is obvious that, when the network becomes denser, the same message will be rebroadcasted more redundantly. The limited wireless channel bandwidth is wasted. Moreover, the packet collision problem becomes severe since a large number of vehicles in the same vicinity may rebroadcast the message at the same time. To solve the broadcast storm problem, the common method is adjusting the broadcast delay or probability. In this way, the channel contention brought by the flooding broadcast is mitigated.

Generally, in the delay-based broadcasting protocols, the farthest candidate forwarding vehicle is given the shortest broadcast delay. An efficient 802.11-based protocol called urban multihop broadcast (UMB) is proposed in [13]. UMB assigns each node with the specified broadcasting delay, which is determined by the distance between the vehicle and the transmitter. The lower broadcasting delay is assigned to the vehicle that is farther away from the transmitter. Therefore, the vehicle with the lowest delay has the highest priority to rebroadcast the message. At the same, when receiving the rebroadcasted message, the other vehicles cancel the retransmission process.

As another typical delay-based protocol, ReC [14] also uses geographical information to select the forwarding vehicles. In ReC, the selected forwarder is the nearest vehicle to the centroid of neighboring vehicles that have not received the message. Once receiving the message, the selected forwarder retransmits immediately. Thus, the unnecessary retransmission is reduced, while at the same time the forwarder can retransmit the message without delay. However, due to the vehicle's high mobility, there exists a great practical difficulty that ReC requires a complete and continuously updated knowledge of the neighboring vehicles.

To avoid the inaccurate transmission range estimation in highly mobile environment, JIVCA [15] utilizes the hello messages to get the real-time position information of other vehicles around. With the neighbor position information, JIVCA

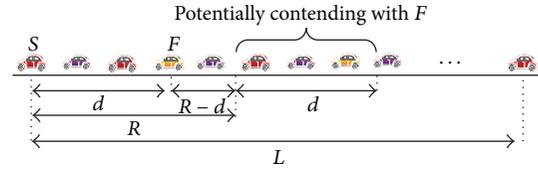


FIGURE 1: Network scenario.

continuously updates its transmission range estimation. Similar to the other delay-based protocols, JIVCA prioritizes farther vehicles in forwarding messages. In particular, the broadcast delay is computed based on the contention window (CW) in IEEE 802.11p MAC protocols. The farther vehicle has smaller CW and thus will have lower waiting delay before relaying the message.

In probabilistic-based broadcasting, a different rebroadcast probability is assigned to each receiving vehicle [16–18]. In [12], the weighted  $p$ -persistence protocol is proposed. The vehicle that receives message computes its own rebroadcast probability based on the distance between itself and the transmitter. The rebroadcast probability becomes larger as the distance between the vehicle and the transmitter increases. The farthest node from the sender has the highest chance of rebroadcasting the message firstly. Since only some vehicles will rebroadcast the message, the number of redundant messages and channel collisions is decreased.

To maximize the forwarding speed of safety messages, both delay-based and probabilistic-based broadcasting protocols give the highest priority to farthest vehicle to relay message. Though the number of forwarding hop between the sender and the last receiver is reduced, there exists the contention problem between the forwarder vehicle and the vehicles outside of the transmission range. Since the farthest vehicle is selected as the forwarder, when the forwarder rebroadcasts the message, large number of vehicles may contend with the forwarder, which enlarges the backoff delay for the usage of wireless channel. Furthermore, when the vehicle density increases, the contention delay will become larger and the end-to-end multihop broadcast performance is degraded.

## 3. Problem Analysis

In this section, we first describe the system assumption. Then, the mathematical model and analysis simulation are used to demonstrate contention problem when the farthest vehicle is selected as the forwarder.

The considered network scenario is in a multilane highway environment as shown in Figure 1. Since the transmission range  $R$  is much larger than the road width, the network scenario can be simplified as a one-dimensional VANET with road length of  $L$ . The vehicles are uniformly distributed on the road and the vehicle density is  $\alpha$ . In each relay, the distance between the sender  $S$  and forwarder  $F$  is hop distance  $d$ . We assume all of the vehicles are equipped with GPS to acquire their own positions.

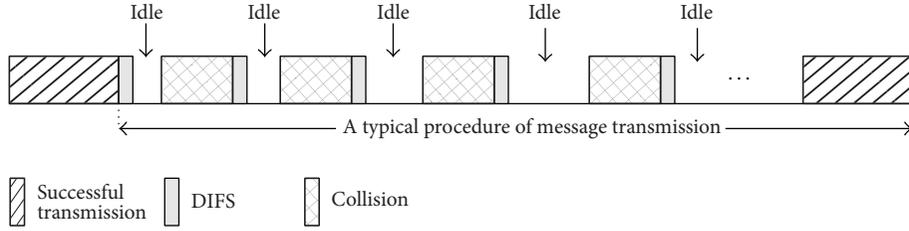


FIGURE 2: A typical procedure of message transmission.

Here, we analyze the end-to-end broadcast delay  $T$  considering the transmission range  $R$ , hop distance  $d$ , and vehicle density  $\alpha$ . We characterize transmission states of the forwarder  $F$ . As shown in Figure 1, since  $F$  is selected as the forwarder,  $F$  will rebroadcast the message first. Hence, we assume that there is no other contending vehicle in the transmission range  $R$  of sender  $S$ . However, the vehicles outside of the transmission range are potentially contending with  $F$ . The number of potential contending vehicles is calculated as  $\alpha d$ .

We model the backoff procedure of the IEEE 802.11p as a  $p$ -persistent CSMA/CA. Different from the binary exponential backoff in IEEE 802.11p, the backoff interval of the  $p$ -persistent CSMA/CA is sampled from a geometric distribution with transmission probability  $p$ . The  $p$ -persistent CSMA/CA provides a very close approximation to the IEEE 802.11 protocol [19, 20], and the memoryless backoff algorithm makes it suitable for mathematical analysis.

We observe the procedure before every successful message transmission. Figure 2 shows that collisions and idle periods may occur before a successful transmission. The collision's reason is that more than one vehicle transmits at the same time slot. The idle period is a time interval (expressed in number of time slots  $\sigma$ ) in which the transmission medium remains free of any transmission due to the backoff algorithm.

Assuming that  $F$  experiences  $n$  collisions before a successful transmission, therefore, we have the probability of successful transmission  $P_s$  and collision probability  $P_c$  as

$$P_s = \frac{\alpha d p (1-p)^{\alpha d - 1}}{1 - (1-p)^{\alpha d}}, \quad (1)$$

$$P_c = \frac{1 - (1-p)^{\alpha d} - \alpha d p (1-p)^{\alpha d - 1}}{1 - (1-p)^{\alpha d}},$$

where the transmission probability  $p$  can be calculated with the minimum value  $CW_{\min}$  of contention window as

$$p = \frac{2}{CW_{\min} + 1}. \quad (2)$$

From (1), the expected value  $E[N_c]$  of collision number  $n$  before a successful transmission can be obtained as

$$E[N_c] = \frac{P_c}{P_s} = \frac{1 - (1-p)^{\alpha d}}{\alpha d p (1-p)^{\alpha d - 1}} - 1. \quad (3)$$

For each transmission collision, the collision time  $T_c$  includes the message's transmission time  $\sigma m$  and the DIFS time  $\sigma D$ . Therefore, the total collision time  $T_{\text{col}}$  before a successful transmission is

$$T_{\text{col}} = E[N_c] T_c = \sigma \left[ \frac{1 - (1-p)^{\alpha d}}{\alpha d p (1-p)^{\alpha d - 1}} - 1 \right] (m + D). \quad (4)$$

As shown in Figure 2, since a collision is just between two idle periods, the expected number  $E[N_i]$  of idle period is

$$E[N_i] = E[N_c] + 1 = \frac{1 - (1-p)^{\alpha d}}{\alpha d p (1-p)^{\alpha d - 1}}. \quad (5)$$

The number of time slots  $T_i$  in each idle period is determined by the transmission probability  $p$  and contending vehicle number  $\alpha d$ . The expected value of  $E[T_i]$  is

$$E[T_i] = \sigma \left[ 1 - (1-p)^{\alpha d} \right] \sum_{i=0}^{\infty} i (1-p)^{\alpha d i} = \frac{\sigma (1-p)^{\alpha d}}{1 - (1-p)^{\alpha d}}. \quad (6)$$

Thus, the total collision time  $T_{\text{idle}}$  before a successful transmission can be calculated as

$$T_{\text{idle}} = E[N_i] E[T_i] = \frac{\sigma (1-p)}{\alpha d p}. \quad (7)$$

The successful transmission time  $T_{\text{trans}}$  is

$$T_{\text{trans}} = \sigma (m + D). \quad (8)$$

Since the one-hop transmission time  $T_{\text{hop}}$  is composed of the collision time  $T_{\text{col}}$ , idle time  $T_{\text{idle}}$ , and successful transmission time  $T_{\text{trans}}$ , we have

$$T_{\text{hop}} = T_{\text{col}} + T_{\text{idle}} + T_{\text{trans}}$$

$$= \frac{\sigma \left[ m + D - (m + D - 1) (1-p)^{\alpha d} \right]}{\alpha d p (1-p)^{\alpha d - 1}}. \quad (9)$$

From the source vehicle of message to the last receiver, the end-to-end multihop broadcast is made up of  $L/d$  relay hops. The end-to-end broadcast delay  $T$  can be expressed as

$$T = \frac{\sigma L \left[ m + D - (m + D - 1) (1-p)^{\alpha d} \right]}{\alpha d^2 p (1-p)^{\alpha d - 1}}. \quad (10)$$

TABLE 1: Road traffic parameters and MAC protocol setting.

Parameter and setting	Value
Vehicle density $\alpha$	25, 100, and 175 vel/km/lane
Vehicle speed $v$	80 km/h
Road length $L$	5000 m
Number of lanes	4
Transmission range $R$	300 m
Channel propagation	Two-ray ground
Time slot $\sigma$	20 $\mu$ s
DIFS time $\sigma D$	50 $\mu$ s
$CW_{\min}$	31 slots
$CW_{\max}$	1023 slots
Wireless transmission rate $r$	1 Mbps
Simulation time	70 sec

We validate our analysis using MATLAB and NS2 [21] simulations. We simulate a VANET scenario as shown in Figure 1, in which all vehicles are using IEEE 802.11p as the MAC-layer protocol. The whole 5 km road segment is composed of four lanes. Along each lane vehicles are uniformly deployed and moving at the constant velocity. The message size  $m$  is 1,000 Bytes, which corresponds to the transmission time of 32 time slots giving the 1 Mbps wireless transmission rate. Table 1 shows the road traffic parameters and MAC protocol settings, which are used for both the simulations and the analysis. It should be noted that the vehicle densities are set as 25, 100, and 175 vel/km/lane for different scenarios. The corresponding values of vehicle density  $\alpha$  are 0.1, 0.4, and 0.7 vel/m, respectively.

In Figure 3, we change the hop distance  $d$  from 5 m to 290 m. The larger hop distance, the farther vehicle is chosen as forwarder. From both mathematical and simulation results, we observe that the end-to-end broadcast delay  $T$  is sensitive to the vehicle density  $\alpha$ . When the vehicle density is small ( $\alpha = 0.1$  vel/m), the end-to-end broadcast delay always decreases if the hop distance becomes larger. The reason is that when the contention is low, the message dissemination will be accelerated if the farthest vehicle is the forwarder. However, when the vehicle density is large, the channel contention for the forwarder becomes heavy if the forwarder is still the farthest vehicle. This, as a consequence, results in high message collision ratio and long contention delay. Figure 3 shows that, when  $\alpha$  is 0.7 vel/m, the end-to-end broadcast delay becomes very high if the hop distance is large.

Based on this observation, the optimal hop distance between the sender  $S$  and forwarder  $F$ ,  $d_{\text{opt}}$ , which minimizes the value of end-to-end broadcast delay  $T$ , can be obtained by equating the first derivative of  $T$  with respect to  $d$  to zero. As presented in the following section, in our proposed protocol, the  $d_{\text{opt}}$  value is used to tune the contention window size to reach the desired performance.

#### 4. Vehicle Density Based Forwarding

The basic ideal of our proposed VDF protocol is to select the forwarder  $F$  with the optimal hop distance  $d_{\text{opt}}$  according

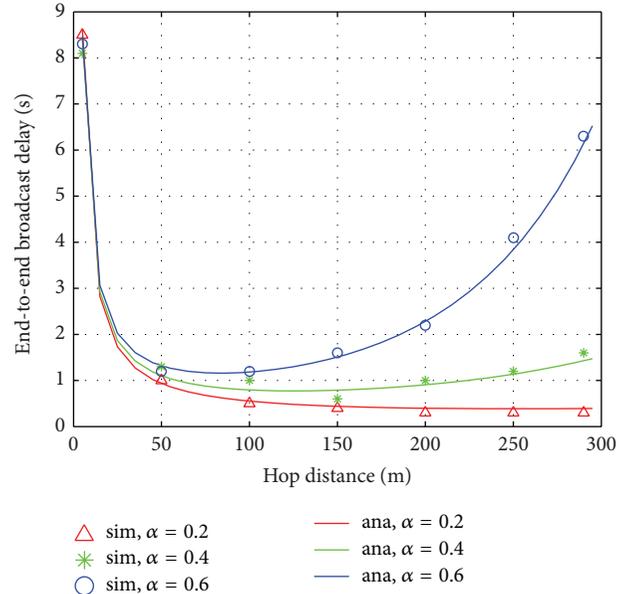


FIGURE 3: End-to-end broadcast delay with different hop distance.

to the vehicle density  $\alpha$ . In the following, we present the design detail of VDF protocol with the pseudocode shown as Pseudocode 1.

In order to sense the vehicle density  $\alpha$  in the transmission range  $R$ , each vehicle utilizes the beacon message to inform its neighboring vehicles. The information in beacon message includes the vehicle's identity and its position. When receiving the beacon message, the vehicle counts the vehicle number in its transmission range and then calculates the vehicle density  $\alpha$ . Moreover, the vehicle can calculate the distance  $d_f$  from the current forwarder to itself with the help of GPS devices. As analyzed in the previous section, the given values of  $\sigma$ ,  $L$ ,  $m$ ,  $D$ ,  $p$ , and  $\alpha$ ,  $d_{\text{opt}}$  can be numerically computed. It should be noted that the maximum value of  $d_{\text{opt}}$  is the transmission range  $R$ .

By assigning different waiting times from the reception to rebroadcasting of the message, VDF prioritizes the best relaying vehicle with hop distance  $d_{\text{opt}}$  in forwarding the message. The waiting time is determined by the contention window in IEEE 802.11p MAC protocols.

Upon receiving the new safety message from the forwarder, each vehicle computes its own contention window  $CW$  as

$$CW = \left\lfloor \frac{d_f - d_{\text{opt}}}{R} \right\rfloor \times (CW_{\max} - CW_{\min}) + CW_{\min}, \quad (11)$$

where  $CW_{\min}$  and  $CW_{\max}$  are the minimum and maximum contention window, respectively. From (11), it is clear that the vehicle with smaller value of  $|d_f - d_{\text{opt}}|$  will have a smaller  $CW$ . That means that the vehicle will have shorter waiting time to transmit the message and, implicitly, higher probability to be selected as the forwarder.

During the waiting, the vehicle may receive the same message again. Then the vehicle stops trying to forward the message, because another vehicle with shorter waiting

```

Parameters:
{
  R: The transmission range;
  CWmax: The maximum contention window;
  CWmin: The minimum contention window;
  neighbour_list: The vehicle set in the transmission range;
  nv: The vehicle number in neighbour_list;
  F: The forwarder vehicle;
  df: The distance from F to the current vehicle;
}
Initialization:
{
  nv = 0;
  F = ∅;
  df = 0;
  neighbour_list = ∅;
}
On receiving a beacon from vehicle v:
{
  if v is not in neighbour_list then
  {
    nv = nv + 1;
    add nv into neighbour_list;
    update the vehicle density α;
  }
  if v is the forwarder F then
  update df as the distance from F to the current vehicle;
}
On receiving a message from the forwarder F:
{
  if the message is new then
  {
    calculate CW with (11);
    start the waiting timer;
  }
  else
  stop the waiting timer;
}
On the waiting time expire:
{
  broadcast the received message;
  the current vehicle becomes the forwarder F;
}

```

PSEUDOCODE 1: Pseudocode of VDF.

time already did it. If the waiting time expires without having received the same message from any other vehicle, the vehicle becomes the forwarder and rebroadcasts the message.

## 5. Performance Evaluation

In this section, we compare the performances of VDF with ReC and 802.11p by using NS2 simulator. As mentioned before, 802.11p protocol employs random backoff scheme without any distance prioritization. In the following, we give the performance metrics, simulation setup, and performance analysis.

### 5.1. Performance Metrics

**5.1.1. Broadcast Delay.** The main focus of our proposed protocol lies in reducing broadcast delay, which is a crucial factor in time-critical safety applications. The first performance metric is message broadcast delay, which is defined as the dissemination delay of the safety message from the source vehicle to the last receiver. The faster the safety message propagates, the more efficient the corresponding protocol is in terms of satisfying the urgent delay requirement of emergency application.

**5.1.2. Broadcast Count.** The message dissemination progress can be measured by the broadcast count, which includes

both the success and failure broadcast. The larger hop distance implies the high channel conflict probability and thus larger number of failure broadcast. However, the smaller hop distance means small coverage range and more relay hops. Only by obtaining good tradeoff between the coverage range and the rebroadcast probability could the fast message dissemination progress be achieved.

**5.2. Simulation Setup.** In the performance evaluation, we model a straight highway with 4 unidirectional lanes, where the vehicles are uniformly deployed. All vehicles send the beacon messages to periodically announce their ID and position with the generation rate of 10 beacon/s. The beacon size is set to 100 B. Moreover, in order to simulate the communication traffic such as web chat applications, we assume all the vehicles send 1.5 KB data packets to their neighbors with the rate of 10 packets/s. The other road traffic parameters and MAC protocol settings are the same as that in Table 1.

To compare the various schemes, we simulate two typical network applications including accident alert and online game. In the accident alert application, the messages are generated only in the case of the abnormal behavior of some vehicles. Specifically, the simulations of this application include single source (Section 5.3) and multiple sources configurations (Section 5.4). In the single source case, only the first vehicle is chosen as the emergency message source. On the other hand, in the multiple sources configuration, a different number of vehicles are randomly selected as the message sources. During simulation, the source vehicles broadcast the emergency message backward to all the following vehicles along the 5 km highway.

The online game is one of the most popular applications in VANET. In the online game application, each player periodically generates multihop transmissions to the other players. The wireless channel is shared by all players in the same application. Moreover, the generation rate is very important for the player experience. Thus, we vary the transmission interval to test the performances under different congestion state of wireless channel (Section 5.5). In the simulations, each test is repeated 10 times. The average value of test results is calculated with 95% confidence intervals.

**5.3. Accident Alert: Impact of Vehicle Density.** We start our evaluation focusing on the performances with varying vehicle density. In this test, the vehicle density is increased from 50 to 250 vehicles per km. Moreover, only the first vehicle is the emergency message source in this test.

Figure 4 shows the message broadcast delay against node density. Initially, when the vehicle density is only 50 vehicles per km, all three protocols obtain small delay. VDF experiences the delay close to ReC but 0.03 s smaller than 802.11p. The performance gap with 802.11p is the consequence of the fact that 802.11p randomly selects the forwarder and results in the larger average hop distance between the forwarders compared with ReC and VDF. As a result, the dissemination speed is reduced by 802.11p.

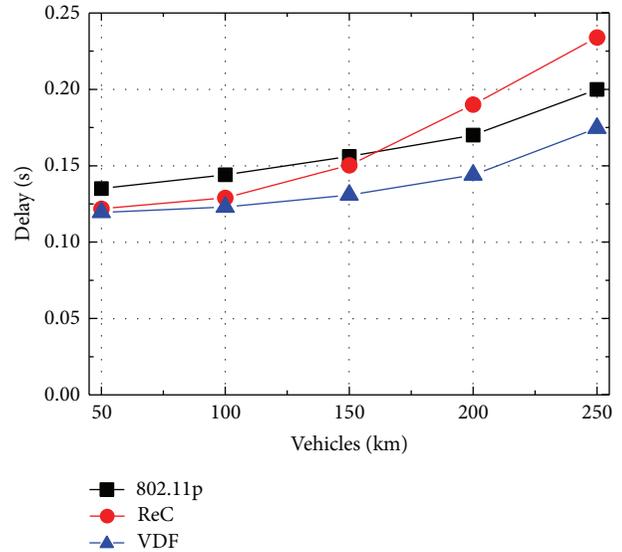


FIGURE 4: Accident alert: delay with different vehicle density.

When the vehicle density increases, the message broadcast delay becomes higher with the larger conflict probability in wireless channel. However, the delay increasing of ReC is much faster than the other two protocols. When the vehicle density is 250 vehicles per km, the delay of ReC is about 34% and 15% higher than that of VDF and 802.11p, respectively. This result is explained by the fact that the larger the vehicle density, the higher the possibility of transmission collision as a single transmission range area hosts more vehicles. Though ReC selects the farthest vehicle as the forwarder to obtain the maximum coverage, it devotes more time to the collision resolution that incurs longer backoff time and thus large delay. Among the three protocols, VDF achieves the best delay performance as it adjusts the hop distance according to the vehicle density. When the vehicle density increases, VDF reduces the hop distance between the forwarders and alleviates the impact of high channel collision rate.

The results of broadcast count required to cover all vehicles are shown in Figure 5. 802.11p has the highest broadcast count as it randomly selects the forwarder vehicle and needs more relay hops to transmit the message to last vehicles. In contrast, with the target of providing the maximum coverage, the broadcast count of ReC is about 20% less than that of 802.11p. Compared with ReC, VDF obtains the nearly same performance though it does not always choose the farthest vehicle as the forwarder. Moreover, when the vehicle density is larger than 150 vehicles per km, VDF is even slightly lower than ReC. This is because that, when the channel contention becomes heavier, VDF reduces the coverage range and then gets the lower rebroadcast probability. Though the relay hops of VDF are more than ReC, VDF still obtains the smaller broadcast count because of the lower number of rebroadcasts.

**5.4. Accident Alert: Impact of Source Number.** In this evaluation, we test the protocol performances in the multiple sources configuration. 5% to 25% of randomly selected

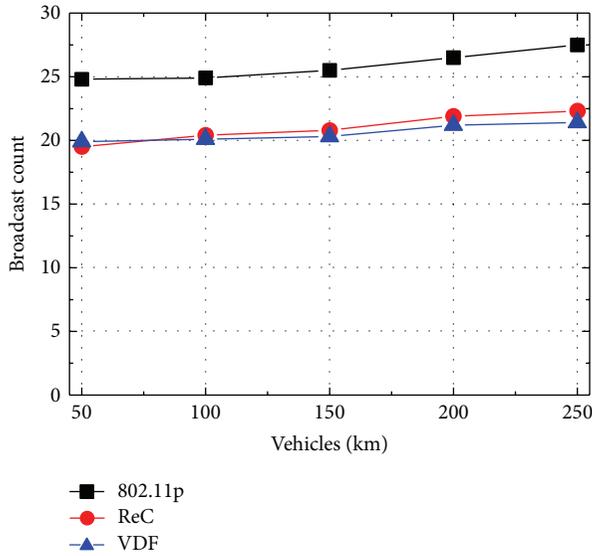


FIGURE 5: Accident alert: broadcast count with different vehicle density.

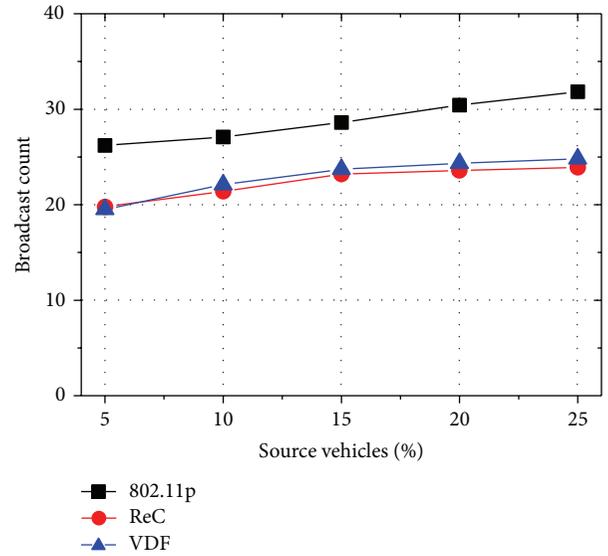


FIGURE 7: Accident alert: broadcast count with different number of source vehicles.

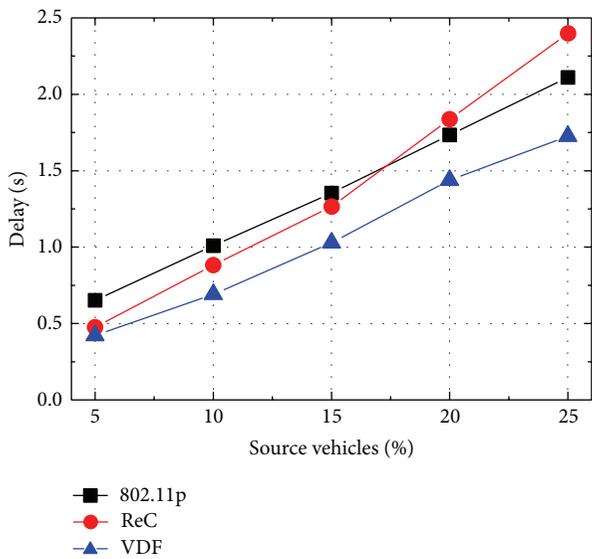


FIGURE 6: Accident alert: delay with different number of source vehicles.

vehicles act as the message sources, which send messages backward to cover all vehicles. The vehicle density is fixed as 125 vehicles per km.

In Figure 6, the message broadcast delay is shown for different number of message sources. For each protocol, the broadcast delay becomes larger with the increasing of number of message sources. This result is supported by the requirement of intensive collision resolution phases with increase in message sources. Moreover, under the heavy channel contention, it is seen that all the broadcast protocols exhibit much larger delay compared with the results of single source configuration. Because of the fixed hop distance in broadcast relay, ReC has the fastest increasing speed of

message broadcast delay among the three protocols. Different from ReC, VDF selects the small coverage range to avoid the intensive channel contention. Therefore, VDF gets the best delay performance in three protocols.

The comparison results for the number of broadcast count are shown in Figure 7. Once again, 802.11p suffers from the highest message broadcast count due to random manner in selecting broadcast forwarders. Both ReC and VDF obtain the lowest broadcast count by considering the coverage speed. It is noticed that, when the number of message sources becomes larger, since VDF experiences a smaller number of rebroadcast, VDF gets a slightly lower broadcast count than ReC.

**5.5. Online Game.** In the test of online game application, the vehicle density is set as 100 vehicles per km. We randomly select 50 vehicles as the players, which periodically generate 2000-Byte-sized packets to the other players. We evaluate the delay performances considering different generation intervals with each player. Specifically, the packets are generated at each vehicle every 10, 50, or 100 ms.

We measure the average value of game event delay, which is defined from the time that the player sends packet to the time that all the other players receive the packet. Figure 8 presents the delay results with different generation intervals. It could be observed that, when the generation rate is 10 ms, the game event delay of all protocol is larger than 3.5 s. These results are attributed to the high packet generation rate and multitude of sources. If the generation interval is 50 ms or 100 ms, the game event delay drops to below 0.8 s. Compared with the case of 50 ms, the game event delay of 100 ms interval is only slightly lower. This indicates that the generation interval is not small enough to saturate the wireless channel. As expected, in the three protocols, VDF gets the lowest delay with all different generation intervals.

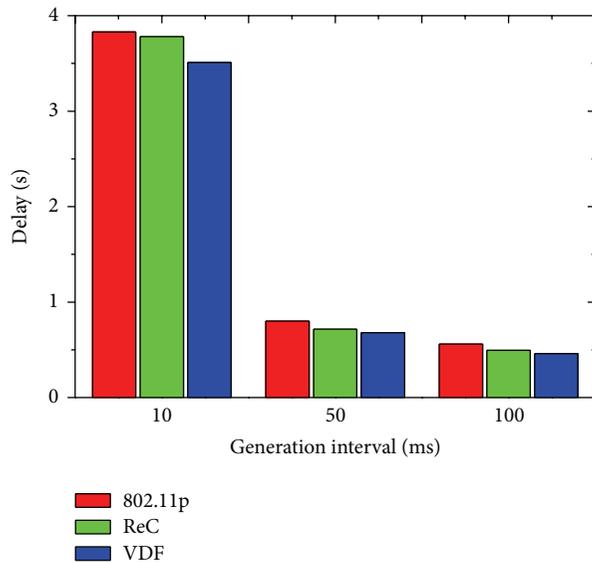


FIGURE 8: Online game: delay with different generation interval.

This is a logic consequence of the fact that VDF decreases the coverage range to adapt to the high transmission interference.

## 6. Conclusion

We design and implement VDF, a vehicle density based forwarding protocol for safety message broadcast in VANET. By adaptively selecting the forwarder node according to the vehicle density, VDF alleviates the heavy wireless channel contention. Thus, VDF achieves the low broadcast delay and small broadcast count in multihop broadcast. By using NS2 simulations, we show that VDF has better performance than the existing message broadcast protocols in two typical network applications including accident alert and online game.

In the future, in order to avoid the impact of highly dynamic circumstances, we will design the backoff algorithm based on motion prediction of vehicle nodes. Moreover, we will test the protocol performance with large scale of testbed experiment.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant nos. 61103204, 61163060). The authors greatly appreciate the anonymous reviewers for their insightful and helpful comments.

## References

- [1] S. Panichpapiboon and W. Pattara-Atikom, "A review of information dissemination protocols for vehicular ad hoc networks," *IEEE Communications Surveys and Tutorials*, vol. 14, no. 3, pp. 784–798, 2012.
- [2] "Safespot," <http://www.safespot-eu.org>.
- [3] "Communications for eSafety," <http://www.comesafety.org/>.
- [4] "eSafety," <http://www.esafetysupport.org>.
- [5] H. A. Omar, W. Zhuang, and L. Li, "VeMAC: a TDMA-based MAC protocol for reliable broadcast in VANETs," *IEEE Transactions on Mobile Computing*, vol. 12, no. 9, pp. 1724–1736, 2013.
- [6] M. Asefi, J. W. Mark, and X. Shen, "A mobility-aware and quality-driven retransmission limit adaptation scheme for video streaming over VANETs," *IEEE Transactions on Wireless Communications*, vol. 11, no. 5, pp. 1817–1827, 2012.
- [7] Z. Liao, J. Wang, S. Zhang, and X. Zhang, "A deterministic sensor placement scheme for full coverage and connectivity without boundary effect in wireless sensor networks," *Ad-Hoc and Sensor Wireless Networks*, vol. 19, no. 3-4, pp. 327–351, 2013.
- [8] J. Huang, J. Wang, and J. Ye, "A buffer management algorithm for improving up/down TCP fairness in IEEE 802. 11 WLANs," *International Journal of Communication Systems*, 2012.
- [9] "IEEE, 1609: Family of Standards for Wireless Access in Vehicular Environments (WAVE)," <http://www.standards.its.dot.gov>.
- [10] W. Luo, J. Wang, J. Guo, and J. Chen, "Parameterized complexity of Max-lifetime Target Coverage in wireless sensor networks," *Theoretical Computer Science*, vol. 518, pp. 32–41, 2014.
- [11] J. Wang, W. Luo, Q. Feng, and J. Guo, "Parameterized complexity of Min-power multicast problems in wireless ad hoc networks," *Theoretical Computer Science*, vol. 508, pp. 16–25, 2013.
- [12] B. Williams and T. Camp, "Comparison of broadcasting techniques for mobile ad hoc networks," in *Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC '02)*, pp. 194–205, New York, NY, USA, June 2002.
- [13] G. Korkmaz, F. Özgüner, E. Ekici, and Ü. Özgüner, "Urban multi-hop broadcast protocol for inter-vehicle communication systems," in *Proceedings of the 1st ACM International Workshop on Vehicular Ad Hoc Networks (VANET '04)*, pp. 76–85, October 2004.
- [14] J. Liu, Z. Yang, and I. Stojmenovic, "Receiver consensus: on-time warning delivery for vehicular ad-hoc networks," in *Proceedings of the 32nd IEEE International Conference on Distributed Computing Systems (ICDCS '12)*, pp. 386–395, Macau, China, June 2012.
- [15] C. E. Palazzi, M. Roccetti, and S. Ferretti, "An intervehicular communication architecture for safety and entertainment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 90–99, 2010.
- [16] L. Zhou, Y. Zhang, K. Song, W. Jing, and A. V. Vasilakos, "Distributed media services in P2P-based vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 692–703, 2011.
- [17] A. Wegener, H. Hellbrück, S. Fischer, C. Schmidt, and S. Fekete, "AutoCast: an adaptive data dissemination protocol for traffic information systems," in *Proceedings of the 66th Vehicular Technology Conference (VTC '07)*, pp. 1947–1951, Baltimore, Md, USA, October 2007.

- [18] M. Slavik and I. Mahgoub, "Stochastic broadcast for VANET," in *Proceedings of the 7th IEEE Consumer Communications and Networking Conference (CCNC '10)*, pp. 1–5, Las Vegas, Nev, USA, January 2010.
- [19] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Transactions on Networking*, vol. 8, no. 6, pp. 785–799, 2000.
- [20] D.-J. Deng, C.-H. Ke, H.-H. Chen, and Y.-M. Huang, "Contention window optimization for ieee 802.11 DCF access control," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5129–5135, 2008.
- [21] "The Network Simulator," <http://www.isi.edu/nsnam/ns/>.

## Research Article

# Exploring a QoS Driven Scheduling Approach for Peer-to-Peer Live Streaming Systems with Network Coding

Laizhong Cui,<sup>1</sup> Nan Lu,<sup>1</sup> and Fu Chen<sup>2</sup>

<sup>1</sup> College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

<sup>2</sup> Department of Computer Science and Technology, Beijing Foreign Studies University, Beijing 100089, China

Correspondence should be addressed to Laizhong Cui; [cuilazhong@gmail.com](mailto:cuilazhong@gmail.com)

Received 11 April 2014; Accepted 18 June 2014; Published 10 July 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 Laizhong Cui et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Most large-scale peer-to-peer (P2P) live streaming systems use mesh to organize peers and leverage pull scheduling to transmit packets for providing robustness in dynamic environment. The pull scheduling brings large packet delay. Network coding makes the push scheduling feasible in mesh P2P live streaming and improves the efficiency. However, it may also introduce some extra delays and coding computational overhead. To improve the packet delay, streaming quality, and coding overhead, in this paper are as follows. we propose a QoS driven push scheduling approach. The main contributions of this paper are: (i) We introduce a new network coding method to increase the content diversity and reduce the complexity of scheduling; (ii) we formulate the push scheduling as an optimization problem and transform it to a min-cost flow problem for solving it in polynomial time; (iii) we propose a push scheduling algorithm to reduce the coding overhead and do extensive experiments to validate the effectiveness of our approach. Compared with previous approaches, the simulation results demonstrate that *packet delay*, *continuity index*, and *coding ratio* of our system can be significantly improved, especially in dynamic environments.

## 1. Introduction

P2P has become a dominant solution for distributing live video content to large populations of users in recent years, by leveraging clients' resources to serve each other. To provide the robustness and meet the streaming bandwidth requirement, most existing large-scale P2P live streaming systems organize peers into mesh. However, the streaming quality of them is not so satisfactory, especially in dynamic environments [1]. The performance bottleneck is due to the lack of a proper and optimal scheduling design.

Given a number of neighbors, a peer needs to decide which packets are transmitted by which neighbors, which is called scheduling. Existing scheduling can be broadly divided into two categories: pull and push. In the pull scheduling [2], the streaming is divided into blocks and a data structure called buffer map is periodically exchanged to reflect a peer which has blocks in its cache. According to the received buffer maps, each peer explicitly requests the desired blocks from its neighbors using some pull strategies, such as rarest-first

or sequence-first. However, the *packet delay* is very large for this receiver-driven delivery and the bandwidth is not fully utilized since each block is only served by a peer at a time. To reduce the *packet delay*, the tree push scheduling [3] is introduced. Nevertheless, this strategy is not suitable for large-scale P2P live streaming. The reason lies in the following aspects: (1) the tree structure has large maintenance and repair costs in dynamic P2P environment; (2) since the leaf peers of the tree will not deliver content to any other peers, the bandwidth of the leaf peers is wasted. In the mesh push scheduling [4], the video streaming is divided into some substreams, and each peer reassembles all the substreams through receiving packets pushed by different neighbors. This method cannot solve the problems about performance degradation when the intensive system dynamic happens.

Network coding has been shown to be an effective way to improve the performance of P2P streaming by maximizing the network throughput and making the push scheduling feasible in mesh P2P [5]. Moreover, a missing segment of a peer can be served by multiple neighbors simultaneously. Mea

and Baochun proposed a random push scheduling for network coding based P2P live streaming system called  $R^2$  [6]. Through reducing the complexity of coordinated scheduling and improving the bandwidth resources,  $R^2$  improves the system performance. However, they do not optimize the QoS metrics of transmission performance and the streaming quality so that they do not reach the optimal solution. Moreover, in  $R^2$ , before pushing a block, a peer has to produce a new coding block, which brings a lot of coding overhead since the operation of encoding consumes the computational resources. To generate new coding blocks by reencoding, a peer must receive enough coding blocks, which increases the extra packet delay. Since current scheduling cannot fully take advantages of network coding, it is necessary to redesign a QoS driven push scheduling approach for network coding based P2P live streaming system.

In this paper, we propose a novel QoS driven scheduling approach for network coding based P2P live streaming system with the following contributions. First, we introduce a new coding method, through combing the substreams with network coding. Second, we formulate the scheduling optimization problem and transform it to an equivalent min-cost flow problem for solving it in polynomial time. Third, we design a simple yet effective push scheduling algorithm to reduce the coding overhead and improve the robustness in dynamic environments. To evaluate the performance and effectiveness, we implement our approach on an event-driven P2P streaming simulator [7] and compare it with CoolStreaming [2] and  $R^2$  [6]. The simulation results show that our approach improves the transmission performance and streaming quality by reducing the *packet delay* and improving the *continuity index*. Meanwhile, the coding overhead is lower by reducing the *coding ratio*, which reflects the better robustness of our approach in dynamic environments.

The rest of the paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we present the details of our analysis model and introduce a new coding method. We propose the optimization and algorithm of our scheduling approach in Section 4. The simulation results are discussed in Section 5. Finally, Section 6 concludes our work.

## 2. Related Work

The tree-based P2P live streaming systems, such as [3, 8, 9], could reduce the playback delay. In the tree topology, the root is the streaming servers and all other peers are organized into one or more multicast trees. The streaming content of the server is decomposed into substreams that are pushed through corresponding trees from the server to all nodes in that tree. Although such tree-based push scheduling algorithm is beneficial in reducing the delays of transmitting data, they are not suitable to deploy in real-world large-scale streaming systems. The main reason is its complexity and cost involved in maintaining the tree topology in dynamic P2P environment.

CoolStreaming [2] is a mesh based P2P live streaming system and it can serve large-scale users. In the mesh based systems, the streaming content is divided into a series of

segments and each represents a short duration content of playback. A new concept, called buffer map, is introduced to represent the segment available information of each peer. To know which segments each neighbor has, the buffer map is periodically exchanged among peers. CoolStreaming proposes a rarest-first pull scheduling algorithm, which means the rarest segment among its neighbors is transmitted preferentially. Although the systems with the mesh topology and pull scheduling algorithm are more robust to peers dynamic than the systems with the tree topology and push scheduling algorithm, they inevitably increase the delay of data transmission from servers to all participating peers. These delays mainly come from the periodic exchange of buffer map and explicit segment request. Zhang et al. [4] propose Grid Media to improve the delay of CoolStreaming, which is also a mesh based P2P live streaming system. It combines a hybrid scheduling algorithm, consisting of the pull and push modes. In Grid Media, a peer requests the streaming packets with the pull mode at startup and then relays streaming packets in the push mode. Essentially, to utilize the push mode, the streaming content is divided into multiple substreams, each of which is pushed in a different tree structure. However, it also needs to exchange the segment available information among peers and is not robust due to the dynamic environment.

For dealing with the defects of mesh based P2P streaming system, some coding methods are introduced into P2P streaming system, the most representatives of which are rateless codes and network coding. The rateless fountain codes, including LT codes [10], Raptor codes [11], and online codes [12], can be readily used in peer-to-peer streaming with substantial advantage. The typical P2P system with rateless codes is rStream. In rStream [13], the Raptor codes are used in P2P streaming system to eliminate the coordination of the content available information. The peer selection and rate allocation are formulated as an optimization problem and the algorithm is also proposed to solve the optimization problem. Although it improves the end-to-end latency, it neglects the other QoS metrics, such as throughput and redundancy.

Recently, network coding [14–16] has been widely used to improve the performance of P2P systems. Gkantsidis and Rodriguez [17, 18] have proposed that randomized network coding can significantly reduce the downloading times in P2P content distribution and file downloading systems. Lava [5] fairly evaluates the feasibility and effectiveness of random network coding [19] for P2P live streaming systems. While Lava has focused on a fair comparison study without improving the P2P live streaming traditional mechanism, the advantages of network coding have not been fully explored. Inspired by Lava, Mea and Baochun [6] redesign the scheduling algorithm and propose a random push with random network coding scheduling algorithm called  $R^2$  to take full advantage of network coding. The random push scheduling algorithm of  $R^2$  is revised to be suitable for UUSee [20], which is a popular P2P VoD system. It demonstrates that network coding with random push scheduling algorithm can also improve the P2P VoD system. Sarkar and Wang [21] give the details of the setup through a measurement study of network coding

in real P2P VoD system. Sarkar and Wang [22] propose a prefetch strategy for network coding based P2P VoD systems. Nguyen and Nakazato [23] discuss that the rare-first scheduling algorithm is not enough for P2P streaming with network coding. Sheikh et al. [24] propose a distributed media-aware scheduling algorithm for P2P streaming with network coding, which considers the feedback information of neighbors, including loss rate and decoding ratio.

Although the network coding in P2P streaming is effective and practical, the research of the scheduling algorithm in network coding based P2P streaming systems is still an open research area, especially in optimizing several QoS metrics. Hsu [25] produces a knowledge sharing method and Mishra and Srivastava [26] discuss the information spreading behavior in the distributed systems, both of which enlighten our work indirectly. Our idea of QoS driven scheduling algorithm is partly inspired by the previous research, but we introduce a new coding method, formulate an optimization for the scheduling problem, and propose the corresponding distributed solution.

### 3. The Analysis Model and Coding Method

We let  $R$  bits/s be the streaming rate of the live stream. To realize the push scheduling in mesh [4], we also divide the live stream into several substreams. Let each substream's rate be  $\gamma$ . It means the live stream is divided into  $N$  substreams,  $N = R/\gamma$  (assuming that  $R$  is divisible by  $\gamma$ ). On the other hand, as for the traditional P2P live streaming with network coding, the live stream is divided into several segments, and each segment is divided into several blocks. The network coding is only used in each segment to generate coding blocks, without encoding blocks across different segments.

In our approach, we propose a network coding method by combining the substream with network coding, called coding substream. The details of the coding method of coding substream are described as follows. We let the live stream be divided into segments. Each segment has a sequence number called *segment\_id* and is divided into  $M$  blocks further. The network coding is used in each segment to generate  $M'$  coding blocks. We directly utilize random network coding and progressive decoding method [5]. Whenever a peer wants to encode a block, it first independently and randomly chooses a set of coding coefficients  $[e_1, e_2, \dots, e_m]$  in the Galois field  $GF(2^8)$  and then produces a coding block  $x$ , using the following equation:

$$x = \sum_{i=1}^M e_i \cdot b_i. \quad (1)$$

When a peer receives  $M$  linearly independent coding blocks  $x = [x_1, x_2, \dots, x_M]$ , it can decode the original segment as follows. It extracts the coefficients of each encoded block  $x_i$  to form the  $M \times M$  coefficient matrix  $E$ . Then, it recovers the original segment  $b = [b_1, b_2, \dots, b_M]$  as (2). We utilize Gaussian elimination [5] to solve this equation. Consider the following:

$$b = E^{-1} x^T. \quad (2)$$

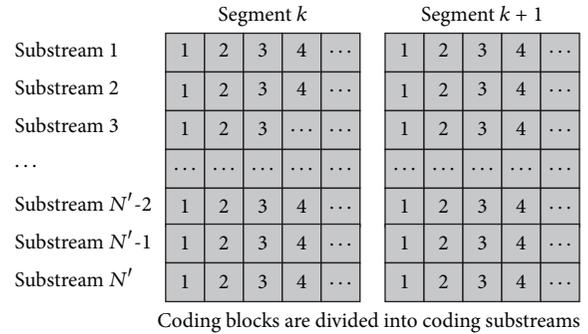


FIGURE 1: The coding method and the coding substreams.

The  $M'$  coding blocks are interleaved into  $N'$  coding substreams,  $N' = \lfloor B_s/\gamma \rfloor$ ,  $M' = M * (N'/N)$  ( $B_s$  is the upload bandwidth of source server). Each substream has a sequence number called *sub-stream\_id*. Each coding block in one coding substream has a sequence number called *block\_id*. Thus, each coding block is identified by a triple,  $\langle segment\_id, sub-stream\_id, block\_id \rangle$ . A diagram of this coding method and coding substreams is shown in Figure 1.

We take an example to explain this method. Assume that  $N = 3, M = 6$ , and  $B_s = 2R$ , which means the original segment includes 6 blocks and each segment is divided into 3 substreams; namely, each substream has 2 blocks. To generate the coding substreams, it will produce  $M' = 2M = 12$  coding blocks (with the *block\_id* 1,2,3,11,12), for a segment and these 12 coding blocks are divided into  $N' = 2N = 6$  coding substreams, that is, coding substreams 1 with coding blocks {1, 7}, coding substreams 2 with coding blocks {2, 8}, ..., and coding substreams 6 with coding blocks {6, 12}.

In our approach, the network coding operation is also employed both on the source server side and on the peers side. However, the encoding operations mainly happen on the server and occasionally happen on the peers when necessary (this situation will be discussed later), which is different from the traditional network coding based P2P streaming systems, such as Lava [5] and  $R^2$  [6]. This design could effectively reduce the coding overhead.

In the traditional push scheduling and non-network coding systems [4], each peer needs to collect all  $N$  substreams for smooth playback. This brings the limitation that the peers may fail to get enough substreams since the system is highly dynamic. However, in our approach, we apply the network coding to the  $M$  original blocks of a segment for producing  $M'$  coding blocks with little probability of duplication, which also means that it can produce more available substreams, that is, from  $N$  substreams to  $N'$  substreams. On the one hand, our approach increases the diversity of the content and robustness of the system. On the other hand, it decreases the complexity of the substreams scheduling. Each peer could subscribe to any  $N$  substreams from these  $N'$  substreams. As long as the peer collects  $M$  linearly independent coding blocks from  $N$  substreams, it can decode the original content.

We formulate this simple coding substreams scheduling as the following model. In the transmission process based on coding substreams, each peer subscribes to  $N$  coding

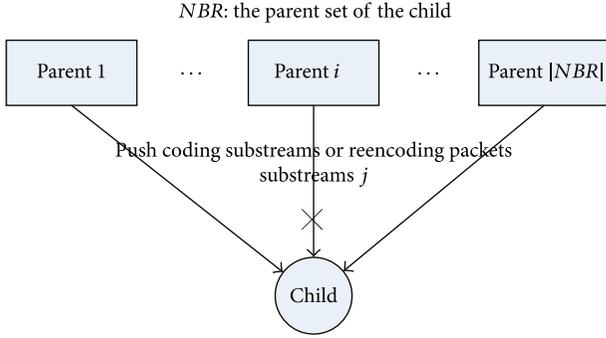


FIGURE 2: The scheduling process of a child and its parents.

substreams via its neighbors. Once a peer receives coding blocks from its subscribed substreams, it also relays the content to its downstream neighbors that requested the corresponding substreams. We can decompose the transmission process of the whole system into several transmission units. A typical transmission unit is illustrated in Figure 2, in which the focused peer is called child and the peers serving substreams are called parents. Without loss of generality, we can solve the substreams scheduling problem by focusing on a certain child and its parents in Figure 2. The child has a set of parents denoted by  $NBR$ . For each parent  $i$  in  $NBR$ , it will allocate a certain upload bandwidth  $B_i$  bits/s to the child for the coding substreams transmission. Let  $h(i, j) \in \{0, 1\}$  denote whether parent  $i$  could provide substream  $j$ .  $h(i, j)$  has value 1 if parent  $i$  could provide the substreams  $j$  and 0 otherwise. When a child first joins the system, it requests a neighbor list as its parent set  $NBR$  from the bootstrap server. Actually, any neighbor selection algorithm can be used in our approach. After obtaining the parent set, the child asks each parent for their buffer information about the coding substreams (i.e., the vector  $[h(i, j)] = (h(i, 1), h(i, 2), \dots, h(i, N'))$ ). Upon receiving each parent's vector  $[h(i, j)] = (h(i, 1), h(i, 2), \dots, h(i, N'))$ , the child solves the scheduling problem about arranging the coding substreams that means it needs to decide to subscribe to which coding substream via which parent for obtaining the  $N$  coding substreams.

#### 4. The Optimization and Algorithm for the Scheduling Problem

We first formulate the scheduling problem as a cost optimization problem and introduce a polynomial time solution in Section 4.1. And then, we give the details of the push scheduling algorithm of our approach in Section 4.2.

**4.1. The Optimization of the Scheduling Problem.** Let  $o_i$  be the maximal substreams which can be pushed to the child from parent  $i$ ,  $i \in NBR$ . It is calculated by (3), in which  $B_i$  is the upload bandwidth that parent  $i$  allocated to the child and  $\gamma$  is the streaming rate of each coding substream. Consider

$$o_i = \left\lfloor \frac{B_i}{\gamma} \right\rfloor. \quad (3)$$

Before formulating the optimal scheduling problem, we define a cost function  $C(i, j)$  as (4), which represents the average transmission delay of a packet considering the packet loss probability when the substream  $j$  is assigned to parent  $i$ . The larger  $C(i, j)$  means higher transmission delay.  $\rho_{i,j}$  represents the child's packet loss probability of coding substream  $j$  from parent  $i$ .  $D_{i,j}$  denotes the child's link latency of substream  $j$  from parent  $i$ . We use an example to explain our design of the cost function  $C(i, j)$ . For two parents  $i$  and  $k$ , we assume that  $D_{i,j} = 3$  seconds,  $D_{k,j} = 1$  second,  $\rho_{i,j} = 0.1$ ,  $\rho_{k,j} = 0.3$ , and 10 packets are transmitted in each sending period. For parent  $i$ , the child receives 9 packets in 3 seconds, namely, using 1/3 second per packet, and for parent  $k$ , the child receives 7 packets in 1 second, namely, using 1/7 second per packet. Apparently, the child should prefer to subscribe to coding substream  $j$  via parent  $k$ . Consider the following:

$$C(i, j) = D_{i,j} \times \frac{1}{\rho_{i,j}}. \quad (4)$$

Given the coding substreams vector  $[h(i, j)] = (h(i, 1), h(i, 2), \dots, h(i, N'))$  from parent  $i$  and the maximal substreams allocated to the child  $o_i$ ,  $\forall i \in NBR$ , the goal of the scheduling problem is to find a solution of subscribing to which coding substream via which parent, achieving the minimum total transmission cost of the  $N$  subscribed coding substreams for the child. We formulate it as the optimization scheduling problem with some given restrictions in the following equation:

$$\begin{aligned} & \text{Minimize} && \sum_{j=1}^{N'} \sum_{i \in NBR} x_{ij} h(i, j) C(i, j) \\ & \text{subject to} && \text{(a) } x(i, j) \in 0, 1, \quad i \in NBR, \\ & && \quad \quad \quad j \in \{1, 2, \dots, N'\} \\ & && \text{(b) } \sum_{i \in NBR} x(i, j) = 1, \quad \forall j \\ & && \text{(c) } \sum_{j=1}^{N'} x(i, j) \leq o_i, \\ & && \text{(d) } \sum_{j=1}^{N'} \sum_{i \in NBR} x_{ij} = N. \end{aligned} \quad (5)$$

From constraint (a),  $x_{ij}$  is a decision variable, which has value 1 if the substream  $j$  is assigned to parents  $i$  and 0 otherwise. It indicates this scheduling optimization is a 0-1 programming problem. The constraint (b) means the child can only subscribe to one coding substream via one parent. The constraint (c) means parent  $i$ 's number of subscribed coding substreams must be smaller than its upper bound  $o_i$ . The constraint (d) means the child only needs  $N$  coding substreams.

The classical min-cost flow problem could be formulated as (6). The min-cost flow problem is a well-known optimization problem. Since the min-cost flow problem is a convex

problem, which could be used by several algorithms to get the solution in polynomial time. Consider

$$\begin{aligned} \min \quad & \sum_{(i,j) \in A} c(i,j) x(i,j) \\ \text{subject to} \quad & \text{(a) } \sum_{j:(i,j) \in A} x(i,j) - \sum_{j:(j,i) \in A} x(j,i) = d_i, \\ & \forall i \in V \\ & \text{(b) } 0 \leq x(i,j) \leq u(i,j), \quad \forall (i,j) \in A. \end{aligned} \quad (6)$$

For solving this optimization problem in polynomial time, we propose some transformation rules to transform this scheduling optimization problem to an equivalent min-cost flow problem. The transformation rules are described in Table 1, which includes two aspects: vertexes and edges. The key idea of our transformation rules lies in two aspects: (1) we use two type vertexes to represent the parents and coding substreams; (2) besides the edges between the vertex  $p_i$  and vertex  $ss_j$ , the cost of other edges is 0.

Figure 3 shows the transform result. We use the double scaling algorithm [27] to solve this min-cost flow problem in polynomial time. In the optimal solution, the flow amount on edges  $(p_i, ss_j)$  is the value of  $x(i, j)$ . With the optimal solution, we can also get the scheduling decision of the child; that is, for each  $x(i, j) = 1$ , the substream  $j$  is assigned to parent  $i$ .

**4.2. Push Scheduling Algorithm.** We also design a push scheduling algorithm to ensure the coding blocks transmission of substreams with low delay and overhead. As traditional P2P live streaming, each peer has a buffer with a limited constant length. When a child receives a coding block from the subscribed coding substream, it puts this block into its buffer and does one step progressive decoding operation by applying Gauss-Jordan elimination [5], which can reduce the decoding time. For serving the child, a parent has two modes to push its content: Forwarding Push and Reencoding Push. According to the status of child's buffer, the parent makes a choice from these two modes.

**4.2.1. Forwarding Push.** The parent directly forwards the coding block in its buffer to the child only if the child lacks the block in the corresponding and subscribed substreams. When a child subscribes to a coding substream via a parent, it should inform the largest *block.id* of the block it has received of each segment in that substream. If the parent has some blocks, the *block.id* of which is larger than its child's largest *block.id*, it can directly forward them to its child and at the same time it updates the child's largest *block.id*.

**4.2.2. Reencoding Push.** If the parent cannot find any block in its buffer that can be directly forwarded to the child (it means that the child has received the whole coding blocks of its subscribed substreams from this parent), it will try to produce a new coding block through reencoding the received coding blocks in its buffer and push the new coding block to the child. To ensure the new produced coding block is linearly

independent of the child, we import the concept of *coding aggressiveness*  $\alpha$  ( $0 < \alpha < 1$ ) [5]. A parent can produce a new coding block by re-encoding the received coding blocks only if the percentage of received coding blocks in that segment is larger than  $\alpha$  (the segment is called segment exists). After producing a new coding block by re-encoding, the parent will push it to the child. Upon receiving this block, the child will insert it into a missing substream and assign it with the largest *block.id* in that substream.

This design of Reencoding Push makes our system more robust in dynamic environment. We explain it through an example in Figure 2. Assume the child subscribes to coding substream  $j$  via parent  $i$  and the parent  $i$  suddenly leaves the system. At this time, the child needs to search for another parent by contacting the bootstrap server or any other membership mechanisms. Anyway, this parent repair process will take some time. During this period, the child may not receive the content in time and have to suffer from the incomplete streaming that is decreasing the streaming quality. However, with Reencoding Push, the child's other parents will discover coding substream  $j$  is missing. Although the child does not subscribe to the coding substream  $j$  via any parent of them, they will reencode to produce new coding blocks, in order to supply the missing substream  $j$ . After receiving enough reencoding blocks, the child could decode the original content and provide smooth playback. Besides, in traditional push-mesh system, the child has to find a new parent which can provide the exact same substream  $j$ . In our approach, the child need not find the exact same coding substream  $j$ , and it needs only to find a coding substream  $j'$ , which has not been subscribed before. Since the coding substream  $j'$  is linear independence with the remaining  $N - 1$  coding substreams of the child, it can replace substream  $j$  to be used for decoding and therefore the parent repairing process could be shortened effectively.

We summarize our QoS driven scheduling approach by describing the flowchart as in Figure 4. In the beginning of a new child joining the system, it contacts the server to get the neighbor list as its parent set. Then it asks for parents' buffer information (i.e., the vector  $[h_{i,j}]$ 's). Upon receiving the vector  $h_{i,j}$  from all parents, the child computes a new scheduling for these substreams, through solving the optimization as (5). The child recomputes the scheduling when it meets large changes about network conditions, such as the departure of parents or congestion in a certain connection. After deciding the substreams subscription, the child will receive content from parents and push the content to its child via two modes of our push scheduling algorithm.

## 5. Evaluation

**5.1. Simulation Setup and Metrics.** We utilize a discrete event-driven packet level simulator [7] and realize the network coding operation and our scheduling approach on it. We conduct a series of extensive simulations to study the impacts of our scheduling approach. For comparison, we simulate two conventional systems: the classic system CoolStreaming

TABLE I: Transformation rules.

Transformation rules of vertexes	
(1)	We add two virtual vertexes, source vertex $S$ and terminal vertex $T$ .
(2)	We add $ NBR $ parent vertexes, each of which is represented by $p_i, i \in NBR$ .
(3)	We add the vertex $ss_j$ to express the substream $j, j \in \{1, 2, \dots, N'\}$ .
(4)	We add the vertex $C$ to express the child node, which makes the scheduling strategy.
Transformation rules of edges	
(5)	We add the edge between vertex $S$ and vertex $p_i$ , the cost of which is 0 and capacity of which is $o_i$ .
(6)	For $\forall j \in \{1, 2, \dots, N'\}$ , if $h(i, j) = 1$ , we add the edge between the vertex $p_i$ and vertex $ss_j$ , the cost of which is $C(i, j)$ and the capacity of which is 1.
(7)	For $\forall j \in \{1, 2, \dots, N'\}$ , we add the edge between the vertex $ss_j$ and the vertex $C$ , the cost of which is 0 and the capacity of which is 1.
(8)	We add the edge between vertex $C$ and vertex $T$ , the cost of which is 0 and capacity of which is $N$ .

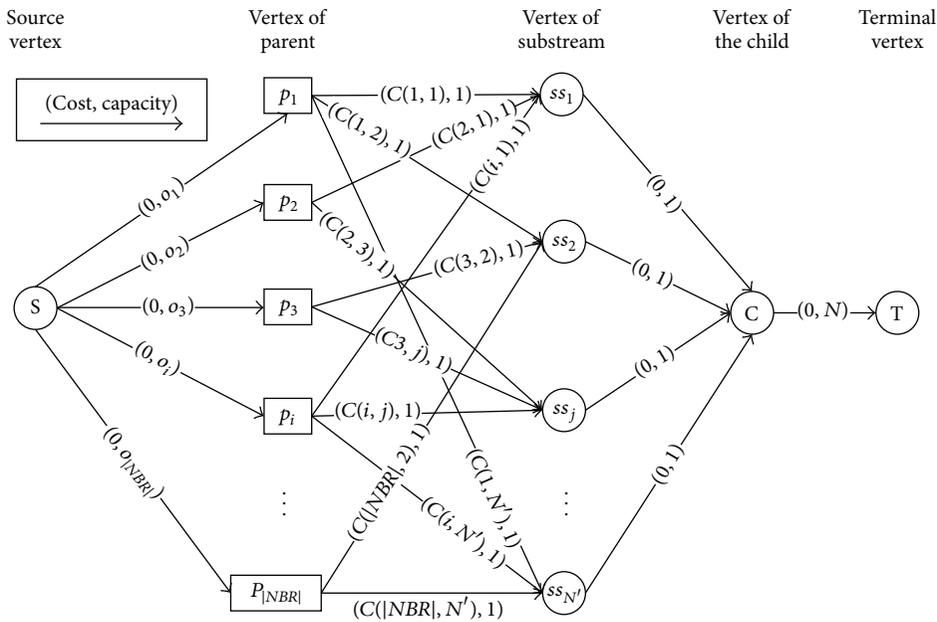


FIGURE 3: A min-cost-transformation example.

[2] with traditional pull and state-of-the-art network coding based system  $R^2$  [6] with random push.

In our simulation, all streaming and control packets including the sending and receiving buffer of each peer are carefully simulated. In all experiments, unless specified otherwise, we set the original streaming rate  $R = 400$  Kbit/s and the substream rate  $\gamma = 50$  Kbit/s. Each segment represents 1 second of the playback, which means each segment is 400 Kbit/s long. The segments are divided into  $M = 320$  blocks. According to literature [5], the *coding aggressiveness*  $\alpha = 0.5$ . Each peer has 15 parents. The upload capacity of the source server is  $B_s = 5$  Mbit/s, which is a reasonable ratio in practice. The default number of peers is 200. The default parameters of the simulated system are set as Table 2. We employ real-world end-to-end latency matrix ( $2500 \times 2500$ ) measured on Internet [28] and the transmission loss rate for data packets between two peers as uniformly distributed from 0.02 to 0.1, which are typical in Internet. To simulate the bandwidth heterogeneity of peers, we define three different

typical ADSL peers. The default of the peer bandwidth distribution is presented in Table 3, which is measured on Internet [29].

Besides the simulation with default parameters values in Table 2, we also change the peer number and streaming rate to do extra experiments. As the values of the peer number and streaming rate change, the peer bandwidth distribution has to be reset, in order to meet the increase of system demand for peers bandwidth. The details of the corresponding peer bandwidth distribution are described in Table 4.

We focus on the following metrics in our evaluation.

*Packet Delay.* It refers to the delay between the time when the packet is sent out from the source server and when it is received at a peer after several hops.

*Continuity Index.* It is defined as the fraction of the segments that could be received and decoded before their playback deadlines.

TABLE 2: The default parameters of the simulation.

Category	Parameter value
Peer number	200
Substream rate $\gamma$	50 Kbit/s
Streaming rate $R$	400 Kbit/s
Segment length	1 s (320 blocks)
Parent count	15
Upload bandwidth of streaming server $B_s$	5 Mbit/s
Coding aggressiveness $\alpha$	0.5

TABLE 3: Peer bandwidth distribution.

Category	Downlink	Uplink	Ratio (default)
A	768 Kbit/s	128 Kbit/s	30%
B	1.5 Mbit/s	384 Kbit/s	40%
C	3 Mbit/s	1 Mbit/s	30%

**Coding Ratio.** It is defined as the percentage of the transmitted blocks produced by encoding operation over all the transmitted blocks that are sent by peers.

### 5.2. Simulation Results

**5.2.1. Packet Delay.** Figure 5 shows the average *packet delay* of the three systems versus the number of peers. Generally, the average *packet delay* increases with the number of peers, since the network scale of systems becomes large. The average *packet delay* of CoolStreaming is the largest, as the pull scheduling algorithm accumulates large *packet delay* along the transmission path. Since network coding makes the push scheduling feasible in mesh P2P systems,  $R^2$  reduces *packet delay*. However, it uses random push scheduling without considering the link transmission delay and the frequent reencoding operations make the *coding ratio* too high (in Figure 9). So, its *packet delay* is still larger than ours. Our system achieves the smallest *packet delay*. The reasons are that our scheduling optimization chooses the parents with low link transmission delay to push coding substreams in a timely mode and our scheduling algorithm reduces the reencoding operations as much as possible.

The average *packet delay* of the three systems versus streaming rate is illustrated in Figure 6. In general, the average *packet delay* becomes larger with the increase of streaming rate, as the quantity of transmitted data increases. The reason is that with larger streaming rate, a segment has more blocks and it has to take more times to receive these blocks and decode them. Overall, as the streaming rate increases, although the packet delay inevitably increases, our approach can keep it at a more reasonable and lower level than those of CoolStreaming and  $R^2$ , which means our algorithm has better scalability.

**5.2.2. Continuity Index.** The streaming quality is measured by *continuity index*. Figure 7 shows the average *continuity index* of the three systems versus simulation duration. To simulate the dynamic environment, we let 50 peers and 20

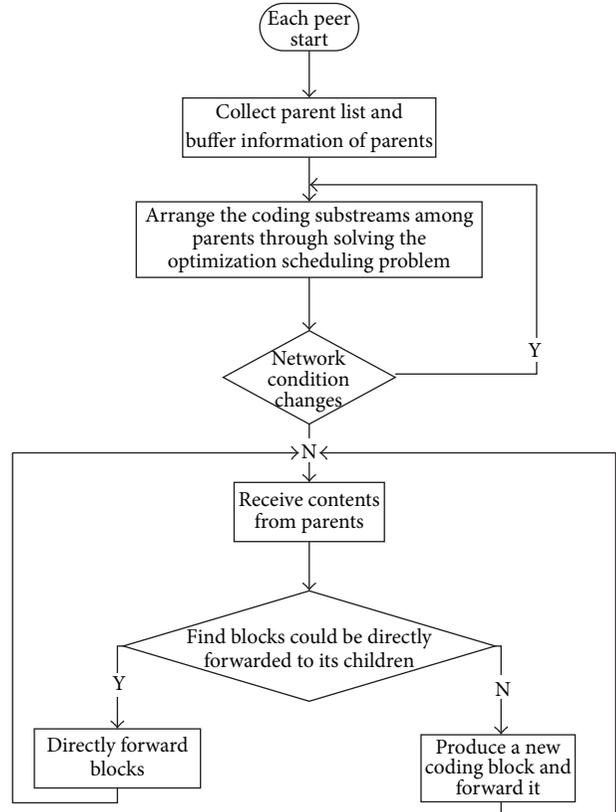


FIGURE 4: The flowchart of our QoS driven scheduling approach.

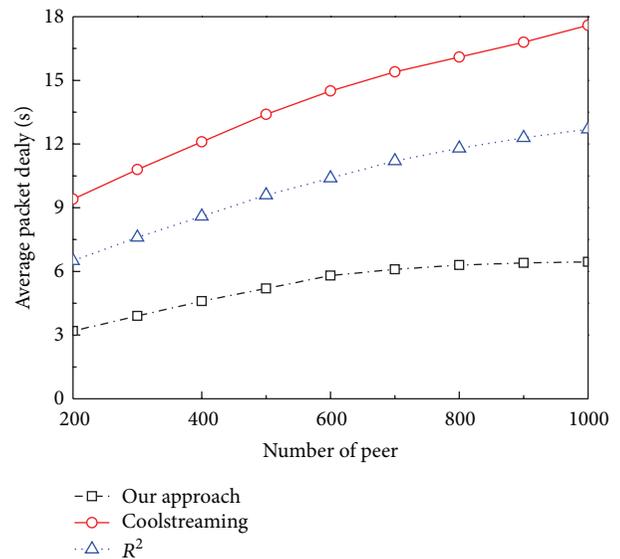
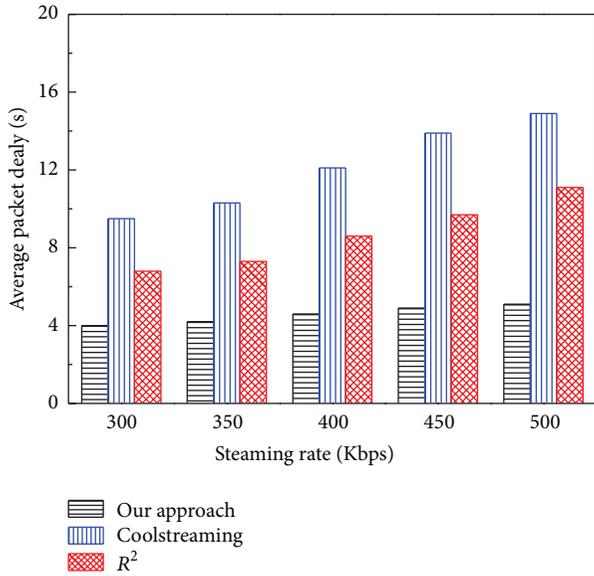


FIGURE 5: Average *packet delay* versus number of peers.

peers leave the system, respectively, at 500 s and 800 s of the duration. Our system achieves the highest *continuity index*. The *continuity index* of ours has the least reduction, which means our systems have the best robustness in dynamic environment. The reason mainly lies in two aspects. (i) Our approach considers the packet loss probability

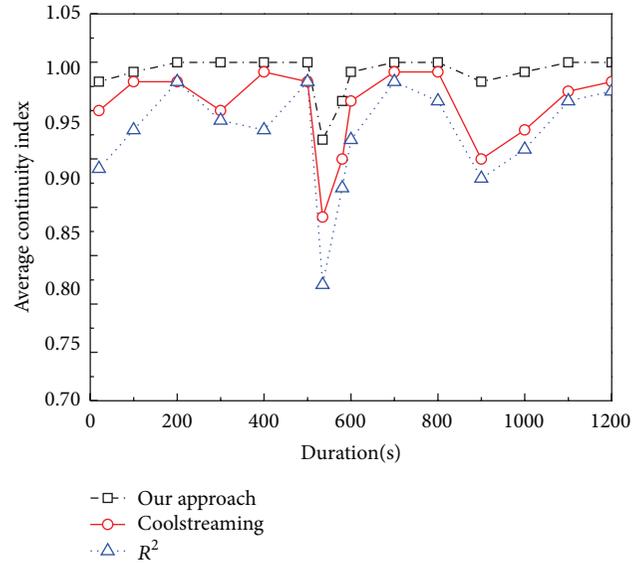
TABLE 4: Peer bandwidth distribution versus peers number and streaming rate.

The variation of peers number		The variation of streaming rate	
Peer number	Category ratio (A/B/C)	Streaming rate (Kbit/s)	Category ratio (A/B/C)
300	35%/45%/20%	300	40%/40%/20%
400	35%/40%/25%	350	35%/45%/20%
500	30%/45%/25%	400	30%/40%/30%
600	40%/30%/30%	450	30%/35%/35%
700	20%/35%/45%	500	20%/35%/45%
800	20%/30%/50%		

FIGURE 6: Average *packet delay* versus streaming rate.

and transmission delay in the optimization scheduling problem. This optimization solution ensures the packet delay can be reduced as much as possible; that is, most segments could be received and decoded before their playback deadlines. (ii) The design of Reencoding Push mode ensures the child can still receive enough blocks in dynamic environment.

The average *continuity index* of the three systems versus the number of peers is illustrated in Figure 8. In general, the average *continuity index* becomes smaller with the increase of system scale. However, the decrease amplitude of our *continuity index* is slight and the *continuity index* of our approach can still keep at a high level, which demonstrates that our approach can keep good scalability. The reason mainly lies in two aspects. (i) The *packet delay* of CoolStreaming and  $R^2$  becomes larger (described in Figure 5), so that, as for CoolStreaming, some segments cannot arrive at the peers before the playback deadline of these segments, and as for  $R^2$ , the peers also cannot receive enough blocks to decode the segments in time. (ii) In our approach, the segments, not decoded and close to the playback deadline, still have chances to obtain the absent coding blocks through our Reencoding Push mode, which will increase the decoding probability of these segments. Our approach has improved CoolStreaming

FIGURE 7: Average *continuity index* versus simulation duration.

and  $R^2$  in terms of these two aspects. Thus, more segments could be decoded before their playback deadline.

**5.2.3. Coding Ratio.** The encoding operation of network coding not only consumes the computing resources but also increases the packet transmission delay. Since encoding operations bring obvious coding overhead, it should be reduced as much as possible. We use *coding ratio* to measure the coding overhead. To simulate the dynamic environment, we let 20 peers leave the system at 600 s of the duration. Figure 9 shows the average *coding ratio* of the two systems, our system and  $R^2$  versus simulation duration. For  $R^2$ , the *coding ratio* is nearly always 100% since each peer has to produce a new coding block by encoding operation before serving a packet to its neighbor. In our push algorithm, most coding blocks are directly forward to the child and a new coding block will be produced for the child by Reencoding Push only if necessary. Therefore, the *coding ratio* of our system is less than 30% most of the time. When some peers leave the system, the reencoding operations briefly increase to supply missing substreams in the parent repair process. So, our system achieves better overhead control and robustness in dynamic environment.

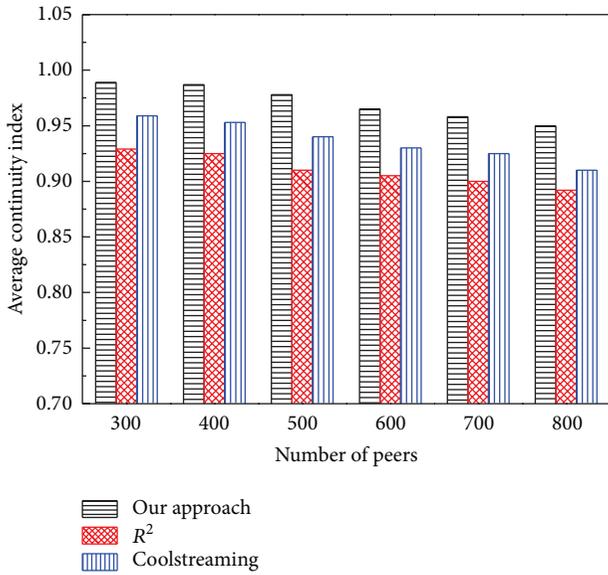


FIGURE 8: Average continuity index versus peer number.

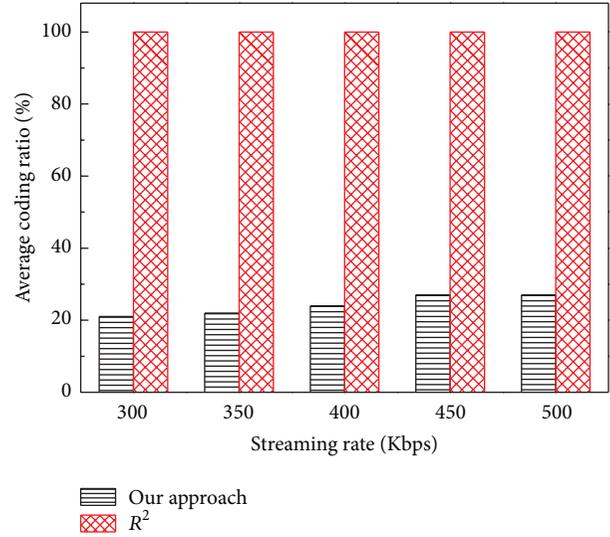


FIGURE 10: Average coding ratio versus streaming rate.

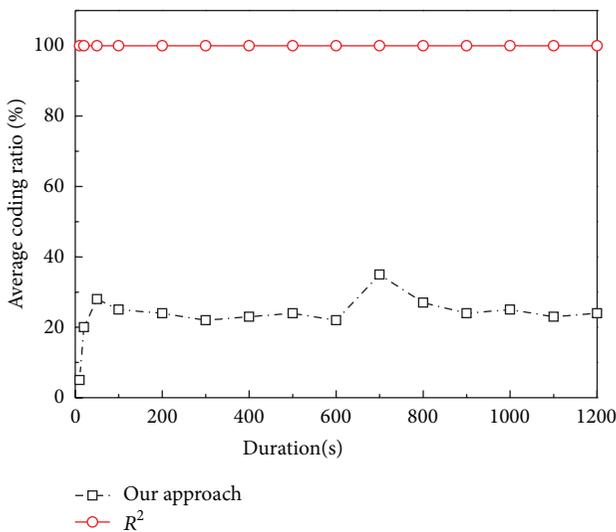


FIGURE 9: Average coding ratio versus simulation duration.

The average coding ratio of the two systems, our system and  $R^2$ , versus the streaming rate is illustrated in Figure 10. The average coding ratio of  $R^2$  is also nearly always 100%. Whatever the streaming rate is, each peer of  $R^2$  needs to generate a new coding block before transmitting a packet to any of its neighbors. The average coding ratio of our approach increases as the streaming rate becomes larger. The reason is that both the quantities of transmitted data and packet delay increase, so that the peers have to use longer time to obtain all the blocks and the quantity of loss packets becomes larger. This situation leads to the probability that the parent cannot forward any block in its buffer to its child will increase a little. According to our approach, the reencoding operations will increase slightly. The increase amplitude of our coding ratio is slight and the coding ratio of our algorithm can still keep at

a low level, which demonstrates that our algorithm can keep good scalability.

## 6. Conclusions

In this paper, we study and propose a QoS driven scheduling approach for network coding based P2P live streaming system. Through introducing a new network coding method for substreams, we reduce the complexity of the scheduling problem, which is formulated as an optimization problem. Furthermore, we transform the optimization problem to an equivalent min-cost flow problem to solve it in polynomial time and propose a push scheduling algorithm to reduce the coding overhead. We conducted extensive simulation to validate the performance and effectiveness of our approach compared with other traditional and state-of-the-art schemes. Experimental results show that our approach achieves better transmission performance and streaming quality with substantially much lower overhead in dynamic environments.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This work was supported in part by Guangdong Natural Science Foundation (Grant no. S2013040012895), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (Grant no. S2013040012895), Major Fundamental Research Project in the Science and Technology Plan of Shenzhen (Grant nos. JCYJ20120613104215889, JCYJ20130329102017840, and JCYJ20130329102032059), National Natural Science Foundation of China (Grant no.

61170209), and Program for New Century Excellent Talents in University (Grant no. NCET-13-0676).

## References

- [1] C. Wu, B. Li, and S. Zhao, "Multi-channel live P2P streaming: refocusing on servers," in *Proceedings of the 27th IEEE Communications Society Conference on Computer Communications (INFOCOM '08)*, pp. 2029–2037, April 2008.
- [2] X. Zhang, J. Liu, B. Li, and Y.-S. Yum, "CoolStreaming/DONet: a data-driven overlay network for peer-to-peer live media streaming," in *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM '05)*, vol. 3, pp. 2102–2111, Miami, Fla, USA, March 2005.
- [3] J. Li and P. Chou, "Mutualcast: an efficient mechanism for one-to-many content distribution," in *Proceedings of SIGCOMM ASIA*, Beijing, China, April 2005.
- [4] M. Zhang, J.-H. Luo, L. Zhao, and S.-H. Yang, "A peer-to-peer network for live media streaming using a push-pull approach," in *Proceedings of the 13th ACM International Conference on Multimedia*, pp. 287–290, ACM, New York, NY, USA, November 2005.
- [5] M. Wang and B. Li, "Lava: a reality check of network coding in peer-to-peer live streaming," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 1082–1090, May 2007.
- [6] W. Mea and L. Baochun, "R2: random push with random network coding in live peer-to-peer streaming," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 9, pp. 1655–1666, 2007.
- [7] "Peer-to-Peer streaming simulator," 2008, <http://media.cs.tsinghua.edu.cn/~zhangm/>.
- [8] M. Castro, P. Druschel, A. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: high-bandwidth multicast in cooperative environments," in *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP '03)*, pp. 298–313, October 2003.
- [9] V. Venkataraman, P. Francis, and J. Calandrino, "Chunky-spread: multi-tree unstructured peer-to-peer multicast," in *Proceedings of 5th International Workshop on Peer-to-Peer Systems (IPTPS '06)*, February 2006.
- [10] M. Luby, "LT codes," in *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS '02)*, pp. 271–280, Vancouver, Canada, November 2002.
- [11] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2551–2567, 2006.
- [12] P. Maymounkov and D. Mazieres, "Rateless codes and big downloads," in *Proceedings of the 2nd International Workshop Peer-to-Peer Systems (IPTPS '03)*, February 2003.
- [13] C. Wu and B. Li, "rStream: resilient and optimal peer-to-peer streaming with rateless codes," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 1, pp. 77–92, 2008.
- [14] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.
- [15] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, 2003.
- [16] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 782–795, 2003.
- [17] C. Gkantsidis and P. Rodriguez, "Network coding for large scale content distribution," in *Proceedings of the 24th IEEE International Conference on Computer Communications (INFOCOM '05)*, March 2005.
- [18] C. Gkantsidis, J. Miller, and P. Rodriguez, "Anatomy of a p2p content distribution system with network coding," in *Proceedings of the 5th International Workshop on Peer-to-Peer Systems (IPTPS '06)*, February 2006.
- [19] T. Ho, M. Medard, J. Shi, M. Effros, and D. R. Karger, "On randomized network coding," in *Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Ill, USA, October 2003.
- [20] Z. Liu, C. Wu, B. Li, and S. Zhao, "UUSee: large-scale operational on-demand streaming with random network coding," in *Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM '10)*, March 2010.
- [21] S. Sarkar and M. Wang, "A measurement study of network coding in peer-to-peer video-on-demand systems," in *Proceedings of the 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS '12)*, pp. 423–430, August 2012.
- [22] S. Sarkar and M. Wang, "Mitigating the asymmetric interests among peers in peer-to-peer video-on-demand systems," in *Proceedings of the International Conference on Computing, Networking and Communications (ICNC '13)*, pp. 653–659, San Diego, Calif, USA, January 2013.
- [23] D. Nguyen and H. Nakazato, "Rarest-first and coding are not enough," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '12)*, pp. 2683–2688, Anaheim, Calif, USA, December 2012.
- [24] A. M. Sheikh, A. Fiandrotti, and E. Magli, "Distributed media-aware scheduling for p2p streaming with network coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13)*, pp. 3597–3601, Vancouver, Canada, May 2013.
- [25] H.-C. Hsu, "A knowledge sharing simulation of team learning," *Transaction on IoT and Cloud Computing*, vol. 1, no. 1, pp. 26–38, 2013.
- [26] B. K. Mishra and S. K. Srivastava, "A quarantine model on the spreading behavior of worms in wireless sensor network," *Transaction on IoT and Cloud Computing*, vol. 2, no. 1, pp. 1–13, 2014.
- [27] R. K. Ahuja, T. L. Magnanti, and J. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, 1993.
- [28] *Meridian node to node latency matrix (2500 - 2500)*, 2005, <http://www.cs.cornell.edu/People/egs/meridian/data.php>.
- [29] S. Saroiu, P. Gummadi, and S. Gribble, "A measurement study of peer-to-peer file sharing systems," in *Proceedings of ACM Multimedia Computing and Networking*, January 2002.

## Research Article

# An Emergency Packet Forwarding Scheme for V2V Communication Networks

**Faika Hoque and Sungoh Kwon**

*School of Electrical Engineering, University of Ulsan, Ulsan 680-749, Republic of Korea*

Correspondence should be addressed to Sungoh Kwon; [sungoh@ulsan.ac.kr](mailto:sungoh@ulsan.ac.kr)

Received 18 March 2014; Accepted 7 June 2014; Published 25 June 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 F. Hoque and S. Kwon. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes an effective warning message forwarding scheme for cooperative collision avoidance. In an emergency situation, an emergency-detecting vehicle warns the neighbor vehicles via an emergency warning message. Since the transmission range is limited, the warning message is broadcast in a multihop manner. Broadcast packets lead two challenges to forward the warning message in the vehicular network: redundancy of warning messages and competition with nonemergency transmissions. In this paper, we study and address the two major challenges to achieve low latency in delivery of the warning message. To reduce the intervehicle latency and end-to-end latency, which cause chain collisions, we propose a two-way intelligent broadcasting method with an adaptable distance-dependent backoff algorithm. Considering locations of vehicles, the proposed algorithm controls the broadcast of a warning message to reduce redundant EWM messages and adaptively chooses the contention window to compete with nonemergency transmission. Via simulations, we show that our proposed algorithm reduces the probability of rear-end crashes by 70% compared to previous algorithms by reducing the intervehicle delay. We also show that the end-to-end propagation delay of the warning message is reduced by 55%.

## 1. Introduction

According to the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA), vehicle crashes are the leading cause of death in the United States [1]. Rear-end collisions account for almost one-third of all traffic crashes [2]. Driver inattention is a major factor in 91% of rear-end crashes, as reported in [3]. Most of the rear-end collisions resulted from insufficient time for drivers to react in an emergency situation. A driver typically depends on the visual information of the immediately preceding vehicle. In road emergency situations, if a vehicle does not notice the visual information in time or there is not enough given time to brake or react, a chain collision could result. According to [4], an extra 0.5 seconds of warning time can prevent about 60% of rear-end collisions.

The following simplified example illustrates a chain collision due to inadequate time to react in an emergency situation. We assume that three vehicles are travelling in the same lane at the identical speed of 110 km/hr and are spaced

by 30 m. It is also assumed that the vehicles decelerate at  $4 \text{ m/s}^2$  and follow the visual information of the immediately preceding vehicle. After observing an emergency situation, the first vehicle brakes abruptly. In general, driver reaction time (the duration time between when an event is observed and when the driver brakes) is in the range of 0.75 s to 1.5 s [5]. Suppose that the second vehicle takes 0.75 s and the third vehicle takes 1.5 s to react; then although the second vehicle can stop before colliding with first vehicle, the third vehicle crashes into the second vehicle.

The intelligent transportation system (ITS) can help drivers to react in an emergency situation. ITS provides a framework to alleviate traffic congestion and improves public safety goals such as collision avoidance. The allocation of 75 MHz in the 5.2 GHz band for licensed dedicated short range communication (DSRC) delivers high media contents to vehicle-to-vehicle (v2v) communications [6, 7]. To enhance highway public safety, a cooperative collision avoidance (CCA) system can be used via v2v wireless

communication [8]. In such cooperative systems, vehicles determine the emergency situation and make a decision whether to warn others based on the information given by the neighbor vehicles.

As soon as an emergency situation is detected, a vehicle warns its neighboring vehicles via an emergency warning message (EWM), and the recipients of the warning message announce the warning to their neighbors. Since acknowledgement of one-to-one transmission produces more delay and all the neighbors should be aware of the warning situation immediately, EWMs are broadcast to neighbors [9, 10]. Moreover, to improve the delivery success, EWMs are periodically rebroadcast.

Although such cooperative v2v communication systems can improve highway safety by using emergency warning messages, reliable transmissions and stringent delay are the main requirements to deliver a warning message in an emergency situation. If the EWM delivery latency is large, then a chain collision will occur. Therefore, an efficient EWM forwarding protocol is necessary to reduce redundant broadcasting and packet collisions, which impede the warning message delivery, at the media access control (MAC) layer.

Previous work has taken effort to reduce the delivery latency of EWM by improving the warning message forwarding protocol and MAC enhancement. To reduce redundant broadcasting, intelligent broadcasting with implicit acknowledgement (I-BIA) is proposed in [9, 10]. In this proposed method, implicit acknowledgement is adopted to reduce redundant broadcasting, in which reception of the duplicate EWM acts as implicit acknowledgment. After receiving a duplicate EWM from one of vehicles behind itself, the vehicle stops broadcasting the EWM packet. Note that I-BIA uses only one-way implicit acknowledgement to stop the redundant broadcasting assuming that all vehicles in the vehicle chain received the warning messages. If an intermediate vehicle in the chain cannot receive the warning message due to transmission collisions, the proposed algorithm delays the recognition of emergency at the vehicle until an EWM arrives from a preceding vehicle. If the intermediate vehicle receives the EWM after more than 500 ms delay, then a chain collision will occur [4]. The algorithm also ignores the warning message delay in the dense network. Moreover, the algorithm uses only binary exponential back-off (BEB) for EWM. Binary exponential back-off induces a large delay to propagate the warning message as the senders enter into a long inhibition period before transmission.

To reduce the EWM propagation delay due to exponential random back-off, a fixed contention window (CW) is proposed in [11] with implicit acknowledgment. While a fixed contention window induces a small delay in a sparse network, it causes an extremely large delay in a dense network due to the large number of transmissions among EWMs. An adaptable offset slot (AOS) mechanism is proposed in [12] to reduce the delay by initializing different contention window sizes depending on the number of neighbor vehicles. Since the algorithm linearly increases the contention window as the number of vehicles increases, an emergency message has a large delay to overcome the nonemergency transmissions in a dense network.

To overcome nonemergency transmissions, an adaptable distance-dependent random back-off algorithm is proposed in [13]. The algorithm adopts heterogeneous random back-off based on the location and the number of neighbor vehicles. However, the proposed algorithm has no scheme to reduce the redundant EWMs, which results in collisions among EWMs and induces delay in propagation of the warning message.

In this paper, we propose a two-way intelligent broadcasting scheme to reduce the redundant broadcasting of EWMs and avoid a chain collision among intermediate vehicles. Moreover, a composite random back-off is adopted for EWM to compete with nonemergency transmissions and emergency transmissions in a dense network. The performance is evaluated via simulations in different environments.

The rest of this paper is organized as follows. In Section 2, we describe the system model and problem of the IEEE 802.11 DCF-based MAC protocol for EWM in a vehicle environment. In Section 3, we propose an efficient forwarding strategy with an adaptable random back-off for EWM transmission. In Section 4, we analyze the performance of the proposed algorithm via simulation results. Finally, we conclude the paper in Section 5.

## 2. System Model and Challenge

*2.1. System Model.* All vehicles on the highway are assumed to be equipped with wireless communication devices based on the IEEE 802.11p standard. The IEEE 802.11p uses the basic mechanism of the distributed control function (DCF) as the fundamental protocol for medium access control. The protocol adopts the carrier sense multiple access with collision avoidance (CSMA/CA) as a random access scheme for all vehicles [14]. We also assume that all vehicles are equipped with global positioning system (GPS) [9, 15] to estimate location. In a normal situation, vehicles establish a mobile ad hoc network via a routing protocol and exchange nonemergency data with each other.

When noticing an emergency situation, an event-detecting vehicle broadcasts an EWM to immediately warn the neighbor vehicles. Due to limited transmission range, all vehicles receiving the warning message relay the message to neighbor vehicles, which then forward the message in a multihop manner. Before transmitting the EWM, vehicles sense the medium to check whether it is idle or busy via CSMA/CA scheme [16]. If the medium is idle during interframe space (IFS), the scheme waits for a random back-off period time before transmitting data. A random back-off period is a waiting slot time, which is randomly chosen between 0 and the contention window (CW), before transmitting the packet to avoid the collision. The back-off time is decreased from the randomly chosen time as long as the medium is sensed to be idle. When the scheme reaches time 0, it transmits the packet. If the medium is busy, the vehicle defers message transmission until the end of the current transmission, and the CSMA/CA procedure starts again. Since broadcasting does not support acknowledgement, to improve delivery efficiency, EWM is periodically rebroadcast.

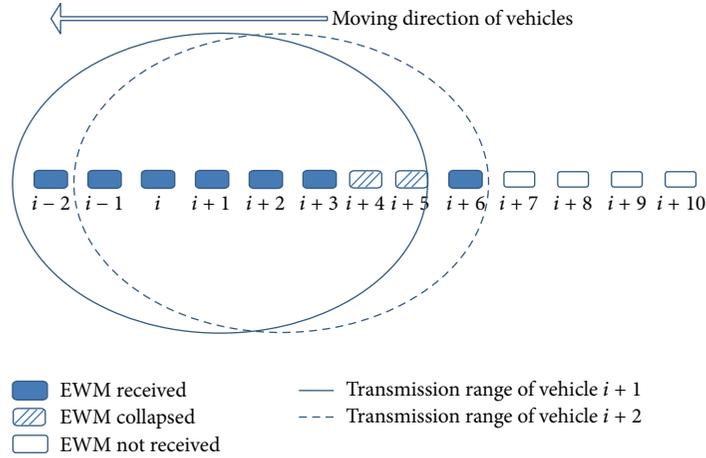


FIGURE 1: EWM collisions in the vehicle chain.

In such CSMA/CA-based v2v networks, redundant EWMs induce high collisions to access the channel. Moreover, EWMs compete with two categories of transmission: EWMs and nonemergency transmissions. In next subsection, we discuss two major challenges in a v2v network in detail and propose an effective EWM forwarding scheme in the following section.

### 2.2. Challenges in Transmitting EWM

**2.2.1. Redundant Broadcast of EWM.** Since EWM is periodically broadcast to improve the delivery success rate, vehicles which have already successfully forwarded the message to the vehicles behind will keep broadcasting the message. Moreover, overlapping transmission ranges of vehicles also induce collision among EWMs. The primary limitation of this network is a broadcasting storm in the network, which results in packet collisions for channel access and a hindrance of EWM forwarding to neighboring vehicles. The effect of redundant EWMs may be serious hazard situations in emergency cases.

For example, let us assume that an emergency-detecting vehicle broadcasts an EWM to warn the neighbors of an emergency and that all vehicles up to  $i + 3$  received the warning message in a multihop manner, as shown in Figure 1. If vehicles  $i + 1$  and  $i + 2$  transmit the message after the random back-off, vehicle  $i + 6$  receives the message from vehicle  $i + 2$ , but due to EWM transmission collision, vehicles  $i + 4$  and  $i + 5$  do not receive the warning message. Such EWM-unaware vehicles in the chain have greater possibility for rear-end crashes.

**2.2.2. Competition with Nonemergency Messages.** Vehicles in the chain can be categorized into two groups: vehicles that did not receive the EWM and vehicles that received the EWM, as shown in Figure 2. Vehicles which did not receive the warning message due to limited transmission range are unaware of the emergency situation and continue to send their existing

nonemergency messages. The vehicles that received the warning message relay the message by periodically rebroadcasting. However, nonemergency transmissions prohibit further warning message forwarding. Therefore, vehicles close to the boundary region, such as vehicles  $i + 2$  and  $i + 3$  in Figure 2, are more affected by nonemergency transmissions. If vehicles that did not receive EWM are present among vehicles receiving the EWM, as in Figure 1, the interference with nonemergency transmissions will significantly affect the EWM forwarding. Hence, the EWM-unaware vehicles among EWM-aware vehicles not only are in a hazardous situation due to EWM unawareness, but also jeopardize other vehicles due to hindrance to disseminate the EWM.

Vehicles receiving EWM should have a small contention window for EWM to compete with nonemergency transmissions, although these same vehicles should contradictorily have a large contention window to reduce collisions among EWM transmissions.

### 3. Proposed Algorithm

In this section, two-way intelligent broadcasting with implicit acknowledgement (2I-BIA) is proposed to reduce redundant EWMs where vehicles receive the acknowledgement from both directions. In addition, to compete with nonemergency transmissions, vehicles close to the boundary region use a small contention window. To reduce the collisions among EWMs, vehicles in the near region of the corresponding EWM sender use a conventional contention window.

To that end, we employ the EWM format proposed in [10]. The EWM contains an original vehicle ID, an event ID, a message type, and a sender's location. The origin vehicle ID and the event ID uniquely identify a message across the network. The message type informs others of an emergency situation. With the sender's location and the receiver's location, the receiver recognizes the direction of the message, from a preceding vehicle or a following vehicle, and measures the distance between the sender and itself.

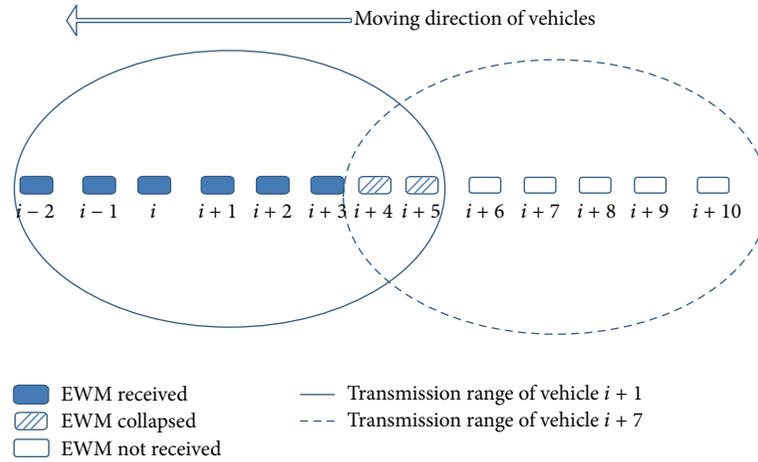


FIGURE 2: Interference of EWM with nonemergency transmissions.

```

if New EWM packet then
    Store the EWM and periodically rebroadcast the EWM with adaptable CW;
else
    Compare the direction of the arrival packet with the previous EWM packet;
    if From the opposite direction then
        Ignore the packet and stop broadcasting;
    else
        Periodically broadcast the EWM with adaptable CW;
    end
end

```

ALGORITHM 1: Two-way intelligent broadcast with implicit acknowledgment (2I-BIA).

When receiving an EWM, a vehicle decides if the EWM is a new message based on the originating vehicle ID and the event ID. If the EWM is a new message, the vehicle stores the message and periodically rebroadcasts the message with an adaptable contention window, which will be described later.

If the EWM is a duplicate, the vehicle determines whether the message is from a sender that is in the opposite direction of the previous EWM sender. For example, the previous EWM came from a vehicle that is behind (or in front of) the receiver, and the present EWM arrived from a vehicle that is in front of (or behind) the receiver. In such a case, the EWM-receiving vehicle recognizes the EWM as an implicit acknowledgement of the previous EWM so that the vehicle stops periodic broadcasting of the EWM. If the duplicate EWM comes from a vehicle that is in the same direction as the previous EWM sender with respect to the receiver, then the receiver keeps periodically broadcasting the EWM with an adaptable contention window.

When a vehicle periodically broadcasts an EWM, in order to compete with nonemergency transmissions and reduce the possibility of collisions among EWMs, an adaptable contention window is employed, which is introduced in [13]. To employ an adaptable contention window, a threshold value is chosen as a function of neighbor vehicles and contention window size. If a vehicle receiving an EWM is close to

vehicles that do not receive an EWM, the vehicle chooses a short and fixed contention window size for random back-off. Otherwise, an ordinary binary random back-off is used. In other words, if a receiver is behind the corresponding EWM sender with respect to the driving direction and the distance from the sender to the receiver is greater than a threshold, the receiver is determined to be in the area of competition with nonemergency messages and chooses a fixed small contention window to compete. Otherwise, assuming that the collisions among EWMs are more dominant, an ordinary binary random back-off is chosen. The algorithm is summarized in Algorithm 1.

#### 4. Simulations

The proposed system is implemented using the OPNET 16.0 network simulator [17]. The simulation results are the average of 20 runs using different random seeds unless stated otherwise. The physical characteristics follow the specifications of 802.11 with 11 MHz of bandwidth. The transmission range of each vehicle is 300 m, as specified by the DSRC [6, 18]. The underlying MAC protocol is based on the 802.11 DCF function. To employ the proposed adaptable random back-off algorithm for the EWM, the value of the  $CW_{\text{fix}}$  is set

to 7 for the fixed random back-off, and the values of the  $CW_{min}$  and  $CW_{max}$  are set to 15 and 1023, respectively, for the ordinary binary random back-off algorithm. We consider nonemergency transmissions such as communications between vehicles and control messages for routing [9, 11].

For simplicity, we assume that vehicles are driving in one direction in a single lane. To maintain a safe distance between vehicles, the intervehicle spaces are based on the driving speed. If there is no traffic congestion, vehicles move at high speed, and the intervehicle space will be large. If traffic congestion is present, vehicles move slowly and have a high density. Hence, we assume that the vehicles are uniformly distributed over 1 km and that the number of vehicles ranges from 20 to 140.

When an emergency event occurs, the event-detecting vehicle broadcasts an EWM packet to warn the neighbor vehicles. The length of the packet is 1024 bits. Since the vehicle does not wait for acknowledgement, the EWM is rebroadcast every 50 ms to ensure successful transmission. After receiving the warning message, the neighboring vehicles rebroadcast the packet until it reaches the end of the vehicle chain. In this way, drivers become aware of the emergency situation and start to decelerate to avoid the collision.

For performance measurement, we consider intervehicle delay and end-to-end delay. The intervehicle delay is the time difference of receiving the EWM between two adjacent vehicles. The end-to-end delay is defined as the elapsed time from when the first vehicle transmits an EWM to the time when the last vehicle in the lane receives the EWM. We compare the performance of our proposed algorithm 2I-BIA with the original CSMA/CA, I-BIA proposed in [9] and adaptable distance-dependent back-off algorithm (ADDB) proposed in [13].

**4.1. Interverhicle Delay.** In this section, we study the receiving time of an EWM at each vehicle after the first emergency-detecting vehicle begins broadcasting. If the receiving time of an EWM between two adjacent vehicles is large, then a chain collision may occur because the vehicles will not have enough time to brake or react to the emergency situation.

Figure 3 shows the receiving time of an EWM packet at each vehicle when the first of 100 vehicles initiates the warning message in the presence of 100 kbps nonemergency message traffic. Up to the 40th vehicle, the receiving time is almost the same at each vehicle. However, in the region far from the first vehicle, when the number of transmissions increases, the receiving times at each vehicle are different due to transmission collisions among redundant EWMs and competition of the EWM with the nonemergency transmissions. In the case of the original CSMA/CA, the receiving time is abruptly increased between the 79th vehicle and the 80th vehicle, and the 80th vehicle produces a discontinuity between the 79th and 81th vehicles. Although I-BIA and ADDB reduce the intervehicle delay, discontinuity between adjacent vehicles in a vehicle chain still occurs. However, our proposed 2I-BIA algorithm reduces the probability of discontinuity in the vehicle chain by reducing the intervehicle delay of adjacent vehicles. Since the proposed two-way

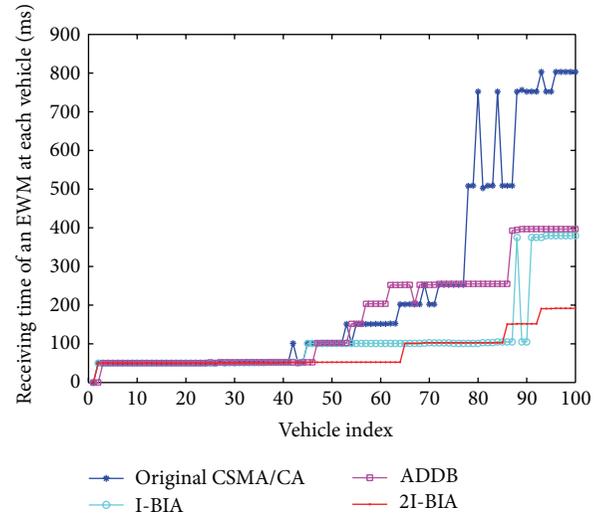


FIGURE 3: Receiving time of first EWM at each vehicle.

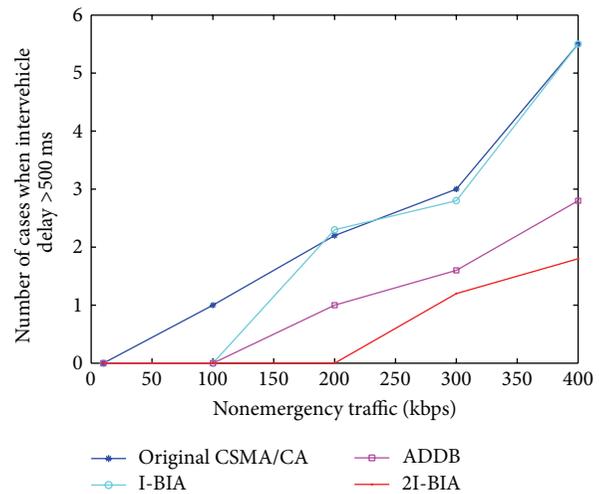


FIGURE 4: Number of cases where intervehicle delay is  $\geq 500$  ms.

acknowledgement reduces the redundant EWMs and ensures successful delivery in both directions, all vehicles in the chain receive the EWM after less than 100 ms intervehicle delay.

Greater intervehicle delay induces a higher possibility of chain collision. According to [4], if driver has an extra 500 ms to react in an emergency situation, 60% of chain collisions are avoidable. Hence, we discuss the number of cases in which intervehicle delays are greater than 500 ms. Figure 4 shows the number of cases when intervehicle delay is greater than 500 ms with the presence of different levels of nonemergency transmissions. Simulations are performed with 100 vehicles in the presence of different levels of non-emergency transmissions. Simulation results show that, with other algorithms, at least one vehicle experiences a 500 ms intervehicle delay when the nonemergency traffic is 200 kbps, while our proposed algorithm gives an intervehicle delay less than 500 ms. As the number of nonemergency transmissions increases, the number of cases where intervehicle delay is

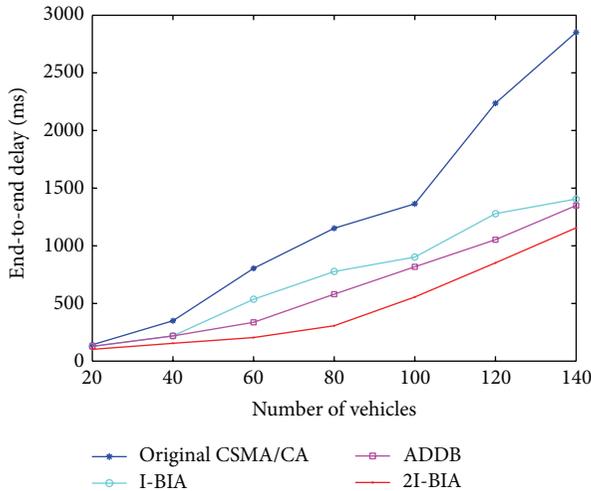


FIGURE 5: End-to-end delay of an EWM packet with various numbers of vehicles.

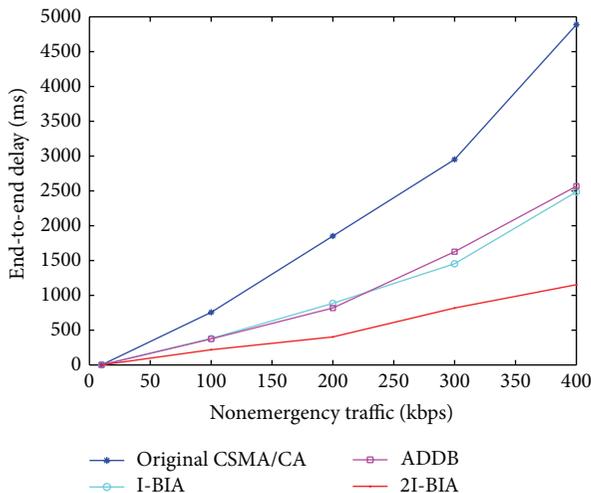


FIGURE 6: End-to-end delay of an EWM packet with different nonemergency traffic levels.

greater than 500 ms also increases. However, our proposed algorithm reduces the collision probability by at least 70% compared with the I-BIA algorithm.

**4.2. End-To-End Delay.** In this section, we compare the average end-to-end delay as performance measurement in the presence of nonemergency traffic as the number of vehicles varies. Figure 5 shows the end-to-end delay with 200 kbps nonemergency traffic. As the number of vehicles increases, the number of retransmissions also increases, which results in more delay in all the algorithms. However, our proposed algorithm uses short CW to overcome non-emergency transmissions and two-way acknowledgement to reduce redundant EWMs to produce a shorter delay than the other methods.

As nonemergency transmissions increase, the end-to-end delay also increases, as shown in Figure 6. For all cases, our

algorithm outperforms the other algorithms and improves the performance by approximately 55% compared with I-BIA and ADDB.

## 5. Conclusions

In this paper, we proposed an emergency warning message forwarding scheme to improve message delivery efficacy in vehicular environments. In a v2v network, two challenges prevent warning message forwarding: redundant EWM transmissions from EWM recipients and non-emergency transmissions from the emergency-unaware vehicles. To overcome these challenges, the proposed algorithm intelligently broadcasts emergency warning messages using duplicates as implicit acknowledgements and adopting an adaptable contention window size. If a vehicle is recognized to be between EWM recipients based on duplicate EWMs, then the vehicle stops relaying the EWM. Moreover, a vehicle broadcasts EWM using heterogeneous contention window sizes to reduce EWM transmission collision and compete with non-emergency transmissions. Simulation results show that our proposed algorithm reduces the intervehicle delay by 70% and the end-to-end delay by 55% compared to previous algorithms, thus reducing the possibility of a chain collision.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was supported by the 2014 Research Funds of Hyundai Heavy Industries for University of Ulsan.

## References

- [1] The National Highway Traffic Safety Administration (NHTSA), "National center for statistics and analysis(NCSA)," 2010.
- [2] Washington State Department of Transportation (WSDOT), "2005 Annual State Highway Collision Data Summary," 2007.
- [3] H. Lum and J. A. Reagan, "Interactive highway safety design model: accident predictive module," *Public Roads*, vol. 59, no. 2, 1995.
- [4] The National Highway Traffic Safety Administration (NHTSA), *Final Report of Automotive Collision Avoidance Systems (ACAS) Program*, 2000.
- [5] N. D. Lerner, "Brake perception-reaction times of older and younger drivers," in *Proceedings of the 37th Annual Meeting the Human Factors and Ergonomics Society*, pp. 206–210, October 1993.
- [6] ASTM E2213-03, "Standard specification for telecommunications and informations exchange between roadside and vehicle systems-5 GHz band dedicated short range communications (DSRC) media access control (MAC) and physical layer (PHY) specifications," 2003.
- [7] Q. Xu, R. Sengupta, D. Jiang, and D. Chrysler, "Design and analysis of highway safety communication protocol in 5.9 GHz dedicated short range communication spectrum," in

- Proceedings of the 57th IEEE Semiannual Vehicular Technology Conference (VTC '03)*, pp. 2451–2455, April 2003.
- [8] T. ElBatt, S. K. Goel, G. Holland, H. Krishnan, and J. Parikh, “Cooperative collision warning using dedicated short range wireless communications,” in *Proceedings of the 3rd ACM International Workshop on Vehicular Ad Hoc Networks (VANET '06)*, pp. 1–9, September 2006.
- [9] S. Biswas, R. Tatchikou, and F. Dion, “Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety,” *IEEE Communications Magazine*, vol. 44, no. 1, pp. 74–82, 2006.
- [10] R. Tatchikou, S. Biswas, and F. Dion, “Cooperative vehicle collision avoidance using inter-vehicle packet forwarding,” in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '05)*, pp. 2762–2766, December 2005.
- [11] F. Ye, M. Adams, and S. Roy, “V2V wireless communication protocol for rear-end collision avoidance on highways,” in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '08)*, pp. 375–379, Beijing, China, May 2008.
- [12] C. Hsu and H. Tseng, “MAC channel congestion control mechanism in IEEE 802.11p/WAVE vehicle networks,” in *Proceedings of the 74th Vehicular Technology Conference (VTC '11)*, San Francisco, Calif, USA, September 2011.
- [13] F. Hoque, *Efficient algorithm for emergency message transmission in a vehicle environment [M.S. thesis]*, University of Ulsan, Ulsan, Republic of Korea, 2013.
- [14] H. Menouar, F. Filali, and M. Lenardi, “A survey and qualitative analysis of MAC protocols for vehicular ad hoc networks,” *IEEE Wireless Communications*, vol. 13, no. 5, pp. 30–35, 2006.
- [15] S. Kwon and N. B. Shroff, “Geographic routing in the presence of location errors,” *Computer Networks*, vol. 50, no. 15, pp. 2902–2917, 2006.
- [16] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.
- [17] “Application and Network Performance Management,” <http://www.opnet.com>.
- [18] X. Yang, L. Liu, N. Vaidya, and F. Zhao, “A vehicle to vehicle communication protocol for cooperative collision warning,” in *Proceedings of the 1st Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MOBIQUITOUS '04)*, pp. 114–123, 2004.

## Research Article

# Towards Accurate Node-Based Detection of P2P Botnets

Chunyong Yin<sup>1,2</sup>

<sup>1</sup> School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China

<sup>2</sup> Jiangsu Engineering Center of Networking Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China

Correspondence should be addressed to Chunyong Yin; [yinchunyong.nju@gmail.com](mailto:yinchunyong.nju@gmail.com)

Received 4 April 2014; Accepted 15 May 2014; Published 24 June 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Chunyong Yin. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Botnets are a serious security threat to the current Internet infrastructure. In this paper, we propose a novel direction for P2P botnet detection called node-based detection. This approach focuses on the network characteristics of individual nodes. Based on our model, we examine node's flows and extract the useful features over a given time period. We have tested our approach on real-life data sets and achieved detection rates of 99-100% and low false positives rates of 0-2%. Comparison with other similar approaches on the same data sets shows that our approach outperforms the existing approaches.

## 1. Introduction

*1.1. Background and Motivation.* Botnets are groups of computers which are linked to each other through similar network processes which perform coordinated tasks like information crawling, Internet Relay Chatting (IRC), and information sharing. While botnets can be used for benign purposes, over the past decade, there has been a drastic increase in the number of malicious botnets [1] which have become a serious concern to the security of Internet applications and networking infrastructure. To build a malicious botnet, the attacker, known as botmaster, uses one or more central servers to compromise and acquire control of vulnerable computers using malware. These compromised computers, known as “bots,” link to the central server, which acts as the command and control center (C&C) using protocols like IRC or HTTP. The botmaster uses these channels to deliver additional malware and instructions to the bots for launching different kinds of attacks. Malicious botnets are capable of a wide range of attacks including e-mail spam, keystroke logging, packet sniffing, DoS attacks, and identifying new targets for enlisting in the botnet, among others [2-6]. In this paper, unless explicitly stated, we use the term “botnet” to refer to a malicious botnet.

Figure 1 illustrates a botnet of five bots connected to two C&C servers controlled by a botmaster. This botnet

is operating in a centralized mode; that is, one or two servers control the bots in the network. Such centralized botnets can be shutdown or blocked if the C&C servers are identified, thereby rendering the botnet ineffective. To increase resiliency to detection, recent botnets are built using peer-to-peer networking principles where any node can act as a client as well as a server. Accordingly, in a P2P botnet any node can act as a bot as well as the C&C server. In Figure 2, we show an example P2P network and in Figure 3 we show a corresponding P2P botnet. The botmaster can connect to any P2P bot in the network and operate it as the C&C server. Compared to the server-client botnet, the P2P botnet has the ability to realize highly scalable and extensible network structure which is resilient to firewall sanctions and node/path failures.

In this paper, we focus on the problem of detecting P2P bots in a distributed network. Mitigating the threat of a P2P botnet is a challenging task as the botnet has no central C&C server which can be blocked and the botmaster uses the overlay structure of the P2P network to stay connected to the bots. As shown in Figure 3, even if some bots are blocked by firewalls, the botmaster can continue communication with these bots along alternate routing paths as long as the blocked nodes are connected to at least one other P2P bot. This implies that it is essential to detect the bots in a systematic and comprehensive manner. Furthermore, the malware programs

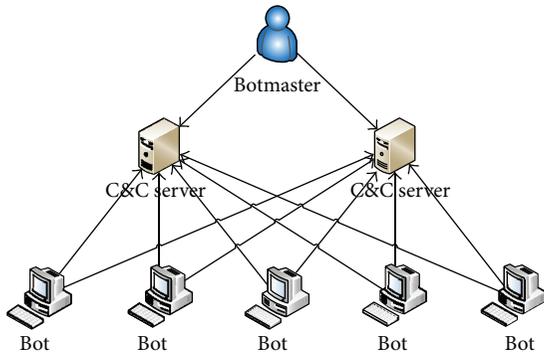


FIGURE 1: C&amp;C botnet.

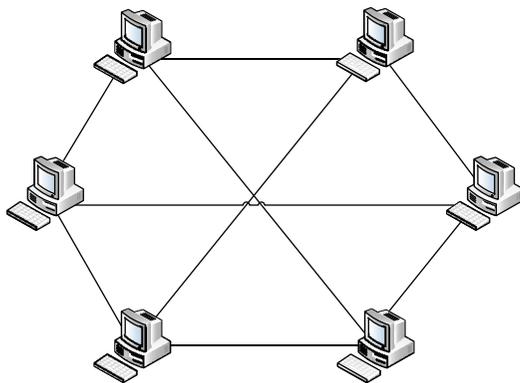


FIGURE 2: P2P network.

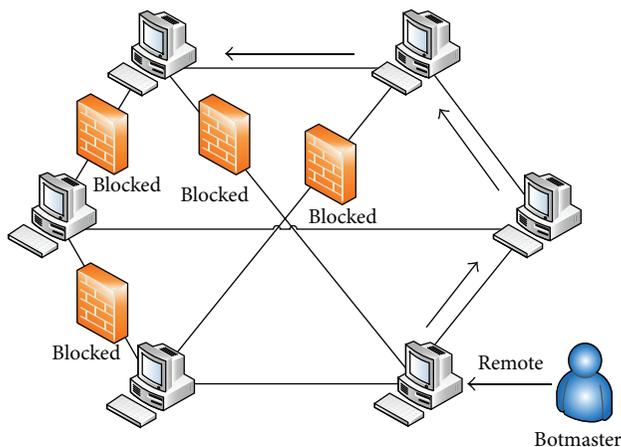


FIGURE 3: P2P botnet bypassing firewalls.

used in P2P bots are, typically, self-propagating—which help in discovering new peers and mutating as well—to avoid signature-based detection. Due to their resilience, P2P botnets represent a major threat to the Internet applications and infrastructure. Therefore, to safeguard the Internet from strategic coordinated attacks, there is an urgent need to devise solutions to detect P2P bots and render the P2P botnets ineffective.

*1.2. Limitation of Prior Art.* The existing solutions for P2P botnet detection can be broadly classified into signature [7–13] and flow-based detection [14, 15, 15–18]. Signature-based P2P bot detection is based on inspecting each packet in the network traffic, entering or leaving the Internet gateway of the network, for the presence of special features such as port numbers, byte sequences in the payload, and blacklisted IP address. These special features, known as signatures, are extracted from known botnet infections in the past and stored in a signature database. While signature-based detection has good detection rate and is easily deployed, it has two major limitations. First, it is deterministic and relies only on detecting known botnet infections and cannot detect unknown bots. Even known bots can evade signature detection by changing ports of communication or use packet payload encryption to hide the bot specific features. Second, inspection of each packet results in performance degradation especially when the traffic consists of a large amount of benign data.

Flow analysis based bot detection examines network flows between two nodes where a flow is defined as a set of packets which have the same source address, source port, destination address, and destination port. The intuition in these approaches is that the flow features, such as the count of the packets in the flow, the order of the packet arrivals, and the interval between packets, can model the botnet communication patterns more accurately than direct packet inspection. The extracted features are used to construct a classifier that can differentiate normal flows and malicious bot flows. Since classifiers use statistical profiling, flow-based analysis is capable of detecting unknown bots which exhibit behavioral similarities to known bots. However, flow-based techniques suffer from two key limitations. First, there are several flows between any two network nodes which need to be analyzed and, usually, most of these flows belong to normal network processes. Second, the flow features need to be extracted at runtime which implies that flow-based analysis requires considerable computational overhead at runtime. At any given instant, there are a significant number of flows in the network which exaggerate the impact of these limitations further.

*1.3. Proposed Approach.* In this paper, we describe a novel P2P bot detection approach, called node-based bot detection, in which we analyze the network profile of nodes to detect bot characteristics. A sample network profile of a node may comprise the different protocols used by the node, the number of flows in a particular time period, packet statistics, and so on. Our approach is based on the intuition that P2P bots exhibit a distinct network profile due to the various P2P network maintenance related tasks they are required to perform. A P2P bot will be more active in communicating with other P2P bots and exchanging various instructions related to control and command. Also, unlike normal P2P nodes, the P2P bots exhibit nearly uniform network activity based on the instructions of the current C&C server. Based on these observations, our approach consists of identifying and quantifying the network profile features that are typical of a

P2P bot. To extract these features, we monitor the network flows at each node and generate the node's network profile. The final network profile of a node is a combination of the features typical of P2P bots and the features observed from the network flows at the node. Finally, we use machine learning based classification techniques to detect whether the network profile of a node corresponds to the network profile of a P2P bot.

*1.4. Technical Challenges and Solutions.* There are several technical challenges in our approach. First, the process of constant flow monitoring at a node results in a major computational and storage overhead. We address this issue by using a sampling approach, wherein we periodically sample a network flow at different time intervals. Although sampling may not detect the same number of bots as those detected by constant flow monitoring in the same time interval, due to the cyclic nature of P2P botnets, the sampling approach eventually detects all the bots in the P2P botnet. Second, quantifying the network profile of P2P bots is nontrivial as different botnets exhibit different semantics and use variable protocols. To address this concern, we abstract and model the general network features of a P2P botnet using the profiles of few existing P2P botnets. We focus on the communication patterns of P2P botnets and do not consider the individual protocol and payload features. By avoiding the payload inspection we are able to overcome the difficulty of handling encrypted payloads and also avoid compromising the privacy of individual users. We combine these unique bot specific features with the flow statistics of the node to obtain the network profile of a node. Third, differentiating between the behavior of a normal node and a P2P bot is a complex problem. Towards this, we use machine learning techniques to cluster and classify the collected network profile features. We use the decision tree technique because of its efficiency and ease of implementation. To evaluate our approach, we use real-life data sets which contain a mix of malicious and nonmalicious data. We ensure that the nonmalicious data dominates the malicious content in order to estimate the sensitivity of our approach.

*1.5. Key Contributions.* The key contributions of our work are as follows. (a) We describe the first node-based approach to detect P2P bots in a P2P botnet. Our approach is a significant deviation from the signature-based and flow-based detection approaches. (b) We describe a sampling technique to reduce the overhead of monitoring for the network administrator. (c) We abstract and quantify the network profile features of a P2P bot. Our abstraction technique avoids dealing with issues like packet encryption and user privacy. (d) We describe the use of efficient machine learning algorithms to classify the network profile into normal and P2P bot profile. (e) We have evaluated our approach on real-life malicious traffic datasets and obtained a detection accuracy of 99-100% with extremely low false positives in the range of 0-2%. We also show that existing state-of-the-art techniques perform poorly on the same data set when compared to our approach.

*1.5.1. Organization.* In Section 2, we describe the related research in this domain. In Section 3, we describe our node-based detection approach. In Section 4, we perform a detailed evaluation of our approach. We compare our scheme with other existing approaches in this domain. We summarize our paper and describe future directions in Section 5.

## 2. Related Research

Signature-based bot detection approach has been widely studied [7-13]. This approach is effective to detect known bots, for example, Phatbot. Kolbitsch et al. [19] proposed a signature-based malware detection system which uses special graphs to detect different kinds of bots. However, the detection rate in this approach is only 64%. The utility of signature-based methods is limited as they are not capable of detecting unknown bots or variants of known bots. In the current Internet scenario numerous new bot variants are increasing rapidly, thereby necessitating the need for more adaptive approaches for bot detection.

Flow-based analysis for bot detection has better detection rate. These techniques [14, 15] were proposed to model a wider range of bot behaviors than those covered in signature-based techniques. Livadas et al. [16] developed a system to detect C&C traffic of botnets based on flow analysis. This system consists of two stages: the first stage extracts several per-flow traffic features including flow duration, maximum initial congestion window, and average byte count per packet; and the second stage uses a Bayesian network classifier to train and detect bots. However, the observed false positive rate is still very high, 15.04%, as it fails to capture botnet specific network profiles effectively. Choi et al. [17] proposed a botnet detection mechanism based on the monitoring of DNS traffic during the connection stage of bots. However, a botnet can easily evade this mechanism, if it rarely uses DNS at its initialization and limits or avoids DNS usage at latter stages.

Wang et al. [20] presented a detection approach of P2P botnets by observing the stability of control flows in the botnet initialization time intervals. However, this approach suffers from high performance and storage overhead while achieving similar detection accuracy as earlier approaches. Kang et al. [15] proposed a novel real-time detection model named the multistream fused model, in which they process different types of packets in a graded manner. However, this model cannot achieve desirable detection accuracy when deployed in a large-scale network environment. Liu et al. [18] presented a general P2P botnet detection model based on macroscopic features of the network streams by utilizing cluster techniques. The proposed method is unreliable or ineffective if only a single infected machine is present on the network.

To the best of our knowledge, there has been no research focusing on the application of node-based analysis for P2P bot detection. The node-based approach has distinct advantages that separate it from signature-based and flow-based techniques.

### 3. Node-Based P2P Bot Detection

In this section, we describe our node-based P2P bot detection approach. Our approach consists of four important steps: P2P bot quantification, efficient flow monitoring, classification, and evaluation. In Section 3.1, we describe our methodology for modeling P2P bots which is the most important step of our detection approach. Using this model, we identify the features to quantify a P2P bot. In Section 3.2, we describe our approach for reducing the complexity of the flow monitoring at the network nodes. In Section 3.3, we describe our classification approach, for identifying P2P profiles from the set of network profiles of all nodes, and describe the evaluation metrics of the classification approach.

*3.1. Our Model for Quantifying P2P Bot Features.* In our node-based P2P bot detection approach, we monitor the communication flows at every node in the network to check for bot infection. Since each flow can exhibit many features, it is important to identify and isolate features which are unique to P2P bots. Our model of P2P bots is based on two key observations. First, since a P2P bot is part of a P2P network, it exhibits the communication behavior of a normal P2P node but with some distinguishable differences. Second, a P2P bot exhibits different types of network activity compared to regular P2P nodes. Now, using these observations, we identify several important features of a P2P bot. We group the features into two categories, P2P bot communication model and P2P bot behavior model, respectively, and describe them as follows.

*3.1.1. P2P Bot Communication Model.* In a P2P network, a node might attempt to connect to one or more network peers periodically in order to maintain the connection status or to query for data of interest. A P2P bot performs a similar activity but with the key difference being that the P2P bot attempts such connections more actively so as to ensure the connectivity across the P2P botnet. This behavior is uniform across all P2P bots in the P2P botnet. Furthermore, unlike regular P2P communication, where the P2P node attempts connections based on responses received from other peers, the P2P bot attempts to initiate connections proactively. Therefore, at the beginning of its activity, a bot sends connection requests to other bot nodes according to the peer list. A certain amount of such requests fails, because some peers are shutdown or not infected. On the contrary, the success rate is usually high when normal P2P applications send connection requests. Thus, the success rate of connection requests is an important criterion for P2P bot detection. From the observed P2P botnet data, we note that if the success rate of a node connection attempt is below 50%, the node tends to be a bot.

*3.1.2. P2P Bot Behavior Model.* To understand the unique features of P2P botnets, we chose four kinds of real P2P bots available in the wild. Using a controlled virtual environment, with the help of VMware technology, we analyzed these bots. A summary of the results is shown in Table 1. Using this data, we identify the following network features of interest specific

to P2P bot behavior. A feature represents a characteristic of a node in a given time window  $T$ , which is chosen by the system analyst performing the P2P bot detection.

*Number of Different Protocols.* A majority of the P2P bots utilize both UDP and TCP packets in the same flow, that is, send and receive to the same destination port and address from the same source port and source address. For instance, Bot 2 makes SMTP connection using TCP and sends several UDP packets in the same flow. The more important aspect is that the UDP packets outnumber the TCP packets, which is indicative of P2P bot behavior.

*Large Number of Flows.* The number of flows in a P2P bot is higher than for a normal P2P node. For instance, in Bot 3, there is a high amount of ICMP traffic from the P2P bot towards port 53 of random IP addresses on the Internet. These packets correspond to discovery packets intended to locate new targets for the P2P bot infection. The number of flows reflects the degree of extensive connections with other nodes.

*Large Number of Packets.* In P2P botnets, due to the P2P topology, there are a large number of packets exchanged among the P2P nodes. This differs from a server-client botnet where the communication among client nodes is minimal or nonexistent. Therefore, a P2P botnet typically generates a higher average in the number of packets exchanged. This behavior is observed uniformly across all the bots we have analyzed.

*Average Packet Length.* Since the P2P bots need to exchange updates or instructions regularly with other bots, the size of the packets is necessarily small to avoid detection by the IDS system. However, this results in many continuous uniform small packets which is a useful feature for P2P bot detection. For instance in Bots 1, 2, and 3, there is constant communication among peers regarding P2P status and other features.

*Ratios of Packets Exchanged.* The number of packets sent and received by P2P bots exhibits certain uniformity across all the bots. These values differ considerably from the regular P2P communication as observed from our analysis of the sample bots. One important feature of interest is ratio of number of packets sent to the number of packets received, RNP, where a higher value of RNP indicates that these nodes are more active than other nodes. Another feature of importance is ratio of the average of length of packets sent to the average of length of packets received, RLP, where the value of RLP is an indicator to the peering relationship between nodes, a lower value indicating a normal P2P node and a higher value indicating a P2P node controlled by some other peer nodes.

Table 2 lists the seven features we have selected for the purposes of P2P bot detection. In this list, the features, such as the source and destination IP addresses, are extracted directly from the TCP/UDP headers. Other features, such as the number of protocols used, require additional processing and computation. Therefore, we perform flow monitoring on

TABLE 1: Sample bot analysis and behavior.

Name	Host behavior	Network behavior	Remark
Bot 1. Phatbot	(1) Modify the registry (2) Add startup item (3) Modify a file (4) Terminate antivirus thread	(1) Start the IRC thread (2) Start the P2P server thread (3) Start the P2P client thread	(1) Modify a file named host in system directory (2) Start the thread of IRC client, and connect to IRC server. (3) In order to improve the communication of p2p, start both client thread and server thread
Bot 2. Zhelatin .zy	(1) Modify the registry (2) Add a startup item (3) Copy file	(1) Connect to SMTP server (2) UDP connection	(1) In order to a bot's propagation, copy the bot itself to the shared directory (2) Connect to SMTP Server by SMTP thread (3) A lot of UDP connections with both the same source port and the random target port
Bot 3. Sinit		(1) UDP protocol (2) A high ICMP traffic (3) Sending packets to port 53	(1) Sending special discovery packets to port 53 of random IP addresses on the Internet.
Bot 4. Nugache	(1) Modify the registry	(1) Open TCP port 8 (2) Encrypted data transmission	(1) Modify the registry and install the list with hosts into Windows's registry. (2) Has a static list of IP addresses (20 initial peers) to which it will try to connect on TCP port 8. (3) The exchanged data is encrypted.

TABLE 2: Features for node-based analysis.

Feature	Description
(1) Node	Computer address for transmitting information
(2) NP	Number of protocols used for time interval
(3) NF	Number of flows used for time interval
(4) NPS	Number of packets sent for time interval
(5) ALPS	Average length of packets sent
(6) RNP	Ratio of number of packets sent to number of packets received for time interval
(7) RLP	Ratio of average sending packets length to average receiving packets length for time interval

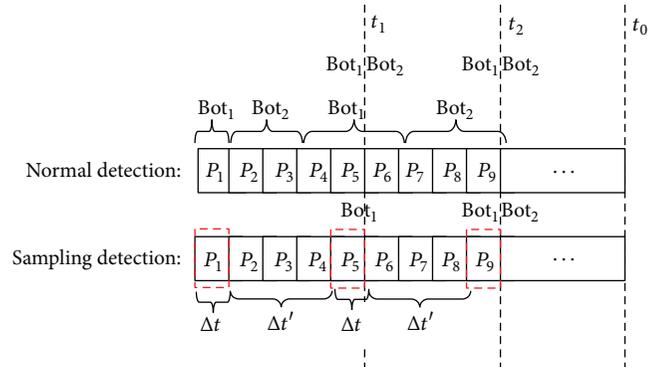


FIGURE 4: Effect of sampling on bot detection.

individual nodes to extract the desired features. We describe our flow monitoring approach next.

**3.2. Flow Monitoring.** Our node-based detection approach monitors the flows at each node to extract the network features identified in Table 2. We note that, even with this requirement, the node-based analysis is still much more efficient than signature-based analysis. But there are two challenges in flow monitoring. First, analyzing each flow at a node requires capturing each packet. The process of packet capturing is known to suffer from high packet losses. Specialized hardware might be required to handle the packet loss which may prove to be an expensive option. Second, as some features, like NFS, NP, RNP, RLP and so forth, cannot be obtained from packet headers directly, the packets need to be stored and processed. This results in a large storage overhead for the bot detection process. To overcome these two challenges, we adopt the sampling approach proposed in [21, 22].

Specifically, for each flow at the node, we sample the packets in a periodic manner, thereby reducing the number of packets that need to be captured. However, reducing the number of captured packets can reduce accuracy of bot detection. To evaluate the impact of sampling on bot detection, we compare continuous packet capturing against sampling and show the results in Figure 4. This figure shows that the normal capturing can possibly detect more bots than sampling detection within the same time window. For example, at around time  $t_1$ , the normal packet capturing detects two bots while the sampling detects one bot. However, eventually the two methods detect the same number of bots after a few time windows. For example, at around time  $t_2$ , both normal capturing and the sampling detect two bots. The asymptotic results are possible due to the cycle limit, that is, the constant reassignment of the C&C server to different P2P bots.

Therefore, using sampling in combination with our bot detection approach can reduce the overhead of flow

monitoring without sacrificing the detection accuracy when considered over a time period. We note that the trade-off in terms of detection time is reasonable as our approach detects all P2P bots within acceptable time windows.

*3.3. Classification Technique and Evaluation Metrics.* For our P2P bot detection approach, we require classification techniques which have high performance in order to support real-time detection goals and at the same time have high detection accuracy. Machine learning classification techniques attempt to cluster and classify data based on feature sets. We have selected the decision tree classifier technique for our evaluation. Decision tree based classifiers exhibit desirably low computational complexity with high performance. In a decision tree, interior nodes represent input features with edges extending from them which correspond to possible values of the features. These edges eventually lead to a leaf node which represents an output variable corresponding to a decision. For our approach, the decision tree is trained based on the real-life P2P bot data using the feature set from Table 2. During the detection phase, the feature set extracted from node's flow information is given to the classifier which essentially classifies this feature set into malicious or nonmalicious feature set.

We consider the standard metrics true positive, TP, true negative, TN, false positive, FP, and false negative, FN, with respect to the classification of feature set into malicious or nonmalicious. The TP and TN values indicate the number of feature sets correctly classified as malicious and benign, respectively. The values FP and FN indicate the number of feature vectors incorrectly classified as malicious and benign, respectively. The true positive rate, TPR, and the false positive rate, FPR, are calculated using the following equations. We define the detection accuracy of our technique using the term, detection rate, DR, and set it to TPR. The true positive rate, TPR, estimates the performance of our P2P botnet detection technique in terms of the probability of a suspicious feature set correctly classified as malicious. On the contrary, the false positive rate, FPR, estimates the probability of a normal traffic being classified as malicious. Finally, we use the standard variable precision to indicate the probability of detection precision of our technique:

$$\begin{aligned} DR = TPR &= \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{FP + TN} \\ \text{Precision} &= \frac{TP}{TP + FP}. \end{aligned} \quad (1)$$

We note that the detection rate, DR, approaches 1 if the false negatives, FN, tend to zero. Similarly, the precision approaches 1 if the false positives, FP, tend to zero. Therefore, both the detection rate and the precision have equivalent importance in the P2P bot detection process.

## 4. Experimental Evaluation

In this section, we describe the evaluation results of our node-based P2P bot detection approach. First, we present the performance results of our scheme showing the detection rate and the precision achieved. Next, we compare our approach with general flow-based detection approach and a state-of-the-art detection tool Bothunter [23] which uses event correlation based analysis.

*4.1. Experimental Methodology and Implementation.* We construct our experimental dataset by combining two separate datasets, one containing malicious traffic related to the Storm botnet and the other containing malicious traffic related to the Waledac botnet, obtained from the French chapter of the honeynet project [24]. Waledac is currently one of the most prevalent P2P botnets and has a highly decentralized communication structure than the Storm botnet. While Storm uses Overnet for P2P communication, Waledac utilizes HTTP communication and a fast-flux based DNS network extensively. The highly distributed nature of Waledac makes it resilient to bot detection approaches. Next, we incorporated two benign datasets into our experimental dataset; one is from the Traffic Lab at Ericsson Research in Hungary [25] and the other is from the Lawrence Berkeley National Lab (LBNL) [26]. The Ericsson Lab dataset contains a large amount of traffic from different applications, including HTTP web browsing behaviors, World of Warcraft gaming packets, and from popular bit-torrent clients such as Azureus. The LBNL trace data provides additional nonmalicious background traffic. As the LBNL is a research institute with a medium-sized enterprise network, their trace data provides a different variety of benign traffic such as web, email, network backup, and streaming media data. This variety of traffic serves as a good example of modeling the day-to-day use of enterprise networks.

We implemented our approach in Java. Our program extracts all node information from a given packet capture (pcap) file, and parses the individual node information into relevant features for use in classification. For classification, we utilized the popular Weka machine learning framework [21] with the decision tree instantiation on our data. We used the standard training approach for training and testing our solution. The key intuition is that the feature vectors correspond to individual node flows and the analysis is done based on these features.

*4.2. Performance of Our Approach.* To evaluate our approach, we used different time windows to represent the amount of flow data analyzed; that is, in a wider time window we analyze more flow data. The time window attempts to align to the bot life cycle, that is, to capture the entire bot specific network activity. While it might seem that larger time windows are preferred, our results show that with reasonable time windows we are able to achieve 99-100% detection rates. We have tested our approach using regular and sampled monitoring approaches for flow capture. We have tabulated the values of detection rate, DR, false positive rate, FPR, and

TABLE 3: Detection rate and precision of node-based detection.

Time interval	Time window	Detection rate	FPR	Precision
0	10	0.997	0.026	0.997
	20	0.998	0.024	0.998
	30	0.998	0.007	0.997
	60	0.999	0	0.999
	180	1	0	1
10	10	0.998	0.004	0.998
	20	0.998	0	0.997
	30	0.999	0	0.998
	60	0.995	0	0.995
	180	0.996	0	0.996
20	10	0.998	0.005	0.998
	20	0.997	0.035	0.996
	30	0.999	0.015	0.998
	60	0.997	0	0.996
	180	1	0	1
30	10	0.998	0.016	0.998
	20	0.998	0.007	0.998
	30	0.997	0.053	0.997
	60	0.999	0	0.998
	180	1	0	1
60	10	0.998	0.053	0.997
	20	0.999	0	0.998
	30	0.997	0	0.997
	60	0.999	0.01	0.998
	180	0.999	0	0.998
180	10	1	0	1
	20	1	0	1
	30	1	0	1
	60	0.999	0.026	0.999
	180	1	0	1

precision in Table 3. We selected time windows of 10, 20, 30, 60, and 180 seconds. For regular flow analysis, the time interval of capture is set to 0 seconds; that is, all packets are captured and analyzed. For sampled flow capturing, we set the time intervals of capturing to 10, 20, 60, and 180 seconds. With large time windows, our sampled approach reduces the processing overhead considerably while retaining the detection accuracy near to optimal.

From Table 3, we make an important observation that for both regular and sampled monitoring the average bot detection rate of our approach is 99%. For regular monitoring, which is the row corresponding to time interval of 0 seconds, there is a gradual reduction in the false positive rate as the time window increases and for the time window of 180 seconds, the FPR falls to 0 and the DR and precision values achieved are 1. This result shows that our approach is capable of detecting P2P bots with 100% accuracy given a sufficient time window. This result combined with the 99% accuracy achieved in other time windows validates our node-based bot detection approach.

For sampled monitoring we chose the time intervals of 0, 10, 20, 30, 60, and 180 seconds. From the results in Table 3, we observe that the sampling has limited effect on DR and precision while the FPR shows a slight increase for smaller time windows. However, the impact of FP on precision is very slight. These results show that our bot detection approach works equally well with sampled flow monitoring and hence can scale to real-time detection for high performance systems. The main advantage of sampling is that it reduces more than 60% of the input raw packet traces while retaining the high detection rates and very low false positive rates 0–2% when viewed in the absolute terms. The number of bots found in different time windows and the length of packets captured are illustrated in Figure 5. We can see from this figure that the amount of data processed, denoted by the length of packets, is considerably smaller for sampling detection while detecting the same number of bots.

*4.3. Comparison with Flow-Based Detection.* We compared our node-based approach with the generic flow-based bot

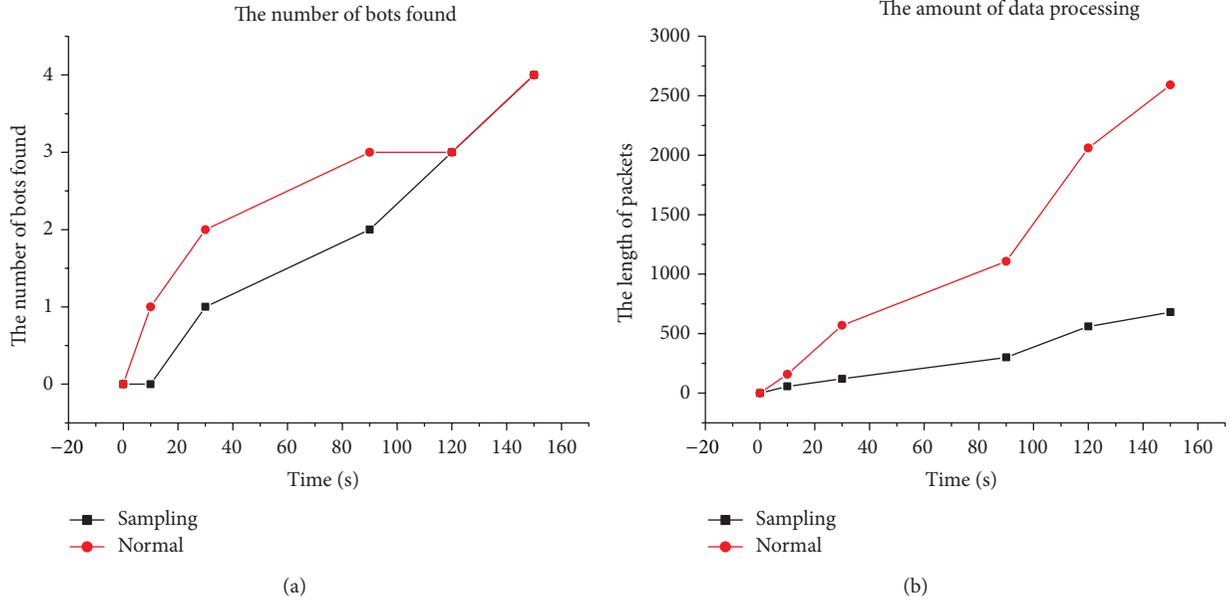


FIGURE 5: Bots detected in different time windows and data processed for detection.

detection approaches [15–18, 20, 27]. Since our node-based approach has broader adaptability to new bot behaviors, we expected our approach to perform better than flow-based approaches. To verify this expectation, we implemented flow-based detection by extracting 12 features from the network flows as shown in Table 4. The summary of our experimental results is shown in Table 5. Our approach has lower false positive rate than flow-based approach and the detection rate is higher. More importantly, our approach has better performance since the sampled detection approach reduces the processing and storage overhead considerably.

**4.4. Comparison with BotHunter.** BotHunter [23] is one of the few botnet detection tools relevant to our work and is publicly accessible. BotHunter mainly consists of a correlation engine that creates associations among alerts generated by Snort [22]. For generating Snort alerts, Bothunter uses two custom plugins, the SLADE plugin for detecting payload anomalies and the SCADE plugin for detecting in/out bound scanning of the network. In addition to regular Snort rule sets, Bothunter uses an enhanced rule set that is specifically designed to detect malicious traffic related to botnet activities, such as egg downloads and C&C traffic. The correlation engine analyzes all the alerts, creates associations among them, and generates a report for botnet infections.

When we tested BotHunter on our dataset, the generated alerts indicated that there is a spambot in the dataset. More specifically, there were three alerts with priority 1 that reported the presence of botnet traffic. But all three alerts pointed to the same IP address corresponding to a machine infected with the Waledac botnet. Moreover BotHunter failed to detect the other machine that was infected with the Storm botnet. Finally, among the 97,043 unique malicious flows in

TABLE 4: Features for flow-based detection comparison.

Attribute	Description
SrcIp	Flow source IP address
SrcPort	Flow source port address
DstIp	Flow destination IP address
DstPort	Flow destination port address
Protocol	Transport layer protocol or “mixed”
APL	Average payload packet length for time interval
PV	Variance of payload packet length for time interval
PX	Number of packets exchanged for time interval
PPS	Number of packets exchanged per second in time interval $T$
FPS	The size of the first packet in the flow
TBP	The average time between packets in time interval
NR	The number of reconnections for a flow
FPH	Number of flows from this address over the total number of flows generated per hour

TABLE 5: Comparison of flow-based and our approach.

Attribute	Flow-based	Our approach
True positive rate	98.3%	100%
False positive rate	0.01%	0%

the system, BotHunter was able to detect only 56 flows which is a very small percentage of the total flows. These results show that our approach performs better than BotHunter in terms of both performance and detection accuracy.

## 5. Conclusion and Future Work

We described node-based detection, a novel direction to detect P2P botnets. Our approach is node-centric and focuses on modeling the network behavior of individual nodes. Our model is constructed by using a combination of the P2P communication model and the observed behavior of real-life P2P botnets. We identified useful features that are indicative of bot behavior and extract these features using the flows at individual nodes. Due to the generality of our P2P bot model, we are able to use sampling to reduce the effort of flow monitoring at individual nodes while retaining high detection accuracy. Since our model is based on observed behavior, our approach is resilient to variations in protocols and payload obfuscations usually employed by P2P botnets. Our experimental results over different time windows show that by choosing an appropriate time window we can achieve 99-100% accuracy of detection. We have also shown that our approach outperforms existing approaches considerably.

We note that it is very important and necessary to design a system that can evaluate the performance of the detection online instead of the present off-line mechanism in our work. It is also important to train the detection system online, instead of an off-line training process so that it is suitable for live deployment. Such a system is ideal for identifying new threats such as zero-day malware. We will explore these issues in our future work. Further, we will explore the use of AIS, Artificial Immune System, to solve the huge number of behavior problems and identify key influence factors in the future.

## Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was funded by the National Natural Science Foundation of China under Grant no. 61373134.

## References

- [1] B. Stone-Gross, M. Cova, B. Gilbert, R. Kemmerer, C. Kruegel, and G. Vigna, "Analysis of a botnet takeover," *IEEE Security and Privacy*, vol. 9, no. 1, pp. 64–72, 2011.
- [2] X. Ma, X. Guan, J. Tao et al., "A novel IRC botnet detection method based on packet size sequence," in *Proceedings of the IEEE International Conference on Communications (ICC '10)*, pp. 1–5, Cape Town, South Africa, May 2010.
- [3] W. Liao and C. Chang, "Peer to peer botnet detection using data mining scheme," in *Proceedings of the International Conference on Internet Technology and Applications (ITAP '10)*, pp. 1–4, Wuhan, China, August 2010.
- [4] C. Mazzariello, "IRC traffic analysis for botnet detection," in *Proceedings of the 4th International Symposium on Information Assurance and Security (IAS '08)*, pp. 318–323, Naples, Italy, September 2008.
- [5] D. A. L. Romana, Y. Musashi, R. Matsuba, and K. Sugitani, "Detection of bot worm-infected PC terminals," *Information*, vol. 10, no. 5, pp. 673–686, 2007.
- [6] P. Wang, S. Sparks, and C. C. Zou, "An advanced hybrid peer-to-peer botnet," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 2, pp. 113–127, 2010.
- [7] X. Dong, F. Liu, X. Li, and X. Yu, "A novel bot detection algorithm based on API call correlation," in *Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '10)*, vol. 3, pp. 1157–1162, Yantai, China, August 2010.
- [8] W. Zilong, W. Jinsong, H. Wenyi, and X. Chengyi, "The detection of IRC botnet based on abnormal behavior," in *Proceedings of the 2nd International Conference on MultiMedia and Information Technology (MMIT '10)*, pp. 146–149, Kaifeng, China, April 2010.
- [9] S. Wang, Q.-J. Du, G.-X. Yu et al., "Method of choosing optimal characters for network intrusion detection system," *Computer Engineering*, vol. 36, no. 15, pp. 140–144, 2010.
- [10] B. Al-Duwairi and L. Al-Ebbini, "BotDigger: a fuzzy inference system for botnet detection," in *Proceedings of the 5th International Conference on Internet Monitoring and Protection (ICIMP '10)*, pp. 16–21, Barcelona, Spain, May 2010.
- [11] Y. Al-Hammadi and U. Aickelin, "Detecting bots based on key logging activities," in *Proceedings of the 3rd International Conference on Availability, Security, and Reliability (ARES '08)*, pp. 896–902, Piscataway, NJ, USA, March 2008.
- [12] M. Crotti, F. Gringoli, P. Pelosato, and L. Salgarelli, "A statistical approach to IP-level classification of network traffic," in *Proceedings of the IEEE International Conference on Communications (ICC '06)*, vol. 1, pp. 170–176, Istanbul, Turkey, July 2006.
- [13] X. Wang, F. Liu, J. Ma, and Z. Lei, "Research of automatically generating signatures for botnets," *Journal of Beijing University of Posts and Telecommunications*, vol. 34, no. 4, pp. 109–112, 2011.
- [14] K. Wang, C. Huang, S. Lin, and Y. Lin, "A fuzzy pattern-based filtering algorithm for botnet detection," *Computer Networks*, vol. 55, no. 15, pp. 3275–3286, 2011.
- [15] J. Kang, Y. Song, and J. Zhang, "Accurate detection of peer-to-peer botnet using multi-stream fused scheme," *Journal of Networks*, vol. 6, no. 5, pp. 807–814, 2011.
- [16] C. Livadas, R. Walsh, D. Lapsley, and W. T. Strayer, "Using machine learning techniques to identify botnet traffic," in *Proceedings of the 31st Annual IEEE Conference on Local Computer Networks (LCN '06)*, pp. 967–974, Tampa, Fla, USA, November 2006.
- [17] H. Choi, H. Lee, and H. Kim, "Botnet detection by monitoring group activities in DNS traffic," in *Proceedings of the 7th IEEE International Conference on Computer and Information Technology (CIT '07)*, pp. 715–720, Aizuwakamatsu, Japan, October 2007.
- [18] D. Liu, Y. Li, Y. Hu, and Z. Liang, "A P2P-botnet detection model and algorithms based on network streams analysis," in *Proceedings of the International Conference on Future Information Technology and Management Engineering (FITME '10)*, vol. 1, pp. 55–58, Changzhou, China, October 2010.
- [19] C. Kolbitsch, P. M. Comparetti, C. Kruegel, E. Kirda, X. Zhou, and X. Wang, "Effective and efficient malware detection at the end host," in *Proceedings of 18th USENIX Security Symposium*, pp. 351–366, USENIX Association, Montreal, Canada, 2009.
- [20] B. Wang, Z. Li, H. Tu, and J. Ma, "Measuring peer-to-peer botnets using control flow stability," in *Proceedings of the*

*International Conference on Availability, Reliability and Security (ARES '09)*, pp. 663–669, Fukuoka, Japan, March 2009.

- [21] I. H. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 3rd edition, 2011.
- [22] M. Roesch, “Snort—lightweight intrusion detection for networks,” in *Proceedings of the 13th USENIX Conference on System Administration (USENIX LISA '99)*, pp. 229–238, USENIX Association, Berkeley, Calif, USA, 1999.
- [23] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee, “BotHunter: detecting malware infection through IDS-driven dialog correlation,” in *Proceedings of the 16th USENIX Security Symposium*, pp. 167–182, 2007.
- [24] French Chapter of Honenynet, <http://www.honeynet.org/chapters/france>.
- [25] G. Szabó, D. Orincsay, S. Malomsoky, and I. Szabó, “On the validation of traffic classification algorithms,” in *Proceedings of the 9th International Conference on Passive and Active Network Measurement (PAM '08)*, pp. 72–81, Cleveland, Ohio, USA, 2008.
- [26] LBNL Enterprise Trace Repository, 2005, <http://www.icir.org/enterprise-tracing>.
- [27] L. Braun, G. Münz, and G. Carle, “Packet sampling for worm and botnet detection in TCP connections,” in *Proceedings of the 12th IEEE/IFIP Network Operations and Management Symposium (NOMS '10)*, pp. 264–271, IEEE, Osaka, Japan, April 2010.

## Research Article

# The Collaborative Search by Tag-Based User Profile in Social Media

Haoran Xie,<sup>1</sup> Xiaodong Li,<sup>1</sup> Jiantao Wang,<sup>2</sup> Qing Li,<sup>2,3</sup> and Yi Cai<sup>4</sup>

<sup>1</sup> Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

<sup>2</sup> Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

<sup>3</sup> Multimedia Software Engineering Research Centre, City University of Hong Kong, Kowloon, Hong Kong

<sup>4</sup> School of Software Engineering, South China University of Technology, Guangzhou 510006, China

Correspondence should be addressed to Yi Cai; [yicai.scut@gmail.com](mailto:yicai.scut@gmail.com)

Received 7 March 2014; Revised 14 May 2014; Accepted 14 May 2014; Published 11 June 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Haoran Xie et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recently, we have witnessed the popularity and proliferation of social media applications (e.g., Delicious, Flickr, and YouTube) in the web 2.0 era. The rapid growth of user-generated data results in the problem of information overload to users. Facing such a tremendous volume of data, it is a big challenge to assist the users to find their desired data. To attack this critical problem, we propose the collaborative search approach in this paper. The core idea is that similar users may have common interests so as to help users to find their demanded data. Similar research has been conducted on the user log analysis in web search. However, the rapid growth and change of user-generated data in social media require us to discover a brand-new approach to address the unsolved issues (e.g., how to profile users, how to measure the similar users, and how to depict user-generated resources) rather than adopting existing method from web search. Therefore, we investigate various metrics to identify the similar users (user community). Moreover, we conduct the experiment on two real-life data sets by comparing the *Collaborative* method with the latest baselines. The empirical results show the effectiveness of the proposed approach and validate our observations.

## 1. Introduction

With the rapid development of web communities, we have witnessed the popularity and proliferation of social media applications in web 2.0 era, which allow the user to annotate and share various kinds of resources like web pages (Delicious (<http://www.delicious.com/>)), movies (MovieLens (<http://www.movielens.org/>)), and images (Flickr (<http://www.flickr.com/>)). On one hand, the tremendous user-generated data provides the opportunity to easily communicate and share information with each other; on the other hand, such a big volume of data results in the problem of information overload to the users. Facing such a tremendous volume of data, it is a big challenge to assist the users to find their desired data.

To attack this critical problem, we propose the collaborative search approach in this paper. The core idea is that similar users may have common interests so as to help users to find their demanded data. Similar research [1–3]

has been conducted on the user log analysis in web search. However, the rapid growth and change of user-generated data in social media require us to discover a brand-new approach to address the unsolved issues (e.g., how to profile users, how to measure the similar users, and how to depict user-generated resources) rather than adopting existing method from web search. Therefore, we investigate the following research questions in this paper:

- (i) how to depict users and resources in the social media;
- (ii) how to measure the user similarity in the social media;
- (iii) how to assist users to find their interested data (resources) by similar users in the social media.

The remaining parts of this paper are structured as follows. In Section 2, the related works on collaborative search and social media are reviewed. We introduce the framework of the collaborative search for social media in Section 3. The experiments are conducted and the corresponding results

are analyzed in Section 4. Finally, we summarize our work and discuss the potential directions for future research in Section 5.

## 2. Related Works

In this section, we review relevant works in the areas of collaborative search and social media.

*2.1. Collaborative Search.* Collaborative search (a.k.a., social search) has been intensively studied to facilitate the search performance in the web by incorporating similar user search behaviors. In [4], R. B. Almeida and V. A. F. Almeida devised a community-aware search engine, which incorporated community information as another evidence of relevance and improved the conventional content-based ranking strategies up to 48% of the average precision. Park and Ramamohanarao [1] proposed a popularity score for multiresolution community to generalize and improve PageRank algorithm for web search. Smyth [2] deployed a community-based search engine to collect search behavior for a community (e.g., a department) and found that the search quality can be significantly improved by the community members. Moreover, McNally et al. [3] described the results of real user study, which demonstrated the benefits of a collaborative search method (called HeyStaks) making use of personalization and social networking. In [5], HeyStaks was further enhanced and improved by the recommendation method and the reputation model in terms of the click-through rate. Morris et al. [6] explored the design space for collaborative search systems on interactive tabletops. Boydell and Smyth [7] described a technique for summarizing search results that harnesses the collaborative search behavior of communities of like-minded searchers to produce snippets that are more focused on the preferences of the searchers. Ju and Xu [8] proposed a novel collaborative recommendation approach based on users' clustering by using artificial bee colony algorithm. Xue et al. [9] developed a user language model for the personalized collaborative search, so that the behaviors of the group users can be utilized to improve the search performance. Fu et al. [10] exploited the characteristics of local communities to facilitate collaborative recommendations. Cai et al. [11] further improve the conventional collaborative filtering methods by borrowing the idea of "object typicality" in cognitive science.

*2.2. Social Media.* In this subsection, we mainly focus on one mainstream of social media applications: collaborative tagging systems. Previous research on the collaborative tagging system can be mainly divided into two classes. One is trying to find the main patterns and characteristics of the user-generated tags and resources in such social media communities. In [12], the tag generation and usage patterns were investigated and analyzed by Golder and Huberman. To reveal the power of the tags, Bischoff et al. [13] studied various aspects of the social tagging throughout a comprehensive survey on many real tagging data sets. Moreover, Gupta et al. [14] investigated and summarized the main patterns of tagging

behaviors and the popular tagging techniques. Carmel et al. [15] presented a folksonomy-based term extraction method, called tag-boost, which boosts terms that are frequently used by the public to tag content. Wei et al. [16] analyzed and studied the cooperation rate in cooperation social networks via a two-phase Heterogeneous Public Goods Game (HPGG) model. Ye et al. [17] studied the feasibility of social network research technologies on process recommendation and built a social network system of processes based on the features' similarities. The other class is to apply these characteristics and patterns in various applications (e.g., social media resource search or recommendation). Bao et al. [18] presented two novel algorithms called SocialSimRank (SSR) and SocialPageRank (SPR) by incorporating social annotation to facilitate web search. Three approaches (naive, cooccurrence, and adaptive) were proposed by Michlmayr and Cayzer [19] to construct tag-based profiles and assist in information access. Xu et al. [20] measured the semantic relatedness between Flickr images from the tag-based perspectives. Balali et al. [21] presented a supervised approach to predicting and reorganizing the hierarchical structure of conversation threads for user-generated text in social media. The tag-based profiles were further studied and investigated to facilitate personalized search [22–24]. Furthermore, a source-initiated on-demand routing algorithm, which can assist users to communicate in mobile wireless sensor network, was proposed by Mao and Zhu [25].

## 3. Methodology

In this section, we will introduce and discuss the proposed collaborative search method for social media. First of all, the research problem is formulated so that the clear picture of the methodology is given. Then, the methodology can be further divided into three subprocesses, which are user and resource profiling, user community discovery, and collaborative ranking.

*3.1. Problem Formulation.* Intuitively, the collaborative search is to consider the search history of similar users as evidence of relevance and rerank the resources [3, 5]. Specifically, the research problem of collaborative search can be formulated as a mapping function  $\theta$  as follows:

$$\theta : U \times Q \times C \times R \longrightarrow S, \quad (1)$$

where  $U$  is the set of users,  $Q$  is the set of queries,  $C$  is the set of user communities (similar user clusters), and  $R$  is the set of resources; the ultimate goal of function  $\theta$  is to map the above four elements to ranking score  $S$ . In the next three subsections, we will detail how to model user and resource, discover similar users, and perform the collaborative search.

*3.2. User and Resource Profiling.* To depict the user and resource, we adopt the bag-of-tags (BOT) paradigm to construct user and resource profiles, which is similar to our previous research in [23]. The paradigm is mainly based on the assumption that the tags used by user (or annotated to resource) reflect the user's interest (or resource feature) to

some extent. Formally, the user and resource profiles are defined as follows.

*Definition 1.* The *user profile* of user  $a$  is a vector of tag : value pairs, which is denoted by  $\vec{U}_a$  as follows:

$$\vec{U}_a = (t_{a,1} : v_{a,1}, t_{a,2} : v_{a,2}, \dots, t_{a,n} : v_{a,n}), \quad (2)$$

where  $t_{a,x}$  is a tag used by user  $a$ ,  $n$  is the total amount of tags used by this user, and  $v_{a,x}$  means the degree of interest for user  $a$  on this tag  $t_{a,x}$ . Similarly, the resource profile is also defined by the BOT paradigm below.

*Definition 2.* The *resource profile* of resource  $i$  is also a vector of tag : value pairs, which is denoted by  $\vec{R}_i$  as follows:

$$\vec{R}_i = (t_{i,1} : w_{i,1}, t_{i,2} : w_{i,2}, \dots, t_{i,m} : w_{i,m}), \quad (3)$$

where  $t_{i,y}$  is a tag annotated to the resource  $i$ ,  $m$  is the total number of tags annotated to this resource, and  $w_{i,y}$  indicates the degree of relevance for tag  $t_{i,y}$  to the resource. The weight of each in both user and resource profiles can be obtained by various methods (e.g., tag-frequency (TF) [26], tag-frequency and inverse resource frequency (TF-IRF) [22], best match 25 (BM 25) [27], and normalized tag-frequency (NTF) [23]). To compare and find the best paradigm for our problem, we compare these different paradigms in the experiment (see Section 4).

**3.3. User Community Discovery.** Users have similar interests and/or intentions usually form user groups (communities) explicitly and implicitly; the community-based information can be adopted and utilized to improve the efficiency and effectiveness of various user navigational behaviors [24, 28]. There are many existing techniques (e.g., the topic model [29], semantic space [30], and Gaussian mixture model [31]) that can be employed to discover user community. However, the main shortage for these community discovering approaches is that the time complexities of method are exponentially increased (e.g.,  $\Theta(n^k)$  in [29], where  $n$  is the number of tags and  $k$  is the number of communities), which are very time consuming [24] and unapplicable in current big data era. To tackle this problem, we propose a lightweight method to discover the user community with the acceptable level of time complexity. The core idea is to precluster off-line firstly and then discover the user community for the user according to his/her current issued query and user profile.

**3.3.1. Off-Line Clustering.** The purpose of off-line stage is to precluster the similar users and classify them into some user communities. The existing clustering approaches [29–31] can be adopted in this step as it performs off-line. However, the performance of various clustering methods has been studied in [24]. Therefore, we employ a conventional clustering method K-means to clearly investigate the performance of

various user similarity measurements. Intuitively, a straightforward method is to adopt Jaccard and Ochiai coefficient as follows:

$$\text{Sim}_J(\vec{U}_a, \vec{U}_b) = \frac{|\vec{U}_a \cap \vec{U}_b|}{|\vec{U}_a \cup \vec{U}_b|}, \quad (4)$$

$$\text{Sim}_O(\vec{U}_a, \vec{U}_b) = \frac{|\vec{U}_a \cap \vec{U}_b|}{\sqrt{|\vec{U}_a| \times |\vec{U}_b|}}.$$

For the purpose of clustering by K-means, the above similarities are required to be converted to distance (e.g., using  $\text{Dist}() = (1/\text{Sim}()) - 1$ ). Since these measurements focus on the tag and neglect the relevance of each tag in the user profile, they are named as *tag-level distance*. If we focus on the degree of relevance, the Euclidean distance and Manhattan distance (named as *value-level distance*) can be used as follows:

$$\text{Dist}_E(\vec{U}_a, \vec{U}_b) = \sqrt{\sum_{i,j=(1,1)}^{(n_a, n_b)} (v_{a,i} - v_{b,j})^2}, \quad (\forall (i, j), t_{a,i} = t_{b,j}), \quad (5)$$

$$\text{Dist}_M(\vec{U}_a, \vec{U}_b) = \sum_{i,j=(1,1)}^{(n_a, n_b)} |v_{a,i} - v_{b,j}|, \quad (\forall (i, j), t_{a,i} = t_{b,j}).$$

In our earlier work [23], we have found that matching in both tag-level and value-level can contribute to finding relevant resources. Thus, we propose *hybrid-level distance* by integrating distances in tag-level and value-level as follows:

$$\text{Dist}_H(\vec{U}_a, \vec{U}_b) = e^{(1 - \text{Sim}_J(\vec{U}_a, \vec{U}_b)) / \text{Sim}_J(\vec{U}_a, \vec{U}_b) + \text{Dist}_E(\vec{U}_a, \vec{U}_b)}, \quad (6)$$

where  $(1 - \text{Sim}_J(\vec{U}_a, \vec{U}_b)) / \text{Sim}_J(\vec{U}_a, \vec{U}_b)$  and  $\text{Dist}_E(\vec{U}_a, \vec{U}_b)$  are the distances in tag-level and value-level, respectively (there are other combinations for the hybrid-level distance and we select one of them to illustrate the usefulness of the hybrid of both tag-level and value-level distances). After selecting a particular distance (similarity) measurement above, K-means is then performed to discover  $k$  user communities (clusters). Formally, the community profile is depicted by members and their relevance as follows.

*Definition 3.* The *community profile* of community  $k$  is a vector of member : value pairs, which is denoted by  $\vec{C}_k$  as follows:

$$\vec{C}_k = (\vec{U}_{k,1} : d_{k,1}, \vec{U}_{k,2} : d_{k,2}, \dots, \vec{U}_{k,p} : d_{k,p}), \quad (7)$$

where  $\vec{U}_{k,z}$  is the user profile of community member (user)  $z$  and  $d_{k,z}$  is the distance of the user to the centroid of community  $k$ .

**Input:** The cluster set  $\{\vec{C}_1, \vec{C}_2, \dots, \vec{C}_k\}$ , their center set  $\{\vec{U}_1^*, \vec{U}_2^*, \dots, \vec{U}_k^*\}$ , the current user  $\vec{U}_{x,i}$ , issued query  $\vec{q}$  and the threshold  $t_\alpha$  and  $t_\beta$ .

**Output:** The updated user profile  $\vec{U}'_{x,i}$  and the on-line discovering community  $\vec{C}_o$ .

- (1) **if**  $\text{Dist}(\vec{U}_{x,i}, \vec{q}) > t_\alpha$  **then**
- (2)    $\vec{U}_{x,i} \leftarrow \vec{U}_{x,i} \cap \vec{q}$
- (3)    $\vec{U}'_{x,i} \leftarrow \vec{U}_{x,i}$
- (4)   Update the degree of relevance by the selected paradigm.
- (5)   **if**  $\text{Dist}(\vec{U}'_{x,i}, \vec{U}_x^*) > t_\beta$  **then**
- (6)     **for**  $j = 1$  to  $k$  **do**
- (7)       Find the community  $\vec{C}_{\min}$  by calculate  $\text{Dist}(\vec{U}'_{x,i}, \vec{U}_j^*)$
- (8)     **end for**
- (9)      $\vec{C}_o \leftarrow \vec{C}_{\min}$
- (10)  **end if**
- (11) **else**
- (12)    $\vec{C}_o \leftarrow \vec{C}_x$
- (13)    $\vec{U}'_{x,i} \leftarrow \vec{U}_{x,i}$
- (14) **end if**

ALGORITHM 1: Algorithm of on-line community discovery.

TABLE 1: The details of FMRS and Movielens data set.

	Users number	Resources number	Tags number	Domain
FMRS	203	500	7889	Cooking recipes
Movielens	71567	10681	10000054	Movies

**3.3.2. On-Line Discovering.** In off-line clustering stage, a user is classified to a particular user community according to his/her user profile. While in on-line discovering stage, we cannot fix the user to his/her preallocated community as the search context may be different or even totally irrelevant to it [32]. To avoid this case, we firstly compare the query with the user profile to examine whether the current query context is relevant to the user profile or not. Then, we discover a new user community for the user if the current issued query is not relevant to his/her current community. Finally, the user profile is updated by the query terms (tags) accordingly. The detailed algorithm is shown in Algorithm 1. Note that the time complexity of the algorithm is quite acceptable as it only has the time complexity  $\Theta(k)$  and is much faster and more scalable than the on-line methods of  $\Theta(n^k)$ .

**3.4. Collaborative Ranking.** The last stage is to obtain the ranking score for the resource. Since the user ( $U$ ), query ( $Q$ ), community ( $C$ ), and resource ( $R$ ) are obtained and defined, we can adopt cosine measurement as the ranking function  $\theta$  as

$$\theta(\vec{U}_i, \vec{q}, \vec{C}_o, \vec{R}_x) = \frac{\vec{R}_x \cdot \vec{U}_i}{\|\vec{R}_x\| \|\vec{U}_i\|} \cdot \frac{\vec{R}_x \cdot \vec{q}}{\|\vec{R}_x\| \|\vec{q}\|} \cdot \frac{\vec{R}_x \cdot \vec{U}_o^*}{\|\vec{R}_x\| \|\vec{U}_o^*\|}, \quad (8)$$

where  $\vec{U}_i$  and  $\vec{R}_x$  are given in Definitions 1 and 2,  $\vec{C}_o$  and  $\vec{U}_o^*$  are obtained in Algorithm 1, and  $\vec{q}$  is the query.

The greater value of  $\theta$  function for a resource indicates the higher relevance to user interests and his/her current search intentions.

## 4. Experiment

In this section, we conduct the experiment on FMRS [32] and Movielens (<http://www.grouplens.org/node/73>) data sets to evaluate the performance of the proposed method.

**4.1. Data Sets.** The details of the two data sets are shown in Table 1. The main reason for selecting these two data sets is that they are in different domains (cooking recipes and movies) and they have different scales ( $10^4$  versus  $10^7$  tags) so that we can examine the performance of the proposed method in both large and small scales for multiple domain applications in social media. To evaluate the proposed method, we split the data sets into 80% and 20% as training and testing sets, respectively. In training stage, the profiles and models are learned, while we examine whether the learned models can predict the right target resource by the given query terms (tags) from the testing set in the testing stage.

**4.2. Metrics.** Two widely adopted metrics are used in the experiments, which are  $P@N$  (Precision @ $N$ ) [33] and MRR (Mean Reciprocal Rank).  $P@N$  is mainly to measure

the accuracy of a particular search strategy, which is given as follows:

$$P@N = \sum_{i=1}^n \frac{p(q_i)}{n}, \quad (9)$$

$$p(q_i) = \begin{cases} 1, & \text{if } p(R_i^q) \leq N \\ 0, & \text{if } p(R_i^q) > N, \end{cases}$$

where  $p(R_i^q)$  is the position of target resource for query  $q_i$  and  $n$  is the total number of tuples in the testing set. The metric MRR reflects how quickly a search strategy can assist users in finding their desired resources, which is given as follows:

$$\text{MRR} = \frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{\text{rank}(r_i^q)}. \quad (10)$$

**4.3. Baselines.** To verify the effectiveness of the proposed method, there are three state-of-the-art baselines for comparison. We denote the proposed method as “*Collaborative*” to simplify the notations. The abbreviations and details of the three baselines are introduced as follows.

**Profile-Based.** The profile-based personalized method was proposed in [27], which neglects the community-based information and only considers the relationships among the user profile, the resource profile, and the query.

**Community-Aware.** The community-aware resource search method [24] takes not only the user community but also the user and resource profiles into consideration. The main shortage of this approach is that the community discovering stage was performed on-line so that it is quite time consuming as discussed in Section 3.3.

**Social.** The social search method was proposed in [5] using similar user queries and the users’ clicked resources as the evidence of relevance. The main difference between the social method and the collaborative search is that they are focusing on the tag-level, while the latter one takes both tag-level and value-level into consideration.

**4.4. Overall Performance.** The overall performance of the metric  $P@N$  in FMRS and Movielens data sets is shown in Figures 1 and 2. We can find that the community-aware baseline achieves the best performance among all four methods with all  $N$  values (from 1 to 30), while the *Collaborative* method performs the second best (less accuracy from 0.7% to 2.3% with community-aware). This is mainly because the community-aware adopts the on-line clustering method which timely updated the user communities and the most relevant one can be obtained by the user. Note that the cost of on-line clustering is quite expensive ( $\Theta(n^k)$ ). Meanwhile, the off-line clustering of *Collaborative* only has the complexity with  $\Theta(k)$ . Moreover, it is observed that the social baseline, which only replies the tag-level profiles, is less accurate than both the *Collaborative* and community-aware ones. Therefore, we argue that considering both tag-level and

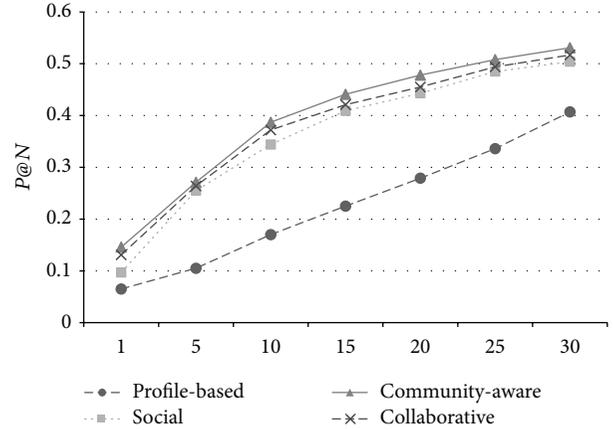


FIGURE 1: The performance of  $P@N$  on FMRS data set.

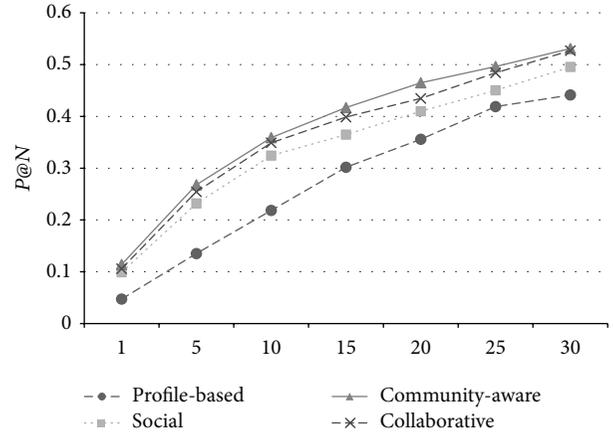


FIGURE 2: The performance of  $P@N$  on Movielens data set.

TABLE 2: The performance of MRR on FMRS and Movielens data sets.

	Profile-based	Social	Community-aware	Collaborative
FMRS	0.183	0.221	<b>0.240</b>	0.229
Movielens	0.109	0.178	<b>0.213</b>	0.194

value-level of the resource and user profiles will improve the search quality. Last but not least, the profile-based, which neglects the user community, has the worst achievement among all methods. It implies that the community-based information is quite useful to assist to find their relevant resources. Furthermore, we observe that the metric MRR has a similar trend to  $P@N$ , as shown in Table 2.

**4.5. Alternative Paradigms and Distances.** As we discussed in Sections 3.1 and 3.2, there are some existing alternative paradigms for user and resource profiling and other distance measurements (tag-level, value-level, and hybrid-level) for user similarity measurement. To investigate the impact of these alternative techniques, we compare their MRR values with various settings in *Collaborative* method. As shown in

TABLE 3: The performance with different paradigms in *Collaborative* on two data sets.

	TF	TF-IRF	BM 25	NTF
FMRS	0.188	0.215	0.196	<b>0.229</b>
Movielens	0.167	0.189	0.173	<b>0.194</b>

TABLE 4: The performance with different metrics in *Collaborative* on two data sets.

	Tag-level		Value-level		Hybrid-level
	Jaccard	Ochiai	Euclidean	Manhattan	Hybrid
FMRS	0.218	0.220	0.226	0.224	<b>0.229</b>
Movielens	0.175	0.179	0.186	0.191	<b>0.194</b>

Table 3, the paradigm of NTF (the values with bold) has the best MRR performance. This result is consistent with our previous study in paradigm comparison [23]. Furthermore, we investigate the various distance measurements for user similarity. According to Table 4, we can observe that the hybrid distance is the most suitable one (with the bold values). It verifies our observation that the hybrid distance is a good tradeoff between tag-level and value-level distances. The value-level distance (Euclidean and Manhattan) gains the second best performance, which indicates that value-level distances are more precise than tag-level ones. We can further observe that the MRR value in tag-level distance (Jaccard and Ochiai) has a similar performance with social baseline, which also focuses on tag-level only.

## 5. Conclusion

In this research, we have proposed a lightweight user clustering method to find similar users in social media. The performance of collaborative search based on this clustering method is a bit less accurate than the on-line clustering community approach. However, the trade-off here is that we have gained much more scalability with less time complexity (from  $\Theta(n^k)$  to  $\Theta(k)$ ). Moreover, the various distance measurements in three levels (tag-level, value-level, and hybrid-level) have been investigated. We believe that the proposed hybrid distance metric is the most suitable to measure the user similarity. Furthermore, we have confirmed the performance with NTF paradigm, which is the proper one to construct user and resource profiles. In the future study, we plan to find out the most important feature of score function  $\theta$  so as to further improve the search quality in social media.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The research described in this paper has been supported by a Strategic Research Grant of the City University of Hong Kong

(Project no. 7004046), the National Natural Science Foundation of China (Grant no. 61300137), the Guangdong Natural Science Foundation of China (no. S2013010013836), and the Fundamental Research Funds for the Central Universities, SCUT (no. 2014ZZ0035).

## References

- [1] L. A. F. Park and K. Ramamohanarao, "Mining web multi-resolution community-based popularity for information retrieval," in *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07)*, pp. 545–554, ACM, November 2007.
- [2] B. Smyth, "A community-based approach to personalizing web search," *Computer*, vol. 40, no. 8, pp. 42–50, 2007.
- [3] K. McNally, M. P. O'Mahony, B. Smyth, M. Coyle, and P. Briggs, "Social and collaborative web search: an evaluation study," in *Proceedings of the 15th ACM International Conference on Intelligent User Interfaces (IUI '11)*, pp. 387–390, ACM, February 2011.
- [4] R. B. Almeida and V. A. F. Almeida, "A community-aware search engine," in *Proceedings of the 13th International World Wide Web Conference (WWW '04)*, pp. 413–421, ACM, May 2004.
- [5] B. Smyth, M. Coyle, and P. Briggs, "Heystaks: a real-world deployment of social search," in *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12)*, pp. 289–292, ACM, September 2012.
- [6] M. R. Morris, D. Fisher, and D. Wigdor, "Search on surfaces: exploring the potential of interactive tabletops for collaborative search tasks," *Information Processing and Management*, vol. 46, no. 6, pp. 703–717, 2010.
- [7] O. Boydell and B. Smyth, "Social summarization in collaborative web search," *Information Processing and Management*, vol. 46, no. 6, pp. 782–798, 2010.
- [8] C. Ju and C. Xu, "A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm," *The Scientific World Journal*, vol. 2013, Article ID 869658, 9 pages, 2013.
- [9] G.-R. Xue, J. Han, Y. Yu, and Q. Yang, "User language model for collaborative personalized search," *ACM Transactions on Information Systems*, vol. 27, no. 2, article a11, 2009.
- [10] Y. Fu, Q. Liu, and Z. Cui, "A collaborative recommend algorithm based on bipartite community," *The Scientific World Journal*, vol. 2013, Article ID 295931, 14 pages, 2013.
- [11] Y. Cai, H. F. Leung, Q. Li, H. Min, J. Tang, and J. Li, "Typicality-based collaborative filtering recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 766–779, 2014.
- [12] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, 2006.
- [13] K. Bischoff, C. S. Firan, W. Nejdil, and R. Paiu, "Can all tags be used for search?" in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pp. 193–202, ACM, October 2008.
- [14] M. Gupta, R. Li, Z. Yin, and J. Han, "Survey on social tagging techniques," *ACM SIGKDD Explorations Newsletter*, vol. 12, pp. 58–72, 2010.
- [15] D. Carmel, E. Uziel, I. Guy, Y. Mass, and H. Roitman, "Folksonomy-based term extraction for word cloud generation," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, article 60, 2012.

- [16] G. Wei, P. Zhu, A. V. Vasilakos, Y. Mao, J. Luo, and Y. Ling, "Cooperation dynamics on collaborative social networks of heterogeneous population," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1135–1146, 2013.
- [17] Y. Ye, J. Yin, and Y. Xu, "Social network supported process recommender system," *The Scientific World Journal*, vol. 2014, Article ID 349065, 8 pages, 2014.
- [18] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pp. 501–510, ACM, May 2007.
- [19] E. Michlmayr and S. Cayzer, "Learning user profiles from tagging data and leveraging them for personal (ized) information access," in *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW '07)*, 2007.
- [20] Z. Xu, X. Luo, L. Mei, and C. Hu, "Measuring semantic relatedness between flickr images: from a social tag based view," *The Scientific World Journal*, vol. 2014, Article ID 758089, 12 pages, 2014.
- [21] A. Balali, H. Faili, and M. Asadpour, "A supervised approach to predict the hierarchical structure of conversation threads for comments," *The Scientific World Journal*, vol. 2014, Article ID 479746, 23 pages, 2014.
- [22] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, pp. 155–162, ACM, July 2008.
- [23] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile in collaborative tagging systems," in *Proceedings of the 19th International Conference on Information and Knowledge Management and Co-Located Workshops (CIKM '10)*, pp. 969–978, ACM, October 2010.
- [24] H.-R. Xie, Q. Li, and Y. Cai, "Community-aware resource profiling for personalized search in folksonomy," *Journal of Computer Science and Technology*, vol. 27, no. 3, pp. 599–610, 2012.
- [25] Y. Mao and P. Zhu, "A source-initiated on-demand routing algorithm based on the thorup-zwick theory for mobile wireless sensor networks," *The Scientific World Journal*, vol. 2013, Article ID 283852, 9 pages, 2013.
- [26] M. G. Noll and C. Meinel, "Web search personalization via social bookmarking and tagging," in *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference (ISWC'07/ASWC '07)*, pp. 367–380, Springer, 2007.
- [27] D. Vallet, I. Cantador, and J. M. Jose, "Personalizing web search with folksonomybased user and document profiles," pp. 420–431, *Proceedings of the 32nd European conference on Advances in Information Retrieval (ECIR '10)*, 2010.
- [28] J. Teevan, M. R. Morris, and S. Bush, "Discovering and using groups to improve personalized search," in *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM '09)*, pp. 15–24, ACM, February 2009.
- [29] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*, pp. 1055–1060, IEEE, Pisa, Italy, December 2008.
- [30] W. Xian, Z. Lei, and Y. Yong, "Exploring social annotations for the semantic web," in *Proceedings of the 15th International Conference on World Wide Web (WWW '06)*, pp. 417–426, ACM, May 2006.
- [31] H. Zhang, C. Lee Giles, H. C. Foley, and J. Yen, "Probabilistic community discovery using hierarchical latent gaussian mixture model," in *Proceedings of the 22nd AAAI Conference on Artificial Intelligence and the 19th Innovative Applications of Artificial Intelligence Conference (AAAI-07/IAAI '07)*, pp. 663–668, July 2007.
- [32] H. Xie, Q. Li, and X. Mao, "Context-aware personalized search based on user and resource profiles in folksonomies," in *Proceedings of the Web Technologies and Applications*, pp. 97–108, 2012.
- [33] R. W. White, P. Bailey, and L. Chen, "Predicting user interests from contextual information," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pp. 363–370, ACM, July 2009.

## Research Article

# The Application of Baum-Welch Algorithm in Multistep Attack

Yanxue Zhang,<sup>1</sup> Dongmei Zhao,<sup>2</sup> and Jinxing Liu<sup>3</sup>

<sup>1</sup> College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050000, China

<sup>2</sup> College of Information Technology, Hebei Normal University, Shijiazhuang 050000, China

<sup>3</sup> The First Aeronautics College of PLAAF, Xinyang 464000, China

Correspondence should be addressed to Dongmei Zhao; zhaodongmei666@126.com

Received 8 April 2014; Accepted 6 May 2014; Published 28 May 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Yanxue Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The biggest difficulty of hidden Markov model applied to multistep attack is the determination of observations. Now the research of the determination of observations is still lacking, and it shows a certain degree of subjectivity. In this regard, we integrate the attack intentions and hidden Markov model (HMM) and support a method to forecasting multistep attack based on hidden Markov model. Firstly, we train the existing hidden Markov model(s) by the Baum-Welch algorithm of HMM. Then we recognize the alert belonging to attack scenarios with the Forward algorithm of HMM. Finally, we forecast the next possible attack sequence with the Viterbi algorithm of HMM. The results of simulation experiments show that the hidden Markov models which have been trained are better than the untrained in recognition and prediction.

## 1. Introduction

Currently, the network security situation is increasingly sophisticated and the multistep network attack has become the mainstream of network attack. 2012 Chinese Internet network security reports released by the National Computer Network Emergency Response Technical Team Coordination Center of China (CNCERT/CC) show that the two typical multistep attacks: warms and distributed denial of service (DDOS) [1] account for 60% of overall network attacks. Multistep attack [2] means that the attacks apply multiple attack steps to attack the security holes of the target itself and achieve the devastating blow to the target. There are three features of attack steps of multistep attack. (1) In the multistep attack, there is a casual relationship between multiple attack steps. (2) The attack steps of multistep attack have the property of time sequence [3]. (3) The attack steps of multistep attack have the characteristics of uncertainty [4].

Multistep attack is one of the main forms of network attack behaviors, recognizing and predicting multistep attack that laid the foundation of active defense, which is still one of the hot spots nowadays. Literature (application of

hidden Markov models to detect multistep network attacks) proposed a method to recognize multistep attack based on hidden Markov model.

Markov model literature (improving the quality of alerts and predicting intruder's next goal with hidden colored Petri-net) introduced the concept of attack "observation," but both stayed in the specific attack behaviors, which have some limitations. Current research on the approaches to forecast multistep attack behaviors mainly includes four types: (1) the approach to forecasting multistep attack based on the antecedents and consequences of the attack [5]. It applies the precursor subsequent relationship of the event, to forecast the attacker wants to implement attacks in the near future. Because of the complexity and the diversity of the attack behaviors, this approach is difficult to achieve. (2) The approach to forecasting multistep attack based on hierarchical colored Petri-nets (HCPN) applies the raw alerts by Petri-nets and considers that the attack intention is inferred by raw alerts [4]. But this approach focuses on the intrusion detection of multistep attack behaviors. (3) The approach to forecasting multistep attack based on Bayes game theory could forecast the probability that the attackers choose to

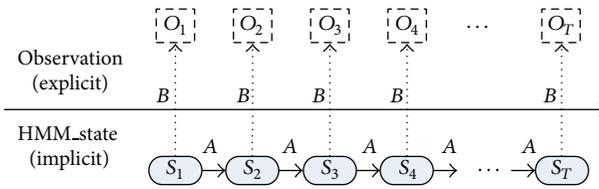


FIGURE 1: Model of recognizing and forecasting multistep attack based on hidden Markov model.

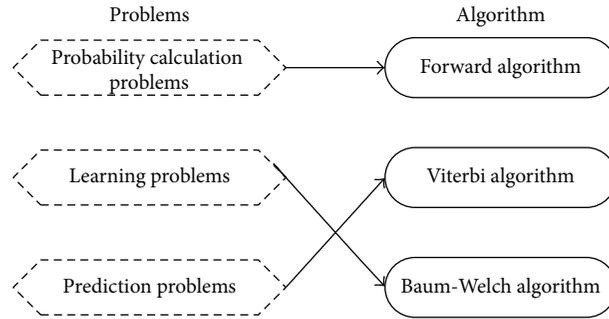


FIGURE 2: Correspondence between the problems and algorithms of hidden Markov model.

attack and the probability that the defenders choose to defend in the next stage rationally [6, 7]. However, in current study, only two-person game model is established, so this approach has some limitations. (4) The approach to forecasting multistep attack based on attack intention [3, 8] uses extended-directed graph to describe the logical relationship between attack behaviors and forecasts the next stage by logical relationship. The shortcoming of this approach is that it is difficult to determine the matching degree of the multistep attack. At the same time, there exists a certain degree of subjectivity in recognizing and forecasting multistep attack. In this regard, we integrate the attack intentions and hidden Markov model and propose a method to forecast multistep attack based on hidden Markov model. Firstly, we train the existing hidden Markov model(s) by the Baum-Welch algorithm of HMM. Then we recognize the alert belonging to attack scenarios with the Forward algorithm of HMM. Finally, we forecast the next possible attack sequence with the Viterbi algorithm of HMM. Simulation experiments results show that the hidden Markov models which have been trained are better than the untrained in recognition and prediction.

**2. Hidden Markov Model**

Hidden Markov model was first proposed by Baum and Petrie in 1966. It is a statistical model, which is used to describe a Markov process which contains a hidden parameter [9]. The research object of this model is a data sequence; each value of this data sequence is called an observation. Hidden Markov model assumes that there still exists another sequence which hides behind this data sequence; the other sequence consists of a series of states. Each observation occurs in a state, the state cannot be observed directly, and the features of the state can only be inferred from the observations.

A complete hidden Markov model (HMM) is usually represented by a triple  $\lambda = (A, B, \pi)$ , which includes the following five elements:

- (1) a finite state, which is represented by the set  $S$ , where  $S = \{s_1, s_2, \dots, s_N\}$  and, at time  $t$ , the state is denoted by  $q_t$ ;
- (2) the set of observations, which is represented by the set  $O$ , where  $O = \{o_1, o_2, \dots, o_T\}$ ;
- (3) the state transition matrix, which is represented by the matrix  $A$ , where  $a_{ij} = p[q_{t+1} = s_j \mid q_t = s_i]$  and  $1 \leq i, j \leq N$ ;
- (4) the probability distribution of matrix  $A$ , which is represented by the matrix  $B$ , where  $b_j(k) = p[o_k \mid q_t = s_j]$  and  $1 \leq j \leq N, 1 \leq k \leq T$ ;
- (5) the set of initial state probability distribution of HMM, which is represented by the set  $\pi$ , where  $\pi_i = p[q_1 = s_i]$  and  $1 \leq i \leq N$ .

The model of recognizing and forecasting multistep attack based on hidden Markov model is shown in Figure 1.

There are three problems which can be solved by hidden Markov model well.

- (1) *Probability Calculation Problems.* Calculate the probability  $p(O \mid \lambda)$  under a given hidden Markov model  $\lambda = (A, B, \pi)$  and the observation sequence  $O = \{o_1, o_2, \dots, o_T\}$ .
- (2) *Learning Problems.* Estimate the parameters of  $\lambda = (A, B, \pi)$  when the observation sequence  $O = \{o_1, o_2, \dots, o_T\}$  is known, to maximize the probability  $p(O \mid \lambda)$ .
- (3) *Prediction Problems.* Calculate the state sequence  $I = \{i_1, i_2, \dots, i_T\}$  under the maximum probability, when the hidden Markov model  $\lambda = (A, B, \pi)$  and observation sequence  $O = \{o_1, o_2, \dots, o_T\}$  are given.

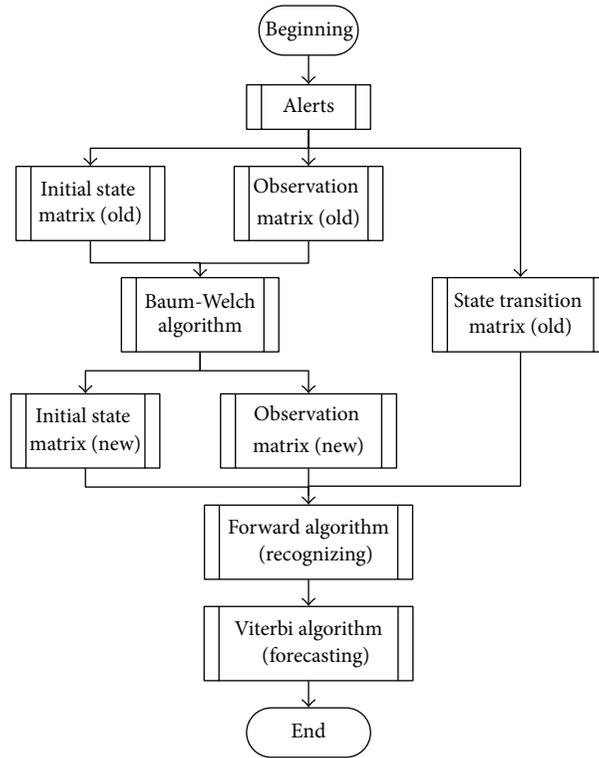


FIGURE 3: Flow chart of recognizing and forecasting multistep attack.

Input: alert sequence.

$$O = \{o_1, o_2, \dots, o_T\};$$

Output: the parameters of hidden Markov model.

$$\lambda^{(n+1)} = (A^{(n+1)}, B^{(n+1)}, \pi^{(n+1)}).$$

Step 1. Initialization.

for  $n = 0$ , select  $a_{ij}^{(0)}, b_j(k)^{(0)}, \pi_i^{(0)}$ , we can obtain the initial model  $\lambda^{(0)} = (A^{(0)}, B^{(0)}, \pi^{(0)})$ .

Step 2. Iterative calculation.

for  $n = 1, 2, \dots$ ,

$$a_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)};$$

$$b_j(k)^{(n+1)} = \frac{\sum_{t=1, o_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)};$$

$$\pi_i^{(n+1)} = \gamma_1(i).$$

$$\text{where } \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{p(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)};$$

$$\xi_t(i, j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{p(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}.$$

Step 3. Termination. We can obtain the parameters of hidden Markov model.

$$\lambda^{(n+1)} = (A^{(n+1)}, B^{(n+1)}, \pi^{(n+1)}).$$

ALGORITHM 1

Forward\_Algorithm ( $\lambda, O$ ):

Input: (1) alert sequence  $O = \{alert_1, alert_2, \dots, alert_T\}$ ;

(2) hidden Markov model (HMM)  $\lambda$ .

Output: the probability  $p(O | \lambda)$  generated by alert sequence  $O = \{alert_1, alert_2, \dots, alert_T\}$  of hidden Markov model.

Begin:

(1)  $\forall \text{int } ent_i \in \lambda, 1 \leq i \leq N$ .

//  $N$  is the number of attack intentions.

calculate the probability of  $alert_1$  generated by  $\text{int } ent_i: \alpha_1(i) = \pi_i b_i(alert_1)$

(2) calculate the probability of alert sequence  $\{alert_1, alert_2, \dots, alert_T\}$  and  $q_{t+1} = \text{int } ent_j$ .

(a) at time  $t$ , calculate the probability of alert sequence  $\{alert_1, alert_2, \dots, alert_T\}$  and  $q_t = \text{int } ent_j: \alpha_t(j)$ .

(b) at time  $t + 1$ , calculate the probability of intent sequence  $\{alert_1, alert_2, \dots, alert_T\}$  generated by hidden Markov model (HMM):  $\lambda$  and

$q_t = \text{int } ent_j: \alpha_{t+1}(j) = \left[ \sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(alert_{t+1})$  where  $1 \leq t \leq T - 1; 1 \leq j \leq N$ .

(3) calculate the probability of the intent sequence  $O = \{alert_1, alert_2, \dots, alert_T\}$  generated by hidden Markov model (HMM):  $\lambda$ .

$p(O | \lambda) = \sum_{i=1}^N \alpha_T(i)$ .

(4) Return  $p(O | \lambda)$ .

End;

ALGORITHM 2

Viterbi\_Algorithm( $\lambda, O$ ):

Input: alert sequence  $O = \{alert_1, alert_2, \dots, alert_T\}$ ;

Output: (1) intent sequence:  $Q = \{\text{int } ent_1, \text{int } ent_2, \dots, \text{int } ent_T\}$ .

(2) the completed intent sequence and the next likely intent.

Begin:

for  $i = 1$  to HMM. $m$

// HMM. $m$  is the number of hidden Markov model(s)

{

    Prob = Forward\_Algorithm(hmm. $i, O$ );

    // calculate the probability of alert sequence generated by each hidden Markov

    // model(s)

}

Most\_likely\_multi-step\_attack\_intention = maximum(Prob);

$Q = \text{Viterbi\_Algorithm}(\text{hmm.}i', O)$ ;

//  $Q$  is the completed intent sequence

//  $\text{hmm.}i'$  is the maximum(Prob) of  $\text{hmm.}i$

$Q' = S - Q$  // the next likely intent

//  $S$  is the intent sequence of  $\text{hmm.}i'$

End;

ALGORITHM 3

Correspondence between the problems and algorithms of hidden Markov model are shown in Figure 2.

Hidden Markov model is usually used to deal with the problems related to the time sequence and it has been widely used in speech recognition, signal processing, bioinformatics, and other fields. Based on the characteristics of the attack steps of hidden Markov model and the problems that hidden Markov model can be solved, we apply the hidden Markov model to the field of recognizing and forecasting multistep attack. Firstly, the improved Baum-Welch algorithm is used to train the hidden Markov model  $\lambda$ , and we get a new hidden Markov model  $\lambda'$ . Then we recognize the alert belonging to attack scenarios with the Forward algorithm of

hidden Markov model. Finally, we forecast the next possible attack sequence with the Viterbi algorithm of hidden Markov model.

### 3. The Approach to Recognizing and Forecasting Multistep Attack

The steps of the approach to recognizing and forecasting multistep attack are as follows.

*Step 1.* Obtain the initial state matrix (old), state transition matrix (old), and observation matrix (old) of HMM ( $\lambda$ ).

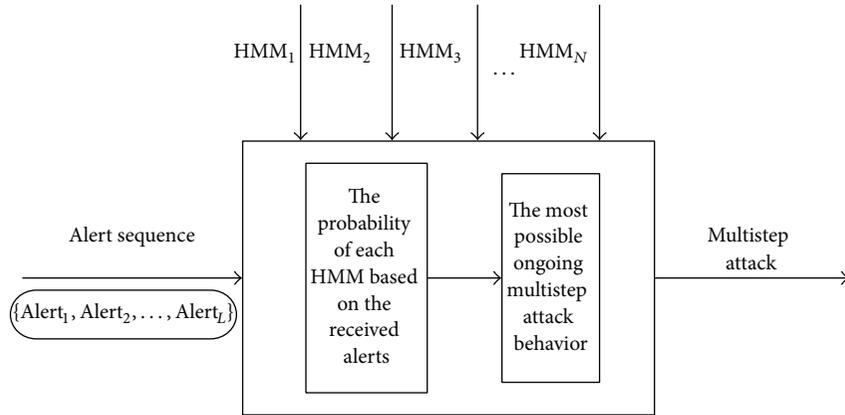


FIGURE 4: The structure of recognizing multistep attack with Forward algorithm.

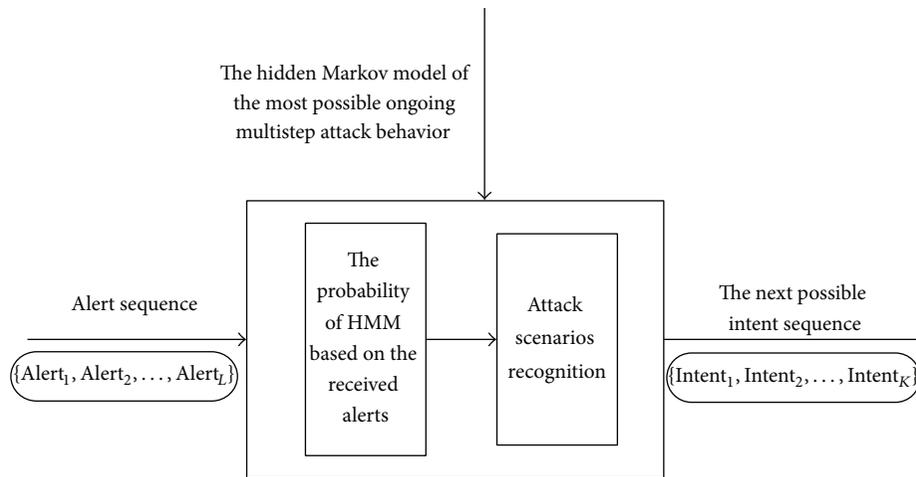


FIGURE 5: Forecasting multistep attack with Viterbi algorithm.

TABLE 1: The initial state matrix of DDoS\_HMM.

State <sub>1</sub>	State <sub>2</sub>	State <sub>3</sub>	State <sub>4</sub>	State <sub>5</sub>
0.250	0.750	0.000	0.000	0.000

TABLE 2: The state transition matrix of DDoS\_HMM.

	State <sub>1</sub>	State <sub>2</sub>	State <sub>3</sub>	State <sub>4</sub>	State <sub>5</sub>
State <sub>1</sub>	0.000	1.000	0.000	0.000	0.000
State <sub>2</sub>	0.000	0.177	0.823	0.000	0.000
State <sub>3</sub>	0.000	0.228	0.688	0.028	0.056
State <sub>4</sub>	0.000	0.000	0.000	0.750	0.250
State <sub>5</sub>	0.000	0.000	0.000	0.000	0.000

Step 2. Use the improved Baum-Welch algorithm to train the initial state matrix (old) and observation matrix (old), and we get an initial state matrix (new), observation matrix (new), and a new HMM ( $\lambda'$ ).

Step 3. Recognize the alert belonging to attack scenarios with the Forward algorithm.

Step 4. Forecast the next possible attack sequence with the Viterbi algorithm.

The flow chart is shown in Figure 3.

3.1. *The Introduction of Baum-Welch Algorithm.* If we want to apply the hidden Markov model to the multistep attack, the biggest problem is to determine the observations of HMM. A better parameter can improve the efficiency of

calculation. Meanwhile, if the selection of observation is improper, this may result in a longer training time and even not complete the training. In this regard, we apply the Baum-Welch algorithm to train the given hidden Markov model. From the result of literature (accurate Baum-Welch algorithm free from overflow), we can learn that the most reliable algorithm to train the HMM is Baum-Welch algorithm. Baum-Welch algorithm can train the given hidden Markov model ( $\lambda$ ) by an observation sequence and generate a new hidden Markov model ( $\lambda'$ ) for detection.

The steps of Baum-Welch algorithm are as in Algorithm 1.

TABLE 3: The observation matrix of DDoS\_HMM.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
S1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S2	0.000	0.490	0.490	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S3	0.000	0.000	0.000	0.000	0.200	0.200	0.200	0.200	0.200	0.000	0.000	0.000	0.000
S4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
S5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.660	0.170	0.170

TABLE 4: The initial state matrix of DDoS\_HMM'.

State <sub>1</sub>	State <sub>2</sub>	State <sub>3</sub>	State <sub>4</sub>	State <sub>5</sub>
0.599	0.401	0.000	0.000	0.000

TABLE 5: The observation matrix of DDoS\_HMM'.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
S1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S2	0.000	0.499	0.499	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S3	0.000	0.000	0.000	0.000	0.387	0.000	0.387	0.000	0.226	0.000	0.000	0.000	0.000
S4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
S5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.998	0.001	0.001

TABLE 6: DDoS\_HMM.

STATE	ALERT
State <sub>1</sub>	{Alert <sub>1</sub> }
State <sub>2</sub>	{Alert <sub>2</sub> , Alert <sub>3</sub> , Alert <sub>4</sub> }
State <sub>3</sub>	{Alert <sub>5</sub> , Alert <sub>6</sub> , Alert <sub>7</sub> , Alert <sub>8</sub> , Alert <sub>9</sub> }
State <sub>4</sub>	{Alert <sub>10</sub> }
State <sub>5</sub>	{Alert <sub>11</sub> , Alert <sub>12</sub> , Alert <sub>13</sub> }

TABLE 7: FTP Bounce\_HMM.

State	Alert
State <sub>1</sub>	{Alert <sub>1</sub> ' , Alert <sub>2</sub> '}
State <sub>2</sub>	{Alert <sub>3</sub> ' , Alert <sub>4</sub> '}
State <sub>3</sub>	{Alert <sub>5</sub> ' , Alert <sub>6</sub> ' , Alert <sub>7</sub> '}
State <sub>4</sub>	{Alert <sub>8</sub> '}
State <sub>5</sub>	{Alert <sub>9</sub> ' , Alert <sub>10</sub> '}

3.2. *Forward Algorithm.* The pseudocode of Forward algorithm is as in Algorithm 2.

Recognizing multistep attack is mainly based on the alert sequence. First, we calculate the probability of alert sequence generated by the given HMM(s). Then we decide that the attack which has the maximum is likely to be the ongoing attack. The structure of recognizing multistep attack with Forward algorithm is shown in Figure 4.

3.3. *Viterbi Algorithm.* The pseudocode of Viterbi algorithm is as in Algorithm 3.

Predicting the behavior of multistep attack is mainly to determine the intentions that the attackers have been completed and forecast the next possible attack intentions. The structure of forecasting multistep attack with Viterbi algorithm is shown in Figure 5.

#### 4. The Simulation Experiment and Analysis

4.1. *Baum-Welch Algorithm: Train the Given HMM(s).* Based on the literature (approach to forecast multistep attack based on fuzzy hidden Markov model), we can obtain the initial state matrix, state transition matrix, and observation of DDoS\_HMM, as is shown from Tables 1, 2, and 3.

The data set which is used in the simulation experiment is an attack scenario testing data set LLDOS1.0 (inside) provided by DARPA (Defense Advanced Research Projects Agency) in 2000. We extract two kinds of multistep attack from it; they are DDoS multistep attack and FTP Bounce multistep attack. While the calculation of the state transition matrix is completely the statistical calculations on data, we only train the initial state matrix and observation matrix of HMM. We can see that there are a large number of zeros in observation matrix clearly and the observation matrix is the sparse matrix. So we train the matrix(s) by block. We suppose that the number of observation sequences is  $S$  and the length of  $S$  is 32, where  $S$  multiplied by 32 equals the number of training data. And there is no corresponding sequence of state. In this regard, we can obtain the initial state matrix (new) and the observation matrix (new) of the DDoS\_HMM' ( $\lambda'$ ), as is shown in Tables 4 and 5.

4.2. *Forward Algorithm: Recognize the Alert Belonging to Attack Scenarios.* The attack intentions and alerts of DDoS\_HMM and FTP Bounce\_HMM are shown in Tables 6 and 7, respectively.

When the alerts “Alert<sub>1</sub>” and “Alert<sub>3</sub>” were received, according to the Forward algorithm of hidden Markov model,

TABLE 8: The comparison of results.

	$p(\text{alerts} \mid \text{DDoS\_HMM})$	$p(\text{alerts} \mid \text{FTP Bounce\_HMM})$	$\frac{p(\text{alerts} \mid \text{DDoS\_HMM})}{p(\text{alerts} \mid \text{FTP Bounce\_HMM})}$
Before training	0.1225	0.0079	15.5
After training	0.2989	0.0036	83.0

we will obtain the probability based on DDoS\_HMM' and FTP Bounce\_HMM', respectively:

$$p(\text{alerts} \mid \text{DDoS\_HMM}) = 0.2989,$$

$$p(\text{alerts} \mid \text{FTP Bounce}) = 0.0036.$$

We can see from the above results,  $p(\text{alerts} \mid \text{DDoS\_HMM}) > p(\text{alerts} \mid \text{FTP Bounce})$ . That is to say, the ongoing multistep attack behavior is likely to be DDoS\_HMM.

**4.3. Viterbi Algorithm: Forecast the Next Possible Attack Sequence.** When the alert sequence  $\{\text{Alert}_1, \text{Alert}_3, \text{Alert}_7, \text{Alert}_8, \text{Alert}_{10}\}$  was received by the console, we can obtain the completed intent sequence  $\{\text{State}_1, \text{State}_2, \text{State}_3, \text{State}_4\}$ . That is to say, now completed intentions are the previous four attack intentions; the next intention will be  $\text{state}_5$ .

**4.4. Comparison of Results.** We compare the results between the untrained HMM(s) and the trained HMM(s) by Baum-Welch algorithm; the comparison of results are shown in Table 8.

## 5. Conclusion

The biggest difficulty of hidden Markov model applied in multistep attack is the determination of observations. Now the research of the determination of observations is still lacking, and it shows a certain degree of subjectivity. In this regard, we train the existing hidden Markov model(s) by the Baum-Welch algorithm of HMM based on several groups of observation sequence. And we can obtain a new hidden Markov model which is more objectively. Simulation experiments results show that the hidden Markov models which have been trained are better than the untrained in recognition and prediction.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The authors would like to thank the reviewers for their detailed reviews and constructive comments, which have helped in improving the quality of this paper. This work was supported by the National Natural Science Foundation of China no. 60573036, Hebei Science Fund under Grant no. F2013205193, and Hebei Science Supported Planning Projects no. 12213514D.

## References

- [1] B. L. Xie, S. Y. Jiang, and Q. S. Zhang, "Application-ialer DDoS attack detection based on request keywords," *Computer Science*, vol. 40, no. 7, pp. 121–125, 2013.
- [2] C. Yuan, *Research on Multi-Step Attack Detection Method Based on GCT*, Jilin University, Jilin, China, 2010.
- [3] C. Chen and B. Q. Yan, "Network attack forecast algorithm for multi-step attack," *Computer Engineering*, vol. 5, no. 37, pp. 172–174, 2011.
- [4] G. Q. Zhai and S. Y. Zhou, "Construction and implementation of multistep attacks alert correlation model," *Journal of Computer Applications*, vol. 31, no. 5, pp. 1276–1279, 2011.
- [5] Z. L. Wang and X. P. Cheng, "An Attack predictive algorithm based on the correlation of intrusion alerts in intrusion response," *Computer Science*, vol. 32, no. 4, pp. 144–146, 2005.
- [6] H. Cao, Q. Q. Wang, Z. Y. Ma et al., "Attack Prediction model based on dynamic bayesian games," *Computer Applications*, vol. 27, no. 6, pp. 1545–1547, 2007.
- [7] H. Cao, Q. Q. Wang, Z. Y. Ma et al., "Attack prediction model based on static Bayesian game," *Application Research of Computers*, vol. 24, no. 10, pp. 122–124, 2007.
- [8] J.-W. Zhuge, X.-H. Han, Z.-Y. Ye, and W. Zou, "Network attack plan recognition algorithm based on the extended goal graph," *Chinese Journal of Computers*, vol. 29, no. 8, pp. 1356–1366, 2006.
- [9] S. H. Zhang, *Research on Network Security Early Warning Technology Based on Hidden Markov Model*, PLA Information Engineering University, Henan, China, 2007.

## Research Article

# A New Seamless Transfer Control Strategy of the Microgrid

Zhaoyun Zhang, Wei Chen, and Zhe Zhang

State Key Laboratory of Advanced Electromagnetic Engineering and Technology, Huazhong University of Science & Technology, Wuhan 430074, China

Correspondence should be addressed to Zhaoyun Zhang; [zzy\\_zhaoyun@163.com](mailto:zzy_zhaoyun@163.com)

Received 24 January 2014; Accepted 20 February 2014; Published 22 May 2014

Academic Editors: Y. Mao, X. Meng, J. Zhou, and Z. Zhou

Copyright © 2014 Zhaoyun Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A microgrid may operate under two typical modes; the seamless transfer control of the microgrid is very important. The mode conversion controller is installed in microgrid and the control logic of master power is optimized for microgrid mode conversion. In the proposed scheme, master power is very important. The master-power is under the PQ control when microgrid is under grid-connected. And it is under V/F control when the microgrid is under islanding. The microgrid mode controller is used to solve the planned conversion. Three types of conversion are simulated in this paper. The simulation results show the correctness and validity of the mode control scheme. Finally, the implementation and application of the operation and control device are described.

## 1. Introduction

A microgrid is a low-voltage distribution grid comprising various controllable loads, storage devices, and distributed generators as a controlled entity that can either be isolated from or operate interconnectedly with the main grid. Distributed generation (DG) and the microgrid (MG) system have received increasing research attention [1–6]. At the same time, many demonstration projects of microgrids have been constructed in China, such as Dongfushan Island microgrid in Zhejiang and Zhangbei microgrid in Hebei. In general, demonstration projects involve the power such as photovoltaic, wind, and storage battery [7–9].

The microgrid may operate under two typical modes: it can connect with the main grid, known as grid-connected mode (GM), and it can operate without main grid, called islanding mode (IM). Mode conversion is one of the core issues of the microgrid control. The researches have focused on the grid-connected mode inverter control [10–12], but few research has been done on the mode conversion.

The microgrid control may be implemented under the master-slave control mode, droop control mode [13, 14], and so forth. However, many microgrids have been built or under construction adopting master-slave mode, mainly because this type of microgrids can keep the voltage and frequency of the microgrid near nominal point. On the other hand, the active power of the solar and wind energy usually is not

controllable continuously, and PV systems and wind turbines often work at the maximum power point (MPP). So, control of solar and wind energy is in PQ control mode whether it is under islanding mode or grid-connected mode. Control logic is relatively simple. The reactive and active power of energy storage can be adjusted, and the energy storage system becomes the “master” power of microgrid when the microgrid is under islanding mode, and it is the frequency and voltage support of microgrid.

Frequency and voltage of the “master-slave” architecture microgrid can remain near the nominal point, control structures is clear, the control logic of the “slave” power is simple, and the “slave” power has the plug and play features. Because of these advantages, this microgrid architecture has been used in a wide range of applications. In papers [15–17], dual-mode inverter for this type of microgrid was researched, and the system is under PQ control when grid-connected mode is adopted; the system is under V/F control when islanding mode is adopted. The authors also proposed a microgrid mode conversion as a preliminary study but did not give more specific solutions.

The mode conversion of microgrid with “master-slave” architecture is discussed in this paper. The microgrid mode conversion includes the following four types:

- (1) planned conversion from grid-connected mode to islanding mode,

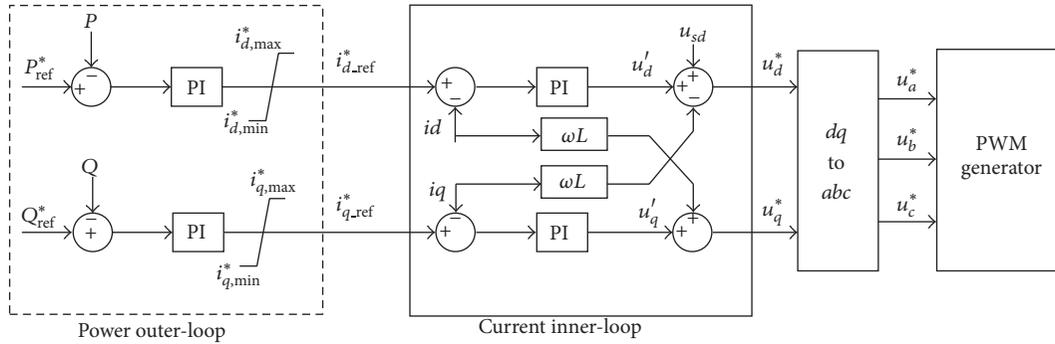


FIGURE 1: Schematic diagram of the PQ controller.

- (2) unplanned conversion from grid-connected mode to islanding mode,
- (3) planned conversion from islanding mode to grid-connected mode,
- (4) unplanned conversion from islanding mode to grid-connected mode.

The fourth conversion can be avoided, but others cannot. In this study, for the microgrid mode conversion, microgrid sets a centralized mode controller and optimizes the master power's control logic. The unplanned mode conversion is solved by the logic optimization of the master power, and the planned mode conversion is solved by the mode controller and the logic optimization of the master power.

A detailed program of the mode controller and an optimization scheme of the master power converter control system are presented. The energy storage system as an example of master power is described. The master power operates under PQ mode when microgrid works under grid-connected mode and master power operates under V/F mode when microgrid operates under islanding mode. The microgrid operating mode is detected through the microgrid information such as current, voltage, and digital input. The master power will change the operating mode, when the microgrid changes its operating mode. In the mode conversion process, a series of programs will be used to ensure microgrid stability.

## 2. Design of Control System

In a large number of the latest microgrid demonstration projects, a microgrid includes the photovoltaic power generation system, wind power systems, and energy storage system in which the "wind-solar-storage" mode is adopted. There are the photovoltaic power generation system, wind power systems, and energy storage system. Therefore, this study focuses on this type of microgrids. Without loss of generality, all of the PV systems are equivalent to one photovoltaic power generation system; all the wind systems in parallel are equivalent to one wind power system; all of the energy storage systems in parallel are equivalent to one storage system; load is distributed in microgrid.

In photovoltaic systems and direct drive wind power generation system of the microgrid, which are under PQ

control mode, the maximum power tracking is always used and reactive power output is always 0.

Storage system is under the PQ control when microgrid is under grid-connected mode. The system is under V/F control, when the microgrid is under islanding.

To achieve the smooth transition between islanding mode and grid-connected mode, the control of the microgrid mode conversion includes two parts: one is the conversion control system between PQ control and V/F control for the storage system and the other is the mode controller for the mode conversion of microgrid.

*2.1. Mode Conversion for the Storage System.* The mode conversion between the PQ control and the V/F control for the storage system is a key component for the mode conversion of the microgrid.

*2.1.1. PQ Control of the Storage System.* When the storage system is under PQ control mode, the double loop control is used, which includes power outer-loop control and current inner-loop control. Figure 1 is the control diagram.

In power outer loop, the energy management system of the microgrid provides the active power reference value  $P_{ref}^*$  and reactive power reference value  $Q_{ref}^*$ , which depend on the state of the storage system and the load balance of microgrid.

The difference between the reference value and the actual value of active power is the input of the outer-loop PI regulator, for which the output is the reference values of the  $d$ -axis current  $i_d^*$  in the inner loop. Value of the  $q$ -axis current  $i_q^*$  is similar to  $i_d^*$ .

In current inner loop, the difference between reference value and actual value of the  $d$ -axis current is the input of the PI regulator, for which the output is the reference value of the  $d$ -axis voltage ( $u_d^*$ ) of the inverter. The difference between reference value and actual value of the  $q$ -axis current is the input of the PI regulator, for which the output is the reference value of the  $q$ -axis voltage ( $u_q^*$ ) of the inverter.

*2.1.2. V/F Control of Storage System.* When storage system is under V/F control mode, the voltage reference value is set. And frequency reference value is set, too. Voltage difference between reference value and the actual value is the input of PI regulator, for which the output is the reference values of the control voltage of the inverter. Figure 2 is the control diagram.

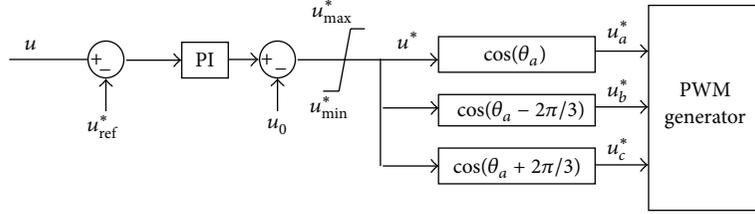


FIGURE 2: Schematic diagram of the V/F controller.

Where,  $u$  is the actual value of the voltage of the storage system output,  $u^*$  is the reference of microgrid voltage,  $f$  is the microgrid frequency, which can be set at 50 Hz (or 60 Hz),  $\theta_0$  is the initial phase angle, and  $u_0$  is the initial voltage value.

To keep frequency and voltage of the microgrid stable, angle  $\theta_0$  and voltage  $u_0$  will be adjusted.

**2.1.3. Mode Conversion of the Storage System.** Whether the system is under PQ control mode or V/F control mode, the last step is to get  $u_a^*$ ,  $u_b^*$ , and  $u_c^*$  to PWM driver circuit and control IGBT turn-off and turn-on. In order to ensure a smooth transition, the reference voltages  $u_a^*$ ,  $u_b^*$ , and  $u_c^*$  must be changed smoothly.

To ensure the continuity of the PWM reference voltage, it is necessary to ensure continuous amplitude and phase.

When the microgrid switches from grid-connected mode to islanding mode, storage system switch control includes the following.

- (a) The amplitude  $u_a^*$  and phase angle  $\theta_0$  of PWM reference voltage are recorded when system is under grid-connected mode.
- (b) When the microgrid switches from grid-connected mode to islanding mode, storage system is from the PQ control mode to the V/F control mode. The voltage reference value  $u_0$  in V/F control mode is equal to the recording amplitude, when system is under PQ control mode, and the phase angle is

$$\theta_a = \theta_0 + 2\pi f \Delta t, \quad (1)$$

where  $\theta_0$  is the recording phase angle when the storage system is under PQ mode,  $\Delta t$  is time after the storage system switches into the V/F control mode, and  $f$  is the voltage frequency value of the microgrid.

When microgrid operates continuously under V/F mode, the recursive algorithm is used in the phase angle calculation; the formula is

$$\theta_t = \theta_{t-\tau} + 2\pi f \tau, \quad (2)$$

where  $\theta_t$  is the phase angle of the current sampling point,  $\theta_{t-\tau}$  is the phase angle of the latest sampling point,  $\tau$  is the sampling interval, and  $f$  is frequency value of microgrid.

When microgrid switches from islanding mode to grid-connected mode, the synchronization function is fulfilled by the mode controller of microgrid. The storage system only

needs to change the response time of the PQ control mode (increasing  $K_i$  and reducing  $K_p$ ) and the system can complete the smooth transition from the V/F control mode to the PQ control mode.

**2.1.4. Inverter Parameters.** The storage system inverter control includes the following parameters:

- (a) the control mode: the PQ mode or V/F mode, which can through an external input, active detection, or the set of the inverter,
- (b) the active power reference value: valid under PQ mode,
- (c) the reactive power reference value: valid under the PQ mode,
- (d) the voltage reference value: valid under V/F mode,
- (e) the frequency reference value: valid under V/F mode,
- (f) the phase angle reference values: valid under V/F mode; this generally is not for the external input, instead of the internal automatic continuous calculation.

**2.2. The Mode Controller of Microgrid.** With mode conversion of storage system, microgrid can smoothly switch between grid-connected mode and islanding mode. The microgrid mode controller is used to solve two problems and accomplish two goals as follows:

- (a) planned conversion from grid-connected mode to islanding mode,
- (b) planned conversion from islanding mode to grid-connected mode.

**2.2.1. Planned Conversion from Grid-Connected Mode to Islanding Mode.** Planned conversion from grid-connected mode to islanding mode does not trip the PCC's breaker immediately. A series of control methods are adopted via the mode controller of the microgrid, and the appropriate time for the trip breaker is selected, and then the system switches from the grid-connected mode to the islanding mode.

Control logic is as follows.

(a) After splitting command is produced, the storage system continues to work under PQ control mode. The mode controller begins to change the reference power of the storage

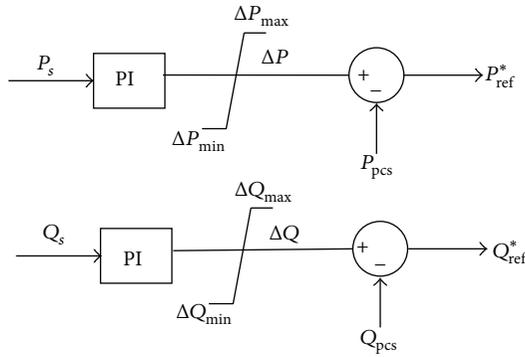


FIGURE 3: Schematic diagram of the power adjustment.

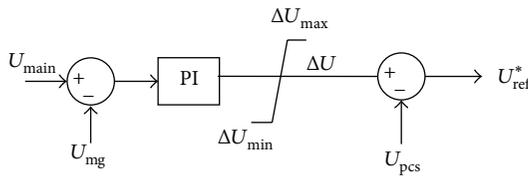


FIGURE 4: The voltage amplitude adjustment diagram.

system until the apparent power reaches 0. The control logic is as shown in Figure 3.

In Figure 3,  $P_s$  is the output active power from microgrid to main grid,  $Q_s$  is the output reactive power from microgrid to main grid,  $P_{pcs}$  is the actual output active power of PCS (storage system),  $Q_{pcs}$  is the actual output reactive power of PCS,  $P_{ref}^*$  is reference value of output active power of PCS, and  $Q_{ref}^*$  is reference value of output reactive power of PCS.

(b) If the active power and reactive power of microgrid are lower than the given values, the microgrid mode controller trips PCC breaker.

The microgrid mode controller monitors apparent power amplitude  $S_s = P_s + jQ_s$  in real time. When the amplitude is continuously lower than the threshold value for a period of time (e.g., 0.1 s), the mode controller of the microgrid sends the command to trip the PCC breaker. The storage system continues to work under PQ control mode.

(c) After the trip command is received, PCC breaker turns off, and the microgrid works under the islanding mode. The storage system detects the islanding mode and starts to work under V/F control mode. Planned conversion is completed.

**2.2.2. Planned Conversion from Islanding Mode to Grid-Connected Mode.** With planned conversion from islanding mode to grid-connected mode, the synchronization problem between microgrid and main grid is resolved. Frequency and voltage adjustment is adjusted to allow microgrid to eventually meet the synchrony conditions.

The control logic is as follows:

(a) voltage adjustment: through adjusting the voltage reference value of storage system, the microgrid voltage becomes closer to main grid's voltage; the control logic is shown in Figure 4;

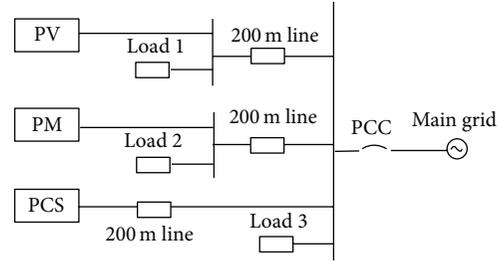


FIGURE 5: Schematic diagram of the microgrid.

(b) frequency adjustment: the mode controller detects the main grid's frequency and phase difference between the microgrid and main grid; then microgrid frequency is set as:

$$f_{ref} = f_{main} \pm \Delta f, \quad (3)$$

where  $f_{ref}$  is the storage system's reference frequency,  $f_{main}$  is the main grid's frequency, and  $\Delta f$  is the minor adjustment frequency, which can be set to 0.05 Hz;

the selection of addition or subtraction depends on the initial phase angle difference, which can be set as subtraction;

(c) synchronization check: mode controller makes checking synchronization feature to take effect; when the voltage difference and phase angle difference meet synchronization requests, the mode controller sends a close PCC breaker command;

(d) after the close command is received, the PCC breaker turns on, and microgrid operate under grid-connected mode; storage system detects this and works in PQ control mode; planned conversion is completed.

### 3. Simulation Analysis

**3.1. Simulation Parameters.** The simulation system is shown in Figure 5. Microgrid has three distributed sources, photovoltaic "PV," direct-drive wind power "PM," and battery energy storage system "PCS," and three groups of loads which are voltage-sensitive.

The photovoltaic generation "PV" works in PQ control mode, where the rated power factor is 1.0, the maximum active power track generation is adopted, and the rated active power is 150 kW. Direct-drive wind power "PM" works in PQ control, where the rated active power is 200 kW, the rated power factor is 1.0, and the maximum active power track is adopted.

In this example, all lines are 380 V line,  $R = 0.642 \Omega/\text{km}$ ,  $X = 0.102 \Omega/\text{km}$ . Load is constant impedance load, expressed as  $Z_{ld} = R_{ld} + jX_{ld}$ . Load parameters are as follows:

$$\begin{aligned} Z_{ld1} &= 0.922 + 0.218j \Omega, \\ Z_{ld2} &= 1.296 + 0.466j \Omega, \\ Z_{ld3} &= 1.294 + 0.466j \Omega. \end{aligned} \quad (4)$$

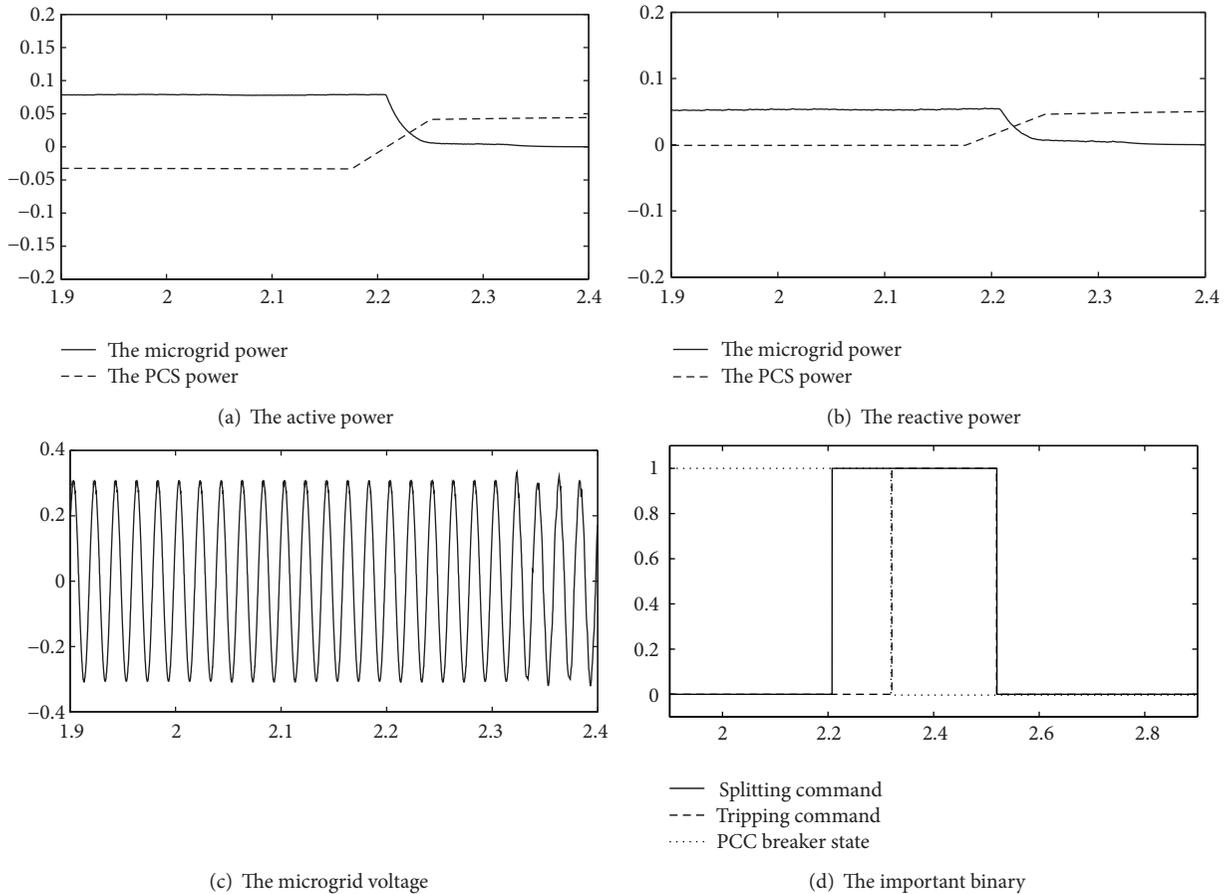


FIGURE 6: Operation results (1).

### 3.2. Simulation Results

#### 3.2.1. Planned Conversion from Grid-Connected Mode to Islanding. Figure 6 shows the operation results.

The microgrid works under the grid-connected mode before 2.2 seconds, and the switching reactive power is 50 kVA; the switching active power is 80 kW.

At 2.2 seconds, the mode controller received the splitting commands. The mode controller began to change the reference power of the storage system until the apparent power reached 0, as shown in Figures 6(a) and 6(b).

After adjustment for 150 milliseconds, the amplitude of apparent power is continuously lower than the threshold value for a period of time, and it meets the tripping conditions; the mode controller of the microgrid sends the command to trip the PCC breaker, as shown in Figure 6(d).

Before and after splitting, the voltage of the microgrid remained unchanged, as shown in Figure 6(c).

Through the simulation results, the mode controller of the microgrid is useful in this conversion and maintains stability of power system voltage before and after splitting.

#### 3.2.2. Unplanned Conversion from Grid-Connected Mode to Islanding. Operation results are shown in Figure 7.

The microgrid works under grid-connected mode before 4.5 seconds, and the switching reactive power is 110 kVA; the switching active power is 170 kW.

At 4.5 seconds, the microgrid disconnected from the main grid.

After 200 milliseconds, storage system detected islanding mode, operated under V/F control mode, and adjusted the output power.

At moment of splitting, the voltage will have some changes. In this case, the voltage dropped a little. After the storage system works from PQ control mode to VF control model, the voltage of the microgrid gets gradually regeneration and ultimately achieves stable operation, as shown in Figure 7(c).

Through the simulation results, the storage system control logic is useful in this conversion and maintains stability of power system voltage before and after splitting.

#### 3.2.3. Planned Conversion from Islanding Mode to Grid-Connected Mode. Operation results are shown in Figure 8.

The microgrid operates in islanding mode before 5.0 seconds.

The microgrid sends the reclosing command at 5.0 seconds. The voltage of microgrid and the main grid is shown

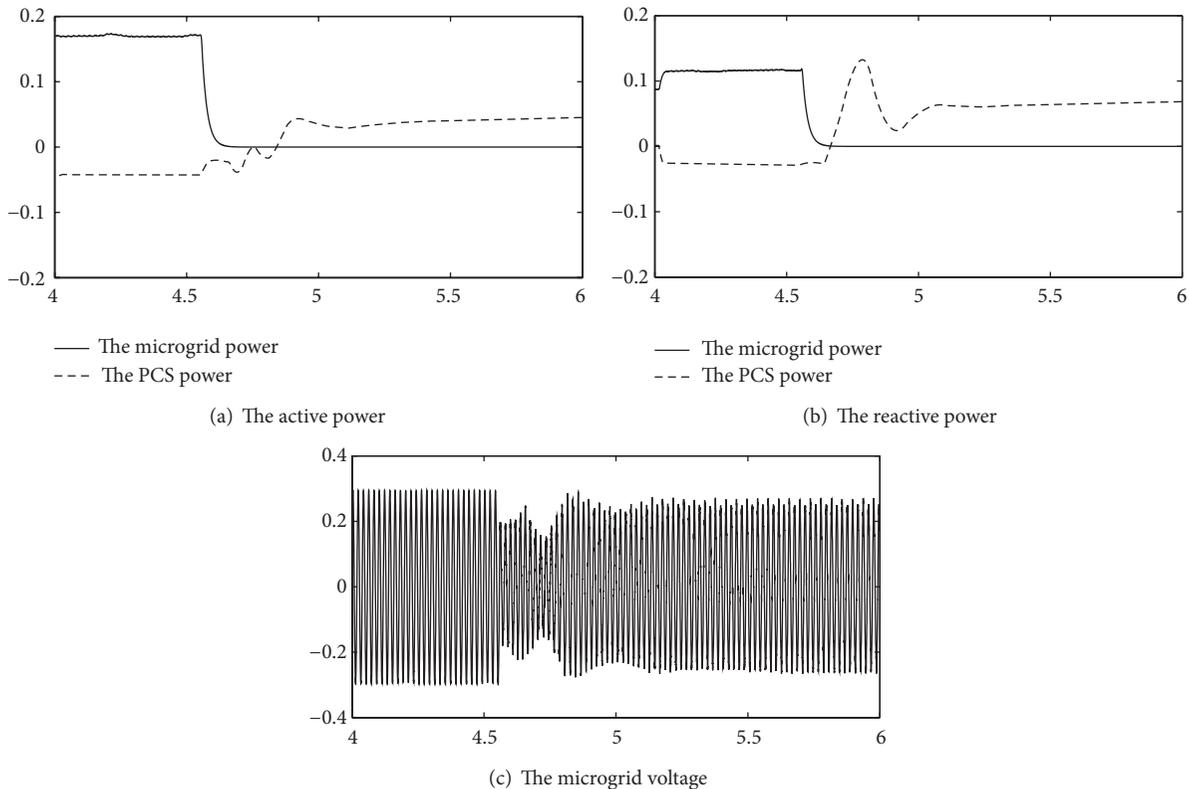


FIGURE 7: Operation results (2).

in Figure 8(a). In Figure 8(a), the phase difference between microgrid and main grid is about  $40^\circ$ .

After adjustment about 0.6 seconds, synchronization condition is met at 5.7 seconds, as shown in Figure 8(b). After the mode controller sends a close command, complete the conversion, as shown in Figure 8(e). Figures 8(c) and 8(d) are phase difference and amplitude difference in the process.

The simulation results show that the mode controller of microgrid is valid for synchronous reclosing.

Asynchronous reclosing is not detailed in this paper.

#### 4. The Implementation of Device

The operation and control device is the important device of the microgrid, and the mode controller function is one of the important functions in the operation and control device. The characteristic of the operation and control device can influence the stability of the microgrid.

*4.1. The Architecture of the Operation and Control Device.* The architecture of the operation and control device is as shown in Figure 9.

The operation and control device included three layers as follows:

- (1) hardware layer which includes CPU (Inter D525) and the necessary interface, such as RS-485, Ethernet, GPS, binary input, and binary output,

- (2) operating system layer which is the operation and control device used LINUX operating system,

- (3) software layer which includes operation and control function, mode controller function, power control function, energy management function, communication function, and so forth.

The core control strategies of the operation and control device including:

- (1) microgrid black start,
- (2) microgrid power control,
- (3) microgrid optimization control,
- (4) microgrid mode controller,
- (5) microgrid energy management.

In order to adapt to different grid structures, different micropower types, multiterminals, and other different applications, the operation and control device needed be programmed by xml file, and then it is consistent with the energy storage device and the secondary control equipment to achieve and mode conversion smoothly.

*4.2. The Communication of the Operation and Control Device.*

With the purpose of implementing control function, the device needs to communicate with the intelligent terminal in microgrid. The IEC61850 protocol, such as 9-2, GOOSE, and MMS, can be support by purposed device. The device

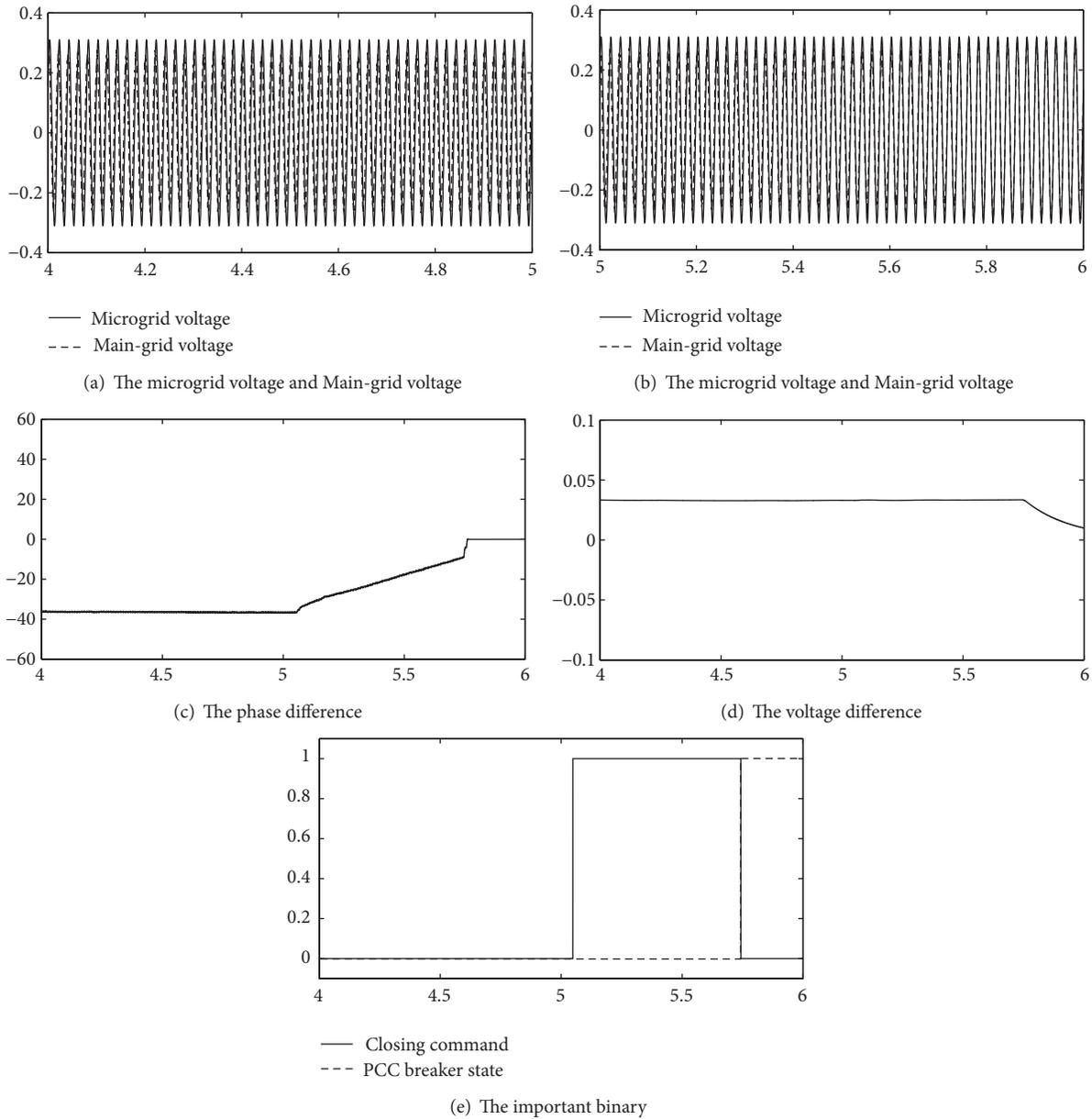


FIGURE 8: Operation results (3).

adopts the 61850 communication architecture, and uses the SMV protocol and GOOSE protocol. The communication of microgrid can adopt the networking mode or the point-to-point mode between intelligent terminal and the operation and control device. The communication of the microgrid involves three kinds of network services, which are SMV, GOOSE, and time synchronization network, each of which can use a separate network, or share a common network. In practical applications, the IEEE 1588 Synchronous Ethernet technologies are adopted to build a shared network for SMV, GOOSE, and time synchronization, and optimized data flow distribution is achieved by using VLAN. A typical communication architecture of the microgrid based on IEC61850 is shown in Figure 10.

As shown in Figure 10, the operation and control system of microgrid mainly consists of the following:

- (1) the operation and control device: each microgrid sets up one operation and control device,
- (2) intelligent terminal: the intelligent terminal is mainly responsible for collecting AC quantities and acquiring information from local circuit breakers and executing the control commands issued from the central unit,
- (3) synchronization clock source: the synchronization clock source mainly provides clock synchronization of the intelligent terminal.

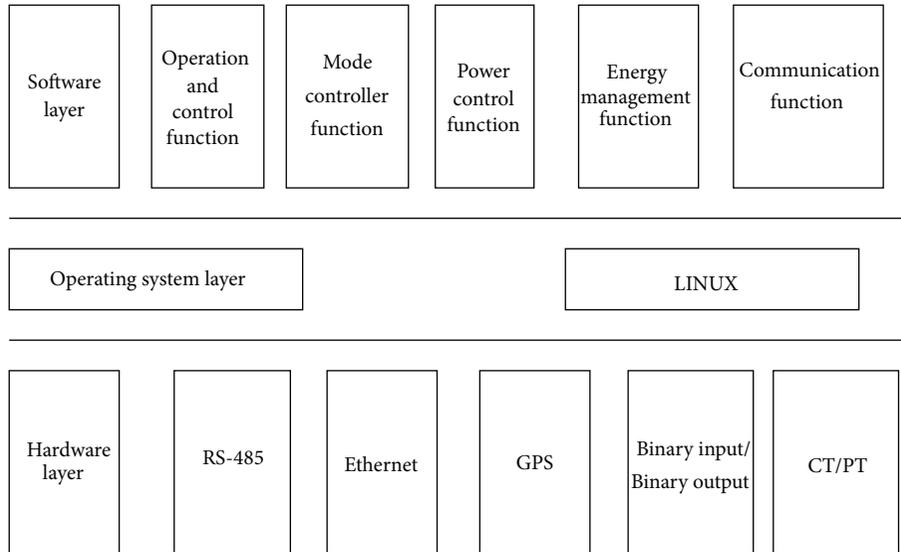


FIGURE 9: The operation and control device of microgrid.

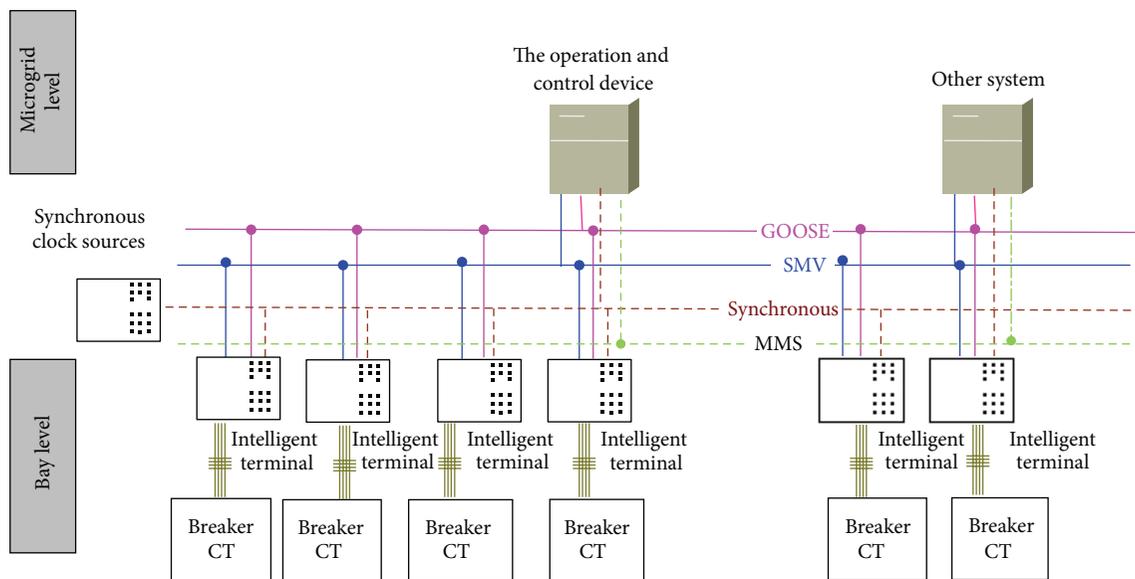


FIGURE 10: Communication architecture of the microgrid based on IEC61850.

### 5. Conclusions

Microgrid operates under two typical modes. Microgrid mode conversion has been an important part of microgrid research, and the seamless transfer of the microgrid is the control goal. The “master-slave” architecture microgrid which is widely used in engineering is selected as the research focus. The mode conversion function is fulfilled by the mode controller of microgrid and the inverter of the energy storage system. The simulation results show that the mode controller and energy storage system inverter operation mode conversion logic is valid. For further research, characteristic analysis of the equipment in the microgrid can make the microgrid more standardized and more reasonable.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgment

This paper was supported by the National Natural Science Foundation of China (no. 51177058 and no. 51277085).

### References

[1] E. Sortomme, S. S. Venkata, and J. Mitra, “Microgrid protection using communication-assisted digital relays,” *IEEE Transactions on Power Delivery*, vol. 25, no. 4, pp. 2789–2796, 2010.

- [2] R. H. Lasseter, J. H. Eto, B. Schenkman et al., "CERTS microgrid laboratory test bed," *IEEE Transactions on Power Delivery*, vol. 26, no. 1, pp. 325–332, 2011.
- [3] C. Ramonas and V. Adomavicius, "Research of the reactive power control possibilities in the grid-tied PV power plant," *Electronics and Electrical Engineering*, vol. 19, no. 1, pp. 31–34, 2013.
- [4] V. Pilkauskas, R. Plestys, G. Vilutis, and D. Sandonavicius, "Improvement of WMS functionality, aiming to minimize processing time of jobs in grid computing," *Electronics and Electrical Engineering*, vol. 113, no. 7, pp. 111–116, 2011.
- [5] G. N. Tsouehnikas, N. L. Soultanis, A. I. Tsouehnikas et al., "Dynamic modeling of microgrids," in *Proceedings of the International Conference on Future Power Systems*, pp. 1–7, Amsterdam, The Netherlands, 2006.
- [6] H. R. Lasseter, "Microgrids," in *Proceedings of the Power Engineering Society Winter Meeting*, pp. 305–308, New York, NY, USA, 2002.
- [7] K. Sheng, L. Kong, Z.-P. Qi, W. Pei, H. Wu, and P. Xi, "A survey on research of microgrid—a new power system," *Relay*, vol. 35, no. 12, pp. 75–81, 2007.
- [8] C.-S. Wang, F. Gao, P. Li, and F. Ding, "Review on the EU research projects of integration of renewable energy sources and distributed generation," *Southern Power System Technology*, vol. 2, no. 6, pp. 1–6, 2008.
- [9] C. Li, *Study on micro grid modeling and its connecting operation mode [M.S. dissertation]*, Tianyuan University of Technology, 2010.
- [10] Z. Yang, C. Wang, and Y. Che, "A small-scale microgrid system with flexible modes of operation," *Automation of Electric Power Systems*, vol. 33, no. 14, pp. 89–92, 2009.
- [11] J. Zeng, *Construction and control of energy storage systems used in renewable energy and micro grid [Ph.D. dissertation]*, Huazhong University of Science and Technology, 2009.
- [12] Z. Xiao, *Control and operation characteristic analysis of a micro grid [Ph.D. dissertation]*, Tian University, 2008.
- [13] H.-Z. Yang and X.-M. Jin, "Research on grid-connected photovoltaic inverter of maximum power point tracking," *Journal of Northern Jiaotong University*, vol. 28, no. 2, pp. 65–68, 2004.
- [14] C. Zhang, M.-Y. Chen, and Z.-C. Wang, "Study on control scheme for smooth transition of micro-grid operation modes," *Power System Protection and Control*, vol. 39, no. 20, pp. 1–10, 2011.
- [15] Z. Wang, L. Xiao, Z.-L. Yao, and Y.-G. Yan, "Design and implementation of a high performance utility-interactive inverter," *Proceedings of the Chinese Society of Electrical Engineering*, vol. 27, no. 1, pp. 54–59, 2007.
- [16] Z.-L. Yang, C.-S. Wu, and H. Wang, "Design of three-phase inverter system with double mode of grid-connection and stand-alone," *Power Electronics*, vol. 44, no. 1, pp. 14–16, 2010.
- [17] J. Jiang, S. Duan, and Z. Chen, "Research on control strategy for three-phase double mode inverter," *Transactions of China Electrotechnical Society*, vol. 27, no. 2, pp. 52–58, 2012.

## Research Article

# Reputation-Based Secure Sensor Localization in Wireless Sensor Networks

Jingsha He,<sup>1</sup> Jing Xu,<sup>2</sup> Xingye Zhu,<sup>1</sup> Yuqiang Zhang,<sup>2</sup> Ting Zhang,<sup>2</sup> and Wanqing Fu<sup>3</sup>

<sup>1</sup> School of Software Engineering, Beijing University of Technology, Beijing 100124, China

<sup>2</sup> College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China

<sup>3</sup> Information Center, SINOPEC Research Institute of Petroleum Processing, Beijing 100083, China

Correspondence should be addressed to Jingsha He; [jhe@bjut.edu.cn](mailto:jhe@bjut.edu.cn)

Received 7 March 2014; Accepted 25 April 2014; Published 20 May 2014

Academic Editor: Yuxin Mao

Copyright © 2014 Jingsha He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Location information of sensor nodes in wireless sensor networks (WSNs) is very important, for it makes information that is collected and reported by the sensor nodes spatially meaningful for applications. Since most current sensor localization schemes rely on location information that is provided by beacon nodes for the regular sensor nodes to locate themselves, the accuracy of localization depends on the accuracy of location information from the beacon nodes. Therefore, the security and reliability of the beacon nodes become critical in the localization of regular sensor nodes. In this paper, we propose a reputation-based security scheme for sensor localization to improve the security and the accuracy of sensor localization in hostile or untrusted environments. In our proposed scheme, the reputation of each beacon node is evaluated based on a reputation evaluation model so that regular sensor nodes can get credible location information from highly reputable beacon nodes to accomplish localization. We also perform a set of simulation experiments to demonstrate the effectiveness of the proposed reputation-based security scheme. And our simulation results show that the proposed security scheme can enhance the security and, hence, improve the accuracy of sensor localization in hostile or untrusted environments.

## 1. Introduction

The technologies of wireless sensor networks (WSNs) are becoming popular along with the rapid advancement of wireless communication technology, more remarkable performance of integrated circuits as well as decrease in cost and increase in functionality of sensor nodes. Since WSNs are a kind of intelligence networks that are able to integrate data collection, fusion, and transmission, such networks have been widely used in fields such as military defense, industrial and agricultural control, urban management, environment monitoring, health care, emergency rescue, and disaster relief. In addition, sensor networks also have a broad prospect of applications in tracking logistics management and space exploration. Depending on different application scenarios in the above areas, researchers have put forward some new technology and strategy, such as sensor deployment methods suitable for underwater detection [1] and intelligent monitoring technologies in Smart Home scenarios [2]. In short,

the applications of WSNs are being developed to achieve ubiquity that can bring more convenience for human beings in many areas.

In most applications, sensor nodes are used to collect physical data, such as temperature, humidity, water level, pressure, and wind speed, that are sent along with the location information to the data center to ensure that the collected data have spatial meaning. Furthermore, the location information of sensor nodes can also serve as the basis for some network functions, such as network configuration and real-time statistics of network coverage. Therefore, in massively deployed WSNs, location information of sensor nodes is very important for enabling many applications, which makes sensor localization one of the basic services and a core technology for WSNs.

Since sensor localization in wireless sensor networks (WSNs) is a fundamental technical issue and is critical for monitoring applications and for most location-based routing protocols and services, research in sensor localization

technology has generated a wide spread interest and various issues on different aspects have been studied, which include efficiency [3], accuracy [4], and security [5], among many hot issues in sensor localization.

Current algorithms for sensor localization fall into two categories: range-free algorithms [6] and range-based algorithms [7]. In a range-free algorithm, such as Centroid [8] or CTDV-Hop [9], a node estimates its location using information of connectivity between different nodes. In a range-based algorithm, a sensor node estimates its own location based on information about distances or angles between sensor nodes and through using techniques such as time of arrival (TOA) [10], time difference of arrival (TDOA) [11], received signal strength indicator (RSSI) [12], and angle of arrival (AOA) [13] as well as methods such as trilateration, triangulation, or maximum likelihood estimation [14]. Among the many different sensor localization algorithms, RSSI-based positioning technology is perhaps the most popular due to its low cost and easy implementation. On the other hand, sensor localization results can be greatly affected by malicious nodes in hostile or untrusted environments. This is because sensor nodes can hardly perform accurate localization if they use location information that is provided by untrusted beacon nodes. Security in sensor localization has thus received a great deal of attention along with the development of sensor localization technologies for WSNs.

In the past few years, researchers have proposed several security strategies for sensor localization from different aspects. Some of the methods implement verification measures to reduce the impact of using unreliable or false location information [15] while some others apply a series of schemes in which temporal, spatial, and consistent properties are considered to deal with distance-consistent spoofing attacks [16]. However, in these schemes, sensor nodes are divided into just two types: secure and insecure sensor nodes through the mechanisms of comparing the nodes and their behavior against normal situations. However, such an approach cannot be very objective, which could cause many false positive and false negative results.

Meanwhile, some other researchers have proposed localization methods that are able to fight against attacks launched by compromised sensor nodes, a problem that is more difficult to deal with. Liu et al. proposed robust computing algorithms to improve the reliability of localization schemes [17]. Park and Shin proposed an attack-tolerant localization protocol that would perform adaptive management of a profile for normal localization behavior [18]. However, the limitation of these schemes is that they did not consider the security of sensor localization when sensor nodes are joining and leaving the network along with the passage of time. In addition, they did not pay enough attention to secure sensor localization in dynamic wireless networks.

As an effective means of ensuring security, the notion of reputation has been introduced and some reputation-based schemes have since been proposed for sensor localization. Srinivasan et al. proposed a distributed reputation-based beacon trust system [19] and Xu et al. proposed a reputation-based revising scheme for sensor localization which would incur high computation cost [20]. However, complicated

reputation evaluation in the above schemes for sensor localization makes it necessary to further improve the efficiency of evaluation for beacon nodes. Any sensor localization method that can achieve good performance should ensure the reliability of location information before such information can be actually used for sensor localization.

In real applications, there may be other types of sensor localization methods to fit different application scenarios. Therefore, specific localization methods in real applications need to be continuously developed and improved based on orientation methods in order to adapt basic sensor localization schemes to the many different network scenarios. Consequently, in order to develop effective sensor localization methods, we should analyze and understand the main characteristics of specific networks and develop proper performance metrics that can be used to measure the performance of sensor localization schemes. In addition, we should also consider limitations of wireless sensor networks such as constrained energy supply in the sensor nodes as well as the complexity of network environments in the development of effective sensor localization methods.

In this paper, we propose a novel reputation-based secure sensor localization scheme to improve the accuracy of sensor localization for WSNs in hostile environments. In the proposed reputation model, the reputation of each beacon node is evaluated by each other to ensure that sensor nodes will get credible location information to perform sensor localization. The proposed scheme can therefore effectively reduce the impact of malicious beacon nodes on the localization of regular sensor nodes by relying on the security mechanism of beacon node evaluation. Our simulation results show that the proposed reputation-based secure sensor localization scheme can improve the accuracy of sensor localization in hostile or untrusted environments. In addition, the proposed secure sensor localization scheme possesses the desirable characteristics of expandability and flexibility since it can be used in both static and dynamic networks.

The remainder of this paper is structured as follows. In Section 2, we present a reputation model in which we first describe the network model and then propose a reputation evaluation model. In Section 3, we present our sensor localization scheme which is based on the evaluation of the reputation of beacon nodes. In Section 4, we describe the simulation that we have performed and present the simulation results. Finally, in Section 5, we conclude this paper in which we also discuss some future work.

## 2. The Reputation Model

In hostile network environments, which most current WSN deployments would assume, regular sensor nodes need to be confronted with security threats during the process of sensor localization. If a sensor node can identify the security and credibility of location information that it receives and subsequently use the information appropriately, the accuracy of sensor localization can be greatly improved or ensured in such environments. Therefore, in order to develop effective sensor localization schemes, we should understand the main

characteristics of the specific networks as well as the performance goals of the localization schemes. To achieve the above objective, we need to consider such characteristics as resource constraints in the sensor nodes and the complexity of the environment where the sensor nodes are deployed. Any sensor localization scheme must be effectively working in a specific WSN after the above-mentioned factors are considered in the design.

To achieve the above goal, we first propose a reputation scheme to be used in the sensor localization scheme we will propose later in this paper to deal with a hostile deployment environment in which malicious nodes can be dropped into the network at will and regular sensor nodes can also be easily compromised to make them behave in a malicious manner. We call the scheme that we propose the reputation-based localization scheme (RBL). The main characteristics of the reputation model and the RBL are as follows.

- (1) The proposed secure sensor localization scheme is developed based on a reputation model and on the evaluation of all the beacon nodes for deriving a reputation value for each and every beacon node.
- (2) In the reputation model, the reputation of each beacon node is evaluated and consequently used by regular sensor nodes to determine the credibility of the location information provided by the beacon node.
- (3) In the reputation model, the reputation of each beacon node is updated continuously with the passage of time if sensor localization needs to be carried out from time to time.

In the following sections, we will first describe the network model followed by the threat model and the reputation model.

**2.1. The Network Model.** The WSN under consideration is composed of beacon nodes and regular sensor nodes. Beacon nodes are capable of positioning themselves (e.g., by determining their positions through GPS) while the regular sensor nodes need to locate their own positions based on position information from other nodes, especially from the beacon nodes.

Our sensor localization method in this paper requires that a regular sensor node first estimates its relative position to some of the beacon nodes through the means of receiving signals from creditable beacon nodes and by computing the distances between them using a signal attenuation formula. Then, the sensor node estimates its position using the maximum likelihood estimation method [21] after it has collected enough position information.

**2.2. The Threat Model.** An analysis of the network model described above indicates that position information received from beacon nodes and the estimation of relative positions between a regular sensor node and the referenced beacon nodes can determine the accuracy of sensor localization. There are, however, two primary types of security threats for the network model as described below.

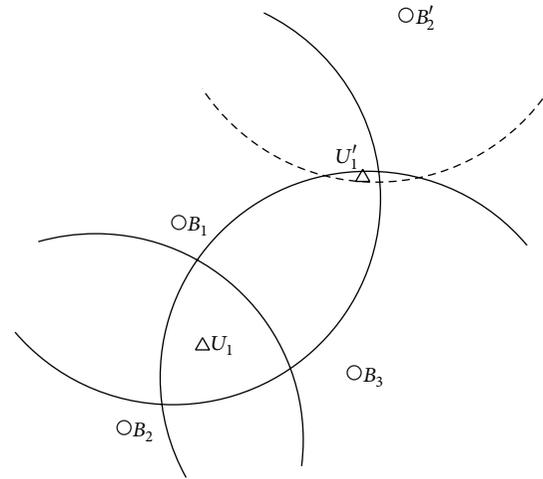


FIGURE 1: An example of sensor localization in hostile environments.

- (1) Sending false beacon information: if malicious beacon nodes send false position information, such information received by regular sensor nodes may not be accurate. Then, the estimated position of a regular sensor node cannot be guaranteed to be accurate and will lose its credibility because it is calculated based on received false information from beacon nodes. The impact to sensor localization from this type of attacks is shown in Figure 1. We can see from the figure that  $B'_2$  is the false position of beacon node  $B_2$ , which makes sensor node  $U_1$  receive a false localization result  $U'_1$ .
- (2) Obstructing physical property: if malicious nodes interfere with normal signals from beacon nodes, no regular sensor node would be able to estimate its relative position to the beacon nodes accurately by the means of signal attenuation, leading to reduced accuracy for sensor localization.

Consequently, the scheme that we propose in this paper needs to deal with the potential threats that result from the above two types of attacks in order to improve the accuracy of sensor localization in WSNs.

**2.3. The Proposed Reputation Model.** To deal with the above security threats, we propose a novel reputation model for sensor localization in WSNs. In the reputation model, beacon nodes evaluate each other using information such as the characteristics about the perception of positions and provide the evaluation results to the regular sensor nodes. The regular sensor nodes use the evaluation results provided by the beacon nodes to rank the beacon nodes and base the credibility of the location information provided by beacon nodes on such ranking.

First, let us make the following assumptions in our reputation model.

- (1) The reputation value for each and every beacon node is a number between 0 and 1, indicating values from the lowest to the highest reputations.

- (2) The reputation value for each and every beacon node is initialized to be 0.5, a medium value to start with.
- (3) In the reputation model, each beacon node performs evaluation only on its neighboring beacon nodes, that is, the beacon nodes that are one hop away from it.

Pseudocode 1 contains the pseudocode for our proposed reputation model for beacon node  $B_j$  and sensor node  $U_m$ .

The details of the evaluation procedure in the proposed reputation model are as follows.

- (1) Beacon node  $B_i$  sends its coordinate  $(x_i, y_i)$  to its neighboring beacon nodes.
- (2) Each neighboring beacon node to  $B_i$  will calculate its distance to  $B_i$  using the received coordinate information and the signal strength information independently. Let  $l_{B_{ji}}$  denote the distance between  $B_j$  and  $B_i$  based on the coordinate information and let  $d_{B_{ji}}$  denote the distance based on the signal strength information.  $B_j$  can then calculate  $l_{B_{ji}}$  using the coordinate information from  $B_i$  and calculate  $d_{B_{ji}}$  through a signal strength ranging algorithm based on the strength of the signals received from  $B_i$ .
- (3) All the neighboring beacon nodes evaluate the reputation of  $B_i$ . The value of reputation evaluation is determined using (1) in which  $R_{B_{ji}}^t$  and  $R_{B_{ji}}^{t+\Delta t}$  denote the reputation values on  $B_i$  by  $B_j$  at times  $t$  and  $t + \Delta t$ , respectively, and  $\Delta t$  denotes the time interval of two reputation values. Let  $\Delta d$  be the threshold for the distance, that is, the error that can be tolerated for the distance, and let  $\alpha$  be the weight of the evaluation value which is determined using (2)

$$R_{B_{ji}}^{t+\Delta t} = \alpha \times R_{B_{ji}}^t + (1 - \alpha), \quad \left| l_{B_{ji}} - d_{B_{ji}} \right| \leq \Delta d \quad (1)$$

$$R_{B_{ji}}^{t+\Delta t} = (1 - \alpha) \times R_{B_{ji}}^t, \quad \left| l_{B_{ji}} - d_{B_{ji}} \right| > \Delta d,$$

$$\alpha = \frac{\left| l_{B_{ji}} - d_{B_{ji}} \right|}{l_{B_{ji}} + d_{B_{ji}}}. \quad (2)$$

- (4) The neighboring sensor nodes get the reputation values for  $B_i$  from the beacon nodes. Each regular sensor node collects the evaluation values from all the neighboring beacon nodes and computes the average reputation value using (3) in which  $R_{U_m, B_i}^{t+\Delta t}$  and  $R_{B_{ki}}^{t+\Delta t}$  denote the reputation value on beacon node  $B_i$  from a sensor node  $U_m$  and that on  $B_i$  evaluated by  $B_k$  at time  $t + \Delta t$ , where  $B_k$  is a neighboring beacon node to  $B_i$  and  $n$  is the number of such neighboring nodes. Consider

$$R_{U_m, B_i}^{t+\Delta t} = \frac{\sum_{k=1}^n R_{B_{ki}}^{t+\Delta t}}{n}. \quad (3)$$

- (5) Every regular sensor node ranks the neighboring beacon nodes from high to low based on the received reputation values.

### 3. The Sensor Localization Scheme

The sensor localization scheme in this paper uses the proposed reputation evaluation scheme described above in which the reputation model is relied upon by the beacon nodes to evaluate each other. In the illustration below, we use the RSSI ranging technology for sensor localization although the same reputation scheme can be applied equally to TOA, TDOA, and AOA ranging methods in practical applications.

After receiving the evaluation results, a regular sensor node will select credible beacon nodes based on the reputation values. Afterwards, the sensor node will measure the distance to the credible beacon nodes using the RSSI ranging technology and estimate its location through maximum likelihood estimation. The main steps in our localization scheme are described as follows.

- (i) Every beacon node provides its location information to all the neighbor nodes. As shown in Figure 2, beacon node  $B_1$  sends its position coordinate  $(x_1, y_1)$  to all the neighboring nodes.
- (ii) Beacon nodes in the network will evaluate each other using the proposed reputation model and each will send its evaluation results to all the neighboring nodes. As shown in Figure 2, beacon nodes  $B_2, B_3, B_4$ , and  $B_5$  evaluate the reputation of  $B_1$  after receiving the location information from  $B_1$  using the proposed reputation model and each will send the evaluation result to their neighboring sensor nodes including node  $U_1$ .
- (iii) Each regular sensor node will select credible beacon nodes based on the results from the reputation evaluation. Sensor node  $U_1$  computes the reputation value for  $B_1$  and collects the reputation values from neighboring beacon nodes using (3). Then,  $U_1$  ranks the neighboring beacon nodes according to the reputation values in the order of high to low, based on which it selects the credible beacon nodes accordingly.
- (iv) Regular sensor nodes estimate their relative positions to the credible beacon nodes using the signal attenuation formula in RSSI [12].
- (v) Regular sensor nodes calculate their coordinates using maximum likelihood estimation. Suppose that the number of credible neighboring beacon nodes around  $U_1$  is  $p$  with coordinates  $(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)$ , respectively, and the distances between  $U_1(x_{U_1}, y_{U_1})$  and the beacon nodes are  $d_1, d_2, \dots, d_p$ , respectively; then the position of  $U_1$  can be calculated using the following:

$$(x_{U_1} - x_i)^2 + (y_{U_1} - y_i)^2 = d_i^2, \quad i = 1, 2, \dots, p. \quad (4)$$

In addition,  $p$  distance equations about  $U_1$  and the  $p$  beacon nodes are listed as in (5) that result from subtracting

```

// Beacon nodes evaluate each other between neighbor beacon nodes
RBjit = 0.5
while(true)
    α =  $\frac{|l_{B_{ji}} - d_{B_{ji}}|}{l_{B_{ji}} + d_{B_{ji}}}$ 
    if  $|l_{B_{ji}} - d_{B_{ji}}| \leq \Delta d$ 
        RBjit+Δt = α × RBjit + (1 - α)
    else
        RBjit+Δt = (1 - α) × RBjit
    sleep(Δt)
// Beacon nodes send the reputation value to their neighbor sensor nodes
Send (node Bj, node Um, reputation value)
// Sensor nodes compute the reputation value of their neighbor beacon nodes
RUm,Bit+Δt =  $\frac{\sum_{k=1}^n R_{B_{ki}}^{t+\Delta t}}{n}$ 
    
```

PSEUDOCODE 1: Pseudocode for the reputation model.

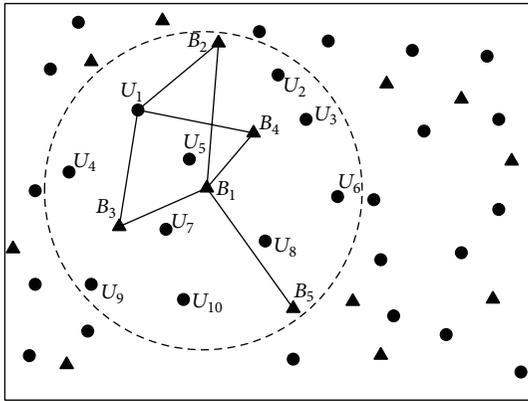


FIGURE 2: The network topology.

the last equation from each of the first  $p - 1$  equations. Consider

$$\begin{aligned}
 &x_1^2 - x_p^2 - 2(x_1 - x_p)x_{U_1} + y_1^2 - y_p^2 \\
 &\quad - 2(y_1 - y_p)y_{U_1} = d_1^2 - d_p^2 \\
 &\quad \dots
 \end{aligned} \tag{5}$$

$$\begin{aligned}
 &x_{p-1}^2 - x_p^2 - 2(x_{p-1} - x_p)x_{U_1} + y_{p-1}^2 - y_p^2 \\
 &\quad - 2(y_{p-1} - y_p)y_{U_1} = d_{p-1}^2 - d_p^2.
 \end{aligned}$$

$U_1(x_{U_1}, y_{U_1})$  can then be calculated using the following:

$$U_1 = A^{-1}b. \tag{6}$$

The matrices in (6) can then be expressed as the following expressions:

$$\begin{aligned}
 A &= 2 \begin{bmatrix} x_1 - x_p & y_1 - y_p \\ \dots & \dots \\ x_{p-1} - x_p & y_{p-1} - y_p \end{bmatrix}, \\
 b &= \begin{bmatrix} x_1^2 - x_p^2 + y_1^2 - y_p^2 - d_1^2 + d_p^2 \\ \dots \\ x_{p-1}^2 - x_p^2 + y_{p-1}^2 - y_p^2 - d_{p-1}^2 + d_p^2 \end{bmatrix}, \\
 U_1 &= \begin{bmatrix} x_{U_1} \\ y_{U_1} \end{bmatrix}.
 \end{aligned} \tag{7}$$

The final solution to (6) can be obtained using the following:

$$U = (A^T A)^{-1} A^T b. \tag{8}$$

From the above steps, we can see that a regular sensor node would treat the location information from neighboring beacon nodes differently according to the result of reputation evaluation. There is no need to determine the position relationship between regular sensor nodes and beacon nodes that have low reputation values, which are required in the signal attenuation formula, resulting in reducing a certain amount of computational overhead.

#### 4. Simulation and Analysis

We have performed some simulation on wireless sensors localization with the proposed RBL to evaluate the performance of the scheme.

The network configuration for the simulation is set up as follows. The regular sensor nodes and beacon nodes are deployed randomly in an area of 650 m × 600 m. The transmission radius of each beacon and sensor node is set

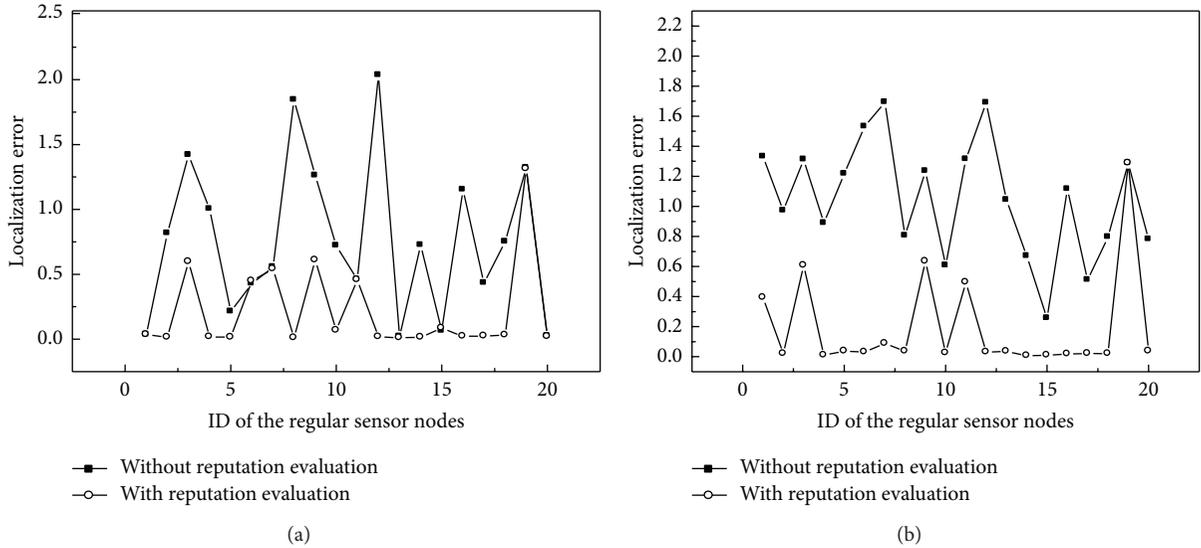


FIGURE 3: Sensor localization error with a different number of malicious beacon nodes: (a) 10 normal and 5 malicious beacon nodes; (b) 10 normal and 10 malicious beacon nodes.

at 200 m. There exist some malicious beacon nodes that randomly send out false location information.

Localization error is one important indicator of the performance in sensor localization for WSNs, which is calculated using (9). In the formula,  $(x_{U_m}, y_{U_m})$  and  $(x'_{U_m}, y'_{U_m})$  denote the measured coordinates and the actual coordinates for node  $U_m$ , respectively,  $R$  denotes the transmission radius of the nodes, and  $e_m$  is the localization error. Consider

$$e_m = \frac{\sqrt{(x_{U_m} - x'_{U_m})^2 + (y_{U_m} - y'_{U_m})^2}}{R}. \quad (9)$$

The localization error from the simulation for 20 sensor nodes is shown in Figure 3. We can see from the figure that reputation evaluation is effective for reducing localization error in hostile environments and the improvement is more significant as the number of malicious beacon nodes increases.

There are two types of threats in sensor localization: attacks targeted at the nodes and attacks targeted at the location information. RBL evaluates the credibility of beacon nodes by evaluating the location information that beacon nodes provide in order to reduce the influence of compromised beacon nodes on localization results and to resist the threat of location information tampering by the malicious beacon nodes. To measure the capability of RBL on countering the above security threats, in our evaluation, we deploy 40 regular sensor nodes to expand the scale of our experiment in which we measure the average localization error using (10) where  $N$  denotes the number of regular sensor nodes in the network. The average localization error from our simulation on a network in which there exist one or more compromised beacon nodes is shown in Figure 4. We can see from the figure that although the average localization error fluctuates with the number and the locations of the regular sensor nodes,

the result of RBL is much better than that of the primary localization scheme (PLS) using RRSI in which no evaluation of beacon nodes is performed. Consider

$$\bar{e} = \frac{\sum_{i=1}^N e_i}{N}. \quad (10)$$

Since WSNs possess the characteristics of dynamic network topology, an advanced secure sensor localization scheme should not only be able to ensure the security of sensor localization in a static network, but also be able to handle the cases of nodes joining, leaving, and removing from the network. We have performed some simulations on sensor localization for the above scenarios.

In order to expand the coverage of beacon nodes in a network so as to make more regular sensor nodes the neighbors of the beacon nodes in the network, consequently improving the utilization of beacon information, we can increase the signal transmission power of the beacon nodes to effectively expand the signal transmission radius of the beacon nodes.

In the simulations, we first deploy 20 regular nodes and 4 normal beacon nodes in the area. Then, we add more beacon nodes into the network at the rate of one node per minute starting at the moment of 1.5 min with normal and malicious beacon nodes being added alternately. Malicious beacon nodes that are added into the network would send out false position information randomly while normal beacon nodes always send out their real position information. Figure 5 shows the average localization error for regular sensor nodes during the first seven minutes from which we can see that the average localization error for regular sensor nodes fluctuates noticeably in the primary localization scheme but exhibits a good performance in our proposed RBL.

We have also performed some simulations to evaluate the impact of nodes leaving the network on sensor localization.

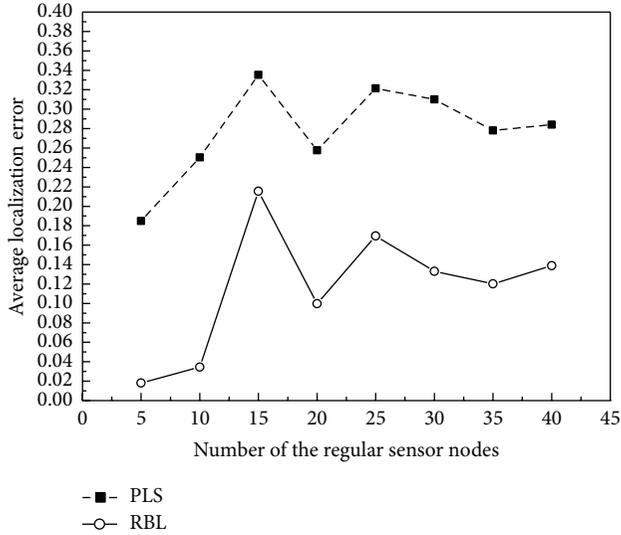


FIGURE 4: Average sensor localization error with a varying number of regular sensor nodes.

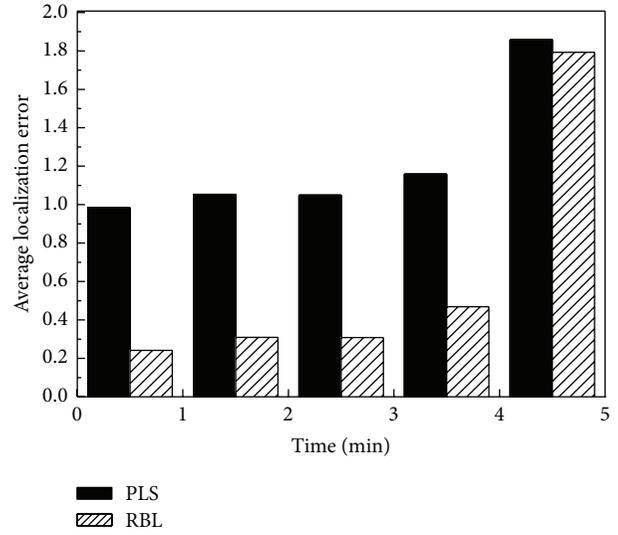


FIGURE 6: Average sensor localization error as beacon nodes are removed from the network.

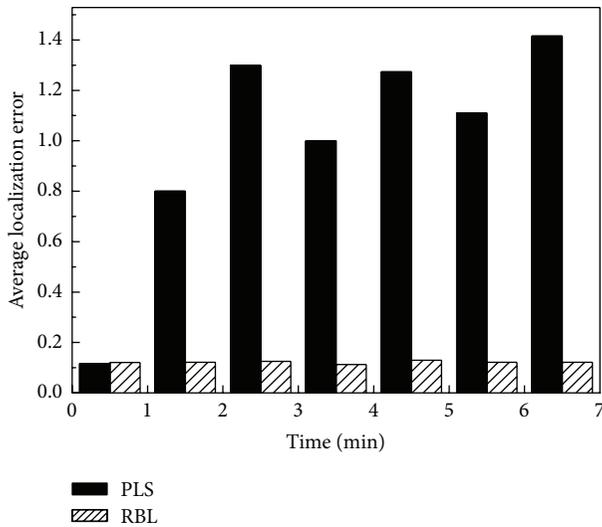


FIGURE 5: Average sensor localization error as beacon nodes are added into the network.

In the simulation, we first deploy 20 regular nodes, 4 normal beacon nodes, and 4 malicious beacon nodes in the area. Then, we remove the beacon nodes from the network at the rate of one node per 1 minute starting at the moment of 1.5 min with normal and malicious beacon nodes being removed alternately. Again, normal beacon nodes always claim their real positions while malicious beacon nodes would send out false position information randomly. Figure 6 shows the average sensor localization error for regular sensor nodes during the process from which we can see that the average localization error of RBL is noticeably lower than that of PLS in most cases. However, when the number of normal beacon nodes falls below three in the whole network,

TABLE 1: State situations for the beacon nodes.

Categories	1	2	3	4
Situations	$RP_1 \cap CP_2$	$RP_2 \cap CP_1$	$RP_2 \cap CP_2$	$RP_2 \cap CP_3$

the advantage would disappear, which seems to be a limitation of the current RBL.

Lastly, we evaluate the impact of status change among the existing beacon nodes on regular sensor nodes under the assumption that the total number of beacon nodes remains the same. Four possibilities exist for such status change as illustrated in Table 1 in which  $RP_1$  and  $RP_2$  represent the cases in which a beacon node does not change its real position and changes its real position, respectively, and  $CP_1$ ,  $CP_2$ , and  $CP_3$  represent the cases in which a beacon node does not change its claimed position information, changes its claimed position information randomly, and changes its claimed position information consistently, respectively.

We perform evaluations for four scenarios. In the first evaluation, we deploy 20 regular sensor nodes and 8 normal beacon nodes in the area and then change the status of 4 beacon nodes to the state that corresponds to situation 1 in Table 1 gradually during a 4-minute time period starting at the moment of 1.5 min. In the second evaluation, we deploy 20 regular sensor nodes and 8 normal beacon nodes in the area and then change the status of 4 beacon nodes to the state that corresponds to situation 2 in Table 1 gradually during a 4-minute time period starting at the moment of 1.5 min. In the third evaluation, we deploy 20 regular sensor nodes and 8 normal beacon nodes in the area and then change the status of 4 beacon nodes to the state that corresponds to situation 3 in Table 1 gradually during a 4-minute time period starting at the moment of 1.5 min. In the last evaluation, we deploy 20 regular sensor nodes and 8 normal beacon nodes in

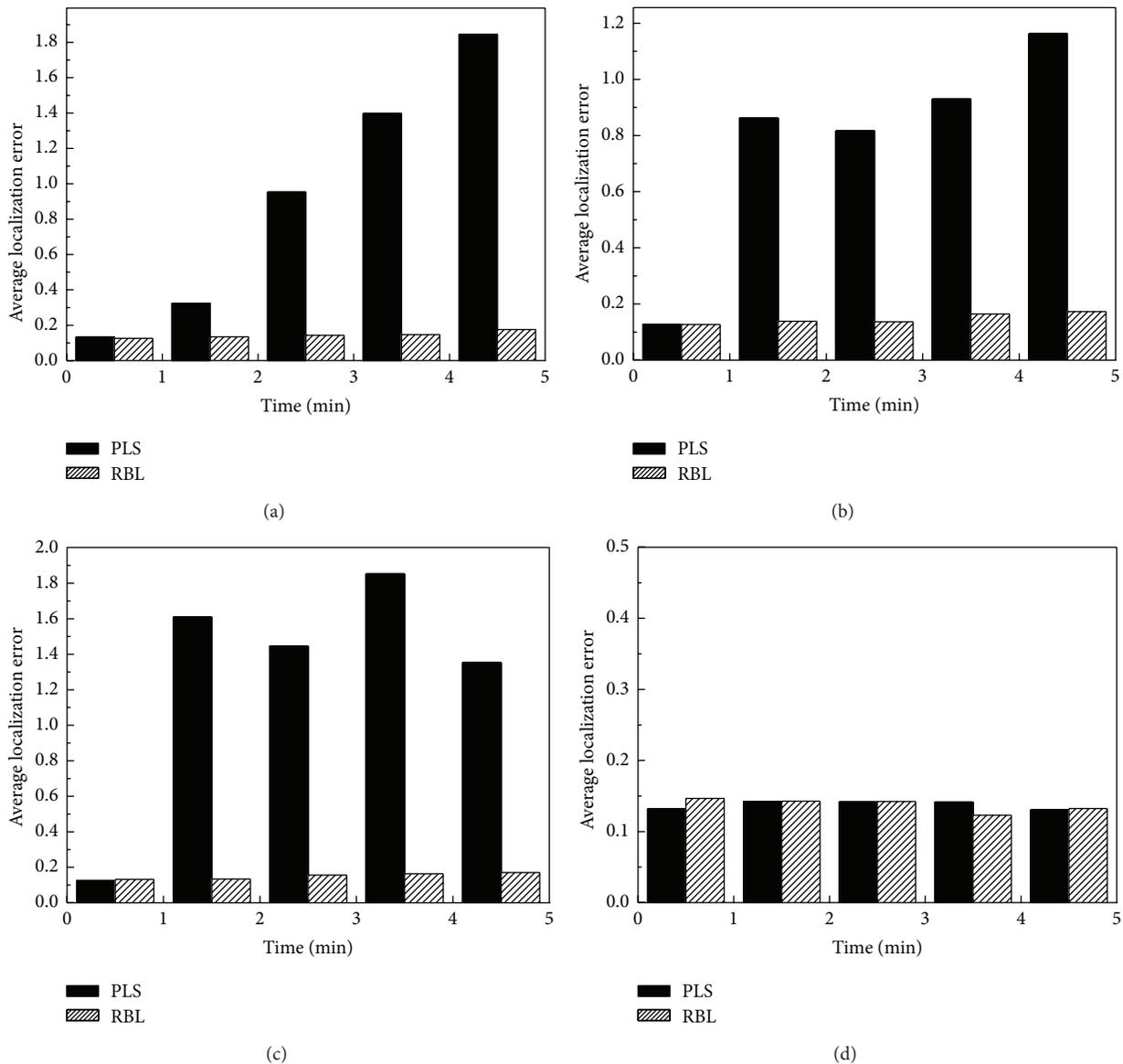


FIGURE 7: Average sensor localization error as beacon nodes change their status in various scenarios.

the area and then change the status of 4 beacon nodes to the state that corresponds to situation 4 in Table 1 gradually during a 4-minute time period starting at the moment of 1.5 min.

Figures 7(a), 7(b), 7(c), and 7(d) show the results of the evaluations that correspond to the above four evaluation scenarios. We can see from the figure that RBL can effectively filter out abnormal (or malicious) beacon nodes when some of the beacon nodes change their status in an unpredictable manner, which demonstrates that RBL is an effective scheme for secure sensor localization for WSNs, which clearly shows that RBL can improve the accuracy of sensor localization in hostile or untrusted environments.

In summary, so far, we have performed three sets of simulation experiments to verify the performance and

the effectiveness of the proposed security scheme for sensor localization in hostile or untrusted environments. In the first one, we evaluated the performance of localizing regular sensor nodes in the presence of a varying number of malicious beacon nodes. In the second one, we evaluated the average sensor localization error for different numbers of regular sensor nodes in hostile or untrusted environment. In the third one, we evaluated sensor localization results; when new beacon nodes join the network, existing beacon nodes leave the network and existing beacon nodes change their status that determines how they would make claims on their positions. It is clear that the purpose of the last experiment is to evaluate the influence on the localization of regular sensor nodes due to changes on the credibility of the beacon nodes. That is, the first two experiments are

mainly aimed at showing the performance of RBL on secure sensor localization in static WSNs while the third one is aimed at verifying the effectiveness of RBL on secure sensor localization in dynamic WSNs. All the simulation results that we have obtained clearly show that RBL can reduce the effect of malicious beacon nodes on the localization of regular sensor node, thus allowing us to conclude that RBL can effectively improve the security and the accuracy of sensor localization in WSNs. The experiments also indicate that RBL can scale well with the size of the network and can be applied in dynamic WSNs, especially when new sensor nodes can join and existing sensor nodes can leave networks with the passage of time.

## 5. Conclusion

In this paper, we proposed a novel reputation model for regular sensor nodes to evaluate the credibility of beacon nodes in sensor localization. In the model, beacon nodes first evaluate each other and then provide the evaluation results to regular sensor nodes for them to determine the credibility of beacon nodes to ensure that they will receive and use credible position information from the beacon nodes in locating their own positions. The proposed security scheme can improve the accuracy of sensor localization in hostile or untrusted environments. The scheme can help to ensure the reliability of received location information under the scenario of signal attenuation by minimizing the effects of false location information as well as interfering signals caused by malicious beacon nodes.

In the future, we will extend our security scheme to counter other types of malicious attacks in sensor localization without incurring too much additional computational cost and communication overhead and to apply our reputation-based sensor localization scheme to different network environments to further verify and improve the scheme. We will also study the impact on evaluation due to other factors of sensor nodes to further improve the performance and usability of our secure sensor localization scheme in WSNs.

## Notations

$B_i$ :	Beacon node $i$
$U_m$ :	Sensor node $m$
$\Delta t$ :	The time interval of the two reputation values
$l_{B_{ji}}$ :	The distance between beacon node $j$ and $i$ based on the coordinate information
$d_{B_{ji}}$ :	The distance between beacon node $j$ and $i$ based on the ranging techniques (such as received signal strength indicator)
$R_{B_{ki}}^t$ :	The reputation value on beacon node $i$ from a beacon node $k$
$R_{U_m, B_i}^t$ :	The reputation value on beacon node $i$ from a sensor node $m$ at time $t$
$(x_p, y_p)$ :	The coordinate of beacon node $p$
$(x_{U_m}, y_{U_m})$ :	The coordinate of sensor node $m$ .

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

The work in this paper has been supported by National Natural Science Foundation of China (Grant no. 61272500) and Beijing Natural Science Foundation (Grant no. 4142008).

## References

- [1] E. F. Golen, B. Yuan, and N. Shenoy, "An evolutionary approach to underwater sensor deployment," *International Journal of Computational Intelligence Systems*, vol. 2, no. 10, pp. 184–201, 2009.
- [2] J. C. Augusto, J. Liu, P. McCullagh, H. Wang, and J.-B. Yang, "Management of uncertainty and spatio-temporal aspects for monitoring and diagnosis in a smart home," *International Journal of Computational Intelligence Systems*, vol. 1, no. 4, pp. 361–378, 2008.
- [3] S.-K. Yang and K.-F. Ssu, "An energy efficient protocol for target localization in wireless sensor networks," *World Academy of Science, Engineering and Technology*, vol. 56, no. 8, pp. 398–407, 2009.
- [4] M. Boushaba, A. Hafid, and A. Benslimane, "High accuracy localization method using AoA in sensor networks," *Computer Networks*, vol. 53, no. 18, pp. 3076–3088, 2009.
- [5] R. Sugihara and R. K. Gupta, "Sensor localization with deterministic accuracy guarantee," in *Proceedings of the IEEE INFOCOM*, pp. 1772–1780, April 2011.
- [6] J. Park, Y. Lim, K. Lee, and Y.-H. Choi, "A polygonal method for ranging-based localization in an indoor wireless sensor network," *Wireless Personal Communications*, vol. 60, no. 3, pp. 521–532, 2011.
- [7] Y. W. E. Chan and B. H. Soong, "A new lower bound on range-free localization algorithms in wireless sensor networks," *IEEE Communications Letters*, vol. 15, no. 1, pp. 16–18, 2011.
- [8] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," *IEEE Personal Communications*, vol. 7, no. 5, pp. 28–34, 2000.
- [9] H. Wu, M. Deng, L. Xiao, W. Wei, and A. Gao, "Cosine theorem-based DV-hop localization algorithm in wireless sensor networks," *Information Technology Journal*, vol. 10, no. 2, pp. 239–245, 2011.
- [10] I. Güvenç and C.-C. Chong, "A survey on TOA based wireless localization and NLOS mitigation techniques," *IEEE Communications Surveys and Tutorials*, vol. 11, no. 3, pp. 107–124, 2009.
- [11] A. Savvides, C.-C. Han, and M. B. Strivastava, "Dynamic fine-grained localization in ad-hoc networks of sensors," in *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, pp. 166–179, July 2001.
- [12] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies*, pp. 775–784, March 2000.
- [13] D. Niculescu and B. Nath, "Ad hoc positioning system (APS) using AOA," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies*, pp. 1734–1743, April 2003.

- [14] Y. Zhang, L. Bao, S.-H. Yang, M. Welling, and D. Wu, "Localization algorithms for wireless sensor retrieval," *Computer Journal*, vol. 53, no. 10, pp. 1594–1605, 2010.
- [15] S. Čapkun, K. B. Rasmussen, M. Čagalj, and M. Srivastava, "Secure location verification with hidden and mobile base stations," *IEEE Transactions on Mobile Computing*, vol. 7, no. 4, pp. 470–483, 2008.
- [16] H. Chen, W. Lou, and Z. Wang, "A novel secure localization approach in wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, pp. 1–12, 2010.
- [17] D. Liu, P. Ning, A. Liu, C. Wang, and W. K. Du, "Attack-resistant location estimation in wireless sensor networks," *ACM Transactions on Information and System Security*, vol. 11, no. 7, pp. 22–39, 2008.
- [18] T. Park and K. G. Shin, "Attack-tolerant localization via iterative verification of locations in sensor networks," *Transactions on Embedded Computing Systems*, vol. 8, no. 12, pp. 1–24, 2008.
- [19] A. Srinivasan, J. Teitelbaum, and W. Jie, "DRBTS: distributed reputation-based beacon trust system," in *Proceedings of the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing (DASC '06)*, pp. 277–283, October 2006.
- [20] X. Xu, H. Jiang, L. Huang, H. Xu, and M. Xiao, "A reputation-based revising scheme for localization in wireless sensor networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '10)*, pp. 1–6, April 2010.
- [21] R.-I. Rusnac and A. Ş. Gontean, "Maximum Likelihood Estimation Algorithm evaluation for wireless sensor networks," in *Proceedings of the 12th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC '10)*, pp. 95–98, September 2010.

## Research Article

# Supporting Seamless Mobility for P2P Live Streaming

Eunsam Kim,<sup>1</sup> Sangjin Kim,<sup>2</sup> and Choonhwa Lee<sup>3</sup>

<sup>1</sup> Department of Computer Engineering, Hongik University, Seoul 121-791, Republic of Korea

<sup>2</sup> Hyundai Autoever Corporation, 576 Sam, Uiwang, Gyeonggi 437-040, Republic of Korea

<sup>3</sup> Division of Computer Science and Engineering, Hanyang University, Seoul 133-791, Republic of Korea

Correspondence should be addressed to Choonhwa Lee; [lee@hanyang.ac.kr](mailto:lee@hanyang.ac.kr)

Received 3 March 2014; Accepted 28 April 2014; Published 19 May 2014

Academic Editor: Jingjing Zhou

Copyright © 2014 Eunsam Kim et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With advent of various mobile devices with powerful networking and computing capabilities, the users' demand to enjoy live video streaming services such as IPTV with mobile devices has been increasing rapidly. However, it is challenging to get over the degradation of service quality due to data loss caused by the handover. Although many handover schemes were proposed at protocol layers below the application layer, they inherently suffer from data loss while the network is being disconnected during the handover. We therefore propose an efficient application-layer handover scheme to support seamless mobility for P2P live streaming. By simulation experiments, we show that the P2P live streaming system with our proposed handover scheme can improve the playback continuity significantly compared to that without our scheme.

## 1. Introduction

With the widespread deployment of high speed broadband networks such as FTTH, the IPTV services converging broadcasting and communication technologies have emerged. So far most commercial IPTV systems have employed the client/server architecture where video data are transmitted only from servers to clients. To support a huge number of IPTV subscribers at the same time, however, the client/server architecture should employ CDN (content distribution networks) structures to reduce data transmission delay. As the number of IPTV subscribers increases, the client/server architecture thus causes high expense for expanding network capacity by adding proxy servers to accommodate all the increasing subscribers [1]. The personalized IPTV services including time-shifted TV may make it even more difficult for IPTV systems to manage network traffic efficiently [2, 3].

Many research efforts have therefore been made on peer-to-peer (P2P) live streaming since it is a cost-effective and scalable alternative to client/server architectures on the Internet [4–6]. In P2P live streaming systems, peers exchange distributed video data with each other on virtual overlay networks. Thus, the better performance can be achieved as the number of participating peers increases.

On the other hand, with recent advance in wireless networks and advent of powerful mobile devices such as smart phones, the video streaming services have become feasible in mobile platforms [7–9]. Since these mobile IPTV services can provide users with the mobility and portability in wireless networks, the users' demand to enjoy IPTV services with mobile devices has been increasing rapidly. One of the most challenging issues when designing mobile IPTV systems is that users may experience the degradation of service quality due to data loss caused by the handover occurring when mobile devices are moving across APs. To get over this problem, it is thus essential to provide seamless mobility for P2P live streaming.

To minimize the transmission delay during the handover period, many schemes have been proposed at different layers of the protocol stack including data link [10, 11], network [12, 13], and transport layer [14, 15] depending on the characteristics of each layer. In the handover schemes at the layers below the application layer, however, the data loss inherently occurs while the network is being disconnected to switch APs. To avoid playback jitter in P2P live streaming systems, it is thus important to compensate for the amount of data that could not be received during the handover period. We thus need a new handover scheme at the application layer

apart from that at lower layers. In fact, several application-layer handover schemes have been proposed recently but they did not consider P2P streaming structures, all based on the client/server architecture using CDN structures with low scalability and high cost.

In this paper, we therefore propose an efficient application-layer handover scheme to provide seamless mobility in the presence of handover by considering mobile peers' limited resources and the unstable characteristics of wireless networks. In our proposed scheme, to receive data from neighbor peers at a faster speed, neighbor peers transmit data to a mobile peer through a push manner for the period around the handover. To further improve the performance, an agent peer for the mobile peer is selected among stationary peers. The agent peer receives data in place of the mobile peer before and during the handover and then transmits the data through a new AP after the handover. It can transmit data at a faster speed compared to the other neighbor peers because it is selected depending on its RTT value from a new AP and the appropriateness of its buffered period from a mobile peer's perspective.

Through extensive simulations, we demonstrate the effectiveness of our proposed handover scheme. The simulation results show that the mobile P2P streaming system with our handover scheme improved the playback continuity significantly compared to that without our scheme. We also show that we can further improve the performance by adjusting the weight value used when selecting an agent peer depending on peers' current situation.

The remainder of this paper is organized as follows. Section 2 describes related work to mobile IPTV systems and handover schemes. Section 3 describes our design considerations to develop an efficient handover scheme. Section 4 proposes an efficient handover scheme for mobile P2P live streaming systems. Section 5 presents extensive simulation results. Finally, Section 6 offers conclusions.

## 2. Related Work

So far most streaming systems have employed the client/server architecture based on content delivery networks (CDNs) that consist of many proxy servers geographically distributed on the Internet. However, this architecture requires tremendous cost to expand the network capacity for the rapidly increasing number of IPTV subscribers. To solve this scalability problem, many P2P streaming systems have thus been proposed. They can be broadly classified into tree-push and mesh-pull structures. The tree-push structures require much overhead to rebuild tree structures whenever peers join or leave [4]. On the other hand, the mesh-pull structures provide robust structure against peers' churn while creating long startup delay and requiring a large number of data exchanges among peers [5]. Thus, several hybrid push-pull architectures such as mTreebone [6] have been proposed to offer a good tradeoff between two structures. However, these P2P structures did not consider mobile platforms in a wireless network environment, only having focused on constructing overlay networks using stationary peers in a wired network environment.

With recent bandwidth improvement in wireless networks and advent of various mobile devices, mobile IPTV systems have become feasible. So far most of research efforts on mobile P2P streaming systems have been made in MANET [16] or iMANET [17]. However, they are not suitable for large-scale P2P systems due to their characteristics of high energy consumption and low scalability caused by direct communication between peers. To provide IPTV services with mobile devices, some systems have attempted to simply add mobile devices to the existing P2P streaming structure based on wired networks via APs [18]. As a result, they did not consider the characteristics of mobile computing environment such as low bandwidth, unstable wireless signal, and peers' high mobility when designing the mobile P2P live streaming systems.

On the other hand, many handover schemes have been proposed at each protocol layer to support mobile peers' mobility. The handover schemes at the data link layer have focused on performing the fast handover between two APs according to their signal strength using RSSI (received signal strength indication) [10, 11]. At the network layer, the proposed schemes have attempted to reduce the handover period by receiving CoA (care of address) as quickly as possible [12, 13]. To do so, they predict mobile peers' moving directions when they move from one subnet to another subnet. At the transport layer, new protocols such as mSCTP and mDCCP to add mobility features to the existing TCP and UDP protocols have been proposed [14, 15]. In these handover schemes operating at the layers below the application layer, however, it is not possible to avoid data loss because the network is physically disconnected during the handover period. To compensate for such data loss at lower layers, several application layer handover schemes have therefore been developed apart from lower layer schemes [19–21]. However, those application layer handover schemes have been developed based only on CDN structures, not considering the P2P live streaming structures.

## 3. Design Considerations for Seamless Mobility

We describe several considerations to reflect the characteristics of mobile devices and wireless networks when designing an efficient application-layer handover scheme for P2P live streaming systems.

*3.1. Data Transmission Manners for P2P Live Streaming.* In a mesh-based P2P streaming architecture, the data unit for data delivery and display is a video block. Each video is divided into small blocks, which are distributed to other peers through the mesh structure. Each peer displays video blocks after buffering and sequencing received blocks in memory. Peers periodically exchange their status using buffer maps that represent the blocks' availability in peers' buffers. After obtaining buffer maps from its neighbors, a peer can determine to which neighbor peers it will request missing blocks. In such a mesh-based streaming structure, where data

are transmitted in a pull manner, playback should be delayed until a peer can receive the sufficient data to start to playback a video.

On the other hand, in a tree-based P2P streaming architecture, peers receive video data from an origin server or parent peers only in a push manner. This structure enables peers to transmit data at a faster speed because they can keep transmitting data without any specific requests from their child peers once the tree structure is constructed.

In general, a mobile peer tends to experience data loss while communicating with others due to unstable wireless network environment. The mesh structure is thus more suitable for mobile P2P streaming architecture since a mobile peer can receive data more stably by requesting the retransmission of lost blocks. However, a peer has the longer delay when receiving data in a pull manner compared to that in a push manner. This is because, in a pull manner, it can receive the desired blocks from neighbor peers by specifically requesting them after exchanging buffer maps. Moreover, since a mobile peer cannot receive any data block during the handover, the transmission delay can be much longer after the handover. In our P2P live streaming system, a mobile peer therefore receives data in a push manner only for a short period around the handover to receive data at a faster speed.

**3.2. Criteria for Selecting Neighbor Peers for a Mobile Peer.** Even though a mobile peer is not able to receive any data only for a short period due to network condition, especially during the handover, it must continue to playback the video, keeping consuming the data that have been buffered before that period. As a result, a mobile peer may experience playback jitter unless it can quickly obtain the required data as soon as the network is available. To avoid such degradation of playback quality in our P2P live streaming system, we consider the proximity to a mobile peer as the first criterion when selecting neighbor peers. This can reduce the network latency through the shortened transmission route.

On the other hand, peers buffer the data corresponding to a specific period around their current playback positions. Since lag times between an origin server and peers are getting large as the number of peers increases, however, peers' buffering periods also become widely distributed. Furthermore, when supporting VCR operations, peers' playback positions become distributed more widely. Note that a mobile peer can receive data at a faster speed as neighbor peers are buffering more data required for its immediate playback. The other criterion is therefore how much data required by a mobile peer a candidate peer is currently buffering.

**3.3. Handover Prediction.** If a mobile peer cannot receive sufficient data before the handover, they may not be able to continue playing back the video due to lack of buffered data even though they receive the data at a fast speed after the handover. To prevent this situation, it is necessary to receive as much data as possible by predicting the handover before it actually happens. In our P2P live streaming system, we adopt the most common technique using signal strength of APs, that is, RSSI, to predict the handover. In other words, a mobile peer predicts that handover will occur soon when

the difference of signal strength between the current and the target AP becomes smaller than the given threshold. Once the handover is predicted, neighbor peers can transmit data to the mobile peer at a faster speed by switching their transmission manner to a push one.

## 4. An Efficient Handover Scheme for P2P Live Streaming

In this Section, we propose a new application-layer handover scheme to provide the seamless mobility in P2P live streaming systems. We first describe our handover behavior model according to the state of each peer. We then explain our agent peer selection policy to minimize playback jitter.

### 4.1. Handover Behavior Model

**4.1.1. State Transition for Handover Behavior.** In our proposed handover scheme, there are four states for each mobile peer: normal ( $N$ ), prediction ( $P$ ), handover ( $H$ ), and after-handover ( $A$ ) state. When a mobile peer is triggered by one of several events, it changes its state after taking the corresponding action depending on its current state as follows.

- (i) The  $N$  state represents the state where the signal of the current AP measured by a mobile peer is stronger than those of other adjacent APs by the threshold value for handover prediction.
- (ii) The  $P$  state indicates the period between after the handover is predicted to occur and before the handover actually occurs.
- (iii) The  $H$  state indicates the period when a mobile peer is switching its current AP to a new one while the network is being disconnected due to the handover.
- (iv) The  $A$  state represents the period when a mobile peer is receiving data through a newly connected AP for a short period until going to  $N$  state.

Figure 1 shows a state transition diagram of a mobile peer for its handover behavior. A mobile peer can move to another state depending on the events relating to the handover. First, a mobile peer usually enters into  $N$  state when it joins. When the difference of RSSI values of the current and target AP from the mobile peer becomes smaller than the threshold, it transits to  $P$  state. When the handover actually occurs, its state moves to  $H$  state. If the signal of the target AP from the mobile is getting weaker at  $P$  state, it returns to  $N$  state. Once the mobile peer is connected to the target AP, it transits to  $A$  state. Its state is changed to  $N$  state when the following two requirements are met: one is that the amount of buffered data should reach the same buffering level as that for initial playback and the other is that the difference of RSSI values of the current and each of other adjacent APs should become larger than the threshold value. If the handover occurs successively at  $A$  state, it returns to  $H$  state.

**4.1.2. Data Transmission Route and Manner.** In our P2P streaming system, a mobile peer has different data transmission route and manner according to its current state. As

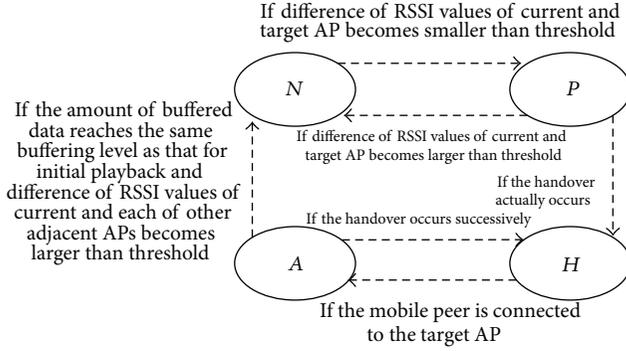


FIGURE 1: State transition diagram for handover behavior.

shown in Figure 2(a), at  $N$  state, it receives data from neighbor peers in a pull manner ( $N2M\_PL$ : neighbors to mobile in a pull manner). To maximize the amount of data that can be received at  $P$  state before the handover, the mobile peer receives data from neighbor peers after switching the transmission manner to a push one ( $N2M\_PS$ : neighbors to mobile in a push manner) as shown in Figure 2(b). To transmit data to the mobile peer at a faster speed after the handover, an agent peer also receives as much data as possible at this state. The agent peer thus receives data from neighbor peers in a push manner ( $N2A\_PS$ : neighbors to agent in a push manner).

As shown in Figure 2(c), at  $H$  state, the mobile peer cannot receive any data during the handover while the agent peer maintains  $N2A\_PS$  because it can still receive data from neighbor peers. The mobile peer transits to  $A$  state once it is connected to a new AP. It receives data as fast as possible from the agent peer as well as from neighbor peers in a push manner to minimize playback jitter ( $N2M\_PS$ ,  $A2M\_PS$ : agent to mobile in a push manner) as shown in Figure 2(d). After returning to  $N$  state, the mobile peer receives data from newly selected neighbor peers in a pull manner ( $N2M\_PL$ ) as shown in Figure 2(e).

**4.2. Agent Peer Selection Policy.** To further improve the playback continuity of a mobile peer when the handover occurs, a tracker server selects an agent peer for the mobile peer among stationary peers. In our P2P live streaming system, the agent peer plays an important role in reducing playback jitter caused by the handover. The agent peer receives data in place of the mobile peer from the moment the handover is predicted until the handover ends. It then transmits the buffered data to the mobile peer as fast as possible through a new AP so that the mobile peer cannot experience buffer starvation. To select the most suitable peer as an agent peer to perform this task, we thus consider a couple of criteria: RTT value and the appropriateness of the buffered period. The RTT value is considered to measure the transmission delay between a new AP and a candidate peer. The appropriateness of buffered period is considered to estimate how appropriate the period of the data buffered in a candidate peer is for the immediate playback of a mobile peer. In other words, it indicates how much required data from a mobile peer's perspective a candidate peer is currently

TABLE 1: Summary of simulation parameter values.

Parameter	Default value
Moving speed of mobile peers	5~20 Km/h
Handover latency	0.5~1 second
Bandwidth of backbone networks	100 Gbps
Bandwidth of wired networks	100 Mbps
Bandwidth of wireless networks	20 Mbps
Video playback rate	750 Kbps
Number of peers	1000
Number of neighbor peers	5
RTT values	1~200 ms
Buffering interval	3 seconds
$W$	0.5

buffering. The following equation represents the criteria to select an agent peer for a mobile peer:

$$\text{MIN} \{W \times \text{RTT}_i + (1 - W) \times \text{ABP}_i\}. \quad (1)$$

In (1),  $i$  is an index for a specific candidate peer,  $\text{RTT}_i$  and  $\text{ABP}_i$  denotes the RTT value from a new AP and the appropriateness of the buffered period of a candidate peer with an index of  $i$ , respectively, and  $W$  is the weight value between  $\text{RTT}_i$  and  $\text{ABP}_i$ . The candidate peer with a minimum value of (1) is selected as an agent peer for the corresponding mobile peer.

It is noted that  $W$  can be adjusted according to peers' current situation. If the network latency affects the performance more significantly in some situation, we need to increase  $W$ . On the contrary, in the situation where the appropriateness degree of buffered data for the playback of the mobile peer is a more important factor to improve the performance, it is necessary to decrease  $W$ .

## 5. Experimental Evaluation

To show the effectiveness of our proposed handover scheme for mobile P2P live streaming systems, we have performed extensive simulations using a QualNet network simulator. The default values of simulation parameters are shown in Table 1. They are used throughout our simulations unless otherwise indicated. It is assumed that mobile peers are moving at a speed of the range from 5 to 20 Km/h and the latency range of the handover is from 0.5 to 1 second. The bandwidths of backbone, wired, and wireless networks are set to 100 Gbps, 100 Mbps, and 20 Mbps, respectively. Each video has 750 Kbps playback rate. The numbers of peers are 1000 and each peer can have at most 5 neighbor peers. The RTT values between peers range from 1 to 200 ms and the average buffering interval starting from the current playback positions of peers is 3 seconds. The weight value of (1), that is,  $W$ , is set to 0.5.

**5.1. Effectiveness of Our Handover Scheme.** Figure 3 shows the comparison of playback continuity ratios for 7 seconds around the handover in two cases: with and without our

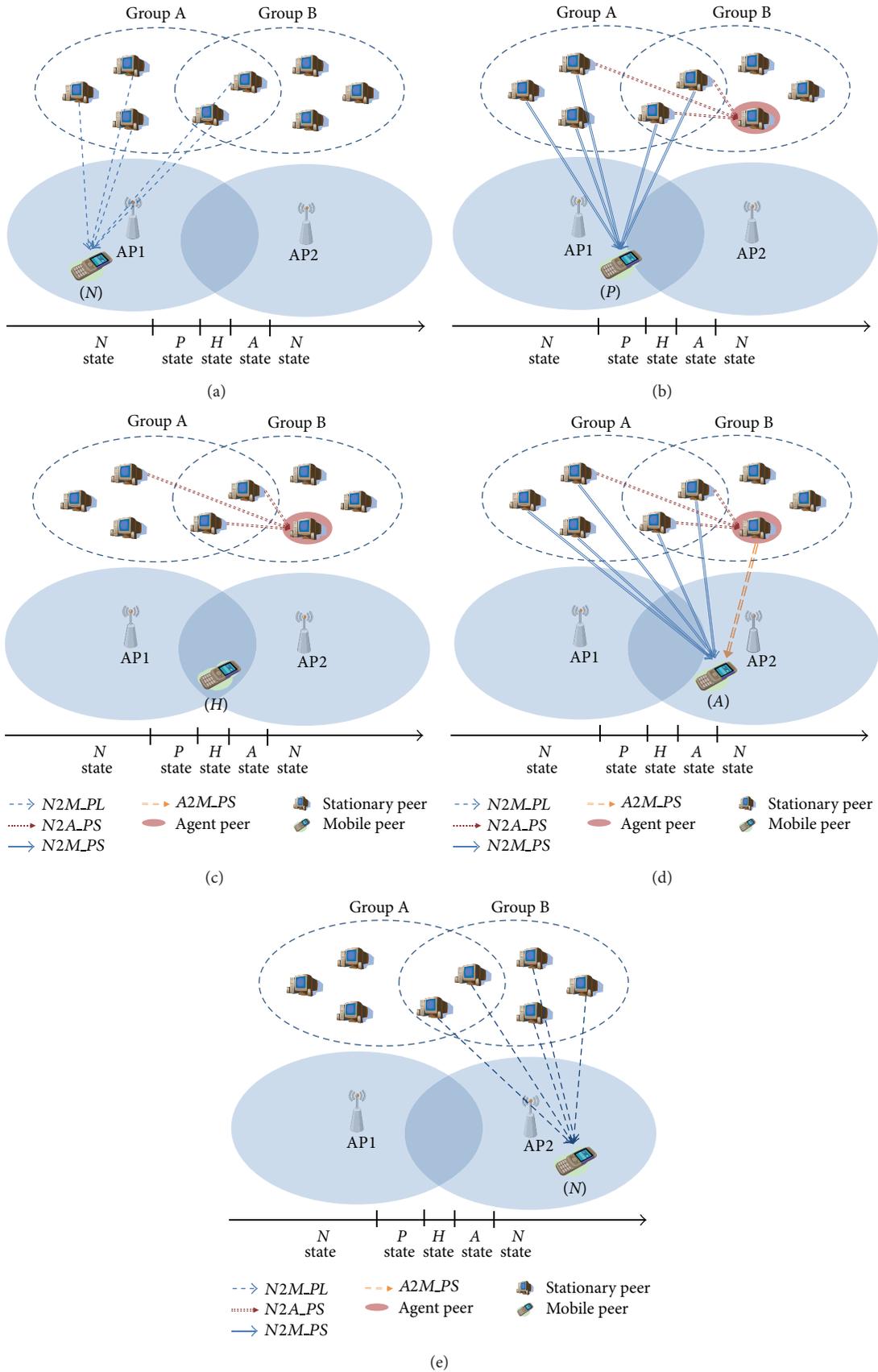


FIGURE 2: Data transmission route and manner according to a mobile peer's state: (a) *N* state before handover, (b) *P* state, (c) *H* state, (d) *A* state, and (e) *N* state after handover.

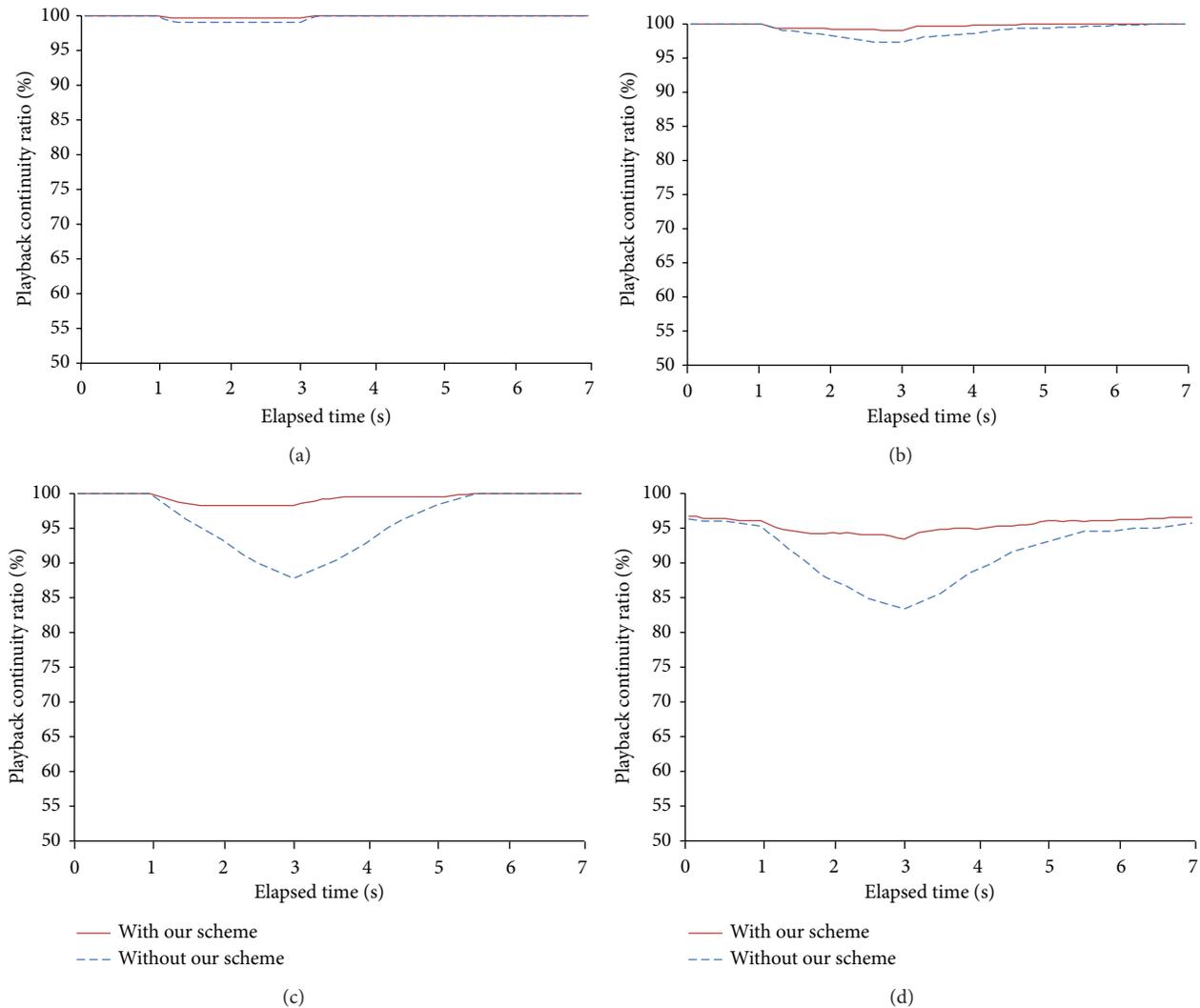


FIGURE 3: Playback continuity ratios of two cases according to effective wireless network bandwidth: (a) with effective bandwidth of 16 Mbps, (b) with effective bandwidth of 14 Mbps, (c) with effective bandwidth of 12 Mbps, and (d) with effective bandwidth of 10 Mbps.

proposed scheme. The case without our proposed scheme indicates the existing P2P live streaming systems based on mesh structure that do not perform any particular operation for the handover. To examine the impact of effective wireless network bandwidths on the performance, we varied them from 16 Mbps to 10 Mbps by generating background traffic from 4 to 10 Mbps. You can see that the handover occurs immediately after one second position from the beginning of the  $x$ -axis in Figure 3.

The experimental results show that the case with our handover scheme improved the playback continuity ratios significantly compared to the case without our scheme. Especially, in case of effective bandwidth of 10 Mbps in Figure 3(d), the difference of the minimum playback continuity ratios between two cases was 10.1%. That is, the minimum playback continuity ratio of the case with our scheme was 93.4% while that of the case without our scheme was only 83.3%. This implies that our handover scheme performs effectively no matter how much bandwidth the wireless network provides.

This is possible because our handover scheme can obtain the sufficient amount of data required by a mobile peer in advance through handover prediction and an agent peer. The mobile peer can also rush to receive data after the handover by adopting a push transmission manner.

It can also be seen from Figure 3 that, as the effective bandwidth of wireless networks decreases, that is, as the background traffic becomes heavier, the playback continuity ratios of two cases also decrease. It is noted that, however, the performance degradation degree in the case with our handover scheme is much lower than that without our scheme. As the background traffic increases from 4 Mbps to 10 Mbps, the minimum playback continuity ratio of the case with our scheme was reduced by only 6.3% while that without our scheme was reduced by 15.7%. This result indicates that our handover scheme can utilize the decreased network bandwidth efficiently. That is, a mobile peer can overcome the shortage of the network bandwidth after the handover by receiving as much data as possible before and during

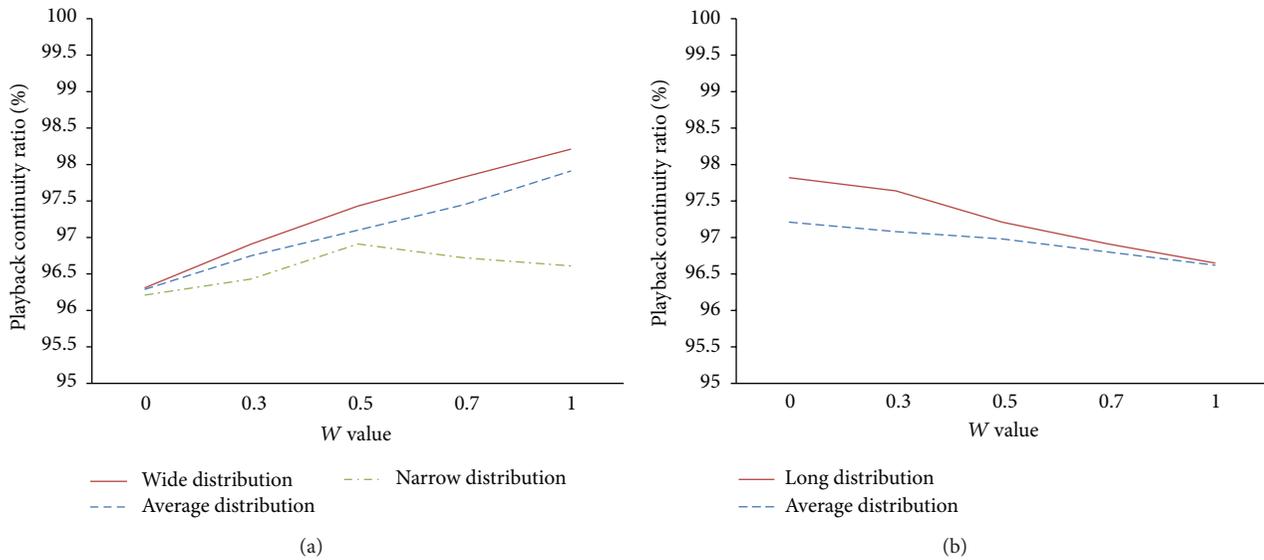


FIGURE 4: Impact of two criteria for agent peer selection on the playback continuity ratio: (a) RTT values and (b) lengths of average buffering interval.

the handover together with an agent peer. It can also reduce the transmission latency and the number of control messages considerably by receiving data in a push manner for the period around the handover.

**5.2. Impact of Agent Peer Selection Criteria.** To investigate the impact of two criteria in (1) on the performance when selecting an agent peer, we have made three different distributions for each criterion. The RTT values between a mobile peer and other candidate peers are distributed as follows: narrow distribution (5–100 ms), average distribution (1–200 ms), and wide distribution (0.1–1000 ms). The lengths of peers' average buffering intervals starting from the current playback positions are also distributed as follows: short distribution (1 second), average distribution (3 seconds), and long distribution (5 seconds). In our simulations, peers' RTT values and buffering interval lengths are randomly generated within the given range of each distribution. We also set the effective wireless network bandwidth to 12 Mbps for these simulations.

Figure 4(a) shows the playback continuity ratios according to the distribution degree of RTT values while fixing the distribution degree of buffering interval lengths to a short one. In case we employ wide and average distribution for buffering interval lengths, we achieved the highest playback continuity ratio when  $W$  is 1 while achieving the lowest one when  $W$  is 0. The differences between the highest and lowest ratio in case of wide and average distribution are 1.9% and 1.6%, respectively. Note that, as  $W$  increases, the playback continuity ratios also keep increasing almost linearly. This implies that, as peers are distributed more widely, that is, as the differences in RTT values are getting larger, it is more advantageous to select the peer with shorter RTT value from the mobile peer as an agent peer as indicated in (1). When using the narrow distribution for buffering interval lengths, the playback continuity ratio was highest when  $W$  is 0.5 while it was lowest when  $W$  is at both ends 0 and 1. This indicates

that the short distribution of RTT values and the narrow distribution of buffered interval lengths have similar degree from the perspective of the agent peer selection criteria. In such peers' situation, we thus need to consider two criteria to the same degree to maximize the performance.

Figure 4(b) shows the performance according to the distribution degree of buffering interval lengths when the distribution degree of RTT values is fixed to a narrow one. The simulation results show similar trends to those in Figure 4(a) except that the performance improved with the decreased value of  $W$ . That is, in case of the long and average distribution for RTT values, we achieved the highest playback continuity ratio when  $W$  is 0 while achieving the lowest one when  $W$  is 1. The improvement ratios in the long and average distribution are 1.2% and 0.6%, respectively. It can also be seen that the performance keeps improving with the decreased value of  $W$ . This indicates that, as average buffering interval lengths of peers are getting longer, it is more beneficial to select the peer that are buffering more data required by the mobile peer. In this case, we thus need to put more weight on the appropriateness of buffering period as indicated in (1).

From the simulation results in Figure 4, we have observed the impact of two criteria including RTT values and the appropriateness of the buffered periods when selecting an agent peer. Note that we can make the agent peer selection policy flexible in different peers' situation by adjusting  $W$  value, thereby further improving the performance.

## 6. Conclusions

We have presented an efficient application-layer handover scheme in mobile P2P live streaming systems to improve the playback continuity ratio significantly even though the handover occurs. This improvement was possible because a mobile peer can receive the sufficient amount of data required for the video playback in advance through handover

prediction and an agent peer. It can also receive data at a faster speed for the period around the handover by employing a push transmission manner. As the video contents requiring the higher playback rate, such as 3D and UD (ultrahigh definition) TV, are emerging, our handover scheme is expected to be widely applied to many applications in mobile platforms to provide seamless mobility even though the handover occurs.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013RIA1A2009913) and the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1017) supervised by the NIPA (National IT Industry Promotion Agency).

## References

- [1] X. Cheng and J. Liu, "NetTube: exploring social networks for peer-to-peer short video sharing," in *Proceedings of the 28th IEEE Conference on Computer Communications (INFOCOM '09)*, pp. 1152–1160, Rio de Janeiro, Brazil, April 2009.
- [2] E. Kim and C. Lee, "An on-demand TV service architecture for networked home appliances," *IEEE Communications Magazine*, vol. 46, no. 12, pp. 56–63, 2008.
- [3] C. Lee, E. Hwang, and D. Pyeon, "A popularity-aware prefetching scheme to support interactive P2P streaming," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 382–388, 2012.
- [4] M. Castro, P. Druschel, A. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: high-bandwidth multicast in cooperative environments," in *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP '03)*, pp. 298–313, Bolton Landing, NY, USA, October 2003.
- [5] X. Hei, C. Liang, J. Liang, Y. Liu, and K. Ross, "A measurement study of a large-scale P2P IPTV system," *IEEE Transactions on Multimedia*, vol. 9, no. 8, pp. 1672–1687, 2007.
- [6] F. Wang, Y. Xiong, and J. Liu, "mTreebone: a hybrid tree/mesh overlay for application-layer live video multicast," in *Proceedings of the 27th IEEE International Conference on Distributed Computing Systems (ICDCS '07)*, Toronto, Canada, June 2007.
- [7] D. Kim, E. Kim, and C. Lee, "Efficient peer-to-peer overlay networks for mobile IPTV services," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2303–2309, 2010.
- [8] Q. Qi, Y. Cao, T. Li, X. Zhu, and J. Wang, "Soft handover mechanism based on RTP parallel transmission for mobile IPTV services," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2276–2281, 2010.
- [9] J. Koo and K. Chung, "Adaptive channel control scheme to reduce channel zapping time of mobile IPTV service," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 357–365, 2011.
- [10] M. Portoles, Z. Zhong, S. Choi, and C. Chou, "IEEE 802.11 link-layer forwarding for smooth handoff," in *Proceedings of the 14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '03)*, pp. 1420–1424, Beijing, China, September 2003.
- [11] P. Khadivi, T. Todd, and D. Zhao, "Handoff trigger nodes for hybrid IEEE 802.11 WLAN/cellular networks," in *Proceedings of the 1st International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE '04)*, pp. 164–170, Dallas, Tex, USA, October 2004.
- [12] S. Speicher and C. Cap, "Fast layer 3 handoffs in AODV-based IEEE 802.11 wireless mesh networks," in *Proceedings of the 3rd IEEE International Symposium on Wireless Communication Systems (ISWCS '06)*, pp. 233–237, September 2006.
- [13] S. Sharma, N. Zhu, and T. Chiueh, "Low-latency mobile IP handoff for infrastructure-mode wireless LANs," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 4, pp. 643–652, 2004.
- [14] K. Brown and S. Singh, "M-UDP: UDP for mobile networks," *ACM SIGCOMM Computer Communication Review*, vol. 26, no. 5, pp. 60–78, 1996.
- [15] H. Elaarag, "Improving TCP performance over mobile networks," *ACM Computing Surveys*, vol. 34, no. 3, pp. 357–374, 2002.
- [16] T. Hara, "Replica allocation in ad hoc networks with period data update," in *Proceedings of the 3rd International Conference on Mobile Data Management (MDM '02)*, pp. 79–86, Singapore, January 2002.
- [17] F. Sailhan and V. Issarny, "Cooperative caching in ad-hoc networks," in *Proceedings of the 4th International Conference on Mobile Data Management (MDM '03)*, pp. 13–28, Melbourne, Australia, January 2003.
- [18] T. Zahariadis, O. Negru, and F. Alvarez, "Scalable content delivery over P2P convergent networks," in *Proceedings of the 12th IEEE International Symposium on Consumer Electronics (ISCE '08)*, pp. 1–4, Vilamoura, Portugal, April 2008.
- [19] T. Hong, K. Kang, D. Ahn, and H. Lee, "Adaptive buffering scheme for streaming service in intersystem handover between terrestrial and satellite systems," in *Proceedings of the 12th IEEE International Symposium on Consumer Electronics (ISCE '08)*, pp. 1–4, Vilamoura, Portugal, April 2008.
- [20] C. Huang and C. Lee, "Layer 7 multimedia proxy handoff using anycast/multicast in mobile networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 4, pp. 411–422, 2007.
- [21] B. Ciubotaru and G. M. Muntean, "SASHA—a quality-oriented handover algorithm for multimedia content delivery to mobile users," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 437–450, 2009.

## Research Article

# Goodness-of-Fit Based Secure Cooperative Spectrum Sensing for Cognitive Radio Network

**Hiep Vu-Van and Insoo Koo**

*The School of Electrical Engineering, University of Ulsan, San 29, Muger 2-dong, Ulsan 680-749, Republic of Korea*

Correspondence should be addressed to Insoo Koo; [iskoo@ulsan.ac.kr](mailto:iskoo@ulsan.ac.kr)

Received 7 March 2014; Accepted 25 April 2014; Published 18 May 2014

Academic Editor: Yuxin Mao

Copyright © 2014 H. Vu-Van and I. Koo. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cognitive radio (CR) is a promising technology for improving usage of frequency band. Cognitive radio users (CUs) are allowed to use the bands without interference in operation of licensed users. Reliable sensing information about status of licensed band is a prerequisite for CR network. Cooperative spectrum sensing (CSS) is able to offer an improved sensing reliability compared to individual sensing. However, the sensing performance of CSS can be destroyed due to the appearance of some malicious users. In this paper, we propose a goodness-of-fit (GOF) based cooperative spectrum sensing scheme to detect the dissimilarity between sensing information of normal CUs and that of malicious users, and reject their harmful effect to CSS. The empirical CDF will be used in GOF test to determine the measured distance between distributions of observation sample set according to each hypothesis of licensed user signal. Further, the DS theory is used to combine results of multi-GOF tests. The simulation results demonstrate that the proposed scheme can protect the sensing process against the attack from malicious users.

## 1. Introduction

Nowadays, more bandwidth and higher bit-rates have been required to meet usage demands due to an explosion in wireless communication technology. According to the Federal Communications Commission's spectrum policy task force report [1], the actual utilization of the licensed spectrum varies from 15% to 80%. In some cases, the utilization is only a small percentage of the total capacity. Cognitive radio (CR) technology [2] has been proposed to solve the problem of ineffective utilization of spectrum bands. Both unlicensed and licensed users, termed the cognitive radio user (CU) and primary user (PU), respectively, operate in CR networks. In CR network, CUs are allowed to access the frequency assigned to PU when it is free. But CU must vacate the occupied frequency when the presence of PU is detected. Therefore, reliable detection of the PU's signal is a requirement of CR networks.

In order to ascertain the presence of a PU, CUs can use one of several common detection methods, such as matched filter, feature, and energy detection [2, 3]. Energy detection is the optimal sensing method if the CU has limited

information about PU's signal (e.g., only the local noise power is known) [3]. In energy detection, frequency energy in the sensing channel is received in a fixed bandwidth  $W$  over an observation time window  $T$  to compare with the energy threshold and determine whether or not the channel is utilized. However, the received signal power may fluctuate severely due to multipath fading and shadowing effects. Therefore, it is difficult to obtain reliable detection with only one CU. Better sensing performance can be obtained by allowing some CUs to perform cooperative spectrum sensing [4–6].

CSS can use some combination methods such as equal gain combination (EGC) and maximum gain combination (MGC) [7] to combine sensing information of all CUs in the network and make a global decision about status of PU signal. Since EGC gives the same weight for all CUs in the network, it is easy to execute but with limited performance. MGC is known as the optimal combination rule. However, it requires information about the SNRs of the sensing channel, which is difficult to obtain in practice. In addition, MGC is sensitive to attack by malicious users who send false sensing data to the fusion center (FC) [8]. The research presented in [8, 9]

determined that the presence of a few malicious users can severely reduce the performance of a CSS scheme. Algorithms used to identify the malicious users have been proposed in the studies of [8, 9]. In previous research, a simple technique (i.e., outlier-detection) is used to detect less damage malicious CUs such as *always No* or *always Yes* CU. In addition, the technique is unable to protect the CSS in the event of a large number of malicious users in the network.

In this paper, we utilize multi-goodness-of-fit (GOF) tests to design a robust CSS, in which the event detection technique [10, 11] will be used to provide the combination of different evidence of each type of GOF test which are supported by a particular hypothesis of PU signal. The proposed scheme considers two types of GOF tests, Kolmogorov-Smirnov (KS) and Cramer-von Mises (CM) tests. The proposed scheme can distinguish the sensing information of normal CUs and that of malicious users and reject the harmful effect of malicious user to sensing combination process. Three common types of malicious users including *always Yes*, *always No*, and *opposite* are considered in this paper.

## 2. Background

**2.1. Goodness-of-Fit Test.** The GOF test summarizes the discrepancy between the observed samples with theoretical distributions or empirical distributions and the reference distribution. For the  $n$  independent and identical distributed observation, the sample is first arranged in ascending order such that  $s_1 \leq s_2 \leq \dots \leq s_n$ . The GOF test is used to determine whether or not the samples set was drawn from the same distribution with a cumulative distribution function (CDF)  $F_0$ . The testing hypothesis can be formulated as follows:

$$\begin{aligned} F(s) &= F_0(s) : H_0, \\ F(s) &\neq F_0(s) : H_1, \end{aligned} \quad (1)$$

where  $F(s)$  is the empirical CDF of the sample. It can be calculated as follows:

$$F(s) = \frac{1}{n} \sum_{i=1}^n I\{s_i \leq s\}, \quad (2)$$

where  $I\{\cdot\}$  is the indicator of event  $\{\cdot\}$ .

There are many types of GOF tests, for instance, Cramer-von Mises (CM), Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Hosmer-Lemeshow (HL) tests. In this paper, we consider two types of GOF tests, CM and KS tests, which can run well with a low number of samples.

- (1) Kolmogorov-Smirnov (KS) test: the KS test, which is based on the empirical CDF of the samples set and the reference CDF, can be calculated according to the largest difference of two distributions as follows:

$$D_{KS} = \sup \{|F(s_i) - F_0(s_i)| : i = 1, \dots, n\}, \quad (3)$$

where  $\sup\{\cdot\}$  is supremum function, which indicates the greatest element of the set. If the sample comes from distribution  $F_0(x)$ , then  $D_{KS}$  will converge to 0.

- (2) Cramer-von Mises (CM) test: CM test is used for judging the goodness-of-fit of the sample set's CDF  $F(s)$  and reference distribution's CDF  $F_0(s)$ . The test statistic is given by

$$D_{CM} = n \int_{-\infty}^{+\infty} [F(s) - F_0(s)]^2 dF_0(s) \quad (4)$$

and can be approximated as

$$D_{CM} = \frac{1}{12n} + \sum_{i=1}^n \left[ \frac{2i-1}{2n} - F_0(s_i) \right]^2. \quad (5)$$

If this value,  $D_{CM}$ , is larger than the threshold, the hypothesis that the sample data come from the reference distribution  $F_0$  can be rejected.

**2.2. Combination of Evidence in Dempster-Shafer Theory.** Dempster-Shafer (DS) theory was first introduced by Dempster and was later extended by Shafer. This is a potentially valuable tool for the evaluation of risk and reliability in engineering applications when it is not possible to obtain a precise measurement from experiments or when knowledge is obtained from expert elicitation. An important aspect of this theory is the combination of evidence obtained from multiple sources and the modeling of conflict between them.

In DS theory [12], a representation of ignorance is provided by assigning a nonzero mass function to hypothesis  $m$ , also called the basic probability assignment (BPA), and is defined for every hypothesis  $A$  such that the mass value  $m(A)$  belongs to the interval  $[0, 1]$  and satisfies the following conditions:

$$\begin{aligned} m(\phi) &= 0, \\ \sum m(A) &= 1, \quad A \subseteq \Omega, \end{aligned} \quad (6)$$

where  $\Omega$  is the frame of discernment, which is a fixed set of  $q$  mutually exclusive and exhaustive elements.

By assigning a nonzero mass in a compound hypothesis,  $A \cup B$  means that there exists the option to not make a decision between  $A$  and  $B$  but to leave the formulation in the  $A \cap B$  class. In DS theory, two functions, belief (Bel) and plausibility (Pls), are defined to characterize the uncertainty and support of certain hypotheses. Bel measures the minimum or necessary support, whereas Pls reflects the maximum or potential support for that hypothesis [13]. These two measures, derived from mass values, are defined as a map from a set of hypotheses to interval  $[0, 1]$  as follows:

$$\begin{aligned} \text{Bel}(A) &= \sum_{B \subseteq A} m(B), \\ \text{Pls}(A) &= \sum_{A \cap B \neq \emptyset} m(B). \end{aligned} \quad (7)$$

The sum of mass functions from different information source,  $m_j$  ( $j = 1, 2, \dots, M$ ), combined with the DS rule is known as the orthogonal sum, which is commutative and

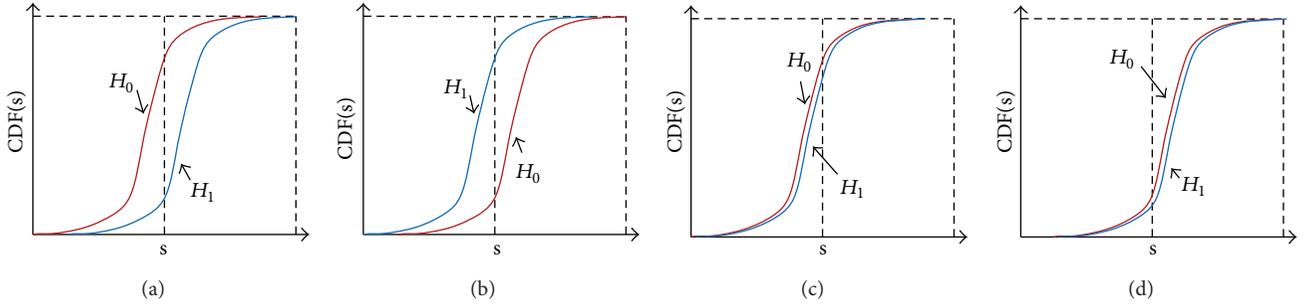


FIGURE 1: The CDF of received signal energy at CU under absence and presence hypothesis of PU signal for (a) normal CU, (b) *opposite* malicious CU, (c) *always Yes* malicious CU, and (d) *always No* malicious CU.

associative. The result is a new mass function,  $m(A_k) = (m_1 \oplus m_2 \oplus \dots \oplus m_d)(A_k)$ , which incorporates the joints information provided by the sources as follows:

$$m(A_k) = \frac{1}{1-K} \sum_{A_1 \cap A_2 \dots A_d = A_k} \left( \prod_{1 \leq j \leq M} m_j(A_j) \right), \quad (8)$$

$$K = \sum_{A_1 \cap A_2 \dots A_d = \phi} \left( \prod_{1 \leq j \leq M} m_j(A_j) \right),$$

where  $K$  is the measure of conflict between the different sources and is introduced as a normalization factor.

### 3. The Proposed Secure Cooperative Spectrum Sensing Based on GOF Test

There is a definite difference between the CDF of received signal energy of normal CU and that of the malicious users as shown in Figure 1. The CDF of the received signal energy of normal CU corresponding to the presence of the PU is always “*under*” that one corresponding to the absence of the PU. On the contrary, the *opposite malicious* CU has the CDF corresponding to the presence of PU to be “*above*” the CDF corresponding to the absence of PU. The *always Yes* and *always No* malicious CUs have a similar CDF corresponding to presence and absence of PU. Due to the difference between CDF of normal and malicious CUs, we utilize GOF test to detect the appearance of malicious users in the network, so that their harmful effect can be rejected out of CSS process. Multi-GOF tests including KS and CM tests will be applied for adaptive robust CSS. The DS theory will be used to combine results of multi-GOF tests.

In this paper, we consider a CR network including  $N$  CUs who cooperate to sense the signal from a PU. There are  $p < N$  malicious CUs appearing in the network which can be classified as three common types: *always Yes*, *always No*, and *opposite* malicious CUs. All CUs use energy detectors to perform spectrum sensing and send their sensing data to the FC through a control channel. Based on the sensing data obtained from the CUs, the FC makes a global decision concerning the presence or absence of the PU signal by using

the proposed data fusion scheme. The proposed scheme has 3 steps as follows.

*Step 1.* All CUs perform spectrum sensing by using energy detection method to determine received signal energy  $E_j = \{e_{1,j}, e_{2,j}, \dots, e_{M,j}\}$ , where  $M$  is the number of sensing samples that the  $j$ th CU takes in the sensing interval.

*Step 2.* At the FC, GOF test statistics of each CU will be computed according to hypothesis of the PU as given in (11). After that, BPA and final BPA for current sensing data will be estimated based on the “*reputation level*” of each CU, which is updated from previous sensing interval. Based on final BPA, a global decision rule will be proposed to make global decision about status of PU signal.

*Step 3.* Update “*reputation level*” of each CU according to the global decision.

The detailed description of each step will be given in the following subsections.

*3.1. Energy Detection.* At the sensing interval for the  $j$ th CU, the local spectrum sensing is to decide between the two following hypotheses:

$$H_0 : s_j(k) = n_j(k),$$

$$H_1 : s_j(k) = h_j p(k) + n_j(k), \quad (9)$$

where  $H_0$  and  $H_1$  correspond to the hypothesis of the absence and presence of the PU signal, respectively,  $h_j$  denotes the amplitude gain of the channel,  $s(k)$  is the signal transmitted from the PU,  $n_j(k)$  is the additive white Gaussian noise, and  $k$  is index of sensing sample at each sensing interval.

A received signal energy of a sensing sample,  $e_{k,j}$ , is given as

$$e_{k,j} = \begin{cases} |n_j(k)|^2, & H_0 \\ |h_j s(k) + n_j(k)|^2, & H_1. \end{cases} \quad (10)$$

*3.2. BPA Estimation.* The GOF test statistics of the current sensing data  $e_{k,j}$  ( $k = 1, \dots, M$ ) of the  $j$ th CU will be

TABLE 1: Reputation ranges according to each type of CUs.

Status of PU	Normal CU	Always Yes CU	Always No CU	Opposite CU
$H_1$	$D_{0,j}^t \gg D_{1,j}^t$ $D_{1,j}^t \approx 0$ $r_{0,j}^t(i) \gg 0$	$D_{0,j}^t \gg D_{1,j}^t$ $D_{1,j}^t \approx 0$ $r_{0,j}^t(i) \gg 0$	$D_{0,j}^t \ll D_{1,j}^t$ $D_{0,j}^t \approx 0$ $r_{0,j}^t(i) \ll 0$	$D_{0,j}^t \ll D_{1,j}^t$ $D_{0,j}^t \approx 0$ $r_{0,j}^t(i) \ll 0$
$H_0$	$D_{0,j}^t \ll D_{1,j}^t$ $D_{0,j}^t \approx 0$ $r_{1,j}^t(i) \gg 0$	$D_{0,j}^t \gg D_{1,j}^t$ $D_{1,j}^t \approx 0$ $r_{1,j}^t(i) \ll 0$	$D_{0,j}^t \ll D_{1,j}^t$ $D_{0,j}^t \approx 0$ $r_{1,j}^t(i) \gg 0$	$D_{0,j}^t \gg D_{1,j}^t$ $D_{1,j}^t \approx 0$ $r_{1,j}^t(i) \ll 0$

calculated according to each hypothesis of PU signal based on (3) and (5) as follows, respectively:

$$D_{h,j}^{\text{KS}} = \sup \left\{ \left| F(e_{k,j}) - F_h(e_{k,j}) \right| : k = 1, 2, \dots, M \right\},$$

$$D_{h,j}^{\text{CM}} = \frac{1}{12M} + \sum_{i=1}^M \left[ \frac{2i-1}{2M} - F_h(e_{k,j}) \right]^2, \quad (11)$$

where  $h = \{0, 1\}$  is index of hypothesis  $H_h$  of PU signal,  $e_{k,j}$  is the received signal energy of  $k$ th sensing sample of the  $j$ th CU,  $F(\cdot)$  and  $F_h(\cdot)$  are empirical CDF of observed sensing sample and CDF of  $H_h$  hypothesis of PU, and  $M$  is the number of samples for each sensing interval.

It is noteworthy that normal CU and malicious CU have different characteristics of  $D_{1,j}^t$  and  $D_{0,j}^t$  as shown in Table 1, where  $t$  indexes types of GOF tests: KS and CM.

Based on the values of  $D_{1,j}^t$  and  $D_{0,j}^t$ , we will estimate BPA of current sensing data of each CU and their "reputation level" for robust CSS as follows:

$$\Delta_{1,j}^t = \frac{D_{0,j}^t}{D_{1,j}^t + D_{0,j}^t} R_{0,j}^t, \quad (12)$$

$$\Delta_{0,j}^t = \frac{D_{1,j}^t}{D_{1,j}^t + D_{0,j}^t} R_{1,j}^t,$$

where  $R_{h,j}^t$  is "reputation level" of the  $j$ th CU according to hypothesis  $h$ , and it can be determined based on history observation of the  $j$ th CU as follows:

$$R_{h,j}^t(i) = \frac{r_{h,j}^t(i-1)}{\sum_j r_{h,j}^t(i-1)}, \quad (13)$$

where  $i$  is the index of current sensing interval and  $r_{1,j}^t(i-1)$  and  $r_{0,j}^t(i-1)$  are updated from the previous sensing interval according to global decision:

$$r_{1,j}^t(i-1) = r_{1,j}^t(i-2) + (D_{1,j}^t(i-1) - D_{0,j}^t(i-1)), \quad (14)$$

$$r_{0,j}^t(i-1) = r_{0,j}^t(i-2) + (D_{0,j}^t(i-1) - D_{1,j}^t(i-1)). \quad (15)$$

By using  $r_{1,j}^t$  and  $r_{0,j}^t$ , types of CUs will be easily distinguished. The normal CU has positive value of both  $r_{1,j}^t$  and  $r_{0,j}^t$  that will be increased after updating step.  $r_{1,j}^t$  of *always*

*Yes* and  $r_{0,j}^t$  of *always No* malicious CUs are almost negative and tend to decrease after updating step. On the other hand, both values of opposite CU are negative and have a tendency to decrease. We define "malicious threshold" as  $\rho$  to reject the attack of malicious CR in CSS, so that the CU, which has either  $r_{1,j}^t < 0$  or  $r_{0,j}^t < 0$ , will be determined as malicious CU. The sensing data of malicious CUs will not be considered to make global decision by giving them  $r_{1,j}^t = 0$  and  $r_{0,j}^t = 0$ .

The BPA of all CUs will be combined with their reputation levels as

$$\Delta_1^t = \frac{1}{n_\Omega} \sum_{j \in \Omega} \frac{D_{0,j}^t}{D_{1,j}^t + D_{0,j}^t} R_{0,j}^t, \quad (16)$$

$$\Delta_0^t = \frac{1}{n_\Omega} \sum_{j \in \Omega} \frac{D_{1,j}^t}{D_{1,j}^t + D_{0,j}^t} R_{1,j}^t,$$

where  $\Omega$  and  $n_\Omega$  are set of normal CUs and number of members of the set, respectively.

Because the error in estimating  $\Delta_0^t$  and  $\Delta_1^t$ ,  $\Delta_0^t + \Delta_1^t$  can be bigger than 1, we need to normalize those values as

$$\Delta_0^{t*} = \frac{\Delta_0^t}{\Delta_0^t + \Delta_1^t}, \quad (17)$$

$$\Delta_1^{t*} = \frac{\Delta_1^t}{\Delta_0^t + \Delta_1^t}.$$

**3.3. DS Theory Combination.** The DS theory will be used to combine the BPA of both GOF tests according to each hypothesis as follows:

$$\begin{aligned} \Delta_1 &= \Delta_1^{\text{KS}*} \oplus \Delta_1^{\text{CM}*} \\ &= \frac{\Delta_1^{\text{KS}*} \Delta_1^{\text{CM}*}}{1 - (\Delta_1^{\text{KS}*} \Delta_0^{\text{CM}*} + \Delta_0^{\text{KS}*} \Delta_1^{\text{CM}*})}, \\ \Delta_0 &= \Delta_0^{\text{KS}*} \oplus \Delta_0^{\text{CM}*} \\ &= \frac{\Delta_0^{\text{KS}*} \Delta_0^{\text{CM}*}}{1 - (\Delta_1^{\text{KS}*} \Delta_0^{\text{CM}*} + \Delta_0^{\text{KS}*} \Delta_1^{\text{CM}*})}. \end{aligned} \quad (18)$$

Finally, the global decision will be made as follows:

$$G = H_1, \quad \text{if } \frac{\Delta_1}{\Delta_0} \geq \eta, \quad (19)$$

$$G = H_0, \quad \text{otherwise,}$$

where  $\eta$  is the threshold for global decision.

According to the global decision,  $r_{1,j}^t$  or  $r_{0,j}^t$  will be updated for the next sensing interval as follows, respectively.

(i) If the global decision is  $G(i) = 0$ , we update  $r_{1,j}^t(i)$  by using (14).

(ii) Otherwise, we update  $r_{0,j}^t(i)$  by using (15).

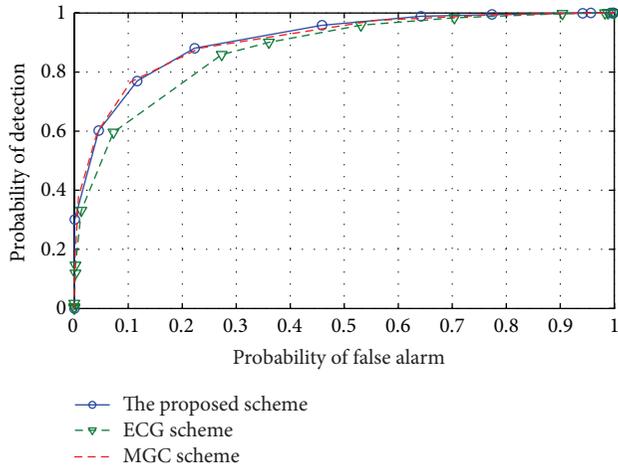


FIGURE 2: ROC of the proposed scheme and reference schemes when no malicious CU is considered.

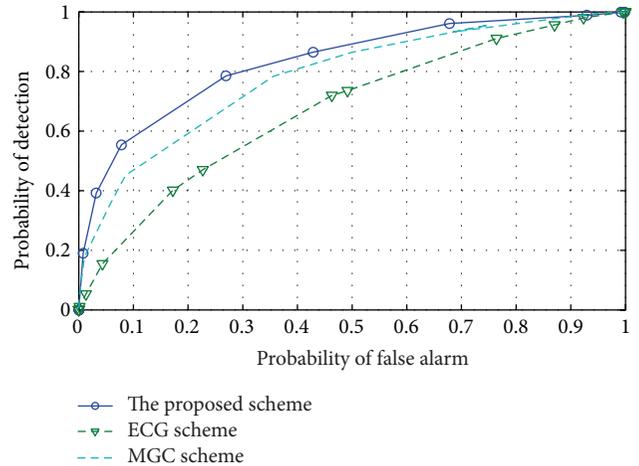


FIGURE 4: ROC of the proposed scheme and reference schemes when 4 *always Yes* malicious CUs are considered.

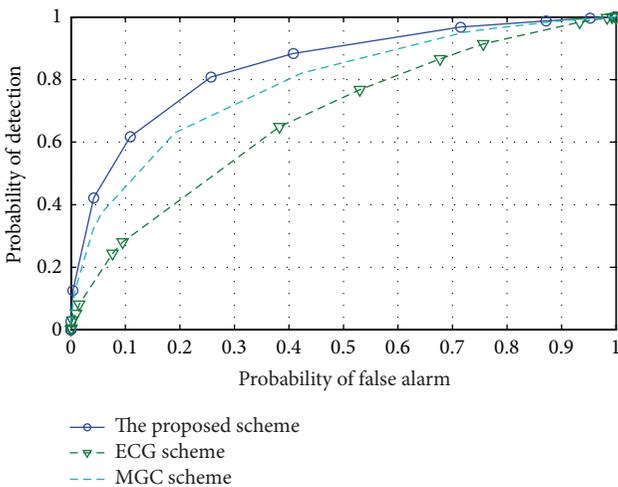


FIGURE 3: ROC of the proposed scheme and reference schemes when 4 *always No* malicious CUs are considered.

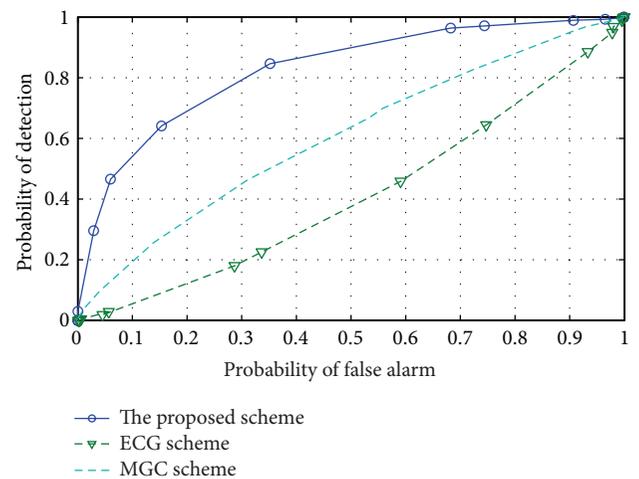


FIGURE 5: ROC of the proposed scheme and reference schemes when 4 *opposite* malicious CUs are considered.

### 4. Simulation Results

In this section, simulation results of the proposed scheme and other soft combination schemes such as maximum gain combination (MGC) and equal gain combination (EGC) are provided. The network is considered in which 5 CUs exist and some of them can be malicious CUs.

In order to verify the reliability of the proposed combination scheme, we perform a simulation without considering malicious CU. The sensing results in Figure 2 show that the proposed scheme can obtain better sensing performance in comparison with EGC scheme and obtain a similar sensing performance to that of the MGC scheme when no malicious CU is considered.

The robustness of the proposed scheme will be investigated in the network with the appearance of *always No*, *always Yes*, and *opposite* malicious CUs in the network. Figures 3 and 4 show performance of the proposed scheme when 4 CUs

are *always No* or *always Yes* malicious CUs among 5 CUs in the network. The results show that the proposed scheme with all CUs can achieve much better sensing performance than that one of the MGC and EGC schemes. This means that, by applying GOF test to CSS, the proposed scheme can detect the presence of those types of malicious CUs and reject their harmful effects to sensing process.

Opposite malicious CU causes the most damage to sensing performance. However, the proposed scheme is expected to protect CSS against this type of malicious CU. Figure 5 shows the sensing performance of the network when 4 CUs are *opposite* malicious among 5 CUs. MGC and EGC with all CUs provide very low performance due to the attack of *opposite* malicious CU. However, the proposed scheme can defend their attacks and achieve high sensing performance.

## 5. Conclusion

In this paper, multi-GOF tests are proposed to measure the difference between sensing data of normal CU and that of malicious CU. Further, the DS theory is used to combine results of multi-GOF tests. The proposed scheme considers the appearance of the most common types of malicious CU: *always Yes*, *always No*, and *opposite* types. The simulation results prove that the proposed scheme can reject almost harmful effect from those malicious CUs to protect CSS.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was supported by the KRF funded by the MEST (NRF-2012R1A1A2038831).

## References

- [1] Federal Communications Commission, "Spectrum policy task force," Tech. Rep. 02-135, ET Docket, 2002.
- [2] Y. Hur, J. Park, W. Woo et al., "A wideband analog Multi-Resolution Spectrum Sensing (MRSS) technique for cognitive radio (CR) systems," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '06)*, pp. 4090–4093, Island of Kos, Greece, May 2006.
- [3] A. Sahai, N. Hoven, and R. Tandra, "Some fundamental limits on cognitive radio," in *Proceedings of the Allerton Conference on Communications, Control, and Computing*, Monticello, Va, USA, 2004.
- [4] G. Ganesan and Y. Li, "Cooperative spectrum sensing in cognitive radio networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 137–143, Baltimore, Md, USA, November 2005.
- [5] S. M. Mishra, A. Sahai, and R. W. Brodersen, "Cooperative sensing among cognitive radios," in *Proceedings of the IEEE International Conference on Communications (ICC '06)*, vol. 5, pp. 1658–1663, July 2006.
- [6] R. Deng, J. Chen, C. Yuen, P. Cheng, and Y. Sun, "Energy-efficient cooperative spectrum sensing by optimal scheduling in sensor-aided cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 2, pp. 716–725, 2012.
- [7] J. Ma and Y. Li, "Soft combination and detection for cooperative spectrum sensing in cognitive radio networks," in *Proceedings of the 50th Annual IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 3139–3143, November 2007.
- [8] P. Kaligineedi, M. Khabbaziyan, and V. K. Bhargava, "Secure cooperative sensing techniques for cognitive radio systems," in *Proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 3406–3410, May 2008.
- [9] P. Kaligineedi, M. Khabbaziyan, and V. K. Bhargava, "Malicious user detection in a cognitive radio cooperative sensing system," *IEEE Transactions on Wireless Communications*, vol. 9, no. 8, pp. 2488–2497, 2010.
- [10] J. Li, J. Liu, and K. Long, "Reliable cooperative spectrum sensing algorithm based on Dempster-Shafer theory," in *Proceedings of the 53rd IEEE Global Communications Conference (GLOBECOM '10)*, pp. 1–5, December 2010.
- [11] X. Zheng, J. Wang, Q. Wu, and J. Chen, "Cooperative spectrum sensing algorithm based on Dempster-Shafer theory," in *Proceedings of the 11th IEEE Singapore International Conference on Communication Systems (ICCS '08)*, pp. 218–221, November 2008.
- [12] N. Nguyen-Thanh and I. Koo, "Empirical distribution-based event detection in wireless sensor networks: an approach based on evidence theory," *IEEE Sensors Journal*, vol. 12, no. 6, pp. 2222–2228, 2012.
- [13] K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," Sandia Report SAND2002-0835, Sandia National Laboratories, Albuquerque, NM, USA, 2002.

## Research Article

# Mobility-Assisted on-Demand Routing Algorithm for MANETs in the Presence of Location Errors

**Trung Kien Vu and Sungoh Kwon**

*School of Electrical Engineering, University of Ulsan, Ulsan 680-749, Republic of Korea*

Correspondence should be addressed to Sungoh Kwon; [sungoh@ulsan.ac.kr](mailto:sungoh@ulsan.ac.kr)

Received 7 March 2014; Accepted 30 April 2014; Published 18 May 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 T. K. Vu and S. Kwon. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose a mobility-assisted on-demand routing algorithm for mobile ad hoc networks in the presence of location errors. Location awareness enables mobile nodes to predict their mobility and enhances routing performance by estimating link duration and selecting reliable routes. However, measured locations intrinsically include errors in measurement. Such errors degrade mobility prediction and have been ignored in previous work. To mitigate the impact of location errors on routing, we propose an on-demand routing algorithm taking into account location errors. To that end, we adopt the Kalman filter to estimate accurate locations and consider route confidence in discovering routes. Via simulations, we compare our algorithm and previous algorithms in various environments. Our proposed mobility prediction is robust to the location errors.

## 1. Introduction

A mobile ad hoc network (MANET) [1] consists of a set of wireless mobile nodes that dynamically exchange data among themselves without relying on any fixed infrastructure. Because of their easy deployment and extension, MANET application scenarios include emergency and rescue operations, conference settings, car networks, and personal networking. Due to limited transmission ranges and infrastructure-free networks, each node in such networks has the responsibility not only to discover new routes but also to relay messages.

The most challengeable problem of MANETs [2] is how to adapt the topology changing that affects the performance of the network [3, 4]. Due to changeable topology, routes from sources to destinations may be suddenly broken and nodes have to discover other available routes to deliver data. The ad hoc on-demand distance vector routing algorithm (AODV) was proposed as a reactive routing algorithm to allow mobile nodes to quickly adapt to topology changes and link breaks in mobile ad hoc networks [5]. To find a possible route, the AODV makes a source flood a routing request message over the network and discovers a route based on the principle of the shortest path. The amount of overhead messages for route

discovery and route maintenance depends on the longevity of routing paths. The awareness of link and path durations can improve routing performance in such mobile networks [6–8].

In [9, 10], the authors modeled the distribution of path duration and analyzed the relation between path duration and other factors such as relative speed, transmission range, and number of hops. Their analysis shows that routing protocol with higher path duration can improve the network performance. In [11], the authors also investigate the distribution of path duration and then design a scheme to select a route with the largest expected duration and provide reliable network services in MANETs.

Location information enables nodes to predict mobility and estimate path durations more accurately. In [12–14], the authors proposed schemes to improve routing performance with location awareness. The proposed algorithms in [12, 13] anticipate the link expiration time (LET) based on measured locations and velocities and were applied to routing protocols to reduce overheads in [12] or to select the most reliable route that has the longest path duration [13]. In [14], the link duration time is adaptively applied to route maintenance in order to reduce unnecessary overhead. However, lifetime of link may be incorrectly calculated due to location errors that lead to incorrect hello frequency setting.

In practical deployment scenarios, location errors intrinsically occur in measurement [15], even if locations are measured by the global positioning system (GPS) receiver. Such imperfect location information leads imperfect mobility prediction, which results in performance degradation. However, the previous work assumed error-free location information and developed routing algorithms. In [12], the impact of location errors on routing performance was provided only by simulations, but there is no effort to improve routing performance in such noisy information environments. Therefore, it is necessary to develop an efficient routing that is robust to location errors.

In this paper, we proposed a mobility-assisted on-demand routing algorithm in the presence of location errors in order to mitigate the impact of location errors on routing performance. To that end, the algorithm adopts the Kalman filter to compensate for the measurement location errors and estimate link durations to reduce overheads and select reliable routes. We also consider the confidence level of route in selecting the best route. Via simulations, we compare our proposed algorithm with previous algorithms.

The rest of this paper is organized as follows. In Section 2, we describe the system model and problem. In Section 3, we propose the Kalman filter based routing algorithm with mobility prediction for location correction and route selection. In Section 4, we provide numerical results to analyze the impact of location errors and the efficient of our proposal in the presence of location errors, and we conclude the paper in Section 5.

## 2. System Model and Problem

*2.1. System Model.* In this paper, we consider a mobile wireless network that supports multihop routing. The network is modeled as a set  $\mathcal{N}$  of mobile nodes with transmission range  $r$  and a set  $\mathcal{L}$  of communication links  $(i, j)$  between nodes  $i$  and  $j$  in  $\mathcal{N}$ .

Link  $(i, j)$  is called *valid* or *connected* link at time  $t_k$  when the distance between nodes  $i$  and  $j$  at time  $t_k$  is less than or equal to the transmission range  $r$ ; that is,

$$\|X_i(t_k) - X_j(t_k)\| \leq r, \quad (1)$$

where  $X_i(t_k)$  and  $X_j(t_k)$  are locations of nodes  $i$  and  $j$ , respectively, and  $\|X\|$  stands for a Euclidian distance of vector  $X$ . Otherwise, link  $(i, j)$  is considered *broken* or *disconnected*, because the two nodes are out of their communication range. The link duration of link  $(i, j)$  is defined as the time interval for which the link is valid.

Due to a limited transmission range, packets are delivered from a source to a destination in a multihop manner via a route, which is defined as a set of links. For given source and destination nodes,  $s$  and  $d$ , respectively,  $H$  possible routes at time  $t_k$  are denoted as  $R_{(s,d)}^{(h)}(t_k)$  for  $h \in \mathcal{H} = \{1, \dots, H\}$ , which consists of  $|R_{(s,d)}^{(h)}(t_k)|$  links.

To find a route from a source to a destination and maintain routes, each mobile node employs the AODV routing algorithm, which is one of the reactive routing protocols and frequently adopted in mobile ad hoc networks.

*2.2. Overview of AODV.* The AODV [5] routing algorithm consists of two main operations: route discovery and route maintenance. Route discovery is initiated by a source node that has data to send a destination node and does not have an active route in its routing table. To find a valid route to the destination, the source node broadcasts a route request (RREQ) message, including a sequence number, to neighboring nodes. The RREQ message is flooded through the entire network until the message reaches the destination or an intermediate node that has a valid route to the destination. Each node that receives the RREQ message stores a reverse route to the source and then broadcasts the message to their neighboring nodes if the node is not the destination and the RREQ message is not a duplicate. When the RREQ message arrives at a destination node or at an intermediate node that has a valid route to the destination, the node sends a route reply (RREP) message to the neighboring node in a reverse route in a unicast manner. The RREP message contains the number of hops to reach the destination node and the sequence number for the destination. A node receiving the RREP message sends this message to the source via the stored reverse route and then creates or updates a forward route to the destination.

Route maintenance is performed by nodes after route discovery operation, in order to maintain local connectivity and routes. Nodes periodically send a hello message to their neighbors to check if links are connected. If a node does not receive any hello message from its neighbors during a certain time period, referred to as the lifetime of hello message, the node assumes that the link to the neighbor is currently disconnected and reports the link failure to the source corresponding to the link via a route error (RRER) message.

*2.3. Location Awareness and Performance Enhancement.* In a mobile ad hoc network, the location information of nodes helps to improve routing performance, such as packet delivery rate and overhead by estimating node mobility. In a route discovery operation, the route with the longest lifetime can be selected to reduce the number of transmission failures and the number of overheads to find a new route [13]. To reduce overhead messages, instead of a fixed period for hello message, the adaptive period is proposed using link lifetime to achieve high protocol efficiency in [14].

To predict mobility, the previous work proposed a location prediction scheme [12], which is defined as

$$\widehat{X}_i(t_k + \Delta t) = X_i'(t_k) + \vec{V}_i(t_k) \Delta t, \quad (2)$$

where  $\widehat{X}_i(t_k + \Delta t)$ ,  $X_i'(t_k)$ , and  $\vec{V}_i(t_k)$  are the predicted location of node  $i$  at time  $t_k + \Delta t$ , a measured location at time  $t_k$ , and a measured velocity at time  $t_k$ , respectively. If individual velocities of nodes are not available in (2), the nodes can approximately estimate their velocities using the previously stored location information [15] as follows. For

$t_k > t_{k-1}$ , the velocity of node  $i$  at time  $t_k$  is approximately expressed as

$$\vec{V}_i(t_k) \approx \frac{X'_i(t_k) - X'_i(t_{k-1})}{t_k - t_{k-1}}. \quad (3)$$

Based on the mobility prediction, nodes estimate link durations corresponding to adjacent nodes, and destination nodes choose the longest lifetime route among candidates. Since a link between two nodes is connected only if the distance between the two nodes is less than or equal to their transmission range, the estimated link duration  $LDT_{(i,j)}$  between nodes  $i$  and  $j$  is defined as

$$LDT_{(i,j)} = \max \Delta t \quad (4)$$

$$\text{subject to } \widehat{D}_{(i,j)}(t_k + \Delta t) \leq r, \quad (5)$$

where  $\widehat{D}_{(i,j)}(t_k + \Delta t)$  is the estimated distance between nodes  $i$  and  $j$  elapsed time  $\Delta t$  from current time  $t_k$ . A route consists of ordered links and is disconnected if one of the links is broken. Hence, the route expiration time  $RET_{(s,d)}^{(h)}$  of a route  $R_{(s,d)}^{(h)}$  between nodes  $s$  and  $d$  is expressed as

$$RET_{(s,d)}^{(h)} = \min_{(i,j) \in R_{(s,d)}^{(h)}} LDT_{(i,j)} \quad (6)$$

for  $h \in \mathcal{H}$ . The most reliable route can be chosen among candidate routes based on (6).

**2.4. Location Errors and Estimation Problem.** In practice, location errors inevitably exist in measurement. However, in previous work, mobility prediction used perfect location information receiving from the GPS devices or other techniques [16, 17]. The imperfect location information induces erroneous mobility estimate, which results in performance degradation.

For example, let  $X_i(t_k)$  and  $X'_i(t_k)$  be the real location and the measured location of node  $i$  at time  $t_k$ . Then, based on measured locations  $X'_i(t_k)$  and  $X'_j(t_k)$  of nodes  $i$  and  $j$ , respectively, after elapsed time  $\Delta t$  from time  $t_k$ , the estimated distance  $\widehat{D}'_{(i,j)}(t_k + \Delta t)$  between the two nodes is less than the transmission range  $r$  and the link between two nodes is considered *connected*, even though node  $j$  locates out of the transmission range of node  $i$ ; that is, the communication link between two nodes is *disconnected*, as shown in Figure 1. Hence, we propose a routing algorithm in the presence of location errors in measurement to mitigate the impact of imperfect location information.

### 3. Proposed Algorithm

In this section, we proposed an on-demand routing algorithm robust to location errors with mobility prediction. In MANETs, the mobility prediction plays a great role in predicting the link lifetime and the route lifetime, which can reduce overhead messages and improve routing performance [13]. However, as shown in Figure 1, location errors in measurement provide an incorrect mobility prediction, which

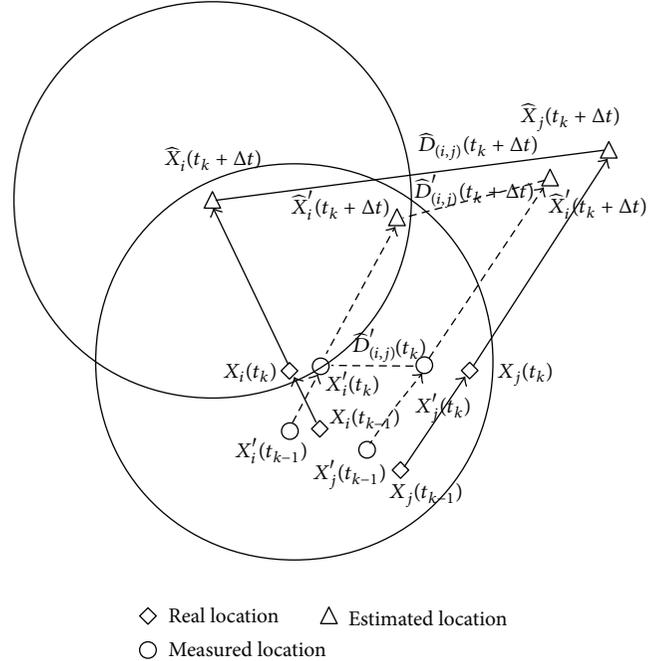


FIGURE 1: Estimated link duration.

induces wrong decision for routing. To mitigate the impact of such errors on mobility prediction and routing decision, we adopt two schemes: location error correction and route confidence.

**3.1. Location Correction and Mobility Prediction.** We employ the discrete Kalman filter, which is a set of recursive mathematical equations and supports the estimation of states in such way that minimizes the variance of estimation errors. The recent updates with previous measured location compensate current location for measurement errors. In this paper, the process errors are ignored and the main focus is the measurement errors. A detail of the discrete Kalman filter is presented in [18].

From (2), the current or future location depends on the previous location. The location errors are defined as the difference between the actual location and the measurement location. Let  $W_i$  be the location errors at node  $i$ , which is the additive noise; then, the measurement location of node  $i$  at time  $t_k$  can be expressed as  $X'_i(t_k) = X_i(t_k) + W_i(t_k)$ .

For each node  $i \in \mathcal{N}$ , let state matrix  $x$  be defined as  $x(t_k) = [X(t_k) \ \vec{V}(t_k)]^T$  with real location  $X$  and velocity  $\vec{V}$ ; then,  $x(t_k)$  denotes the actual state at time  $t_k$ . In the same way, we define the measurement state  $x'(t_k)$  at time  $t_k$  as  $x'(t_k) = [X'(t_k) \ \vec{V}'(t_k)]^T$ .

During time interval  $\Delta T$ , which is the elapsed time from the previous updated time  $t_{k-1}$  until current time  $t_k$ , that is,  $\Delta T = t_k - t_{k-1}$ , the node moves from  $X(t_{k-1})$  to  $X(t_k)$  such

that  $X(t_k) = X(t_{k-1}) + \Delta T \vec{V}(t_k)$ . Hence, the measured velocity  $\vec{V}'_i(t_k)$  is

$$\begin{aligned} \vec{V}'_i(t_k) &= \frac{X'_i(t_k) - X'_i(t_{k-1})}{\Delta T} \\ &= \vec{V}_i(t_k) + \frac{1}{\Delta T} W_i(t_k, t_{k-1}), \end{aligned} \quad (7)$$

where  $W_i(t_k, t_{k-1})$  is the sum of measurement errors at times  $t_k$  and  $t_{k-1}$ .

Suppose that during elapsed time  $\Delta T$  the velocity is constant; that is,  $\vec{V}(t_k) = \vec{V}(t_{k-1})$ . The actual state  $x(t_k)$  and measurement state  $x'(t_k)$  can be written as

$$\begin{aligned} x(t_k) &= \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix} x(t_{k-1}) \\ x'(t_k) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(t_k) + w(t_k), \end{aligned} \quad (8)$$

where  $w(t_k) = [W(t_k) \ (1/\Delta T) \ W(t_k, t_{k-1})]^T$ . Denote that matrix  $A(t_{k-1}) = \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix}$  and that matrix  $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ . The matrix  $A(t_{k-1})$  represents the state change and the matrix  $B$  describes the relation between the actual state and measurement state. The above equation can be rewritten as

$$\begin{aligned} x(t_k) &= A(t_{k-1}) x(t_{k-1}) \\ x'(t_k) &= Bx(t_k) + w(t_k). \end{aligned} \quad (9)$$

Since the actual state  $x(t_k)$  cannot directly be acquired, we define  $\hat{x}^-(t_k)$  as *a priori* estimate at time  $t_k$  for a given state prior to time  $t_k$ , and  $\hat{x}(t_k)$  as *a posteriori* estimate state at time  $t_k$  for a given measurement state  $x'(t_k)$ . Let  $P^-(t_k)$  and  $P(t_k)$

be *a priori* estimate error covariance and *a posteriori* estimate error covariance, respectively, and they can be expressed by

$$P^-(t_k) = E \left[ (x(t_k) - \hat{x}^-(t_k)) (x(t_k) - \hat{x}^-(t_k))^T \right] \quad (10)$$

$$P(t_k) = E \left[ (x(t_k) - \hat{x}(t_k)) (x(t_k) - \hat{x}(t_k))^T \right]. \quad (11)$$

To find the best estimate of the current state, we apply the Kalman filter. The operation of the Kalman filter includes two mechanisms: time update and measurement update. The time update process is responsible for predicting the current estimate state based on the previous state by computing  $\hat{x}^-(t_k)$  and  $P^-(t_k)$  as follows:

$$\hat{x}^-(t_k) = A_{t_{k-1}} \hat{x}(t_{k-1}), \quad (12)$$

$$P^-(t_k) = A(t_{k-1}) P(t_{k-1}) A^T(t_{k-1}).$$

After the time update operation, the measurement update corrects the measurement state as follows:

$$\begin{aligned} K(t) &= P^-(t_k) B^T (B P^-(t_k) B^T + R)^{-1} \\ \hat{x}(t_k) &= \hat{x}^-(t_k) + K(t_k) (x'(t_k) - B \hat{x}^-(t_k)) \end{aligned} \quad (13)$$

$$P(t_k) = (I - K(t_k) B) P^-(t_k),$$

where  $K(t_k)$  and  $R$  are the Kalman gain and the measurement error covariance, respectively. After that, the operation is repeated and the estimate state is measured based on the previous state and measurement state. Each node updates and tracks its current location based on periodically or eventually measured locations as the process of the discrete Kalman filter algorithm, which is summarized in Figure 2.

In implementation, the measurement error covariance  $R$  is measured prior to the operation of the Kalman filter. The measurement error covariance is determined by the variance of measurement noise by obtaining some off-line sample measurement [18]. The initial value for each state  $\hat{x}(t_0)$  is set to the measured information at the beginning.

In addition, we can obtain the confidence level of a link duration from the *a posteriori* estimate error covariance matrix  $P(t_k)$ . The *a posteriori* estimate error covariance matrix in (11) can be reexpressed as

$$P(t_k) = \begin{bmatrix} E[e_X^2(t_k)] & E \left[ \frac{1}{\Delta T} (e_X^2(t_k) - e_X(t_k) e_X(t_{k-1})) \right] \\ E \left[ \frac{1}{\Delta T} (e_X^2(t_k) - e_X(t_k) e_X(t_{k-1})) \right] & E \left[ \frac{1}{\Delta T^2} ((e_X(t_k) - e_X(t_{k-1})))^2 \right] \end{bmatrix}, \quad (14)$$

where  $e_X(t_k) \equiv X(t_k) - \hat{X}(t_k)$ . The square root of the expected square error  $E[e_X^2(t_k)]$  is equivalently considered as the standard deviation in the engineering community [19]. Hence, the root-mean-square error,  $\sqrt{E[e_X^2(t_k)]}$ , is equivalently the standard deviation of errors, and  $\sqrt{E[e_{X_i}^2(t_k) + e_{X_j}^2(t_k)]} / \vec{V}_{(i,j)}$ ,

denoted as  $\varepsilon$ , becomes the confidence level of link duration of link  $(i, j)$ .

**3.2. The Enhanced Mobility Prediction Routing Protocol.** In this subsection, we develop a mobility prediction-based routing protocol in the presence of location errors. Our goal

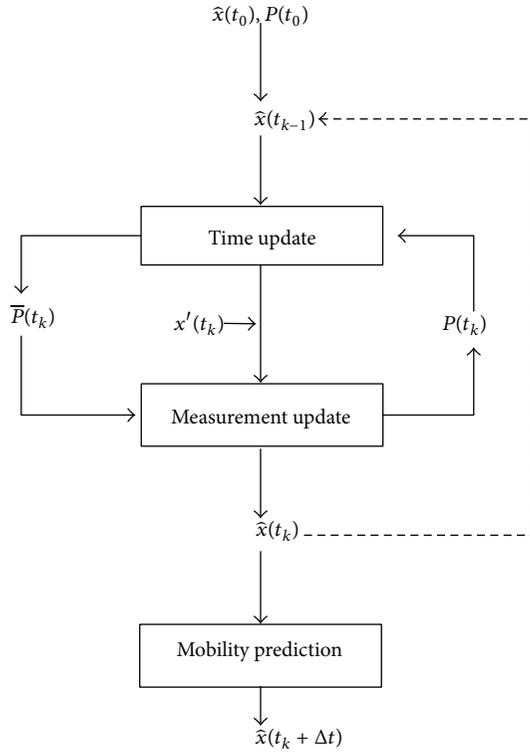


FIGURE 2: The Kalman filter based location correction process.

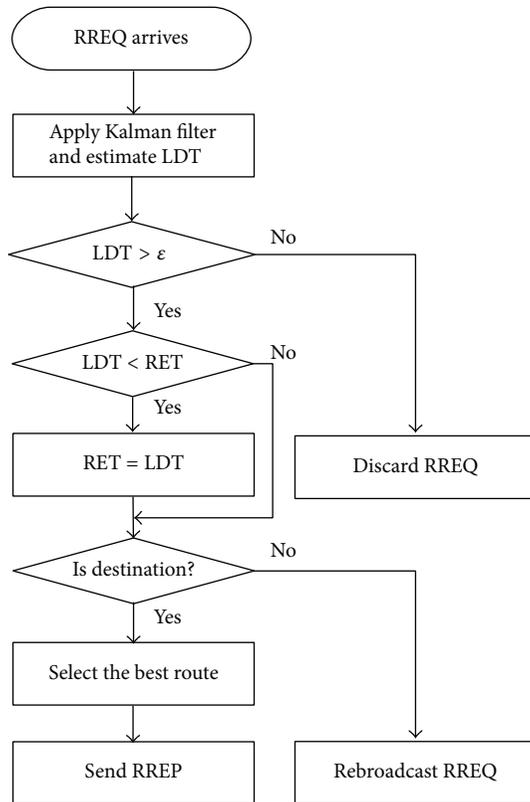


FIGURE 3: The Kalman filter based enhanced mobility prediction (EMP).

of mobility prediction is to find the longest RET and to avoid the risky link. The risky link that is defined as a link with vulnerable link duration time (LDT) seems to be dead or to be no longer alive in a short time after discovering.

When new data arrive at a node, the source node finds an active route associated with the corresponding destination in its routing table, as in Section 2.2. If no active route exists, the source node initiates route discovery to find a route to the destination node by broadcasting a RREQ message with recently updated location information and the standard deviation  $\sqrt{E[e_{\hat{x}}^2(t_k)]}$  of location error to neighboring nodes. The RET field and the hop count field in the RREQ message are initially set to infinity and one, respectively.

Upon reception of RREQ, a node computes the link duration time between the RREQ sender and itself, which implies the estimated lifetime of the link, from (5). To compute link durations in (5), nodes use the compensated location information  $\hat{x}(t_k)$  instead of the measured location information  $x'(t_k)$ . To exclude the risky link, the node compares LDT value with the confidence level  $\epsilon$  of LDT, which is computed from the standard deviations of the RREQ sender and itself. If the LDT value is less than  $\epsilon$ , the node discards the RREQ. Otherwise, the LDT value updates a RET value in the RREQ. If the LDT is smaller than the RET in the RREQ, the receiving node replaces the RET value by the new LDT. If the RREQ receiver is not the destination of the RREQ, the node broadcasts the receiving RREQ to other nodes after increasing the hop count by one until the RREQ reaches the destination.

In the case when a node is the destination of RREQ, the node waits for time interval  $T_w$  and collects RREQs whose destination is the node. After the time interval  $T_w$ , the destination chooses the longest route among the received routes and replies a RREP message after setting the lifetime field as the corresponding RET. RREP receivers relay the RREP message in a unicast manner until the RREP reaches the source, as described in Section 2.2. The details of proposed algorithm, AODV with enhanced mobility prediction (EMP), are described in Figure 3.

For route maintenance, we adopt the adaptive period for hello messages as in [14, 20], referred to as hello interval adjustment (HIA), to reduce the overheads instead of a fixed period. When receiving a RREQ from node  $i$ , node  $j$  estimates link duration  $LDT_{(i,j)}$  in Figure 3 and set the period for hello frequency to

$$\max \left\{ T_{\min}, \frac{\min_{i \in N_j} LDT_{(i,j)}}{\beta} \right\}, \quad (15)$$

where  $T_{\min}$  is the minimum value for the hello period,  $N_j$  is a set of the nodes that establish active links with node  $j$ , and  $\beta$  is a control parameter. The value of  $\beta$  is greater than or equal to 1, which aims to adjust the hello frequency.

#### 4. Performance Evaluation

We evaluate the performance of our proposed algorithms by using the network simulator NS-2 [21]. For simulations, there

TABLE 1: Parameter settings.

Parameter	Values
Network simulator	NS-2.34
Simulation area	2 km × 1.5 km
Number of mobile nodes	100
Simulation time	900 s
Mobility model	Random way point
Pause time	0 s
Packet generation rate	4 packets/s
Packet size	512 bytes
Transmission range	250 m

are 100 nodes initially distributed in an area of 2 km by 1.5 km and the transmission range of each node is set to 250 m. We run simulations with ten different random seeds and average the simulation results.

The random waypoint mobility (RWP) [22] model is used as a referenced mobility model, in which mobile nodes move from their current locations to new locations by randomly choosing directions and speeds. Upon arrival at a destination, after a pause time, they choose another random destination in the simulation area and travel toward the destinations with a uniformly distributed speed between the maximum speed and minimum speed. We set the pause time to zero to represent constant mobility.

The constant bit rate (CBR) traffic under the user datagram protocol (UDP) is used to accurately compare different routing protocols with a sending rate of 4 packets per second and 512 bytes of packet size. The parameter settings are listed in Table 1.

Two metrics are used for evaluating the network performance: the packet delivery rate and the normalized routing load. The packet delivery ratio is defined as the ratio of the number of generated packets to the number of packets received at the corresponding destinations. For the amount of overhead packets, we count the number of packets used for route discovery and route maintenance. For comparison, the total number of overhead packets is normalized by the number of packets successfully delivered to destinations.

To evaluate the performance improvement, our EMP routing protocol is compared with mobility prediction-based AODV routing protocol with route discovery mechanism [13] and the conventional AODV routing protocol in various noisy environments. For simplicity, the mobility prediction-based AODV routing protocol is denoted by the classic mobility prediction (MP). For simulations, we assume that the location errors of each node  $i$  are Gaussian random variables with zero mean and standard deviation  $\sigma_i$ .

Firstly, we compare the performance of our enhanced mobility prediction EMP with the previous work MP in the presence of location errors by varying the standard deviation of location errors from 3 m (1.12% of transmission range) to 50 m (20% of transmission range).

Secondly, we fix the standard deviation of location errors to 20 m (8% of transmission range) and show the network

performance under different impact of network environments, such as the impact of node velocity, traffic load, and node density. For each scenario, the HIA mechanism is enabled or disabled to show the impact of adaptive hello period.

*4.1. The Performance of the Kalman Filter Based Enhanced Mobility Prediction in the Presence of Location Errors.* To compare our EMP routing protocol with the MP routing protocol, ten source-destination pairs generate 4 packets per second during the simulation time. For mobility, each node follows the RWP mobility model with randomly selected speed between 1 m/s and 20 m/s.

Our proposal incorporates the Kalman filter to remove the location errors in order to reduce the impact of location errors and predicts the link duration time more accurately. The EMP can also improve the network performance by limiting the number of route discovery due to the dangerous link with an uncertain link duration time. The node establishing the uncertain link duration time does not allow forwarding the RREQ messages. Therefore, the discovered route becomes a better candidate for route selection and the number of overhead messages is significantly decreased.

In Figure 4(a), the packet delivery rates of EMP, MP, and AODV routing protocols are compared. As the standard deviation of location errors increases, the packet delivery rate of the MP routing protocol is decreased faster than that of EMP. When the standard deviation of location errors is behind a certain level (20 m in this case), the packet delivery rate of the MP routing protocol is lower than that of the AODV routing protocol. The large location errors lead to poor mobility prediction, which results in performance degradation. However, the packet delivery rate of our proposed routing protocol EMP outperforms MP and AODV routing protocols in all the cases and is robust to the location errors.

Figure 4(b) shows the normalized routing loads of EMP, MP, and AODV routing protocols. As the standard deviation of location errors increases, the normalized routing loads of MP and EMP increase due to inaccurate prediction. The normalized routing load of MP increases faster than that of EMP and is even greater than that of the conventional AODV. However, the EMP just slightly increases the routing overhead, which demonstrates that our proposed algorithm is robust to location errors.

Figures 5(a) and 5(b) show the packet delivery rate and the normalized routing load when the HIA is enabled for the mobility prediction-based routing protocol. The HIA mechanism is used for reducing the unnecessary hello messages. The AODV routing protocol sets the hello frequency to 1 second and the AODV-I sets the hello frequency to 20 seconds. As the location errors increase, the performance of MP is degraded. It is because the MP routing cannot estimate the true value of link duration that leads to incorrect route selection. Therefore, the selected route is unreliable and unstable so that the source node has to handle the route more frequently. When the standard deviation of location errors is larger than 40 m, the performance of the MP routing is

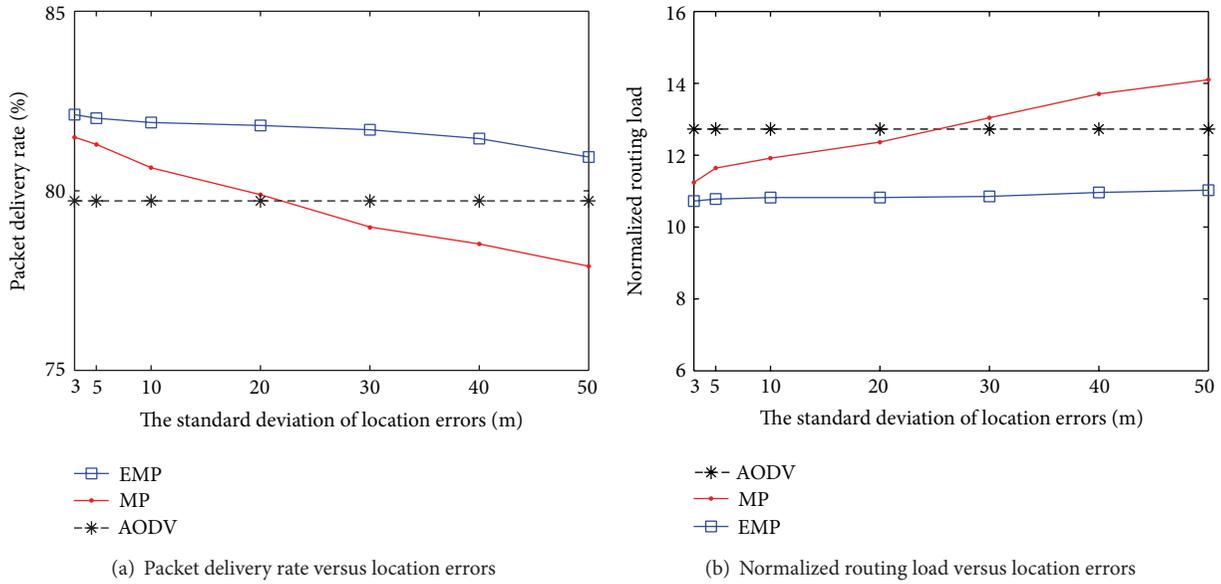


FIGURE 4: Impact of location errors—fixed hello interval.

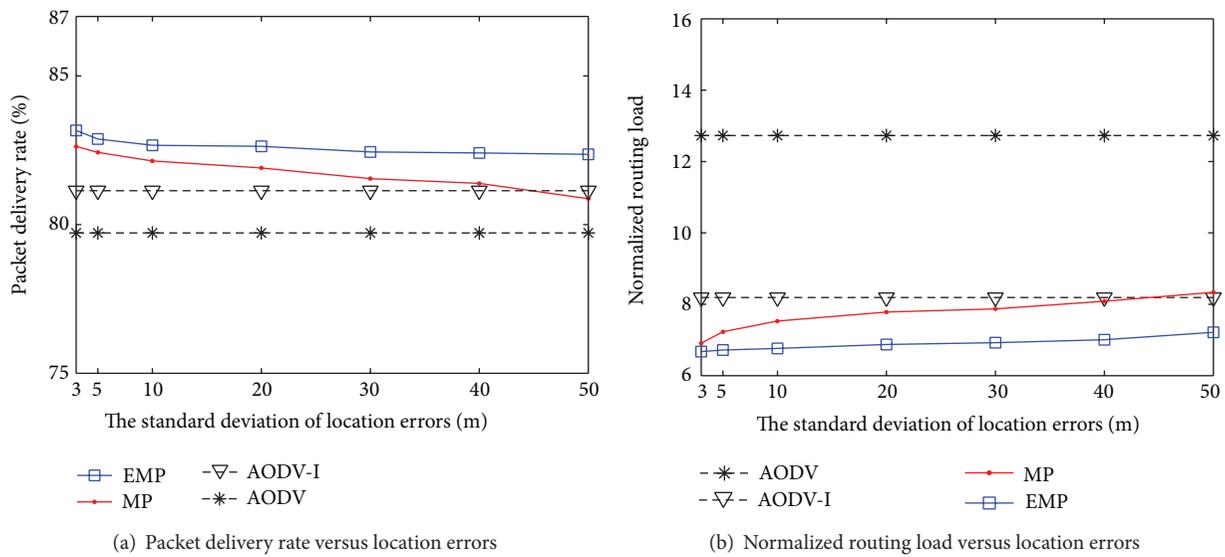


FIGURE 5: Impact of location errors—flexible hello interval.

lower than the AODV-I routing. The inaccurate link duration for selecting the route and setting the hello interval causes the performance degradation of mobility prediction-based routing protocol without location error compensation.

4.2. *The Impact of Node Velocity.* We study the impact of node velocity on routing performance in various network environments. The node mobility has a great impact on network performance [23, 24] since the change of topology leads to more exchanging messages in order to find and maintain new routes. During simulations, performances are compared in three different mobile environments: low mobility, medium mobility, and high mobility. For the low

mobility environment, we set the speed for RWP to 1 m/s, which is a pedestrian speed (3.6 km/h). We also set 10 m/s and 20 m/s (72 km/h) as the node speeds for the medium mobility and the high mobility environments, respectively.

Figure 6 shows that the packet delivery rate decreases as the node velocity increases since routes are more frequently broke and more overhead messages are necessary due to fast topology change, as shown in Figure 7. Whether hello interval for route maintenance is fixed or adaptive to mobility, AODV with mobility prediction is better than the conventional AODV in the presence of location errors, as shown in [12]. Our algorithm, which compensates for location errors, outperforms the others and is close to the case (EMP-wo) when location information is error-free. Therefore, our

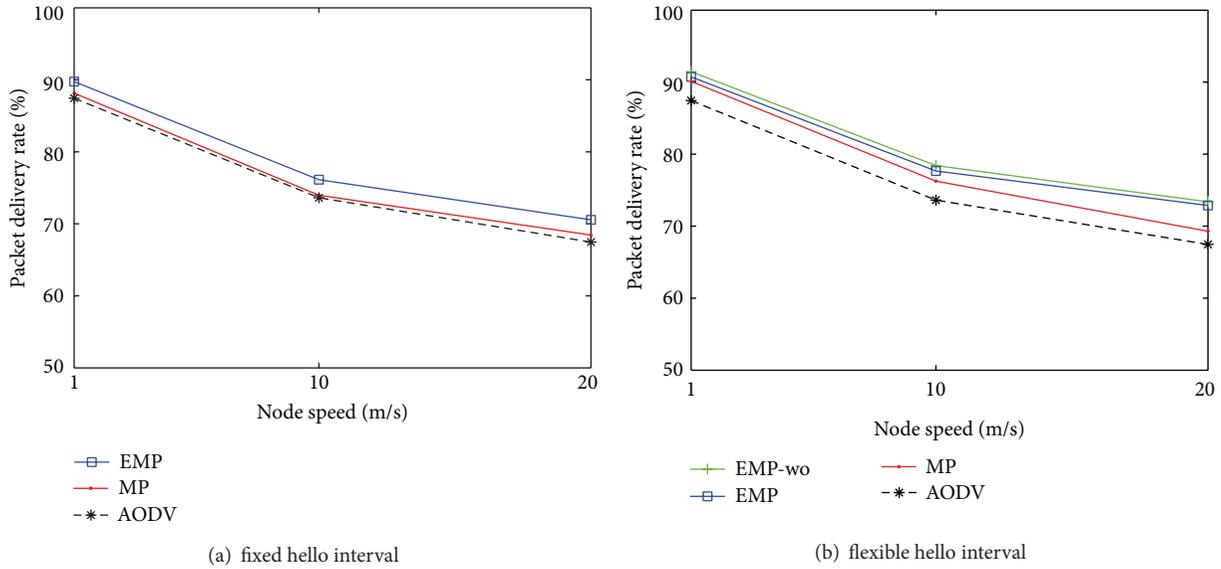


FIGURE 6: Packet delivery rate versus node velocity.

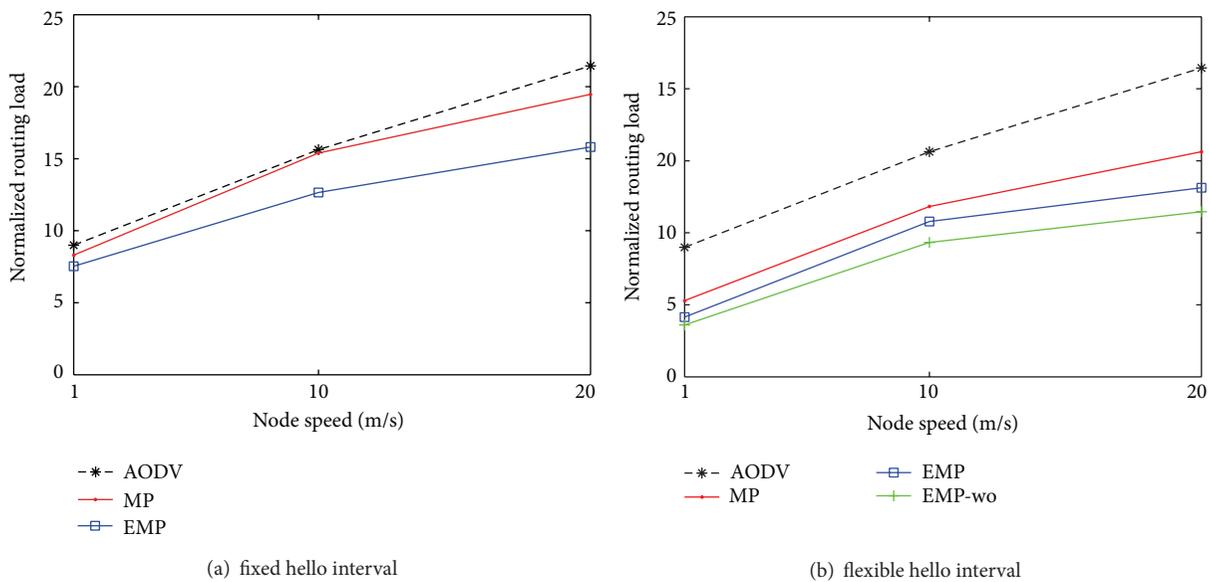


FIGURE 7: Normalized routing load versus node velocity.

proposed routing protocol EMP can adapt to the scalability network even in the presence of location errors.

4.3. *The Impact of Traffic Load.* Traffic load can affect the performance of routing protocols. To study the impact of traffic load, we vary the number of source-destination pairs to deliver generated data. For mobility, each node also follows the RWP mobility model with randomly selected speed between 1 m/s and 20 m/s.

Figure 8 shows the packet delivery rates. By increase of the number of source-destination pairs in the network, due to transmission collision and congestion, the packet delivery

rates are reduced. In Figures 8(a) and 8(b), our algorithm outperforms the others and is almost close to the EMP-wo, which assumes no location errors in measurement and is an upper bound of the performance. That means that our proposed algorithm EMP is robust to the location errors.

Figure 9 reports the normalized routing load when increasing the traffic load. In Figure 9(a), the HIA mechanism is disabled, the MP needs to exchange more routing messages caused by the location errors, whereas the EMP can reduce the amount of routing overhead as compared to the MP and the original AODV. When the HIA mechanism is enabled, a large number of hello messages are reduced, but the hello message still contributes well to the local connectivity

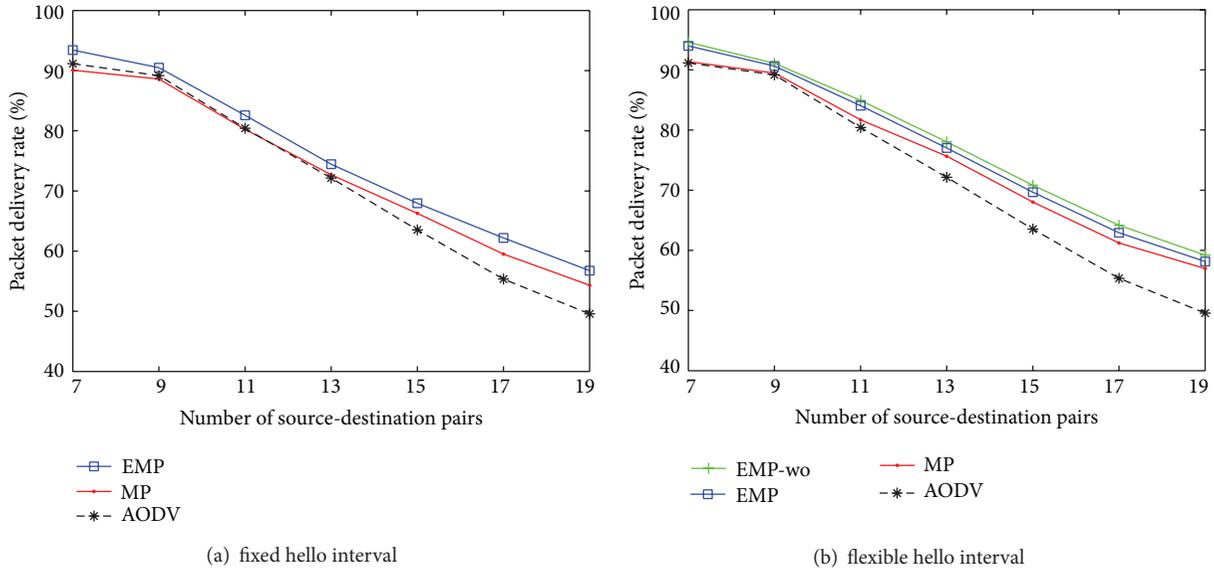


FIGURE 8: Packet delivery rate versus traffic load.

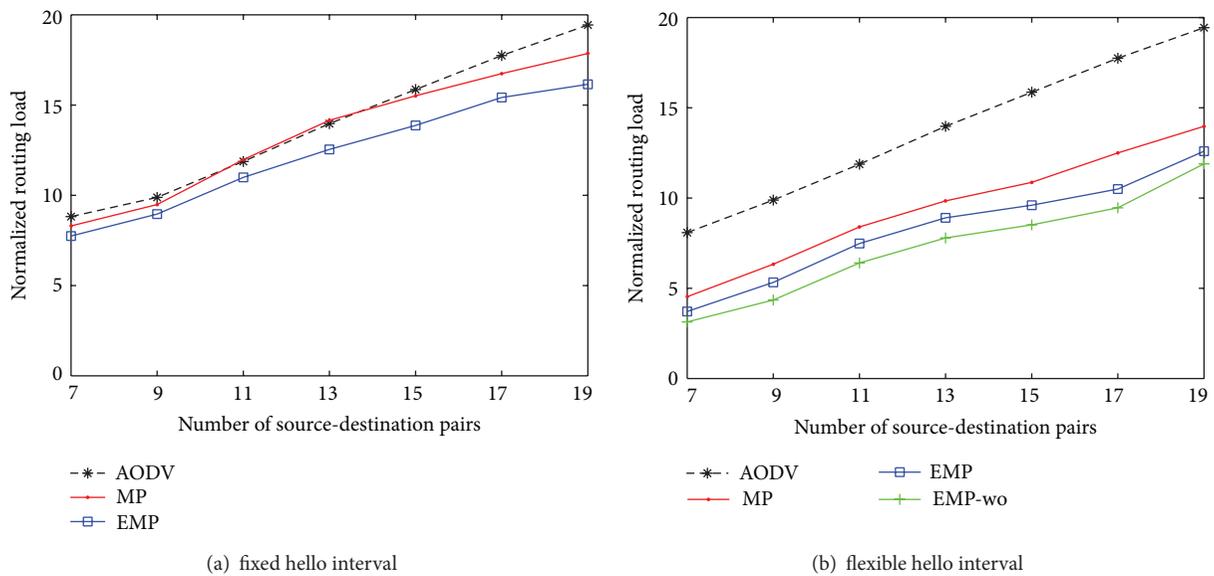


FIGURE 9: Normalized routing load versus traffic load.

management. The EMP and EMP-wo routing protocol can sharply reduce a great number of overheads as compared with the MP and the original AODV routing protocol.

4.4. *The Impact of Node Density.* In this subsection, we study the impact of node density on routing performance by varying the number of nodes from 75 nodes to 200 nodes as shown in Figures 10 and 11. If the number of nodes is too small, feasible routes between sources and destinations may not exist in the network so that the routing performance improves as the number of nodes increases in the network. However, above a certain number of nodes, the larger number of node hinders packet delivery due to larger overhead

messages required to maintain and discover routes. The EMP still outperforms the MP with respect to the packet delivery rate and the overhead in the presence of location errors.

## 5. Conclusion

This paper proposed an on-demand routing algorithm with enhanced mobility prediction that takes into account the location errors. Imperfect location information induces the performance degradation, but location errors in measurement were ignored in previous work. In the presence of location errors, we develop an on-demand routing algorithm collaborating to the Kalman filter to predict node mobility.

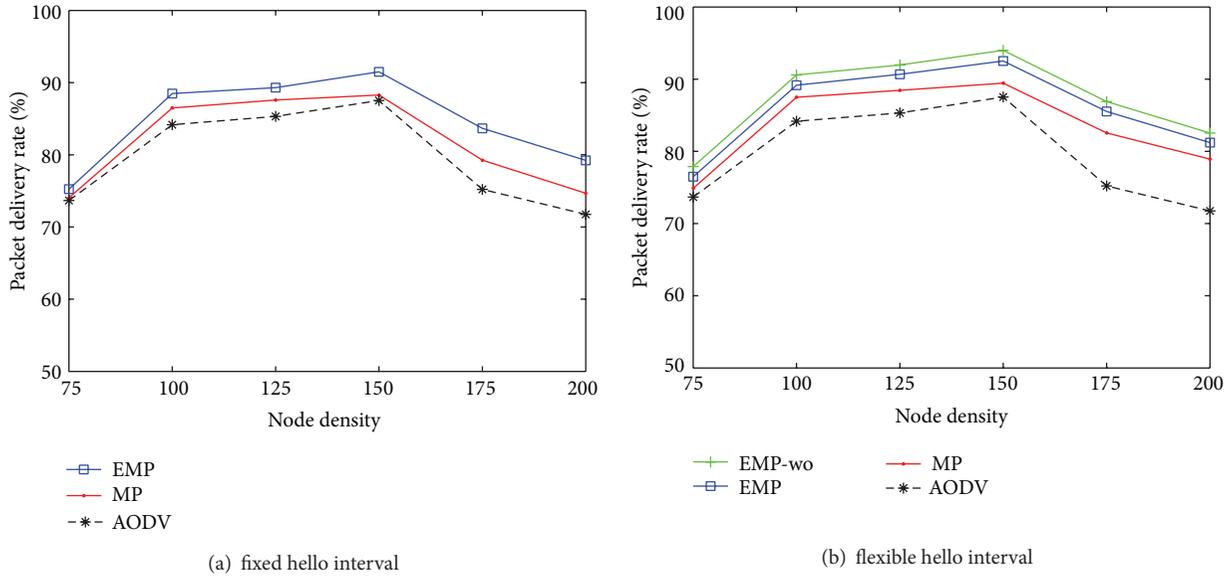


FIGURE 10: Packet delivery rate versus node density.

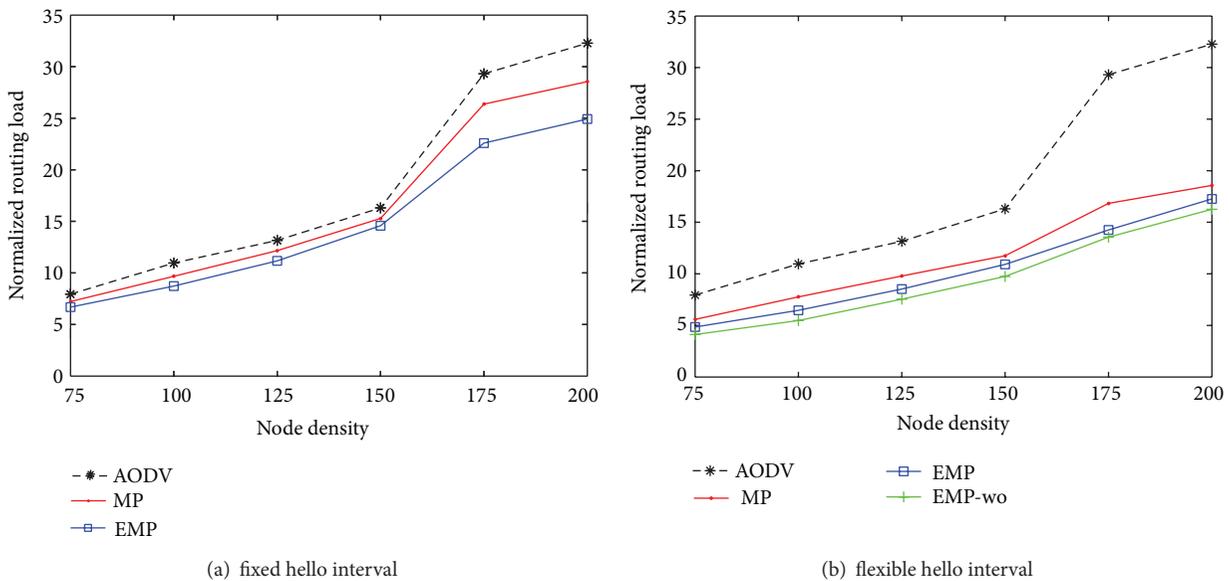


FIGURE 11: Normalized routing load versus node density.

Since the Kalman filter provides the root-mean-square error between the actual location and estimated location, the proposed algorithm excludes unreliable links considering the confidence levels of links. The estimated link duration adapts to the route maintenance period to reduce overheads. Via simulations, our proposed algorithm is robust to location errors and outperforms the previous algorithms.

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**Acknowledgment**

This work was supported by the 2013 Research Fund of University of Ulsan.

**References**

- [1] D. Ismail and M. Jaafar, "Mobile ad hoc network overview," in *Proceedings of the Asia-Pacific Conference on Applied Electromagnetics*, pp. 1-8, December 2007.
- [2] S. Corson and J. Macker, *Rfc 2501 routing protocol performance issues and evaluation considerations on mobile ad hoc networking*, 1999.

- [3] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.
- [4] F. Bai, N. Sadagopan, and A. Helmy, "A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 2, pp. 825–835, April 2003.
- [5] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, pp. 90–100, February 1999.
- [6] C. L. Tsao, Y. T. Wu, W. Liao, and J. C. Kuo, "Link duration of the random way point model in mobile ad hoc networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference*, vol. 1, pp. 367–371, April 2006.
- [7] S. Jiang, D. He, and J. Rao, "A prediction-based link availability estimation for routing metrics in MANETs," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1302–1312, 2005.
- [8] H. Luo and D. I. Laurenson, "Link-duration-oriented route lifetime computation for AODV in MANET," in *Proceedings of the International Conference on Wireless Communications and Signal Processing*, pp. 1–4, October 2010.
- [9] F. Bai, N. Sadagopan, B. Krishnamachari, and A. Helmy, "Modeling path duration distributions in MANETs and their impact on reactive routing protocols," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 7, pp. 1357–1373, 2004.
- [10] K. Namuduri and R. Pendse, "Analytical estimation of path duration in mobile ad hoc networks," *IEEE Sensors Journal*, vol. 12, no. 6, pp. 1828–1835, 2012.
- [11] R. J. La and Y. Han, "Distribution of path durations in mobile ad hoc networks and path selection," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 993–1006, 2007.
- [12] J. T. Hicks and J. Q. Walker II, "Mobility prediction and routing in ad hoc wireless networks," *International Journal of Network Management*, vol. 11, no. 1, pp. 3–30, 2001.
- [13] X. Hu, J. Wang, and C. Wang, "Mobility-adaptive routing for stable transmission in mobile ad hoc networks," *Journal of Communications*, vol. 6, no. 1, pp. 79–86, 2011.
- [14] L. Chao and H. Aiqun, "Reducing the message overhead of aodv by using link availability prediction," in *Proceedings of the 3rd International Conference on Mobile Ad-hoc and Sensor Networks (MSN '07)*, pp. 113–122, Springer, 2007.
- [15] S. Kwon and N. B. Shroff, "Geographic routing in the presence of location errors," *Computer Networks*, vol. 50, no. 15, pp. 2902–2917, 2006.
- [16] P. Misra and P. Enge, *Global Positioning System: Signals, Measurements and Performance*, Ganga-Jamuna Press, 2010.
- [17] N. Drawil, H. Amar, and O. Basir, "Gps localization accuracy classification: a context-based approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 262–273, 2013.
- [18] G. Welch and G. Bishop, *An Introduction to the Kalman Filter*, Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
- [19] E. W. Weisstein, Standard deviation, from MathWorld—a Wolfram Web Resource.
- [20] N. Hernandez-Cons, S. Kasahara, and Y. Takahashi, "Dynamic hello/timeout timer adjustment in routing protocols for reducing overhead in MANETs," *Computer Communications*, vol. 33, no. 15, pp. 1864–1878, 2010.
- [21] K. Fall and K. Varadhan, *The ns manual (formerly ns notes and documentation)*, 2005.
- [22] E. Hyttiä, H. Koskinen, P. Lassila, A. Penttinen, J. Roszik, and J. Virtamo, *Random Waypoint Model in Wireless Networks*, Networks and Algorithms: Complexity in Physics and Computer Science, Helsinki, Finland, 2005.
- [23] Y. T. Wu, W. Liao, C. L. Tsao, and T. N. Lin, "Impact of node mobility on link duration in multihop mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 5, pp. 2435–2442, 2009.
- [24] K. Amjad and A. J. Stocker, "Impact of node density and mobility on the performance of AODV and DSR in MANETS," in *Proceedings of the 7th International Symposium on Communication Systems, Networks and Digital Signal Processing*, pp. 61–65, July 2010.

## Research Article

# Effects of ADC Nonlinearity on the Spurious Dynamic Range Performance of Compressed Sensing

Rongzong Kang, Pengwu Tian, and Hongyi Yu

Zhengzhou Information Science and Technology Institute, Zhengzhou 450002, China

Correspondence should be addressed to Rongzong Kang; rongzongkang@outlook.com

Received 12 March 2014; Accepted 19 April 2014; Published 7 May 2014

Academic Editor: Zhongmei Zhou

Copyright © 2014 Rongzong Kang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Analog-to-information converter (AIC) plays an important role in the compressed sensing system; it has the potential to significantly extend the capabilities of conventional analog-to-digital converter. This paper evaluates the impact of AIC nonlinearity on the dynamic performance in practical compressed sensing system, which included the nonlinearity introduced by quantization as well as the circuit non-ideality. It presents intuitive yet quantitative insights into the harmonics of quantization output of AIC, and the effect of other AIC nonlinearity on the spurious dynamic range (SFDR) performance is also analyzed. The analysis and simulation results demonstrated that, compared with conventional ADC-based system, the measurement process decorrelates the input signal and the quantization error and alleviate the effect of other decorrelates of AIC, which results in a dramatic increase in spurious free dynamic range (SFDR).

## 1. Introduction

Traditional approaches to acquiring and sampling signal are based on Nyquist sampling theory, which states that the sampling rate must be at least twice the maximum frequency of the input signal. The increasing demand for ADC with both wider bandwidth and higher quantization bits seems to contradict with each other. The new theory of compressed sensing (CS) [1, 2] introduced an alternative data acquisition framework, which states that CS enables the acquisition and recover of sparse signals in some transform domains at a rate proportional to their information content that is much below the Nyquist rate.

Analog-to-information converter (AIC) is designed to acquire samples at a lower rate for compressed sensing system, and various architectures have been proposed of the recent work in this area, such as the random demodulator sampling architecture [3], the modulated-wideband converter [4], and others [5–8]. However, in the view of practical hardware implementation, the basic components constitute an AIC consists of mixer, integrator/low passed filter and ADC, and so forth. Among these components, the ADC commonly has the lowest dynamic range; an A/D converter's deviation from its ideal "linear" performance is

commonly characterized by the spurious-free dynamic range (SFDR) [9], which is defined as the difference in decibel, between the full-scale fundamental tone and the largest spurious harmonic component in the output spectrum. In order to make this notion precise, we will ignore the effects of any noise or nonlinearities from the other components except of ADC, since the SFDR of an AIC is typically dominated by the nonlinear process of ideal quantization and circuit-based (e.g., buffer, sample-and-hold) nonlinearities of ADC.

However there have been little literatures for characterizing and calculating the dynamic range performance of compressed sensing. In [10] a deterministic approach to dynamic range of a CS-based acquisition system is proposed, and the parameter of signal-to-quantization noise ratio is presented, whereas the dynamic parameter, that is, SFDR, is not considered. In [11] the quantization noise and dynamic range are considered for compressive imaging (CI) systems design and evaluate the quantization depth requirements for CI, while the quantization error and the SFDR performance of CS are still undiscussed. In [12] the impact of ADC nonlinearity in a mixed-signal CS system is studied, without considering the effect of ADC quantization error.

In this paper, we use an analytical approach couple with simulation results to formulate the SFDR performance of

a compressed sensing system when considered with the quantization and nonlinearity of ADC. The background of compressed sensing is introduced firstly; then the power spectrum of quantization noise of AIC is analyzed numerically and the SFDR of ideal AIC-based system is derived. Furthermore, a detailed analysis of the other ADC nonlinear effects in SFDR performance of AIC-based system is presented. Finally, the behavioral simulations results are presented that clearly verify the accuracy of the analysis.

## 2. Background

*2.1. Quantization-Limited SFDR of ADC-Based System.* Quantization changes a sine wave from a smooth function to a staircase signal; due to this nonlinear effect, the output signal is composed of a large number of nonlinear distortion products. The most important contribution to the output distortion comes from the quantization process, because this is an inherently nonlinear process. In an ideal quantizer, suppose that there is no nonlinearity and noise exists except for the nonlinearity due to quantization. In this case, the spurious signal of ADC output is only produced by the quantization. When a sine wave is passed through the ideal quantizer, the Fourier series of the output signal leads to the closed-form expression for the magnitudes of the harmonic as [13]:

$$A_p = \delta_{p,1}A + \sum_{m=1}^{\infty} \frac{2}{m\pi} J_p(2m\pi A), \quad (1)$$

where  $A_p$  is the output amplitude of the  $p$ th harmonic,  $\delta_{p,1}$  is the Kronecker delta function,  $A$  is the input amplitude, and  $J_p$  is the  $p$ th-order Bessel function of the first kind. Although the largest harmonic is always located roughly at  $2\pi A$  when the quantization levels are larger than 20, we consider the third harmonic as the largest and the power of the largest harmonic as a function of the number of bits. As a result, the quantization-limited SFDR performance of an ideal ADC-based system is approximated by [13]:

$$\text{SFDR} = 8.07b + 3.29 \text{ dB}. \quad (2)$$

*2.2. Nonlinear-Limited SFDR of ADC-Based System.* Besides the nonlinearity produced by quantization, the circuit imperfections such as capacitor mismatches and finite opamp DC gains are considered. These nonlinearities of ADC would also influence the SFDR performance of the system. The simplest form of a nonlinear system is the memoryless power series, which is based on normal polynomials:

$$z = \sum_{i=0}^L a_i y^i, \quad (3)$$

where  $z$  is the output signal,  $y$  is the input signal, and  $L$  is the order of the circuit nonlinearity. If the input signal is a single tone signal given by

$$y = A \cos(\omega t + \varphi_0), \quad (4)$$

then the amplitudes of the harmonic terms can be computed from (3). However in the case of analog circuits, the order

of a polynomial expression is mostly limited to third order, polynomial coefficients of the 4th order and higher, and the nonlinearity caused by saturation at full scale are neglected.

When substituting (4) into (3), we get

$$\begin{aligned} z &= \sum_{i=0}^L a_i y^i \\ &= a_1 [A \cos(\omega t + \varphi_0)] + a_2 [A \cos(\omega t + \varphi_0)]^2 \\ &\quad + a_3 [A \cos(\omega t + \varphi_0)]^3, \quad (5) \\ z &\cong a_1 \cos(\omega t + \varphi_0) + \frac{1}{2} a_2 A^2 [\cos(2\omega t + 2\varphi_0)] \\ &\quad + \frac{1}{4} a_3 A^2 [\cos(3\omega t + 3\varphi_0)]^2. \end{aligned}$$

The specific relation between polynomial coefficients and harmonic power can be expressed in general [14]. The same analysis can be done when the input signal is supposed to be two tone signals; it will cause the production of more terms, the specific terms harmonics, and intermodulation. Furthermore, the dynamic range performance of ADCs is specified in terms of one-tone and two-tone SFDR [15].

While in practice, tests which have been developed to measure the performance mostly rely on Fourier analysis using discrete Fourier transform (DFT) and the fast Fourier transform (FFT). The input analog signal is first sampled at Nyquist rate; the harmonics and intermodulation distortion are calculated through the input signal spectrum, which is estimated from the time-domain samples with nonlinear distortion via DFT. However the DFT-based method needs to avoid the leakage of the input frequency and the number of periods of the input waveform in the sample record should not be a nonprime integer submultiple of the record length, further the ADC needs to have a high resolution, which limits the maximum achievable sampling rate.

As an alternative solution to high-speed ADCs, AIC-based system enables high resolution at high frequencies while only using low frequency, sub-Nyquist ADCs [3–8]. In this work, we investigate the effect of nonlinearity induced by quantization and other circuit's nonidealities of ADC on the AIC-based system and examine the SFDR performance in the presence of these nonlinearities.

*2.3. Analog-to-Information Converter (AIC).* There have been many theoretical discussions on AIC system in the literature [3–8], in this work, the block diagram of a typical AIC implementation [3] called the random demodulator shown in Figure 1 is considered to compare with the conventional ADCs. In this architecture, the input signal  $x(t)$  is mixed by a different pseudorandom number  $p_c(t)$  waveform; then the mixer output is integrated over a time period of  $1/M$ . Finally, the integrator outputs are sampled and quantized, by a traditional integrate-and-dump ADC at  $M$  Hz.

Note that this AIC architecture employs sub-Nyquist rate ADCs, and the input signal is mixed with the PN sequence and sent to integrator before sampling. As a result, the spectrum of the signal sent to the ADC is relatively flat

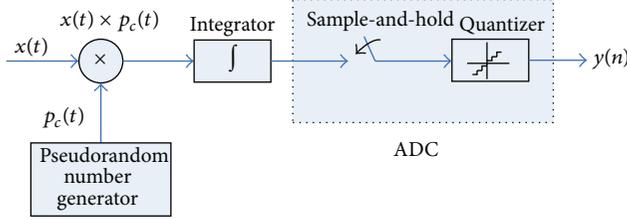


FIGURE 1: Block diagram of the random demodulator.

within the filter pass band, and then the harmonic and intermodulation energy due to the nonlinearity of ADC is spread along the signal bandwidth rather than concentrate on a few tones, which can lead to a better SFDR performance after reconstruction. In the following section we present our framework for investigating the impacts of nonlinearity caused by quantization and other circuits induced on the SFDR performance of AIC-based system.

### 3. SFDR Performance of AIC-Based System

**3.1. Quantization-Limited SFDR of AIC-Based System.** In an ideal case the dynamic range performance is mainly limited by quantization error; the spectra of the AIC quantization output is analyzed in this section.

As we know, the time-domain expression of the measuring process of AIC is given by

$$y_i = \langle x, \phi_i \rangle = \sum_{j=1}^N \phi_{i,j} x_j, \quad (6)$$

where  $\phi_{ij}$  is the element of measurement matrix and  $x_j$  is the element of the input signal. Suppose that the measurement matrix is sub-Gaussian random matrix; then the element  $\phi_{ij}$  is independent centered sub-Gaussian random variables with variance  $1/M$ , given  $S_{i,j} = \phi_{i,j} x_j$ ; then

$$y_i = \sum_{j=1}^N S_{ij}. \quad (7)$$

Then we can get the mean and variance of  $S_{i,j}$ :

$$\begin{aligned} E[S_{i,j}] &= E[\phi_{i,j} x_j] = x_j E[\phi_{i,j}] = 0, \\ D[S_{i,j}] &= E[S_{i,j}^2] = x_j^2 E[\phi_{i,j}^2] = \frac{x_j^2}{M}. \end{aligned} \quad (8)$$

According to the central limit theorem, when  $N \rightarrow \infty$ , the  $y_i$  subject to Gaussian distribution with mean 0 and variance  $\sum_{j=1}^N (x_j^2/M) = \|X\|_2^2/M$ .

As we know that when the input signal subjected to Gaussian distribution with mean 0, then the relation between autocorrelation function  $R_e(m)$  of quantization error and input signal can be expressed as follows:

$$R_e(m) = \frac{\Delta^2}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \exp\left[-4\pi^2 \frac{\sigma^2}{\Delta^2} k^2 (1 - r_y(m))\right], \quad (9)$$

where  $\Delta$  is the quantization step size and  $\sigma^2$  is the variance of the input signal.  $r_y(m) = R_y(m)/R_y(0)$  represents the normalized autocorrelation function. While the autocorrelation function of the measurement value can be expressed as

$$\begin{aligned} R_y(m) &= E[y_i y_{i+m}] \\ &= E\left[\sum_{j=1}^N \phi_{i,j} x_j \sum_{k=1}^N \phi_{i+m,k} x_k\right] \\ &= E\left[\sum_{j=1}^N \sum_{k=1}^N \phi_{i,j} \phi_{i+m,k} x_j x_k\right] \\ &= \sum_{j=1}^N \sum_{k=1}^N E[\phi_{i,j} \phi_{i+m,k}] x_j x_k. \end{aligned} \quad (10)$$

Because the element of the measurement matrix is independent, then  $R_y(0) = \|X\|_2^2/M$ , when  $j = k$  and  $m = 0$ , for others  $R_y(m)$  equal to 0, so normalized autocorrelation function is

$$r_y(m) = \begin{cases} 1, & m = 0, \\ 0, & \text{else.} \end{cases} \quad (11)$$

So, the autocorrelation function  $R_e(m)$  of quantization error is

$$R_e(m) = \begin{cases} \frac{\Delta^2}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2}, & m = 0, \\ \frac{\Delta^2}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} \exp\left[-4\pi^2 \frac{\sigma^2}{\Delta^2} k^2\right], & \text{else,} \end{cases} \quad (12)$$

where  $\sigma/\Delta \geq 1$ , and when  $\sigma/\Delta = 1$  and  $m \neq 0$ ,

$$R_e(m) = \frac{\Delta^2}{2\pi^2} \left[ \frac{e^{-4\pi^2}}{1} + \frac{e^{-16\pi^2}}{4} + \frac{e^{-36\pi^2}}{9} + \dots \right]. \quad (13)$$

For  $e^{-4\pi^2} \approx 7 \times 10^{-18}$ ,  $e^{-16\pi^2} \approx 2 \times 10^{-69}$ , we get  $R_e(m) \approx 0$ , when  $m \neq 0$ , and  $R_e(0) = (\Delta^2/2\pi^2) \sum_{k=1}^{\infty} (1/k^2) = (\Delta^2/12)$ .

From the above analysis, we know that  $R_e(m)$  is approximated to  $\delta$  function, and according to the Fourier transform relationship between power spectrum and autocorrelation function, the power spectrum of quantization noise is white noise spectrum. As a result, the spurious energy due to the quantization effect of ADC is spread to the whole bandwidth, and we can get a better SFDR performance of AIC-based system compared with the conventional ADC-based system.

**3.2. Nonlinear-Limited SFDR of AIC-Based System.** Compared with the analysis of the conventional ADC-based system, in AIC-based system, the input signal goes through random projection, filtering, and sampling.

A signal  $x$  can be viewed as a  $N \times 1$  column vector in  $\mathbb{R}^N$  with elements  $x[n]$ ,  $n = 1, 2, \dots, N$ . Let the

matrix  $\Psi = [\psi_1, \psi_2, \dots, \psi_N]$  have columns which form a basis of vectors in  $\mathbb{R}^N$ . And then, any signal  $x$  can be expressed as

$$x = \sum_{i=1}^N s_i \psi_i \quad \text{or} \quad x = \Psi s, \quad (14)$$

where  $s$  is the  $N \times 1$  column vector of weighting coefficients  $s_i = \langle x, \psi_i \rangle$ .

Consider a generalized linear measurement process of a signal  $x$  which is  $K$ -sparse. When we say that  $x$  is  $K$ -sparse, we mean that it is well reconstructed or approximated by a linear combination of just  $K$  basis vectors from  $\Psi$ , with  $K \ll N$ . That is, there are only  $K$  of the  $s_i$  in (1) that are nonzero and  $(N - K)$  are zero. Let  $\Phi$  be an  $M \times N$  measurement matrix,  $M \ll N$  where the rows of  $\Phi$  are incoherent with the columns of  $\Psi$ . The incoherent measurements can be obtained by computing  $M$  inner products between  $x$  and the rows of  $\Phi$  as in  $y_j = \langle x, \phi_j \rangle$ . It can also be expressed as

$$y = \Phi x = \Phi \Psi s = \Theta s, \quad (15)$$

where  $\Theta = \Phi \Psi$  is a  $M \times N$  matrix. It is proved that  $\Phi$  does not depend on the signal  $x$  and it can be constructed as a random matrix such as Gaussian matrix.

Then according to the nonlinearity model of ADC, we substitute the transform-domain samples into (3), and then we can get the measurement output of the AIC with nonlinear effect as follows:

$$z = \sum_{i=0}^L a_i y^i = a_1 [\Phi \Psi s] + a_2 [\Phi \Psi s]^2 + a_3 [\Phi \Psi s]^3. \quad (16)$$

**3.3. Reconstruction of Frequency Sparse Signal.** After quantization and sampling of ADC, we get the measurement in discrete values, in order to evaluate the SFDR performance of the AIC-based system, we need to compute the spectrum of the reconstruction signal. So, in this section, we frame the reconstruction problem for the AIC-based system with the nonlinearity effect.

Furthermore, the spectrum of the input signal is estimated from the measurement value  $z$  with nonlinear distortion via solving the following optimization problem:

$$\hat{s} = \operatorname{argmin} \|s\|_1 \quad \text{s.t.} \quad \|y - \Phi \Psi s\|_2 \leq \varepsilon_n + \varepsilon_d, \quad (17)$$

where  $\varepsilon_n$  is the error due to the noise and  $\varepsilon_d$  is the error due to the nonlinear distortion.

Up to now, there are many mature algorithms to resolve this convex optimization problem, including interior-point algorithms [16, 17], gradient projection [18], iterative thresholding [19, 20], and greedy approaches such as orthogonal matching pursuit (OMP) [21, 22]. Here we use the algorithm of basis pursuit with denoising [23] to resolve the reconstruction problem for evaluation of SFDR performance of AIC-based system.

## 4. Simulation Results

Figure 2 shows the SFDR performance of conventional ADC-based system and AIC-based system of ideal ADC with

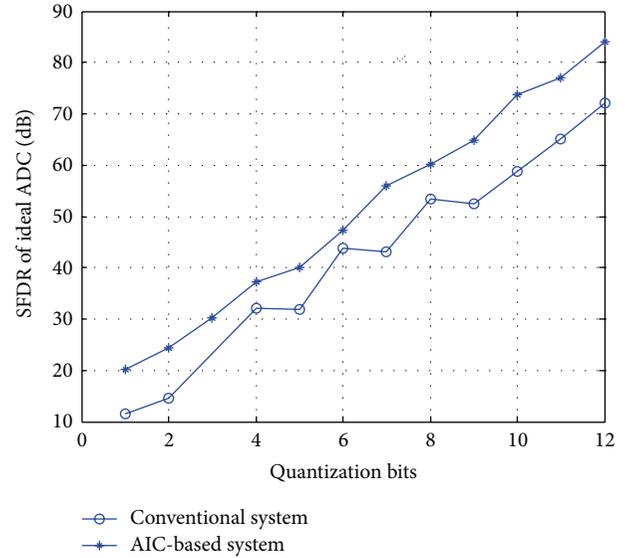


FIGURE 2: SFDR performance of AIC-based system and ADC-based with ideal ADC.

a single sinusoidal input for different quantization bits.  $\Phi$  is set to an  $M \times N$  Gaussian random measurement;  $M = 256$ , and  $N = 1024$ . The input signal frequency is  $f_0 = 64$  Hz,  $f_s = (5 * 64 - 1)$  Hz, and use BPDN [23] as the reconstruction algorithm. Every measurement was repeated 300 times to test the reproducibility.

As shown in Figure 2, the SFDR performance of AIC-based system outperforms that of conventional ADC-based system. That is because in the conventional ADC-based system, noise spectrum of sinusoid signals consists of discrete components, and the harmonic is concentrated in the odd multiple of its fundamental frequency, while in the AIC-based system the spectrum of quantization error is uniformly distributed. However the total quantization noise power represented by the area under the noise spectrum is approximately equal to  $\Delta^2/12$ , for AIC-based system the spurious energy is spread along the whole signal bandwidth; then each harmonic of the quantization error is thereby pulled downward into a more dense portion of the noise spectrum leading to increasing in SFDR performance. The observation from this simulation was intuitively illustrated in Figure 2.

Figure 3 shows a snapshot of the single-tone reconstructed error spectrum for conventional system and CS-based system. The second-order  $a_2 = 0.1$  and third-order distortion coefficients  $a_3 = 0.1$ . As we can see, in the conventional ADC-based system, the spurious harmonic due to the ADC nonlinearity concentrates on the multiple of fundamental frequency, whereas, in the CS-based system, the spurious energy is spread along the whole signal bandwidth. Meanwhile the amplitude of the spurious harmonic of AIC-based is lower than that of ADC-based system.

Figures 4 and 5 show the SFDR performance of ADC-based system and AIC-based system for two-tone input with quantization and different nonlinear distortion coefficients.  $f_1 = 16$  Hz,  $f_2 = 256$  Hz,  $f_s = 1024$  Hz, and

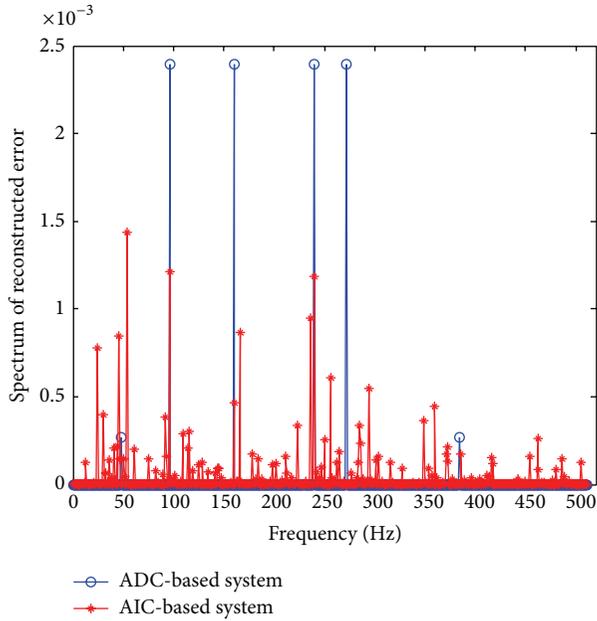


FIGURE 3: Spectrum of reconstruction error comparison between the conventional ADC-based system and the AIC-based system.

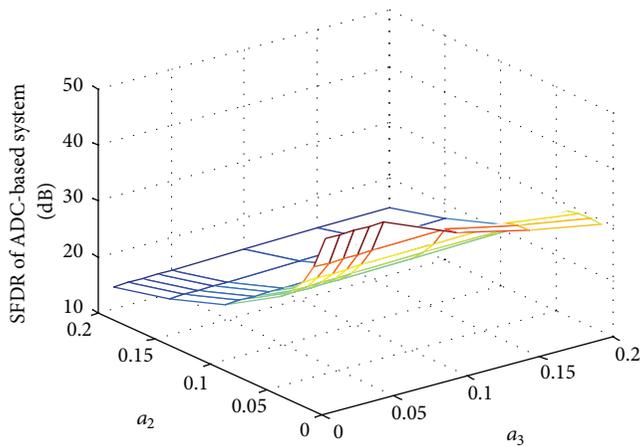


FIGURE 4: SFDR performance of conventional ADC-based system with other nonlinear effects.

quantization bits  $N = 4$ . The reconstruction algorithm and other simulation conditions are set the same as in Figure 2.

As we can see, both of the SFDR performances decrease when nonlinear distortion becomes large with the second- and third-order distortion coefficients increase. The simulation results also indicate that the second-order distortion influences the SFDR performance more seriously than that of the third-order distortion.

Comparing the results of Figure 4 with Figure 5, we can see that the SFDR performance of AIC-based system outperforms that of the ADC-based system when introducing the nonlinearity with both the quantization and circuit nonideality. This is because the randomization in AIC-based system changes the distribution of the error power from ADC nonlinear distortion; the signals sent to ADCs in

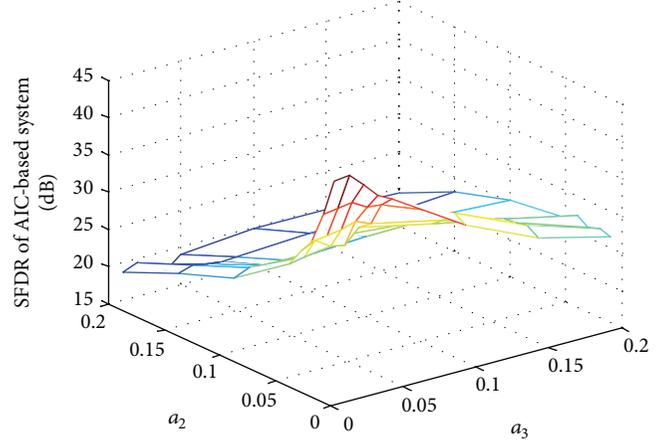


FIGURE 5: SFDR performance of for AIC-based system with other nonlinear effects.

the conventional Nyquist sampling architecture are original sinusoid signals, whereas those in the AIC-based system have relatively flat spectrum. As a result, by spreading the spurious energy along the signal bandwidth, the CS randomization relaxes the requirement on the ADC SFDR specification.

### 5. Conclusions

In this paper, we compare the SFDR performance of AIC-based system and conventional ADC-based system when considering both nonlinearity due to quantization and other circuit nonideality of ADC. We demonstrate that the quantization noise of AIC is spectrally white and uniformly distributed, and the quantization harmonics of AIC-based system is spread to the whole bandwidth, which means an improvement of SFDR performance. We show that AIC-based systems are less sensitive to the nonlinearity of ADC because of the CS randomization, which provides improvement of SFDR performance compared with conventional ADC-based system. Our results suggest that the second- and third-order distortion coefficients and quantization bits are the main factors that affect the SFDR performance of compressed sensing. The results presented in this paper can also be easily extended to the case when the signals input to AIC are multisinusoids.

### Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

### References

- [1] E. J. Candès and J. Romberg, “Quantitative robust uncertainty principles and optimally sparse decompositions,” *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 227–254, 2006.
- [2] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [3] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, “Beyond Nyquist: efficient sampling of sparse

- bandlimited signals,” *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 520–544, 2010.
- [4] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, “Xampling: analog to digital at sub-Nyquist rates,” *IET Circuits, Devices and Systems*, vol. 5, no. 1, pp. 8–20, 2011.
- [5] J. N. Laska, S. Kirolos, M. F. Duarte, T. S. Ragheb, R. G. Baraniuk, and Y. Massoud, “Theory and implementation of an analog-to-information converter using random demodulation,” in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '07)*, pp. 1959–1962, May 2007.
- [6] S. Pfetsch, T. Ragheb, J. Laska et al., “On the feasibility of hardware implementation of sub-Nyquist random-sampling based analog-to-information conversion,” in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '08)*, pp. 1480–1483, May 2008.
- [7] S. Kirolos, J. Laska, M. Wakin et al., “Analog-to-information conversion via random demodulation,” in *Proceedings of the IEEE Dallas ICAS Workshop on Design, Applications, Integration and Software (DCAS '06)*, pp. 71–74, October 2006.
- [8] J. Yoo, S. Becker, and A. Emami-Neyestanak, “Design and implementation of a fully integrated compressed-sensing signal acquisition system,” in *Proceedings of the IEEE International Conference on Acoustics, Speech Signal Process*, pp. 5325–5328, 2012.
- [9] J. Goodman, B. Miller, M. Herman, M. Vai, and P. Monticciolo, “Extending the dynamic range of RF receivers using nonlinear equalization,” in *Proceedings of the IEEE International Waveform Diversity and Design Conference (WDD '09)*, pp. 224–228, February 2009.
- [10] M. A. Davenport, J. N. Laska, J. R. Treichler et al., “The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range,” *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4628–4642, 2012.
- [11] A. Stern, Y. Zeltzer, and Y. Rivenson, “Quantization error and dynamic range considerations for compressive imaging systems design,” *Journal of the Optical Society of America A*, vol. 30, no. 6, pp. 1069–1077, 2013.
- [12] Z. Yu, J. Zhou, M. Ramirez, S. Hoyos, and B. M. Sadler, “The impact of ADC nonlinearity in a mixed-signal compressive sensing system for frequency-domain sparse signals,” *Physical Communication*, vol. 5, no. 2, pp. 196–207, 2012.
- [13] M. S. Oude Alink, A. B. J. Kokkeler, E. A. M. Klumperink, K. C. Rovers, G. J. M. Smit, and B. Nauta, “Spurious-free dynamic range of a uniform quantizer,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 56, no. 6, pp. 434–438, 2009.
- [14] N. Björnsell and P. Händel, “Achievable ADC performance by postcorrection utilizing dynamic modeling of the integral nonlinearity,” *Eurasip Journal on Advances in Signal Processing*, vol. 2008, Article ID 497187, 2008.
- [15] W. Kester, “Intermodulation Distortion Considerations for ADCs,” Analog Devices Tutorial MT-012, 2009.
- [16] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [17] D. E. N. van Ewout Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.
- [18] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.
- [19] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [20] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.
- [21] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” in *Proceedings of the 27th Asilomar Conference on Signals, Systems & Computers*, pp. 40–44, November 1993.
- [22] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [23] M. Friedlander and E. van den Berg, “Toolbox SPGL1 [EB/OL],” 2011, <http://www.cs.ubc.ca/labs/scl/spgl1>.

## Research Article

# Modeling and Analysis of Mobility Management in Mobile Communication Networks

Woon Min Baek,<sup>1</sup> Ji Hyun Yoon,<sup>2</sup> and Chesoong Kim<sup>3</sup>

<sup>1</sup> Kongju National University High School, Kongju, Chungnam 314-801, Republic of Korea

<sup>2</sup> Korean Minjok Leadership Academy, Hoengseong, Kangwon 225-823, Republic of Korea

<sup>3</sup> Sangji University, Wonju, Kangwon 220-702, Republic of Korea

Correspondence should be addressed to Chesoong Kim; [dowoo@sangji.ac.kr](mailto:dowoo@sangji.ac.kr)

Received 7 February 2014; Accepted 12 March 2014; Published 5 May 2014

Academic Editors: Y. Mao and Z. Zhou

Copyright © 2014 Woon Min Baek et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Many strategies have been proposed to reduce the mobility management cost in mobile communication networks. This paper studies the zone-based registration methods that have been adopted by most mobile communication networks. We focus on two special zone-based registration methods, called two-zone registration (2Z) and two-zone registration with implicit registration by outgoing calls (2Zi). We provide a new mathematical model to analyze the exact performance of 2Z and 2Zi. We also present various numerical results, to compare the performance of 2Zi with those of 2Z and one-zone registration (1Z), and show that 2Zi is superior to 2Z as well as 1Z in most cases.

## 1. Introduction

The number of mobile subscribers has been increasing, and the more accelerated growth of smart phone subscribers is expected with 4G networks. In a mobile communication network, since a mobile (mobile phone) is continually moving due to its basic characteristic, mobility management of the mobile is essential, to provide communication services with high quality.

One of the most important issues in mobility management is the location tracking. The location of a mobile must be maintained to connect an incoming call to the mobile if required. Location registration and paging are two basic functions to locate a mobile. Location registration is the series of processes to register a mobile's new location information in the system database, and paging is the series of processes to page the mobile in current location, to find mobile's exact cell and connect an incoming call, when an incoming call arrives. Since there is a tradeoff between location registration cost and paging cost, it is essential to analyze location registration cost and paging cost, in order to find the optimal location tracking method.

Various location registration methods have been proposed for mobile communication networks [1–9]. However,

the most important location registration method is zone-based registration [5–8], since it is adopted by most mobile communication networks.

In this study, zone-based registration is considered. In zone-based registration, each mobile has a *zone\_list*, where the visited zone is stored. If a mobile moves to a new zone, which is not in its current *zone\_list*, the new zone is stored in the *zone\_list*, and the mobile registers its new location information in the system database. A mobile may have more than one zone in zone-based registration [5], but most mobile communication networks adopt one-zone registration (1Z) because of ease of operation.

Lin [6] suggested a precise mathematical model, to analyze the performance of the case where a mobile has two zones, and compared the performance of the two cases, in which the number of zones is one and two. In addition, Jang et al. [7] considered implicit registration effects by outgoing calls, to improve the performance of the case where a mobile has two zones. However, Jang et al. [7] assumed a special mobility model and provided just a rough approximation of the performance.

Even Lin [6] provided a precise mathematical model for the performance of the case where a mobile has two zones, but his model is too complex to be applied to Jang et al.'s study [7],

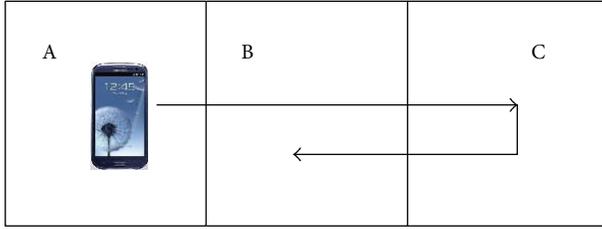


FIGURE 1: Movement of a mobile.

which considers implicit registration by outgoing calls under two registered zones, to improve the performance.

In this study, we derive a new mathematical model to analyze the exact performance of two-zone registration (2Z) and two-zone registration with implicit registration by outgoing calls (2Zi). Section 2 briefly describes general zone-based registration methods. Section 3 describes the mathematical model that we derived, to analyze the performance of 2Z and 2Zi. Section 4 presents the computational results of the signaling cost on radio channels, using our model. Section 5 summarizes the results and suggests future research directions.

## 2. Zone-Based Registration and 2Z

In zone-based registration, whenever a mobile moves to a new zone that is not in its current *zone\_list*, this new zone is stored in its *zone\_list*, and the mobile registers new location information in the system database. If a mobile can have only one registered zone (i.e., one-zone registration, 1Z), the mobile stores the newly entered zone in its *zone\_list* every time it moves from one zone to another. Thus, the system knows the zone in which the mobile is located, and the paging process always meets with success for incoming calls. Figure 1 shows the movement of a mobile.

However, if the mobile can have two registered zones (i.e., two-zone registration, 2Z), the system sometimes does not know the exact zone in which the mobile is located. For example, two zones, A and B, are stored in the *zone\_list*, and the mobile is currently in zone B. Let the left-hand zone denote the most recently registered zone in the following *zone\_list* and *system\_DB*:  $\boxed{\text{zone\_list} \mid B \mid A}$   $\boxed{\text{system\_DB} \mid B \mid A}$

When the mobile enters a new zone, C, then the *zone\_list* and *system\_DB* are changed as follows:  $\boxed{\text{zone\_list} \mid C \mid B}$   $\boxed{\text{system\_DB} \mid C \mid B}$

If an incoming call arrives in this situation, then the system pages the mobile in zone C and this paging succeeds (this is referred to as *zone hit*).

Consider another case. If the mobile reenters zone B, then *zone\_list* is changed as below, but location registration does not occur, because zone B is already stored in the current *zone\_list*:  $\boxed{\text{zone\_list} \mid B \mid C}$   $\boxed{\text{system\_DB} \mid C \mid B}$

In this case, the system does not know the correct zone (B) in which the mobile is located, and the paging process

is somewhat complicated. If an incoming call arrives, the system pages the mobile in zone C, since the mobile is known to be in this zone. If there is no response to paging after a predetermined time, the system recognizes that the mobile is not in zone C, but in zone B, and next pages the mobile in zone B (this is referred to as *zone miss*; note that the second paging always succeeds). This is one of the disadvantages of 2Z. That is, when the mobile reenters a previously visited zone and an incoming call occurs, the system must make two-step paging because of *zone miss*.

However, two-step paging can be avoided, if two-zone registration with implicit registration by outgoing calls (2Zi) is employed. For example, if the mobile enters zones in the order  $A \rightarrow B \rightarrow C \rightarrow B$  and the outgoing call occurs in the last zone, B, then the call setup messages of an outgoing call can provide the system with the exact zone, B, and the system can successfully page the mobile in zone B at this time. In other words, when the mobile makes the outgoing call, call setup messages can provide the correct location information of the mobile implicitly, without an additional location registration message. This is termed implicit registration [5, 7].

Henceforth, for convenience, we refer to location registration by entering a new zone as regular location registration (RR) and the location registration effect by an outgoing call as implicit registration (IR).

2Zi was considered and analyzed by Jang et al. [7], assuming the following 4-direction mobility model.

- (i) A mobile moves in a straight line, until it reaches a turning point.
- (ii) When it reaches the turning point, it can choose one of 4 directions with equal probability.
- (iii) The distance between two consecutive turning points is exponentially distributed.

When the above mobility model is assumed, it is impossible to express the exact equations for the performance measures such as registration cost and paging cost. Thus, for convenience, Jang et al. [7] assumed that, once a mobile enters a zone and makes one direction change, it is located at a random point in the zone. This wild assumption makes it possible to obtain some performance measures, but they are inherently rough approximations of the exact performance. Lin [6] provided a precise model for the performance of 2Z, but Lin's model is too complex to be applied to Jang et al.'s study, which considers implicit registration by outgoing calls under two registered zones.

## 3. New Mathematical Model and Performance Analysis

In this section, we propose a new mathematical model, to analyze the exact performance of 2Z and 2Zi. The radio channel is the most important resource determining the network performance in mobile communication networks. Although, thanks to technological enhancements, the capacity of mobile communication systems has been greatly improved, radio channels still have their own capacity and technological limit. Thus, the signaling cost on radio channels determines

the performance of the entire mobile communication system. The performance analysis of 2Z and 2Zi is conducted from this viewpoint.

3.1. *Notations and Assumptions.* The following notations are defined, to analyze the signaling cost on radio channels:

$N_1(Z)$ : number of location registrations between two incoming calls in the location registration method Z;

$C_1(Z)$ : total location registration cost between two incoming calls in the location registration method Z;

$N_2(Z)$ : number of cells for paging for an incoming call in the location registration method Z;

$C_2(Z)$ : total paging cost per zone for an incoming call in the location registration method Z;

$C_p$ : signaling cost for paging per cell on radio channels;

$C_u$ : signaling cost for one location registration on radio channels;

$\theta$ : probability of returning to the registered zone;

$n$ : number of cells per zone;

$t_c$ : interval between two incoming calls (r. v.);

$t_{oc}$ : interval between two outgoing calls (r. v.);

$t_m$ : sojourn time in a zone (r. v.);

$\lambda_c$ : arrival rate of incoming calls;

$\lambda_{oc}$ : arrival rate of outgoing calls;

$1/\lambda_m$ : mean of sojourn time in a zone;

$\rho$ : call-to-mobility ratio (CMR),  $\rho = \lambda_c/\lambda_m$ ;

$f_m^*(s)$ : Laplace-Stieltjes Transform for  $t_m$  ( $= \int_{t=0}^{\infty} e^{-st} f_m(t)dt$ );

$\Lambda$ : probability that an outgoing call occurs before an incoming call,  $\Lambda = \Pr[t_{oc} \leq t_c] = \lambda_{oc}/(\lambda_c + \lambda_{oc})$ ;

$p$ : probability that an outgoing call occurs while in the zone,  $p = \Pr[t_{oc} \leq t_m] = (\lambda_m/\lambda_{oc})[1 - f_m^*(\lambda_{oc})]$ ;

$q$ : probability that an outgoing call does not occur while in the zone,  $q = 1 - p$ .

In addition, the following assumptions are necessary to analyze the signaling cost on radio channels:

- (i) the incoming calls to a mobile form a Poisson process with  $\lambda_c$ ;
- (ii) the outgoing calls from a mobile form a Poisson process with  $\lambda_{oc}$ ;
- (iii) the sojourn time in a zone follows a general distribution with a mean of  $1/\lambda_m$ ;
- (iv) the first paging is applied to the most recently registered zone. If there is no response, the second paging is applied to the other zone.

3.2. *Performance Analysis of 2Z and 2Zi.* This section studies the performance analysis for 2Z and 2Zi. We first estimate registration cost and paging cost between two incoming calls, which constitute the total signaling cost.

To find registration cost and paging cost between two incoming calls, let us introduce the probability  $\alpha(K)$  that the mobile moves across  $K$  zones between two incoming calls. We use Lin's result on  $\alpha(K)$  [6], because this is closely related to our study:

$$\alpha(K) = \begin{cases} 1 - \frac{1}{\rho} [1 - f_m^*(\lambda_c)], & K = 0, \\ \frac{1}{\rho} [1 - f_m^*(\lambda_c)]^2 [f_m^*(\lambda_c)]^{K-1}, & K \geq 1. \end{cases} \quad (1)$$

3.2.1. *Registration Cost.* The number of RRs between two incoming calls is [6]

$$N_1(2Z) = \sum_{k=0}^{\infty} \sum_{i=0}^k i \binom{k}{i} \theta^{k-i} (1-\theta)^i \alpha(k) = \frac{(1-\theta)}{\rho}. \quad (2)$$

The number of RRs between two incoming calls in 2Zi is the same as that in 2Z:

$$N_1(2Zi) = \frac{(1-\theta)}{\rho}. \quad (3)$$

Thus, the location registration cost between two incoming calls is

$$C_1(2Z) = C_1(2Zi) = \frac{(1-\theta)C_u}{\rho}. \quad (4)$$

3.2.2. *Paging Cost.* Next, consider the paging cost. The paging cost can be derived by multiplying the number of cells to page by the paging cost per cell. In the case of 2Z and 2Zi, the number of cells to page for an incoming call is the sum of the cases where the system has correct location information (first paging success or *zone hit*) and incorrect location information (first paging failure or *zone miss*):

$$\begin{aligned} N_2(2Z) &= n + n \cdot \Pr(\text{zone miss in } 2Z) \\ &= n + n [1 - \Pr(\text{zone hit in } 2Z)], \end{aligned} \quad (5)$$

$$N_2(2Zi) = n + n [1 - \Pr(\text{zone hit in } 2Zi)].$$

To obtain the probability of the first paging success in (5), let us define the conditional probability  $S_Z(K)$  that, given that the mobile moves across  $K$  zones between two incoming calls, the first page succeeds in the location registration method Z. Then, the probabilities that the first page succeeds in 2Z and 2Zi are, respectively,

$$\begin{aligned} \Pr[\text{zone hit in } 2Z] &= \sum_{K=0}^{\infty} S_{2Z}(K) \alpha(K), \\ \Pr[\text{zone hit in } 2Zi] &= \sum_{K=0}^{\infty} S_{2Zi}(K) \alpha(K). \end{aligned} \quad (6)$$

To obtain the above probabilities, we need to derive the general expressions of  $S_{2Z}(K)$  and  $S_{2Zi}(K)$ .

For the sake of convenience, we first derive the probability  $S_{2Zi}(K)$  that the first page succeeds in  $2Zi$ , given that the mobile moves across  $K$  zones between two incoming calls.

(1) *Derivation of the Conditional Probability  $S_{2Zi}(K)$ .* Let us consider the following sequential procedure for deriving the general expression of  $S_{2Zi}(K)$ .

- (i) For  $K = 0$ , no movement occurs between two incoming calls, so the first page always succeeds, and  $S_{2Zi}(0) = 1$ .
- (ii) For  $K = 1$ , the probability that the first page succeeds can be obtained in the following two cases.

- (1) Case 1. The mobile moves to a new zone, with probability  $(1 - \theta)$ . In this case, the *zone.list* is updated by RR; thus, the probability that the first page succeeds is the same as that when  $K = 0$ .
- (2) Case 2. The mobile moves back to the zone from whence it came, with probability  $\theta$ . In this case, the system has incorrect information as to which zone the mobile is located in, but the *zone.list* can be updated by IR, with probability  $\Lambda$  that an outgoing call occurs before an incoming call. Therefore,

$$S_{2Zi}(1) = (1 - \theta) S_{2Zi}(0) + \theta \Lambda. \tag{7}$$

- (iii) For  $K = 2$ , the probability that the first page succeeds can also be obtained in the following two cases.

- (1) Case 1. The first movement of the mobile is to a new zone, with probability  $(1 - \theta)$ . In this case, the *zone.list* is updated; thus, the probability that the first page succeeds is the same as that when  $K = 1$ .
- (2) Case 2. The first movement of the mobile is back to the zone from whence it came, with probability  $\theta$ . In this case, if the *zone.list* is updated by IR, with probability  $p$  that an outgoing call occurs while the mobile is in the zone, then the probability that the first page succeeds is the same as that when  $K = 1$ . If an outgoing call does not occur while the mobile is in the zone, then the system has incorrect location information, and the probability that the first page succeeds is the same as that when  $K = 0$ , because the system will have correct location information when the mobile either enters a new zone or moves back to the zone from whence it came, in the second movement. Therefore,

$$S_{2Zi}(2) = (1 - \theta) S_{2Zi}(1) + \theta [p S_{2Zi}(1) + (1 - p) S_{2Zi}(0)]. \tag{8}$$

- (iv) The process can be generalized, when  $K = k$ . In the first movement after an incoming call, the mobile

moves to a new zone with probability  $(1 - \theta)$  or moves back to the zone from whence it came, with probability  $\theta$ .

- (1) Case 1. In the case where the mobile moves to a new zone, the probability that the first page succeeds is the same as that when  $K = k - 1$ .
- (2) Case 2. In the case where the mobile moves back to the zone from whence it came, if an outgoing call occurs while the mobile is in the zone with probability  $p$ , then the *zone.list* is updated by IR, and the probability that the first page succeeds is the same as that when  $K = k - 1$ . Otherwise, the system will have the correct location information after the mobile makes one more movement, by either entering a new zone or moving back to the zone from whence it came. Thus, the probability that the first page succeeds is the same as that when  $K = k - 2$ . Therefore, we get a recurrence formula for  $S_{2Zi}(k)$  as follows:

$$S_{2Zi}(k) = (1 - \theta) S_{2Zi}(k - 1) + \theta [p S_{2Zi}(k - 1) + (1 - p) S_{2Zi}(k - 2)], \tag{9}$$

for  $k = 2, 3, \dots$ ,

$$S_{2Zi}(0) = 1, \quad S_{2Zi}(1) = (1 - \theta) + \theta \Lambda.$$

Note that  $\Lambda = 0$  and  $p = 0$  in the case of  $2Z$ , because IR by an outgoing call is not employed. Therefore,

$$S_{2Z}(k) = (1 - \theta) S_{2Z}(k - 1) + \theta S_{2Z}(k - 2), \tag{10}$$

for  $k = 2, 3, \dots$ ,

$$S_{2Z}(0) = 1, \quad S_{2Z}(1) = (1 - \theta).$$

(2) *Paging Cost for an Incoming Call.* Using (9) and (10), for  $2Z$  and  $2Zi$ , respectively, (5) can be written by

$$N_2(2Z) = n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Z}(K) \alpha(K) \right), \tag{11}$$

$$N_2(2Zi) = n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Zi}(K) \alpha(K) \right).$$

Using the above results, we can compute the number of cells required when an incoming call occurs, and we can find the paging cost, by multiplying this number by the paging cost per cell. Finally, the total paging costs for an incoming call for  $2Z$  and  $2Zi$  are, respectively,

$$C_2(2Z) = \left[ n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Z}(K) \alpha(K) \right) \right] C_p, \tag{12}$$

$$C_2(2Zi) = \left[ n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Zi}(K) \alpha(K) \right) \right] C_p.$$

3.2.3. *Total Signaling Cost.* The total signaling cost on radio channels is derived by combining the registration cost and the paging cost as follows:

$$C(2Z) = \frac{(1-\theta)C_u}{\rho} + \left[ n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Z}(K) \alpha(K) \right) \right] C_p, \tag{13}$$

$$C(2Zi) = \frac{(1-\theta)C_u}{\rho} + \left[ n + n \left( 1 - \sum_{K=0}^{\infty} S_{2Zi}(K) \alpha(K) \right) \right] C_p.$$

3.2.4. *Propositions for Explicit Expressions of Costs*

**Proposition 1.** *The general solution of (9) is*

$$S_{2Zi}(n) = 1 + \frac{-\theta(1-\Lambda)[1 - (-\theta q)^n]}{1 + \theta q}. \tag{14}$$

*Proof.* For convenience, let us omit subscripts. Rearranging the above equation, we can get

$$S(n) - S(n-1) = -\theta q [S(n-1) - S(n-2)], \tag{15}$$

for  $n = 2, 3, \dots$

It can be seen that differences of the progression form a geometric progression with equal ratio  $(-\theta q)$ . Then, the general term of the progression  $S(n)$  can be easily obtained as follows:

$$S(n) = S(0) + \sum_{i=1}^n (-\theta q)^i$$

$$= 1 + \frac{-\theta(1-\Lambda)[1 - (-\theta q)^n]}{1 + \theta q}$$

$$= \begin{cases} 1 + \frac{-\theta(1-\Lambda)[1 - (\theta q)^n]}{1 + \theta q} & \text{if } n \text{ is even number} \\ 1 + \frac{-\theta(1-\Lambda)[1 + (\theta q)^n]}{1 + \theta q} & \text{if } n \text{ is odd number.} \end{cases} \tag{16}$$

□

**Proposition 2.** *The general solution of (10) is  $S_{2ZR}(n) = [1 - (-\theta)^{n+1}]/(1 + \theta)$ .*

*Proof.* For convenience, let us omit subscripts. In the case of 2Z, since IR by an outgoing call is not employed,  $\Lambda = 0$  and  $p = 0$ . Inserting these values into (10), we have

$$S(n) = \frac{[1 - (-\theta)^{n+1}]}{1 + \theta}$$

$$= \begin{cases} \frac{1 + \theta^{n+1}}{1 + \theta} & \text{if } n \text{ is even number} \\ \frac{1 - \theta^{n+1}}{1 + \theta} & \text{if } n \text{ is odd number.} \end{cases} \tag{17}$$

□

**Proposition 3.** *Equation (17) gives the same  $P[\text{zone hit in } 2Z]$  as in Lin's study [6].*

*Proof.* From (18), (19), and (20) of Lin's study [6],

$$P[\text{zone hit in } 2Z] = \omega_1 + \omega_2 + \omega_3$$

$$= \alpha(0) + \sum_{K=1}^{\infty} \omega_2(K) \alpha(K) + \sum_{i=1}^{\infty} \theta^{2i} \alpha(2i)$$

$$= \alpha(0) + \sum_{K=1}^{\infty} \frac{1 - \theta^{2[(K-1)/2]+2}}{1 + \theta} \alpha(K)$$

$$+ \sum_{i=1}^{\infty} \theta^{2i} \alpha(2i). \tag{18}$$

If we express the previous equation as  $\sum_{i=1}^{\infty} z(K) \alpha(K)$ , it is easy to show that  $z(K)$  is as follows:

$$z(K) = \begin{cases} 1, & \text{if } K = 0, \\ \frac{1 - \theta^K}{1 + \theta} + \theta^K = \frac{1 + \theta^{K+1}}{1 + \theta}, & \text{if } K \text{ is even number,} \\ \frac{1 - \theta^{K+1}}{1 + \theta} + 0 = \frac{1 - \theta^{K+1}}{1 + \theta}, & \text{if } K \text{ is odd number.} \end{cases} \tag{19}$$

Since  $z(K) = S(K)$  for all  $K \geq 0$ , the proof is complete. □

Note that the above proposition implies that our model includes Lin's model on 2Z [6].

As shown in the appendix, the probability that the first paging succeeds in 2Zi, given that the mobile moves across  $k$  zones between two incoming calls, is composed

of three probabilities for three exclusive cases,  $\omega_1(k)$ ,  $\omega_2(k)$ , and  $\omega_3(k)$ , and their sum,  $\omega(k)$ , is

$$\omega(k) = \begin{cases} (1-\theta)(1-\Lambda) \frac{[1-\theta^{k+1}q^{k+1}]}{1-\theta^2q^2} + p(1-\Lambda)\theta^2q \frac{[1-\theta^{k-1}q^{k-1}]}{1-\theta^2q^2} + \Lambda, & \text{if } K \text{ is odd number } (k \geq 3), \\ (1-\theta)(1-\Lambda) \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \theta^kq^k(1-\Lambda) + p(1-\Lambda)\theta^2q \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \Lambda, & \text{if } K \text{ is even number } (k \geq 2), \\ (1-\theta)(1-\Lambda) + \Lambda = 1 - \theta + \theta\Lambda, & \text{if } k = 1, \\ 1, & \text{if } k = 0. \end{cases} \quad (20)$$

**Proposition 4.** Equation (14),  $S(k) = 1 + (-\theta(1-\Lambda)[1 - (-\theta q)^k]/(1+\theta q))$ , is the same as (20).

*Proof.* (i) When  $k = 0$  and  $k = 1$ , it is trivial.  
 (ii) When  $k$  is even number ( $k \geq 2$ ),

$$\begin{aligned} \omega(k) &= (1-\theta)(1-\Lambda) \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \theta^kq^k(1-\Lambda) \\ &\quad + p(1-\Lambda)\theta^2q \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \Lambda \\ &= (1-\theta)(1-\Lambda) [1 + \theta^2q^2 + \theta^4q^4 + \dots + \theta^{k-2}q^{k-2}] \\ &\quad + \theta^kq^k(1-\Lambda) + p(1-\Lambda)\theta^2q \\ &\quad \times [1 + \theta^2q^2 + \theta^4q^4 + \dots + \theta^{k-2}q^{k-2}] + \Lambda \\ &= 1 - \theta(1-\Lambda) [1 - \theta q + \theta^2q^2 - \theta^3q^3 + \theta^4q^4 + \dots \\ &\quad + \theta^{k-2}q^{k-2} - \theta^{k-1}q^{k-1}] \frac{1+\theta q}{1+\theta q} \\ &= 1 - \theta(1-\Lambda) \frac{[1-\theta^kq^k]}{1+\theta q}. \end{aligned} \quad (21)$$

TABLE 1: Signaling cost with respect to CMR.

CMR ( $=\lambda_c/\lambda_m$ )	0.125	0.25	0.5	1	1.5	2
C(1Z)	40.00	24.00	16.00	12.00	10.68	10.00
C(2Z)	26.46	18.29	14.00	11.60	10.68	10.14
C(2Zi)	24.74	16.54	12.48	10.38	9.65	9.26
Reduction ratio (%) $=100 \times [1 - C(2Zi)/C(1Z)]$	38.15	31.08	22.00	13.49	9.63	7.38
Reduction ratio (%) $=100 \times [1 - C(2Zi)/C(2Z)]$	6.50	9.54	10.86	10.51	9.60	8.69

(iii) When  $k$  is odd number ( $k \geq 3$ ),

$$\begin{aligned} \omega(k) &= (1-\theta)(1-\Lambda) \frac{[1-\theta^{k+1}q^{k+1}]}{1-\theta^2q^2} + p(1-\Lambda)\theta^2q \\ &\quad \times \frac{[1-\theta^{k-1}q^{k-1}]}{1-\theta^2q^2} + \Lambda \\ &= 1 - \theta(1-\Lambda) [1 - \theta q + \theta^2q^2 - \theta^3q^3 + \theta^4q^4 + \dots \\ &\quad - \theta^{k-2}q^{k-2} + \theta^{k-1}q^{k-1}] \frac{1+\theta q}{1+\theta q} \\ &= 1 - \theta(1-\Lambda) \frac{[1+\theta^kq^k]}{1+\theta q}. \end{aligned} \quad (22)$$

□

**Proposition 5.** The explicit expressions of  $C(2Z)$  and  $C(2Zi)$  are

$$\begin{aligned} C(2Z) &= \frac{(1-\theta)C_u}{\rho} + \left[ n + n \left( \frac{\theta [1 - f_m^*(\lambda_c)]}{\rho [1 + \theta f_m^*(\lambda_c)]} \right) \right] C_p, \\ C(2Zi) &= \frac{(1-\theta)C_u}{\rho} + \left[ n + n \left( \frac{\theta [1 - f_m^*(\lambda_c)] (1-\Lambda) (1+\theta q)}{\rho (1+\theta q) [1 + \theta q f_m^*(\lambda_c)]} \right) \right] C_p. \end{aligned} \quad (23)$$

(24)

*Proof.* The result follows from

$$\begin{aligned}
 & \sum_{k=0}^{\infty} S_{2Zi}(k) \alpha(k) \\
 &= 1 - \frac{(1 - f_m^*(\lambda_c))}{\rho} \\
 & \quad + \sum_{k=1}^{\infty} \left\{ 1 + \frac{-\theta(1 - \Lambda) [1 - (-\theta q)^k]}{1 + \theta q} \right\} \alpha(k) \\
 &= 1 - \frac{(1 - f_m^*(\lambda_c))}{\rho} + \frac{1 + \theta(q + \Lambda - 1)}{1 + \theta q} \\
 & \quad \times \sum_{k=1}^{\infty} \alpha(k) - \frac{\theta(\Lambda - 1)}{1 + \theta q} \sum_{k=1}^{\infty} (-\theta q)^k \alpha(k) \\
 &= 1 - \frac{(1 - f_m^*(\lambda_c))}{\rho} \\
 & \quad + \frac{[1 + \theta(q + \Lambda - 1)] [1 - f_m^*(\lambda_c)]^2}{\rho(1 + \theta q)} \\
 & \quad \times \left[ \frac{1}{1 - f_m^*(\lambda_c)} \right] - \frac{\theta(\Lambda - 1) [1 - f_m^*(\lambda_c)]^2}{\rho(1 + \theta q)} \\
 & \quad \times \left[ \frac{-\theta q}{1 + \theta q f_m^*(\lambda_c)} \right] \\
 &= 1 - \frac{\theta [1 - f_m^*(\lambda_c)] (1 - \Lambda) (1 + \theta q)}{\rho (1 + \theta q) [1 + \theta q f_m^*(\lambda_c)]}, \\
 & \quad \sum_{k=0}^{\infty} S_{2Z}(k) \alpha(k) = 1 - \frac{\theta [1 - f_m^*(\lambda_c)]}{\rho [1 + \theta f_m^*(\lambda_c)]}. \tag{26}
 \end{aligned}$$

(25)

□

### 4. Numerical Results

In this section, the performances of 1Z, 2Z, and 2Zi are investigated through various numerical results for the signaling cost on radio channels. The signaling cost of 1Z can be obtained by substituting  $\theta = 0$  in (24). The performance of 2Z is analyzed, using both our proposed model and Lin's model [6], and it can be seen that the results of both models are the same, in every case, as shown in Proposition 3. The performance of 2Zi is analyzed, using our proposed model, and is compared with those of 1Z and 2Z.

We obtain the numerical results, assuming the following environments [2, 6, 7]:

$$\begin{aligned}
 C_p &= 1, & C_u &= 4, & \theta &= 0.5, \\
 n &= 8, & \lambda_c &= 1, & \lambda_{oc} &= 4, & \lambda_m &= 4.
 \end{aligned} \tag{27}$$

In our examples, the sojourn time in a zone ( $t_m$ ) is assumed to follow an exponential distribution, for convenience. However, since the foregoing equations were derived

TABLE 2: Paging cost with respect to  $\lambda_m$ .

$\lambda_m$	0.5	1	2	4	8
$C_2(1Z)$	8.00	8.00	8.00	8.00	8.00
$C_2(2Z)$	9.14	9.60	10.00	10.29	10.46
$C_2(2Zi)$	8.26	8.38	8.48	8.54	8.74
Reduction ratio (%) = $100 \times [1 - C_2(2Zi)/C_2(1Z)]$	9.64	12.70	15.20	16.97	16.44
Reduction ratio (%) = $100 \times [C_2(2Z)/C_2(1Z) - 1]$	14.29	20.00	25.00	28.57	30.77
Reduction ratio (%) = $100 \times [C_2(2Zi)/C_2(1Z) - 1]$	3.27	4.76	6.00	6.76	9.27

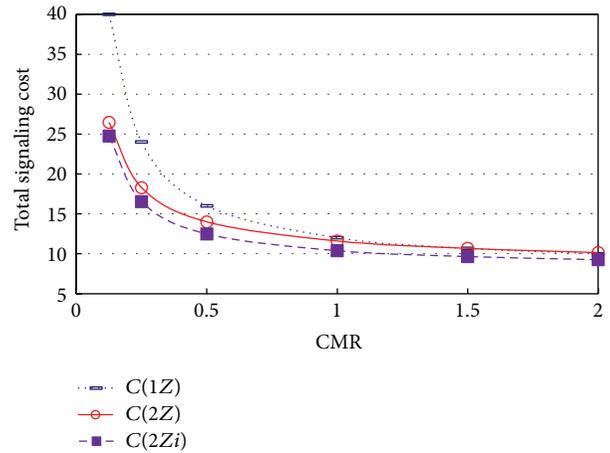


FIGURE 2: Signaling cost with respect to CMR.

under the assumption that  $t_m$  has a general distribution, any distribution can be assumed.

Figure 2 shows the signaling cost with respect to CMR ( $= \lambda_c/\lambda_m$ ). It shows the signaling cost when  $\lambda_c = 1$ , with different levels of  $\lambda_m$  from 0.5 to 8.0. The same results are shown in Table 1. As shown in Figure 2 and Table 1, the signaling cost of 2Z is lower than that of 1Z, in most cases, and the signaling cost of 2Zi is lower than that of 2Z. Table 1 shows that the signaling cost of 2Zi is 22% lower than that of 1Z and 10.86% lower than that of 2Z, when  $CMR = 1/2$ . In fact, the signaling cost of 2Zi is lower than those of the other two methods, 2Z and 1Z, in most cases.

Table 1 also shows that, as CMR increases ( $\lambda_m$  decreases), the signaling cost of 2Zi always remains lower than that of 1Z, but the reduction ratio of the signaling cost decreases. Conversely, the signaling cost of 2Zi is always lower than that of 2Z, for all CMR values, but the largest reduction of the signaling cost occurs when  $CMR = 1/2$ . Another notable feature of Table 1 is that the signaling cost of 2Zi is lower than that of 1Z, whereas the signaling cost of 2Z is greater than that of 1Z, when  $CMR = 2$ . When CMR is very large (i.e.,  $\lambda_m$  is very small), there are very few location registrations, and 2Z, which has an increasing paging cost, may have a disadvantage, compared to 1Z. Although it is not shown in Table 1, we can infer that 2Zi also may have a disadvantage, compared to 1Z, when CMR is very large.

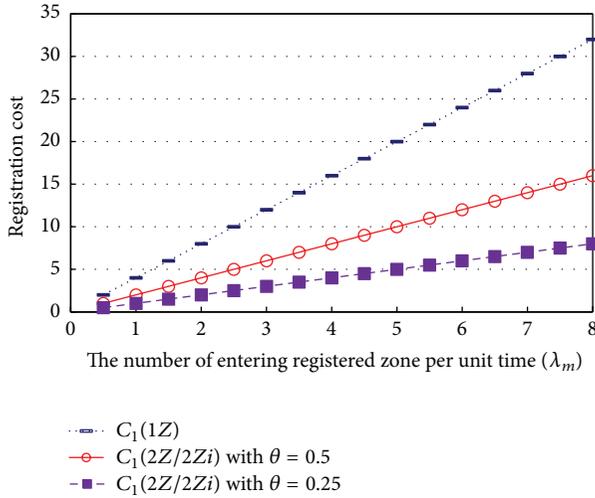


FIGURE 3: Location registration cost with respect to  $\lambda_m$ .

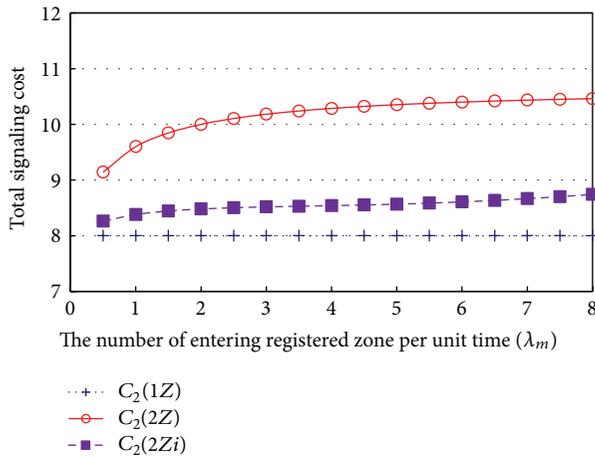


FIGURE 4: Paging cost with respect to  $\lambda_m$ .

The signaling cost of  $2Zi$  and  $2Z$  is lower than that of  $1Z$ , because  $2Zi$  and  $2Z$  have lower location registration cost than  $1Z$ . To show this feature clearly, we present the location registration cost with respect to  $\lambda_m$ , when  $\lambda_c = 1$ , in Figure 3. As shown in Figure 3, the increase of the location registration cost is exactly proportional to the increase of  $\lambda_m$ . In addition, the location registration cost is directly related to  $\theta$ , which is the probability of returning to the previous zone. The location registration cost of  $2Z$  and  $2Zi$  is 25% lower than that of  $1Z$ , when  $\theta$  is 0.25, and 50% lower, when  $\theta$  is 0.5.

The location registration cost of  $2Z$  and  $2Zi$  is lower, but the paging cost is greater, than that of  $1Z$ . To show this feature clearly, we present the paging cost with respect to  $\lambda_m$ , when  $\lambda_c = 1$ , in Figure 4.

To show this feature clearly, we present the paging cost with respect to  $\lambda_m$ , when  $\lambda_c = 1$ , in Figure 4 and Table 2. As shown in Figure 4 and Table 2, the paging cost of  $1Z$  remains constant, while that of  $2Z$  and  $2Zi$  increases, as  $\lambda_m$  increases. One of the notable results of this study is that, when  $\lambda_m = 8$ , the paging cost of  $2Z$ , which is 10.46 (30.77% greater than 8.00

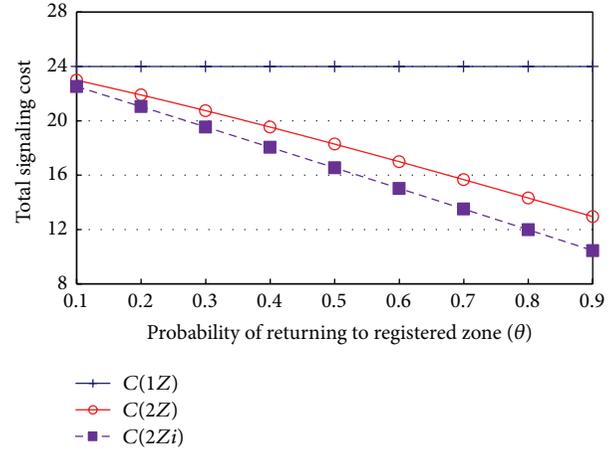


FIGURE 5: Signaling cost with respect to  $\theta$ .

of  $1Z$ ), can be reduced to 8.74 (9.25% greater than 8.00 of  $1Z$ ) if  $2Zi$  is adopted, which causes the total signaling cost of  $2Zi$  to be lower than that of  $2Z$ , to an extent corresponding to this reduction, as shown in Table 2.

Figure 5 shows the signaling cost of each method with respect to  $\theta$ , the probability of returning to the previous zone. As shown in Figure 5, the signaling cost of  $2Z$  and  $2Zi$  decreases, as  $\theta$  increases. In particular, the signaling cost of  $2Zi$  decreases more than that of  $2Z$ . Even though it is clear that the signaling cost of  $2Z$  and  $2Zi$  decreases, as  $\theta$  increases, it seems to be unreasonable to assume that  $\theta$  is larger than 0.5, in a real-world mobile communication environment.

Finally, Figure 6 shows the signaling cost with respect to  $n$ , the number of cells in a zone. In this case, since the location registration cost remains constant, the overall amount of the signaling cost will increase, as the number of cells in a zone increases, due to the increase of the paging cost. As shown in Figure 6, the signaling costs of  $1Z$ ,  $2Z$ , and  $2Zi$  all increase, as the number of cells in a zone increases, but the increased ratios of  $2Zi$  and  $1Z$  are lower than that of  $2Z$ . That is, if the other conditions are the same,  $2Z$  is more superior to  $1Z$ , and  $2Z$  is more superior to  $2Zi$ , respectively, as the paging cost increases.

## 5. Conclusion

Many efficient mobility management methods have been suggested, to minimize the signaling cost on radio channels. This study considered the zone-based registration methods that are widely used in the majority of mobile communication networks.

We provided a new mathematical model to analyze the performance of the zone-based registration methods,  $2Z$  and  $2Zi$ , by considering implicit registration effects of outgoing calls from a mobile, which were not considered properly in the previous studies. It should be noted that our mathematical model is simple, compared to the previous studies, but provides the exact performance of  $2Zi$  for the first time. Also, our model can easily be applied to  $2Z$  and  $1Z$  and provides the same results as Lin's previous study on  $2Z$  and  $1Z$ . From

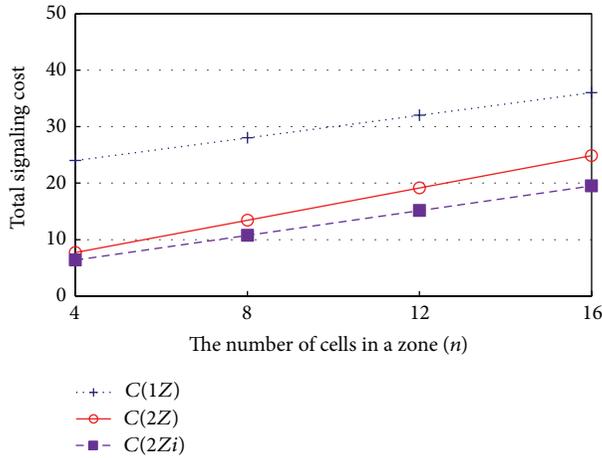


FIGURE 6: Signaling cost with respect to the number of cells in a zone ( $\lambda_{oc} = \lambda_m = 5$ ).

various numerical results by using our model, we showed that 2Zi is superior to 2Z as well as 1Z, in most cases.

Our results are helpful in considering which registration scheme should be adopted. For further study, we will consider the case where a mobile can have multiple zones, to get the performance of every type of zone-based registration.

### Appendix

#### Derivation of $\omega(k)$

**Proposition 6.** *The probability that the first paging succeeds in 2Zi, given that the mobile moves across k zones between two incoming calls, is composed of three probabilities for three exclusive cases,  $\omega_1(k)$ ,  $\omega_2(k)$ , and  $\omega_3(k)$ , and their sum,  $\omega(k)$ , is given by*

$$\omega(k) = \begin{cases} (1-\theta)(1-\Lambda) \frac{[1-\theta^{k+1}q^{k+1}]}{1-\theta^2q^2} + p(1-\Lambda)\theta^2q \frac{[1-\theta^{k-1}q^{k-1}]}{1-\theta^2q^2} + \Lambda, & \text{if } k \text{ is odd number } (k \geq 3), \\ (1-\theta)(1-\Lambda) \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \theta^kq^k(1-\Lambda) + p(1-\Lambda)\theta^2q \frac{[1-\theta^kq^k]}{1-\theta^2q^2} + \Lambda, & \text{if } k \text{ is even number } (k \geq 2), \\ (1-\theta)(1-\Lambda) + \Lambda = 1 - \theta + \theta\Lambda, & \text{if } k = 1, \\ 1, & \text{if } k = 0. \end{cases} \quad (A.1)$$

*Proof.* If the probability  $\Pr[\text{zone hit in } 2Zi]$  that the first paging succeeds in 2Zi is composed of the three probabilities,  $\omega_1$ ,  $\omega_2$ , and  $\omega_3$ , for three exclusive cases, then we have

$$\Pr[\text{zone hit in } 2Zi] = \omega_1 + \omega_2 + \omega_3. \quad (A.2)$$

Each probability for the three exclusive cases can be derived as follows.

(i) Case 1. If, between two incoming calls, the last registration is followed by an even number of movements with no registration and no outgoing call, then the first paging succeeds (see Figure 7(a)).

Letting  $\omega_1$  be the probability of Case 1, we have

$$\omega_1 = \sum_{k=1}^{\infty} \omega_1(k) \alpha(k). \quad (A.3)$$

In the above,  $\omega_1(k)$  is the conditional probability that, given that the mobile moves across k zones between two incoming calls, the last registration is followed by an even number of movements, with no registration and no outgoing call. This can be derived by

$$\omega_1(k) = \begin{cases} (1-\theta)(1-\Lambda) \frac{[1-\theta^kq^k]}{1-\theta^2q^2}, & \text{if } k \text{ is even number,} \\ (1-\theta)(1-\Lambda) \frac{[1-\theta^{k-1}q^{k-1}]}{1-\theta^2q^2}, & \text{if } k \text{ is odd number,} \end{cases} \quad (A.4)$$

since, for  $k \geq 1$ ,

$$\begin{aligned} \omega_1(k) &= (1-\theta) \sum_{i=0}^{\lfloor (k-1)/2 \rfloor} \theta^{2i} (1-P[t_{oc} < t_m])^{2i} P[t_c < t_{oc}] \\ &= (1-\theta) \sum_{i=0}^{\lfloor (k-1)/2 \rfloor} \theta^{2i} q^{2i} (1-\Lambda) \\ &= (1-\theta)(1-\Lambda) \frac{[1-\theta^{2\lfloor (k-1)/2 \rfloor + 2} q^{2\lfloor (k-1)/2 \rfloor + 2}]}{1-\theta^2q^2}. \end{aligned} \quad (A.5)$$

(ii) Case 2. If, between two incoming calls, the last outgoing call is followed by an even number of movements, with no registration and no further outgoing calls, then the first paging succeeds (see Figure 7(b)).

Letting  $\omega_2$  be the probability of Case 2, we have

$$\omega_2 = \sum_{k=1}^{\infty} \omega_2(k) \alpha(k). \quad (A.6)$$

In the above,  $\omega_2(k)$  is the conditional probability that, given that the mobile moves across k zones between two incoming calls, the last outgoing call is followed by an even

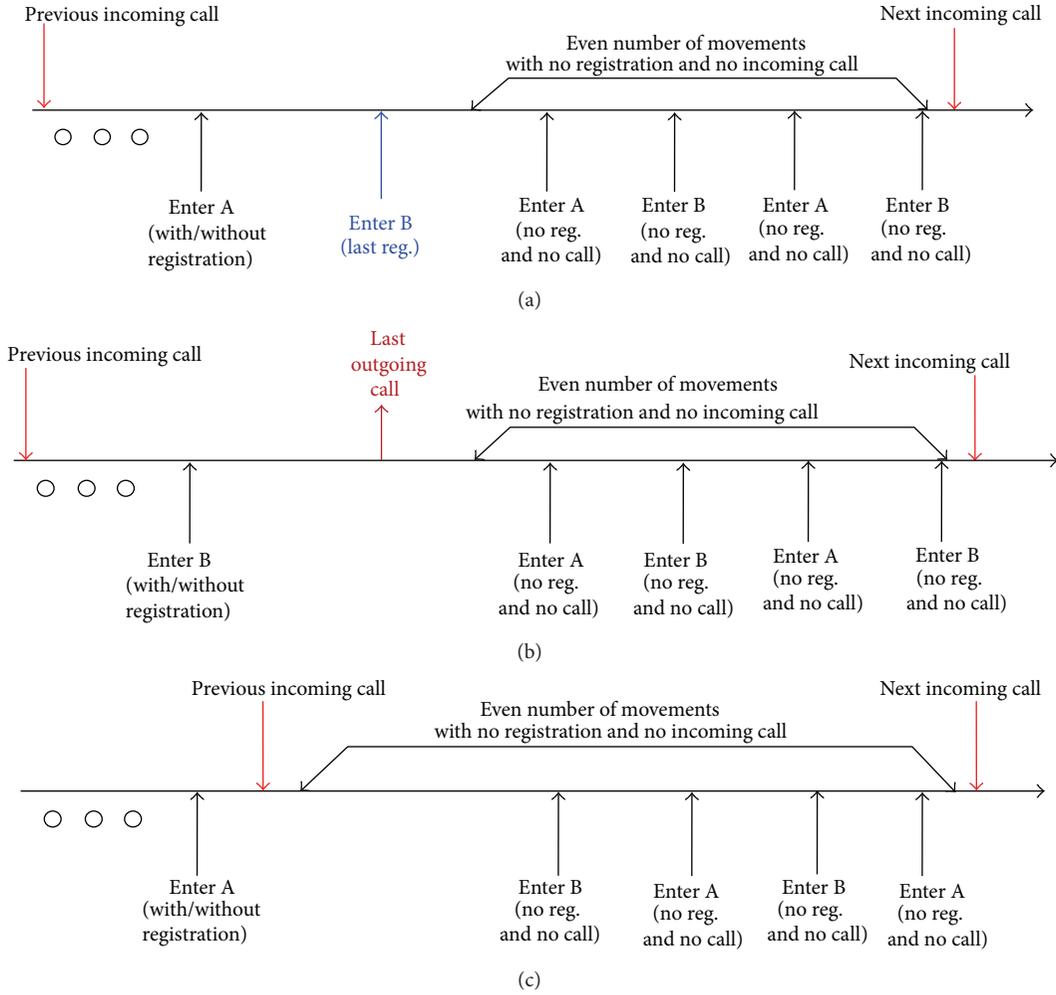


FIGURE 7: Situations when the system has the correct view of the latest visited zone. (a) Case 1: the last registration is followed by an even number of movements. (b) Case 2: the last outgoing call is followed by an even number of movements with no registration and no outgoing call. (b) Case 3: there are an even number of movements with no registration and no outgoing call.

number of movements, with no registration and no further outgoing call. This can be derived by

since, for  $k \geq 2$ ,

$$\omega_2(k) = \begin{cases} \Lambda + p(1-\Lambda)\theta^2 q \frac{[1-\theta^k q^k]}{1-\theta^2 q^2}, & \text{if } k \text{ is even number,} \\ \Lambda + p(1-\Lambda)\theta^2 q \frac{[1-\theta^{k-1} q^{k-1}]}{1-\theta^2 q^2}, & \text{if } k \text{ is odd number } (k \geq 3), \\ \Lambda, & \text{if } k = 1, \end{cases} \quad (A.7)$$

$$\omega_2(k) = P[t_c > t_{oc}] + P[t_m > t_{oc}] \times \sum_{i=1}^{\lfloor k/2 \rfloor} \theta^{2i} (1 - P[t_{oc} < t_m])^{2i-1} P[t_c < t_{oc}]$$

$$= \Lambda + p \sum_{i=1}^{\lfloor k/2 \rfloor} \theta^{2i} q^{2i-1} (1 - \Lambda) \quad (A.8)$$

$$= \Lambda + p(1-\Lambda)\theta^2 q \frac{[1-\theta^{2\lfloor k/2 \rfloor} q^{2\lfloor k/2 \rfloor}]}{1-\theta^2 q^2}.$$

(iii) Case 3. If, between two incoming calls, there are an even number of movements with no registration and no outgoing call, then the first paging succeeds (see Figure 7(c)).

Letting  $\omega_3$  be the probability of Case 3, we have

$$\omega_3 = \sum_{k=1}^{\infty} \omega_3(k) \alpha(k). \tag{A.9}$$

In the above,  $\omega_3(k)$  is the conditional probability that, given that the mobile moves across  $k$  zones between two incoming calls, there are an odd number of movements, with no registration and no outgoing call. This can be derived by

$$\omega_3(k) = \begin{cases} \theta^k q^k (1 - \Lambda), & \text{if } k \text{ is even number,} \\ 0, & \text{if } k \text{ is odd number,} \end{cases} \tag{A.10}$$

since, for even number  $k$ ,

$$\begin{aligned} \omega_3(k) &= \sum_{i=1}^{k/2} \theta^{2i} (1 - P[t_{oc} < t_m])^{2i} P[t_c < t_{oc}] \\ &= \theta^k q^k (1 - \Lambda). \end{aligned} \tag{A.11}$$

Finally, we have

$$\begin{aligned} \Pr[\text{zone hit in } 2Zi] &= \omega_1 + \omega_2 + \omega_3 \\ &= \sum_{k=1}^{\infty} \omega_1(k) \alpha(k) + \sum_{k=1}^{\infty} \omega_2(k) \alpha(k) \\ &\quad + \sum_{k=1}^{\infty} \omega_3(k) \alpha(k) = \sum_{k=1}^{\infty} \omega(k) \alpha(k), \end{aligned} \tag{A.12}$$

where

$$\omega(k) = \begin{cases} (1 - \theta)(1 - \Lambda) \frac{[1 - \theta^{k+1} q^{k+1}]}{1 - \theta^2 q^2} \\ \quad + p(1 - \Lambda) \theta^2 q \frac{[1 - \theta^{k-1} q^{k-1}]}{1 - \theta^2 q^2} + \Lambda, \\ \quad \text{if } k \text{ is odd number } (k \geq 3), \\ (1 - \theta)(1 - \Lambda) \frac{[1 - \theta^k q^k]}{1 - \theta^2 q^2} + \theta^k q^k (1 - \Lambda) \\ \quad + p(1 - \Lambda) \theta^2 q \frac{[1 - \theta^k q^k]}{1 - \theta^2 q^2} + \Lambda, \\ \quad \text{if } k \text{ is even number } (k \geq 2), \\ (1 - \theta)(1 - \Lambda) + \Lambda = 1 - \theta + \theta\Lambda, \\ \quad \text{if } k = 1, \\ 1, \quad \text{if } k = 0. \end{cases} \tag{A.13}$$

□

**Conflict of Interests**

The authors declare that there is no conflict of interests regarding the publication of this paper.

**Acknowledgment**

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant no. 2011-0015214).

**References**

- [1] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: to update or not to update?" *Wireless Networks*, vol. 1, no. 2, pp. 175–185, 1995.
- [2] I. F. Akyildiz, J. S. M. Ho, and Y.-B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 629–638, 1996.
- [3] J. Li, H. Kameda, and K. Li, "Optimal dynamic mobility management for PCS networks," *IEEE/ACM Transactions on Networking*, vol. 8, no. 3, pp. 319–327, 2000.
- [4] R. H. Liou, Y. B. Lin, and S. C. Tsai, "An investigation on LTE mobility management," *IEEE Transactions on Mobile Computing*, vol. 12, no. 1, pp. 166–176, 2013.
- [5] TIA/EIA/IS-95-B, MS-BS compatibility standard for dual-mode wideband spread, 1999.
- [6] Y.-B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 25–33, 1997.
- [7] H.-S. Jang, H. Hwang, and K.-P. Jun, "Modeling and analysis of two-location algorithm with implicit registration in CDMA personal communication network," *Computers and Industrial Engineering*, vol. 41, no. 1, pp. 95–108, 2001.
- [8] Z. Liu, J. Almhana, and R. McGorman, "Markov mobility model and registration area optimization in cellular networks," *Wireless Communications and Mobile Computing*, vol. 9, no. 12, pp. 1608–1617, 2009.
- [9] Z. Mao and C. Douligeris, "A location-based mobility tracking scheme for PCS networks," *Computer Communications*, vol. 23, no. 18, pp. 1729–1739, 2000.

## Research Article

# Historical Feature Pattern Extraction Based Network Attack Situation Sensing Algorithm

Yong Zeng,<sup>1,2</sup> Dacheng Liu,<sup>1</sup> and Zhou Lei<sup>1,3</sup>

<sup>1</sup> Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

<sup>2</sup> Bengbu Automobile NCO Academy, Bengbu 233011, China

<sup>3</sup> Tianjin Port (Group), Ltd., Tianjin 300456, China

Correspondence should be addressed to Dacheng Liu; liudacv@mail.tsinghua.edu.cn

Received 20 February 2014; Accepted 18 March 2014; Published 27 April 2014

Academic Editors: Y. Mao and Z. Zhou

Copyright © 2014 Yong Zeng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The situation sequence contains a series of complicated and multivariate random trends, which are very sudden, uncertain, and difficult to recognize and describe its principle by traditional algorithms. To solve the above questions, estimating parameters of super long situation sequence is essential, but very difficult, so this paper proposes a situation prediction method based on historical feature pattern extraction (HFPE). First, HFPE algorithm seeks similar indications from the history situation sequence recorded and weighs the link intensity between occurred indication and subsequent effect. Then it calculates the probability that a certain effect reappears according to the current indication and makes a prediction after weighting. Meanwhile, HFPE method gives an evolution algorithm to derive the prediction deviation from the views of pattern and accuracy. This algorithm can continuously promote the adaptability of HFPE through gradual fine-tuning. The method preserves the rules in sequence at its best, does not need data preprocessing, and can track and adapt to the variation of situation sequence continuously.

## 1. Introduction

With attacks becoming more prevalent, the traditional static passive defense and whole system consolidation are hard to keep up with the changing rhythms, which have huge amounts of investment and affect the network performance. In this case, the dynamic, proactive, and targeted defending measures have been presented, most of which rely on attack situation forecast, that is, network attack situation sensing (NASS) [1, 2]. NASS aims at forecasting future evolution trend of network attack situation based on historical features and current attack indications, guiding dynamic defense, and allowing administrators to take corresponding measures in advanced, and effective manner to quickly respond to the complex and ever-changing attack threats [3, 4].

Rarely studying attack situation forecast, previous researches mostly using existing methods, such as autoregressive moving average model (ARMA), grey model (GM), and radial basis function neural network (RBFNN) [5–8]. ARMA identifies the dependence relationship and autocorrelation of situation sequences and establish mathematical prediction

model [9]. It requests that situation sequences or their certain step difference satisfies the steady suppose, which is too strict to increase suitable scope. As one of GMs, GM(1,1) firstly weakens the randomness of situation sequences by using accumulation, secondly fits the born sequence through index curve, and then does regressive restitution after prediction, which can embody monotonously and slowly changing trend but hardly reflect some characteristics such as random rove and periodic fluctuation [10, 11]. Grey Verhulst is suitable to describe the situation sequences with swing development according to “S” or anti-“S” form [12], and the method dividing the changing line into several stages does not lack rationality, but the difficulty is how to predict the occurrence moment and lasting time of each stage [13]. RBFNN utilizes the nonlinear characteristic to describe the regulation contained in situation sequences [14]. However, evolving regulation of attack situation is infinite and changeable; a practical type neural network with small scale cannot solve well [15, 16].

Situation sequence contains massive complex and inconstant evolution trends, beyond the expression and prediction

capability of traditional methods only by some formulas, functions or via some training [17, 18]. Most traditional methods suffer from the confliction among training samples, rely on data preprocessing and artificial intervention heavily, do not support incremental training, and need to rebuild model once situation sequence changes [19–21]. Therefore, a situation prediction method based on historical feature pattern extraction (HFPE) is presented. The method measures the similarity between historical feature from the aspects of pattern and accuracy and utilizes multiple order difference operation to discriminate trends. It searches similar indications from recorded historical situation sequence, measures the link intensity of occurred indication upon subsequent effect, and infers the recurrence possibilities of some effects according to current indication. An evolution algorithm is introduced to measure prediction deviation and improve the adaptability of prediction algorithm continuously via gradual fine-tuning.

This paper proceeds as follows: Section 2 discusses algorithm principle for HFPE. Section 3 clarifies algorithm establishment and analysis. Section 4 presents the experiment results and Section 5 concludes the paper.

## 2. Algorithm Principle

**2.1. Basic Definition.** Looking from mathematical form, the continuous time-varied curve,  $z = f(t)$ , is commonly applied to describe the evolving process of attack situation. This curve is carried out by computer through sampling method, that is, to sample situation values with time interval  $\tau$ , and then obtains discrete time sequences composed by  $(t_k, z_k)$ , where  $z_k$  represents the situation value at moment  $t_k$ . To facilitate the research, a basic definition is made as follows: let  $G(i, m)$  be the segmental subimage with  $m$  neighboring segments from moment  $t_i$ , let  $q_k$  be the segmental gradient, let  $Q(i, m)$  be the gradient sequence, let  $(q_i, q_{i+1}, \dots, q_{i+m-1})$ ,  $L(i, m)$  be the characteristic spectrum of  $Q(i, m)$ , let  $O$  be zero vector; then

$$q_k = \frac{z_{k+1} - z_k}{t_{k+1} - t_k},$$

$$L(i, m) = O, \quad Q(i, m) = O, \quad (1)$$

$$L(i, m) = \frac{Q(i, m)}{\|Q(i, m)\|}, \quad Q(i, m) \neq O.$$

For the  $k$ th component product of  $L(i, m)$ ,  $l_{i+k}$ , the angle of inclination,  $\theta_{i+k}$ , can be defined as

$$\theta_{i+k} = \arctan l_{i+k}, \quad -\frac{\pi}{2} < \theta_{i+k} < \frac{\pi}{2}. \quad (2)$$

The stretch rate from  $Q(i, m)$  to  $Q(j, m)$  can be calculated by  $f_\sigma(i, j, m)$ , which is defined as

$$f_\sigma(i, j, m) = 1, \quad Q(i, m) = O, \quad Q(j, m) = O,$$

$$f_\sigma(i, j, m) = \frac{\|Q(j, m)\|}{\|Q(i, m)\|}, \quad Q(i, m) \neq O, \quad Q(j, m) \neq O,$$

not exist, other conditions. (3)

$\gamma[i, m, \rho]$  is utilized to adjust the stretch rate, where  $\rho$  is the prediction steps.

The following three theorems through further analysis can be easily obtained: (1)  $L(i, m)$  does not change with  $G(i, m)$ ; (2) if  $L(i, m) = L(j, m)$ , then  $Q(i, m)$  is linear correlative with  $Q(j, m)$ ; (3) if  $L(i, m) = L(j, m)$ , then  $G(i, m)$  can be the same with  $G(j, m)$  through translating and magnifying.

**2.2. Prediction Principle.** Looking from probability theory and statistics, similar situation curve shapes are more probably derived from similar origin, mechanism, and impact, subsequently resulting in a similar subsequent effect. From the point of view of statistics, when the precedence relations of sequences in time appear frequently, it usually meant that the logical causal relationship exists in a certain degree.

It is supposed that  $G(i, m)$  and  $G(j, m)$  are known historical feature subpatterns, from the same pattern,  $t_i < t_j$ , and the further trend after  $t > t_{j+m}$  is unknown and needed to be predicted. If  $G(i, m)$  is similar with  $G(j, m)$ , then it can be deduced that the origin, mechanism, and impact in  $[t_j, t_{j+m})$  are similar with those in  $[t_i, t_{i+m})$ , and the history after  $t_{i+m}$  may be repeated after  $t_{j+m}$  with some differences. According to this principle, the slope of the line segment behind can be forecasted by

$$\hat{q}_{j+m+k} = f_\sigma(i, j, m) \times q_{i+m+k}. \quad (4)$$

$\rho$  is utilized to control the predicting steps. When  $k = 0, 1, 2, \dots, \rho - 1$ , the trend prediction curve can be recurred by  $\tau$  and  $\hat{q}_{j+m+k}$ .

## 2.3. Measurement System

**2.3.1. Fitting Degree.** Firstly calculate the angle cosine similarity between slope vectors, secondly introduce more order difference operators to obtain the trend difference of qualitative change and quantitative change, and then acquire the narrowing fitting degree by the difference of similarity degree and trend difference.

Let  $\phi_\theta(i, j, m)$  represent the angle cosine similarity between  $Q(i, m)$  and  $Q(j, m)$ ; then

$$\phi_\theta(i, j, m) = 1, \quad Q(i, m) = O, \quad Q(j, m) = O,$$

$$\phi_\theta(i, j, m) = \frac{Q(i, m) \times Q^T(j, m)}{\|Q(i, m)\| \times \|Q(j, m)\|},$$

$$Q(i, m) \neq O, \quad Q(j, m) \neq O \quad (5)$$

$$\phi_\theta(i, j, m) = 0, \quad Q(i, m) = O, \quad Q(j, m) \neq O$$

or  $Q(i, m) \neq O, \quad Q(j, m) = O.$

The trend differences of qualitative change and quantitative change are denoted by  $\phi_1(x)$  and  $\phi_2(x)$ , respectively, and the former of which stands for the pattern difference,

and the latter stands for the accuracy difference. The above two parameters can be derived by

$$\phi_1(x) = \begin{cases} +1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0, \end{cases} \quad (6)$$

$$\phi_2(x) = \sin x.$$

Thus the composite trend,  $\phi_{\perp}(x)$ , can be defined by

$$\phi_{\perp}(x) = 0.2 \times \phi_1(x) + 0.8 \times \phi_2(x). \quad (7)$$

Let  $\nabla$  represent backward difference operator, and define

$$\nabla^0 \theta_k = \theta_k, \quad (8)$$

and then the  $\alpha$  order differential recursive equation can be obtained by

$$\nabla^{\alpha} \theta_k = 0.5 \times (\nabla^{\alpha-1} \theta_k - \nabla^{\alpha-1} \theta_{k-1}), \quad (9)$$

in which  $\alpha$  is a positive integer, and (9) meets

$$\frac{\pi}{2} < \nabla^{\alpha} \theta_k < \frac{\pi}{2}. \quad (10)$$

Let  $\phi_{\nabla}(i, j, m)$  denote the trend difference between the feature patterns  $L(i, m)$  and  $L(j, m)$ ; then

$$\phi_{\nabla}(i, j, m) = \frac{2 \sum_{\alpha=0}^{m-1} \sum_{k=\alpha}^{m-1} |\phi_{\perp}(\nabla^{\alpha} \theta_{i+k}) - \phi_{\perp}(\nabla^{\alpha} \theta_{j+k})|}{m(m+1)}. \quad (11)$$

The fitting degree function,  $\phi(i, j, m)$ , can be defined by

$$\phi(i, j, m) = \phi_{\theta}(i, j, m) - \phi_{\nabla}(i, j, m), \quad (12)$$

where the large value of  $\phi(i, j, m)$  represents a fine fitting, and for  $-1 < \phi_{\theta}(i, j, m) \leq 1$  and  $0 < \phi_{\nabla}(i, j, m) \leq 2$ , it can be derived that

$$-3 < \phi_{\theta}(i, j, m) \leq 1. \quad (13)$$

The occurrence probability of  $\phi_{\theta}(i, j, m) > 0$  may be 50% statistically, which is too big. Therefore, it is necessary to subtract the penalty term,  $\phi_{\nabla}(i, j, m)$ , and filter  $\phi(i, j, m)$  by the threshold  $\varepsilon_{\phi}$  ( $0 < \varepsilon_{\phi} < 1$ ) to narrow the fitting degree.

**2.3.2. Universality Degree.** Divide the attack situation sub-sequence into two parts, that is, occurred indication and subsequent effect; the values of the domination intensity of the former to the latter (or call link intensity between the two parts) may be high or low, some of which have a far-ranging representative, and some just have rare earth especially instance. If all the values are treated evenly, then the prediction accuracy will be affected seriously, so it is important to outstand inevitable link of the high intensity and weaken accidental link of the low intensity.

Let  $\chi[k, m, \rho]$  be the universality value of  $Q(k, m + \rho)$  in the historical feature pattern  $G(0, n)$ , where  $\chi_{\max}$  can be derived by

$$\chi_{\max} = \max \{ \chi [k, m, \rho] \mid 0 \leq k < n - m \}. \quad (14)$$

The value of  $\chi_{\max}$  will be updated with the change of  $\chi[k, m, \rho]$  and can be accessed directly without waiting to calculate.

The universality value can be shined upon to universality degree in  $(0, n - m]$  by function  $f_{\chi}(k, m, \rho)$ , which is shown as follows:

$$f_{\chi}(k, m, \rho) = (n - m) \times (1 + 2\pi^{-1} \arctan(\chi[k, m, \rho] - \chi_{\max})). \quad (15)$$

The larger value of universality degree reflects finer representativeness of  $Q(k, m)$  and its extension and more exact patterns predicted by  $Q(k, m + \rho)$ . Otherwise,  $Q(k, m + \rho)$  is just a special example, and the prediction effect is worse.

**2.3.3. Contrast Degree.** The predication results of situation are usually impacted by link intensity of several different weights. The function mechanism often changes; that is, sometimes they work with a community decision and sometimes with an individual domination. Therefore, it is necessary to trace and adjust between outstanding statistics effect and showing individual advantage.

It is supposed that  $\bar{w}_1, \bar{w}_2, \dots, \bar{w}_n$  are not normalized weights, which can be adjusted to  $\bar{w}_1^{\eta}, \bar{w}_2^{\eta}, \dots, \bar{w}_n^{\eta}$  by sensitization index  $\eta$  ( $\eta > 0$ ). Then the standardized weight  $w_k$  can be derived by

$$w_k = \frac{\bar{w}_k^{\eta}}{\sum_{k=1}^n \bar{w}_k^{\eta}}, \quad (16)$$

and comparison degree  $w_i/w_j$  can be obtained by

$$\frac{w_i}{w_j} = \frac{\bar{w}_i^{\eta}}{\bar{w}_j^{\eta}}. \quad (17)$$

If  $\bar{w}_i \neq \bar{w}_j$ , we can suppose  $\bar{w}_i < \bar{w}_j$ ; then five generalized cutoff points can be acquired; that is,  $0 < \bar{w}_i/\bar{w}_j < 1 < \bar{w}_j/\bar{w}_i < \infty$ .

Equation (16) can be derived into the form of  $\bar{w}_k^{\eta} = w_k \times \xi$  by

$$\frac{\bar{w}_k^{\eta \times x}}{\sum_{k=1}^n \bar{w}_k^{\eta \times x}} = \frac{(w_k \times \xi)^x}{\sum_{k=1}^n (w_k \times \xi)^x} = \frac{w_k^x}{\sum_{k=1}^n w_k^x}, \quad (18)$$

where  $\eta$  is utilized to adjust comparison degree; that is,  $0 < \eta < 1$  outstands statistics effect,  $\eta > 1$  shows individual advantage, and  $\eta = 1$  maintains the present status.

### 3. Algorithm Establishment and Analysis

**3.1. Prediction Algorithm.** The prediction algorithm flow is given in Figure 1. According to historical feature pattern

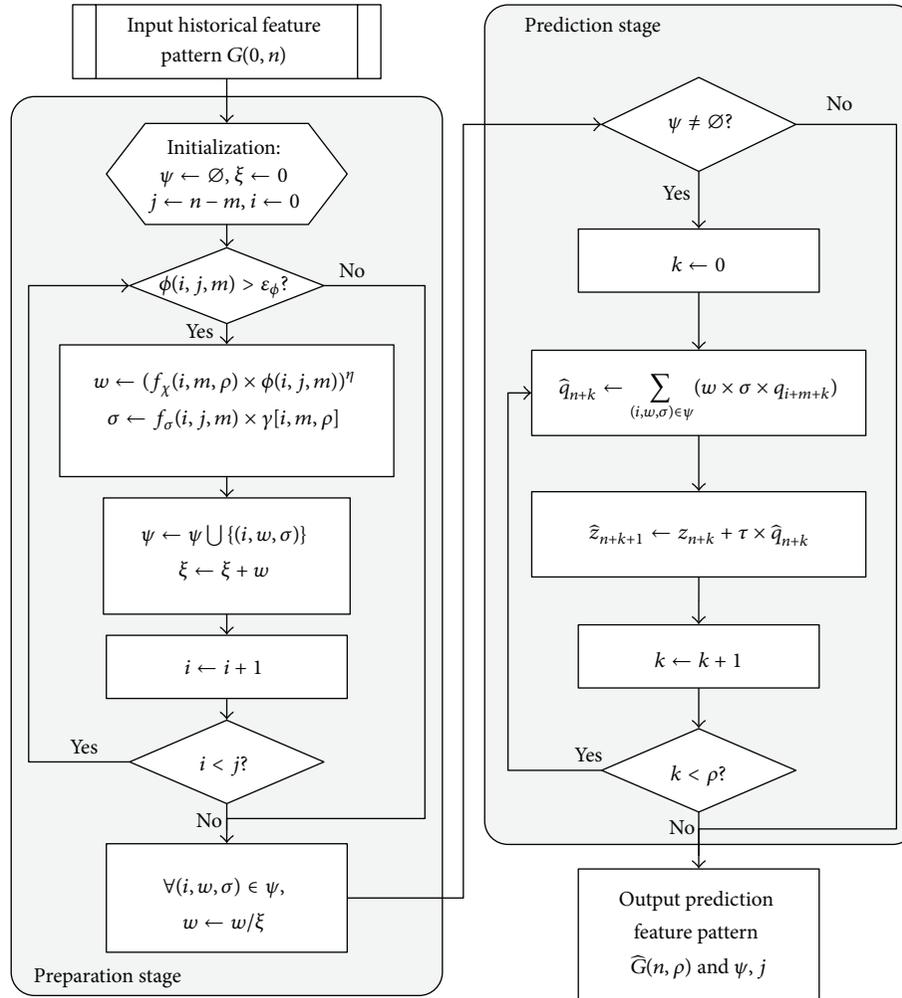


FIGURE 1: Prediction algorithm flow chart.

$G(0, n-1)$  and occurred indication  $G(n-m, m)$ , the algorithm can predict subsequent effect  $\widehat{G}(n, \rho)$  through two main stages, that is, preparation stage and prediction stage, which are marked in the chart.

As shown in the figure, the preparation part circularly promotes the sliding window  $G(i, m)$ , selects poor values of fitting degree  $Q(i, m)$  to reject, and sensitizes the product of universality degree  $f_\chi(i, m, \rho)$  and fitting degree  $\phi(i, j, m)$ , which is assigned to  $w$ . The prediction part first checks whether historical feature pattern set  $\psi$  has record. What calls for special attentions is that the value of  $Q(i, m)$  in the sliding window or the fitting degree value of it with  $Q(j, m)$  in the occurred indication cannot be too small, because the smaller the above value, the poorer the contribution to prediction value  $\widehat{q}_{n+k}$ .

**3.2. Evolution Algorithm.** Evolution algorithm is introduced to measure predicting deviation from the views of pattern and accuracy, which can be fine-tuned to raise the adaptability of prediction algorithm.

The accuracy of adjusting  $\eta$  to  $\eta \times x$  can be derived by

$$f_k(x) = \frac{\sum_{(i,w,\sigma) \in \psi} (w^x \times \phi(i+m, n, \rho))}{\sum_{(i,w,\sigma) \in \psi} w^x}, \quad (19)$$

which is based on current weight set and (18) and meets  $-3 < f_k(x) \leq 1$ . And the definition can be popularized to  $f_\Lambda(x_1, x_2, \dots)$ , only when  $f_k(x_i)$  is the largest value first met in  $\{f_k(x_1), f_k(x_2), \dots\}$ , which is in ascending order by  $k$  of  $x_k$ ; it can be obtained that

$$f_\Lambda(x_1, x_2, \dots) = x_i. \quad (20)$$

As shown in Figure 2, the evolution algorithm carries on the variables and results of the prediction algorithm and works to promote adaptability after acquiring measured value.  $\Delta\epsilon_\phi$  is an adjustment variable for  $\epsilon_\phi$  and meets  $-n^{-1} \leq \Delta\epsilon_\phi \leq n^{-1}$ . If  $n$  rises or the distance between  $|\psi|$  and  $\ln n$  drops, then the adjustment amplitude becomes lower, else becomes higher. If  $|\psi| < \ln n$ , then decrease the threshold to soften the terms, else increase the threshold.  $\Delta\chi$  is calculated

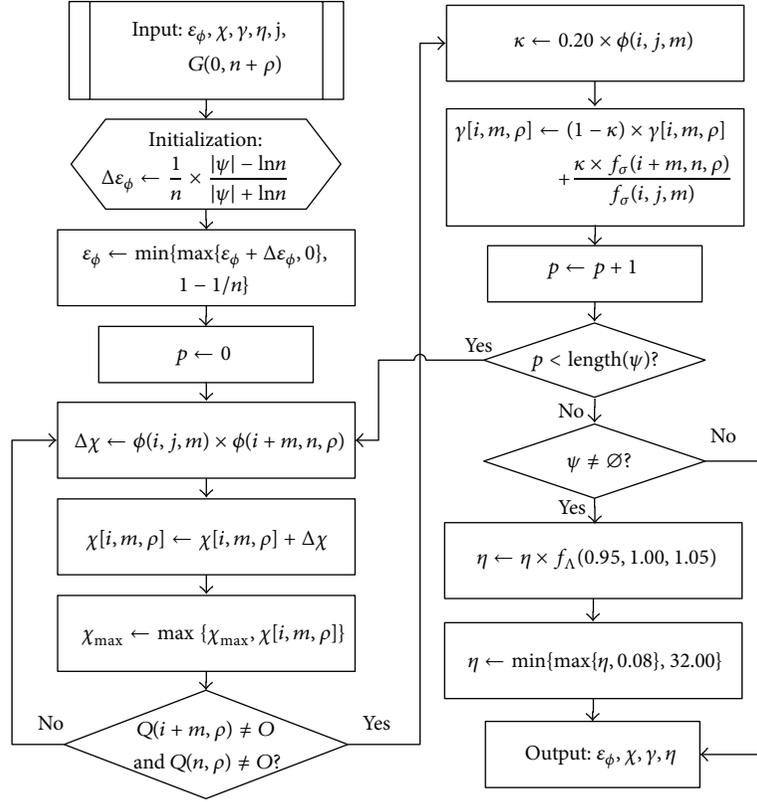


FIGURE 2: Evolution algorithm flow chart.

through fitting degree,  $\phi(i, j, m)$ , selected by the prediction algorithm, and meeting  $0 < \phi(i, j, m) \leq 1$ . When the fitting of  $Q(i, m)$  and  $Q(j, m)$  is poor, the adjustment to  $\chi[i, m, \rho]$  needs to be cautious; that is, if  $\phi(i + m, n, \rho) > 0$ , then it needs to raise the value of  $\chi[i, m, \rho]$ , and if the extension value  $Q(i + m, \rho)$  approximates  $Q(n, \rho)$ , then the prediction according to  $Q(i + m, \rho)$  is accurate, and the value raising can be large. To determine  $\eta$ , select the best one among value lowering by 5%, current value, and value rising by 5%, and restrict it by a reasonable range to prevent passivating or sharpening.

**3.3. Example Analysis.** Figure 3 gives an example of predicting  $\widehat{G}(13, 2)$  according to the historical feature pattern  $G(0, 13)$ , in which the value of  $(n, m, \rho)$  is  $(13, 3, 2)$ ,  $\varepsilon_\phi = 0.2$ ,  $\chi[i, m, \rho] = 0.0$ ,  $\chi_{\max} = 0.0$ ,  $\gamma[i, m, \rho] = 1.0$ , and  $\eta = 1.0$ .

According to the prediction algorithm, it can be known through comparing all the values of  $Q(i, 3)$  with  $Q(10, 3)$  that when  $i = 0$  and  $\phi(0, 10, 3) = 1.00$ , so  $Q(0, 3)$  is selected, and when  $i = 5$  and  $\phi(5, 10, 3) = 0.42$ , so  $Q(5, 3)$  is also selected, and other values are excluded for they meet  $\phi(i, 10, 3) \leq \varepsilon_\phi$ , which are partly listed in Table 1. When  $i = 5$ , the slope becomes larger at  $t = 7$  and smaller at  $t = 12$ , which are reflected through  $\nabla^1 \theta_7 > 0$  and  $\nabla^1 \theta_{12} < 0$  derived by (9), and the relative penalty value is recorded by  $\phi_\nabla(5, 10, 3)$ . And through normalization process, the elements of set  $\psi$  are  $(0, 0.705, 0.67)$  and  $(5, 0.295, 0.98)$ . Thus, the prediction value  $\widehat{q}_{13}$  is equal to  $0.705 \times 0.67 \times (-1.29) + 0.295 \times 0.98 \times 0.50$ ,

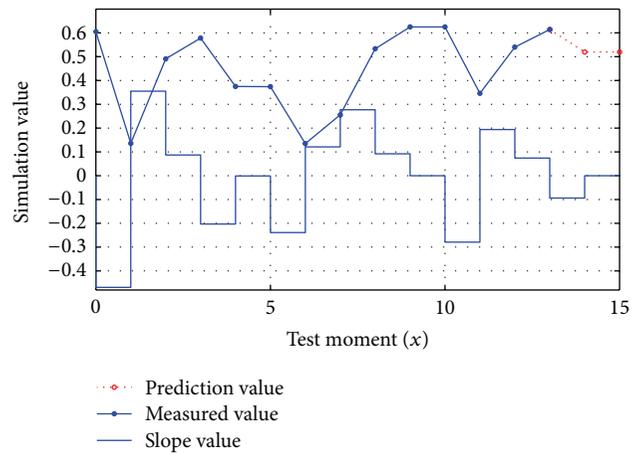


FIGURE 3: Trend prediction illustration.

TABLE 1: Similarity metric and punishment value.

$i$	0	2	3	5	6	8
$\phi_\ominus$	1.00	-0.85	0.42	0.72	0.32	-0.32
$\phi_\nabla$	0.00	1.01	0.34	0.30	0.30	0.68

and so forth, and  $\rho$  step trend can be predicted. It can be seen that this scheme has the ability to identify multiple long-range correlation contained in the same situation sequence.

TABLE 2: Effects of prediction and evolution.

	1	3	5	7	9
$\omega_0$	0.71	0.95	0.98	0.99	1.00
$\omega_5$	0.29	0.05	0.02	0.01	0.00
$\hat{q}_{13}$	-0.46	-0.78	-0.83	-0.85	-0.86
$\hat{q}_{14}$	0.00	0.00	0.00	0.00	0.00
$\varepsilon_\phi$	0.19	0.17	0.15	0.13	0.11
$f_\chi(0, 3, 2)$	10.00	10.00	10.00	10.00	10.00
$f_\chi(5, 3, 2)$	3.27	1.18	0.72	0.51	0.40
$\gamma[0, 3, 2]$	1.00	1.00	1.00	1.00	1.00
$\gamma[5, 3, 2]$	1.06	1.17	1.27	1.35	1.41
$\eta$	1.05	1.16	1.28	1.41	1.55

This part is analyzed according to evolution algorithm. Assuming that  $q_{13} = -0.86$  and  $q_{14} = 0.00$ , so  $|\psi| = 2$ , which is smaller than  $\ln 13$ ; thus, the value of  $\varepsilon_\phi$  needs to be lower, and once  $|\psi| > 2$ , then raise the value of  $\varepsilon_\phi$ . It can be known that the changing value of universality degree  $\chi[0, 3, 2]$  is  $1.00 \times 1.00$ , and raising this degree can strengthen the role of vector  $Q(0, 5)$ . The changing value of  $\chi[5, 3, 2]$  is  $0.42 \times (-1.85)$ , and lowering this value can weaken the interference of vector  $Q(5, 5)$ . To bridge the gap between epitaxial scale and measured scale,  $\gamma[0, 3, 2]$  is adjusted to  $(1 - 0.20) \times 1.00 + 0.20 \times 0.67/0.67$ , and  $\gamma[5, 3, 2]$  is adjusted to  $(1 - 0.08) \times 1.00 + 0.08 \times 1.72/0.98$ . And for  $(f_\kappa(0.95), f_\kappa(1.00), f_\kappa(1.05)) = (0.13, 0.16, 0.19)$ , the value of  $\eta$  needs to rise. Table 2 shows that  $f_\chi(5, 3, 2)$  becomes smaller, and  $\eta$  becomes larger with continued evolution, which results in rapid rise of  $w_0/w_5$ , and approach between prediction value and measured value.

With the passage of time,  $m$  and  $\rho$  keep unchanged,  $n$  grows linearly, and the algorithm can delete stale data, save recent data, and correct fitting threshold and universality degree. The above process can be complicated not only by autonomous evolution, but also by artificially modified parameters.

#### 4. Experiment Results Analysis

The traditional indexes utilized to measure the prediction accuracy include mean absolute error (MAE), standard deviation error (SDE), and mean absolute percentage error (MAPE) [21] derived by

$$\text{MAPE} = \frac{1}{\rho} \sum_{k=n+1}^{n+\rho} \left| \frac{\hat{z}_k - z_k}{z_k} \right|. \quad (21)$$

This section selects MAPE to obtain the relative error between prediction pattern and measured pattern, which is denoted by  $E_r$ . The standard deviation of  $\rho$  relative error components is denoted by  $E_{\text{std}}$ .

**4.1. Experiment 1.** Figure 4 is a critical subsequence selected from actual network attack situation records, which includes various features such as ascent trend, saturation trend, decline trend, periodic fluctuation, and stochastic disturbance.

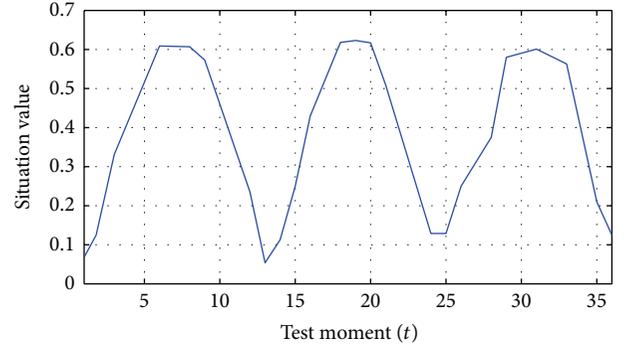


FIGURE 4: Historical records of network attack situation.

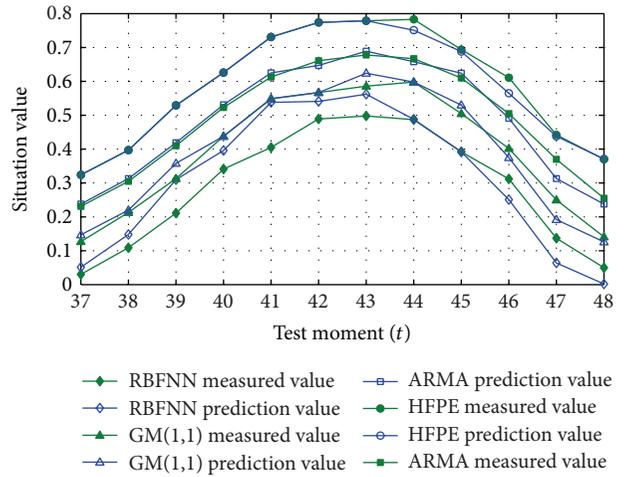


FIGURE 5: Prediction results of network attack situation.

From the view of the experimental prediction results, the relative errors of HFPE, ARMA, GM(1,1), and RBFNN are 3.28%, 5.89%, 7.18%, and 16.11%, respectively. As shown in Figure 5, in the experiment, ARMA, GM(1,1), and RBFNN need to be artificially identified and protected against cyclical situation fluctuations. The difference transformation utilizes 12 as the distance and is restored after prediction to prevent poor prediction effect; otherwise, the relative errors of GM(1,1) and RBFNN may reach 59.67% and 73.99%, respectively. However, the above method is special, cannot be spread for that data preprocessing of these algorithms does not exist in universal law. On the contrary, HFPE can maintain adaptation to complicated and changeable trends but does not need data preprocessing or artificial cognition.

**4.2. Experiment 2.** This experiment is to randomly choose subsequences with similar parts, repeat 20 times, and then calculate the average value.

From the view of the experimental prediction results, the relative errors of HFPE, ARMA, GM(1,1), and RBFNN are 8.09%, 20.89%, 44.89%, and 34.75%, respectively. If the situation sequence selected does not exist in any principle, then the relative errors will be 3.96%, 21.72%, 37.47%, and 53.54%, respectively. Figure 6 shows one group of data, in

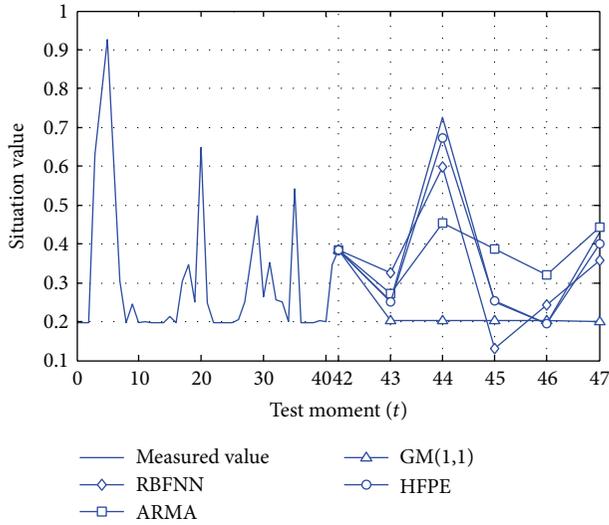


FIGURE 6: Prediction results of network attack situation.

which  $t = 42$  is a boundary for historical feature pattern and prediction feature pattern.

If put all groups of the historical feature patterns into a new long sequence, and repeat above prediction, then the performance of ARMA and GM(1,1) drops rapidly, and that of HFPE does not change much for that longer sequence containing more correlation is benefit to prediction.

**4.3. Experiment 3.** To compare differences among four algorithms, random data are utilized to simulate situation sequences. First, extract random data with  $\xi$  bits from the entropy pool of Windows 7 system. Then randomly gather subsequence with 16 bits, the former 8 bits of which are occurred indication and the latter 8 bits are subsequent effect. Thirdly, splice occurred indication behind the random sequence to form a historical feature pattern and treat the subsequent effect as a prediction feature pattern. Let us make 100 groups of experiments to test each algorithm's capacity in resisting random interference and in identifying the correlation with far distance. The average results are listed in Table 3.

It can be found from the table data that HFPE has the best performance among the four algorithms. When the scale of experiment is large, this conclusion can be repeated well. And ARMA and RBFNN cannot deal with the random sequences with long bits, while HFPE can perform smoothly.

## 5. Conclusion

This paper proposes a prediction method based on historical feature pattern, that is, HFPE. The main principle of this algorithm is shown as follows. Fitting degree is introduced to measure the similarity among subsequences from the views of pattern and accuracy. Universality degree is utilized to test the representation of subsequence and its epitaxy. Contrast of the weight system is adjusted by sensitized index, which gives prominence to statistical effect in passivation

TABLE 3: Prediction effects of network attack situation.

	$\xi$	$\varepsilon_\phi$	$E_r$	$E_{std}$
HFPE	$1.2 \times 10^2$	0.930	2.49	0.09
	$1.2 \times 10^3$	0.980	1.75	0.09
	$1.0 \times 10^6$	0.998	1.88	0.09
ARMA	$1.2 \times 10^2$	—	32.96	0.15
GM(1,1)	$1.2 \times 10^2$	—	49.32	0.20
	$1.2 \times 10^3$	—	51.06	0.18
RBFNN	$1.2 \times 10^2$	—	27.87	0.23

area and highlights individual strengths in sharpening area. Prediction algorithm and evolution algorithm are proposed to predict situation results according to historical feature patterns. HFPE algorithm maximally reserves the rules in the situation sequences, does not need data preprocessing, and can adapt to the situation changes automatically. The experiment results prove the performance of HFPE.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

This work was supported by China Postdoctoral Science Foundation (no. 2013M540107).

## References

- [1] J. Wang, Z.-G. Qin, and L. Ye, "Research on prediction technique of network situation awareness," in *Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems (CIS '08)*, pp. 570–574, Chengdu, China, September 2008.
- [2] H. Zhu, R. Lu, X. Shen, and X. Lin, "Security in service-oriented vehicular networks," *IEEE Wireless Communications*, vol. 16, no. 4, pp. 16–22, 2009.
- [3] G. Jakobson, J. Buford, and L. Lewis, "Situation management," *IEEE Communications Magazine*, vol. 48, no. 3, pp. 110–111, 2010.
- [4] F. Lan, W. Chunlei, and M. Guoqing, "A framework for network security situation awareness based on knowledge discovery," in *Proceedings of the 2nd International Conference on Computer Engineering and Technology (ICCET '10)*, vol. 1, pp. 226–231, Chengdu, China, April 2010.
- [5] W. Lou and K. Ren, "Security, privacy, and accountability in wireless access networks," *IEEE Wireless Communications*, vol. 16, no. 4, pp. 80–87, 2009.
- [6] A. Feng, M. Knieser, M. Rizkalla, B. King, P. Salama, and F. Bowen, "Embedded system for sensor communication and security," *IET Information Security*, vol. 6, no. 2, pp. 111–121, 2012.
- [7] B. Magoutas, G. Mentzas, and D. Apostolou, "Proactive situation management in the future internet: the case of the smart power grid," in *Proceedings of the 22nd International Workshop on Database and Expert Systems Applications (DEXA '11)*, pp. 267–271, Toulouse, Greece, September 2011.

- [8] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1292–1304, 2011.
- [9] X. Chen, H. Gao, and Y. Fu, "Situation analysis and prediction of web public sentiment," in *Proceedings of the International Symposium on Information Science and Engineering (ISISE '08)*, vol. 2, pp. 707–710, Shanghai, China, December 2008.
- [10] N. Zhang, N. Cheng, N. Lu, H. Zhou, J. W. Mark, and X. Shen, "Risk-aware cooperative spectrum access for multi-channel cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 516–527, 2014.
- [11] R. K. Iyer, Z. Kalbarczyk, K. Pattabiraman et al., "Toward application-aware security and reliability," *IEEE Security & Privacy*, vol. 5, no. 1, pp. 57–62, 2007.
- [12] K. Janac, "Control of large power systems based on situation recognition and high speed simulation," *IEEE Transactions on Power Apparatus and Systems*, vol. 98, no. 3, pp. 710–715, 1979.
- [13] R. K. Iyer, Z. Kalbarczyk, K. Pattabiraman et al., "Toward application-aware security and reliability," *IEEE Security & Privacy*, vol. 5, no. 1, pp. 57–62, 2007.
- [14] A. Kitadai, M. Nakagawa, H. Baba, and A. Watanabe, "Similarity evaluation and shape feature extraction for character pattern retrieval to support reading historical documents," in *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS '12)*, pp. 359–336, Gold Cost, Australia, March 2012.
- [15] J. R. Almeida, J. B. Camargo, and P. S. Cugnasca, "Safety and security in critical applications and in information systems ?? A comparative study," *IEEE Latin America Transactions*, vol. 11, no. 4, pp. 1127–1133, 2013.
- [16] M. Panteli, P. A. Crossley, D. S. Kirschen, and D. J. Sobajic, "Assessing the impact of insufficient situation awareness on power system operation," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 2967–2977, 2013.
- [17] M. M. Masud, Q. Chen, L. Khan et al., "Classification and adaptive novel class detection of feature-evolving data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1484–1497, 2013.
- [18] L. Jibao, W. Huiqiang, L. Xiaowu, and L. Ying, "A quantitative prediction method of network security situation based on wavelet neural network," in *Proceedings of the 1st International Symposium on Data, Privacy, and E-Commerce (ISDPE '07)*, pp. 197–202, Chengdu, China, November 2007.
- [19] W. He, G. Hu, and H. Xiang, "Apply anomaly grey forecasting algorithm to cyberspace situation prediction," in *Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems (CIS '08)*, pp. 503–505, Chengdu, China, September 2008.
- [20] R.-F. Wu and G.-L. Chen, "Research of network security situation prediction based on multidimensional cloud model," in *Proceedings of the 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS '12)*, pp. 409–414, Palermo, Italy, 2012.
- [21] X. Cheng and S. Lang, "Research on network security situation assessment and prediction," in *Proceedings of the 4th International Conference on Computational and Information Sciences (ICIS '12)*, pp. 864–867, Chongqing, China, August 2012.

## Research Article

# Compressive Sensing Based Bayesian Sparse Channel Estimation for OFDM Communication Systems: High Performance and Low Complexity

Guan Gui,<sup>1</sup> Li Xu,<sup>2</sup> Lin Shan,<sup>3</sup> and Fumiyuki Adachi<sup>1</sup>

<sup>1</sup> Department of Communications Engineering, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan

<sup>2</sup> Faculty of Systems Science and Technology, Akita Prefectural University, Akita 015-0055, Japan

<sup>3</sup> Wireless Network Research Institute, National Institute of Information and Communications Technology (NICT), Yokosuka 239-0847, Japan

Correspondence should be addressed to Guan Gui; [gui@mobile.ecei.tohoku.ac.jp](mailto:gui@mobile.ecei.tohoku.ac.jp)

Received 12 February 2014; Accepted 24 March 2014; Published 10 April 2014

Academic Editors: J. Lloret and Y. Mao

Copyright © 2014 Guan Gui et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In orthogonal frequency division modulation (OFDM) communication systems, channel state information (CSI) is required at receiver due to the fact that frequency-selective fading channel leads to disgusting intersymbol interference (ISI) over data transmission. Broadband channel model is often described by very few dominant channel taps and they can be probed by compressive sensing based sparse channel estimation (SCE) methods, for example, orthogonal matching pursuit algorithm, which can take the advantage of sparse structure effectively in the channel as for prior information. However, these developed methods are vulnerable to both noise interference and column coherence of training signal matrix. In other words, the primary objective of these conventional methods is to catch the dominant channel taps without a report of posterior channel uncertainty. To improve the estimation performance, we proposed a compressive sensing based Bayesian sparse channel estimation (BSCE) method which cannot only exploit the channel sparsity but also mitigate the unexpected channel uncertainty without scarifying any computational complexity. The proposed method can reveal potential ambiguity among multiple channel estimators that are ambiguous due to observation noise or correlation interference among columns in the training matrix. Computer simulations show that proposed method can improve the estimation performance when comparing with conventional SCE methods.

## 1. Introduction

In broadband wireless communication systems using orthogonal frequency division modulation (OFDM), frequency-selective fading is incurred by the reflection, diffraction, and scattering of the transmitted signals due to the buildings, large moving vehicles, mountains, and so forth. Such fading phenomenon distorts received signals and poses critical challenges in the design of communication systems for high-rate and high-mobility wireless communication applications. Hence, accurate channel estimation becomes a fundamental problem of such communication systems. In last several years, various linear estimation methods have been proposed based on the assumption of rich multipath channel model. However, recently, a lot of physical channel measurements verified that the channel taps exhibit sparse distribution [1–3]

due to the broadband signal transmission. A typical example of sparse multipath channel is shown in Figure 1 where the length is 100 while the number of nonzero taps is 5 only. Note that different broadband transmission may incur different channel structures in wireless communication systems as shown in Table 1.

To improve the estimation performance, extra sparse structure information can be exploited as prior information. Thanks to the development of compressive sensing [4, 5], many sparse channel estimation (CCE) methods have been proposed for exploiting the channel sparsity. In [6], orthogonal matching pursuit (OMP) algorithm with application to sparse multipath channel estimation in the OFDM systems has been proposed. In [7, 8], sparse channel estimation methods have been proposed using compressive sampling

TABLE 1: Channel structures in different mobile communication systems.

Generations of mobile communication systems [21]	2G cellular (IS-95)	3G cellular (WCDMA)	4G/5G cellular (LTE-Advanced~)
Transmission bandwidth	1.23 MHz	10 MHz	20 MHz~100 MHz
Time-delay spread (for example)	0.5 $\mu$ s	0.5 $\mu$ s	0.5 $\mu$ s
Sampling channel length	1	10	20~100
Number of nonzero taps	1	4	2~10
Channel structure model	Dense	Approximate sparse	Sparse

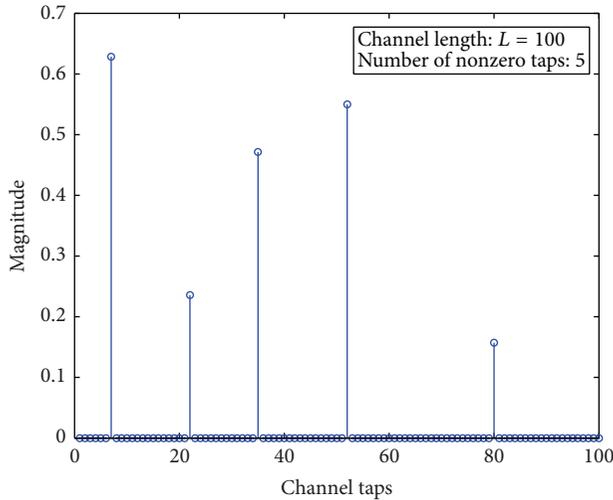


FIGURE 1: A typical example of sparse multipath channel.

matching pursuit (CoSaMP) algorithm [9] in frequency-selective and doubly-selective channel fading communication systems. In [10], to further reduce the computational complexity, sparse channel estimation using smooth  $\ell_0$ -norm (SL0) algorithm [11] has been proposed. Compared to traditional linear methods, sparse channel estimation methods have two obvious advantages: spectral efficiency and lower performance bound. For one thing, improving the spectral efficiency by utilizing less training sequence can achieve the same estimation performance as linear methods. For another, the lower performance bound can be obtained by exploiting channel sparsity due to the fact that less active channel freedom of degree is acquired [12].

Conventional sparse channel estimation methods have a cardinal objective that try to probe the dominant channel taps as accurate as possible, while these methods neglect the posterior information report from additive noise received signal. These proposed channel estimation methods are termed as model selection or basis selection. Unfortunately, their estimation performances are often degraded due to the neglecting channel model uncertainty [13]. To mitigate the unexpected model uncertainty, Bayesian compressive sensing (BCS) [14] and a slight improved Bayesian compressive sensing using Laplace priors (BCS-LAP) [15] could be adopted for estimating sparse channel. The estimation performance could be improved effectively but at the cost of high computational complexity when comparing with existing simple algorithms

(e.g., OMP [6] and SL0 [11]). Hence, it is impractical to employ this algorithm in real communication systems.

Unlike these aforementioned methods, in this paper, we propose an improved Bayesian sparse channel estimation (BSCE) method while its computational complexity is comparable with OMP and SL0. Our proposed Bayesian channel method can be divided into two steps: position detection of dominant channel taps and channel estimation using minimum mean square error (MMSE). In general, our proposed Bayesian estimation method provides model uncertainty which reveals uncertainty among multiple possible position sets of dominant channel taps that are ambiguous due to observation noise or correlation among columns in the training matrix. Furthermore, the complexity of the proposed method is relatively lower due to its smaller search space when compared to conventional methods. Simulation results are given to verify two folds: performance and complexity. Note that estimation performance is evaluated by two metrics: mean-square-error (MSE) and bit-error rate (BER), while computational complexity is measured coarsely by CPU time of computer.

The remainder of this paper is organized as follows. An OFDM system model is described and problem formulation is given in Section 2. In Section 3, the BSCE method is proposed in OFDM systems. Computer simulation results are given in Section 4 in order to evaluate and compare performance of the BSCE method with conventional methods. Finally, we conclude the paper in Section 5.

*Notation 1.* Throughout the paper, matrices and vectors are represented by boldface upper case letters (i.e.,  $\mathbf{X}$ ) and boldface lower case letters (i.e.,  $\mathbf{x}$ ), respectively; the superscripts  $(\cdot)^T$ ,  $(\cdot)^H$ ,  $(\cdot)^{-1}$ , and  $\text{diag}(\cdot)$  denote the transpose, the Hermitian transpose, and the inverse and diagonal operators, respectively;  $E\{\cdot\}$  denotes the expectation operator;  $\|\mathbf{h}\|_0$  is the  $\ell_0$ -norm operator that counts the number of nonzero taps in  $\mathbf{h}$ ; and  $\|\mathbf{h}\|_p$  stands for the  $\ell_p$ -norm operator which is computed by  $\|\mathbf{h}\|_p = (\sum_l |h_l|^p)^{1/p}$ , where  $p \in \{1, 2\}$  is considered in this paper.

## 2. System Model and Problem Formulation

Consider a frequency-selective multipath channel whose impulse response is given by

$$\mathbf{h} = \sum_{l=0}^{L-1} h_l \delta(\tau - \tau_l), \quad (1)$$

where  $L$  is the number of multipaths and  $h_l$  and  $\tau_l$  are the (complex) channel gain and the delay spread, respectively, of path  $l$  at time  $t$ . Hence, the  $L$ -length discrete channel vector can be written as  $\mathbf{h} = [h_0, h_1, \dots, h_{L-1}]^T$ . Let the OFDM system use size- $N$  discrete Fourier transform (DFT), and its number of pilot subcarriers is  $N_p$ . To avoid intersymbol interference (ISI), we assume that the length  $N_g$  of the zero-padding cyclic prefix (CP) in the OFDM symbols is larger than maximum delay spread  $\tau_{\max}$ , where  $\tau_{\max} \geq \tau_l, l = 0, 1, \dots, L - 1$ . Suppose that  $\bar{X}(i)$  denote  $i$ th subcarrier in an OFDM symbol, where  $i = 0, 1, \dots, N - 1$ . If the coherence time of the channel is much larger than the OFDM symbol duration  $T$ , then the channel can be considered quasistatic over an OFDM symbol. Let  $\bar{\mathbf{y}}$  be the vector of received signal samples in one OFDM symbol after DFT; then

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\bar{\mathbf{h}} + \bar{\mathbf{z}} = \bar{\mathbf{X}}\mathbf{F}\mathbf{h} + \bar{\mathbf{z}} = \mathbf{X}\mathbf{h} + \mathbf{z}, \quad (2)$$

where  $\bar{\mathbf{X}} = \text{diag}\{X(0), X(1), \dots, X(N - 1)\}$  denotes diagonal subcarrier matrix,  $\bar{\mathbf{h}}$  is the channel frequency response (CFR) in frequency-domain, and  $\bar{\mathbf{z}}$  is assumed to be additive white Gaussian noise (AWGN) with variance  $\sigma^2$ .  $\mathbf{F}$  is an  $N \times L$  partial DFT matrix with its  $k$ th row which is easily given by  $1/\sqrt{N}[0, e^{-j2\pi k/N}, \dots, e^{-j2\pi k(L-1)/N}]$  and  $\mathbf{X} = \bar{\mathbf{X}}\mathbf{F} = [\mathbf{x}_0, \dots, \mathbf{x}_l, \dots, \mathbf{x}_{L-1}]$  denotes an  $N \times L$  equivalent time-domain signal matrix. In addition,  $\mathbf{h} = [h_0, h_1, \dots, h_{L-1}]^T$  denotes a  $L \times 1$  time-domain channel vector. Since  $\bar{\mathbf{h}} = \mathbf{F}\mathbf{h}$ , hence, the frequency-domain channel impulse response  $\bar{\mathbf{h}}$  lies in the time-delay spread domain.

Assume that a binary random vector  $\mathbf{g} = [g_0, g_1, \dots, g_{L-1}]^T$  denotes a taps' position indicator of sparse channel vector  $\mathbf{h}$  which is generated from a Gaussian mixture density (GMD) function as

$$\{\mathbf{h} \mid \mathbf{g}\} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}(\mathbf{g})), \quad (3)$$

where the covariance matrix  $\mathbf{R}(\mathbf{g})$  is determined by position indicator  $\mathbf{g}$ . For a better understanding, we take  $\mathbf{R}(\mathbf{g})$  to be diagonal element with  $[\mathbf{R}(\mathbf{g})]_{ll} = \sigma_l^2 = \sigma_1^2$  for  $l = 0, 1, \dots, L - 1$ , implying that  $\{h_l \mid g_l\}_{l=0}^{L-1}$  are independent with Gaussian distribution  $\{h_l \mid g_l\} \sim \mathcal{CN}(0, \sigma_1^2)$ . Assume that the position indices  $\{g_l\}_{l=0}^{L-1}$  are satisfied Bernoulli distribution with probability  $p_{1,l}$ ; then the probability of nonzero and zero channel taps of channel vector  $\mathbf{h}$  can be written as

$$\begin{aligned} h_l \neq 0 &\iff \Pr\{g_l = 1\} = p_{1,l}, \\ h_l = 0 &\iff \Pr\{g_l = 0\} = 1 - p_{1,l}, \end{aligned} \quad (4)$$

for  $l = 0, 1, \dots, L - 1$ . According to (4), one can easily find  $\|\mathbf{h}\|_0 = \|\mathbf{g}\|_1$ . In real communication systems, broadband channels are often described by sparse models [16, 17]. Hence, we choose  $\sigma_0^2 = \text{var}\{h_0 \mid g_0\} = 0$  and  $p_1 = \sum_{l=0}^{L-1} p_{1,l} \ll 1$ , so that  $\mathbf{h}$  has relatively few dominant channel taps. In other words, sparseness of channel vector  $\mathbf{h}$  depends on the probability  $p_1$  as shown in Figure 2. Smaller probability  $p_1$  implies sparser channel vector  $\mathbf{h}$  and vice versa.

The research objective of this paper is to estimate the sparse channel vector  $\mathbf{h}$  using received signal vector  $\mathbf{y}$  and

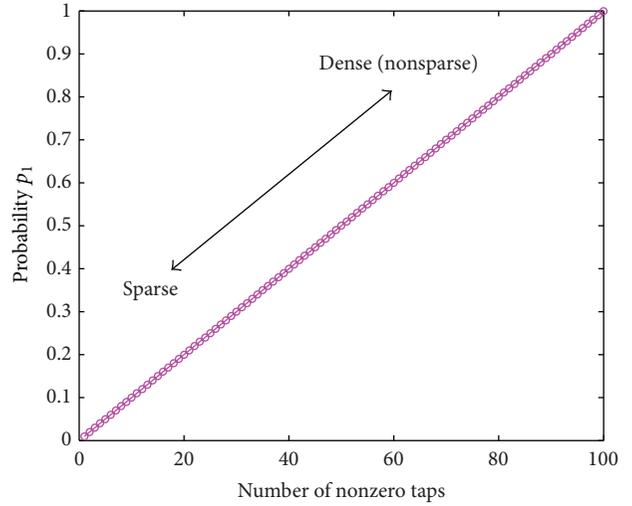


FIGURE 2: Sparseness of channel vector  $\mathbf{h}$  depends on the probability  $p_1$ .

training signal matrix  $\mathbf{X}$ . Hence, the system model can be assumed to satisfy distribution as

$$\begin{aligned} \begin{bmatrix} \mathbf{y} \\ \mathbf{h} \end{bmatrix} \mid \mathbf{g} &\sim \mathcal{CN}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}(\mathbf{g}) & \mathbf{X}\mathbf{R}(\mathbf{g}) \\ \mathbf{R}(\mathbf{g})\mathbf{X}^T & \mathbf{R}(\mathbf{g}) \end{bmatrix}\right) \\ &= \mathcal{CN}\left(\mathbf{0}, \begin{bmatrix} \mathbf{C}(\mathbf{g}) & \sigma_1^2\mathbf{X}\mathbf{I}_L \\ \sigma_1^2\mathbf{I}_L\mathbf{X}^T & \sigma_1^2\mathbf{I}_L \end{bmatrix}\right), \end{aligned} \quad (5)$$

where  $\mathbf{C}(\mathbf{g}) := \mathbf{X}\mathbf{R}(\mathbf{g})\mathbf{X}^T + \sigma_n^2\mathbf{I}_N = \sigma_1^2\mathbf{X}\mathbf{I}_L\mathbf{X}^T + \sigma^2\mathbf{I}_N$  is the covariance matrix of  $\{\mathbf{y} \mid \mathbf{g}\}$ . That is,  $\{\mathbf{y} \mid \mathbf{g}\} \sim \mathcal{CN}(\mathbf{0}, \sigma_1^2\mathbf{X}\mathbf{I}_L\mathbf{X}^T + \sigma^2\mathbf{I}_L)$ .

### 3. Compressive Sensing Based Bayesian Sparse Channel Estimation

In this section, compressive sensing based Bayesian sparse channel estimation is proposed in two steps: (1) *detect the position set of dominant channel taps* and (2) *then estimate sparse channel  $\bar{\mathbf{h}}$  using MMSE algorithm*. Obviously, how to find the dominant channel taps' position is a key technique with low-complexity Bayesian method for estimating sparse channels.

**3.1. Position Detection on Dominant Channel Taps.** According to the well-known Bayesian rules, the posterior of position indicator  $\mathbf{g}$  can be written as

$$P(\mathbf{g} \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid \mathbf{g})P(\mathbf{g})}{\sum_{\mathbf{g}' \in G} P(\mathbf{y} \mid \mathbf{g}')P(\mathbf{g}')}, \quad (6)$$

where  $G = \{0, 1\}^L$  denotes all of possible position index sets of channel taps as shown in Figure 3. Equation (6) implies that estimating  $\{P(\mathbf{g} \mid \mathbf{y})\}_{\mathbf{g} \in G}$  reduces to estimating  $\{P(\mathbf{y} \mid \mathbf{g})P(\mathbf{g})\}_{\mathbf{g} \in G}$ . Due to the extremely computational complexity in (6), the huge size of  $G$  makes it impractical to

compute  $P(\mathbf{g} | \mathbf{y})$  or  $\{P(\mathbf{y} | \mathbf{g}')P(\mathbf{g}')\}$  for all  $\mathbf{g}' \in G$  in the case of high-dimensional broadband channels. By considering sparse structure in channels, only posteriors of dominant taps' position are needed for sparse channel estimation. Assuming that the set  $G_*$  is responsible for position indicator of dominant channel taps, then the search space in  $G_*$  rather than  $G$  can be quite small and therefore practical to compute. Hence, the posteriors of dominant channel taps can be approximated by

$$P(\mathbf{g} | \mathbf{y}) \approx \frac{P(\mathbf{y} | \mathbf{g})P(\mathbf{g})}{\sum_{\mathbf{g}' \in G_*} P(\mathbf{y} | \mathbf{g}')P(\mathbf{g}')}, \quad (7)$$

for dominant channel set  $G_*$ . Hence, exploiting the dominant channel taps set  $G_*$  reduces to the search for  $\mathbf{g} \in G_*$  which only computes the dominant values of  $P(\mathbf{y} | \mathbf{g})P(\mathbf{g})$  in (7). First of all, the probability density function (PDF)  $P(\mathbf{y} | \mathbf{g})$  for position indicator  $\mathbf{g} \in G_*$  can be written as

$$P(\mathbf{y} | \mathbf{g}) = \frac{1}{\sqrt{(2\pi)^L \det(\mathbf{C}(\mathbf{g}))}} \exp\left(-\frac{1}{2}\mathbf{y}^T \mathbf{C}^{-1}(\mathbf{g})\mathbf{y}\right). \quad (8)$$

By transforming it in log-domain for convenience, then the position indicator (PI)  $\text{PI}(\mathbf{g}, \mathbf{y})$  can be given by

$$\begin{aligned} \text{PI}(\mathbf{g}, \mathbf{y}) &\triangleq \ln P(\mathbf{y} | \mathbf{g})P(\mathbf{g}) = \ln P(\mathbf{y} | \mathbf{g}) + \ln P(\mathbf{g}) \\ &= \ln P(\mathbf{y} | \mathbf{g}) + \|\mathbf{g}\|_0 \ln p_1 + (L - \|\mathbf{g}\|_0) \ln(1 - p_1) \\ &= -\frac{L}{2} \ln 2\pi - \frac{1}{2} \ln \det(\mathbf{C}(\mathbf{g})) - \frac{1}{2} \mathbf{y}^T \mathbf{C}^{-1}(\mathbf{g})\mathbf{y} \\ &\quad + \|\mathbf{g}\|_0 \ln \frac{p_1}{1 - p_1} + L \ln(1 - p_1), \end{aligned} \quad (9)$$

which is a metric of position indicator  $\mathbf{g}$ . According to  $\text{PI}(\mathbf{g}, \mathbf{y})$  in (9), one can easily find that the position indicator depends on received signal, channel length, position indicator, and probability of nonzero taps. Due to the positive exponent relationship  $P(\mathbf{g} | \mathbf{y}) = e^{\text{PI}(\mathbf{g}, \mathbf{y})}$ ,  $\text{PI}(\mathbf{g}, \mathbf{y})$  in (9) can also be considered as a measure function of  $P(\mathbf{g} | \mathbf{y})$  on dominant channel taps. However, it is still unfeasible to get the position information of channel in practical system without considering channel estimation. According to [18], the mathematical expectation of  $\text{PI}(\mathbf{g}, \mathbf{y})$  can be given by

$$\begin{aligned} E\{\text{PI}(\mathbf{g}, \mathbf{y})\} &= 2N + Lp_1(1 - p_1) \\ &\quad \times \left( \ln \left[ \left( \frac{\sigma_1^2}{\sigma^2} + 1 \right) \frac{(1 - p_1)}{p_1} \right] \right)^2. \end{aligned} \quad (10)$$

For a given pair  $\{\mathbf{g}', \mathbf{y}\}$ ,  $\text{PI}(\mathbf{g}', \mathbf{y})$  can be used to compare the mean  $E\{\text{PI}(\mathbf{g}', \mathbf{y})\}$  and standard deviation  $\sqrt{\text{var}\{\text{PI}(\mathbf{g}', \mathbf{y})\}}$  in order to get a rough evaluation of  $(\mathbf{g}', \mathbf{y})$ .

To reduce the search space in position set, we resort to an efficient method [13] to determine  $G_*$  as follows. The basic idea is that the position set  $\mathbf{g}$  of unknown channel yielding the dominant values of  $P(\mathbf{g} | \mathbf{y})$  is equivalent to the high probability of  $\text{PI}(\mathbf{g}, \mathbf{y})$ . The search starts with  $\mathbf{g} = \mathbf{0}$  and the

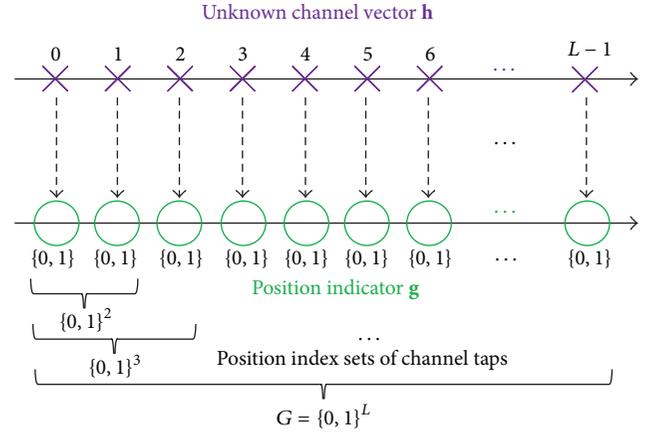


FIGURE 3: Graph illustration for all of possible position index sets of channel taps.

initial position set is set as  $G^{(0)}$ . If we change each element in  $\mathbf{g}$ , then it yields  $L$  position indicators. Consider all of position indicators in a set and refer it to  $G^{(1)}$ . The metrics  $\text{PI}(\mathbf{g}, \mathbf{y})$  for the  $L$  PI vectors in  $G^{(1)}$  are then computed by (9), and elements of  $G^{(1)}$  with the  $D$  largest value of the dominant channel tap are collected in  $G_*^{(1)}$ . For each possible dominant taps' set in  $G_*^{(1)}$ , all positions of a second nonzero tap are then considered, yielding  $\sum_{i=1}^D (L - i) = LD - D(D + 1)/2$  unique binary vectors to store in  $G^{(2)}$ . The  $\text{PI}(\mathbf{g}, \mathbf{y})$  for all possible vectors in  $G^{(2)}$  are then computed, and the elements of  $G^{(2)}$  with the  $D$  largest value are collected in  $G_*^{(2)}$ . Then for each candidate vector in  $G_*^{(2)}$ , all possibilities of a third dominant channel tap are considered, and those with the  $D$  largest channel taps are stored in  $G_*^{(3)}$ . The process continues until  $G_*^{(S)}$  is computed, where  $S$  can be chosen to make  $\text{Pr}(\|\mathbf{h}\|_0 > S)$  sufficiently small to exploit all of channel sparsity. Note that  $G_*^{(S)}$  constitutes the algorithm's final estimate of  $G_*$  and later we denote  $\widehat{G}_*$  as the final estimate. For better understanding of the PI update of dominant channel taps, an intuitive example is given in Figure 4, where the length of position indicator  $\mathbf{g}$  is set as  $L = 5$ ; the number of largest value of PI is chosen as  $D = 1$ , and the maximum number of nonzero taps is set as  $S = 3$ .

For use with the aforementioned Bayesian matching pursuit (BMP) algorithm, we consider a fast metric update which computes the change in  $\text{PI}(\cdot)$  that results from the activation of a position of nonzero tap. More precisely, if we denote by  $\mathbf{g}_l$  the vector identical to  $\mathbf{g}$  except for the  $l$ th coefficient, which is active in  $\mathbf{g}_l$  but inactive in  $\mathbf{g}$  (i.e.,  $[\mathbf{g}_l]_l = 1$  and  $[\mathbf{g}]_l = 0$ ), then it is defined as

$$d_l(\mathbf{g}) \triangleq \text{PI}(\mathbf{g}_l, \mathbf{y}) - \text{PI}(\mathbf{g}, \mathbf{y}), \quad (11)$$

to track the change of active positions. Note that  $\text{PI}(\mathbf{g}, \mathbf{y})$  at the initial step is set as

$$\text{PI}(\mathbf{0}, \mathbf{y}) = -\frac{L}{2} \ln 2\pi - \frac{N}{2} \ln \sigma_1^2 - \frac{1}{2\sigma^2} \|\mathbf{y}\|_2^2 + L \ln(1 - p_1), \quad (12)$$

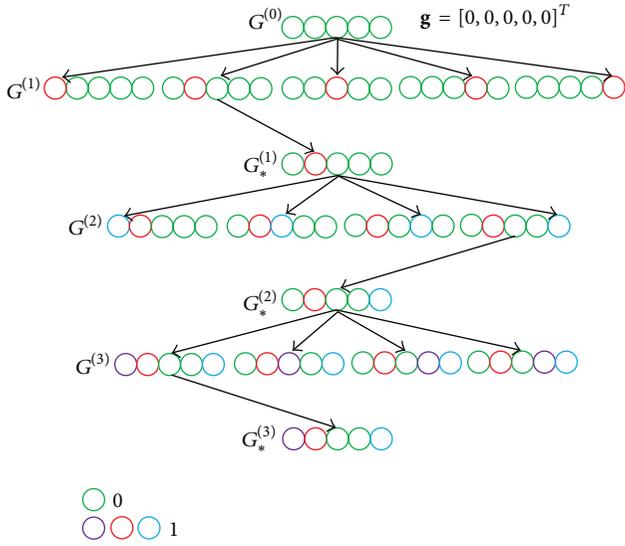


FIGURE 4: An intuitive example of position set selection on dominant channel taps, where the green circle denotes zero while the other colored circles denote one.

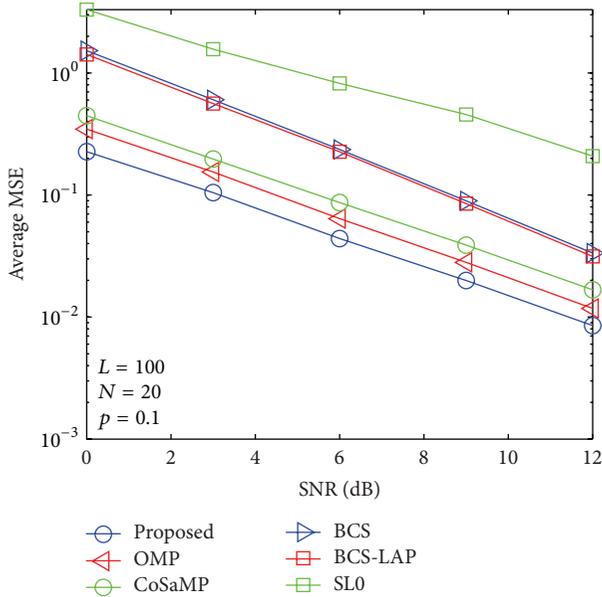


FIGURE 5: Average MSE performance versus SNR when  $p_1 = 0.1$  and  $N = 20$ .

via (9) and the fact that  $\mathbf{C}(\mathbf{0}) = \sigma_1^2 \mathbf{I}_L$ . To obtain the fast PI update, we start with the property that, for any  $l$  and  $\mathbf{g}$ ,

$$\mathbf{C}(\mathbf{g}_l) = \mathbf{C}(\mathbf{g}) + \sigma_1^2 \mathbf{x}_l \mathbf{x}_l^T, \quad (13)$$

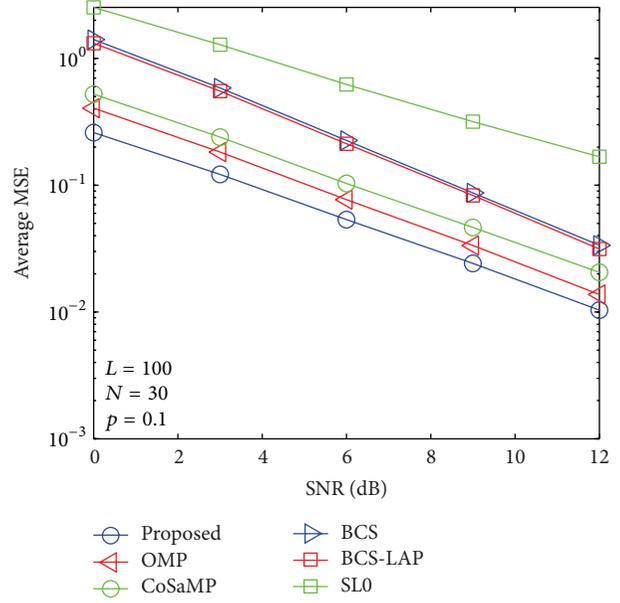


FIGURE 6: Average MSE performance versus SNR when  $p_1 = 0.1$  and  $N = 30$ .

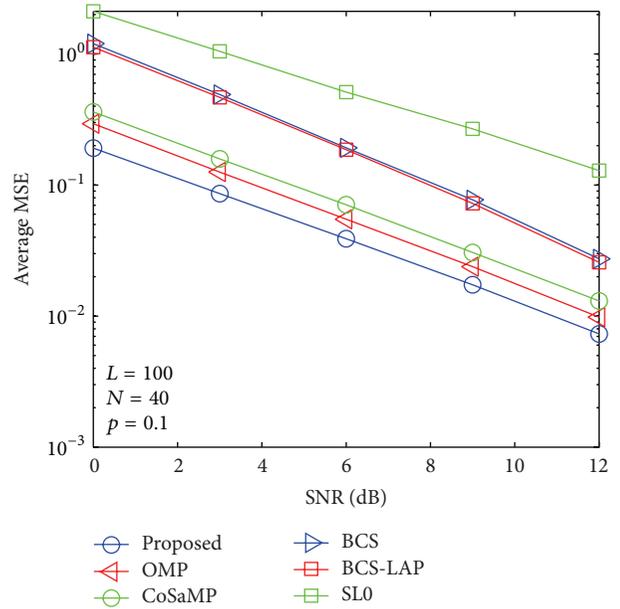


FIGURE 7: Average MSE performance versus SNR when  $p_1 = 0.1$  and  $N = 40$ .

for which the matrix inversion lemma implies

$$\mathbf{C}^{-1}(\mathbf{g}_l) = \mathbf{C}^{-1}(\mathbf{g}) - \sigma_1^2 \beta_l \mathbf{b}_l \mathbf{b}_l^T, \quad (14)$$

$$\mathbf{C}^{-1}(\mathbf{g}) = \frac{1}{\sigma_2^2} \mathbf{I}_N - \sigma_1^2 \sum_{i=1}^p \beta^{(i)} \mathbf{b}^{(i)} (\mathbf{b}^{(i)})^T \quad (15)$$

$$\mathbf{b}_l \triangleq \mathbf{C}^{-1}(\mathbf{g}) \mathbf{x}_l = \frac{1}{\sigma_2^2} \mathbf{x}_l - \sigma_1^2 \sum_{i=1}^p \beta^{(i)} \mathbf{b}^{(i)} (\mathbf{b}^{(i)})^T \mathbf{x}_l, \quad (16)$$

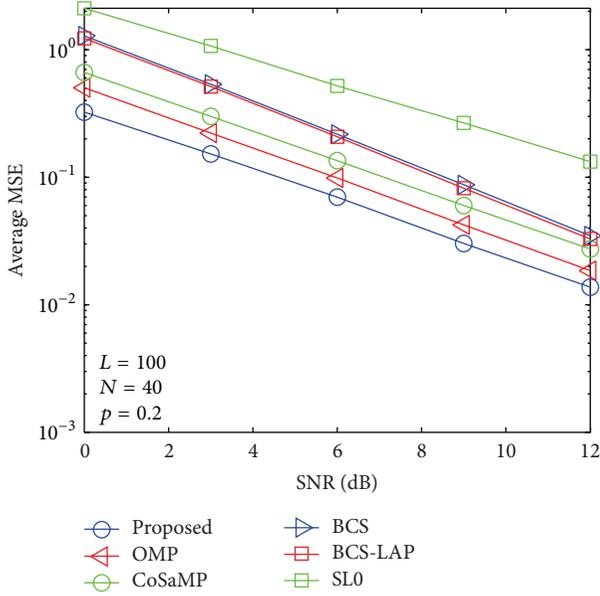


FIGURE 8: Average MSE performance versus SNR when  $p_1 = 0.2$  and  $N = 40$ .

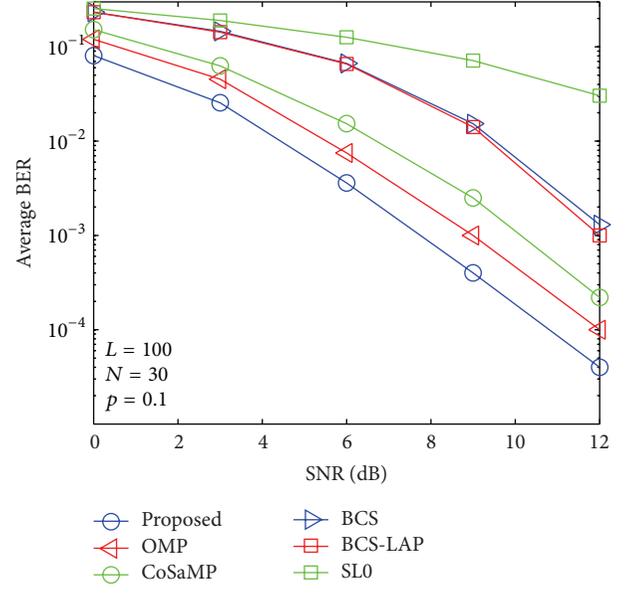


FIGURE 10: Average BER performance versus SNR when  $p_1 = 0.1$  and  $N = 30$ .

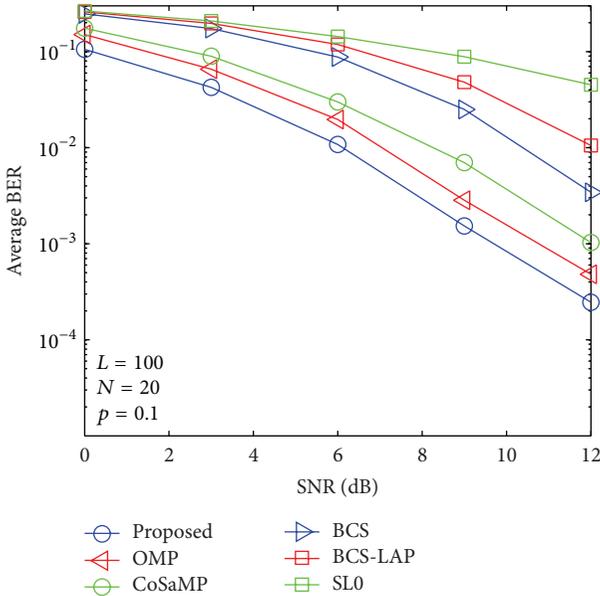


FIGURE 9: Average BER performance versus SNR when  $p_1 = 0.1$  and  $N = 20$ .

where  $\mathbf{b}_l := \mathbf{C}^{-1}(\mathbf{g})\mathbf{x}_l$  and  $\beta_l := (1 + \sigma_1^2 \mathbf{x}_l^T \mathbf{b}_l)^{-1}$ . Notice that the cost of computing  $\beta_l$  in (14) is  $\mathcal{O}(LN^2)$  if standard matrix

multiplication is used [13]. According to previous analysis, we can get

$$\begin{aligned} \mathbf{y}^T \mathbf{C}^{-1}(\mathbf{g}_l) \mathbf{y} &= \mathbf{y}^T (\mathbf{C}^{-1}(\mathbf{g}) - \sigma_1^2 \beta_l \mathbf{b}_l \mathbf{b}_l^T) \mathbf{y} \\ &= \mathbf{y}^T \mathbf{C}^{-1}(\mathbf{g}) \mathbf{y} - \sigma_1^2 \beta_l (\mathbf{y}^T \mathbf{b}_l)^2, \end{aligned} \quad (17)$$

$$\begin{aligned} \text{Indet}(\mathbf{C}(\mathbf{g}_l)) &= \text{Indet}(\mathbf{C}(\mathbf{g}) + \sigma_1^2 \mathbf{x}_l \mathbf{x}_l^T) \\ &= \ln \left[ (1 + \sigma_1^2 \mathbf{x}_l^T \mathbf{C}^{-1}(\mathbf{g}) \mathbf{x}_l) \det(\mathbf{C}(\mathbf{g})) \right] \\ &= \text{Indet}(\mathbf{C}(\mathbf{g})) - \ln \beta_l, \end{aligned} \quad (18)$$

$$\begin{aligned} \|\mathbf{g}_l\|_0 \ln \frac{p_1}{1-p_1} &= (\|\mathbf{g}\|_0 + 1) \ln \frac{p_1}{1-p_1} \\ &= \|\mathbf{g}\|_0 \ln \frac{p_1}{1-p_1} + \ln \frac{p_1}{1-p_1}, \end{aligned} \quad (19)$$

which, combined with (5), yield

$$\text{PI}(\mathbf{g}_l, \mathbf{y}) = \text{PI}(\mathbf{g}, \mathbf{y}) + \underbrace{\frac{1}{2} \ln \beta_l + \frac{\sigma_1^2}{2} \beta_l (\mathbf{y}^T \mathbf{b}_l)^2}_{d_l(\mathbf{g})} + \ln \frac{p_1}{1-p_1}. \quad (20)$$

In summary,  $d_l(\mathbf{g})$  in (18) quantifies the change in  $\text{PI}(\cdot)$  due to the activation of the  $l$ th position of  $\mathbf{g}$ .

Please note that the cost of computing  $\{\beta_l\}_{l=0}^{L-1}$  via  $\mathbf{b}_l := \mathbf{C}^{-1}(\mathbf{g})\mathbf{x}_l$  and  $\beta_l := (1 + \sigma_1^2 \mathbf{x}_l^T \mathbf{b}_l)^{-1}$  is  $\mathcal{O}(LN^2)$ , if standard matrix multiplication is used. As we described, the complexity of this operation can be made linear in  $N$  by exploiting the

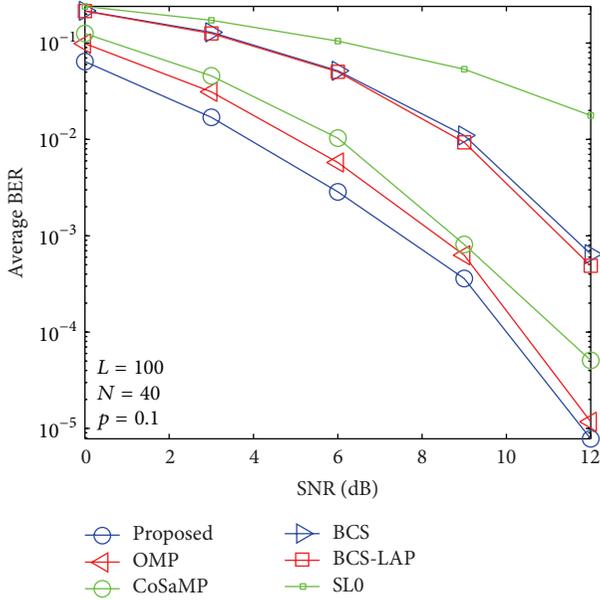


FIGURE 11: Average BER performance versus SNR when  $p_1 = 0.1$  and  $N = 40$ .

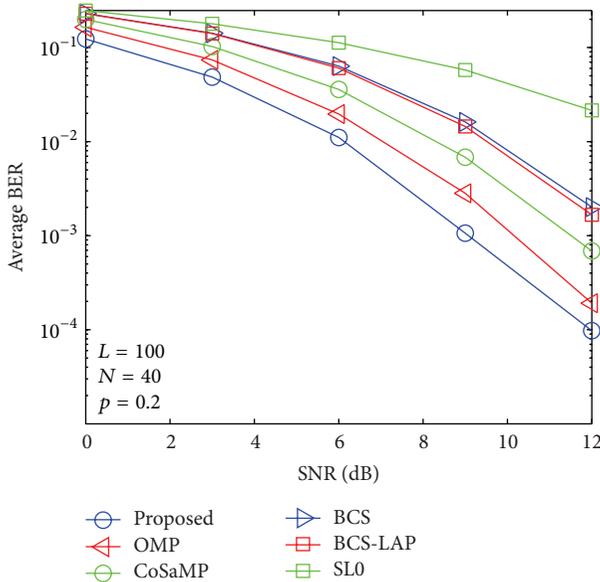


FIGURE 12: Average BER performance versus SNR when  $p_1 = 0.2$  and  $N = 40$ .

structure of  $\mathbf{C}^{-1}(\mathbf{g})$ . Say that  $\mathbf{t} = [t_1, t_2, \dots, t_p]^T$  contains the indices of active elements in  $\mathbf{g}$ . Then from (14), we can get

$$\mathbf{C}^{-1}(\mathbf{g}) = \frac{1}{\sigma^2} \mathbf{I}_N - \sigma_1^2 \sum_{i=1}^p \beta^{(i)} \mathbf{b}^{(i)} \underbrace{\mathbf{b}^{(i)T} \mathbf{x}_i}_{:=c_i^{(i)}} \quad (21)$$

when activating the  $l$ th position in  $\mathbf{g}$ . The key observation is that the coefficients  $\{c_i^{(i)}\}_{i=0}^{L-1}$  need only to be computed once, that is, when index  $t_i$  is active. Furthermore,  $\{c_i^{(i)}\}_{i=0}^{L-1}$  only need to be computed for surviving indices  $t_i$ . According

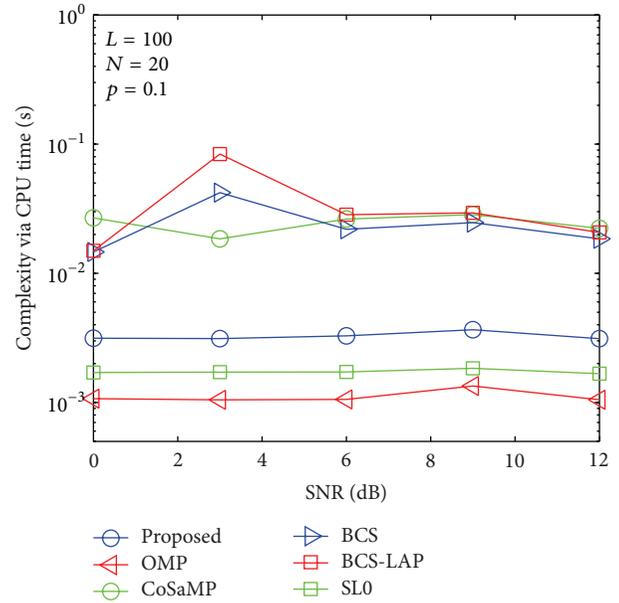


FIGURE 13: Computational complexity comparison via CPU time when  $p_1 = 0.1$  and  $N = 20$ .

to previous analysis in (20), the number of multiplications required by the algorithm is  $\mathcal{O}(\text{LNPD})$  [13]. Moreover, the complexity of the proposed algorithm could be reduced if the smaller  $D$  is adopted.

3.2. MMSE for Estimating Values of Dominant Channel Taps. By utilizing the dominant taps' posteriors, the sparse channel can be estimated readily by MMSE algorithm as

$$\begin{aligned} \tilde{\mathbf{h}} &= E\{\mathbf{h} | \mathbf{y}\} \\ &= \sum_{\mathbf{g} \in G} P(\mathbf{g} | \mathbf{y}) E\{\mathbf{h} | \mathbf{y}, \mathbf{g}\} \\ &\approx \sum_{\mathbf{g} \in G_*} P(\mathbf{g} | \mathbf{y}) E\{\mathbf{h} | \mathbf{y}, \mathbf{g}\}. \end{aligned} \quad (22)$$

According to the above introduction, compressive sensing based Bayesian sparse channel estimation could be implement by (20)–(22) with high estimation performance and low complexity.

## 4. Computer Simulations

In this section, the proposed BSCE estimator adopts 1000 independent Monte Carlo runs for averaging. The length of channel vector  $\mathbf{h}$  is set as  $N = 100$ . Values of dominant channel taps follow Gaussian distribution and their positions are randomly allocated within the length of  $\mathbf{h}$  which is subjected to  $E\{\|\mathbf{h}\|_2^2 = 1\}$ . The received signal-to-noise ratio (SNR) is defined as  $10 \log(E_b/\sigma_n^2)$ , where  $E_b = 1$ .

The proposed method is compared to five conventional sparse channel estimation methods using algorithms OMP [19], CoSaMP [9], BCS [14], BCS-LAP [15], and SL0 [20]. It

TABLE 2: Simulation parameters.

Data modulation		BPSK
Number of subcarriers		$N_d = 256$
Transmitter	Number of pilot symbols	$N \in \{20, 30, 40\}$
	Length of CP	$N_g = 16$
	Pilot sequence	Random Gaussian sequence
Fading		Frequency-selective block
Channel model	Number of channel taps	$L = 100$
	Prob. of nonzero taps	$p \in \{0.1, 0.2\}$
	Power delay profile	Random Gaussian
Receiver	Channel estimation	BSCE
	Data detection	Zero forcing

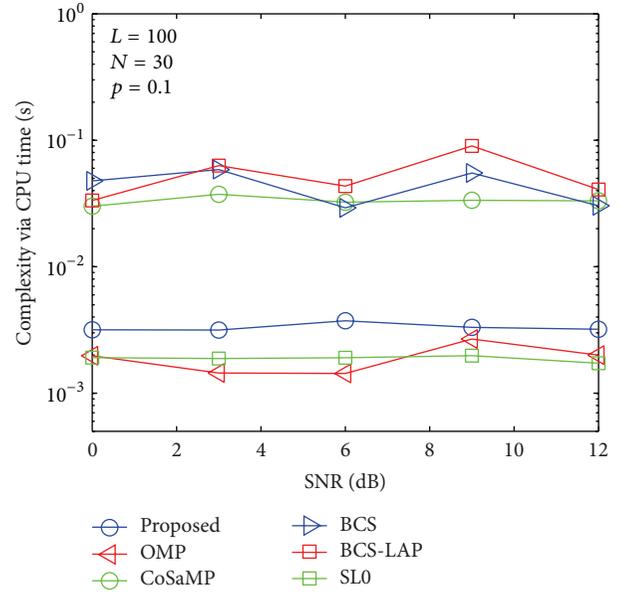
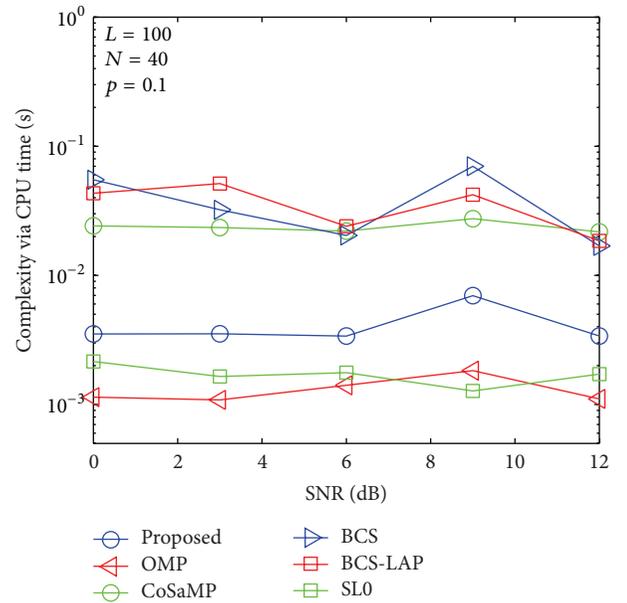
was worth noting that these simulation parameters were chosen in accordance with detailed communication environment in this paper. The stopping error criteria threshold is set as  $10^{-4}$  for all algorithms in Monte Carlo computer simulations. The initial noise variance for BSC and BSC-LAP is set as  $\text{var}(\mathbf{y})/10$ , where  $\text{var}(\mathbf{y}) = (1/(N-1) \sum_{n=1}^N (y_n - \hat{y}))^2$  denotes standard derivation and  $\hat{y} = 1/N \sum_{n=1}^N y_i$ . In addition, the Laplace prior for BCS-LAP is computed automatically which was suggested in [15]. The parameters of FBMP algorithm were initialized as  $\lambda_1 = 0.01$ ,  $\mu_1 = 0$ ,  $\sigma^2 = 0.05$ , and  $\sigma_1^2 = 2$ . Computer simulation parameters are listed in Table 2.

**4.1. MSE versus SNR.** The estimation performance is evaluated by average mean square error (MSE) standard which is defined as

$$\text{MSE} \{\tilde{\mathbf{h}}\} = E \|\mathbf{h} - \tilde{\mathbf{h}}\|_2^2, \quad (23)$$

where  $E\{\cdot\}$  denotes expectation operator and  $\mathbf{h}$  and  $\hat{\mathbf{h}}$  are the actual channel vector and its channel estimator, respectively. In Figures 5, 6, 7, and 8, we compare the average MSE performance of the proposed channel estimator with traditional sparse channel estimators with respect to different channel sparseness,  $p_1 = 0.1$  and  $p_1 = 0.2$ . As the four figures show, our proposed method can achieve better estimation performance than conventional methods. The lower bound is given by least square (LS) method (oracle) which utilized the channel position information. In this figure, it is easily found that the proposed method obtained lower MSE performance than conventional methods. In other words, if the proposed estimator is applied in data detection, smaller BER performance can be achieved when comparing with conventional methods.

**4.2. BER versus SNR.** By using the above channel estimators, signal transmission performances are evaluated as shown in Figures 9, 10, 11, and 12. From the four figures, average BER performance curves are depicted with respect to SNR for binary phase shift keying (BPSK) data. We can see that the BER performance of the proposed method is more close to lower bound which is given by ideal channel estimator whose nonzero taps' positions are known. Here, only low signal

FIGURE 14: Computational complexity comparison via CPU time when  $p_1 = 0.1$  and  $N = 30$ .FIGURE 15: Computational complexity comparison via CPU time when  $p_1 = 0.1$  and  $N = 40$ .

modulation was considered for BER evaluation. It is very easy to predict that our proposed method could improve BER performance in case of high signal modulation.

**4.3. Complexity Evaluation.** To compare the computational complexity of the proposed method with other methods, CPU time is adopted for evaluation standard as shown in Figures 13, 14, 15, and 16. It is worth mentioning that although the CPU time is not an exact measure of complexity, it can give us a rough estimation of computational complexity. Our

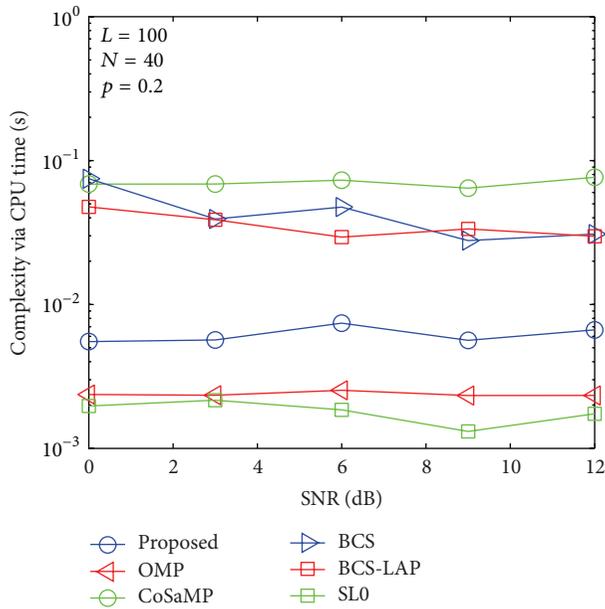


FIGURE 16: Computational complexity comparison via CPU time when  $p_1 = 0.2$  and  $N = 40$ .

simulations are performed in MATLAB 2012 environment using a 2.90 GHz Intel i7 processor with 8 GB of memory and under Microsoft Windows 8 (64 bit) operating system. For comprehensive comparing between our proposed method and other methods in different length of training signal and different channel sparsity, we simulate their comparison results in Figures 13–16. As the four figures shown, the complexity of the proposed method is close to OMP and SL0-based methods and lower than CoSaMP, BCS, and BCS-LAP based methods. It is well known that the complexity of OMP and SL0 is very low on sparse channel estimation [10, 22]. Hence, comparing with traditional methods, our proposed method can achieve better estimation performance and low complexity.

## 5. Conclusion

Traditional sparse channel estimation methods are vulnerable to noise and column coherence interference in training matrix. Their primary aim is an attempt to exploit sparse structure information without a report of posterior channel uncertainty. To improve the estimation performance, fast Bayesian matching pursuit algorithm with application to sparse channel estimation has not only exploited the channel sparsity but also mitigated the unexpected inferences in training matrix. In addition, the proposed method has revealed potential ambiguity among multiple channel estimators that are ambiguous due to observation of noise or correlation among columns in the training signal. Computer simulation results have showed that proposed method improved the estimation performance with comparable computational complexity when comparing with traditional methods.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper. The authors of the paper do not have a direct financial relation that might lead to a conflict of interests for any of the authors.

## References

- [1] L. Dai, Z. Wang, and Z. Yang, "Compressive sensing based time domain synchronous OFDM transmission for vehicular communications," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 460–469, 2013.
- [2] L. Dai, Z. Wang, and Z. Yang, "Next-generation digital television terrestrial broadcasting systems: key technologies and research trends," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 150–158, 2012.
- [3] N. Czink, X. Yin, H. Özcelik, M. Herdin, E. Bonek, and B. H. Fleury, "Cluster characteristics in a MIMO indoor propagation environment," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1465–1475, 2007.
- [4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [5] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [6] G. Z. Karabulut and A. Yongaçoglu, "Sparse channel estimation using orthogonal matching pursuit algorithm," in *Proceedings of the IEEE 60th Vehicular Technology Conference (VTC '04)*, pp. 3880–3884, Los Angeles, Calif, USA, September 2004.
- [7] G. Tauböck, F. Hlawatsch, D. Eiwien, and H. Rauhut, "Compressive estimation of doubly selective channels in multicarrier systems: leakage effects and sparsity-enhancing processing," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 255–271, 2010.
- [8] G. Gui, Q. Wan, W. Peng, and F. Adachi, "Sparse multipath channel estimation using compressive sampling matching pursuit algorithm," in *Proceedings of the 7th IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS '10)*, pp. 10–14, Kaohsiung, Taiwan, May 2010.
- [9] D. Needell and J. A. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.
- [10] G. Gui, Q. Wan, and W. Peng, "Fast compressed sensing-based sparse multipath channel estimation with smooth L0 algorithm," in *Proceedings of the 3rd International Conference on Communications and Mobile Computing (CMC '11)*, pp. 242–245, Qingdao, China, April 2011.
- [11] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "Complex-valued sparse representation based on smoothed  $\ell^0$  norm," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 3881–3884, April 2008.
- [12] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: a new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.
- [13] P. Schniter, L. C. Potter, and J. Ziniel, "Fast bayesian matching pursuit," in *Proceedings of the Information Theory and Applications Workshop (ITA '08)*, pp. 326–333, San Diego, Calif, USA, February 2008.

- [14] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.
- [15] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using laplace priors," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, 2010.
- [16] C. Carbonelli and U. Mitra, "Clustered channel estimation for UWB multiple antenna systems," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 970–981, 2007.
- [17] L. Dai, Z. Wang, and Z. Yang, "Spectrally efficient time-frequency training OFDM for mobile large-scale MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 251–263, 2013.
- [18] P. Schniter, L. C. Potter, and J. Ziniel, "Fast Bayesian matching pursuit: model uncertainty and parameter estimation for sparse linear models," OSU ECE Technical Report, 2009.
- [19] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [20] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed  $\ell^0$  norm," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 289–301, 2009.
- [21] F. Adachi, D. Garg, S. Takaoka, and K. Takeda, "Broadband CDMA techniques," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 8–18, 2005.
- [22] G. Gui, A. Mehbodniya, Q. Wan, and F. Adachi, "Sparse signal recovery with OMP algorithm using sensing measurement matrix," *IEICE Electronics Express*, vol. 8, no. 5, pp. 285–290, 2011.

## Research Article

# Code-Time Diversity for Direct Sequence Spread Spectrum Systems

A. Y. Hassan<sup>1,2</sup>

<sup>1</sup> Benha Faculty of Engineering, Egypt

<sup>2</sup> Faculty of Engineering, Northern Border University, Saudi Arabia

Correspondence should be addressed to A. Y. Hassan; [ayahiahassan@gmail.com](mailto:ayahiahassan@gmail.com)

Received 23 January 2014; Accepted 2 March 2014; Published 10 April 2014

Academic Editors: Y. Mao and Z. Zhou

Copyright © 2014 A. Y. Hassan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Time diversity is achieved in direct sequence spread spectrum by receiving different faded delayed copies of the transmitted symbols from different uncorrelated channel paths when the transmission signal bandwidth is greater than the coherence bandwidth of the channel. In this paper, a new time diversity scheme is proposed for spread spectrum systems. It is called code-time diversity. In this new scheme,  $N$  spreading codes are used to transmit one data symbol over  $N$  successive symbols interval. The diversity order in the proposed scheme equals to the number of the used spreading codes  $N$  multiplied by the number of the uncorrelated paths of the channel  $L$ . The paper represents the transmitted signal model. Two demodulators structures will be proposed based on the received signal models from Rayleigh flat and frequency selective fading channels. Probability of error in the proposed diversity scheme is also calculated for the same two fading channels. Finally, simulation results are represented and compared with that of maximal ration combiner (MRC) and multiple-input and multiple-output (MIMO) systems.

## 1. Introduction

Diversity techniques are used when the channel is in a deep fade. If several replicas of the same information signal are transmitted over independent fading channels, the probability that all signal components will fade simultaneously is reduced considerably. There are different ways in which we can provide the receiver with  $L$  independent fading replicas of the same information signal.

Frequency diversity is a diversity method where the information signal is transmitted on  $L$  carriers. The separation between the successive carriers equals to or exceeds the coherent bandwidth of the channel. Orthogonal frequency division multiplexing (OFDM) transmission is the famous technique that exploits frequency diversity to achieve high data rate and low bit error rate in frequency selective channels [1–3]. OFDM suffers from intercarrier interference (ICI) due to frequency offsets, symbol timing error, and channel estimation errors [4–6]. OFDM systems also suffer from high peak to average power ratio (PAPR) [7, 8].

Another commonly used method for achieving diversity employs multiple antennas. Multiple transmitting antennas are used to transmit the same information signal and multiple receiving antennas are used to receive the independently fading replicas of the transmitted signal through uncorrelated fading paths. A comparative study of space diversity techniques in mobile radio is shown in [9]. Multiple-input and multiple-output (MIMO) system is a well-known system that exploits the antenna diversity to enhance the bit error rate and the channel capacity in fading environments [10, 11].

Time diversity is another diversity method where  $L$  independent fading version of the same information signal is achieved by transmitting the signal in  $L$  different time slots. The separation between the successive time slots equals to or exceeds the coherence time of the channel. Time diversity is used in modern communication system through interleaving of the transmitted symbols and through the using of channel codes [12–14].

Some systems use a combination of more than one diversity technique to enhance their performance in fading

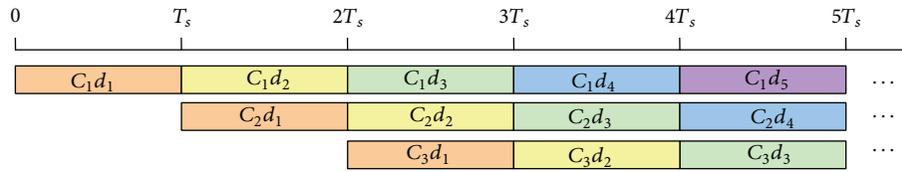


FIGURE 1: The code-time diversity system with 3 orthogonal spreading codes.

channels such as space-time (ST) coding in MIMO systems [15, 16], which use the time and space diversities through encoding the transmitted symbols using space time codes and transmitting the encoded symbols by different antennas in the transmitter. This technique allows the data symbol to be transmitted more than one time through different symbol period and it arrives at the receiver's antennas through different spatial paths. MIMO-OFDM system is another example of multidiversity system where time, frequency, and space diversities are used to enhance the performance of the data transmission over wireless faded channels [17–20]. In MIMO-OFDM system, the symbols are encoded by using space-time-frequency (STF) block codes. The encoded symbol is transmitted more than once over different periods and carrier frequencies using different transmitting antennas. The uncorrelated fading gains that come from the uncorrelated spatial paths, the different transmitting time slots, and the different transmitting carrier achieve the diversity gain at the receiver. This system has large diversity gain and better performance than space-time coded MIMO systems.

The systems that use space diversity such as MIMO systems have the disadvantage of using more than one antenna at the transmitter and the receiver. Multiple antennas need multiple RF drivers (power amplifier at the transmitter and low noise amplifier at the receiver) and this complicates the transmitter and the receiver structure. The spacing between the antennas should be large enough to have uncorrelated fading paths and to reduce the cross correlation and interference between the antennas. The MIMO systems consume more power than single-input and single-output (SISO) systems and they have short lifetime batteries for mobile units. Powerful DSP unit is required for MIMO transceivers because ST and STF encoder and decoder need complex computations.

In this paper, we propose a new diversity technique for SISO spread spectrum systems that can achieve a diversity gain like that of the MIMO system but with a single transmitting antenna and a single receiving antenna. The MIMO system has diversity gain with two degrees of freedom (the number of the transmitting and receiving antennas). The proposed diversity scheme has a diversity gain with two degrees of freedom too. Although the proposed system has a single transmitting antenna and a single receiver antenna, the using of  $N$  spreading codes and  $L$  uncorrelated propagation paths achieves the diversity gain. Section 2 discusses the idea of code-time diversity in detail. In Section 3 the signal model and the new transmit diversity scheme are represented. The receiver structure of the proposed system is shown in Section 4 with the calculations of the probability of error in

the received data. Section 5 shows the simulation results and some implementation issues.

## 2. Code-Time Diversity

In space-time MIMO system, the encoded data symbol is transmitted more than one time through different symbol periods using different transmitting antennas. The transmission of the symbols through independent time slots lets the received symbol to have independent fading gains. The probability of receiving the transmitted symbol with faded gains through the successive time slots is reduced significantly. Also the usage of different antennas allows the transmitted symbol to have independent propagation paths from the transmitter to the receiver and to have independent fading gain through each path. So they are the different time slots and propagation paths that play the main role in the diversity gain enhancement of the time-space MIMO system.

In the proposed code-time diversity technique, we use the same concept of time and space diversities but through another procedure. During each symbol period, the current data symbol and the previous  $(N - 1)$  ones are dispersed in frequency using  $N$  independent spreading codes sequences for each data symbol. The used spreading codes are taken from a set of  $N$  orthogonal codes. The dispersed symbols are added together and transmitted using a single antenna. The orthogonality between the used spreading sequences prevents the interference among the transmitted symbols. The same procedure is repeated for each symbol period so that each symbol can be transmitted  $N$  times through  $N$  successive symbol periods using, each time, a different spreading code from the set of  $N$  orthogonal codes. Figure 1 shows an example of how the modulated data symbols are transmitted three times at three successive symbol periods using three different orthogonal spreading codes.

By this way, time diversity is achieved and the probability of having  $N$  faded gains during  $N$  successive symbols periods is reduced considerably.

The space diversity is achieved by controlling the bandwidth of the dispersed symbols to be greater than the coherent bandwidth of the used wireless channel. This allows uncorrelated multipath propagation from the transmitter to the receiver. The using of direct sequence spread spectrum (DSSS) increases the information message bandwidth by a factor equal to the process gain of the spreading process, which is equivalent to the ratio between the data symbol period and the spreading code chip period. By controlling the process gain, the bandwidth of the spread signal can be equal

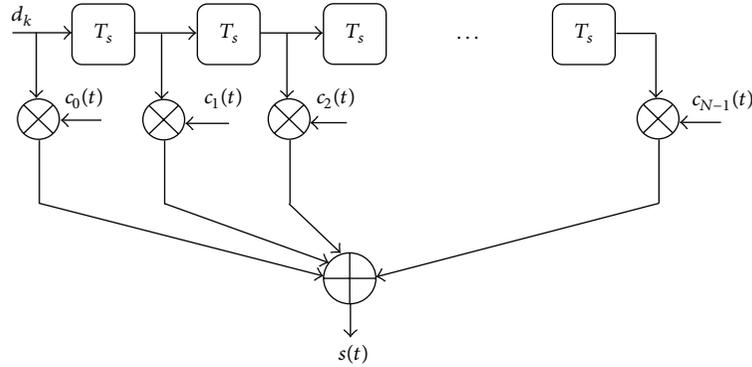


FIGURE 2: The code-time diversity DSSS modulator.

to multiples of the coherent bandwidth of the channel. The number of the uncorrelated paths that can be appeared from the transmitter to the receiver is given by

$$L = \left\lceil \frac{\text{Bandwidth of DSSS signal}}{\text{Channel coherent bandwidth}} + 0.5 \right\rceil. \quad (1)$$

By the same way, the probability of having spatial faded gains through the uncorrelated propagation paths  $L$  is reduced significantly.

The proposed code-time diversity system has a diversity gain with two degrees of freedoms as the space-time MIMO system. Although the diversity gain in the time-space MIMO system is depending on the number of the antennas in the transmitter  $N_t$  and the receiver  $N_r$ , the diversity gain in the code-time diversity system is depending on the number of the used spreading codes  $N$  (which equals to the number of time slots through which each symbol will be repeated) and the number of the uncorrelated propagation paths  $L$ . The proposed diversity scheme uses one antenna and one RF interface unit in the transmitter and in the receiver. However, the code-time diversity seems to have diversity gain similar to that of the time-space MIMO system, and the code-time diversity uses a spread signal with higher bandwidth than the bandwidth of the transmitted signal in the time-space MIMO system. In other words, the increase in the signal bandwidth of the proposed code-time diversity system is the cost that should be paid to improve the diversity gain using a simplified hardware of single antenna and single RF interface in the transmitter and in the receiver. The code-time diversity system needs no space-time codes, and it depends only on the orthogonality between the spreading codes.

### 3. Transmitted Signal Model of Code-Time Diversity DSSS

The code-time diversity system is based on the transmission of the data by using different orthogonal spreading codes at different transmission periods. At each transmission period, the transmitted signal is the summation between the current spread symbol and the previous  $(N-1)$  spread symbols where

$N$  is the number of the used orthogonal spreading codes. Equation (2) shows the  $k$ th symbol transmission:

$$s_k(t) = \sum_{n=0}^{N-1} d_{k-n} c_n(t - kT_s), \quad (2)$$

where  $d_{k-n}$  is the modulated data symbol and  $c_n(t)$  is the spreading code sequence. The used modulation method may be BPSK or  $M$ -QAM. The spreading codes are assumed to be orthonormal through the symbol period  $T_s$ :

$$\int_0^{T_s} c_i(t) \cdot c_j(t) \cdot dt = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \quad (3)$$

Without any loss of generality, the code period is assumed to be equal to the symbol period:

$$T_s = N_c * T_c. \quad (4)$$

$N_c$  is the number of chips on one code period and  $T_c$  is the chip period. So,

$$c_i(t - kT_s) = c_i(t). \quad (5)$$

The transmitted signal of a packet of  $K$  symbols is illustrated in (6) and Figure 2 shows the modulator structure of code-time diversity system:

$$s(t) = \sum_{k=0}^{K-1} s_k(t - kT_s) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} d_{k-n} c_n(t - kT_s). \quad (6)$$

The modulated signal in (6) is transmitted to the channel through single RF interface module and single transmitting antenna.

### 4. Received Signal Model of Code-Time Diversity DSSS Signal in Rayleigh Fading Channel and the Proposed Demodulator Building

4.1. Flat Fading Rayleigh Channel. The impulse response of the flat fading channel is shown in (7). Quasistatic channel is

assumed where the fading gain is fixed during one symbol period and it is changed randomly from one symbol to another:

$$h(t) = \alpha \delta(t), \quad (7)$$

where  $\alpha$  is a complex Gaussian random variable with zero mean and  $\sigma_\alpha^2$  variance. In flat fading channel, the transmitted signal travels from the transmitter to the received through unresolvable propagation paths. Therefore, all frequency components of the signal will experience the same magnitude of fading. In our proposed diversity scheme, the flat fading case looks like the MISO system where multiple antennas transmit the modulated symbols and a single antenna at the receiver picks them up. According to our signal model, the received signal at the demodulator input is

$$r(t) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \alpha_k d_{k-n} c_n(t - kT_s) + w(t), \quad (8)$$

where  $w(t)$  is a sample function of white Gaussian noise process with zero mean and  $\sigma_w^2$  variance.  $\alpha_k$  is the channel random gain at the  $k$ th symbol period. The proposed demodulator for the code-time diversity system consists of three parts. The first part is a bank of correlators that correlate the received signal with the  $N$  spreading codes. The  $n$ th correlator correlates the received signal with the  $n$ th spreading code  $c_n(t)$  through one symbol period. The  $n$ th correlator output at the  $k$ th symbol period is shown in (9a) and (9b):

$$x_n(kT_s) = \int_0^{T_s} r(t) \cdot c_n(t - kT_s) \cdot dt = \alpha_k d_{k-n} + v_{kn}, \quad (9a)$$

$$v_{kn} = \int_0^{T_s} w(t) \cdot c_n(t - kT_s) \cdot dt, \quad (9b)$$

where  $v_{kn}$  is a Gaussian random variable with zero mean and  $\sigma_w^2$  variance. The second part of the proposed demodulator is the combiner. Maximal ration combiner (MRC) is used but with some modifications. In the proposed MRC, the outputs of the  $N$  correlators are multiplied by the conjugate of the channel gain  $\alpha_k$ , which is estimated in the receiver. The multiplications results will independently be delayed according to the spreading code index. The output of the correlation with the code sequence  $c_n(t)$  is delayed  $((N-1)-n)$  symbol periods. The delayed samples are finally added to form a single input to the detector. The output of the proposed MRC in the  $z$ -domain can be represented by

$$Y(z) = \beta_0 X_0(z) z^{-(N-1)} + \beta_1 X_1(z) z^{-(N-1)+1} + \beta_2 X_2(z) z^{-(N-1)+2} + \dots + \beta_{N-1} X_{N-1}(z). \quad (10)$$

The  $\beta_n$  coefficients represent the conjugate of the estimated channel gains according to the following relation:

$$\beta_n(kT_s) = \alpha_{k-(N-1)+n}^*. \quad (11)$$

The output of the combiner at the  $k$ th symbol period is represented by

$$y(kT_s) = \sum_{n=0}^{N-1} \alpha_{k-(N-1)+n}^* x_n((k - (N-1) + n)T_s) \quad (12)$$

$$= \sum_{n=0}^{N-1} \alpha_{k-(N-1)+n}^* \alpha_{k-(N-1)+n} d_{k-(N-1)} + v'_{kn},$$

$$y(kT_s) = \sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 d_{k-(N-1)} + v'_{kn}, \quad (13a)$$

$$v'_{kn} = \sum_{n=0}^{N-1} \beta_n(kT_s) v_{k-(N-1)+n}, \quad (13b)$$

where  $v'_{kn}$  is a Gaussian random variable with zero mean and  $\sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 \sigma_w^2$  variance. Figure 3 shows the receiver block diagram for the code-time diversity scheme. The estimated data at the output of the detector is late  $(N-1)$  symbol periods. This delay represents the time spread at which the transmitted symbol is repeated.

The last part of the demodulator is the detector. The optimum detector computes the Euclidian distance between the received symbol and all the symbols in the symbols constellation diagram. The detector decides that  $d_i$  is transmitted if and only if the distance between  $y(kT_s)$  and  $d_i$  is smaller than the distance between  $y(kT_s)$  and  $d_m$  for all  $m$ :

$$\text{choose } d_i \iff d^2 \langle y(kT_s), d_i \rangle < d^2 \langle y(kT_s), d_m \rangle \quad \forall i \neq m. \quad (14)$$

The combined signals in (12) are equivalent to that of  $N$ -branch MRC receiver. Therefore, the resulting diversity order from the new code-time diversity scheme with  $N$  spreading orthogonal codes and one transmission antenna is equal to that of the  $N$ -branch MRC receiver scheme.

It is important to emphasize on that the combined signals in (12) are similar to that of space-time MIMO system with  $N$  antennas at the transmitter and one antenna at the receiver or a space-time MIMO system with one antenna at the transmitter and  $N$  antennas at the receiver. The proposed code-time diversity system does not use additional encoders or decoders at the transmitter or the receiver such as the space-time encoder and decoder in the space-time MIMO system. No additional RF interface circuits or antennas are used in the code-time diversity. Spreading and despreading circuits are the only used additional hardware. The disadvantage of the proposed code-time diversity system is the extended bandwidth used and the  $N$  symbol period delays that precede the detection of the first transmitted symbol.

*Probability of Symbol Error.* To determine the probability of symbol error in the proposed code-time diversity system, the decision variable is calculated first. The optimum detector calculates the decision variable by multiplying the signal in (12) with the conjugate of all the complex symbols on

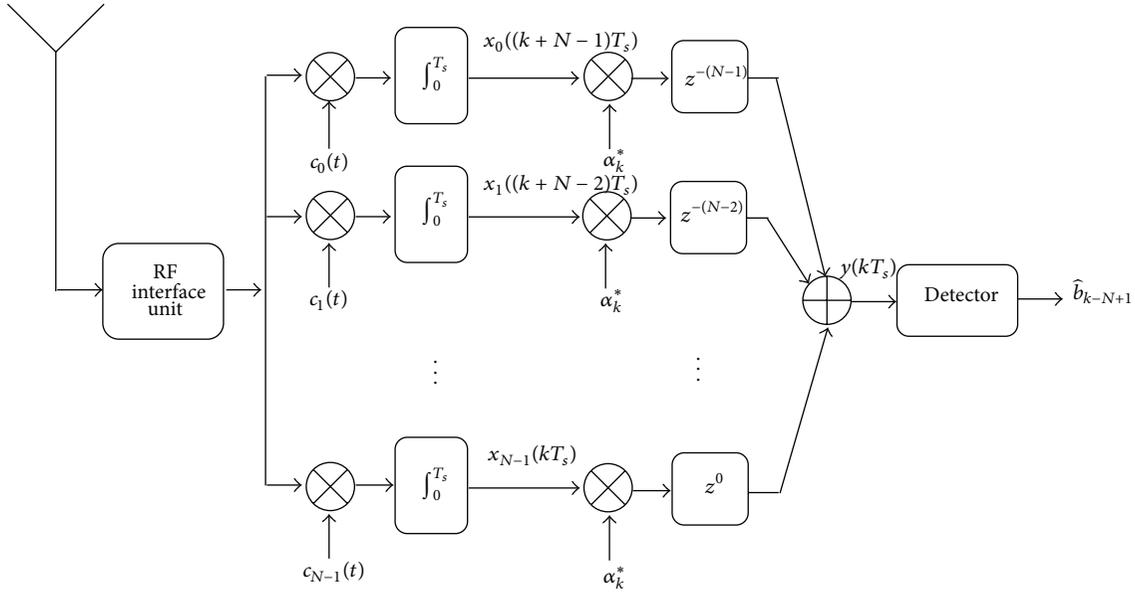


FIGURE 3: The code-time diversity DSSS receiver in Rayleigh flat fading channel.

the constellation diagram. The estimated symbol is the symbol with the largest decision variable. If symbol  $i$  is the symbol transmitted at the  $k$ th symbol period, the largest decision variable will be

$$DV = \langle y(kT_s) \cdot d_i^* \rangle, \quad \text{where } i = k - (N - 1), \quad (15)$$

$$DV = \sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 E_s + v''_{kn}, \quad (16)$$

$$E_s = \langle d_{k-(N-1)} \cdot d_{k-(N-1)}^* \rangle,$$

$$v''_{kn} = \langle v'_{kn} \cdot d_{k-(N-1)}^* \rangle,$$

where  $v''_{kn}$  is a Gaussian random variable with zero mean and  $\sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 \sigma_w^2 E_s$  variance. Therefore, the decision variable (DV) in (16) is also a Gaussian random variable with the following mean and variance:

$$E [DV] = \sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 E_s, \quad (17a)$$

$$\text{Var} [DV] = \sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 \sigma_w^2 E_s. \quad (17b)$$

According to the used modulation method, the probability of symbol error is always a function of the signal to noise ratio. This probability of error will be a random variable because the signal to noise ratio is a random variable. The instantaneous SNR is represented as shown in

$$\text{SNR}(kT_s) = \frac{\sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 E_s}{2\sigma_w^2}, \quad (18)$$

where  $|\beta_n(kT_s)|^2$  is a chi-square random variable. Thus, the conditional probability of symbol error is calculated first for

a certain value of the SNR, and then the average probability of error is calculated by averaging the conditional probability of symbol error over the probability density function of the SNR. For  $M$ -QAM, the probability of symbol error is given by

$$P_{\text{QAM}} = 4 \left( 1 - \frac{1}{\sqrt{M}} \right) Q \left( \sqrt{\frac{3}{M-1} \frac{\sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 E_s}{2\sigma_w^2}} \right) - 4 \left( 1 - \frac{1}{\sqrt{M}} \right)^2 Q^2 \left( \sqrt{\frac{3}{M-1} \frac{\sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 E_s}{2\sigma_w^2}} \right). \quad (19)$$

For simplicity (19) can be approximated by the first term only so that the  $Q^2(x)$  is ignored since  $Q^2(x) \ll Q(x)$  [21]. In the case of nonfading channel, the probability of symbol error in code-time diversity system is

$$P_{\text{QAM}} = 4 \left( 1 - \frac{1}{\sqrt{M}} \right) Q \left( \sqrt{\left( \frac{3}{M-1} \right) \frac{NE_s}{2\sigma_w^2}} \right). \quad (20)$$

For fading channel, the probability density function of the signal to noise ratio is equal to the probability density function (pdf) of a chi-square random variable with  $2N$  degrees of freedom as shown in

$$p(\text{SNR}(kT_s)) = \frac{1}{(N-1)! \text{SNR}^N} (\text{SNR}(kT_s))^{N-1} e^{-(\text{SNR}(kT_s))/\overline{\text{SNR}}}. \quad (21)$$

$\overline{\text{SNR}}$  is the average signal to noise ratio per time slot or by diversity channel and it is given by

$$\overline{\text{SNR}} = \frac{E_s}{2\sigma_w^2} E[|\beta_n(kT_s)|^2] = \frac{\sigma_\alpha^2 E_s}{2\sigma_w^2}. \quad (22)$$

The average probability of symbol error is

$$\begin{aligned} \overline{P}_{\text{QAM}} &= \int_0^\infty P_{\text{QAM}} \cdot p(\text{SNR}(kT_s)) \cdot d\text{SNR} \\ &= 4 \left(1 - \frac{1}{\sqrt{M}}\right) \left(\frac{1}{(N-1)! \overline{\text{SNR}}^N}\right) \\ &\quad \times \int_0^\infty Q\left(\sqrt{\frac{3}{M-1}} \text{SNR}\right) \cdot \text{SNR}^{N-1} e^{-\text{SNR}/\overline{\text{SNR}}} \cdot d\text{SNR}. \end{aligned} \quad (23)$$

After some mathematical manipulations, the exact value of the average probability of symbol error  $\overline{P}_{\text{QAM}}$  in code-time diversity is given by

$$\begin{aligned} \overline{P}_{\text{QAM}} &= 4 \left(1 - \frac{1}{\sqrt{M}}\right) \left(\frac{1}{(N-1)! \overline{\text{SNR}}^N}\right) \\ &\quad \times \left(\frac{2(M-1)}{3}\right)^N \frac{\Gamma(N+1/2)}{\sqrt{\pi}(2N)} \\ &\quad \times {}_2F_1\left(N, \frac{2N+1}{2}; N+1; \frac{-2(M-1)}{3\overline{\text{SNR}}}\right). \end{aligned} \quad (24)$$

${}_pF_q$  is the generalized hyper-geometric function. It is defined in Appendix A.

**4.2. Limitations of the Used Spreading Codes.** The use of DSSS in the proposed code-time diversity system expands the transmitted signal bandwidth more than the bandwidth of the nonspread modulated symbols and also more than the bandwidth of the transmitted signal if a space-time coding MIMO system is used. Although the enlarged bandwidth in the code-time diversity system increases the channel capacity and increases the system resistance to jamming and interference signals, bandwidth efficiency of code-time diversity system is poor.

In order to increase the bandwidth efficiency in code-time diversity system, more than one user are allowed to share the same channel bandwidth but with a different set of orthogonal spreading codes. If  $M$  users share the same channel using the proposed diversity system,  $M \times N$  orthogonal spreading codes are required. This increases the demands on the orthogonal spreading codes.

On the other hand, we can merely assign one PN sequence for each user in the multiuser code-time diversity system and by exploiting the autocorrelation property of the PN sequence, and the rest  $(N-1)$  spreading codes required for the proposed diversity scheme can be generated from the same generator polynomial by cyclically shifting the generated sequence different  $(N-1)$  times. For PN sequence with significant long period, the correlation between the generated

PN sequence and its cyclically shifted sequences is very small but not zero. From [21], if  $c(n)$  is a PN sequence of period  $N_c$ , the autocorrelation between this sequence and its cyclic shift with  $n_i$  chips is equals to

$$\begin{aligned} &\frac{1}{N_c} \sum_{m=0}^{N_c-1} c(m) \cdot c(m-n_i) \\ &= \begin{cases} 1 & \forall n_i = 0, N_c, 2N_c, 3N_c, \dots \\ -\frac{1}{N_c} & \text{else where.} \end{cases} \end{aligned} \quad (25)$$

The use of multiuser DSSS system with code-time diversity enhances the bandwidth efficiency of the system, but in this case a multiuser detector should be used in the receiver shown in Figure 3 instead of a single user detector. This point will be discussed in detail in a separate research, but now we continue with a single user detector case.

Equations (20) and (24) show the probability of error and the average probability of error in the received data for the case of nonfaded and faded channels, respectively, assuming that the used  $N$  spreading codes are mutually orthogonal. On the other hand if nonorthogonal codes are used such as a PN sequence and its cyclic shifted sequences, the correlation between the codes pairs affects the probability of error. This correlation gives rise to the intersymbol interference (ISI) between the transmitted symbols.

In nonorthogonal spreading code case, the output of each correlator with each spreading code consists of the desired signal and  $(N-1)$  interference signals from the previous and proceeding transmitted symbols. The output of the proposed MRC ( $y(kT_s)$ ) will have interference signals from the previous  $(N-1)$  transmitted symbols and interference signals from the proceeding  $(N-1)$  transmitted symbols. The  $n$ th correlator output at the  $k$ th symbol period is illustrated in

$$\begin{aligned} x_n(kT_s) &= \int_0^{T_s} r(t) \cdot c_n(t - kT_s) \cdot dt \\ &= \alpha_k d_{k-n} + \sum_{\substack{m=0 \\ m \neq n}}^{N-1} \alpha_k \rho_{nm} d_{k-m} + v_{kn}. \end{aligned} \quad (26)$$

The first term in the right hand side is the desired signal, the middle term is the ISI from the previous and proceeding symbols according to the value of  $n$ , and the last term is the Gaussian noise component.  $\rho_{nm}$  is the correlation between the spreading codes of index  $n$  and  $m$ :

$$\rho_{nm} = \int_0^{T_s} c_n(t - kT_s) \cdot c_m(t - kT_s) \cdot dt. \quad (27)$$

According to (26), the output of the proposed MRC will be

$$\begin{aligned} y(kT_s) &= \sum_{n=0}^{N-1} |\beta_n(kT_s)|^2 d_{k-(N-1)} \\ &\quad + \sum_{n=0}^{N-1} \sum_{\substack{m=0 \\ m \neq n}}^{N-1} |\beta_n(kT_s)|^2 \rho_{nm} d_{k-(N-1)+n-m} + v'_{kn}. \end{aligned} \quad (28)$$



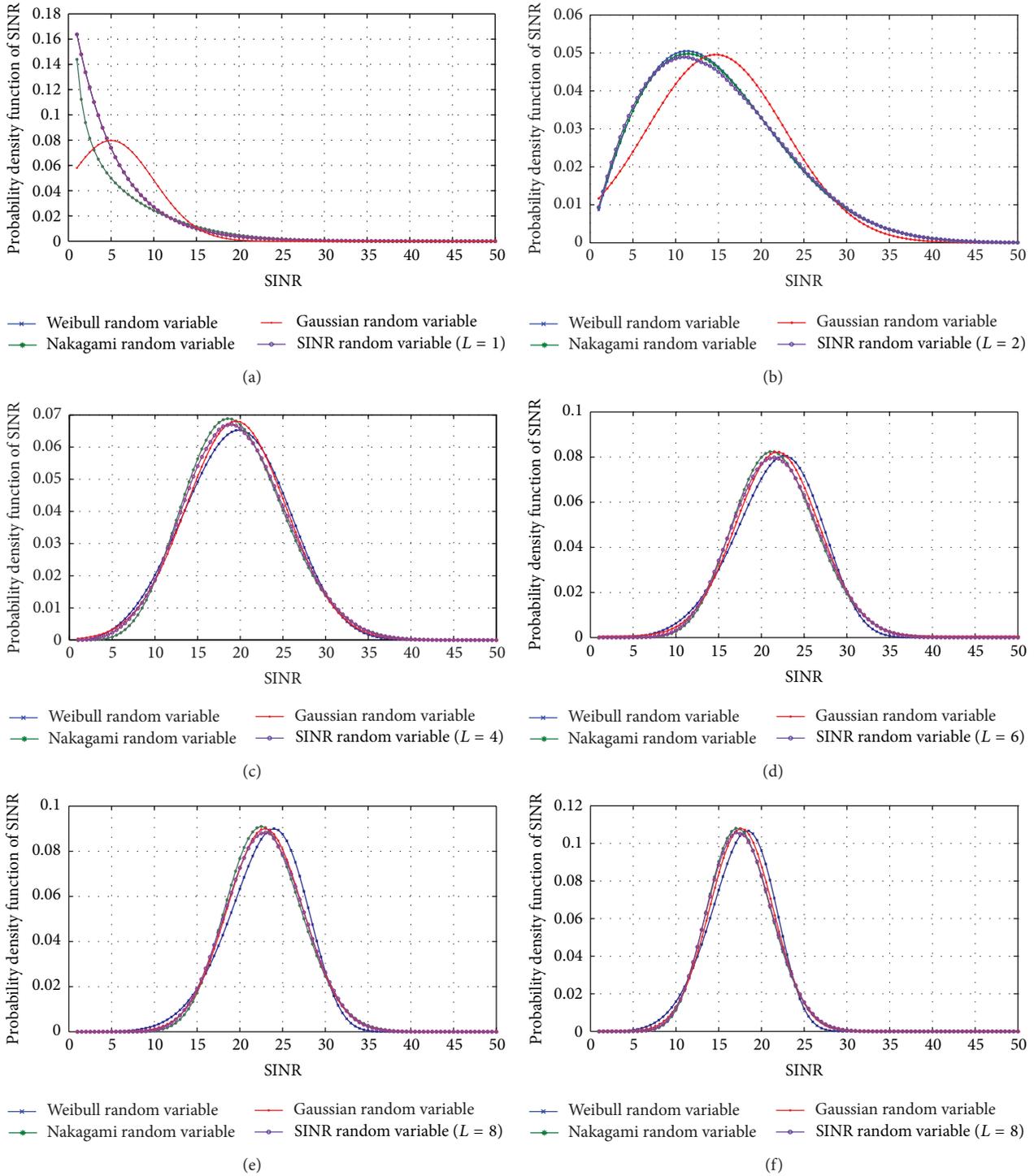


FIGURE 4: Comparison between the pdf of SINR random variable and the pdf of Weibull, Nakagami, and Gaussian random variables at different diversity orders ( $L = N$ ).

variable. This result matches the center limit theory of random variables. Although the unknown pdf of SINR is close to Gaussian pdf at high average SINR, Nakagami pdf will be used to approximate this unknown pdf since Nakagami pdf gives a good approximation of the unknown pdf of SINR at low and high average values of SINR. Equation (36)

is the probability density function of a Nakagami random variable:

$$p_{\text{SINR}}(\text{SINR}) = \frac{2}{\Gamma(m)} \cdot \left(\frac{m}{\Omega}\right)^m \text{SINR}^{2m-1} e^{-(m/\Omega)\text{SINR}^2}, \tag{36}$$

where  $m$  is the shape parameter and  $\Omega$  is the scale parameter. The shape and scale parameters are related to the mean and the variance of the SINR random variable [22]:

$$m = \left( \frac{(E[\text{SINR}^2])^2}{\text{var}[\text{SINR}^2]} \right), \quad \Omega = E[\text{SINR}^2]. \quad (37)$$

Now, the average probability of error will be

$$\begin{aligned} \bar{P}_{\text{QAM}} &= \int_0^\infty P_{\text{QAM}} \cdot p(\text{SINR}(kT_s)) \cdot d\text{SINR} \\ &= 2 \left( 1 - \frac{1}{\sqrt{M}} \right) \frac{2}{\Gamma(m)} \cdot \left( \frac{m}{\Omega} \right)^m \\ &\quad \times \int_0^\infty \text{erfc} \left( \sqrt{\frac{3}{2(M-1)} \text{SINR}} \right) \\ &\quad \times \text{SINR}^{2m-1} e^{-(m/\Omega)\text{SINR}^2} \cdot d\text{SINR}, \end{aligned} \quad (38)$$

$$\begin{aligned} \bar{P}_{\text{QAM}} &= \left( 1 - \frac{1}{\sqrt{M}} \right) \frac{2^{1-m} \Gamma(2m)}{\Gamma(m)} \\ &\quad \cdot \left( \frac{1}{3} \cdot e^{((9/(32(M-1)^2) \cdot (\Omega/m))} \cdot D_{-2m} \left( \sqrt{\frac{9}{8(M-1)^2} \cdot \frac{\Omega}{m}} \right) \right. \\ &\quad \left. + e^{((1/(2(M-1)^2) \cdot (\Omega/m))} \cdot D_{-2m} \left( \sqrt{\frac{2}{(M-1)^2} \cdot \frac{\Omega}{m}} \right) \right). \end{aligned} \quad (39)$$

$D_p(z)$  is the parabolic cylindrical function defined in [23]. The complete derivation of the average probability of symbol error in (39) is represented in Appendix B. If Gaussian pdf is used to model the pdf of SINR in (34), the average probability of symbol error will be

$$\begin{aligned} \bar{P}_{\text{QAM}} &= 2 \left( 1 - \frac{1}{\sqrt{M}} \right) \\ &\quad \cdot \left( \frac{1}{6} e^{((9\sigma^2)/(8(M-1)^2)) - (3\mu/2(M-1))} \right. \\ &\quad \left. + \frac{1}{2} e^{(((2\sigma^2)/(M-1)^2) - (2\mu/(M-1)))} \right), \\ \mu &= E[\text{SINR}], \quad \sigma^2 = E[(\text{SINR} - \mu)^2]. \end{aligned} \quad (40)$$

**4.3. Multipath Rayleigh Channel.** The impulse response of the multipath fading channel is shown in (42). Quasistatic channel is also assumed where the fading gain  $\alpha_l$  of the  $l$ th path is fixed during one symbol period and it is changed randomly from one symbol to another:

$$h(t) = \sum_{l=0}^{L-1} \alpha_l \delta(t - \tau_l), \quad (42)$$

where  $\alpha_l$  is a complex Gaussian random variable with zero mean and  $\sigma_{\alpha_l}^2$  variance.  $\tau_l$  is the  $l$ th path delay.  $L$  is the number of the uncorrelated fading paths from the transmitter to the receiver. The frequency components of the signal will experience different magnitudes of fading. In our proposed diversity scheme, the multipath fading case looks like the MIMO system where multiple antennas transmit the modulated symbols and multiple antennas at the receiver picked them up. According to our signal model, the received signal at the demodulator input is

$$r(t) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \alpha_{kl} d_{k-n} c_n(t - kTs - \tau_l) + w(t). \quad (43)$$

As shown in (1), the number of the paths  $L$  depends on the bandwidth of the DSSS signal and the coherence bandwidth of the channel. By increasing the process gain of the DSSS system, the number of the uncorrelated propagation paths is increased.

Here a new important note should be mentioned. The idea of the proposed diversity scheme is based on transmitting each symbol through more than one symbol period using separate spreading codes. As shown in the previous section, a PN code with different cyclic shifts may be used to encode each data symbol. In multipath fading channel, improper choice of the different shift values increases the ISI because the different channel delays may be equal to the shift values in the used PN code. So a condition should be made on the shift values of the PN code:

$$m_n T_c \neq \tau_l \quad \forall \text{ values of } m_n, \tau_l \in [0 T_m], \quad (44)$$

where  $m_n$  is an integer number, which represents the number of chips by which the PN code is shifted to form the  $n$ th code sequence in the used set of  $N$  spreading codes.  $T_m$  is the multipath delay spread and it represents the maximum delay of the longest path through which the signal propagates. One of the solutions of the inequality in (44) is

$$\max(m_n T_c) < \min(\tau_l). \quad (45)$$

The proposed demodulator for the code-time diversity system with multipath fading channel is more complex than the demodulator shown in Figure 3 for flat fading channel. The demodulator consists of three parts too, but the first part of this demodulator is a bank of  $NL$ -fingers RAKE filters instead of  $N$  correlators. Each  $L$ -fingers RAKE filter correlates the received signal with  $L$  PN sequences. These sequences are generated from one PN code from the used set of  $N$  spreading PN codes according to the set of channel delays  $\tau_l$ . Figure 5 shows the structure of the  $L$  figure RAKE filter that correlates the received data with the  $n$ th PN code in the spreading codes set.

The time resolution between the uncorrelated paths in the used RAKE receiver is  $T_c$ . In each finger of the RAKE filter, the delayed received signal is correlated with the  $n$ th PN code that is shifted with an integer number of chips equal to the path

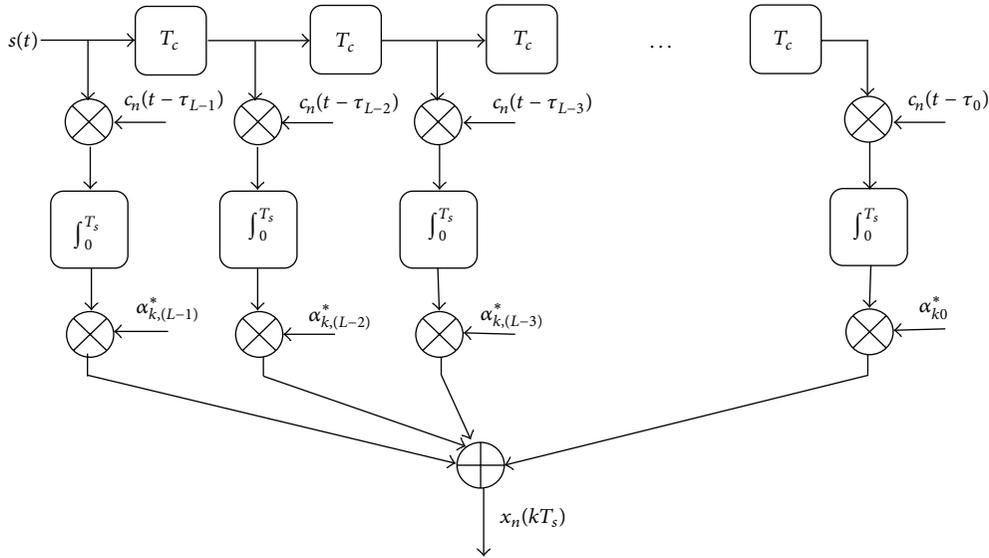


FIGURE 5: The  $n$ th  $L$ -fingers RAKE filter.

delay of that finger. Conventional MRC is used to combine the output of the  $L$ -fingers to form the random variable  $x_n(kT_s)$ . The output of the  $l$ th correlator in the  $n$ th RAKE filter is

$$\begin{aligned}
 x_{nl}(kT_s) &= \int_0^{T_s} r(t) \cdot c_n(t - kT_s - \tau_l) \cdot dt \\
 &= \alpha_{kl}d_{k-n} + \sum_{m \neq n}^{N-1} \alpha_{kl}\rho_{nm}d_{k-m} \\
 &\quad + \sum_{m=0}^{N-1} \sum_{l' \neq l}^{L-1} \alpha_{kl'}\rho_{nm}(l', l)d_{k-m-m_{l'}} + v_{kn}(l)
 \end{aligned} \tag{46}$$

$$m_{l'} = \left\lfloor \frac{\tau_{l'}}{T_s} \right\rfloor$$

$$\rho_{nm} = \int_0^{T_s} c_n(t) \cdot c_m(t) \cdot dt,$$

$$\rho_{nm}(l', l) = \int_0^{T_s} c_n(t - \tau_{l'}) \cdot c_m(t - \tau_l) \cdot dt,$$

$$v_{kn}(l) = \int_0^{T_s} w(t) \cdot c_n(t - kT_s - \tau_l) \cdot dt. \tag{47}$$

The first term in (46) is the desired signal. The second term is the ISI signal that comes from the  $(N-1)$  symbols transmitted through the same symbol period. This ISI signal is due to the correlation between the used spreading codes. The third term is another ISI signals that come from the other  $(L-1)$  fading

paths. The last term is the noise random variable. Using (46), the output of the  $n$ th RAKE filter is

$$\begin{aligned}
 x_n(kT_s) &= \sum_{l=0}^{L-1} |\alpha_{kl}|^2 d_{k-n} + \sum_{l=0}^{L-1} \sum_{m \neq n}^{N-1} |\alpha_{kl}|^2 \rho_{nm} d_{k-m} \\
 &\quad + \sum_{l=0}^{L-1} \sum_{m=0}^{N-1} \sum_{l' \neq l}^{L-1} \beta_{kl'} \rho_{nm}(l', l) d_{k-m-m_{l'}} + v_{kn}
 \end{aligned} \tag{48}$$

$$\beta_{ll'} = \alpha_{kl'} \cdot \alpha_{kl}^*.$$

The last term  $v_{kn}$  is a Gaussian random variable with zero mean and  $\sum_{l=0}^{L-1} |\alpha_{kl}|^2 \sigma_w^2$  variance. The second part of the proposed demodulator is the delayed symbols combiner. The DSC delays the output random variable from each RAKE filter according to the index of the PN sequence used in this RAKE filter. The output of the RAKE filter with the code sequence  $c_n(t)$  is delayed  $((N-1)-n)$  symbol periods. Figure 6 presents the structure of DSC.

The delayed signals are finally added to form a single input to the detector. The output of the DSC in the  $z$ -domain can be represented by

$$\begin{aligned}
 Y(z) &= X_0(z)z^{-(N-1)} + X_1(z)z^{-(N-1)+1} \\
 &\quad + X_2(z)z^{-(N-1)+2} + \dots + X_{N-1}(z).
 \end{aligned} \tag{49}$$

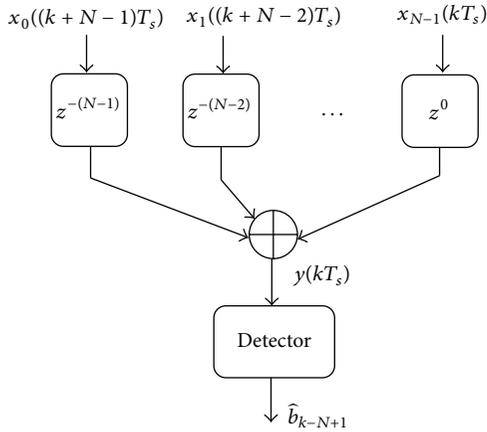


FIGURE 6: The delayed symbols combiner of the outputs of RAKE filters.

The output of the DSC combiner at the  $k$ th symbol period is represented by

$$y(kT_s) = \sum_{n=0}^{N-1} x_n((k - (N - 1) + n) T_s) + v'_{kn}, \quad (50)$$

$$y(kT_s) = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 d_{k-(N-1)} + \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{m \neq n}^{N-1} |\alpha_{nl}(kT_s)|^2 \rho_{nm} d_{k-(N-1)+n-m} + \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{m=0}^{N-1} \sum_{l' \neq l}^{L-1} \beta_{nl'l'}(kT_s) \rho_{nm}(l', l) \times d_{k-(N-1)+n-m-m_{l'}} + v'_{kn}, \quad (51)$$

$$v'_{kn} = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \beta_{nl}(kT_s) v_{k-(N-1)+n}(l), \quad (52)$$

$$\alpha_{nl}(kT_s) = \alpha_{k-(N-1)+n,l}^*$$

$$\beta_{nl'l'}(kT_s) = \alpha_{k-(N-1)+n,l'} \cdot \alpha_{k-(N-1)+n,l}^*$$

where  $v'_{in}$  is a Gaussian random variable with zero mean and  $\sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 \sigma_w^2$  variance. As in flat fading case, the estimated data at the output of the detector is delayed  $(N - 1)$  symbol periods. This delay represents the time spread at which the transmitted symbol is repeated. If the correlation between the used spreading codes is zero, the combined signal in (51) will be

$$y(kT_s) = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 d_{k-(N-1)} + v'_{kn}. \quad (53)$$

The last part of the demodulator is the detector. The estimated symbol at the output of the detector is the symbol with the minimum distance to the detector input as shown in (14).

The combined signals in (53) are equivalent to that of  $(L \times N)$ -branch MRC receiver. Thus, the resulting diversity order of the new code-time transmit diversity scheme with  $N$  spreading orthogonal codes and one transmitting and receiving antenna in frequency selective channel with  $L$  faded paths is equal to that of the  $(L \times N)$ -branch MRC receiver scheme.

The combined signals in (53) are also similar to that of space-time MIMO system with  $N$  antennas at the transmitter and  $L$  antennas at the receiver. The proposed code-time diversity system does not use additional encoders or decoders at the transmitter or the receiver such as the space-time encoder and decoders in the space-time MIMO systems. No additional RF interface circuits or antennas are used in the code-time diversity. Spreading and despreading circuits are the only used additional hardware. Although the code-time diversity has the disadvantage of low bandwidth efficiency due to the usage of DSSS, the extended bandwidth in DSSS increases the channel capacity and the DSSS can resist the jamming and noncochannel interference signals.

The ISI signals that appear in (51) can be eliminated or neglected if the correlation between the spreading codes is zero or very small, respectively. If spreading codes with unavoidable cross-correlation are used as PN sequences, the ISI signals can be minimized by using long codes sequences or by using linear equalizers.

To determine the probability of error in the proposed code-time diversity system in frequency selective channel, the same procedure as that used in flat fading channel case is followed. The decision variable in the detector is calculated first. Then the conditional probability of error is calculated given a fixed set of channel gains. Finally the average probability of error is calculated based on the probability density function of the decision variable. Based on (53), the decision variable for the case of orthogonal spreading codes is

$$DV = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 E_s + v''_{in}, \quad (54)$$

where  $v''_{in}$  is a Gaussian random variable with zero mean and  $\sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 E_s \sigma_w^2$  variance and  $|\alpha_{nl}(kT_s)|^2$  is a chi-square random variable with two degrees of freedom. The instantaneous SNR is

$$SNR = \frac{E [DV]^2}{2 \text{var} [DV]} = \frac{E_s \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2}{2\sigma_w^2}. \quad (55)$$

SNR in (55) is a chi-square with  $2N \times L$  degrees of freedom. The pdf of the SNR random variable is represented in (21) where  $NL$  replaces  $N$ . The average probability of error of code-time diversity system in  $L$ -paths Rayleigh fading channel is

$$\begin{aligned} \bar{P}_{\text{QAM}} &= 4 \left( 1 - \frac{1}{\sqrt{M}} \right) \left( \frac{1}{(N-1)! \text{SNR}^{NL}} \right) \\ &\times \left( \frac{2(M-1)}{3} \right)^{NL} \frac{\Gamma(NL + (1/2))}{\sqrt{\pi} (2NL)} \\ &\times {}_2F_1 \left( NL, \frac{2NL+1}{2}; N+1; \frac{-2(M-1)}{3\text{SNR}} \right). \end{aligned} \quad (56)$$

In nonorthogonal spreading code case, the detector decision variable is

$$\begin{aligned} DV &= \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 E_s \\ &+ \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{m \neq n}^{N-1} |\alpha_{nl}(kT_s)|^2 \rho_{nm} d_{k-(N-1)+n-m} \cdot d_{k-(N-1)}^* \\ &+ \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{m=0}^{L-1} \sum_{l' \neq l}^{L-1} \beta_{nl'l'}(kT_s) \rho_{nm}(l', l) d_{k-(N-1)+n-m-m_l} \\ &\cdot d_{k-(N-1)}^* + v_{in}''. \end{aligned} \quad (57)$$

Following the same procedure as Rayleigh flat fading case, the instantaneous SINR is

$$\begin{aligned} \text{SINR}(kT_s) &= \frac{E_s \left( \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2 \right)^2}{(N-1) \rho^2 E_s \left( \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^4 + \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \sum_{l' \neq l}^{L-1} |\beta_{nl'l'}(kT_s)|^2 \right) + 2\sigma_w^2 \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} |\alpha_{nl}(kT_s)|^2}. \end{aligned} \quad (58)$$

Numerical calculations of the pdf of the SINR random variable in (58) show that the SINR random variable can also be approximated to a Nakagami random variable as represented in Section 4.2. Following the same procedure, the average probability of symbol error of code-time diversity system in  $L$ -paths Rayleigh fading channel with nonorthogonal spreading codes is

$$\begin{aligned} \bar{P}_{\text{QAM}} &= \left( 1 - \frac{1}{\sqrt{M}} \right) \frac{2^{1-mL} \Gamma(2mL)}{\Gamma(mL)} \\ &\cdot \left( \frac{1}{3} \cdot e^{((9/(32(M-1)^2)) \cdot (L\Omega/m))} \cdot D_{-2m} \left( \sqrt{\frac{9}{8(M-1)^2} \cdot \frac{L\Omega}{m}} \right) \right. \\ &\left. + e^{((1/(2(M-1)^2)) \cdot (L\Omega/m))} \cdot D_{-2m} \left( \sqrt{\frac{2}{(M-1)^2} \cdot \frac{L\Omega}{m}} \right) \right). \end{aligned} \quad (59)$$

If Gaussian pdf is used to approximate the pdf of SINR, the average probability of symbol error will be

$$\begin{aligned} \bar{P}_{\text{QAM}} &= 2 \left( 1 - \frac{1}{\sqrt{M}} \right) \\ &\cdot \left( \frac{1}{6} e^{(3L/2)((3\sigma^2)/(4(M-1)^2) - (\mu/(M-1)))} \right. \\ &\left. + \frac{1}{2} e^{2L((\sigma^2/(M-1)^2) - (\mu/(M-1)))} \right). \end{aligned} \quad (60)$$

### 5. Simulations

The proposed code-time diversity system is simulated using a DSSS system. Two different spreading codes are used. Walsh codes simulate the case of orthogonal codes' set; however, PN codes simulate the case of nonorthogonal codes' set. Different number of spreading codes  $N$  is used to achieve transmitter diversity. The used modulation scheme is 16-QAM. The transmitted symbols rate is 5 M symbol/s. The transmitted signal carrier frequency is 10 GHz. The used process gains are 11.78 dB and 15 dB. The transmitted signal bandwidths are 150 MHz and 310 MHz according to the used spreading code and its process gain.

Figures 7 and 8 show the average probability of bit error in the received data when code-time diversity is used in Rayleigh flat fading channel. The simulated system used  $N = 2, 4, 6,$  and  $8$  code sequences. For nonorthogonal code set, PN sequences with 31 chips period and 15 chips period are used. Figure 7 contains the probability of error curves for  $N = 2, 6$  and Figure 8 contains the curves of  $N = 4, 8$ . In orthogonal codes case, the proposed system achieved diversity gain proportional to the number of the used codes  $N$ . Increasing the code diversity by increasing  $N$  will increase the diversity order and enhance the system performance. The orthogonality between the used codes prevents the ISI from appearing. The probability of error curves of the orthogonal codes case in these figures is the same as the probability of error curves of the diversity systems in [24] using the same diversity order. The curves also realize (24) for Rayleigh flat fading channel. The figures likewise show the case of nonorthogonal codes where the ISI appeared. The ISI increases the average probability of error as shown in (39) and

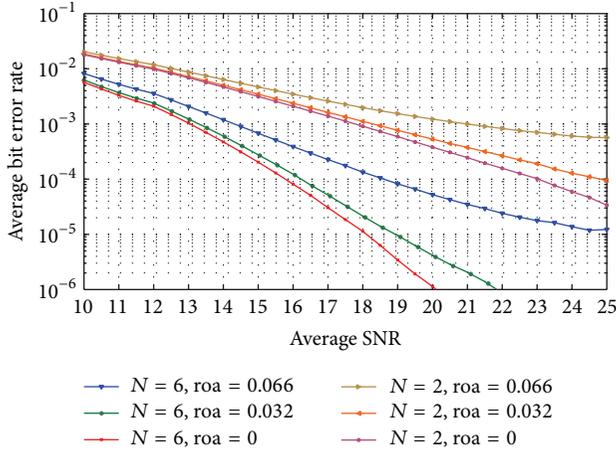


FIGURE 7: The average probability of bit error of code-time diversity system in Rayleigh flat fading channel with  $N = 2, 6$  using orthogonal and nonorthogonal spreading codes.

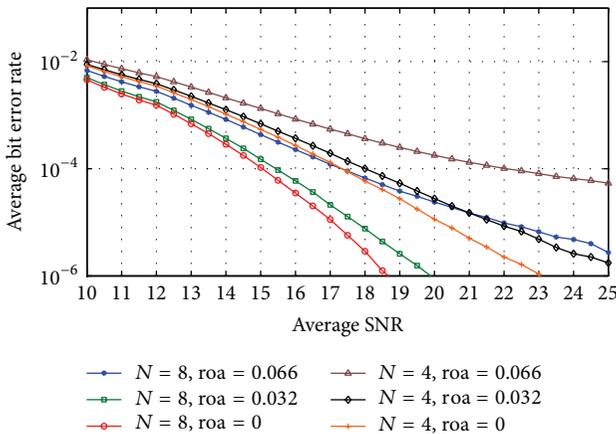


FIGURE 8: The average probability of bit error of code-time diversity system in Rayleigh flat fading channel with  $N = 4, 8$  using orthogonal and nonorthogonal spreading codes.

(40). If the code period of the used code increases, the cross-correlation between the codes pairs decreased and the average probability of error is improved.

The code-time diversity system is simulated in Rayleigh frequency selective channel with  $L = 2$  and  $4$ . As shown before, the performance of the proposed code-time diversity system in frequency selective channel is similar to the performance of the MIMO diversity system. The diversity order in the proposed system will equal the multiplication of the number of used codes ( $N$ ) in the transmitter by the number of the signal paths ( $L$ ) of the channel. Figure 9 shows the average bit error rate in the received data for  $N = 2$  and  $L = 2$ ; that is, diversity order is  $4$ . The proposed code-time diversity system is simulated using orthogonal and nonorthogonal codes. In orthogonal codes case, the average probability of error matches the values of (56) and the average

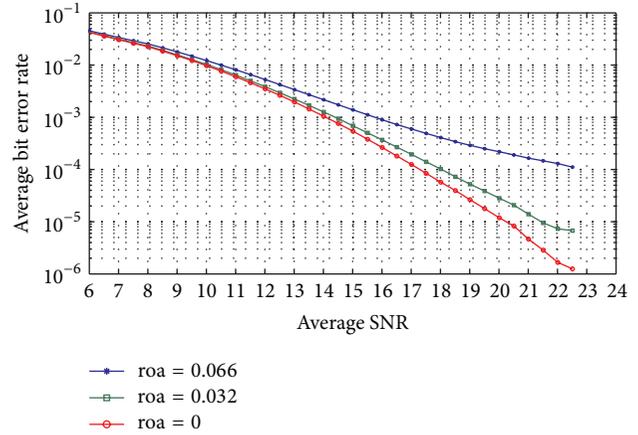


FIGURE 9: The average probability of bit error of code-time diversity system in Rayleigh frequency selective fading channel with  $N = 2$  and  $L = 2$ , using orthogonal and nonorthogonal spreading codes.

probability of error in  $2 \times 2$  MIMO system in [24]. On the other hand, the nonorthogonal codes case gives rise to ISI and the average probability of error will increase. As the code period increases, the correlation between the different codes pairs decreases and so the ISI between the successive symbols. The same results are achieved in Figures 10 and 11 for  $N = 4, L = 2$  and  $N = 4, L = 4$  cases, respectively. Furthermore, the performance of the simulated systems for orthogonal code case in Figures 10 and 11 is the same as the performance of  $4 \times 2$  and  $4 \times 4$  MIMO systems in [24], respectively.

## 6. Conclusions

The proposed code-time diversity is a diversity system suitable for direct sequence spread spectrum. The proposed diversity scheme uses single RF interface unit and single antenna at the transmitter and receiver. The proposed system achieves the benefits of diversity systems as well as the benefits of spread spectrum systems. If orthogonal spreading codes are used, the performance of the code-time diversity system is similar to the performance of the MIMO system with the same diversity order. The code-time diversity can achieve a higher diversity order than the MIMO system, which is limited with the number of the used antennas and the RF interface units. The proposed system is suitable for working in flat and frequency selective channels. The proposed system also gives a good performance if nonorthogonal codes are used as long as the cross-correlation between the used codes pairs is small enough. The paper represents mathematical derivations of the probability of error of the proposed system in nonfaded and Rayleigh faded channels for orthogonal and nonorthogonal spreading codes. The disadvantage of the proposed system is the bandwidth efficiency. This disadvantage can be enhanced if multiusers are allowed to share the same channel bandwidth with different spreading codes set.

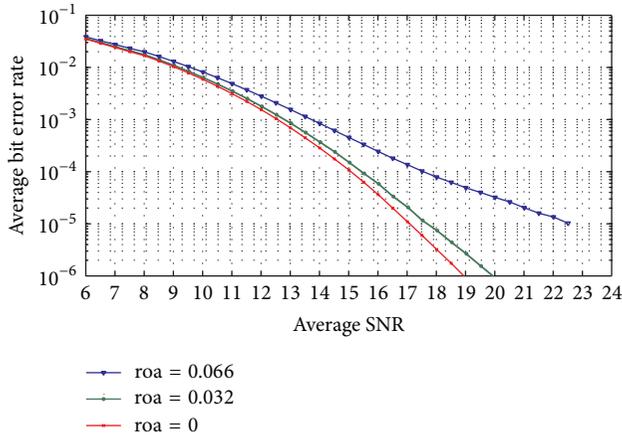


FIGURE 10: The average probability of bit error of code-time diversity system in Rayleigh frequency selective fading channel with  $N = 4$  and  $L = 2$ , using orthogonal and nonorthogonal spreading codes.

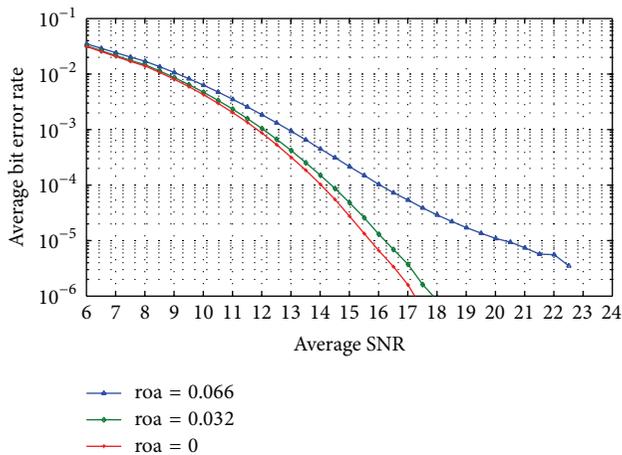


FIGURE 11: The average probability of bit error of code-time diversity system in Rayleigh frequency selective fading channel with  $N = 4$  and  $L = 4$ , using orthogonal and nonorthogonal spreading codes.

## Appendices

### A. Definition of the Generalized Hypergeometric Function

The generalized hypergeometric function  ${}_pF_q$  has a series expansion as shown in the following equation:

$$\begin{aligned}
 & {}_pF_q(\{a_1, a_2, \dots, a_p\}; \{b_1, b_2, \dots, b_q\}; z) \\
 &= \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k} \cdot \frac{z^k}{k!}
 \end{aligned} \tag{A.1}$$

This mathematical function is suitable for both symbolic and numerical manipulation.  $(a)_k$  is the Pochhammer symbol defined as

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{A.2}$$

### B. The Evaluation of the Integration in (38)

Starting from (38),

$$\begin{aligned}
 \bar{P}_{\text{QAM}} &= \int_0^{\infty} P_{\text{QAM}} \cdot p(\text{SINR}(kT_s)) \cdot d\text{SINR} \\
 &= 2 \left(1 - \frac{1}{\sqrt{M}}\right) \frac{2}{\Gamma(m)} \cdot \left(\frac{m}{\Omega}\right)^m \\
 &\quad \times \int_0^{\infty} \text{erfc}\left(\sqrt{\frac{3}{2(M-1)}\text{SINR}}\right) \\
 &\quad \times \text{SINR}^{2m-1} e^{-(m/\Omega)\text{SINR}^2} \cdot d\text{SINR}.
 \end{aligned} \tag{B.1}$$

From [25], the  $\text{erfc}()$  can be approximated to

$$\text{erfc}(x) \approx \frac{1}{6}e^{-x^2} + \frac{1}{2}e^{-(4/3)x^2} \tag{B.2}$$

$$\begin{aligned}
 \bar{P}_{\text{QAM}} &= 2 \left(1 - \frac{1}{\sqrt{M}}\right) \frac{2}{\Gamma(m)} \cdot \left(\frac{m}{\Omega}\right)^m \\
 &\quad \times \left[ \int_0^{\infty} \frac{1}{6} \text{SINR}^{2m-1} e^{-(m/\Omega)\text{SINR}^2} e^{-(a/2)\text{SINR}} \cdot d\text{SINR} \right. \\
 &\quad \left. + \int_0^{\infty} \frac{1}{2} \text{SINR}^{2m-1} e^{-(m/\Omega)\text{SINR}^2} e^{-(2a/3)\text{SINR}} \cdot d\text{SINR} \right],
 \end{aligned} \tag{B.3}$$

where  $a = 3/(M-1)$ .

From 3.462 in [23],

$$\int_0^{\infty} x^{\nu-1} \cdot e^{-Bx^2} \cdot e^{-\gamma x} \cdot dx = (2B)^{-\nu/2} \Gamma(\nu) e^{\gamma^2/8B} D_{-\nu}\left(\frac{\gamma}{\sqrt{2B}}\right) \tag{B.4}$$

Referring to (B.3),  $\nu = 2m$ ,  $B = 2/\Omega$ , and  $\gamma = a/2$  for the first integral and  $\gamma = 2a/3$  for the second one.

By substituting (B.4) into (B.3), the integration can be solved and the final value of the average probability of error can be represented as

$$\begin{aligned}
 \bar{P}_{\text{QAM}} &= \left(1 - \frac{1}{\sqrt{M}}\right) \frac{2^{1-m} \Gamma(2m)}{\Gamma(m)} \\
 &\quad \cdot \left(\frac{1}{3} \cdot e^{((9/(32(M-1)^2) \cdot (\Omega/m))} \cdot D_{-2m} \right. \\
 &\quad \times \left(\sqrt{\frac{9}{8(M-1)^2} \cdot \frac{\Omega}{m}}\right) + e^{((1/(2(M-1)^2) \cdot (\Omega/m))} \\
 &\quad \left. \cdot D_{-2m} \left(\sqrt{\frac{2}{(M-1)^2} \cdot \frac{\Omega}{m}}\right)\right).
 \end{aligned} \tag{B.5}$$

### Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

## Acknowledgment

The author would like to express his special appreciation and thanks to his Advisors Professors Dr. Abdel-Wahab Fayez, Dr. Abdel Aziz M. AL-Bassiouni, and Dr. Khaled Talaat; they have been tremendous mentors for him. The author would like to thank them for encouraging his research and for allowing him to grow as a research scientist. Special thanks are due to Professor Mohamed Saad al-Juhani, the dean of the Faculty of Engineering in Northern Border University, Saudi Arabia, for his encouragement and support to him to finish up this work.

## References

- [1] H. Gacanin, S. Takaoka, and F. Adachi, "BER performance of OFDM combined with TDM using frequency-domain equalization," *Journal of Communications and Networks*, vol. 9, no. 1, pp. 34–42, 2007.
- [2] H. Kun, N. Lim, and M. J. Won, "Performance of coded frequency-hopped OFDM systems in frequency selective channels," in *Proceedings of the 8th International Conference on Signal Processing (ICSP '06)*, vol. 3, November 2006.
- [3] I. M. Arijon and P. G. Farrell, "Performance of an OFDM system in frequency selective channels using Reed-Solomon coding schemes," in *Proceedings of the IEE Colloquium on Multipath Countermeasures*, 1996.
- [4] Y. Liu, Z. Tan, H. Wang, and K. S. Kwak, "Joint estimation of channel impulse response and carrier frequency offset for OFDM systems," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 9, pp. 4645–4650, 2011.
- [5] T. Cui and C. Tellambura, "Joint frequency offset and channel estimation for OFDM systems using pilot symbols and virtual carriers," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1193–1202, 2007.
- [6] Q. Shi, L. Liu, Y. L. Guan, and Y. Gong, "Fractionally spaced frequency-domain MMSE receiver for OFDM systems," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4400–4407, 2010.
- [7] M. S. Ahmed, S. Boussakta, B. S. Sharif, and C. C. Tsimenidis, "OFDM based on low complexity transform to increase multipath resilience and reduce PAPR," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5994–6007, 2011.
- [8] S.-S. Eom, H. Nam, and Y.-C. Ko, "Low-complexity PAPR reduction scheme without side information for OFDM systems," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, 2012.
- [9] R. Krenz, "Comparative study of space-diversity techniques for MLSE receivers in mobile radio," *IEEE Transactions on Vehicular Technology*, vol. 46, no. 3, pp. 653–663, 1997.
- [10] A. Tall, Z. Rezki, and M. -S. Alouini, "MIMO channel capacity with full CSI at low SNR," *IEEE Wireless Communications Letters*, vol. 1, no. 5, 2012.
- [11] M. Matthaiou, N. D. Chatzidiamantis, G. K. Karagiannidis, and J. A. Nossek, "On the capacity of generalized-K fading MIMO channels," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5939–5944, 2010.
- [12] D. Gozávez, D. Gómez-Barquero, D. Vargas, and N. Cardona, "Time diversity in mobile DVB-T2 systems," *IEEE Transactions on Broadcasting*, vol. 57, no. 3, 2011.
- [13] L.-F. Wei, "Coded M-DPSK with built-in time diversity for fading channels," *IEEE Transactions on Information Theory*, vol. 39, no. 6, pp. 1820–1839, 1993.
- [14] W. C. Wong, R. Steele, B. Glance, and D. Horn, "Time diversity with adaptive error detection to combat rayleigh fading in digital mobile radio," *IEEE Transactions on Communications*, vol. 31, no. 3, pp. 378–387, 1983.
- [15] B. Song, N. Kim, and H. Park, "A binary space-time code for MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1350–1357, 2012.
- [16] C.-H. Chen and W.-H. Chung, "Dual diversity space-time coding for multimedia broadcast/multicast service in MIMO systems," *IEEE Transactions on Communications*, vol. 60, no. 11, 2012.
- [17] A. F. Molisch, M. Z. Win, and J. H. Winters, "Space-time-frequency (STF) coding for MIMO-OFDM systems," *IEEE Communications Letters*, vol. 6, no. 9, pp. 370–372, 2002.
- [18] O.-S. Shin, A. M. Chan, H. T. Kung, and V. Tarokh, "Design of an OFDM cooperative space-time diversity system," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 4, pp. 2203–2215, 2007.
- [19] S. Li, D. Huang, K. B. Letaief, and Z. Zhou, "Pre-DFT processing for MIMO-OFDM systems with space-time-frequency coding," *IEEE Transactions on Wireless Communications*, vol. 6, no. 11, pp. 4176–4182, 2007.
- [20] L. Shi, W. Zhang, and X. -G. Xia, "Space-frequency codes for MIMO-OFDM systems with partial interference cancellation group decoding," *IEEE Transactions on Communications*, vol. 61, no. 8, 2013.
- [21] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, NY, USA, 2001.
- [22] E. K. Al-Hussaini and A. A. M. Al-Bassiouni, "Performance of MRC diversity systems for the detection of signals with nakagami fading," *IEEE Transactions on Communications*, vol. 33, no. 12, pp. 1315–1319, 1985.
- [23] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, Elsevier/Academic Press, San Diego, Calif, USA, 7th edition, 2007.
- [24] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.
- [25] M. Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 840–845, 2003.

## Research Article

# Trusted Measurement Model Based on Multitenant Behaviors

Zhen-Hu Ning, Chang-Xiang Shen, Yong Zhao, and Peng Liang

College of Computer Science, Beijing University of Technology, Beijing 100124, China

Correspondence should be addressed to Zhen-Hu Ning; [ning\\_zhenhu@163.com](mailto:ning_zhenhu@163.com)

Received 13 January 2014; Accepted 16 February 2014; Published 30 March 2014

Academic Editors: Y. Mao, X. Meng, J. Zhou, and Z. Zhou

Copyright © 2014 Zhen-Hu Ning et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With a fast growing pervasive computing, especially cloud computing, the behaviour measurement is at the core and plays a vital role. A new behaviour measurement tailored for Multitenants in cloud computing is needed urgently to fundamentally establish trust relationship. Based on our previous research, we propose an improved trust relationship scheme which captures the world of cloud computing where multitenants share the same physical computing platform. Here, we first present the related work on multi-tenant behaviour; secondly, we give the scheme of behaviour measurement where decoupling of multitenants is taken into account; thirdly, we explicitly explain our decoupling algorithm for multitenants; fourthly, we introduce a new way of similarity calculation for deviation control, which fits the coupled multitenants under study well; lastly, we design the experiments to test our scheme.

## 1. Introduction

Cloud computing has recently attracted an important attention and dubbed as the “next best thing” in information and communication technologies (ICT) [1]. As the intrinsic feature of cloud computing, multitenancy brings sharing concept to almost all information technologies such as sharing computing resources, sharing storage resources, and sharing network. Coresident clients might have no preestablished trust relationship and might have no knowledge of the existence or identities of other clients. In such a setting, if one of the coresidents maybe attacks the other coresidents it will be much easier to succeed and be difficult to detect. Therefore, this risk incurred by trusted measurement of multitenant is a barrier to acceptance of cloud computing. Actually cloud computing system, such as Amazon’s Elastic Compute Cloud (EC2), Microsoft’s Azure, and Rackspace’s Mosso, is a large scale system which is studied in cybernetics long before. Here we leverage the generalized predictive control affiliated to cybernetics to solve the problem of behavior measurement of multitenants on the same physical server brought by the new paradigm of cloud computing.

## 2. Background

This section consists of two parts: one is multitenancy threat; the other is the brief introduction of generalized predictive control.

*2.1. Multitenancy Threat.* It is important to consider the unique security risks introduced by multitenancy as intrinsic of the new paradigm of cloud computing in order to be able to derive adequate security solutions. As more and more applications become exported to third-party compute clouds, it becomes increasingly important to quantify any threats to confidentiality that exist in this setting [2, 3]. An obvious threat to these consumers of cloud computing is malicious behavior by the cloud provider, who is certainly in a position to violate customer confidentiality or integrity. However, this is a known risk with obvious analogs in virtually any industry practicing outsourcing. In this work, we consider the provider and its infrastructure to be trusted. This also means we do not consider attacks that rely upon subverting a cloud’s administrative functions, via insider abuse or vulnerabilities in the cloud management systems (e.g., virtual machine monitors).

In our threat model, adversaries are non-provider-affiliated malicious parties. Victims are multitenants running confidentiality-requiring services in the cloud. A traditional threat in such a setting is direct compromise, where an attacker attempts remote exploitation of vulnerabilities in the software running on the system. Of course, this threat exists for cloud applications as well. These kinds of attacks (while important) are a known threat and the risks they present are understood.

We instead focus on where third-party cloud computing gives attackers novel abilities, implicitly expanding the attack

surface of the victim. We assume that, like any customer, a malicious party can run and control many instances in the cloud, simply by contracting for them. Further, Based on the fact the economies offered by third-party compute clouds derive from multiplexing physical infrastructure, we assume (and later validate) that attacker's instances might even run on the same physical hardware as potential victims. From this vantage, an attacker might manipulate shared physical resources (e.g., CPU caches, branch target buffers, network queues, etc.) to learn otherwise confidential information.

*2.2. Generalized Predictive Control.* In general sense, predictive control, regardless of various algorithms, is based on the following three basic principles [4].

(1) *Predictive Model.* Predictive control is also referred to as model-based control where this model is referred to as predictive model. The predictive model can predict the future output of the object based on historical information and input. And the predictive model does not emphasize its structure but emphasizes the function of the model. Therefore, the traditional model such as equation of state and the transfer function can be used as a predictive model. Similarly, nonparametric model such as step response and impulse response can also be used directly as a predictive model.

(2) *Rolling Optimization.* Predictive control is an optimal control algorithm, which determines the future action through an optimal performance index. However, the optimization studied in predictive control is different from optimal control in the traditional sense, and the subtle difference is that optimization in the predictive control is a rolling optimization within the limited time. At each sampling instant, the optimization performance indicators relate only to a limited time since the right moment. Until the next sampling instant, this optimization period moves forward. At different instants, the relative forms of optimization performance indicators are the same, but its absolute form, that is, containing time area, is different. Therefore, during predictive control, optimization is not offline conducted only once but repeated online, which is the core of rolling optimization, that is, the fundamental characteristics of optimal control here is different from the traditional ones.

(3) *Feedback Correction.* Predictive control is a closed-loop control algorithm, where a series of further control actions can be ascertained by optimization. Predictive control does not perform all these actions but perform the present action. So that the deviation from the ideal state can be avoided; this is resulted from either the model mismatch or environmental interference. Until the next sampling time, the first is to detect the actual output of the object; the second is to take advantage of this real-time information to correct the prediction based on the model; and the final is to conduct the new optimization. Therefore, the optimization of the predictive control is not only based on the model, but also the feedback information, which constitutes a closed-loop optimization.

### 3. Related Work

There exist several measurement models such as Tripwire [5], AEGIS [6], and trusted box [7], the trust chain model proposed by the TCG (trusted computing group). These models focus on different measurement aspects of the system or file program, but these approaches belong to static integrity measurement of the resource. As a result, they cannot consider the dynamic trustworthiness in the system.

Further the researchers put forward the following schemes to realize dynamic measurement. In [8], there is a coprocessor-based kernel integrity monitor. The monitor periodically checks system memory and detects whether malicious programs change the host system kernel. Binding instructions and data (BIND) binds with the data and the corresponding block of process in order to provide a basis for the verification side to trace data processing. However, it cannot cope with many attacks when the system is running [9]. Policy reduced integrity measurement architecture (PRIMA) focuses on the flow of information when the system is running [10], but the model trusts flow of information which comes from the trusted subjects in mandatory access control (MAC). However, it is still a role-based privilege. The measure mode is too simple and does not conform to the definition of definition of trust. Behavior based trustworthiness attestation mode (BTAM) is trusted proof model based on the behavior of the system [11]. This model firstly determines whether the system behavior is related to trustworthiness of platform state. For a large number of behaviors that cannot be determined, this model has not yet given the solution. Therefore, the dynamic trusted measure theory and technology is an urgent need for the development of cloud computing [12].

Gong [13] firstly introduces generalized prediction control theory to analyze and measure the tenants' behaviour in the information system. The novel scheme greatly increases the trustworthiness and security of information system and opens a new direction towards behaviour measurement [13]. However, the new features mentioned above brought by the cloud computing were not considered and studied. This paper is to improve that model and to adapt the new feature of multitenancy brought by cloud computing.

### 4. Model Design

Traditional authorization and authentication are to solve the main problem whether the user's identity is trusted, while they are ineffective to solve whether the user's behavior is trusted. The original drive to promote the change of system status is the behavior [14]. Therefore, the trusted measurement of the behavior is more precise than the trusted measurement of the identity when it comes to reflect the trustworthiness of the system. The design of our model is consistent with the trustworthiness defined by Trusted Computing Group (TCG); that is, it is defined as trusted if the behavior can be expected [13]. According to this definition, we propose a measurement model for virtual machine behavior shown in Figure 1.

The first step: the characteristics of the shared resources in cloud computing brings the advantages while leading to

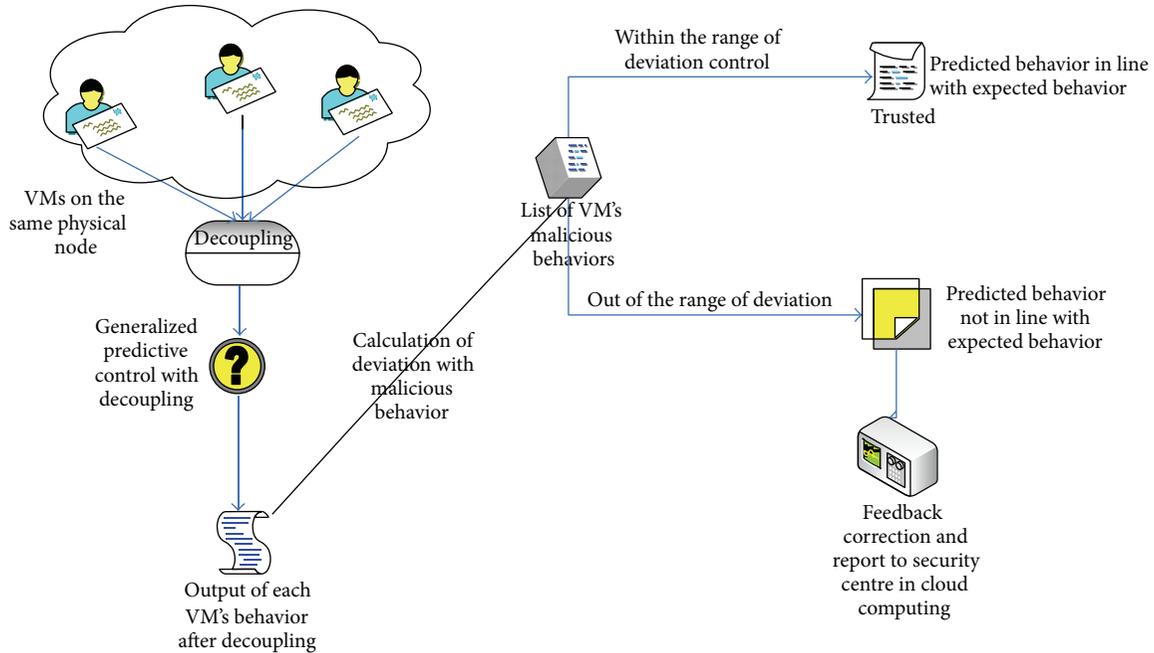


FIGURE 1: Behavior measurement model of virtual machine.

security problems. So it is necessary to conduct decoupling control over the behavior of the virtual machines on the same physical platform. Illustratively, the decoupling control aims to simplify the control over many virtual machines sharing resources of the same physical computing node into a lot of individual control loops for each virtual machine corresponding to individual customers.

The second step: according to the decoupling control algorithm, the inputs and outputs of several virtual machines in the same physical computer can be decoupled. The decoupled inputs and outputs of appropriate virtual machine can be controlled by the generalized predictive control algorithm here. Specifically, through the past and present behavior of the virtual machine, the further behavior can be predicted.

The third step: to match predicted behavior with characteristics list of malicious behaviors so as to obtain the similarity value/deviation value. If the deviation value is less than the threshold value predetermined by the system, then the behavior is trusted, otherwise it is an untrusted behavior.

### 5. Model Implementation

The multiple tenants studied here refer to the ones who share the same physical resource such as network card and bandwidth. Due to the multitenancy sharing, the cloud computing becomes much more complicated. In order to better predict the tenant's complicated behaviors, we utilize the multiple variable generalized predictive control to capture those behaviors. In this section, firstly we depict the cloud computing system in the view of generalized predictive control; secondly, we present the description of behaviors in cloud computing; thirdly, we introduce the establishment of list of malicious tenants' behaviors; fourthly, we give

decoupling algorithm for multitenant behaviors both in private and public clouds using generalized predictive control without coupling; fifthly, we give the similarity calculation used in our scheme for deviation control to confirm whether the suspected behavior is trusted or not finally.

*5.1. Description of Controlled Object.* From the view of control theory, the physical computing nodes where several virtual machines collocate can be taken as a multi-input, multioutput information flow control system. Figure 2 shows a physical computing node collocated by four virtual machines from the perspective of the generalized predictive control theory. Eight behavioral measurement points are as input of the information system; the outputs are four virtual machines captured by eight behavioral measurement points, which are in line with the appropriate expectation, respectively. Each virtual machine is one of outputs of the entire physical computing node, while all four virtual machines are equal to total inputs of the entire physical compute node, such that the total traffic of all four virtual machines should be equal to the traffic of physical computing nodes.

*5.2. Description of Tenant's Behavior.* There exists monitoring components in virtualized trusted computing platform based on dual-system architecture proposed by our research team. These monitoring components can identify measurement indicators of the behavior performance of virtual machine. There are several commonly used monitoring components as follows: (1) memory and CPU monitor: to monitor memory usage and CPU call rate and report monitoring results to the behavioral data collector; (2) port monitor and message analyzer: responsible for monitoring all open TCP or UDP ports of compute nodes and capturing and

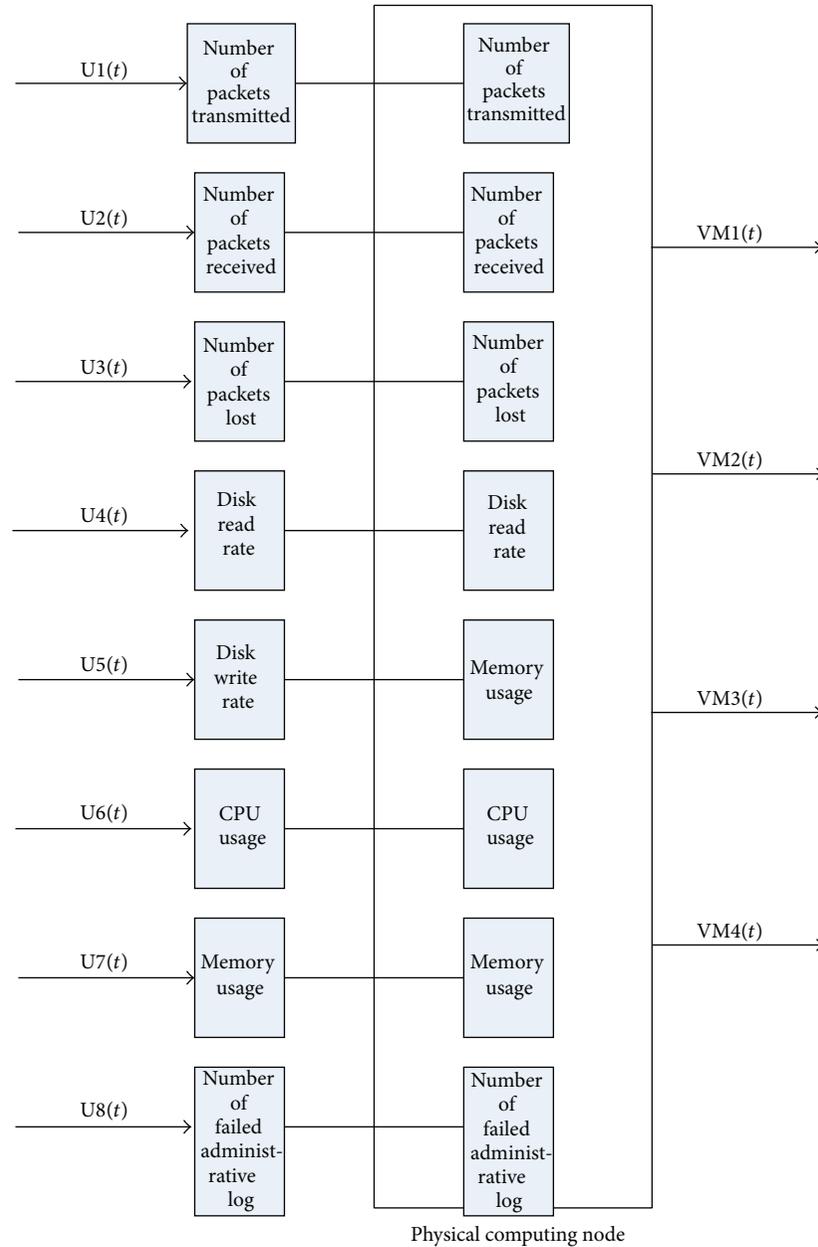


FIGURE 2: Description of physical computing nodes in the view of generalized predictive control.

analyzing communication message packet of the suspicious port. So that we can determine the role of the suspicious port and the corresponding process behavior of this port. If a suspicious user process is found to monitor a suspicious port and to communicate the message frequently, it is necessary to temporarily suspend the implementation of the process and to report to the Cloud Security Management Center; (3) network traffic detector: its role is to monitor the flow of network communication, in particular, the network traffic coming out of a virtual machine. Each virtual node has been deployed the monitor, so that both the denial of service attacks and the worm can be monitored and found. As a matter of fact, DoS and worm attacks will lead to a sharp

rise in network traffic. If it is found that a virtual machine computing task unconventionally and frequently sends out a lot of the packages with the same content, this task needs to be suspended, that is, to prevent the execution of the virtual machine user tasks, and then to be reported to the Cloud Security Management Center.

In the cloud computing model, we studied the related results conducted by both foreign researchers such as Khorshed et al. [15] and local researchers such as Li et al. [16]; we choose the following to depict the virtual machine behavior, which is the number of transmitted packets, the number of received packets, the number of lost packets, disk read speed, disk write speed, memory usage, CPU usage, and the number

TABLE I: Virtual machine behavior metric vector.

Measurement point	Measured object
MP <sub>1</sub>	Number of packets transmitted
MP <sub>2</sub>	Number of packets received
MP <sub>3</sub>	Number of packets lost
MP <sub>4</sub>	Disk read rate
MP <sub>5</sub>	Disk write rate
MP <sub>6</sub>	Memory usage
MP <sub>7</sub>	CPU usage
MP <sub>8</sub>	Number of failed administrative log on attempt

of failed login attempts. Here, these eight performance indicators are named as measurement point, abbreviated as MP. In this paper, the behavior measurement vector of running virtual machine consists of the aforementioned 8 measurement points, see Table I.

**5.3. List of Tenant's Malicious Behaviors.** The researchers from University of California, San Diego, and the Massachusetts Institute of Technology, Cambridge University [17] conducted a thorough experimental study on Amazon's Elastic Compute Cloud [18]. The results show that the cloud infrastructure can be mapped out, and the position of a specific virtual machine can be located. They also point out that the aforementioned information can be exploited to make side channel attacks so as to collect the information of the target virtual machine located on the same physical machine. In a recent study, Rocha and Correia [19] investigated how malicious insiders steal confidential data and demonstrated these attacks using the video and showing insiders can easily obtain passwords, encryption keys, and documents. Chonka et al. [20] reproduced the scenario of some recent attacks happening in the cloud computing and demonstrated how the HTTP-DOS and XML DoS occur in the cloud computing.

Khorshed et al. found that there exists some common factor behind these attack models [17–19], because all the attackers use a similar attack tools and follow a certain attack process. Khorshed et al. firstly collected relevant attack tools such as Hping, socket programming, httping Unix shell script, and side channel attacks. Next they collected a variety of attack scenarios related to network security by browsing relevant website and blog, such as Danchev [21] and Grossman [22] as well as their research work [23–25], and then generated attack script using the aforementioned documents.

Based on above steps, Khorshed et al. designed the experiment to collect data in the cloud computing environment. The type of data will determine the kind of data collection tools. In the attack scenario, most common data types are as follows 8 performance indicators such as the number of transmitted and received data packets, processing time, the round-trip time, and CPU usage. Khorshed et al. adopted machine learning techniques to classify the attacks related to malicious use of resources in the cloud computing. Through a large number of experiments, they obtained 8 measurement points of behavioral performance such as the number of transmitted

packets, the number of received packets, the number of lost packets, disk read speed, disk write speed, memory usage, CPU usage, and the number of failed login attempts. Further, they concluded the behavioral characteristics of the classic attack in conduction of eight measurement points [26].

**5.4. Decoupling Algorithm.** To maximize efficiency, multiple VMs, one VM corresponding to one tenant, may be simultaneously assigned to be executed on the same physical server, which is supported by virtualization technology. As a result, tenants share the physical resources (e.g., CPU caches, branch target buffers, network queues, etc.) to accomplish their computation tasks. From the angle of generalized predictive control (GPC), cloud computing system under study corresponds to multiple inputs and multiple outputs system in cybernetics which is different from the single input and output system that is studied in [13]. The essential difference is the coupling between tenants on the same physical server, which should be studied thoroughly. In this section, first we use GPC theory to capture the multitenant behavior and then to derive the decoupling algorithms that is shown at the end of this part.

The multitenant's behavior in cloud computing can be described by

$$A(z^{-1})y(t) = D(z^{-1})B(z^{-1})u(t-1) + \frac{C(z^{-1})\xi(t)}{\Delta}, \quad (1)$$

where  $A(z^{-1}) = I + A_1z^{-1} + \dots + A_{n_A}z^{-n_A}$ ,  $B(z^{-1}) = I + B_1z^{-1} + \dots + B_{n_B}z^{-n_B}$ ,  $D(z^{-1}) = \text{diag}(z^{-k_i})$ ,  $\Delta = \text{diag}(1 - z^{-1})$ .

$\{u(t)\}$  and  $\{y(t)\}$  indicate coresident tenants' inputs and outputs.  $\xi(t)$  is  $m$ -dimension independent random disturbance vector, and its mean value and variance are zero and  $\sigma I$ , respectively. Without loss of generality, suppose  $A(z^{-1})$  is diagonal matrix.

$B(z^{-1})$  is divided into two parts, namely,

$$B(z^{-1}) = \bar{B}(z^{-1}) + \tilde{B}(z^{-1}), \quad (2)$$

where  $\bar{B}(z^{-1})$  is diagonal matrix polynomials and  $\tilde{B}(z^{-1})$  is a matrix whose diagonal is zero. Equation (2) indicates that  $\bar{B}(z^{-1})$  is the direct relation between tenant's inputs and outputs, and  $\tilde{B}(z^{-1})$  is the mutual coupling part of communication channel.

Using (1) and (2), we have

$$A(z^{-1})\Delta y(t) = D(z^{-1})\bar{B}(z^{-1})\Delta u(t-1) + D(z^{-1}) \times \tilde{B}(z^{-1})\Delta u(t-1) + C(z^{-1})\xi(t). \quad (3)$$

Performance index function is as follows:

$$J = \xi \cdot \left\{ \sum_{j=1}^N \|\phi(t+j) - r_j\omega(t+j) + \bar{S}_j(z^{-1})\Delta u(t+j-1)\|_Q^2 + \sum_{j=1}^N \|\Delta u(t+j-i)\|_{\lambda_j}^2 \right\}, \quad (4)$$

where

$$\phi(t+j) = D(z)\Delta y(t+j) \quad (5)$$

indicates generalized outputs,  $D(z)$  indicates the inverse of  $D(z^{-1})$ ,  $\omega(t+j)$  is fixed vector,  $\|X\|_Q^2$  indicates  $X^T Q X$ , and  $Q$  is symmetric positive definite matrix. There is no such  $\bar{S}_j(z^{-1})u(t+j-1)$ , part of (4), in the performance index of common generalized prediction control.  $\bar{S}_j(z^{-1})$  is a matrix polynomial whose diagonal is zero and  $\bar{S}_j(z^{-1})$  can be used to eliminate the coupling effect between channels. Similarly, weighted constant matrix  $\lambda_i$  can be divided into two  $\bar{\lambda}_j$  and  $\tilde{\lambda}_j$ ;  $\bar{\lambda}_j$  is a diagonal matrix and  $\tilde{\lambda}_j$  is a matrix whose diagonal is zero; the function of  $\tilde{\lambda}_j$  is the same as that of  $\bar{S}_j(z^{-1})$ .

We use the methods in [27] to achieve the decoupling algorithm.

Define Diophantine equation:

$$I = F_j(z^{-1})A(z^{-1}) + z^{-j}D(z^{-1})G_j(z^{-1}), \quad (6)$$

where  $F_j(z^{-1}) = I + F_1^j z^{-1} + \dots + F_{n_D+j-1}^j z^{-n_D-j+1}$ ,  $G_j(z^{-1}) = G_0^j + G_1^j z^{-1} + \dots + G_{n_A-1}^j z^{-n_A+1}$ .

Since  $A(z^{-1})$  and  $D(z^{-1})$  are diagonal matrix,  $F_j(z^{-1})$  and  $G_j(z^{-1})$  are diagonal matrix as well. Equation (6) is left multiplied with  $D(z^{-1})$ :

$$D(z) = D(z)F_j(z^{-1})A(z^{-1}) + z^{-j}G_j(z^{-1}). \quad (7)$$

$D(z)F_j(z^{-1})$  left multiplies with (3), and using the above formula, we obtain the following:

$$\begin{aligned} & D(z)\Delta y(t+j) \\ &= F_j(z^{-1})\bar{B}(z^{-1})\Delta u(t+j-1) \\ & \quad + F_j(z^{-1})\tilde{B}(z^{-1})\Delta u(t+j-1) + G_j(z^{-1})\Delta y(t) \\ & \quad + D(z)F_j(z^{-1})C(z^{-1})\xi(t+j). \end{aligned} \quad (8)$$

Since the term  $D(z)F_j(z^{-1})C(z^{-1})\xi(t+j)$  is unrelated to other terms, optimal prediction of  $\phi(t+j)$  can be represented as follows:

$$\begin{aligned} & \phi^0(t+j) \\ &= F_j(z^{-1})\bar{B}(z^{-1})\Delta u(t+j-1) \\ & \quad + F_j(z^{-1})\tilde{B}(z^{-1})\Delta u(t+j-1) + G_j(z^{-1})\Delta y(t). \end{aligned} \quad (9)$$

Both  $F_j(z^{-1})\bar{B}(z^{-1})$  and  $F_j(z^{-1})\tilde{B}(z^{-1})$  can be divided into two parts:

$$\begin{aligned} F_j(z^{-1})\bar{B}(z^{-1}) &= E_j(z^{-1}) + z^{-j}L_j(z^{-1}), \\ F_j(z^{-1})\tilde{B}(z^{-1}) &= \tilde{E}_j(z^{-1}) + z^{-j}\tilde{L}_j(z^{-1}), \end{aligned} \quad (10)$$

where

$$\begin{aligned} E_j(z^{-1}) &= \sum_{i=1}^j E_i^j z^{-i}, & L_j(z^{-1}) &= \sum_{i=1}^{n_D+n_B-1} L_i^j z^{-i}, \\ \tilde{E}_j(z^{-1}) &= \sum_{i=1}^j \tilde{E}_i^j z^{-i}, & \tilde{L}_j(z^{-1}) &= \sum_{i=1}^{n_D+n_B-1} \tilde{L}_i^j z^{-i}. \end{aligned} \quad (11)$$

Equation (9) can be represented as

$$\begin{aligned} \phi^0(t+j) &= E_j(z^{-1})\Delta u(t+j-1) + \tilde{E}_j(z^{-1})\Delta u(t+j-1) \\ & \quad + G_j(z^{-1})\Delta y(t) + L_j(z^{-1})\Delta u(t+j-1) \\ & \quad + \tilde{L}_j(z^{-1})\Delta u(t+j-1). \end{aligned} \quad (12)$$

Substitute the above formula into (4), and choose  $\bar{S}_j(z^{-1})$  that satisfies

$$\begin{aligned} & \bar{S}_j(z^{-1})\Delta u(t+j-1) + \tilde{E}_j(z^{-1})\Delta u(t+j-1) \\ & \quad + \tilde{L}_j(z^{-1})\Delta u(t+j-1) = \bar{M}_j(z^{-1})\Delta u(t-1), \end{aligned} \quad (13)$$

where  $\bar{M}_j(z^{-1}) = \bar{M}_0^j + \bar{M}_1^j z^{-1} + \dots + \bar{M}_{n_M}^j z^{-n_M}$  is a matrix polynomial whose diagonal is zero, so (4) can be represented as

$$\begin{aligned} J &= \sum_{j=1}^N \left\| E_j(z^{-1})\Delta u(t+j-1) + L_j(z^{-1})\Delta u(t+j-1) \right. \\ & \quad \left. + \tilde{M}_j(z^{-1})\Delta u(t-1) + G_j(z^{-1})\Delta y(t) - r_j \omega(t+j) \right\|_Q^2 \\ & \quad + \sum_{j=1}^N \left\| \Delta u(t+j-i) \right\|_{\lambda_j}^2 \\ &= \|EU + L\Delta u(t-1) + G\Delta y(t) + \bar{M}\Delta u(t-1) \\ & \quad - RW\|_I^2 + \|U\|_{\lambda}^2, \end{aligned} \quad (14)$$

where  $R = \text{diag}(r_j)$ ,  $\lambda = \text{diag}(\bar{\lambda}_j) + \text{diag}(\tilde{\lambda}_j) = \bar{\lambda}_j + \tilde{\lambda}_j$ ,  $j = 1, \dots, N$ ,

$$E = \begin{bmatrix} E_{10} \\ E_{21}E_{20} \\ \vdots \\ E_{NN-1}E_{NN-2} \cdots E_{N0} \end{bmatrix},$$

$$\begin{aligned} U &= [\Delta u(t), \Delta u(t+1), \dots, \Delta u(t+N-1)]^T, \\ W &= [w(t), w(t+1), \dots, w(t+N-1)]^T, \end{aligned} \quad (15)$$

$$G = [G_1(z^{-1}), G_2(z^{-1}), \dots, G_N(z^{-1})]^T,$$

$$L = [L_1(z^{-1}), L_2(z^{-1}), \dots, L_N(z^{-1})]^T,$$

$$\bar{M} = [\bar{M}_1(z^{-1}), \bar{M}_2(z^{-1}), \dots, \bar{M}_N(z^{-1})]^T.$$

Calculate the minimum of  $J$  so that we obtain

$$U = (E^T E + \bar{\lambda})^{-1} E^T \times [RW - G\Delta y(t) - L\Delta u(t-1) - \bar{M}\Delta u(t-1)] - (E^T E + \bar{\lambda})^{-1} \bar{\lambda} U, \quad (16)$$

where the value of matrix  $\bar{M}$  and  $\bar{\lambda}$  can be determined by closed-loop system equation.

The first  $m$  rows of  $(E^T E + \bar{\lambda})^{-1} E^T$  are defined as  $e^T = [e_1, \dots, e_N]$ , and the first  $m$  rows of  $(E^T E + \bar{\lambda})^{-1}$  are defined as  $h^T = [h_1, \dots, h_N]$ .

$u(t)$  can be represented as

$$u(t) = [e_1, \dots, e_N] \times [RW - G(t) - L\Delta u(t-1) - \bar{M}\Delta u(t-1)] - [h_1 \bar{\lambda}_1 + \dots + h_N \bar{\lambda}_N z^{N-1}] \Delta u(t). \quad (17)$$

Substituting above formula into (3), we obtain the closed-loop system equation:

$$\begin{aligned} & \{ [I + z^{-1} (e_1 L_1 + e_2 L_2 + \dots + e_N L_N z^{N-1})] A \\ & + z^{-1} D\bar{B} [e_1 G_1 + \dots + e_N G_N] \} \Delta y(t) \\ & = D\bar{B} [e_1 r_1 + e_2 r_2 z + \dots + e_N r_N z^{N-1}] \omega(t) \\ & - D\bar{T}\Delta u(t-1) + VC\xi(t), \end{aligned} \quad (18)$$

where  $\bar{T}$  indicates the mutual coupling part

$$\begin{aligned} \bar{T} = & \bar{B} [z^{-1} (e_1 \bar{M}_1 + \dots + e_N \bar{M}_N) \\ & + (h_1 \bar{\lambda}_1 + \dots + h_N \bar{\lambda}_N z^{N-1})] \\ & - [I + z^{-1} (e_1 L_1 + \dots + e_N L_N)] \bar{B}. \end{aligned} \quad (19)$$

According to (18), the coupling of closed-loop system is decoupled if and only if  $\bar{T} = 0$ . Because the number of variables is less than that of equations, both  $\bar{M}_j(z^{-1})$  and  $\bar{\lambda}_j$  of (19) can be obtained by least squares method; consequently  $\bar{T}$  is not equal to zero exactly, and further decoupling is approximate.

Moreover, controlled object of formula (1) is CARMA model. Since there is no steady error in outputs of closed-loop system, it is necessary to determine the matrix  $r_j$  of the performance index (4). To be simplified, let  $r_1 = r_2 = \dots = r_N = r$ ; we can obtain  $r$  from formula (18):

$$r = (e_1 + \dots + e_N)^{-1} \times \{ \bar{B}(1)^{-1} [I + e_1 L_1(1) + \dots + e_N L_N(1)] A(1) + e_1 G_1(1) + \dots + e_N G_N(1) \}. \quad (20)$$

After substituting  $\bar{\lambda}_1, \bar{M}_j$ , and  $r_j$  into (16), the following law of decoupling space can be derived:

$$\begin{aligned} \Delta u(t) = & [I, 0, \dots, 0] [E^T E + \bar{\lambda} + \bar{\lambda}]^{-1} \\ & \times E^T [RW - G\Delta y(t) - L\Delta u(t-1) - \bar{M}\Delta u(t-1)]. \end{aligned} \quad (21)$$

Generalized predictive control based decoupling algorithm is as follows.

*Step 1.*  $B(z^{-1})$  can be divided into  $\bar{B}(z^{-1})$  and  $\tilde{B}(z^{-1})$  using (2).

*Step 2.* Least square method on  $\bar{M}_j(z^{-1})$  and  $\bar{\lambda}$  can be computed out using (19).

*Step 3.*  $r_j$  can be calculated using (20).

*Step 4.*  $\Delta u(t)$  can be derived by formula (21).

$u(t)$  is the predicted value of individual virtual machine, after decoupling, on the virtualized platform of cloud computing.

**5.4.1. Decoupling Algorithm for Public Cloud.** The parameters used above are known in the case the user of the virtual machine is fixed, while the aforementioned algorithm with decoupling is not applicable where the users are not fixed. For example, the users in public cloud computing are not fixed, so that the parameters related to users' behavior are unknown. In such public cloud computing, it is necessary to use parameter estimation to obtain the appropriate parameters of the corresponding controlled object and then conduct the predictive control algorithm mentioned above.

$\Delta$  left multiplies with (1); we have

$$\Delta y(t) = Q(z^{-1}) Y(t) + C(z^{-1}) \xi(t), \quad (22)$$

where

$$Q(z^{-1}) = [A(z^{-1}) - I, G(z^{-1})],$$

$$Y(t) = [-\Delta y(t-1), -\Delta y(t-2), \dots, -\Delta y(t-n_A)],$$

$$\Delta u(t-1), \Delta u(t-2), \dots, \Delta u(t-n_B)]. \quad (23)$$

Formula (22) is a multivariate linear equation; we solve  $Q(z^{-1})$  and  $G(z^{-1})$  by the least squares method.

However,  $Q(z^{-1})$  may change slowly with time; a typical equation is

$$G(z^{-1}, t) = G(z^{-1}, t-1) + \frac{M(t-1)q(t)}{\rho + M(t-1)^T M(t-1)}, \quad (24)$$

where  $0.95 \leq \rho \leq 1$  is the forgetting factor and  $q(t) = y(t) - Q(z^{-1})Y(t)$ .

Then the method to deal with (22) may become very complex.

In this paper, to solve the newest ( $z^{-1}$ ), we introduce the least squares method with weighs.

That is,  $Q(z^{-1})$  satisfies

$$F = \min \sum_{i=1}^L \lambda_i q^2(i), \tag{25}$$

where  $L$  is the size of the sample space and  $\lambda_i \geq 0$  is the weight satisfying  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_L$ . Let the derivative of  $F$  be zero. We obtain  $Q(z^{-1})$ . To reduce the error, we can construct  $\{u^*(t)\}_{1 \leq t \leq L}$ ,  $\{y^*(t)\}_{1 \leq t \leq L}$  as follows.

Let  $k, L$  be integers; to obtain  $G(z^{-1})$ , we choose a series of  $\{u(t)\}_{1 \leq t \leq kL}$ ,  $\{y(t)\}_{1 \leq t \leq kL}$  and construct  $\{u^*(t)\}_{1 \leq t \leq L}$ ,  $\{y^*(t)\}_{1 \leq t \leq L}$  as follows:

$$\begin{aligned} \Delta u^*(t) &= \frac{1}{k} \sum_{i=1}^k \Delta u(Lt + i), \\ \Delta y^*(t) &= \frac{1}{k} \sum_{i=1}^k \Delta y(Lt + i). \end{aligned} \tag{26}$$

Since the mean value of  $\xi(t)$  is 0, then we have

$$\frac{1}{k} \sum_{i=1}^k \xi(mt + i) \approx 0. \tag{27}$$

Then using the least squares method with weighs on  $\{u^*(t)\}_{1 \leq t \leq L}$ ,  $\{y^*(t)\}_{1 \leq t \leq L}$ , we obtain  $Q(z^{-1})$ .

**5.5. Deviation Control.** The behavior of the virtual machine can be mapped to a point in the space that consists of eight behavioral measurement points. The model of behavioral trusted measurement can determine whether the behavior of the virtual machine is out of security border, that is, whether the behavior is a malicious one. Mathematically, the aforementioned is to obtain the distance between two points in 8-dimensional space that consists of 8 behavior measurement points. This is actually a problem to calculate the similarity between two different objects.

Similarity calculation is widely used in the intrusion detection technology and other technologies. The typical solutions are like inner product, Dice coefficient, cosine function, and Jaccard coefficient method [28].

In this paper, gray correlation analysis is adopted to calculate the deviation value. Because the predictive value of the virtual machine behavior is unknown, the historical and present behavior of the virtual machine is consistent with the information, so that this known information and corresponding location information constitute a gray system [29]. At present, the gray system theory has been extended to many fields such as the industrial, agricultural, social, economic, energy, geology, and petroleum, successfully solving a large number of practical problems in production, living, and scientific research and making remarkable achievements. The gray relational analysis is a branch of the gray system theory.

The basic idea of gray relational analysis is to determine whether they are similar to each other by the degree of

similarity of curve geometry composed of the appropriate data sequence. In terms of mathematics, gray correlation degree is used here to reflect the degree of similarity. The closer the two curves are, the greater the degree of correlation of the two corresponding data sequences is, and vice versa. When it comes to specific analysis, it is desirable to replace unlimited convergence curve with approximate convergence (data array), so as to provide a great convenience in the case of dealing with a large number of practical problems.

Combined with the characteristics of a distributed computing environment based on virtual architectures, Grey Relational Analysis is adopted in this paper, and the specific calculation steps are as follows.

(1) According to the measurement point of the behavior of the virtual machine, to create the reference sequence of a virtual machine behavior, suppose  $n$  data sequences can form the following matrix:

$$(X'_1, X'_2, \dots, X'_n) = \begin{pmatrix} x'_1(1) & x'_2(1) & \dots & x'_n(1) \\ x'_1(2) & x'_2(2) & \dots & x'_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ x'_1(8) & x'_2(8) & \dots & x'_n(8) \end{pmatrix}. \tag{28}$$

8 indicates the number of behavioral measurement points while  $n$  represents the time series:

$$X'_i = (x'_i(1), x'_i(2), \dots, x'_i(8))^T, \quad i = 1, 2, \dots, n. \tag{29}$$

The data sequence is known as the reference sequence that can reflect the characteristics of the behavior of the system. The data sequence is known as comparison sequence that is composed of the factors that affect behavior of the system.

(2) The goal of the behavior of the virtual machine decides the value of the behavioral measurement point and further determines comparison sequence that has impact on the behavior of the system.

Reference data sequence should be a standard for the comparison. Here, reference data sequence comes from the list of the behavioral characteristics, seen in Table 1, written as

$$X'_0 = (x'_0(1), x'_0(2), \dots, x'_0(8)). \tag{30}$$

(3) Nondimensionalization of the reference sequence and the comparison sequence.

Due to the fact that the factors in the system have various physical meanings, the dimensions involved in the factors are different as well. As a result, it is difficult to compare the factors so as not to obtain a correct conclusion. When it comes to Grey Relational Analysis, generally it is required to carry out nondimensionalization of the appropriate data. The methodologies commonly used for nondimensionalization are as follows, for example, equalization method and the initialization method, seen in (31):

$$x_i(k) = \frac{x'_i(k)}{(1/8) \sum_{k=1}^8 x'_i(k)}, \quad x_i(k) = \frac{x'_i(k)}{x'_i(1)} \tag{31}$$

$$i = 0, 1, \dots, n; k = 1, 2, \dots, 8.$$

After nondimensionalization, data sequence is as follows:

$$(X_0, X_1, \dots, X_n) = \begin{pmatrix} x_0(1) & x_1(1) & \dots & x_n(1) \\ x_0(2) & x_1(2) & \dots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_0(8) & x_1(8) & \dots & x_n(8) \end{pmatrix}. \quad (32)$$

Here the initialization method is adopted to conduct nondimensionalization.

(4) In our scheme, the comparison sequence refers to the behavioral measurement vector of the virtual machine to be measured. For every behavior of the virtual machine, the corresponding absolute difference between the comparison sequence and reference sequence is calculated, respectively; that is,  $|x_0(k) - x_i(k)|$ , where  $k = 1, \dots, 8$ ;  $i = 1, \dots, n$ ,  $n$  is defined as the number of sampling values of the object to be measured during a given period.

(5) Calculate both  $\min_{i=1}^n \min_{k=1}^8 |x_0(k) - x_i(k)|$  and  $\max_{i=1}^n \max_{k=1}^8 |x_0(k) - x_i(k)|$ .

(6) Calculation of the relational coefficient through formula (33), the coefficient of the appropriate elements between every comparison sequence and reference sequence is calculated, respectively. Relational coefficient actually represents the degree of the difference between two curves in terms of geometry. Therefore, the degree of difference can reflect the degree of relationship:

$$\zeta_i(k) = \frac{T_1}{T_2}, \quad (33)$$

where

$$\begin{aligned} T_1 &= \min_i \min_k |x_0(k) - x_i(k)| \\ &+ \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|, \\ T_2 &= |x_0(k) - x_i(k)| \\ &+ \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|, \end{aligned} \quad (34)$$

where  $\rho$  is identification coefficient,  $0 < \rho < 1$ , and usually  $\rho = 0.5$ .

(7) Calculation of the degree of relationship.

Because the relation coefficient reflects the degree of relationship between comparison sequence and reference sequence at each moment. So, obviously there is more than one value and these values are dispersed. Therefore, it is necessary to use one value to reflect all of relation coefficient values moment. Here the average value is chosen to represent the degree of relationship between the comparison sequence and the reference sequence. The corresponding formula is as follows:

$$r_{0i} = \frac{1}{m} \sum_{k=1}^m \zeta_i(k). \quad (35)$$

## 6. Simulation and Results

In this paper, NetLogo simulation is the use of cloud computing mode virtual machines on the virtual platform to analyze

TABLE 2: Simulation parameters.

Items	Meanings
$N$	350 VMs on the physical computing node
$M$	Initial number of infected VM when virus outbreaks
Average-node-degree	Number of interacted VMs with given VM
$r_{0i}$	Deviation between predicted behaviour and expected behaviour
Time	Running time of simulation of the system
$\beta\%$	Percentage of malicious VMs versus the total VMs
Recovery chance	Recovery probability of the virtual computing nodes

our behavior-based trust measurement program. NetLogo is based modeling and integration of multiagent simulation environment, especially for the time evolution of complex systems modeling and simulation. The test environment is Intel Core Duo 2.36 g, and 4 G memory used NetLogo win7 runs to simulate the behavior of the virtual machine on a shared virtual platform, and we measure the effectiveness of the model checking virtual machine malicious behavior tested. Here the basic parameters are shown in Table 2.

A major function of the proposed scheme is to detect a variety of malicious behaviors of the virtual machine. To guarantee the trustworthiness of the group as much as possible, this paper uses the successful detection rate (abbreviated as MSR) of malicious behavior to reflect the detection ability of our scheme against malicious behaviors.

Within  $\Delta t$ , suppose there are  $b(t)$  computing nodes with malicious behavior and  $a(t)$  computing nodes with trusted behavior in the system, so that  $\beta\%$  can be described as follows:

$$\beta\% = \frac{b(t)}{a(t) + b(t)}. \quad (36)$$

This paper will simulate the attack process of “worm” virus, and then to test the effectiveness of our scheme by detecting the behavior of the “worm” virus. As a matter of fact, worm virus has the following characteristics such as breaking into antivirus software, compromising security model of the system, and implantation of Trojan into downloader. The virus typical invasion action [30] is denoted by

$$\text{Attack\_Behavior}. \quad (37)$$

According to the description of the behavior in Table 1, worm virus attacks can be abstracted as a behavioral vector:

$$\begin{aligned} \text{Attack\_Behavior} \\ = (MP_1, MP_2, MP_3, MP_4, MP_5, MP_6, MP_7, MP_8). \end{aligned} \quad (38)$$

In order to verify the effectiveness of the trusted measurement method of the behavior of the virtual machine here, we take the scheme without the decoupling proposed in literature [13] as contrast. In our experiments, the initial ratios of infected virtual machines are set as 30%, 50%, and 70%,

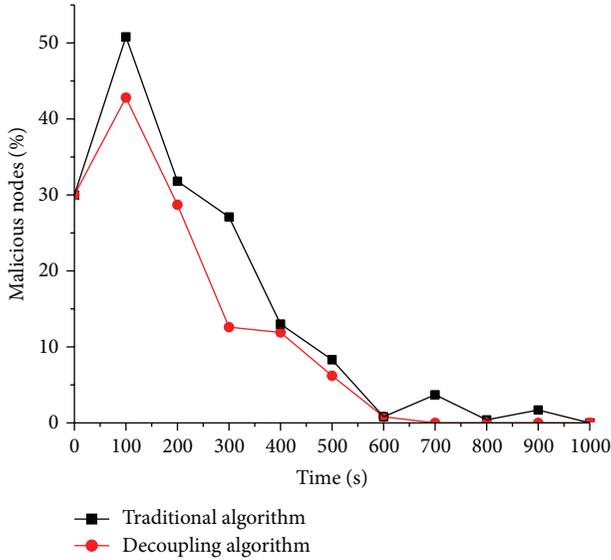


FIGURE 3: Percentage of the malicious suspicious virtual machine node is 30%.

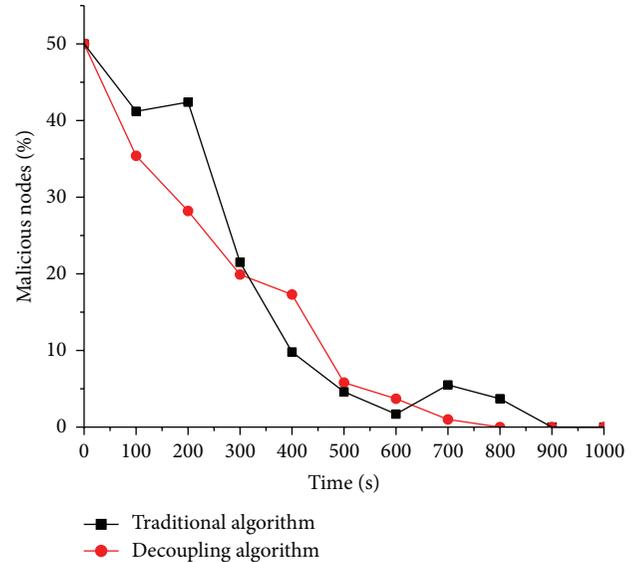


FIGURE 4: Percentage of the malicious suspicious virtual machine node is 50%.

respectively. For behavioral trusted measurement both with decoupling and without decoupling based on generalized predictive control, the experimental simulation are carried out three times.

When the percentages of malicious nodes are 30%, 50%, and 70%, the corresponding experimental results are shown from Figure 3 to Figure 5. After the analysis of Figures 3, 4, and 5, the following conclusions are summarized.

(1) Generally by the analysis of three figures, the simulation system for behavioral measurement model with decoupling can reach a steady state faster than the one without decoupling. The so-called steady state means such state that the number of the malicious nodes within the simulation system is 0. In our experiments, one of the parameters is the recovery chance that indicates the probability that infected node recovers as normal. In practical applications, finally the infected compute nodes recover as normal by various measurements, for example antivirus software. Faster to reach steady state means the corresponding scheme of behavioral trusted measurement is more accurate than the counterpart; that is, the user can detect and stop the spread of malicious worm virus timelier.

(2) In Figures 3, 4, and 5, the red line (decoupling algorithm) is almost below the black line (traditional algorithm), which indicates that, at any time, the scheme with decoupling proposed here can help accurately reflect the trusted state of the virtual machine and further take timely measurement so as to restrict the spread of the worm virus.

(3) In Figure 5, the distance between the red line (decoupling algorithm) and the black line (traditional algorithm) is larger than the previous two figures, which indicates, as the proportion of the malicious nodes in the system goes more, that the behavioral trusted measurement proposed here is better than the scheme in [13].

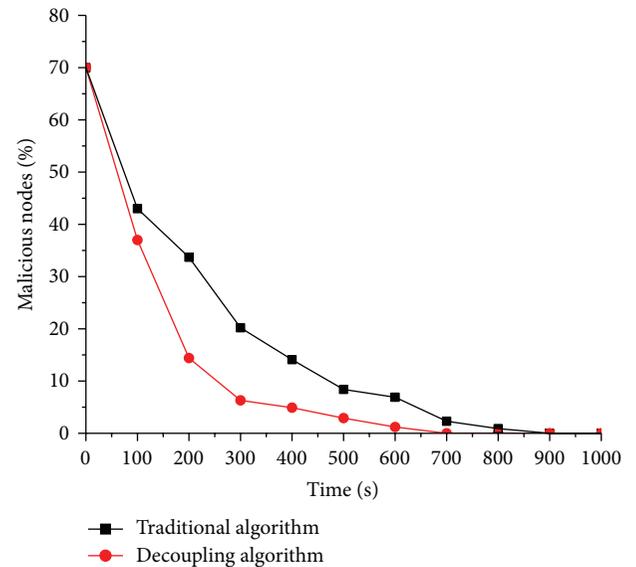


FIGURE 5: Percentage of the malicious suspicious virtual machine node is 70%.

In summary, the experimental simulation shows that trusted measurement scheme here can effectively predict and control the behaviors of the virtual machine. So that such attack behavior that results from the abuse of the resources in cloud computing can be found timely and well restricted; that is, the security of the entire group can be well guaranteed.

### 7. Conclusion

The scheme for trusted measurement over dynamic multi-tenant behavior in cloud computing environment put forward here addresses the problem of resource-sharing existing

in the cloud computing. By extending our previous model to the multiple tenants who share the same resource, we can further use the generalized predictive control to depict complicated behavior in cloud computing. Thanks to the advantages of generalized predictive control such as rolls optimized method and the feedback adjustment, the complicated behaviors of multitenants are well controlled. Further, the problems incurred by coupling between multitenants are solved effectively by the decoupling algorithm of generalized predictive control. As a result, the malicious behaviors between multitenants are restricted in cloud computing platform. In other words, our scheme avoids the threats introduced by multitenancy under cloud computing. In the future, we will refine our scheme and take into account the nonlinear behaviors between multiple tenants in order to deal with the behavior of tenants much more precisely.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgments

This work is partially supported by the program Major Projects of the Wireless Mobile Communications (2012ZX03002003) and the National Science Foundation of China (61003260, 61271275).

### References

- [1] H. Zang, J. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," *Optical Networks Magazine*, vol. 1, pp. 47–60, 2000.
- [2] Amazon Web Services, "Creating HIPAA-Compliant Medical Data Applications with Amazon Web Services," White Paper, 2009, [http://awsmedia.s3.amazonaws.com/AWS\\_HIPAA\\_Whitepaper\\_Final.pdf](http://awsmedia.s3.amazonaws.com/AWS_HIPAA_Whitepaper_Final.pdf).
- [3] E. Kanimozhi, "Trusted cloud—a solution for cloud cartography," *Journal of Global Research in Computer Science*, vol. 3, no. 11, pp. 44–51, 2012.
- [4] D. W. Clarke, C. Mohtadi, and P. S. Tuffs, "Properties of generalized predictive control," *Automatica*, vol. 25, no. 6, pp. 859–875, 1989.
- [5] G. Kim and E. Spafford, "The design and implementation of tripwire:a file system integrity checker," Tech. Rep., Purdue University, West Lafayette, Ind, USA, 1993.
- [6] W. Arbaugh, D. Farber, and J. Smith, "A secure and reliable bootstrap architecture," in *Proceedings of the IEEE Symposium on Security and Privacy*, IEEE Computer Society, Oakland, Calif, USA, 1997.
- [7] P. Iglie, "Trustedbox: a Kernellevel integrity checker," in *Proceedings of the 15th Annual Computer Security Applications Conference (ACSAC '99)*, IEEE Computer Society, Phoenix, AZ, USA, 1999.
- [8] N. Petroni, T. Fraser, J. Molina et al., "Copilot-a coprocessor-based Kernel runtime integrity monitor," in *Proceedings of the 13th Usenix Security Symposium*, USENIX Association, San Diego, Calif, USA, 2004.
- [9] E. Shi, A. Perrig, and L. Doom, "BIND: a fine-grained attestation service for secure distributed systems," in *Proceedings of the IEEE Symposium on Security and Privacy*, IEEE Computer Society, Oakland, Calif, USA, 2005.
- [10] T. Jaeger, R. Sailer, and U. Shankar, "PRIMA: policy-reduced integrity measurement architecture," in *Proceedings of the 11th ACM Symposium on Access Control Models and Technologies (SACMAT '06)*, Association for Computing Machinery, Lake Tahoe, Calif, USA, 2006.
- [11] X.-Y. Li, X.-D. Zuo, and C.-X. Shen, "System behavior based trustworthiness attestation for computing platform," *Chinese Journal of Electronics*, vol. 35, no. 7, pp. 1234–1239, 2007 (Chinese).
- [12] C.-X. Shen, H.-G. Zhang, F. Dengguo et al., "Survey of information security," *Science in China Series:E*, vol. 37, no. 2, pp. 129–150, 2007 (Chinese).
- [13] B. Gong, "The behavior measurement model based on prediction and control of trusted network," *Chinese Journal of Communication*, vol. 9, no. 5, pp. 117–128, 2012.
- [14] L. Zhuang, M. Cai, and C.-X. Shen, "Hierarchical verification of behavior trustworthiness," *Journal of Beijing University of Technology*, vol. 38, no. 9, pp. 1396–1401, 2012.
- [15] M. T. Khorshed, A. B. M. S. Ali, and S. A. Wasimi, "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing," *Future Generation Computer Systems*, vol. 28, no. 6, pp. 833–851, 2012.
- [16] X.-Y. Li, X.-L. Gui, Q. Mao, and D.-Q. Leng, "Adaptive dynamic trust measurement and prediction model based on behavior monitoring," *Chinese Journal of Computers*, vol. 32, no. 4, pp. 664–674, 2009.
- [17] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*, pp. 199–212, 2009.
- [18] Amazon, "Amazon elastic compute cloud (AmazonEC2)," 2011, <http://aws.amazon.com/ec2/>.
- [19] F. Rocha and M. Correia, "Lucy in the sky without diamonds: stealing confidential data in the cloud," in *Proceedings of the IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W '11)*, pp. 129–134, 2011.
- [20] A. Chonka, Y. Xiang, W. Zhou, and A. Bonti, "Cloud security defence to protect cloud computing against HTTP-DoS and XML-DoS attacks," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1097–1107, 2011.
- [21] D. Danchev, "DanchoDanchev'sblog—mind streams of information security knowledge," 2011, <http://ddanchev.blogspot.com/>.
- [22] J. Grossman, "Jeremiah Grossman," 2011, <http://jeremiahgrossman.blogspot.com/>.
- [23] D. Danchev, "Coordinated Russia vs Georgia cyber attack in progress," 2008.
- [24] D. Danchev, "The DDoS attack against CNN.com," 2008.
- [25] Grossman, "Cross-site scripting worms and viruses," Whitehat Security, 2006.
- [26] M. T. Khorshed, A. B. M. S. Ali, and S. A. Wasimi, "Classifying different denial-of-service attacks in cloud computing using rule-based learning," *Security and Communication Networks*, vol. 5, no. 11, pp. 1235–1247, 2012.
- [27] T. Y. Chai, K. Z. Mao, and X. F. Qin, "Decoupling design of multivariable generalised predictive control," *IEE Proceedings*:

*Control Theory and Applications*, vol. 141, no. 3, pp. 197–201, 1994.

- [28] M. Li, X. Chen, M. L. Xin et al., “The similarity metric,” in *Proceedings of the IEEE Transactions Information Theory*, pp. 863–872, 2003.
- [29] J. Deng, *The Control Problems of Grey Systems*, Huazhong University of Science & Technology, 1993.
- [30] B. Gong, *Trusted Network Architecture Supporting Trusted Group Establishment and Key Technologies Research*, Beijing University of Technology, 2012.

## Research Article

# Transition Characteristic Analysis of Traffic Evolution Process for Urban Traffic Network

**Longfei Wang, Hong Chen, and Yang Li**

*School of Highway, Chang'an University, Xi'an 710064, China*

Correspondence should be addressed to Longfei Wang; [longfei.wanglf@163.com](mailto:longfei.wanglf@163.com)

Received 24 January 2014; Accepted 19 February 2014; Published 27 March 2014

Academic Editors: X. Meng and J. Zhou

Copyright © 2014 Longfei Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The characterization of the dynamics of traffic states remains fundamental to seeking for the solutions of diverse traffic problems. To gain more insights into traffic dynamics in the temporal domain, this paper explored temporal characteristics and distinct regularity in the traffic evolution process of urban traffic network. We defined traffic state pattern through clustering multidimensional traffic time series using self-organizing maps and construct a pattern transition network model that is appropriate for representing and analyzing the evolution progress. The methodology is illustrated by an application to data flow rate of multiple road sections from Network of Shenzhen's Nanshan District, China. Analysis and numerical results demonstrated that the methodology permits extracting many useful traffic transition characteristics including stability, preference, activity, and attractiveness. In addition, more information about the relationships between these characteristics was extracted, which should be helpful in understanding the complex behavior of the temporal evolution features of traffic patterns.

## 1. Introduction

Traffic congestion is increasingly becoming a serious problem for densely populated cities throughout the world. Understanding the dynamics of traffic states remains fundamental to solving diverse traffic problems. A traffic state refers to an overall description of the working state of a transportation system or its subsystems at a certain time. The present work on analysis for urban traffic state pays a considerable amount of attention to evolution analysis within region urban network [1–3]. Many researchers want to understand the dynamics of traffic transitions during the traffic evolution process in the traffic network and the conditions in which various traffic transitions emerge.

A common understanding of the fluctuations and transitions of traffic flow between different traffic regimes depends on many factors, including traffic demand, capacity of roads to carry vehicles, and the topology or structure of the network [4]. Recent studies that use differing analytical approaches have found useful characteristics and patterns in dynamical traffic evolution on urban traffic network. In these studies,

the time series is widely introduced to represent the traffic state in a region by a multidimensional traffic flow vector  $F(t)$ . Evaluation of the traffic network state requires analysis of how the time series of  $F(t)$  evolves with time. Reference [5] uses Kohonen's self-organizing maps (SOMs) [6] for multidimensional time series analysis. The SOM serves as a clustering algorithm as well as a dimensionality-reduction technique. Analysis of real-world traffic data shows that the method can capture the nonlinear information of traffic flows data and predict traffic flows on multiple links simultaneously. Another method [7], under the assumption that various time series representing daily cycles of traffic may be nonlinear, uses smooth-transition regression (STR) models to characterize distinct regimes for free flow, congestion, and asymmetric behavior in the transition phases from free flow to congestion and vice versa. Application tests on nonlinearity provided ample evidence of regime-dependent dynamics for all traffic time series examined. Reference [8] proposes an image-based method of observing and analyzing traffic state over large road network. Urban traffic network is mapped to a pseudocolor image to vividly represent the macroscopic

traffic state. The evolutionary patterns of traffic state are determined by calculating and analyzing the optical flow field of consecutive pseudocolor images; thus, the congested regions can be found automatically.

Recent studies on traffic state analysis, including validation of traffic flow models [9], investigation of traffic patterns [10], recognition of traffic congestion propagation [11], and classification of urban traffic state [12], demonstrate that dynamical properties of traffic flow on urban traffic networks have some traffic patterns during different periods of a day. The traffic pattern, generally defined as a set of traffic states with characteristics of traffic flow, average vehicle speed, and occupancy, represents a particular state with the same special and temporal characteristics [13]. Many authors emphasize the importance of finding models by partitioning the traffic states into patterns and using fuzzy reasoning [14], clustering analysis [15, 16], dimensionality reduction [17], and data fusion [18] to investigate their properties. However, previous studies on pattern-oriented traffic state analysis lack attention to the transition characteristic of traffic state pattern during traffic evolution process.

Because of the shortage of studies that consider the transition characteristic between traffic patterns in the evolution process on urban traffic network, the main objective of this study is to investigate the temporal characteristics and distinct regularity in the traffic evolution process from the viewpoint of the whole network. Our particular interest is to understand when and how the traffic state transitions occur on traffic networks with the fluctuations of traffic flow.

In this paper, we construct a transition network model of traffic state pattern and have attempted to gain in-depth insights into the evolution characteristic of traffic time series. Furthermore, a close look at transition characteristic of traffic state pattern not only provides useful information for application in ITS, such as incident detection, extended delay prevention, and traffic control scheme, but also improves upon the shortage of conventional models that lack measurement and quantization of traffic evolution. Besides what was previously stated, we believe that this analysis provides direction for new steps in understanding the complex behavior of the temporal evolution features of traffic patterns. To illustrate the methodology, we use flow rate data of multiple road sections from Network of Shenzhen's Nanshan District, China.

The following content of the paper begins with the rationale for methodology, followed by a brief on data acquisition and a discussion of analytical results. We conclude with an elaboration on the implications of our research and suggestions for a future agenda.

## 2. Traffic State Network Analysis Model

In order to conduct quantitative analysis of the traffic evolution process, a base network model for traffic state analysis of urban regional network is constructed to describe the transition relationship of traffic states. Here we introduce the following notions.

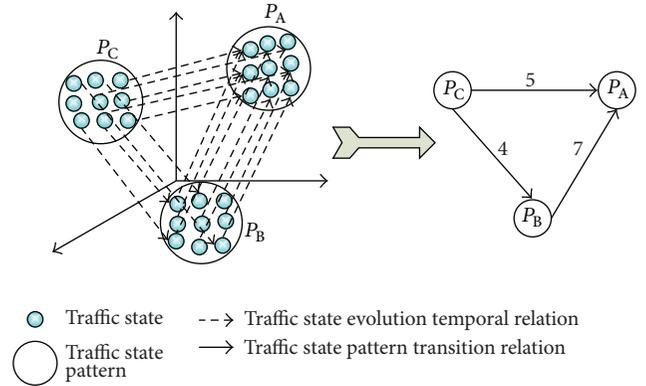


FIGURE 1: The structure diagram of TSPTRN.

*Definition 1.* The regional traffic state during time interval  $t$  can be represented by a  $d$ -dimensional vector, as shown in

$$F(t) = [f_1(t), f_2(t), \dots, f_n(t)]^T, \quad (1)$$

where  $f_i(t)$  represents the flow rate of the  $i$ th section within the time period  $t$  and  $n$  represents the total number of all sections.  $F(t)$  represents the traffic state in this time interval. As the traffic state changes with time  $t$ , the flow rates on the different links vary as reflected in the multidimensional time series  $F(t)$ .

Similar traffic states could repeat occurrences at different times. This leads to the idea of performing clustering on the traffic state to identify large clusters as traffic patterns.

*Definition 2.* According to cluster analysis of the total gathered traffic state data set, multiple traffic state classifications are obtained. Each classification is a set of traffic states, which is defined as traffic state pattern.

*Definition 3.* If  $P_A$  and  $P_B$  are traffic state patterns,  $F(t_1) \in P_A$ ,  $F(t_2) \in P_B$ , and  $t_1$  and  $t_2$  are adjoining times, then there is a traffic evolution temporal relation between  $F(t_1)$  and  $F(t_2)$ . In addition, there is traffic state pattern transition relation (TSPTR) between  $P_A$  and  $P_B$ , denoted by  $R_{A \rightarrow B} = P_A \rightarrow P_B$ .

*Definition 4.* The directed network, with traffic state pattern as vertex and traffic state pattern transition relation as edge, is called traffic state pattern transition relation network (TSPTRN), denoted by  $G = (P, R)$ , where  $P$  represents all traffic state pattern sets, and  $R$  represents all transition relation sets.

TSPTRN can be constructed by traffic state data accumulated within a period of time. The weight  $w_{A \rightarrow B}$  of  $R_{A \rightarrow B} = P_A \rightarrow P_B$  is determined by the total number of evolution temporal relations between all traffic states of  $P_A$  and  $P_B$ . Figure 1 shows the structure of TSPTRN, with transition relations of three traffic state patterns, namely,  $P_A$ ,  $P_B$ , and  $P_C$ .

### 3. Analysis of the Features of Traffic State Pattern Transition

In this section, we use the TSPTRN model to analyze the traffic transition characteristics of state pattern transition in the traffic evolution process, including stability, preference, activity, and attractiveness.

3.1. *The Stability of Traffic State Pattern.* The stability of traffic state pattern consists of the following two parts:

- (1) *balance ability:* the ability to maintain one state,
- (2) *recovery ability:* the ability to transit back to the original pattern in the next period after having transited from one pattern to another.

*Definition 5.* From time period 1 to  $N$ , traffic operation experiences traffic state sequence  $L = S_1, S_2, \dots, S_N$ . If all of these  $N$  states belong to traffic state pattern  $P$ , that is, pattern  $P$  remains the same for  $N$  consecutive time periods, then  $L$  is called the consecutive state sequence of  $P$ , denoted by  $L_P$ ; set  $N$  as the consecutive coefficient of  $L_P$ , denoted by  $C_{L_P}$ .

*Definition 6.* Set  $L = S_1, S_2, \dots, S_N$  as the consecutive state sequence of traffic state pattern  $P$ , if it meets all of the following conditions:

- (1)  $L$  has states that belong to  $P$  and also has states that do not belong to  $P$ ;
- (2)  $S_1$  and  $S_N$  belong to  $P$ ;
- (3) states in  $L$  that do not belong to  $P$  are not consecutive.

Then  $L$  is called the approximate consecutive state sequence of  $P$ , denoted by  $\tilde{L}_P$ . In  $\tilde{L}_P$ , the number of states that do not belong to  $P$  is denoted by  $M$ , and set the consecutive coefficient of  $\tilde{L}_P$  as  $C_{\tilde{L}_P}$ , as shown in

$$C_{\tilde{L}_P} = N \times (1 - \omega)^M. \quad (2)$$

The calculation method for the stability of traffic state pattern  $P$  is as follows: first, isolate all consecutive state sequences and approximate consecutive state sequences of  $P$  in the total data set; second, denote the arithmetic mean of consecutive coefficients of all these sequences by the stability coefficient  $\gamma_S^P$  of  $P$ , as shown in formula (3), where  $N$  and  $M$  are the total number of all consecutive state sequences and all approximate consecutive state sequences, respectively. With greater  $\gamma_S^P$  value,  $P$  is more stable. Consider

$$\gamma_S^P = \frac{\sum_{i=1}^N C_{L_P}^i + \sum_{j=1}^M C_{\tilde{L}_P}^j}{N + M}. \quad (3)$$

3.2. *Preference of Traffic State Pattern.* Preference refers to the tendency of traffic state pattern transiting from one to another. For traffic state pattern  $P_A$ , if the transition time among traffic states within  $P_A$  is  $\hat{n}_P$  and transition time among traffic states between  $P_A$  and  $P_B$ ,  $P_C$ , and  $P_D$  is  $n_{A \rightarrow B}$ ,  $n_{A \rightarrow C}$ ,

and  $n_{A \rightarrow D}$ , respectively, then the probability of  $P_A$  remaining unchanged in the next period is shown in

$$TP_{P_A}^{\text{In}} = \frac{\hat{n}_P}{\hat{n}_P + n_{A \rightarrow B} + n_{A \rightarrow C} + n_{A \rightarrow D}} \times 100\%. \quad (4)$$

The probability for  $P_A$  transiting to other traffic state patterns in the next period is shown in

$$TP_{P_A}^{\text{Out}} = \frac{n_{A \rightarrow B} + n_{A \rightarrow C} + n_{A \rightarrow D}}{\hat{n}_P + n_{A \rightarrow B} + n_{A \rightarrow C} + n_{A \rightarrow D}} \times 100\%. \quad (5)$$

If  $P_A$  transits to another pattern in the next period, the probabilities of transiting to  $P_B$ ,  $P_C$ , and  $P_D$  (i.e., the transition preference of  $P_A$  to  $P_B$ ,  $P_C$ , and  $P_D$ ), taking  $P_B$  as an example, are shown in

$$TP_{P_A \rightarrow P_B|\text{Out}}^{\text{Out}} = \frac{n_{A \rightarrow B}}{n_{A \rightarrow B} + n_{A \rightarrow C} + n_{A \rightarrow D}} \times 100\%. \quad (6)$$

According to a conditional probability formula, the probability of  $P_A$  transiting to  $P_B$  in the next period is shown in

$$TP_{P \rightarrow A}^{\text{Out}} = TP_{P_A \rightarrow P_B|\text{Out}}^{\text{Out}} \times TP_{P_A}^{\text{Out}}. \quad (7)$$

3.3. *The Activity of Traffic State Pattern.* The out-degree of  $P_A$  represents its activity. If the out-degree is higher, then  $P_A$  is easier to transit to other state patterns with higher activity.

Define the external transition times of traffic state pattern  $P_i$  as  $TO_{P_i} = \sum_{j=1}^{M_i} \omega_{P_i}^j$ , where  $\omega_{P_i}^j$  is the weight of each outgoing edge  $P_i$ , and  $M_i$  is the out-degree of  $P_i$ . Rank the times of external transition of all traffic state patterns in an ascending order, and then the distribution probability  $P_{TO_{P_i}}$  for each of external transition times can be obtained, as in

$$P_{TO_{P_i}} = \frac{TO_{P_i} \times n_{P_i}}{\sum_{i=1}^N TO_{P_i} \times n_{P_i}}, \quad (8)$$

where  $N$  is the total number of different external transition times and  $n_{P_i}$  is the number of all traffic state patterns with the same external transition time as  $TO_{P_i}$ ; thus, the expected value of each external transition time could be obtained, as in

$$E(TO_{P_i}) = \sum_{i=1}^N TO_{P_i} \times P_{TO_{P_i}}. \quad (9)$$

Define the activity coefficient of  $P_i$  as  $\gamma_A^{P_i}$ , as in formula (10). The greater is the coefficient value, the stronger is the activity. Consider

$$\gamma_A^{P_i} = e^{(TO_{P_i} - E(TO_{P_i})) / \text{Max}(TO_{P_i})}. \quad (10)$$

3.4. *The Attractiveness of Traffic State Pattern.* The in-degree of  $P_A$  represents the attractiveness of  $P_A$ . If the in-degree of  $P_A$  is higher, then other state patterns are easier to transit to  $P_A$  with attractiveness.

Define the internal transition times of traffic state pattern  $P_i$  as  $TI_{P_i} = \sum_{j=1}^{M_i} \omega_{P_i}^j$ , where  $\omega_{P_i}^j$  is the weight of each incoming edge of  $P_i$ , and  $M_i$  is the in-degree of  $P_i$ . Rank the internal

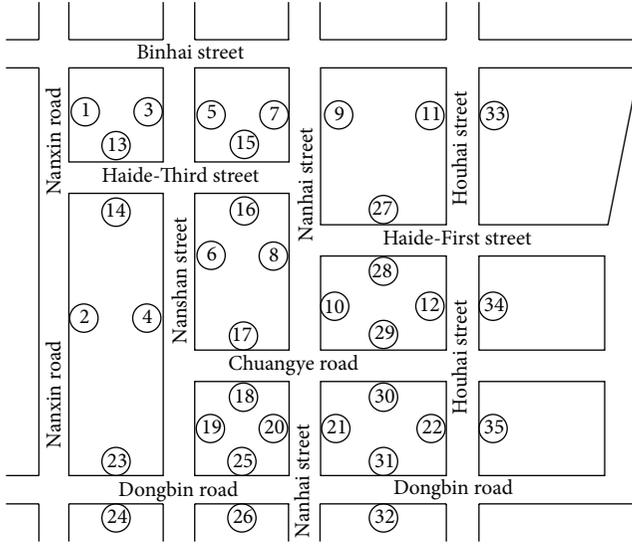


FIGURE 2: Schematic diagram of Nanshan road network topology.

transition times of all traffic state patterns in an ascending order, and then the distribution probability  $P_{TI_{P_i}}$  of each internal transition times could be obtained, as in

$$P_{TI_{P_i}} = \frac{TI_{P_i} \times n_{P_i}}{\sum_{i=1}^N TI_{P_i} \times n_{P_i}}, \quad (11)$$

where  $N$  is the total number of different internal transition times and  $n_{P_i}$  is the number of all traffic state patterns with the same internal transition times as  $TI_{P_i}$ ; thus, the expected value of each internal transition times could be obtained, as

$$E(TI_{P_i}) = \sum_{i=1}^N TI_{P_i} \times P_{TI_{P_i}}. \quad (12)$$

Define the attractiveness coefficient of  $P_i$  as  $\gamma_B^{P_i}$ , as in formula (13). The greater is coefficient value, the bigger is the attractiveness of the activity. Consider

$$\gamma_B^{P_i} = e^{(TI_{P_i} - E(TI_{P_i})) / \text{Max}(TI_{P_i})}. \quad (13)$$

## 4. Experiment

This paper examines the road network of Shenzhen's Nanshan District to record the flow rates (vehicles per hour (veh/h)) of each road segment every 15 min. The road network map is sketched in Figure 2. The network flow rates of 35 road segments ( $d = 35$ ) were measured from June 1st to June 30th, 2011. Thus, the experimental dataset consists of flow rates for 6,480 consecutive time intervals on 35 links, defining a 35-dimensional series with a length of 6,480.

**4.1. Analysis of Traffic State Pattern Transition.** This experiment uses Kohonen's self-organizing maps (SOMs) [5] for the multidimensional time series analysis. A well-trained  $8 \times 8$  SOM network is used to cluster all the data. The SOM here

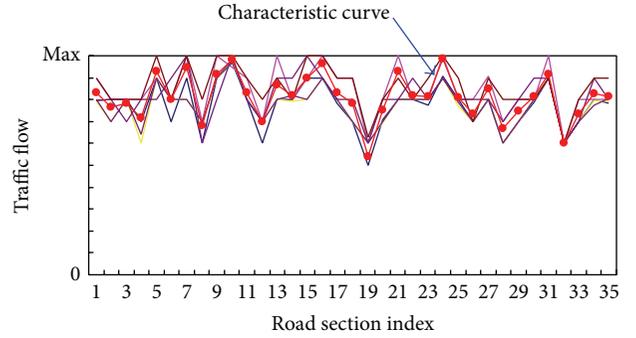


FIGURE 3: The characteristic curve of traffic state mode.

serves as a clustering algorithm as well as a dimensionality-reduction technique when applied to cluster the traffic flow vectors. After the SOM was trained, 64 clusters were obtained. Each cluster represents a traffic state pattern with 35 dimensions. Through fitting all the traffic state curves within each pattern using mean fitting method, we get the characteristic curve of each traffic state pattern, as shown in Figure 3.

The distribution diagram of all traffic state patterns is shown in the SOM topological grid in Figure 4 and each pattern is dyed by a special color and has a unique characteristic curve. It could be observed that adjacent clusters represent similar patterns and each pattern changes gradually along the SOM grid. Four pattern groups A, B, C, and D are shown in this figure, respectively. For pattern group A of "blue clusters," all links have very large flow rates and even traffic distribution. For pattern group B of "pink clusters," the flow rates are a little lower and less evenly distributed among the links, compared to the "blue clusters" of area A. For pattern group C of "yellow clusters," the links have moderate flow rates and uneven traffic distribution, with several links in some clusters having high flow rates in spite of the low flow rates in other links. For pattern group D of "green clusters," the links have small flow rates and even traffic distribution.

The traffic state pattern transition relation network (TSP-TRN) of Shenzhen's Nanshan District is drawn in Figure 5. We can clearly see all the transition relations of total 64 patterns. By statistical analysis to TSP-TRN, all traffic state patterns, transition times, major time, and traffic features of four pattern groups A, B, C, and D are obtained, respectively, as shown in Table 1.

According to Table 1, the distribution time of A, B, C, and D pattern groups is relatively fixed, indicating that macroscopic traffic operation of the road network has strong regularity and the traffic operation is stable within a certain period of time. Because the distribution time directly corresponds to the transition time of traffic state, traffic state transition times of groups A and B are significantly larger than those of groups C and D.

According to Table 1 and Figure 5, pattern transition is mainly distributed in groups A and B. This is because traffic patterns of groups A and B account for 77.8% of that in the whole day, and thus the pattern transition time is relatively larger while that of groups C and D is mainly in the early

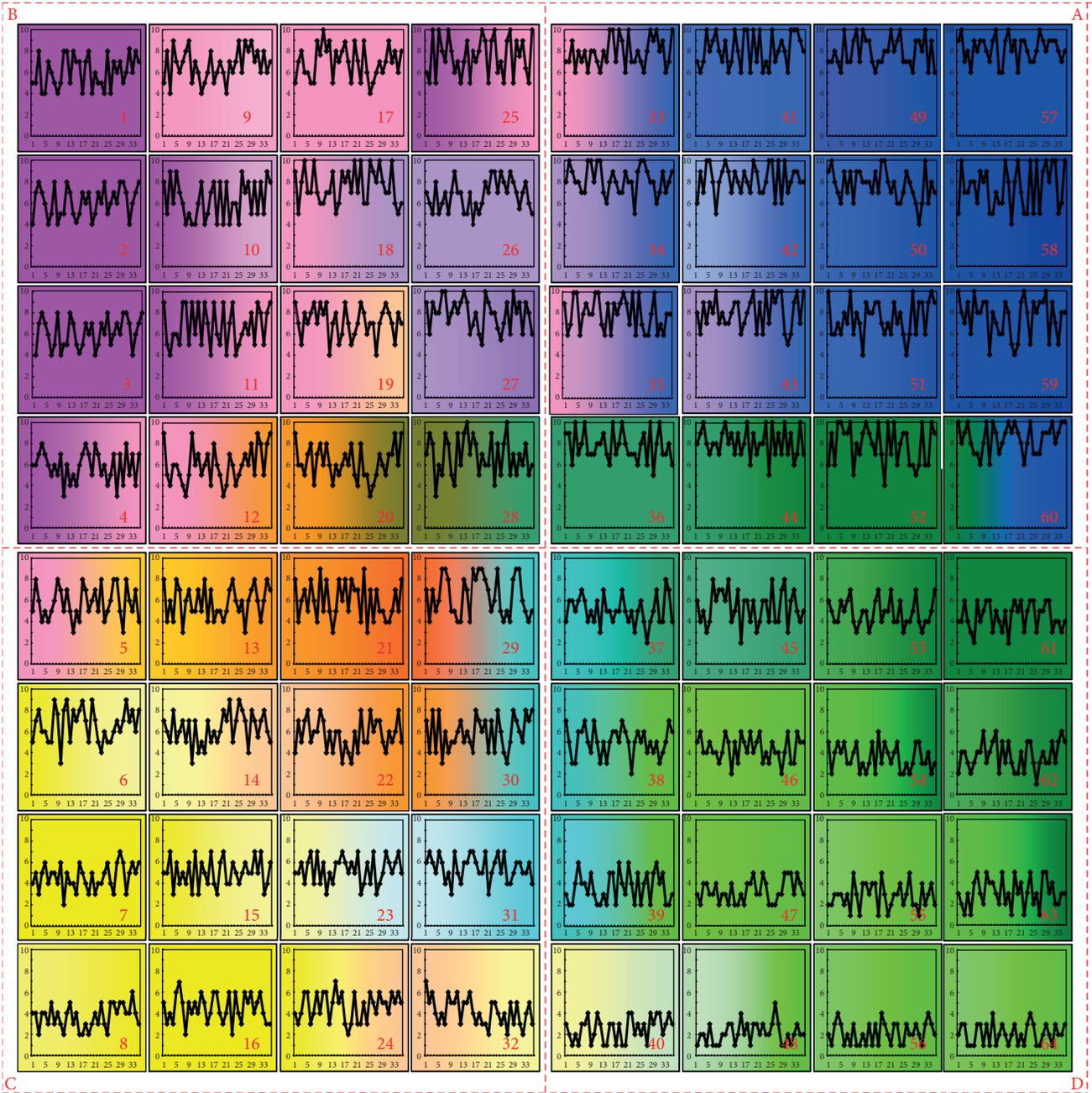


FIGURE 4: Schematic diagram of 64 traffic state patterns.

morning and at night, and thus the pattern transition time is relatively smaller.

However, for the proportion of pattern transition among pattern groups, the proportions of groups B and C are slightly greater than those of groups A and D. This is because groups A and D are, respectively, at peak hours and free travel period with relatively stable traffic demand and traffic distribution without significant disturbance.

In addition, the connection between pattern groups is also an area with frequent transition, accounting for 19.23%

of the total pattern transition, and major transition occurs between groups A and B, accounting for 89.2% of the total critical transition.

4.2. Analysis of the Characteristics of TSPTRN. The detailed characteristics data of each traffic state pattern within TSPTRN are shown in Figure 6. The relations between 4 traffic transition characteristics coefficients are, respectively, shown in Figure 7.

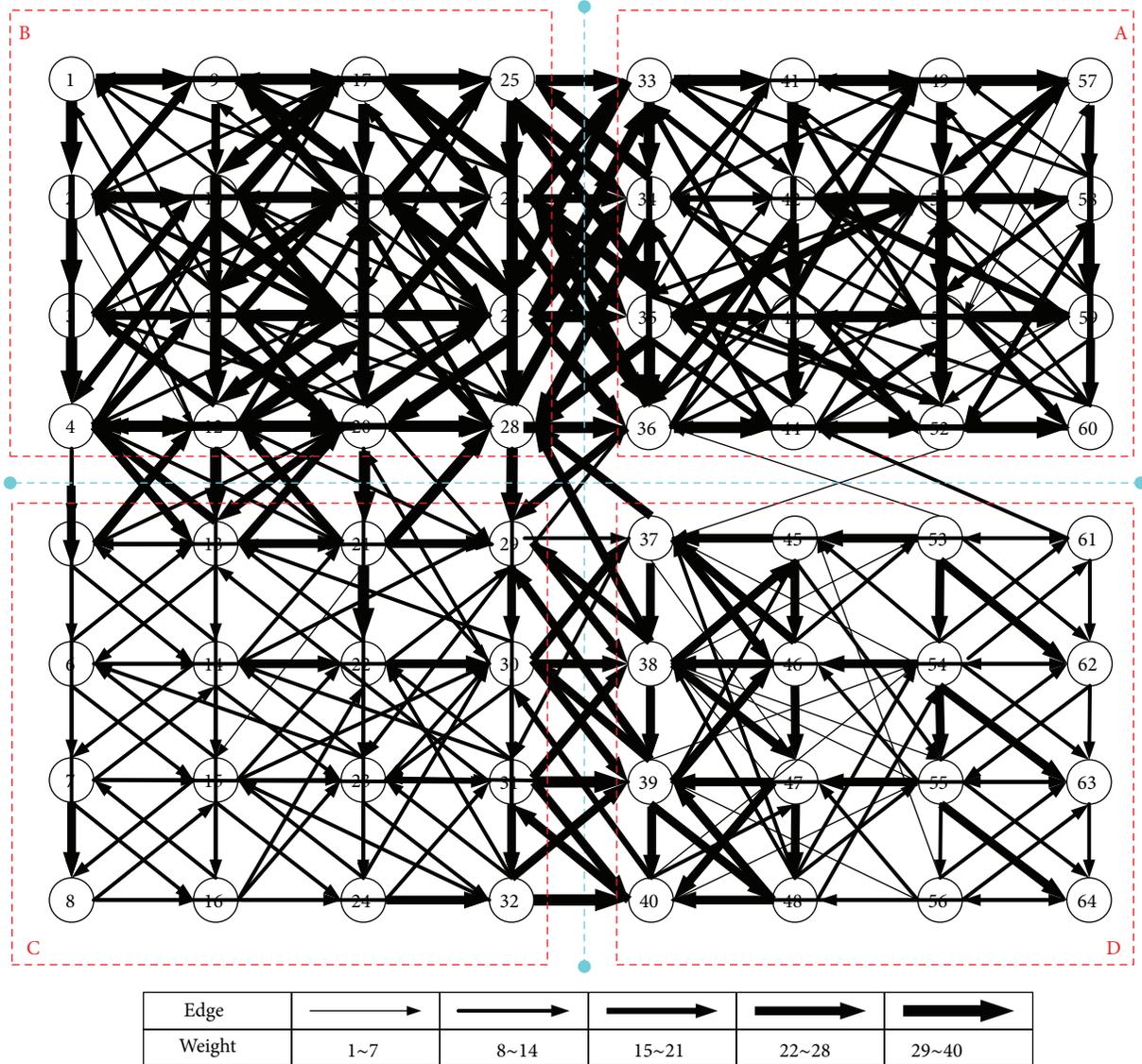


FIGURE 5: Traffic state pattern transition relation network (TSPTRN) of Shenzhen's Nanshan District.

TABLE 1: A detailed table of the transition characteristics of pattern groups.

Pattern group	Traffic state transition times	Pattern transition times	Pattern transition ratio	Occurrence time	Time distribution
A	2325	1528	65.7%	7:30~11:00 16:00~19:30	38.89%
B	2329	1740	74.7%	7:00~7:30 11:00~16:00 19:30~21:00	38.89%
C	850	649	76.3%	6:30~7:00 21:00~22:00	8.33%
D	976	670	68.6%	6:00~6:30 22:00~24:00	13.89%
Total average	6480	4588	70.8%		

TABLE 2: Table of detailed characteristics of pattern group transition.

Pattern group	Average external transition probability	Average stability coefficient	Average activity coefficient	Average attractiveness coefficient
A	0.6737	4.02	0.6243	1.3476
B	0.7703	3.24	1.3518	0.7668
C	0.7526	2.89	1.3452	0.6592
D	0.7039	3.71	0.8347	1.2858
Total average	0.7302	3.46	1.042	1.017

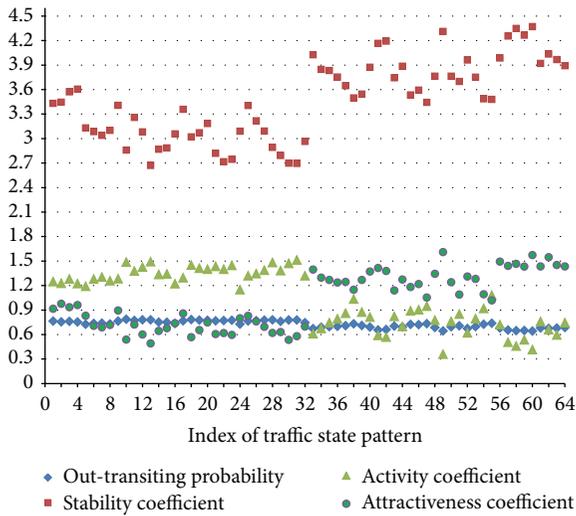


FIGURE 6: Characteristic values distribution of traffic state pattern transition.

There are obvious linear increasing or decreasing relationships between these characteristics. We can clearly see that the stability coefficient increases with the attractiveness coefficient and that both of them decrease with the out-transiting probability and the activity coefficient. This illustrates that, in the traffic state pattern transition process, the stability and attractiveness, which reflect the static characteristic, are in opposition to out-transiting probability and activity, which reflect the dynamic characteristic.

The mean values of each traffic state pattern group within TSPTRN are shown in Table 2, respectively.

According to Table 2, the external transition probability of the traffic state pattern of group A is relatively smaller than that of others, while its stability coefficient and attractiveness coefficient are relatively larger. This is because group A is in the morning peak and evening peak hours, the traffic state of each road is in the “saturated” and “nearly saturated” condition, and the change interval of the traffic state is limited.

The traffic state pattern of group B is near the morning peak and evening peak, with a very unstable traffic operation state. Thus, it has a larger average external transition probability, poorer stability, stronger average activity, and a relatively smaller attractiveness coefficient.

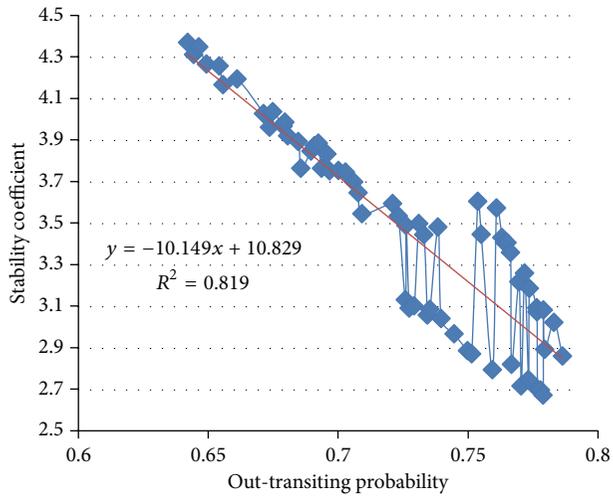
The traffic state patterns of group C are before and after the patterns of group B. During these periods, the traffic operation state is at a transition stage from instability to stability with significant changes. So the traffic operation state has a relatively larger average external transition probability and activity coefficient and a smaller stability and attractiveness coefficient.

The traffic state pattern of group D happens in the early morning and late at night. In these periods, the traffic is basically at the state of free travel, and the traffic operation state is quite smooth and stable. Thus, it has a larger stability coefficient and attractiveness coefficient and a relatively smaller average external transition probability and activity coefficient.

### 5. Conclusions

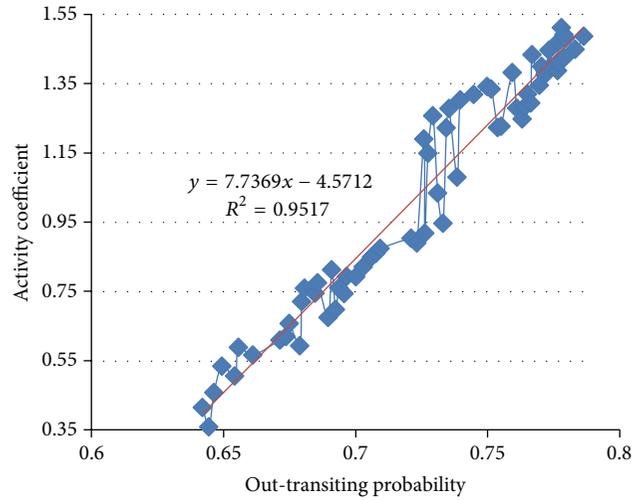
In this paper, we have investigated the transition characteristics and distinct regularity of traffic state pattern during the traffic evolution process from the viewpoint of the whole network. A transition network model of traffic state pattern is constructed, which could facilitate gaining in-depth insights into the evolution characteristic of traffic time series. According to our empirical results, the proposed analytical method permits extracting more information on traffic transition characteristics in the constructed traffic state pattern spaces, particularly including stability, preference, activity, and attractiveness. These favorable features of our method make it a potentially powerful tool for traffic evolution analyses of urban regional networks.

However, in contrast with the temporal transition characteristics of traffic state patterns that occur only in time, spatiotemporal transition characteristics that occur in both time and space can also be investigated. Methodologies taking topology structure and traffic demand into account may gain usefulness, so we intend to further investigate other spatiotemporal transition characteristics in future research.



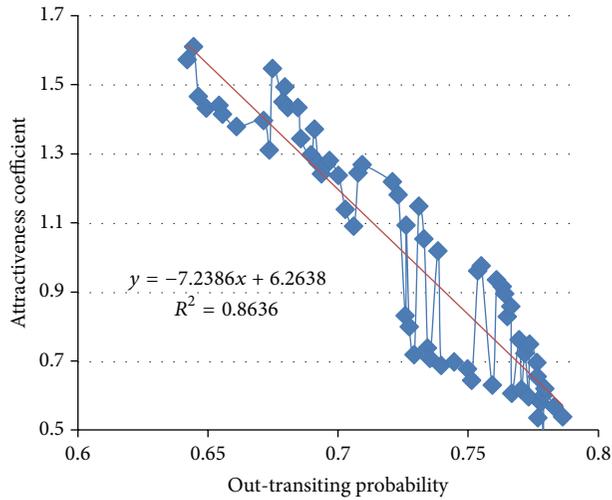
◆ Traffic state pattern  
— Linear (traffic state pattern)

(a)



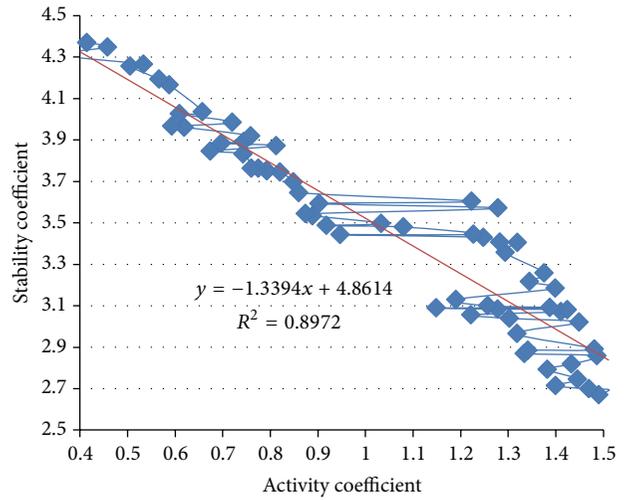
◆ Traffic state pattern  
— Linear (traffic state pattern)

(b)



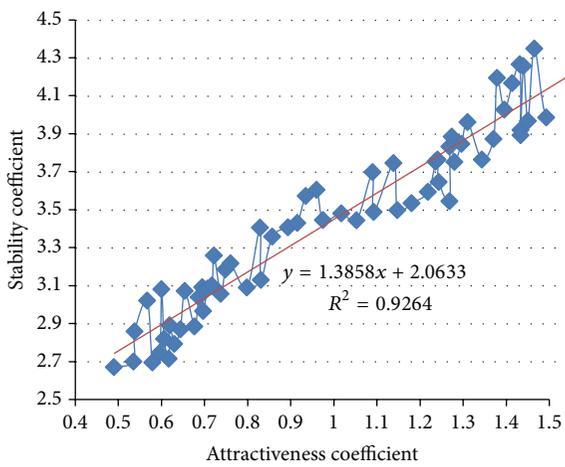
◆ Traffic state pattern  
— Linear (traffic state pattern)

(c)



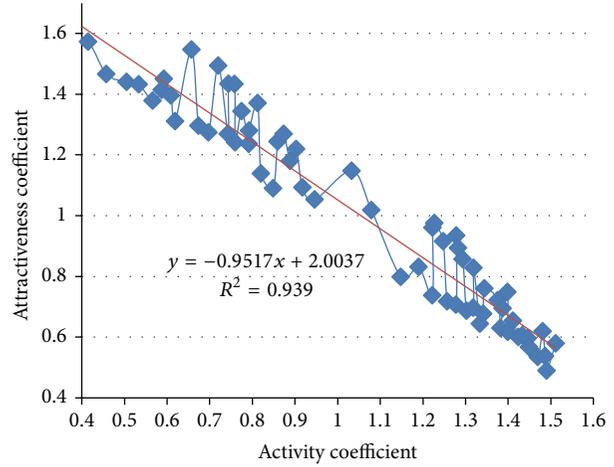
◆ Traffic state pattern  
— Linear (traffic state pattern)

(d)



◆ Traffic state pattern  
— Linear (traffic state pattern)

(e)



◆ Traffic state pattern  
— Linear (traffic state pattern)

(f)

FIGURE 7: The relationships between traffic transition characteristics coefficients.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This research work is supported by the Fundamental Research Funds for the Central Universities under Grant nos. CHD2011JC191, CHD2010JC124, and CHD2011ZY006. It is also supported by the National Natural Science Foundation of China under Grant no. 51208054.

## References

- [1] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Temporal evolution of short-term urban traffic flow: a nonlinear dynamics approach," *Computer-Aided Civil and Infrastructure Engineering*, vol. 23, no. 7, pp. 536–548, 2008.
- [2] Z. Hesheng, Z. Yi, and H. Dongcheng, "Study on method of traffic state analysis for urban traffic network," *Intelligent Transportation System*, vol. 1, pp. 23–27, 2006.
- [3] E. Azimirad, N. Pariz, and M. B. N. Sistani, "A novel fuzzy model and control of single intersection at urban traffic network," *IEEE Systems Journal*, vol. 4, no. 1, pp. 107–111, 2010.
- [4] B. Shen and Z. Y. Gao, "Dynamical properties of transportation on complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 387, no. 5–6, pp. 1352–1360, 2008.
- [5] Y. Chen, Y. Zhang, and J. Hu, "Multi-dimensional traffic flow time series analysis with self-organizing maps," *Tsinghua Science and Technology*, vol. 13, no. 2, pp. 220–228, 2008.
- [6] T. Kohonen, "Self-organizing maps of massive databases," *International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications*, vol. 9, no. 4, pp. 179–185, 2001.
- [7] Y. Kamarianakis, H. Oliver Gao, and P. Prastacos, "Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions," *Transportation Research C: Emerging Technologies*, vol. 18, no. 5, pp. 821–840, 2010.
- [8] H. L. Duan, Z. H. Li, L. Li, Y. Zhang, and S. C. Yin, "Network-wide traffic state observation and analysis method using pseudo-color map," *Journal of Transportation Systems Engineering and Information Technology*, vol. 9, no. 4, pp. 46–52, 2009.
- [9] M. Treiber and A. Kesting, "Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns," *Transportation Research C: Emerging Technologies*, vol. 21, no. 1, pp. 31–41, 2012.
- [10] L. W. Lan, J. B. Sheu, and Y. S. Huang, "Investigation of temporal freeway traffic patterns in reconstructed state spaces," *Transportation Research C: Emerging Technologies*, vol. 16, no. 1, pp. 116–136, 2008.
- [11] B. S. Kerner, H. Rehborn, M. Aleksic, and A. Haug, "Recognition and tracking of spatial-temporal congested traffic patterns on freeways," *Transportation Research C: Emerging Technologies*, vol. 12, no. 5, pp. 369–400, 2004.
- [12] Q. Q. Li, D. Q. Gao, and B. S. Yang, "Urban road traffic status classification based on fuzzy support vector machines," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 39, no. 2, pp. 131–134, 2009.
- [13] Z. H. Li, D. Sun, X. X. Jin, D. Yu, and Z. Zhang, "Pattern-based study on urban transportation system state classification and properties," *Journal of Transportation Systems Engineering and Information Technology*, vol. 8, no. 5, pp. 83–87, 2008.
- [14] Z. H. Li, D. Sun, X. X. Jin, D. Yu, and Z. Zhang, "Pattern-based study on urban transportation system state and properties with fuzzy reasoning methods," *Journal of Transportation Systems Engineering and Information Technology*, vol. 8, no. 5, pp. 83–87, 2008.
- [15] A. Lozano, G. Manfredi, and L. Nieddu, "An algorithm for the recognition of levels of congestion in road traffic problems," *Mathematics and Computers in Simulation*, vol. 79, no. 6, pp. 1926–1934, 2009.
- [16] M. Montazeri-Gh and A. Fotouhi, "Traffic condition recognition using the k-means clustering method," *Scientia Iranica*, vol. 18, no. 4, pp. 930–937, 2011.
- [17] C. Yudong, Z. Yi, H. Jianming, and Y. Danya, "Pattern discovering of regional traffic status with self-organizing maps," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC '06)*, pp. 647–652, September 2006.
- [18] M. Treiber, A. Kesting, and R. E. Wilson, "Reconstructing the traffic state by fusion of heterogeneous data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 26, no. 6, pp. 408–419, 2011.

## Research Article

# Dynamic Resource Allocation in Hybrid Access Femtocell Network

**Afaz Uddin Ahmed,<sup>1</sup> Mohammad Tariqul Islam,<sup>1</sup>  
Mahamod Ismail,<sup>1</sup> and Mohammad Ghanbarisabagh<sup>2</sup>**

<sup>1</sup> Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

<sup>2</sup> Department of Electrical Engineering, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Correspondence should be addressed to Afaz Uddin Ahmed; afazbd@gmail.com

Received 18 January 2014; Accepted 18 February 2014; Published 20 March 2014

Academic Editors: Y. Mao and Z. Zhou

Copyright © 2014 Afaz Uddin Ahmed et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Intercell interference is one of the most challenging issues in femtocell deployment under the coverage of existing macrocell. Allocation of resources between femtocell and macrocell is essential to counter the effects of interference in dense femtocell networks. Advances in resource management strategies have improved the control mechanism for interference reduction at lower node density, but most of them are ineffective at higher node density. In this paper, a dynamic resource allocation management algorithm (DRAMA) for spectrum shared hybrid access OFDMA femtocell network is proposed. To reduce the macro-femto-tier interference and to improve the quality of service, the proposed algorithm features a dynamic resource allocation scheme by controlling them both centrally and locally. The proposed scheme focuses on Femtocell Access Point (FAP) owners' satisfaction and allows maximum utilization of available resources based on congestion in the network. A simulation environment is developed to study the quantitative performance of DRAMA in hybrid access-control femtocell network and compare it to closed and open access mechanisms. The performance analysis shows that higher number of random users gets connected to the FAP without compromising FAP owners' satisfaction allowing the macrocell to offload a large number of users in a dense heterogeneous network.

## 1. Introduction

Cellular operators these days aim to provide higher number of multimedia contents to attract the new generation customers to increase their revenue. Voice services are mostly characterized by a subscriber's number, while the data service is characterized by the use of a vast number of applications and protocols. As the data traffic is increasing exponentially, a need for proper resource management has risen for better quality of service (QoS). Deployment of FAP is primarily supported by the argument of improved indoor coverage for consumers and substantial cost savings for operators due to capacity offload. It is an effective alternative to divert and carry out a big portion of the traffic from the macro-cell/macrobase station (MBS). The key issue that restricts the vast implementation of FAP is the lack of effective schemes to mitigate interferences. FAP causes potential interferences with colocated MBS and neighboring FAPs operating in the

same frequency. Selection of access-control mechanisms in femtocell has a deteriorating effect on network performance. The operation of access-controlled femtocell greatly depends on the mechanism whose sole purpose is to decide whether the user can connect to the cell [1, 2].

Open access, close access, and hybrid access are the three existing access-control methods that decide users' connectivity to the FAP. In open access, whenever the users are within the range of a FAP, they get connected to the FAP easily. This includes a new set of signalling congestion in the network, as the number of handover attempts gets higher compromising the level of sharing and security concerns for the regular user. In the case of closed access, only particular users get access to the FAP, thus avoiding unwanted traffic congestion and possible interferences. In this case, the QoS is guaranteed at the expense of decreasing spectral efficiency. Hybrid access transacts with both challenges by tuning the resource ratio according to the number of femtocell owners

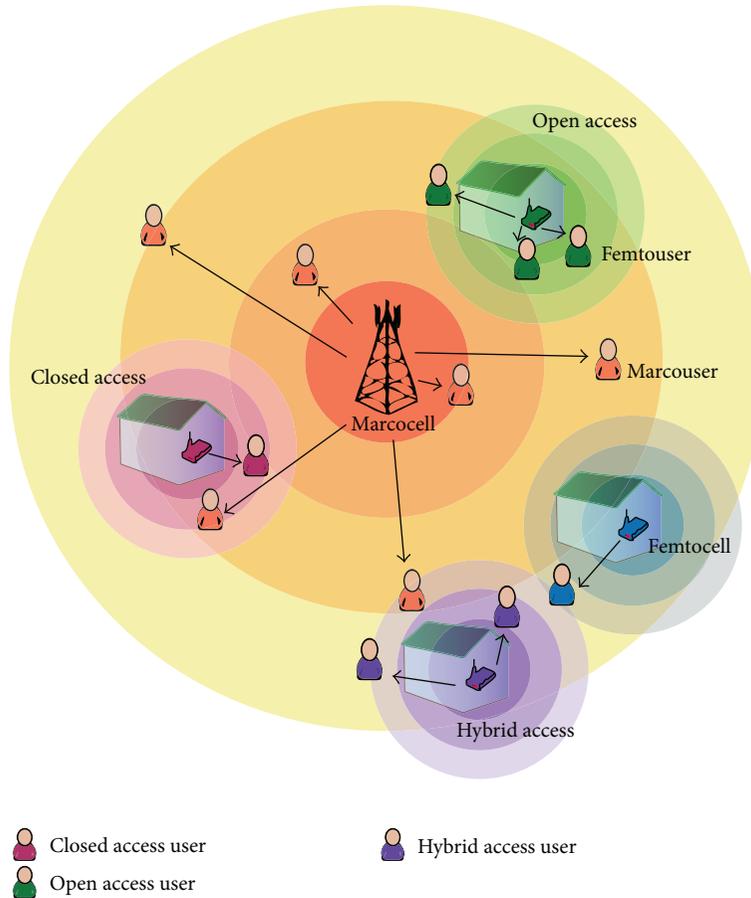


FIGURE 1: Different access control in mechanism femtocell network.

and subscribers. A limited amount of resources is available to the users who are within the coverage range and a “closed subscriber group” possesses the privilege to use the maximum service. Hybrid access reaches a compromise between the impact on the performance of subscribers and level of access granted to nonsubscribers, allowing them to possess a limited amount of features [3, 4]. Figure 1 shows different access control mechanism in a femtocell network.

From the operators’ point of view, open access is suitable for interferences reduction and cost optimization. However, serving to an unknown number of guests may degrade the QoS of the FAP owners. On the contrary, FAPs are especially designed for indoor coverage to ensure the maximum service for their owners. Recent studies have focused on the performance of closed access, open access, and resource-allocation management in heterogeneous network [5–18]. In [19] a resource-allocation technique in OFDMA femtocell network was considered with instantaneous FAPs power control mechanism for a target SINR. Joint resource-allocation and admission control design for dynamic frequency sharing was discussed in [20, 21]. A “3-ON-3 Femtocell Clustering Architecture” concept was developed in [22], where 3 FAPs form a cluster unit that transmits 3 power levels to serve 3 different types of user based on their priority level. In [23] both the centralized and decentralized approaches had

been considered, but the cochannel deployment is avoided which leads to less spectrum efficiency from the frequency reuse prospective, thus placing this approach under dispute. Finally, authors in [24] considered a joined centralized-distributed approach for resource allocation in an open access cochannel femtocell network where the level of users was not distinguished.

Cochannel deployment is preferred by the operators as it contains low cost and better spectral efficiency. Although in shared-bandwidth approaches, majority of which are currently being developed and deployed, the effect of interference is a big concern. Most of the techniques that deal with resource allocation of the frequency and energy at the same time are limited to single-tier networks. For multitier networks, the resource allocation should be more developed to handle the for multitier networks, the resource allocation should be more developed to handle the associate challenges. Resource allocation that is based centrally incurs excessive data transmission. Moreover, the ad hoc nature of femtocell makes the centralized control quite infeasible. However, the local approach provides anatomy in the system that speeds up the performance but suffers from quality degradation due to lack of essential data regarding interferences and resources allocated in the neighboring cells. In this paper, a joined centralized-distributed technique is recommended to

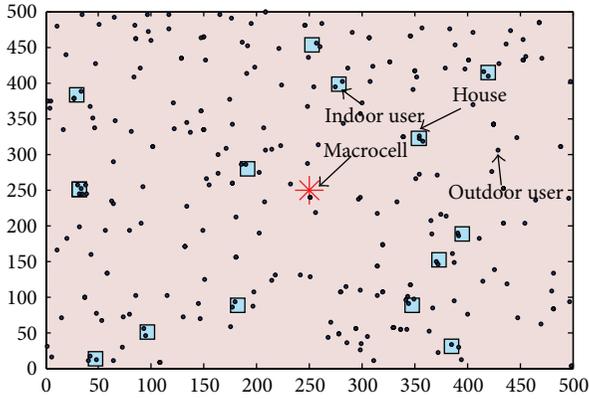


FIGURE 2: Sample layout of the simulation.

design a dynamic resource-allocation scheme for cochannel hybrid access-control femtocell network [25]. The system switches intelligently between the two modes based on the FAP owners' required throughput and users' density in a particular area of the network. It is a combined technique to utilize the unoccupied resources to maximize the QoS for the floating users and offloading MBS during overload condition. The rest of the paper is organized as follows: system analysis and modeling in Section 2, DRAMA-based hybrid access in Section 3, simulation results in Section 4, and conclusion in Section 5.

## 2. System Analysis and Modeling

A spectrum sharing two-tier heterogeneous network is considered with OFDMA in the downlink. An area of  $500\text{ m} \times 500\text{ m}$  is studied that contains one MBS in the middle and variable number of FAPs distributed within the area using homogeneous Poisson Point Distribution (PPP). All the FAPs are in the middle of a rectangular house. The FAPs do not overlap with each other or with the MBS. During the selection, all the users are 1 m apart from the MBS and FAPs. For outdoor users, 3GPP-LTE standard HATA small/large city model is used as it is more realistic for dense urban environment. WINNER II channel models are used for indoor path-loss scenarios for both line-of-sight (LOS) and non-line-of-sight (NLOS) users [26]. Wall penetration losses are also considered for the users who are in NLOS. To simulate the performance of the all the access-control mechanisms, a simulation environment is developed in MATLAB. A sample of the simulation layout is given in Figure 2.

OFDMA in LTE is robust against multipath interference and frequency selectivity [27]. In OFDMA network, FAPs have the advantage of allowing the allocation of orthogonal frequency/time resources to users. In the simulation, the whole bandwidth is divided into certain resource blocks (RBs) and certain numbers of subcarriers are allotted to each of the RB. All randomly deployed FAPs operate in the same bandwidth. The RB assumption is adopted from the 3GPP-LTE standard concept defined in [28]. A narrow channel appears relatively flat in the frequency domain and

the process of equalization is thus simplified. Therefore, channel capacity is relatively higher with narrow bandwidths and it does not change by dividing the frequency band into narrower frequency bins.

For  $x$  number of subcarriers, the signal-to-interference-plus-noise-ratio (SINR) expression for MBS user is

$$\text{SINR}_{m,x} = \frac{P_{M,x} G_{M,m,x} D_M^{-\alpha_m}}{N_0 \Delta f + \sum_{i \in F_i} \phi P_{i,x} G_{i,m,x} D_i^{-\alpha_{mi}}}, \quad (1)$$

where  $N_0$ ,  $\Delta f$ ,  $P_{M,x}$ ,  $G_{M,m,x}$ ,  $D_M$ ,  $P_{i,x}$ ,  $G_{i,m,x}$ , and  $D_i$  are the white noise power spectrum density, subcarrier spacing, transmission power from MBS, random channel gain from MBS to MBS user, transmitting power of interfering FAPs, random channel gain from FAP to MBS user, and distance from FAP to MBS user, respectively.  $\alpha_m$  and  $\alpha_{mi}$  are the path-loss exponents of the link from MBS and from FAP to the user, respectively.  $F_i$  is the set of interfering FAP for that particular MBS user.  $\phi$  is the penetration loss for a single wall.

The SINR expression for the FAP user is

$$\begin{aligned} \text{SINR}_{f,k} &= \frac{P_{F,k} G_{F,f,k} D_F^{-\alpha_f}}{N_0 \Delta f + \phi P_{M,k} G_{M,f,k} D_M^{-\alpha_{mf}} + \sum_{i \in F_i} \phi^2 P_{i,k} G_{i,f,k} D_i^{-\alpha_{fi}}}, \end{aligned} \quad (2)$$

where  $P_{F,k}$ ,  $G_{F,f,k}$ ,  $D_F$ ,  $G_{M,f,k}$ , and  $G_{i,f,k}$  are the transmission power from FAP to FAP user, random channel gain for FAP user, distance from FAP to user, random channel gain from MBS to FAP user, and random channel gain from interfering FAP to targeted FAP user, respectively.  $\alpha_f$ ,  $\alpha_{mf}$ , and  $\alpha_{fi}$  are the path-loss exponent of the link from FAP, MBS, and interfering FAP to the targeted user, respectively.

## 3. Dynamic Resource-Allocation Management Algorithm- (DRAMA-) Based Hybrid Access

FAP includes not just the base station itself but also the controller that enables local radio resource control. This connects back to the mobile operator core at a higher point for central authentication and management, which addresses the scalability concerns above, as the resource is located locally. Within the coverage range of FAP, users give the highest priority to the FAP for better signal quality. MBS wants to transfer "excess load" towards FAP to reduce user congestion and signaling overhead. When a FAP does not allow access; the user tries to stay connected to the MBS to avoid call drop. In hybrid access, FAP allows random users to get a better QoS. Based on the operator's network planning, the service for random users varies. In most of the cases, FAP assigns a ratio of the total resources to the guest users. This degrades owners' satisfactory level of service. In this scheme, the FAP owners are allowed to select the minimum level of uninterrupted service they want to enjoy. Initially, FAP ensures uninterrupted service to the owners. Subsequently it utilizes the rest to serve a particular number of random users. Depending on the user congestion in the network,

radio network controller (RNC) selects a minimum level of service for each FAP to assign to the random users. The trick is that whenever the network has less congestion of users, the FAP serves a fewer number of users, allocating more resources to each of them. However, for higher user congestion, FAP serves more users with fewer resources. The resource allocation must be high enough to cross the required threshold level for handover. The network balances the total traffic ensuring the maximum level of resource efficiency. Bonus usage or reward tariff should be awarded to the owners for offloading a certain percentage of the total load to encourage the sharing of idle resources. Also depending on the pricing policy of the network operators, special tariffs at home can be applied for calls placed under femtocell coverage. As the proposed scheme consists of both centralized and distributed resource management approaches, the radio network controller (RNC) controls the network centrally and FAP controls the rest locally. FAPs connect to the RNC through the backhaul connection. The central controller determines the relative position of the FAPs and MBS. An X2 connection measures and conveys the downlink interference coordination between MBS and FAPs. The DL-HII (Downlink-High Interference Indicator) generates the necessary signal and share through the wired backbone [29]. RNC defines a functioning area for each MBS. Number of active FAPs and users for a constant period of time determines the execution of the scheme for the functioning area.

The network divides all the users in three categories, FAP owners, random users who are under FAP service, and MBS users. Customarily the FAP owners are always within the coverage of their own FAP; otherwise, they are treated as the floating users. Corresponding FAPs allocate adequate resources to the owner(s) so that, overcoming all the losses, FAPs ensure the minimum throughput asked by the owners' (the minimum throughput is shown in Table 2). The rest of the resources are distributed to the random users under the FAPs coverage. The selection of random users is based on the level of service which is subjected to cross-tier interference due to poor MBS coverage and strong FAP interferences. Network allocates resources based on total population of the functioning area. The allocated resources are high enough to insist the MBS users to get handed-over to the nearest FAP. After a predefined interval, the network updates the status of the active FAPs and users and reassigns the resources according to the scheme. The RNC executes certain tasks and the rest will be carried out by the FAPs, avoiding unnecessary delay in assessment. RNC ensures the mobility management undertaking the owners and random user access in the FAPs besides doing its regular function, like, link management, call processing, handling FAPs and MBS, assigning spectrum as RBs, and controlling handover mechanisms. FAPs assign spectrum locally, monitor interference level, get feedback from users in the uplink, and synchronize the strongest signal with the desired macro signal. In addition, hierarchical cell structure (HCS) can also be introduced to distinguish between MBS and FAP and to execute rules for users of different priority level in each layer [30].

The number of active MBS users in the functioning area and the number of owners in each FAP are  $N_m$  and  $N_f$ , respectively.  $N_r$  is the number of MBS users who get access to a FAP service. For users under a FAP's coverage in the network,

$$C_{ff} \propto C_{fr}, \quad \{N_r \in \mathfrak{R}\}, \quad (3)$$

where  $C_{ff}$  and  $C_{fr}$  are the FAP owners throughput and random users throughput who got access to FAP service, respectively.

Assume a hybrid resource distribution constant  $K_r$  as follows:

$$K_r = \frac{C_r}{C_f}, \quad \{x \in K_r : 0 < x < 1\}. \quad (4)$$

If the value of  $K_r = 1$ , it will act as an open access FAP and if  $K_r = 0$ , it will work as a closed access FAP. The cells will assign spectrum as RBs and for a given time interval the transmission power to all the RBs is equal for both MBS and FAPs. Now the throughput of a particular MBS user which gets handed over to FAP can be expressed as

$$C_{fr} = \frac{\delta_{FAP}}{N_f + N_r},$$

$$\delta_{FAP} = N_{RB} C_{RB} \Delta f \log_2 (1 + \alpha \text{SINR}_{f,r}), \quad (5)$$

$$C_{mr} = \frac{\delta_{MBS}}{N_m - \sum_{i=1}^{F-1} N_r},$$

$$\delta_{MBS} = N_{RB} C_{RB} \Delta f \log_2 (1 + \alpha \text{SINR}_{m,r}),$$

where  $C_{mr}$ ,  $C_{fr}$ ,  $N_{RB}$ ,  $C_{RB}$ ,  $\Delta f$ , and  $\alpha$  are the throughput of random users under MBS service, throughput of random users under FAP service, number of resource blocks, sub-carrier per resource block, subcarrier spacing, and  $\alpha$  is the constant for target bit error rate (BER), respectively.

For maximum random users' throughput and maximum number of MBS offloading, the optimization problem can be stated as

$$\sum_{\max} k = \frac{\delta_{FAP}}{C_{ff} (N_f + \sum_{\max} N_r)} \quad (6)$$

$$\text{Subjected to } C_{fr} \geq C_{mr}, \quad N_r < N_m - \sum_{i=1}^{F-1} N_r,$$

where  $F$  is the number of active FAPs in the functioning area under the MBS coverage.

If the users provide continuous feedback in the uplink about the SINR for the assigned RB to its FAP or MBS and if the imperfection of proper feedback and channel estimation is not considered, the value of the hybrid constant and number of random users for a FAP are as follows:

$$\sum_{\max} k = \frac{\delta_{FAP} C_{mr}}{C_{mr} (N_f + N_m) - \delta_{MBS} C_{ff}}, \quad (7)$$

$$\sum_{\max} N_r = \frac{C_{mr} (N_f - N_m - C_{ff} N_f) - \delta_{MBS} C_{ff}}{C_{mr} C_{ff}}.$$

TABLE 1: System parameters.

System parameters	Value/range
Number of MBS	1
Number of FAP	1-25
Number of active user	400
Number of active owners in FAP	3-5
Range of MBS	500 m
Range of FAP	20 m
MBS antenna height	30 m
FAP antenna height	1 m
User equipment height	1 m
Frequency	2 GHz
Bandwidth	10 MHz
Subcarrier spacing	15 KHz
MBS transmission power	46 dBm
Macroantenna gain	13 dBm
FAP transmission power	20 dBm
Distribution time interval	500
FAP arrival intensity	1
Random active user arrival intensity	1.5
Shadow fading std.	6 dB
White noise power density	-174 dBm/Hz
Modulation scheme	64-QAM
Number of resource blocks	50
Subcarrier per resource block	12
Resource block size	180 KHz
BER	$10^{-6}$

TABLE 2: Owners' minimum throughput.

Set	Owners' minimum throughput
Set 1	5 Mbps
Set 2	3.5 Mbps
Set 3	2 Mbps

The FAP will allow  $N_r$  number of random users under its coverage to ensure the satisfactory level of service for both fixed and random users. If the assigned bandwidth for the random users is not high enough, the user might experience a lower service even if they get better coverage. So, any random user who gets better service from MBS, will not change serving cell.

### 4. Simulation Result

The simulations are event-based and developed according to 3GPP standards. The plotted values are an average of 1000 independent simulations. The assumed system parameter for the simulation is given in Table 1.

The standard length of the cyclic prefix in LTE is  $4.69 \mu s$ . This enables the system to tolerate path variations of up to 1.4 km with the symbol length set to  $66.7 \mu s$ . Each subcarrier can carry maximum data rate of 15 Ksps (kilo-symbols per second). Modulation 64-QAM represents 6 bits per symbol.

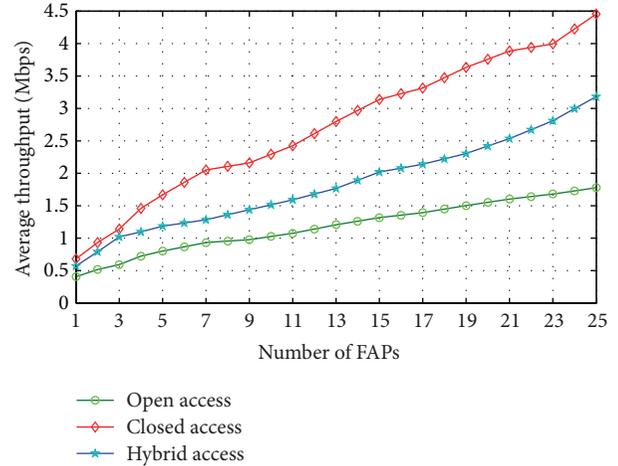


FIGURE 3: Average throughput of total users for variable numbers of FAPs.

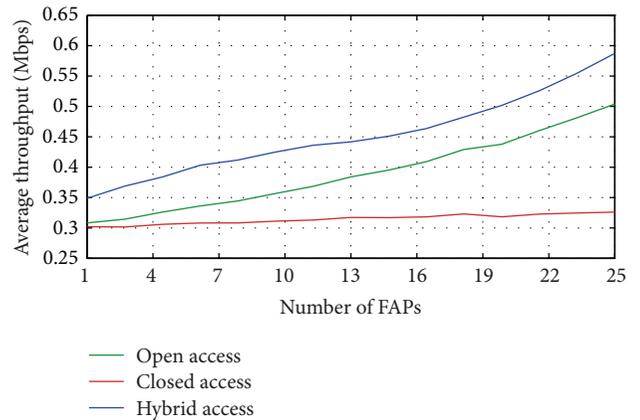


FIGURE 4: Average throughput of random users for variable number of FAPs.

Therefore, 10 MHz can provide a raw symbol rate of 9 Msps or 54 Mbps. This enables the system to compartmentalize the data across standard numbers of subcarriers [31].

The performance of a resource-allocation scheme is evaluated based on different deployment densities of FAPs within the functioning area. The general performance of open access, closed access, and hybrid access is shown in Figure 3. In case of closed access, the throughput of FAP owners is very high compared to the MBS users that boost up the average throughput of the total users of the network. Hybrid access on the other hand using this scheme shows a better performance than the open access. The average throughput of every access mechanism increases along with the FAPs deployment density.

Figure 4 illustrates the average throughput against variable numbers of deployed FAPs. The performance of hybrid access users is better than that of the other two access mechanisms. Cell edge users who are mostly subjected to the interference get under the FAP's service and FAP is only open for a particular number of random users, as it has to

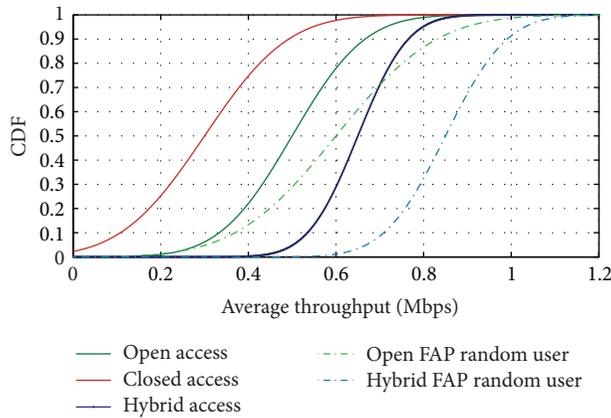


FIGURE 5: Average throughput of random users for different access mechanisms.

ensure a better service to the existing connected users. In contrast, the open access allows any user who grasps a better signal level under its coverage area. It does not have any priority level settings for the owners and for the random users. Random users always look for a better signal level from the neighboring cells which causes increasing numbers of handover attempts. As a result, the average throughput of the open access is lower than the hybrid access.

A resulting cumulative function is given in Figure 5. The deviation of average throughput for random users in hybrid access is much less than both open and closed accesses. In the case of open access and hybrid access, users who get access to the FAP service possess comparatively a higher average throughput. The “-” line shows the average throughput of the random FAP users. FAP only allows a distinct number of users in hybrid access that confirms a higher throughput than the open access. Closed access does not have any mechanism to allow access to the random users. In dense femtocell network, spectrum sharing mandated by the means of co-channel femtocell deployment, has an adverse effect on the system’s throughput and the quality of service.

To see the performance of the hybrid access from the users divesting prospective, three sets of owner throughput are considered as the accessibility of the FAPs’ unoccupied resources mainly depends on the FAP owners’ demand.

Figure 6 shows the number of users occupied by the MBS and FAPs. The owner’s minimum throughput levels are set consequently in the system. For set 3, which is the lowest, the numbers of FAP occupied users are the highest. The number of owners per FAP is fixed and an increasing number of random users get access to the FAP with an increasing number of deployment densities. While for set 2 and set 1, the performance of user offloading decreased correspondingly.

Practically in any dense network, the total number of floating users is always higher than the owners. Thus, the escalation of the overall QoS of the operators depends upon random users’ consumption. Considering the proposed scheme from the perspective of network performance and subscribers’ satisfaction, hybrid access mechanism shows a

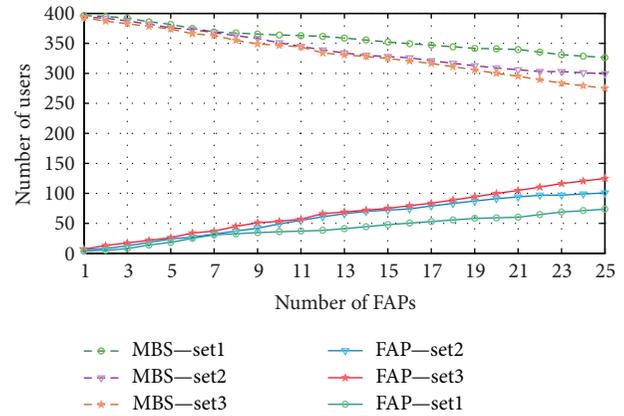


FIGURE 6: Users served by FAPs and MBS.

better performance. The throughput of the FAP owners is above the minimum satisfactory level, while for the random users, it is the maximum. Compared with the open access, the hybrid access confirms 10%–35% better performance for the random users, indicating that the FAP-to-FAP interference is reduced significantly. MBS gets a chance to handover users who are more subjected to cross-tier interference. Because of the reduction of MBS-user congestion, it lessens the chances of cross-tier interference near FAPs. The number of users served by the FAPs is higher, when the owners’ demand is lower. If the number of total users increases, RNC computes a new threshold throughput limit for the floating users in the network.

## 5. Conclusion

The placement of a femtocell has a critical effect on the performance of the wider network. This is one of the key issues to be addressed for successful deployment. The dynamic level resource-allocation scheme presented in this paper changes the environment of the network operation rapidly based of the users’ congestion. Considering the minimum resources required for the FAP owners, the system ensures better utilization of available spectrums. The network can offload a certain amount of excess MBS load by diverting it to FAPs. It utilizes the unused resources and ensures a superior level of service for the roaming users. Based on the capacity optimization, it assists MBS to offload excess traffic that reduces macro-femto-interferences and escalates the network performance which is suitable for higher deployment densities of femtocell. Future research will focus on the spectrum leasing features of hybrid access femtocell network.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## References

- [1] G. de la Roche, A. Valcarce, D. López-Pérez, and J. Zhang, "Access control mechanisms for femtocells," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 33–39, 2010.
- [2] A. U. Ahmed, M. Islam, R. Azim, M. Ismail, and M. F. Mansor, "Microstrip antenna design for femtocell coverage optimization," *International Journal of Antennas and Propagation*. In press.
- [3] P. Xia, V. Chandrasekhar, and J. G. Andrews, "Open versus closed access femtocells in the uplink," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3798–3809, 2010.
- [4] M. Zuair, "Development of an access mechanism for femtocell networks," *Journal of Theoretical and Applied Information Technology*, vol. 51, no. 3, pp. 434–441, 2013.
- [5] H.-S. Jo, P. Xia, and J. G. Andrews, "Open, closed, and shared access femtocells in the downlink," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, article 363, 2012.
- [6] R. M. Radaydeh and M.-S. Alouini, "Switched-Based Interference Reduction Scheme for Open-Access Overlaid Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2160–2172, 2012.
- [7] S. Wang, F. Huang, M. Yuan, and S. Du, "Resource allocation for multiuser cognitive OFDM networks with proportional rate constraints," *International Journal of Communication Systems*, vol. 25, no. 2, pp. 254–269, 2012.
- [8] J. Zhang, Z. Zhang, H. Luo, W. Wang, and G. Yu, "Initial spectrum access control with QoS protection for active users in cognitive wireless networks," *International Journal of Communication Systems*, vol. 25, no. 5, pp. 636–651, 2012.
- [9] K. D. Nguyen, H. N. Nguyen, and H. Morino, "Performance study of channel allocation schemes for beyond 4G cognitive femtocell-cellular mobile networks," in *IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS '13)*, pp. 1–6, Mexico City, Mexico, March 2013.
- [10] W. Zheng, H. Zhang, X. Chu, and X. Wen, "Mobility robustness optimization in self-organizing LTE femtocell networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, article 27, 2013.
- [11] C.-Y. Oh, M. Y. Chung, H. Choo, and T.-J. Lee, "Resource allocation with partitioning criterion for macro-femto overlay cellular networks with fractional frequency reuse," *Wireless Personal Communications*, vol. 68, no. 2, pp. 417–432, 2013.
- [12] S. Wang, J. Wang, J. Xu, Y. Teng, and K. Horneman, "Fairness guaranteed cooperative resource allocation in femtocell networks," *Wireless Personal Communications*, vol. 72, no. 2, pp. 957–973, 2013.
- [13] E. Oh and C. Woo, "Performance analysis of dynamic channel allocation based on the greedy approach for orthogonal frequency-division multiple access downlink systems," *International Journal of Communication Systems*, vol. 25, no. 7, pp. 953–961, 2012.
- [14] C. Olariu, J. Fitzpatrick, P. Perry, and L. Murphy, "A QoS based call admission control and resource allocation mechanism for LTE femtocell deployment," in *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC '12)*, pp. 884–888, Las Vegas, Nev, USA, January 2012.
- [15] J. L. Chen, S. W. Liu, S. L. Wu, and M. C. Chen, "Cross-layer and cognitive QoS management system for next-generation networking," *International Journal of Communication Systems*, vol. 24, no. 9, pp. 1150–1162, 2011.
- [16] Y.-S. Liang, W.-H. Chung, G.-K. Ni, Y. Chen, H. Zhang, and S.-Y. Kuo, "Resource allocation with interference avoidance in OFDMA femtocell networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 5, pp. 2243–2255, 2012.
- [17] W.-S. Lai, M.-E. Chiang, S.-C. Lee, and T.-S. Lee, "Game theoretic distributed dynamic resource allocation with interference avoidance in cognitive femtocell networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '13)*, pp. 3364–3369, Shanghai, China, April 2013.
- [18] D. López-Pérez, X. Chu, A. Vasilakos, and H. Claussen, "Power minimization based resource allocation for interference mitigation in OFDMA femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 333–344, 2014.
- [19] J.-H. Yun and K. G. Shin, "Adaptive interference management of OFDMA femtocells for co-channel deployment," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1225–1241, 2011.
- [20] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 1, pp. 248–2257, 2013.
- [21] L. Le, D. Niyato, E. Hossain, D. Kim, and D. Hoang, "QoS-aware and energy-efficient resource management in OFDMA femtocells," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 180–194, 2013.
- [22] F. Mhiri and G. Pujolle, "Cognitive interference management for autonomic femtocell networks," *International Journal of Applied Information Systems*, vol. 2, no. 2, pp. 40–48, 2012.
- [23] F. Tariq, L. S. Dooley, A. S. Poulton, and Y. Ji, "Dynamic fractional frequency reuse based hybrid resource management for femtocell networks," in *Proceedings of the 7th International Wireless Communications and Mobile Computing Conference (IWCMC '11)*, pp. 272–277, Istanbul, Turkey, July 2011.
- [24] L. Li, C. Xu, and M. Tao, "Resource allocation in open access OFDMA femtocell networks," *IEEE Wireless Communications Letters*, vol. 1, no. 6, pp. 625–628, 2012.
- [25] H. Zhang, X. Chu, and X. Wen, *4G Femtocells: Resource Allocation and Interference Management*, Springer, New York, NY, USA, 2013.
- [26] J. Meinilä, P. Kyösti, T. Jämsä, and L. Hentilä, "WINNER II channel models," in *Radio Technologies and Concepts for IMT-Advanced*, pp. 39–92, 2009.
- [27] E. S. Hassan, X. Zhu, S. E. El-Khany, M. I. Dessouky, S. A. El-Dolil, and F. E. A. El-Samie, "Performance evaluation of OFDM and single-carrier systems using frequency domain equalization and phase modulation," *International Journal of Communication Systems*, vol. 24, no. 1, pp. 1–13, 2011.
- [28] 3GPP Release 8, <http://www.3gpp.org/Release-8>.
- [29] S. Chiochan and E. Hossain, "Adaptive radio resource allocation in OFDMA systems: a survey of the state-of-the-art approaches," *Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 513–527, 2009.
- [30] 3GPP Technical Specification Group, Radio Access Networks, 3G Home NodeB Study Item Technical Report (Release 8), 2008, <http://www.quintillion.co.jp/3GPP/Specs/25820-800.pdf>.
- [31] A. Ghosh, R. Ratasuk, W. Xiao et al., "Uplink control channel design for 3GPP LTE," in *Proceedings of the IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, Athens, Greece, September 2007.

## Research Article

# Analyses of Crime Patterns in NIBRS Data Based on a Novel Graph Theory Clustering Method: Virginia as a Case Study

Peixin Zhao,<sup>1</sup> Marjorie Darrah,<sup>2</sup> Jim Nolan,<sup>3</sup> and Cun-Quan Zhang<sup>2</sup>

<sup>1</sup> School of Management, Shandong University, Jinan, Shandong, China

<sup>2</sup> Department of Mathematics, West Virginia University, Morgantown, WV, USA

<sup>3</sup> Department of Sociology and Anthropology, West Virginia University, Morgantown, WV, USA

Correspondence should be addressed to Peixin Zhao; [pxzhao@126.com](mailto:pxzhao@126.com)

Received 28 January 2014; Accepted 25 February 2014; Published 20 March 2014

Academic Editors: Y. Mao, X. Meng, and J. Zhou

Copyright © 2014 Peixin Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper suggests a novel clustering method for analyzing the National Incident-Based Reporting System (NIBRS) data, which include the determination of correlation of different crime types, the development of a likelihood index for crimes to occur in a jurisdiction, and the clustering of jurisdictions based on crime type. The method was tested by using the 2005 assault data from 121 jurisdictions in Virginia as a test case. The analyses of these data show that some different crime types are correlated and some different crime parameters are correlated with different crime types. The analyses also show that certain jurisdictions within Virginia share certain crime patterns. This information assists with constructing a pattern for a specific crime type and can be used to determine whether a jurisdiction may be more likely to see this type of crime occur in their area.

## 1. Introduction

The National Incident-Based Reporting System (NIBRS) is a crime reporting program for local, state, and federal law enforcement agencies that provides a wealth of incident level data for use in analysis. It is part of the Uniform Crime Reporting (UCR) Program which is administered by the FBI. The UCR Program provides a nationwide view of crime based on data submitted through state programs or directly to the national UCR Program and has been operational for around 70 years. The NIBRS was implemented in the late 1970s to meet law enforcement need for the 21st century. This vast system houses information on offenses, victims, offenders, property, and persons arrested, as well as the incident itself. The data of NIBRS are well structured and readily available for researchers and law enforcement agencies to assist with understanding the intricate nature of crime.

Akiyama and Nolan [1] outlined the structure of the NIBRS data set and provided methods for understanding and analyzing the data. Dunn and Zelenock [2] also describe and test procedures to facilitate the use of this vast system. Building on these initial works, many authors continue

to investigate how this storehouse of data can be turned into useful information for researchers and law enforcement agencies. Much of the work employs descriptive statistics applied in various sophisticated ways to extract information from the files. For example, Thompson et al. [3] apply descriptive statistics to examine intimate partner violence and make connections between this crime and other crimes that occurred in the same incident. They were able to show a link between intimidation and more serious violent crimes and extract information about the relationships between the victim and the offender from the NIBRS data that helps to better understand this type of crime. Snyder [4] used logistic regression techniques to predict the arrest of juvenile robbery offenders. More recently, Addington and Rennison [5] used logistic regression models along with the NIBRS and the National Crime Victimization Survey (NCVS) data for predicting rape cooccurrence to provide a critical initial look at rapes that occur with other crimes.

For criminologists the NIBRS data holds the answers to many long-standing questions about crime, criminal offending, and crime victimization. However, gaining access to some of these answers has remained difficult because of

the size and complexity of the data. Effective techniques, such as data mining and clustering, for criminal justice data are of increasing importance to both the research and law enforcement communities [6]. In recent years, clustering categorical data has gained more importance because it is one of the fundamental methods in data mining [7]. Clustering crime data, as with other categorical data, is unsupervised learning that aims at partitioning a data set into groups of similar items. The goal is to create clusters of data objects where the within-cluster similarity is maximized and the between-cluster similarity is minimized. Over the years, many clustering algorithms have been developed and tested. Some clustering techniques have developed specifically for use with categorical data. Abdu [8] presented three new clustering algorithms that were applied to the clustering of the NIBRS data. Two of his approaches combine spectral analysis and clustering techniques that are scalable to large data sets such as the NIBRS.

Clustering categorical data poses a challenge not encountered in clustering numerical data because the attribute categories are not ordered and defining a metric with which to measure the distance between data objects in a data set becomes a challenge. Many of the algorithms that have emerged for clustering categorical data rely on the occurrence/cooccurrence frequencies of attribute values in the data set to determine clusters of similar data objects. The basic goal is to choose a set of attribute categories that provide a summary of the data objects in a cluster. There are a wide range of clustering algorithms for categorical data, including K-modes [9], STIRR [10], CACTUS [11], ROCK [12], COOLCAT [13], LIMBO [14], and CLICKS [15]. A good summary can be found in [8]. The clustering algorithm used in this research is a mathematically well-defined model implemented in a polynomial time algorithm that guarantees an optimal solution [16].

Due to the lack of well-defined mathematical models and optimization goals, most existing graph theory clustering approaches could not guarantee a proper clustering result in general cases. For example, agglomerative hierarchical clustering methods could not produce proper clusters with larger sizes, while divisive hierarchical methods could not produce clusters with smaller sizes, and clusters with large difference in their sizes and k-core method may produce clusters with small edge-cuts, and so forth. Many papers and articles have mentioned these problems and frustration among users (e.g., see [17–19]). Even the most popular commercial software, SAS, is unable to produce proper outputs for some simple data.

The purpose of this paper is to present a novel multidimensional clustering method for the NIBRS data. We firstly outlines a new measure, called the *likelihood index*, that helps examine quantitatively how likely a crime is to occur in a particular jurisdiction. This measure compares a vector that describes a jurisdiction with a vector that represents a crime type. Then according to the defined distance between these two vectors, we can determine how closely the jurisdiction aligns with that crime type. The data used in this study were obtained from the 2005 NIBRS which is stored at the National Archive for Criminal Justice Data at the University

TABLE 1: NIBRS indexes used.

Segment	Index	Number of subindexes
Victim indexes	Type of victim	3
	Victim age	3
	Victim sex	2
	Victim race	5
	Victim ethnicity	3
	Victim residence status	3
	Aggravated assault/homicide circumstances	10
Offender indexes	Type of injury	6
	Offender age	3
	Offender sex	3
Additional indexes	Offender race	5
	Injury	2
	Juvenile	1
	Violent crime	1
	Juvict	1
	Multiple victims	1
	Multiple offenders	1
	Multiple offenders and victims	1
	Multiple offenders and one victim	1
	One offender and multiple victims	1
One offender and one victims	1	

of Michigan. This work explores the following research questions. Do specific crimes exhibit certain quantifiable characteristics? Do different types of crimes share similar quantifiable characteristics? Do jurisdictions of a state cluster with respect to different crime types? What is the likelihood that if one type of crime is occurring in an area, other types of crime with similar quantifiable characteristic will also occur in that area?

The rest of the paper is organized as follows. In Section 2 we summarize the data unit of analysis and preparation. In Section 3 we introduce the methods to deal with the data matrix and take Virginia as a case study. Section 4 provides some additional results and Section 5 gives the conclusions of the research.

## 2. Data Unit of Analysis and Preparation

The data sets available in the NIBRS provide a wealth of incident level data about each reported crime. As for 2010, approximately 40 states contribute their data to the massive data set. The data and tools are made available by University of Michigan for use by law enforcement agencies and researchers. In order to devise a manageable set of data for preliminary testing of techniques and for preliminary data analysis, only the 2005 data on assaults were explored. From the 2005 assault data, 121 jurisdictions (counties or cities) in

Virginia were selected for examination. These represent all jurisdictions within Virginia with populations greater than 10,000. There were 10,183 incidents reported in these 121 chosen jurisdictions.

For this study, 21 indexes from the NIBRS were chosen from the 246 available indexes. These 21 indexes were deemed important to provide the relevant characteristics of the victim(s), offender(s), and the circumstances of each incident. The selected particular indexes were listed in Table 1.

In order to facilitate the selected analysis techniques, the data was expanded from one column, with many possible entries, to multi columns that contained zero or one. For example, the Offender Segment index contains the sex of the offender and has the possible entries of male, female, or unknown. This index column was split into three individual columns where an entry in the three columns of (1 0 0) means female, (0 1 0) means male, and (0 0 1) means unknown. This turns the column for sex of the offender to three columns. All created columns were binary (0/1) columns that were used to help classify the characteristics of the incident. From the expansion of the original 21 indexes, 57 binary columns were

created. This led to the creation of a  $121 \times 57$  Crime Data Matrix, where each row  $i$  represents the  $i$ th jurisdiction and each column  $j$  represents the  $j$ th parameter related to the incident (e.g., offender sex, victim resident status) or a crime type (e.g., hate crime, drug dealing). The Crime Data Matrix construction is pictured below:

$$\text{Crime Data Matrix} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,57} \\ \vdots & \ddots & \vdots \\ a_{121,1} & \cdots & a_{121,57} \end{bmatrix}, \quad (1)$$

where  $a_{i,j}$  is the entry of the  $j$ th crime index from the  $i$ th jurisdiction.

Normalization of the rows of the matrix was completed by dividing each row entry by the population of the jurisdiction. This gave a per person rate for each crime parameter or each crime type. Normalization of the columns was completed by averaging the columns and subtracting the average from each entry in the column. Then each entry in the column was divided by the vector length of the column. Equation (2) shows these normalization step-by-step operations.

$$\begin{aligned} a_{i,j} &\leftarrow \frac{\text{number of occurrences of crime parameter or crime type}}{\text{population of the jurisdiction}}, \\ a_{i,j} &\leftarrow a_{i,j} - g_j \quad \text{where } g_j \text{ is the average of the } j\text{th column,} \\ a_{i,j} &\leftarrow \frac{a_{i,j}}{n_j} \quad \text{where } n_j \text{ is the vector length of the } j\text{th column.} \end{aligned} \quad (2)$$

### 3. Finding Patterns in the Data

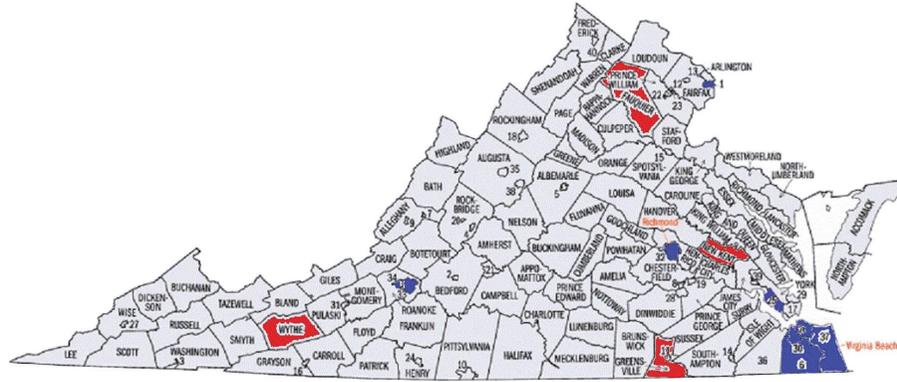
This section explains several different analyses that were performed on the data in the matrix described above in order to attempt to answer the research questions listed in Section 1. These analyses include comparing the columns of the matrix to determine the correlation of crime parameters to crime types and to determine the correlation of different crime types. Also there were two analyses performed comparing the row vectors to develop the likelihood index and to cluster the jurisdictions by crime types.

*3.1. Correlation of Crime Parameters and Crime Types (Comparing the Column Vectors).* The motivation for comparing the different crime parameters to crime types is to determine if there are some characteristics that can tell us about the likelihood of a crime type to occur in a certain jurisdiction. Each crime type may have factors that contribute to a specific crime appearing in a certain place. An overall increase in crime in an area may or may not correlate to an increase in any one particular type of crime, say hate crime, in that area. However, there may be individual parameters whose increase may indicate an increase in a particular type of crime. For example, if juvenile offenders are up in a certain area, this may indicate that hate crimes will also be up in that area. Also crime type vectors were compared against each other

in a similar way. For example, crimes like juvenile gang were compared against hate crime to see if these crime types also have a correlation.

In order to determine the relationship between the  $j$ th parameter ( $j$ th column of the Crime Data Matrix) and any other column (another parameter or another crime type), the correlation coefficient (cosine of the angle between the two column vectors over the norm of the vectors) was calculated. Each of the columns of the crime data matrix forms a vector in a 121-dimensional space and the vectors can be geometrically compared, with the correlation between two parameters represented by the cosine of the angle in this space. For example, the column for the juvenile offender could be compared against the column for hate crime or the column for hate crime can be compared to the column for gang-related crime. These comparisons are made by calculating the angle between these two columns to determine if there is any relationship and how strong that relationship may be. This method can assist in determining whether two columns vary directly, inversely, or separately.

For this comparison of two column vectors, the variation in two vectors must be transformed to eliminate the effects of mean differences. Once the mean deviation is determined, then the correlation can be determined by the cosine of the angle between the vectors. As an example, let  $j = (j_1, j_2, \dots, j_{121})$  be the column for juvenile offender and let



Independent cities

(1) Alexandria	(11) Emporia	(21) Lynchburg	(31) Radford
(2) Bedford	(12) Fairfax	(22) Manassas	(32) Richmond
(3) Bristol	(13) Falls Church	(23) Manassas Park	(33) Roanoke
(4) Buena Vista	(14) Franklin	(24) Martinsville	(34) Salem
(5) Charlottesville	(15) Fredericksburg	(25) Newport News	(35) Staunton
(6) Chesapeake	(16) Galax	(26) Norfolk	(36) Suffolk
(7) Clifton Forge	(17) Hampton	(27) Norton	(37) Virginia Beach
(8) Colonial Heights	(18) Harrisonburg	(28) Petersburg	(38) Waynesboro
(9) Covington	(19) Hopewell	(29) Poquoson	(39) Williamsburg
(10) Danville	(20) Lexington	(30) Portsmouth	(40) Winchester

FIGURE 1: The clustering results of 121 counties in Virginia according to hate crime.

$h = (h_1, h_2, \dots, h_{121})$  be the column for the hate crime indicator. To find the correlation coefficient of this two columns calculate the following:

$$\alpha_j = \cos \theta_j = \frac{\sum_i j_i h_i}{\|j\| \|h\|} \tag{3}$$

The sign of the answer is ignored, since either a strong positive relationship (close to 1 meaning that the two angles are in the same direction and close to one another) or a strong negative relationship (close to -1 meaning that the two angles are in opposite directions, but nearly opposite one another) indicates that the vectors appear to be related in some way.

In doing this comparison, with one vector fixed and comparing it to all other vectors and itself, we form a row vector of size 57. Consider the example that compares the hate crime vector with all other vectors. We construct another vector that we refer to as the Hate Crime Character Vector of size 57, which contains all the cosine  $\theta_j$  (the correlation coefficient of the  $j$ th parameter with respect to hate crime). A new vector is formed from all these correlation coefficients  $(\alpha_1, \alpha_2, \dots, \alpha_{57})$ , where each of the 57 parameters has a correlation coefficient vector associated with it. Now consider the corresponding components in the Hate Crime Character Vector. Table 2 tells us how hate crimes are related to aggravated assault crimes. Among those crimes, juvenile gang (0.4058) has higher correlation with hate crime, while the others, for example, argument (0.1325) and drug dealing (0.0265) have relatively lower correlation.

Table 3 also shows similar evidence that for victim's age, the age group less than 18 is also more correlated to hate crime

than the other two age groups. Many such comparisons can be observed using these correlation coefficient vectors.

3.2. Analyses of Jurisdictions (Comparing the Row Vectors).

By using the correlation coefficients, a 57-dimensional vector, two data analyses can be performed, each of which indicates the likelihood of a particular crime for each jurisdiction. The first one is a numerical index, called the *likelihood index* assigned to each of the 112 jurisdictions. The second one is a clustering analysis of all jurisdictions. Jurisdictions with similar recorded crime patterns (adjusted corresponding to correlation coefficients) form clusters.

Let  $\beta = \{\beta_1, \beta_2, \dots, \beta_{121}\}$  be the *likelihood index vector* where  $\beta_i$  is the likelihood index between jurisdiction  $i$  and a particular crime (e.g., hate crime) or crime parameter (e.g., juvenile offender). For this comparison of two row vectors, the variation in two vectors must be transformed to eliminate the effects of mean differences. Once the mean deviation is determined, then the correlation can be determined by the cosine of the angle between the vectors. As an example, let  $k = (k_1, k_2, \dots, k_{57})$  be the row for the Norfolk jurisdiction and let  $l = (l_1, l_2, \dots, l_{57})$  be the row for the hate crime indicator. To find the correlation coefficient of these two rows calculate the following:

$$\beta_i = \cos \phi_i = \frac{\sum_j k_j l_j}{\|k\| \|l\|} \tag{4}$$

Table 4 gives the top 30 jurisdictions of Virginia with respect to the hate crime likelihood index and also provides the data on the clustering of the jurisdictions discussed in the next section.

TABLE 2: Hate crime correlation coefficient vector values for other crimes.

Crime type	Hate crime correlation coefficient ( $\alpha_j$ )
Argument	0.1325
Assault on law enforcement officers	0.0371
Drug dealing	0.0265
Gangland (organized crime involvement)	0.1565
Juvenile gang	<b>0.4058</b>
Lovers' quarrel	-0.0281
Other felony involved	-0.0375
Other circumstances	0.0961
Unknown circumstance	0.1356

TABLE 3: Hate crime correlation coefficients for age of victim.

Age of victim	Hate crime correlation coefficient ( $\alpha_j$ )
Age < 18	<b>0.4944</b>
18 ≤ Age ≤ 60	0.0540
Age > 60	0.0474

3.3. *Using the Correlation Coefficients to Cluster the Jurisdictions.* To begin the clustering, a weighted complete graph with 121 vertices is formed. The weight on each edge is the correlation coefficient between the jurisdictions. The novel graph theory clustering method we proposed in [20] is used to find all the dense (highly weighted) subgraphs of the complete graph. A distinguished feature of this method, nonbinary hierarchical tree, clearly highlights meaningful clusters which significantly reduces further manual efforts for cluster selections. The results of clustering the jurisdictions with respect to hate crimes are also displayed in Figure 1 and Table 4.

It can be seen that of the top 30 jurisdictions with respect to hate crimes, 12 of the top 15 are in Cluster A and the others are in Cluster B. The remaining 91 jurisdictions also fall into Cluster B. The 12 higher hate crime rate counties (cities) are showed in Figure 1 (in red and blue).

#### 4. Additional Results

Similar analyses were performed with respect to other crime types: drug-dealing, juvenile gang, and gangland (organized crime involvement). The results are summarized in Tables 5–10.

Table 6 summarizes the drug dealing vector comparison with the offender parameter vectors.

Tables 5 and 6 summarize the comparison of the drug dealing vector with other crime parameter vectors. Each of the other crime parameters related to the victim and the offender was compared to the drug dealing vector. A high value for this correlation coefficient implies that there is a correlation between this crime parameter and drug dealing. Again, the motivation for comparing the different crime parameters to crime types is an attempt to determine if there

TABLE 4: Likelihood index and clustering for hate crime.

	Name	Cluster	Hate crime likelihood index
1	NEWPORT NEWS	<b>B</b>	0.867
2	NORFOLK	<b>B</b>	0.837
3	CHESAPEAKE	<b>B</b>	0.810
4	GREENSVILLE	<b>B</b>	0.762
5	PORTSMOUTH	<b>B</b>	0.759
6	RICHMOND	<b>B</b>	0.716
7	WYTHE	<b>B</b>	0.681
8	ALEXANDRIA	<b>B</b>	0.678
9	BRISTOL	<b>B</b>	0.667
10	NEW KENT	<b>B</b>	0.618
11	FAUQUIER	<b>B</b>	0.531
12	ROANOKE	A	0.439
13	RICHMOND	A	0.386
14	WILLIAMSBURG	A	0.377
15	VIRGINIA BEACH	<b>B</b>	0.318
16	CHARLOTTESVILLE	A	0.312
17	PETERSBURG	A	0.280
18	SPOTSYLVANIA	A	0.226
19	HOPEWELL	A	0.205
20	RUSSELL	A	0.121
21	CLARKE	A	0.121
22	WINCHESTER	A	0.118
23	SUFFOLK	A	0.111
24	MARTINSVILLE	A	0.075
25	STAUNTON	A	0.073
26	GALAX	A	0.070
27	SHENANDOAH	A	0.063
28	CAROLINE	A	0.044
29	SURRY	A	-0.020
30	DANVILLE	A	-0.094

are some characteristics that can tell us about the likelihood of a crime type to occur in a certain jurisdiction. For example, the correlation between the drug dealing vector and individual victim is 0.5789, which is much higher than the correlation between drug dealing and victim type business (0.1721) or victim type society/public (0.1729). This would seem to imply that for jurisdictions where the individual victim crimes are evident they may also have the likelihood for drug dealing related crimes. The tables show the correlation coefficient values: the correlation values in bold show a higher correlation with the drug dealing crime type than the other parameters in that category.

Table 7 summarizes the drug dealing vector comparison with the other crime types. The highest correlation is with argument.

Each of the other crime parameters related to the victim and the offender was compared to the gangland (organized

TABLE 5: Drug dealing correlation with other crime parameters related to victim.

	Drug dealing correlation coefficient
Type of victim	
individual	<b>0.5789</b>
Business	0.1721
Society/public	0.1729
Age of victim	
Age < 18	0.2933
Age ≥ 60	0.3872
18 ≤ age < 60	<b>0.5993</b>
Sex of victim	
Male	<b>0.5774</b>
Female	0.4989
Race of victim	
White	0.3906
Black	<b>0.5028</b>
Asia/Pacific Islander	0.0764
Unknown	0.0003
American Indian	0.0393
Ethnicity of victim	
Hispanic origin	0.0688
Not of Hispanic origin	<b>0.6009</b>
Unknown	-0.0616
Resident status of victim	
Nonresident	0.2631
Resident	<b>0.5740</b>
Unknown	0.1290

TABLE 6: Drug dealing correlation with other crime parameters related to offender.

Offender age	
Age < 18	0.3514
Age ≥ 60	0.3762
18 ≤ age < 60	<b>0.5678</b>
Offender sex	
Male	0.5331
Female	<b>0.5685</b>
Unknown	0.1972
Offender race	
White	0.3145
Black	<b>0.5133</b>
American Indian/Alaskan native	-0.0479
Unknown	0.1948
Asian/Pacific Islander	0.113

crime involvement) vector. There were no significant relationships to report from this comparison.

Table 8 gives the top 30 jurisdictions of Virginia with respect to the drug dealing likelihood index and also shows how the jurisdictions are clustered. Given the 121 Virginia jurisdictions being considered, of the top 30 with respect

TABLE 7: Drug dealing correlation with other crime types.

Other crime types	
Argument	<b>0.4854</b>
Assault on law enforcement officer(s)	0.3117
Gangland (organized crime involvement)	0.3456
Juvenile gang	0.0706
Lovers' quarrel	0.2967
Other felony involved	0.2955
Hate crime	0.0265

TABLE 8: Likelihood index and clustering for drug dealing.

	Name	Cluster	Drug dealing likelihood index
1	CHARLOTTESVILLE	<b>B</b>	0.9094
2	NEWPORT NEWS	<b>B</b>	0.9012
3	CHESAPEAKE	<b>B</b>	0.8851
4	PETERSBURG	<b>B</b>	0.8637
5	RICHMOND	<b>B</b>	0.8365
6	PORTSMOUTH	A	0.8281
7	HOPEWELL	A	0.7917
8	ROANOKE	<b>B</b>	0.791
9	SUFFOLK	<b>B</b>	0.7839
10	NORFOLK	<b>B</b>	0.7756
11	BRISTOL	<b>B</b>	0.7365
12	DANVILLE	A	0.707
13	GREENSVILLE	A	0.7032
14	GALAX	A	0.6881
15	CAROLINE	A	0.6852
16	SUSSEX	A	0.6396
17	WINCHESTER	A	0.6327
18	FREDERICKSBURG	<b>B</b>	0.6031
19	CLARKE	A	0.6021
20	FRANKLIN	A	0.5808
21	RICHMOND	A	0.5667
22	LYNCHBURG	A	0.5395
23	MECKLENBURG	A	0.5185
24	RADFORD	A	0.5133
25	GOOCHLAND	A	0.5114
26	MANASSAS	A	0.494
27	HENRY	<b>B</b>	0.4844
28	NORTON	A	0.4456
29	WISE	<b>B</b>	0.4381
30	WILLIAMSBURG	A	0.395

to the likelihood index, twelve are clustered together. Only three others jurisdictions from Cluster B appear outside the top 30; these jurisdictions are Smyth (35), Tazewell (37), and Pittsylvania (50). Each of the other crime parameters related to the victim and the offender was compared to the

TABLE 9: Likelihood index and clustering for gangland (organized crime involvement).

	Name	Cluster	Gangland likelihood index
1	NEWPORT NEWS	B	0.9217
2	ROANOKE	B	0.8786
3	CHESAPEAKE	B	0.8713
4	PETERSBURG	B	0.8421
5	NORFOLK	B	0.8129
6	RICHMOND	B	0.7953
7	LYNCHBURG	B	0.6965
8	FREDERICKSBURG	B	0.689
9	BRISTOL	B	0.6787
10	ALEXANDRIA	B	0.6569
11	NORTHAMPTON	B	0.632
12	LOUDOUN	B	0.4939
13	CHARLOTTESVILLE	A	0.4152
14	STAFFORD	B	0.3853
15	PORTSMOUTH	A	0.3075
16	HOPEWELL	A	0.2724
17	GALAX	A	0.2121
18	CAROLINE	A	0.1372
19	SUFFOLK	A	0.1077
20	GREENSVILLE	A	0.0817
21	CLARKE	A	0.0689
22	DANVILLE	A	0.0519
23	HENRY	A	0.0518
24	RICHMOND	A	0.0074
25	WINCHESTER	A	-0.0071
26	FRANKLIN	A	-0.0508
27	SUSSEX	A	-0.0763
28	MANASSAS	A	-0.0812
29	WILLIAMSBURG	A	-0.1001
30	GOOCHLAND	A	-0.1112

gangland (organized crime involvement) vector. There were no significant relationships to report from this comparison.

Table 9 gives the top 30 jurisdictions of Virginia with respect to the gangland (organized crime involvement) likelihood index and also shows how the jurisdictions are clustered. Given the 121 Virginia jurisdictions being considered, of the top 30 with respect to the likelihood index, thirteen are clustered together. Only one other jurisdiction from cluster B appears outside the top 30; this one jurisdiction is Fairfax County PD (31). Each of the other crime parameters related to the victim and the offender was compared to the juvenile gang vector. There were no significant relationships to report from this comparison.

Table 10 gives the top 30 jurisdictions of Virginia with respect to the juvenile gang likelihood index and also shows how the jurisdictions are clustered. Given the 121 Virginia jurisdictions being considered, of the top 30 with respect to

TABLE 10: Likelihood index and clustering for juvenile gang.

	Name	Cluster	Juvenile gang likelihood index
1	NORFOLK	B	0.8424
2	ROANOKE	B	0.8387
3	RICHMOND	B	0.7987
4	PORTSMOUTH	B	0.7977
5	GREENSVILLE	B	0.7575
6	CHESAPEAKE	B	0.7211
7	NEWPORT NEWS	B	0.7112
8	WILLIAMSBURG	B	0.6639
9	WYTHE	B	0.5719
10	ALEXANDRIA	B	0.5597
11	LYNCHBURG	B	0.5452
12	MARTINSVILLE	B	0.5355
13	PULASKI	B	0.4884
14	HAMPTON	B	0.4848
15	SPOTSYLVANIA	B	0.451
16	POWHATAN	B	0.442
17	PETERSBURG	A	0.4207
18	CHARLOTTESVILLE	A	0.4064
19	HOPEWELL	A	0.3519
20	RICHMOND	A	0.208
21	GALAX	A	0.2025
22	BRISTOL	A	0.2012
23	SUFFOLK	A	0.1925
24	CAROLINE	A	0.121
25	CLARKE	A	0.0939
26	WINCHESTER	A	0.0893
27	DANVILLE	A	0.0834
28	SUSSEX	A	0.051
29	CAMPBELL	A	0.0119
30	FRANKLIN	A	-0.0773

the likelihood index, sixteen are clustered together. No other jurisdiction from Cluster B appears outside the top 30.

### 5. Conclusion

The NIBRS provides a wealth of incident level data for use in analysis. The methods investigated in this research yielded promising preliminary results. The methods were applied only to the assault data from 2005 but can easily be extended to other crime types and to other years to validate these results and also provide longitudinal investigation.

The comparison between the crime type vector and the individual parameters vectors helped in two cases (hate crimes and drug dealing) to determine which factors was more related to those crimes. The different types of analyses that were conducted on the jurisdictions helped to validate one another. The likelihood index looked at whether a certain crime pattern existed in that jurisdiction, while the clustering

method sought to cluster all the jurisdictions based on the crime patterns of that jurisdiction. This information could be useful to assist law enforcement agencies or policy makers in determining which jurisdictions share common challenges that could possibly be addressed through cooperation and sharing resources between jurisdictions.

The next steps would be to utilize this same approach for data from other states or perhaps a larger region to examine if the same information is observed from the analyses. It will be interesting to see if Virginia data and other states have the same patterns or if different patterns emerge. Further research and refinement of these methods should yield tools that would provide researchers, law enforcement agencies, and government officials with a means to find patterns of different crime types and possibly identify jurisdictions that may be likely to experience that type of crime.

### Conflict of Interests

Peixin Zhao, Marjorie Darrah, Jim Nolan, and Cun-Quan Zhang certify that there is no actual or potential conflict of interests in relation to this paper.

### Acknowledgments

First author was partially supported by the China Postdoctoral Science Foundation Funded Project (2011M501149), the Humanity and Social Science Foundation of Ministry of Education of China (12YJCZH303), the Special Fund Project for Postdoctoral Innovation of Shandong Province (201103061), the Informationization Research Project of Shandong Province (2013EI153), the National Statistical Science Project (2013LZ38), and Independent Innovation Foundation of Shandong University (IIFSDU) (IFW12109).

### References

- [1] Y. Akiyama and J. Nolan, "Methods for understanding and analyzing NIBRS data," *Journal of Quantitative Criminology*, vol. 15, no. 2, pp. 225–238, 1999.
- [2] C. Dunn and T. Zelenock, "NIBRS data available for secondary analysis," *Journal of Quantitative Criminology*, vol. 15, no. 2, pp. 239–248, 1999.
- [3] M. P. Thompson, L. E. Saltzman, and D. Bibel, "Applying NIBRS data to the study of intimate partner violence: Massachusetts as a case study," *Journal of Quantitative Criminology*, vol. 15, no. 2, pp. 163–180, 1999.
- [4] H. N. Snyder, "The Overrepresentation of juvenile crime proportions in robbery clearance statistics," *Journal of Quantitative Criminology*, vol. 13, no. 2, pp. 151–161, 1999.
- [5] L. A. Addington and C. M. Rennison, "Rape co-occurrence: do additional crimes affect victim reporting and police clearance of rape?" *Journal of Quantitative Criminology*, vol. 24, no. 2, pp. 205–226, 2008.
- [6] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.
- [7] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," New York, NY, USA, Springer, National Incident-Based Reporting System, National Archive of Criminal Justice Data, 2005, <http://www.icpsr.umich.edu/NACJD/NIBRS/>.
- [8] E. Abdu, *Clustering categorical data using data summaries and spectral techniques [Ph.D. thesis]*, The City University of New York, 2009.
- [9] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.
- [10] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering categorical data: an approach based on dynamical systems," in *Proceedings of the 24th International Conference on Very Large Databases*, 1998.
- [11] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS clustering categorical data using summaries," in *proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and data mining*, pp. 73–83, 1999.
- [12] S. Guha, R. Rastogi, and K. Shim, "ROCK: a robust clustering algorithm for categorical attributes," in *Proceedings of the 15th International Conference on Data Engineering (ICDE '99)*, pp. 512–521, March 1999.
- [13] D. Barbará, J. Couto, and Y. Li, "COOLCAT: an entropy-based algorithm for categorical clustering," in *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM '02)*, pp. 582–589, November 2002.
- [14] P. Andritsos, *Scalable clustering of categorical data and applications [Ph.D. thesis]*, University of Toronto, 2004.
- [15] M. Zaki, M. Peters, I. Assent, and T. Seidl, "Clicks: an effective algorithm for mining subspace clusters in categorical datasets," *Data and Knowledge Engineering*, vol. 60, no. 1, pp. 51–70, 2007.
- [16] Y. Ou and C. Q. Zhang, "A new multimembership clustering method," *Journal of Industrial and Management Optimization*, vol. 3, no. 4, pp. 619–624, 2007.
- [17] R. N. Shepard and P. Arabie, "Additive clustering: representation of similarities as combinations of discrete overlapping properties," *Psychological Review*, vol. 86, no. 2, pp. 87–123, 1979.
- [18] A. V. Lukashin and R. Fuchs, "Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters," *Bioinformatics*, vol. 17, no. 5, pp. 405–414, 2001.
- [19] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics*, vol. 18, no. 4, pp. 536–545, 2002.
- [20] P. Zhao and C. Q. Zhang, "A new clustering method and its application in social networks," *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2109–2118, 2011.

## Research Article

# Channel Estimation in DCT-Based OFDM

**Yulin Wang, Gengxin Zhang, Zhidong Xie, and Jing Hu**

*Institute of Communication Engineering, PLA University of Science and Technology, Yudao Street 14, Nanjing 210007, China*

Correspondence should be addressed to Yulin Wang; [wang\\_yulinsci@126.com](mailto:wang_yulinsci@126.com)

Received 16 January 2014; Accepted 20 February 2014; Published 13 March 2014

Academic Editors: Y. Mao, X. Meng, and Z. Zhou

Copyright © 2014 Yulin Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper derives the channel estimation of a discrete cosine transform- (DCT-) based orthogonal frequency-division multiplexing (OFDM) system over a frequency-selective multipath fading channel. Channel estimation has been proved to improve system throughput and performance by allowing for coherent demodulation. Pilot-aided methods are traditionally used to learn the channel response. Least square (LS) and mean square error estimators (MMSE) are investigated. We also study a compressed sensing (CS) based channel estimation, which takes the sparse property of wireless channel into account. Simulation results have shown that the CS based channel estimation is expected to have better performance than LS. However MMSE can achieve optimal performance because of prior knowledge of the channel statistic.

## 1. Introduction

MULTICARRIER modulation (MCM) is a promising technique for high data rate transmission. It is used in many digital communications standards, such as local area networks (IEEE 802.11a/g/n), metropolitan area networks (IEEE802.16a), and digital audio and terrestrial video broadcast (DAB/DVB-T) in wireless digital communications systems and such as asymmetric digital subscriber loop (ADSL) in wireline digital communications system. All of these systems belong to the class of discrete Fourier transform- (DFT-) based orthogonal frequency-division multiplexing (OFDM) [1], referred to as DFT-OFDM in this paper.

DFT-OFDM system employs the complex exponential functions set as orthogonal basis. It realizes digital modulations and demodulations by the inverse DFT (IDFT) and DFT, respectively [1]. However, in the literature, [2–5] proposed using a single set of cosinusoidal functions as the orthogonal basis to construct baseband multicarrier signals. This MCM scheme can be synthesized using a discrete cosine transform (DCT). Here, we will denote the scheme as DCT-OFDM.

Reference [3] has shown that DCT-OFDM leads to complete elimination of interblock and intercarrier interference at the same guard sequence overhead, compared with DFT-OFDM. Results in [2] have proved that the bit-error

probability (BEP) performance of DCT-OFDM is superior to that of DFT-OFDM in the presence of carrier-frequency offset (CFO), due to the spectral compaction and energy concentration properties of the DCT. Additionally, DCT-OFDM is implemented with a better integrated system design and a reduced overall signal-processing complexity [3].

DCT-OFDM and DFT-OFDM systems are both robust to the multipath induced intersymbol interference (ISI) due to the orthogonal property. However, suffering from the frequency selective fading of the dispersive wireless channel, some subchannels may face deep fading and degrade the overall system performance. On the other hand, multipath can also bring multiplexing/diversity gains which improve the rate/reliability of communication system. As a result, one has to perform accurate channel estimation to obtain channel state information (CSI) before coherent demodulation.

The channel estimation for DFT-OFDM systems has been studied by many researchers [6–9]. However, there is still no investigation on channel estimation for DCT-OFDM system. On the other hand, the radio channel in a wireless communication system is often characterized by multipath propagation, typically with a few distinct paths, resulting in a sparse multipath channel model [10]. A recent advance in compressed sensing (CS) [11] provides a potential solution to reducing the required number of pilot symbols. Compressed channel sensing (CCS) was proposed in [12]

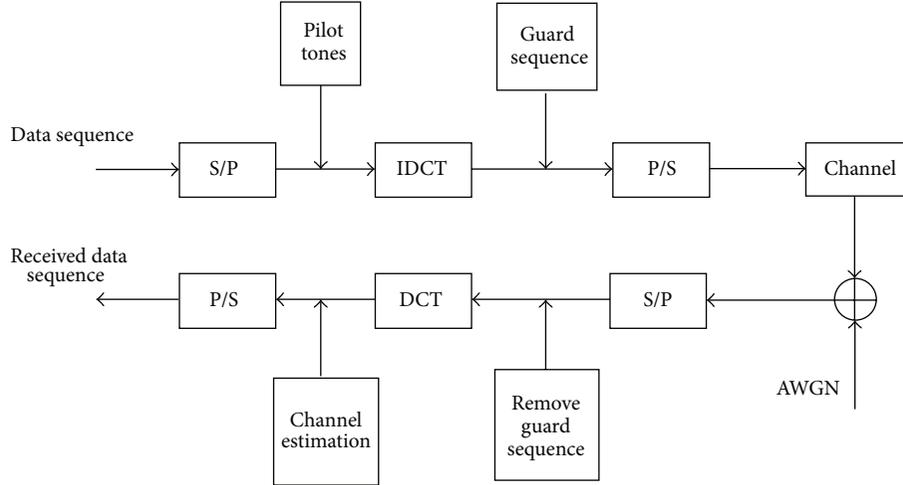


FIGURE 1: DCT-based OFDM systems.

for a frequency selective channel. Reference [13] applied distributed compressive sensing (DCS) theory to a channel estimation application. An efficient pilot design scheme was studied in [14] for seeking optimal pilot placement for sparse channel estimation. In [15], authors designed low-complexity sparse channel estimation and tracking method for time-varying DFT-OFDM channels.

In this paper, we design channel estimation methods for DCT-OFDM system. They are, respectively, least square (LS) estimator, mean square error estimator (MMSE), and compressed sensing estimator (CSE). The CSE directly uses DCT matrix as the orthogonal sparse dictionary. Compared with the two conventional estimators, CSE exploits the inherent sparse property of channels, saving pilots symbols for the same estimation performance. Thereby, CSE leads to economize the key resources, such as energy, latency, and bandwidth in DCT-OFDM system.

The remainder of this paper is organized as follows. In Section 2, we construct the DCT-OFDM system and give the system model. Then we introduce the LS and MMSE estimator for DCT-OFDM system in Section 3. Section 4 gives a brief review of CS theory and its application to sparse channel estimation in DCT-OFDM. In Section 5, we succinctly summarize the performance of the three estimators. We devote the conclusions in Section 6.

## 2. Signal Model

Figure 1 has shown a DCT-OFDM system, in which modulation/demodulation is done by IDCT/DCT. Unlike conventional DFT-OFDM, a single sinusoidal functions set  $\cos(2\pi nF_{\Delta}t)$ ,  $n = 0, 1, \dots, N-1$ , will be used as the orthogonal basis to implement MCM in DCT-OFDM system. The minimum  $F_{\Delta}$  required to satisfy

$$\int_0^T \sqrt{\frac{2}{T}} \cos(2\pi kF_{\Delta}t) \sqrt{\frac{2}{T}} \cos(2\pi mF_{\Delta}t) dt = \begin{cases} 1, & k = m \\ 0, & k \neq m \end{cases} \quad (1)$$

is  $1/2T$  Hz. The continuous-time representation of a baseband DCT-OFDM block  $x(t)$  is

$$x(t) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} d_n \beta_n \cos\left(\frac{n\pi t}{T}\right), \quad (2)$$

where  $d_0, d_1, \dots, d_{N-1}$  are  $N$  independent data symbols obtained from a modulation constellation, and

$$\beta_n = \begin{cases} \frac{1}{\sqrt{2}}, & n = 0 \\ 1, & n = 1, 2, \dots, N-1. \end{cases} \quad (3)$$

Sampling the continuous-time signal  $x(t)$  at time instants  $t_m = T(2m+1)/2N$  gives a discrete time sequence [16]:

$$x_m = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} d_n \beta_n \cos\left(\frac{\pi n(2m+1)}{2N}\right), \quad m = 0, 1, \dots, N-1 \quad (4)$$

which is the inverse DCT (IDCT). Thus, the continuous-time signal  $x(t)$  can be obtained by first performing an IDCT operation on data sequence:

$$\mathbf{d} = [d_0, d_1, \dots, d_{N-1}]^T, \quad (5)$$

where  $[\ ]^T$  represents the transpose operation and then feeds serially the resulting samples  $\mathbf{x} = [x_0, x_1, \dots, x_{N-1}]^T$  through a digital-to-analog (D/A) converter.

Signal is transmitted through a frequency-selective multipath fading channel. We assume the channel impulse response (CIR) is constant during one DCT-OFDM symbol. At the receiver sketched in Figure 1, after matched filtering, the signal is sampled at rate  $1/T_s$  and serial to parallel converted. We indicate with  $\mathbf{h} = [h_0, h_1, \dots, h_{L-1}]^T$  the  $T_s$ -spaced samples of the overall CIR:

$$\mathbf{y} = \mathbf{x} \otimes \mathbf{h} + \mathbf{w}, \quad (6)$$

where  $\otimes$  denotes cyclic convolution and  $\mathbf{w}$  is additive white Gaussian noise (AWGN) with zero mean and variance  $\sigma_w^2$ .

Denoting

$$H_n = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} h_k \cos\left(\frac{\pi k(2n+1)}{2N}\right), \quad n = 0, 1, \dots, L-1 \quad (7)$$

the DCT of  $\mathbf{h}$ , and  $\mathbf{H} = [H_1, H_2, \dots, H_{L-1}]$  is the channel response in DCT domain also called the channel frequency response in this paper. After removing the guard sequence, the received samples are passed to an  $N$ -point DCT unit. The output of the DCT unit is found to be

$$r_n = d_n H_n + \tilde{w}_n, \quad (8)$$

where  $r_n$  is the received signal and  $\tilde{w}_n$  is noise in DCT domain.

In this study, we assume that some known symbols (pilots) are multiplexed into the data stream, and channel estimation is performed at pilots locations. A total of  $N_p$  pilots  $\{c_n; 0 \leq n \leq N_p - 1\}$  are inserted in one DCT-OFDM block at known locations  $\{i_n; 0 \leq n \leq N_p - 1\}$ . The  $N_p$ -dimensional vector containing the DCT output at the pilot locations is denoted by  $\mathbf{r} = [r_0, r_1, \dots, r_{N_p-1}]^T$ ; from (7) and (8), we have

$$\mathbf{r} = \mathbf{D}\mathbf{F}\mathbf{h} + \tilde{\mathbf{w}} = \mathbf{D}\mathbf{H} + \tilde{\mathbf{w}}, \quad (9)$$

where  $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_{L-1}]$  is the AWGN noise vector in DCT domain and  $\mathbf{D}$  is a diagonal matrix containing pilot symbols

$$\mathbf{D} = \text{diag}\{c_0, c_1, \dots, c_{N_p-1}\} \quad (10)$$

and  $\mathbf{F}$  is an  $N_p \times L$  DCT matrix with entries:

$$F(k, n) = \sqrt{\frac{2}{N}} \beta_k \cos\left(\frac{\pi k(2n+1)}{2N}\right). \quad (11)$$

### 3. LS and MMSE Channel Estimation

A conventional approach to estimate the channel is the least square (LS) estimation of channel frequency response on the pilot subcarriers. Then, the frequency domain LS estimation of CIR can be denoted as

$$\hat{\mathbf{H}}_{LS} = \mathbf{D}^\dagger \mathbf{r} = \mathbf{H} + \mathbf{D}^\dagger \tilde{\mathbf{w}}, \quad (12)$$

where  $\dagger$  denotes the pseudoinverse of a matrix and the noise power of LS estimation is calculated as  $\text{tr}\{\sigma_w^2 (\mathbf{D}^\dagger)^H \mathbf{D}^\dagger\}$ , and  $\text{tr}\{\cdot\}$  means the trace of a matrix.

Corresponding to [6], if the channel vector  $\mathbf{h}$  is Gaussian and uncorrelated with the channel noise  $\mathbf{w}$ , the MMSE estimate of  $\hat{\mathbf{h}}_{MMSE}$  could be obtained by

$$\hat{\mathbf{h}}_{MMSE} = \mathbf{R}_{hr} \mathbf{R}_{rr}^{-1} \mathbf{r}, \quad (13)$$

where

$$\begin{aligned} \mathbf{R}_{hr} &= E\{\mathbf{h}\mathbf{r}^H\} = \mathbf{R}_{hh} \mathbf{F}^H \mathbf{D}^H, \\ \mathbf{R}_{rr} &= E\{\mathbf{r}\mathbf{r}^H\} = \mathbf{D}\mathbf{F}\mathbf{R}_{hh}\mathbf{F}^H\mathbf{D}^H + \sigma_w^2 \mathbf{I}_{N_p}, \end{aligned} \quad (14)$$

where  $\mathbf{R}_{hr}$ ,  $\mathbf{R}_{rr}$ , and  $\mathbf{R}_{hh}$  are, respectively, the cross covariance matrix between  $\mathbf{h}$  and  $\mathbf{r}$ , the autocovariance matrix of  $\mathbf{r}$ , and the autocovariance matrix of  $\mathbf{h}$ . Noise variance  $\sigma_w^2$  and  $\mathbf{R}_{hh}$  are assumed to be known. The frequency domain estimate vector  $\hat{\mathbf{H}}_{MMSE}$  is given by

$$\hat{\mathbf{H}}_{MMSE} = \hat{\mathbf{F}}_{MMSE} \mathbf{h}. \quad (15)$$

Both estimators (12) and (15) have their drawbacks. The LS estimator has low complexity, but it performs a high mean square error. The MMSE estimator has its perfect performance. However, it suffers from a high complexity and needs to know the statistics of the channel ( $\sigma_w^2$  and  $\mathbf{R}_{hh}$ ) as a prior information. This condition is usually impossible in practical engineering.

### 4. CS Based Channel Estimation

*4.1. Compressed Sensing Overview.* The CS principles solve the problem of exactly reconstructing an  $N \times 1$  signal from with a small portion of linear measurements. Consider an  $N \times 1$  vector  $\mathbf{x}$ , which can be represented as  $\boldsymbol{\theta}$  in some orthogonal basis  $\boldsymbol{\psi} = [\psi_1, \psi_2, \dots, \psi_N]$ :

$$\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\theta}. \quad (16)$$

$\boldsymbol{\theta}$  is a  $K$ -sparse representation of  $\mathbf{x}$ , where  $K \ll N$ , because only  $K$  coefficients of  $\boldsymbol{\theta}$  are not zero. From CS theory,  $\mathbf{x}$  can be recovery from a linear measurement vector:

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}, \quad (17)$$

where  $\boldsymbol{\Phi}$  is an  $M \times N$  ( $M < N$ ) observation matrix satisfying Restricted Isometry Property (RIP).

*Restricted Isometry Property* [17]: the observation matrix  $\boldsymbol{\Phi}$  is said to satisfy the restricted isometry property of order  $S$  with parameter  $\delta_S \in (0, 1)$ , if

$$(1 - \delta_S) \|\mathbf{z}\|_2^2 \leq \|\boldsymbol{\Phi}\mathbf{z}\|_2^2 \leq (1 + \delta_S) \|\mathbf{z}\|_2^2 \quad (18)$$

holds for all  $S$ -sparse vectors  $\mathbf{z} \in \mathbf{R}^n$ .

If there is no noise in observation  $\mathbf{y}$ , the reconstruction problem can be solved by an  $l_1$ -norm optimization:

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta}. \quad (19)$$

Basis pursuit (BP) is appropriate to solve the basis pursuit problem. If the observation  $\mathbf{y}$  is contaminated with noise, then an additional norm of the residual  $\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta} - \mathbf{y}$  should be minimized as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_1 + \frac{\mu}{2} \|\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta} - \mathbf{y}\|_2^2 \quad (20)$$

or constrained

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\theta} - \mathbf{y}\|_2 \quad \text{s.t.} \quad \|\boldsymbol{\theta}\|_1 \leq \eta, \quad (21)$$

where  $\mu$  and  $\eta$  are constants. Formulation (20) is the  $l_1$ -regularized least squares problem and (21) is the least absolute

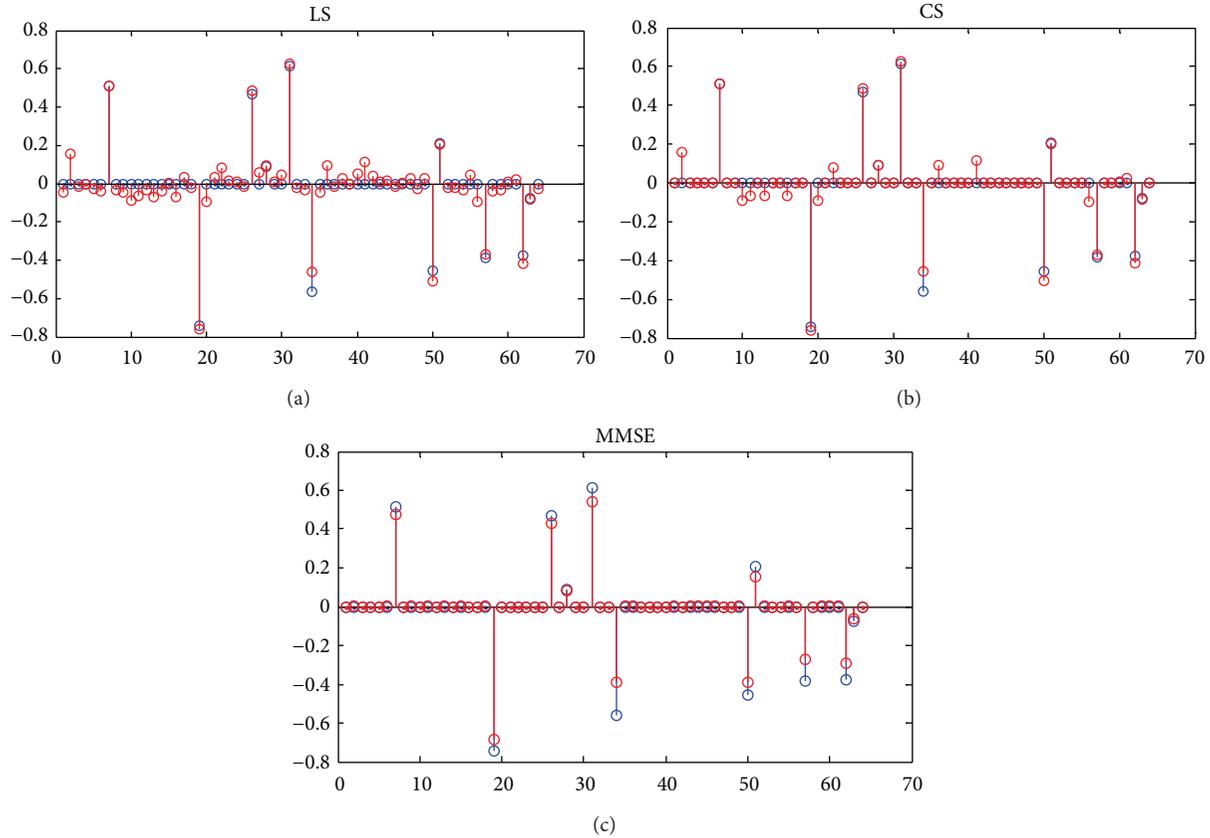


FIGURE 2: Channel estimated by LS, CS, and MMSE estimation.

shrinkage and selection operator (Lasso) problem [18]. On the other hand, there is also a greedy algorithm called Orthogonal Matching Pursuit (OMP) [19] to handle the vector recovery problem. Such method iteratively selects the local optimal solution step by step. The major advantages of OMP algorithm are its ease of implementation and its speed.

**4.2. CS Based Channel Estimation in DCT-OFDM.** Compare (9) with (17), we can find that channel estimation for DCT-OFDM system is able to be settled by CS theory. However, different from CSE in the existed literature, the orthogonal basis in our system model is DCT matrix other than DFT matrix:

$$\mathbf{r} = \mathbf{D}\mathbf{F}\mathbf{h} + \mathbf{w} = \mathbf{D}\mathbf{H} + \mathbf{w}, \quad (22)$$

where pilot data matrix  $\mathbf{D}$  represents as the measure matrix, DCT matrix  $\mathbf{F}$  acts as the sparse dictionary, and  $\mathbf{r}$  is the measure vector. Thus, pilot-aided channel estimation has been formulated as a sparse reconstruction problem discussed above. The computation feasible takes advantage of available fast transform algorithms for DCT [20].

Similar to (21), CIR could be obtained by solving

$$\hat{\mathbf{h}}_{\text{CSE}} = \arg \min_{\mathbf{h}} \|\mathbf{D}\mathbf{F}\mathbf{h} - \mathbf{r}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{h}\|_1 \leq \eta. \quad (23)$$

The frequency domain estimate vector  $\hat{\mathbf{H}}_{\text{CSE}}$  is given by

$$\hat{\mathbf{H}}_{\text{CSE}} = \mathbf{F}\hat{\mathbf{h}}_{\text{CSE}}. \quad (24)$$

Comparing to the LS and MMSE channel estimation, CSE exploits the inherent sparse property of multipath channels. Thereby, CSE leads to economizing the key communication resources of DCT-OFDM system, such as energy, latency, and bandwidth [10].

## 5. Simulation Results

In this section, we present the simulation results to compare the performance of the proposed LS, MMSE, and CS-based channel estimation methods for DCT-OFDM system. The channel  $\mathbf{h}$  is assumed to have  $L = 42$  taps. However, only  $K = 10$  taps have nonzero values and their positions are randomly distributed. The number of DCT-OFDM subcarriers is 256. Data sequences are modulated by binary phase shift keying (BPSK).

Figure 2 shows channel taps estimated by LS estimation, CS based estimation, and MMSE estimation, respectively. Blue taps are original channel taps. The estimation parameters are SNR = 10 dB, pilot symbols are modulated by BPSK, and pilots number  $N_p = 30$ . Channel estimated by CS based estimation is much clearer and less noise contaminated than LS estimation. Channel estimated by MMSE is the clearest and almost no noise contaminated.

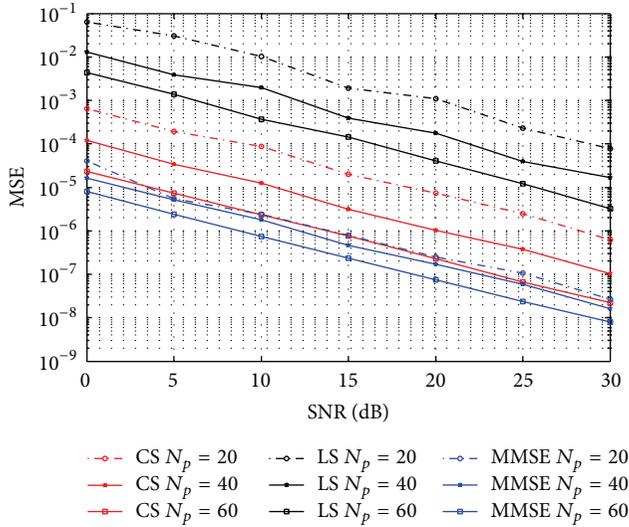


FIGURE 3: MSE performance of LS, CS, and MMSE estimation.

Figure 3 presents the mean square error (MSE) versus the signal-to-noise ratio (SNR). We select the number of pilots to be  $N_p = 14, 28,$  and  $42$ . For LS estimation, when number of pilot symbols is less than channel length  $L$ , performance of estimation is terrible. However, it is seen that the CS estimation significantly outperforms the LS estimation. Channel estimated by CS with pilots much less than LS is more accurate than LS. This phenomenon means that CS can save bandwidth and energy compared with LS. On the other hand, MMSE performs best due to its prior knowledge of channel statistic mentioned in Section 3. This is hardly possible in practical application.

Figure 4 plots the symbol error rate (SER) curves versus SNR. Channel estimations are done under the condition that pilot symbols are modulated by BPSK and pilots number  $N_p = 30$ . We can find that CS channel estimation performs much better than LS. MMSE is always the best among the three methods.

Figure 5 presents the MER curves that compare channel estimation performance for CSE using different classes of pilot symbols. Red curves represent random generated symbols, and blue curves represent BPSK modulated pilots. In the case of  $N_p = 14$ , performances of the two kinds of pilots seem pretty much the same thing. But performance of random modulated pilots is much better in the other two cases.

### 6. Conclusions

In this paper, we have investigated channel estimation in DCT-OFDM system. We have compared performance of LS, MMSE, and CS based estimation. CS has exploited the sparse property of multipath channel. Compared with LS estimation, CS estimation enables accurate channel estimation with less pilots. This means that CS estimation economizes the key communication resources of DCT-OFDM system, such as bandwidth and energy. Hence, CS channel estimation leads to high throughput and data rate. On the other hand,

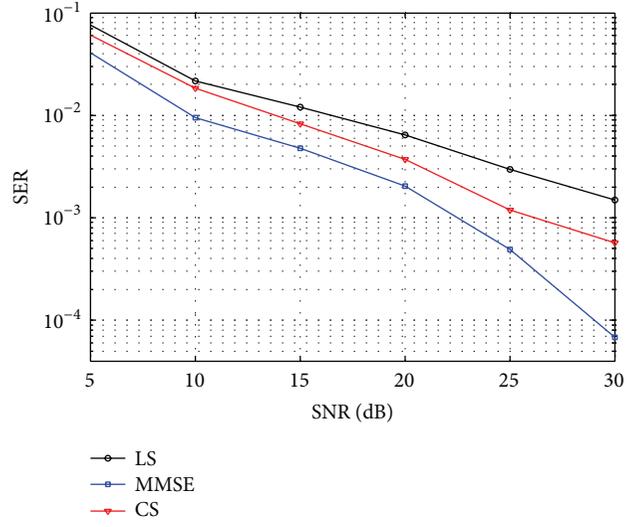


FIGURE 4: SER performance of LS, CS, and MMSE estimation.

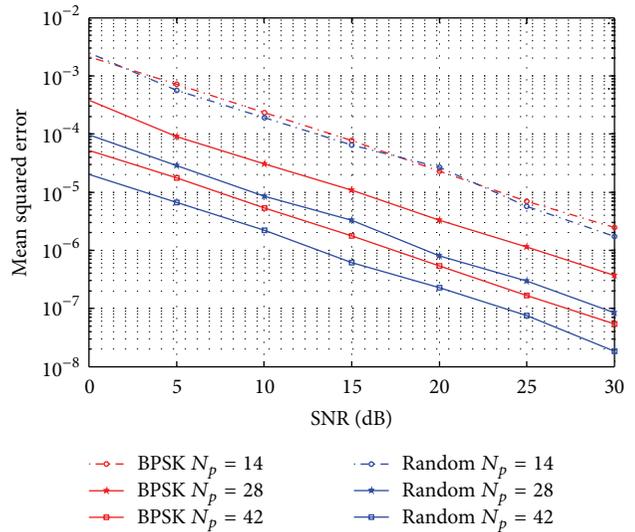


FIGURE 5: MER performance of CS channel estimation for different kind of training sequence.

MMSE can give the optimal channel estimation on the condition of knowing channel statistic as prior information. This condition is unachievable in practical engineering. In our future work, we will focus on CCS scheme in other communication systems.

### Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

### Acknowledgments

This work is supported by the National Natural Science Foundation of China (61032004) and the National

High Technology Research and Development Program of China ("863" Program) (2012AA121605, 2012AA01A503, and 2012AA01A510).

## References

- [1] S. B. Weinstein and P. M. Ebert, "Data transmission by frequency-division multiplexing using the discrete Fourier transform," *IEEE Transactions on Communications*, vol. 19, no. 5, pp. 628–634, 1971.
- [2] P. Tan and N. C. Beaulieu, "A comparison of DCT-Based OFDM and DFT-Based OFDM in frequency offset and fading channels," *IEEE Transactions on Communications*, vol. 54, no. 11, pp. 2113–2125, 2006.
- [3] N. Al-Dhahir, H. Minn, and S. Satish, "Optimum DCT-based multicarrier transceivers for frequency-selective channels," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 911–921, 2006.
- [4] N. Al-Dhahir and H. Minn, "A new multicarrier transceiver based on the discrete cosine transform," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '05)*, pp. 45–50, New Orleans, La, USA, March 2005.
- [5] G. D. Mandyam, "On the discrete cosine transform and OFDM systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 544–547, April 2003.
- [6] J.-J. van de Beek, O. Edfors, M. Sandell, S. K. Wilson, and P. O. Borjesson, "On channel estimation in OFDM systems," in *Proceedings of the IEEE 45th Vehicular Technology Conference*, pp. 815–819, July 1995.
- [7] Y. Li, N. Seshadri, and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 3, pp. 461–471, 1999.
- [8] A. Tomasoni, D. Gatti, S. Bellini, M. Ferrari, and M. Sitti, "Efficient OFDM channel estimation via an information criterion," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1352–1362, 2013.
- [9] X. Xiong, B. Jiang, X. Gao, and X. You, "DFT-based channel estimator for OFDM systems with leakage estimation," *IEEE Communications Letters*, vol. 17, no. 8, pp. 1592–1595, 2013.
- [10] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: a new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.
- [11] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [12] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, "Compressed channel sensing," in *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS '08)*, pp. 5–10, March 2008.
- [13] P. Cheng, Z. Chen, Y. Rui et al., "Channel estimation for OFDM systems over doubly selective channels: a distributed compressive sensing based approach," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4173–4185, 2013.
- [14] J.-C. Chen, C.-K. Wen, and P. Ting, "An efficient pilot design scheme for sparse channel estimation in OFDM systems," *IEEE Communications Letters*, vol. 17, no. 7, pp. 1352–1355, 2013.
- [15] D. Hu, X. Wang, and L. He, "A new sparse channel estimation and tracking method for time-varying OFDM systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 10, pp. 4173–4185, 2013.
- [16] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 90–93, 1974.
- [17] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [18] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 73, no. 3, pp. 273–282, 2011.
- [19] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems & Computers*, pp. 40–44, November 1993.
- [20] W.-H. Chen, C. H. Smith, and S. C. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Transactions on Communications*, vol. 25, no. 9, pp. 1004–1009, 1977.

## Research Article

# Cooperative Search and Rescue with Artificial Fishes Based on Fish-Swarm Algorithm for Underwater Wireless Sensor Networks

Wei Zhao, Zhenmin Tang, Yuwang Yang, Lei Wang, and Shaohua Lan

Computer Department, Nanjing University of Science and Technology, Jiangsu 210094, China

Correspondence should be addressed to Yuwang Yang; [yang-yuwang@163.com](mailto:yang-yuwang@163.com)

Received 23 December 2013; Accepted 29 January 2014; Published 5 March 2014

Academic Editors: Y. Mao, X. Meng, and Z. Zhou

Copyright © 2014 Wei Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper presents a searching control approach for cooperating mobile sensor networks. We use a density function to represent the frequency of distress signals issued by victims. The mobile nodes' moving in mission space is similar to the behaviors of fish-swarm in water. So, we take the mobile node as artificial fish node and define its operations by a probabilistic model over a limited range. A fish-swarm based algorithm is designed requiring local information at each fish node and maximizing the joint detection probabilities of distress signals. Optimization of formation is also considered for the searching control approach and is optimized by fish-swarm algorithm. Simulation results include two schemes: preset route and random walks, and it is showed that the control scheme has adaptive and effective properties.

## 1. Introduction

Recently, with the development of wireless sensor networks, especially the mobile wireless sensor networks, more and more applications emerge and have received general attention. Examples include pollution detection, wildfire monitoring, search and rescue missions, reconnaissance, and surveillance. In this paper, we position the application background at the problem of underwater search and rescue [1–3]. As we all know, underwater search and rescue is a very dangerous activity which requires professionals to conduct, and the rescue workers generally need to receive professional training to do this job. The primary task of search and rescue is to find and position the victim, and due to the complicated underwater environment, this is a very difficult job. With the development of bionics, the emergence of artificial fish [4–6] maybe has provided us with a cooperative search and rescue plan. We can install the searching device onto the artificial fish, and when the fish find the victim, they will send the position signal to the rescuing ship. In addition, through coordinated searching by multiple artificial fishes, the probability to find the victim can be significantly increased, and in the meantime, the searching time can be shortened, which is vital in rescuing tasks. This paper studies

the problem of how to coordinate multiple artificial fishes and conduct efficient and timely search, and it will be a meaningful task to discuss this problem.

The problem proposed in this paper is similar to the fundamental problem of cooperative coverage control or active sensing [7–10]. Cooperative control refers to settings which involve multiple controllable agents cooperating toward a common objective. In [7], the authors propose a coverage control algorithm aimed at maximizing target exposure in some surveillance applications, while in [11–14], heuristic algorithms based on potential fields and virtual forces are applied to push nodes away from each other and disperse them into the unoccupied areas in the mission space to enhance the coverage of a sensor network. The potential fields and virtual forces approach imitates the behavior of electromagnetic particles: when two electromagnetic particles are too close in proximity, a repulsive force pushes them apart. Applied to a sensor network, this method helps move sensors from high-density to low-density areas, thereby minimizing sensing overlap and improving the overall network coverage. In [8], a decentralized coverage control algorithm is proposed based on centroidal Voronoi partitioning, and a dynamic version of the Lloyd algorithm [15] has been used to iteratively find such a configuration. Other related works based on

Voronoi's partitions are included in [16, 17]. However, Lloyd's method suffers from two critical issues when it is used in mobile sensor networks. First, it does not consider the limited sensor communication range. Secondly, it does not optimize sensor movement distance; hence, it can lead to excessive energy consumption, a primary concern in sensor networks. Much of the active sensing literature [10] also concentrates on the problem of tracking specific targets using mobile sensors and the Kalman filter is extensively used to process observations and generate estimates.

In this paper, we consider a setting which involves a team of fish nodes and a set of target points (victims) in a three-dimensional space (e.g., under water). Each target point represents a victim. A mission is defined as the process of controlling the movement of the fish nodes and ultimately assigning them to target points so as to find the victims by visiting points within a given mission time  $T$ . In [18], the authors consider a setting where multiple vehicles form a team cooperating to visit multiple target points and collect rewards associated with them. Related work was proposed in their papers [19, 20]. Different from the papers mentioned above, in this paper, the target points are totally unknown in advance, or just located in an approximate range area. In addition, we pay more attention to find the target points in the shortest time. Because the earlier the victim can be found, the less the losses there will be; thus, we propose a cooperative coverage control scheme aiming at finding the target points in the shortest time under an uncertain environment.

The rest part of this paper is organized as follows: in Section 2, we propose the distributed cooperative searching control scheme for underwater mobile sensor networks. In Section 3, we optimize the scheme mentioned in Section 2 by fish-swarm algorithm. In Section 4, we simulate the distributed cooperative searching control algorithm by using computer software and evaluate its performance. Finally, in Section 5, we reach the main conclusions.

## 2. Fish-Swarm Searching Trajectory Model

We imagine a classic scenario: consider Area  $A$  with  $C(x, y)$  as its center of circle, and it has a radius of  $R$  and a depth of  $d$ . This could be a case of accident in the ocean, the accident center is the circle center  $C$ , the survivors may be scattered in Area  $A$  with a radius of  $R$ , but their specific locations are unknown, and the task objective is to find all the survivors in Area  $A$  with no omission, as shown in Figure 1.

Under the actual situation, the survivors tend to be scattered around the accident center, and the further from the center, the less the survivors there will be. In this paper, the mobile nodes participating in the search task are assumed as the fish swarm, and based on the situation mentioned above, the following two schemes are considered: in one scheme, after the fish swarm (mobile nodes) reaches the accident center, random walk is adopted; in another scheme, cruise is conducted in accordance with a certain preset route. The so-called random walk refers to that the fish swarm moves

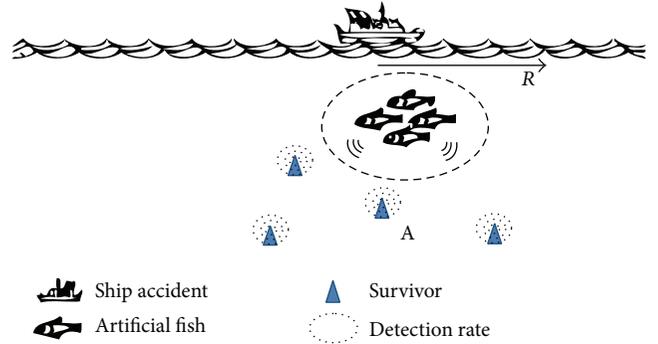


FIGURE 1: Task scenario description.

toward a certain direction; when they reach the task boundary, they randomly choose another direction and continue swimming, but the whole swimming track has to be within the task Area  $A$ . Here, we will discuss the second scheme, and in the experiment part of this paper, the effects of the two schemes will be compared. Assume the trajectory equation of fish swarm's cruise is  $F(r, \theta, z, t)$ , which represents that at any moment  $t$ , the spatial position of fish swarm is  $(r, \theta, z)$ , and Figure 2 has listed the corresponding helixes of several alternative helix equations. Of course, this is only an assumption, any equation will work, and the only difference is the effect.

In accordance with Figure 2, we can see that none of these helix curves satisfies the searching requirement; that is, there is a blind spot in exploration. An ideal track of fish swarm should satisfy the following several conditions. Firstly, it should cover the whole Area  $A$ . Secondly, it only needs to search any position in the area once; that is, there is repeated area, and in this way, it can ensure the shortest searching time. Thirdly, the trajectory equation should not be too complicated, which will sabotage the mechanical realization of robot fish's route. Therefore, in accordance with the above conditions, this paper constructs an applicable helix equation, which can be expressed as the following in a polar coordinate system:

$$F(r, \theta, z, t) = \begin{cases} r = \frac{R}{T}(t - 2nT), \\ 2nT \leq t < (2n + 1)T, \quad n = 0, 1, 2, \dots, \\ r = -\frac{R}{T}(t - 2nT - T) + R, \\ (2n + 1)T \leq t < (2n + 2)T, \quad n = 0, 1, \dots, \\ \theta = \theta_0 + \frac{R}{2R_s} 2\pi t, \\ z = z_0 - 2R_s m, \\ mT \leq t \leq (m + 1)T, \quad m = 0, 1, 2, \dots \end{cases} \quad (1)$$

In Formula (1),  $R$  refers the search radius of Area  $A$ ,  $T$  is a search cycle, the whole search process consists of multiple search cycles  $T$ ,  $R_s$  is the perceived radius of node,  $\theta_0$  and  $z_0$  represent the angle and the initial position of axis  $z$ , respectively, and  $m$  and  $n$  represent integers. In accordance

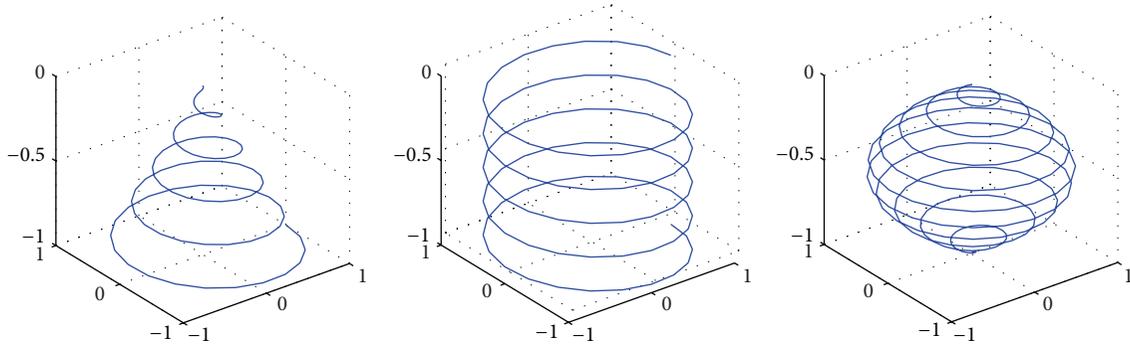


FIGURE 2: Helix curve.

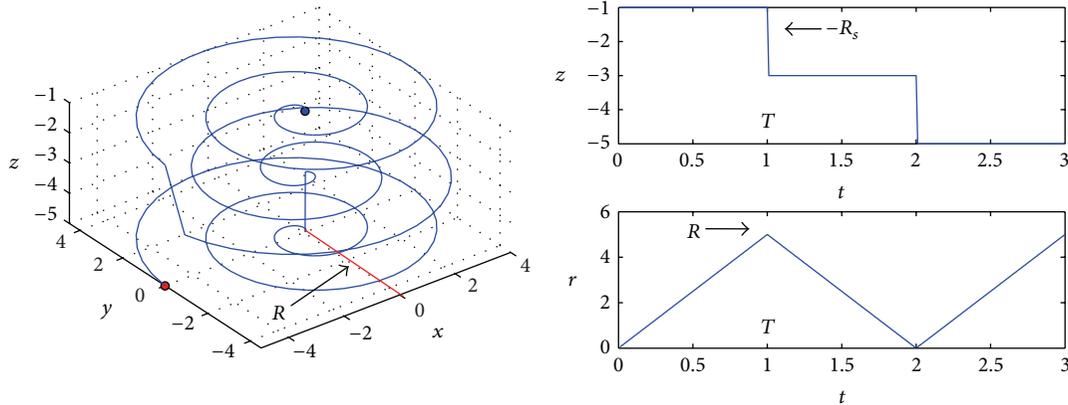


FIGURE 3: Helix constructed in this paper.

with the left diagram in Figure 3, the node moves from the initial position (blue dot) to the final position (red dot) through the helix, during which, 3 cycles ( $T$ ) of search have been conducted, the search of each cycle represents a plane on axis  $z$ , and it gets deeper and deeper. Therefore, the trajectory can cover the whole Area  $A$ , and there is no repeated path.

In accordance with the above right diagram in Figure 3, this is the curve of the search depth of fish swarm changing with time, the initial altitude of search is decided by  $z_0$ , and here,  $z_0 = -R_s$ , which can just cover the sea surface. After a search cycle  $T$  is completed, the fish swarm needs to search in a deeper position; that is,  $z_0 - 2R_s$ , and in this way, it can ensure that there will not be any omission or repeated search area. We call the search plane at the same altitude completed in each cycle  $T$  a search layer, and the whole search process consists of multiple such search layers. The lower part of the right diagram in Figure 3 is the curve of search radius  $r$  changing with time. The initial position is at the accident center, so  $r = 0$ ; as the fish swarm diffuses to the surrounding area,  $r$  reaches the maximum search radius  $R$  in a search layer; then, it will conduct backward search in the next search later, and  $r$  gradually reduces from  $R$  to 0. The benefits of trajectory equation with such design include full coverage of Area  $A$  and there is no area with repeated search, and in the meantime, the fish swarm can smoothly transit between various search layers.

### 3. Optimization Formation with Fish-Swarm Algorithm

If only one artificial fish participates into the search task, there is no need for formation optimization. Of course, during the actual process, in order to increase the searching scale and shorten the searching time, it is more reasonable to adopt collaborative search. Therefore, it will generate the formation optimization problem of fish swarm. Apparently, the formation of searching fish swarm should satisfy the following two requirements: firstly, the search section of fish swarm should be as big as possible; secondly, the fish swarm should maintain connection; that is, they should be within the communication range. Naturally, it is appropriate for us to use the artificial fish swarm optimization algorithm (AFSA) [21, 22] to optimize the formation of fish swarm.

**3.1. Network Coverage.** First of all, let us discuss the calculation of network coverage. As an important index to measure the strategy of sensor network deployment, the network coverage is generally defined as the ratio between the whole area that can be covered by nodes in the monitored area and the total monitored area. Considering the complication of monitoring environment in actual application, this paper has adopted the probability measurement model in the literature

[23] to calculate the network coverage. Assume the total number of nodes is  $N$ ,  $s_i$  representing the  $i$ th node in the network; then, the corresponding node set is  $S = \{s_i \mid i = 1, 2, \dots, N\}$ . Assume  $p(x, y, z)$  is a random point in Area  $A$ ,  $p \in A$ ; then the distance between node  $s_i(x_i, y_i, z_i)$  and point  $P$  is

$$D(s_i, p) = \|s_i - p\| = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}. \quad (2)$$

By adopting the probability measurement model in the literature [23], the detection probability of node  $s_i$  to point  $p$  is

$$P_p(s_i) = \begin{cases} 0, & R_s + R_e \leq D(s_i, p) \\ e^{-\alpha\lambda^\beta}, & R_s - R_e < D(s_i, p) < R_s + R_e \\ 1, & R_s - R_e \geq D(s_i, p), \end{cases} \quad (3)$$

in which  $R_s$  refers to the perceived radius of various nodes in the network,  $R_e$  refers to the uncertain factors within the measurement range of nodes, and  $0 < R_e < R_s$ ;  $\alpha$  and  $\beta$  refer to the measured parameters related to the physical device;  $\lambda$  is the input parameter, which is defined as

$$\lambda = D(s_i, p) - (R_s - R_e). \quad (4)$$

Therefore, we can obtain the joint detection probability of multiple sensor nodes simultaneously conducting measurement to the target point  $p$  as

$$P_p(S) = 1 - \prod_{n_i \in B} (1 - P_p(s_i)), \quad (5)$$

in which  $S$  refers to the sensor node set of the measured target point. Therefore, in order to realize full coverage of the target area, it requires satisfying the following condition:

$$\min \{P_p(S)\} \geq c_{th}, \quad (6)$$

in which  $c_{th}$  ( $0 < c_{th} < 1$ ) refers to the threshold value of detection probability set in accordance with different application requirements. In order to calculate the coverage of sensor network, we need to conduct grid processing of the monitored area (the bigger the distance between adjacent grid points is, the higher the calculation accuracy is), and then, the joint detection probability at each grid point is solved. The percentage of grids which meet  $P \geq c_{th}$  is called network coverage.

As shown in Figure 4, Area  $A$  with grids is the monitored area, and the shadow areas  $U1$  and  $U2$  represent the perceived areas of  $s_1$  and  $s_2$ , respectively. Assume Area  $A$  is divided into  $N$  grids, and  $p$  refers to a random grid; then its network coverage is

$$\frac{N \{ \min \{ P_{p \in (A \cap (U1 \cup U2))} (s_1, s_2) \} \geq c_{th} \}}{N}. \quad (7)$$

In Formula (7), the numerator refers to the number of grids that satisfy the threshold value of detection probability  $c_{th}$ , while the denominator refers to the total number of grids.

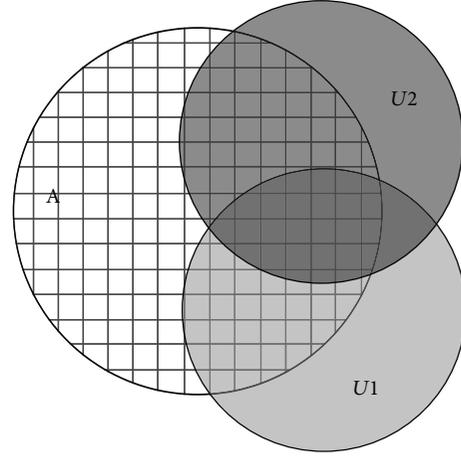


FIGURE 4: Schematic diagram of network coverage.

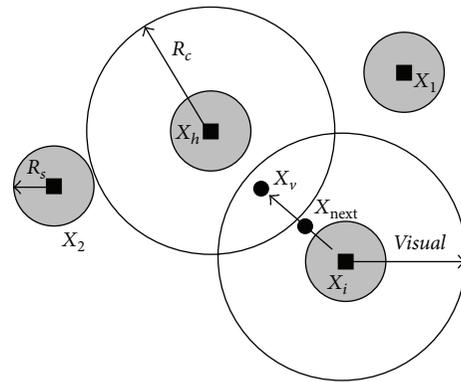


FIGURE 5: Vision concept of the artificial fish.

Because our task model is in a 3-dimensional space, the calculation process is conducted in accordance with the 3D grids, and Figure 5 only shows an example.

Assume under current state, the position vector of all sensor nodes in the network is  $U = [X_n \ Y_n \ Z_n]$ , and the function to calculate the network coverage with  $U$  as the input variable is  $f(U)$ , in which  $X_n = \{x_1, x_2, \dots, x_n\}$ ,  $Y_n = \{y_1, y_2, \dots, y_n\}$ , and  $Z_n = \{z_1, z_2, \dots, z_n\}$  are the node coordinate vectors. We abstract the optimization of network layout into solving the optimization problem with  $f(U)$  as the objective function.

**3.2. Optimization with Fish-Swarm Algorithm.** The artificial fish-swarm optimization algorithm (AFSA) is an optimization algorithm which simulates the behavior of fish swarm, and it uses the preying, gathering, and chasing behavior of fish swarm to find fast and global optimum solution. In this paper, the fish swarm consisting of mobile nodes can also be regarded as a cluster of nodes, in which the cluster head node (the leader of fish swarm) conducts cruise in accordance with the trajectory equation  $F(r, \theta, z, t)$  mentioned in the last section; other nodes are processed in accordance with the fish swarm algorithm, and in this way, it can ensure that the

whole fish swarm maintains reliable communication and a maximum search scale.

Assume in an  $n$ -dimensional target search space, there is a fish swarm consisting of  $N$  artificial fish, and every day, the state of individual artificial fish can be expressed as vector  $\mathbf{X} = (x_1, x_2, \dots, x_n)$ , in which  $x_i (i = 1, \dots, n)$  is the variable that needs to be optimized: the food concentration of current location where the artificial fish are can be expressed as  $Y = f(\mathbf{X})$ , in which  $Y$  is the objective function; the distance between individual artificial fish can be expressed as  $d_{ij} = \|\mathbf{X}_i - \mathbf{X}_j\|$ ; visual refers to the perceived range of artificial fish, step is the moving step length of artificial fish, and  $\delta$  is the crowding degree factor; try\_number represents the maximum trying times each time the artificial fish go preying.

The AF realizes external perception by its vision shown in Figure 5.  $X_i$  is the current position of an AF,  $X_h$  is the cluster head's position,  $Visual (R_c, \text{communication radius})$  is the visual distance, and  $X_v$  is the visual position at some moment. If the position at the visual position is better than the current position, it goes forward a step in this direction, and arrives at the  $X_{\text{next}}$  position; otherwise, it continues an inspecting tour in the vision.

In the sensor network, the process of mobile nodes exploring toward bigger network coverage is similar to the chasing and preying behavior by individual artificial fish, and the food concentration of current location where the artificial fish are can be regarded as the network coverage under current state. Fish usually stay in the place with a lot of food, so we simulate the behaviors of fish based on this characteristic to find the global optimum, which is the basic idea of the AFSA. The basic behaviors of AF are defined as follows.

(1) *AF\_Prey*. This is a basic biological behavior heading for the food; generally the fish perceives the concentration of food in water to determine the movement by vision or sense and then chooses the tendency.

Behavior description: let  $X_i$  be the AF current state and select a state  $X_j$  randomly in its visual distance,  $Y$  is the food concentration (objective function value), the greater the  $Visual$  is, the more easily the AF finds the global extreme value and converges:

$$\mathbf{X}_j = \mathbf{X}_i + Visual.Rand () . \quad (8)$$

If  $Y_i < Y_j$  in the maximum problem, it goes forward a step in this direction; otherwise, select a state  $X_j$  randomly again and judge whether it satisfies the forward condition. If it cannot satisfy after try\_number times, it moves a step randomly. When the try\_number is small in AF\_Prey, the AF can swim randomly, which makes it flee from the local extreme value field:

$$\mathbf{X}_i^{(t+1)} = \mathbf{X}_i^{(t)} + Visual.Rand () . \quad (9)$$

(2) *AF\_Swarm*. The fish will assemble in groups naturally in the moving process, which is a kind of living habits in order to guarantee the existence of the colony and avoid dangers. Behavior description: let  $X_i$  be the AF current state,  $X_c$  the center position, and  $nf$  the number of its companions in the current neighborhood ( $d_{ij} < Visual$ ),  $N$  is the number of

total fishes. If  $Y_c > Y_i$  and  $nf/N < \delta$ , which means that the companion center has more food (higher fitness function value) and is not very crowded, it goes forward a step to the companion center:

$$\mathbf{X}_i^{(t+1)} = \mathbf{X}_i^{(t)} + \frac{\mathbf{X}_c - \mathbf{X}_i^{(t)}}{\|\mathbf{X}_c - \mathbf{X}_i^{(t)}\|} Step.Rand () . \quad (10)$$

Otherwise, it executes the preying behavior. The crowd factor limits the scale of swarms, and more AF only cluster at the optimal area, which ensures that AF moves to optimum in a wide field.

(3) *AF\_Follow*. In the moving process of the fish swarm, when a single fish or several ones find food, the neighborhood partners will trail and reach the food quickly. Behavior description: let  $X_i$  be the AF current state, and it explores the companion  $X_j$  in the neighborhood ( $d_{ij} < Visual$ ), which has the greatest  $Y_j$ . If  $Y_j > Y_i$  and  $nf/N < \delta$ , which means that the companion  $X_j$  state has higher food concentration (higher fitness function value) and the surroundings is not very crowded, it goes forward a step to the companion  $X_j$ :

$$\mathbf{X}_i^{(t+1)} = \mathbf{X}_i^{(t)} + \frac{\mathbf{X}_j - \mathbf{X}_i^{(t)}}{\|\mathbf{X}_j - \mathbf{X}_i^{(t)}\|} Step.Rand () . \quad (11)$$

Otherwise, it executes the preying behavior.

(4) *AF\_Move*. Fish swim randomly in water; in fact, they are seeking food or companions in larger ranges. Behavior description: it chooses a state at random in the vision; then it moves towards this state; in fact, it is a default behavior of AF\_Prey:

$$\mathbf{X}_i^{(t+1)} = \mathbf{X}_i^{(t)} + Visual.Rand () . \quad (12)$$

Therefore, by using the gathering and chasing behavior of fish swarm, it can draw the nodes close to the cluster head. In the meantime, in order to realize maximum search range, the fish swarm should maintain a good formation, which could be controlled through the crowding factor  $\delta$  and the distance between fishes. The formation optimization process based on artificial fish swarm consists of the following specific steps.

- (a) Initialize the wireless sensor network. Ensure the scale of artificial fish swarm  $N$  in accordance with the application requirement, the maximum moving step length of artificial fish is  $step$ , the visible range of artificial fish is  $visual$ , the maximum iteration times are  $k$ , and the crowding degree factor is  $\delta$ .
- (b) Initialize the fish swarm  $X$ , randomly generate  $N$  individual artificial fish within the task space, and in the meantime, set the initial iteration times as  $k = 0$ .
- (c) Calculate the food concentration of current location where the initial individual fish in the swarm is  $Y_i$  (i.e., the network coverage  $Y_i$ ); then put them in sequence, and select the individual artificial fish with the biggest value of  $Y_i$  to enter the billboard  $T$ .

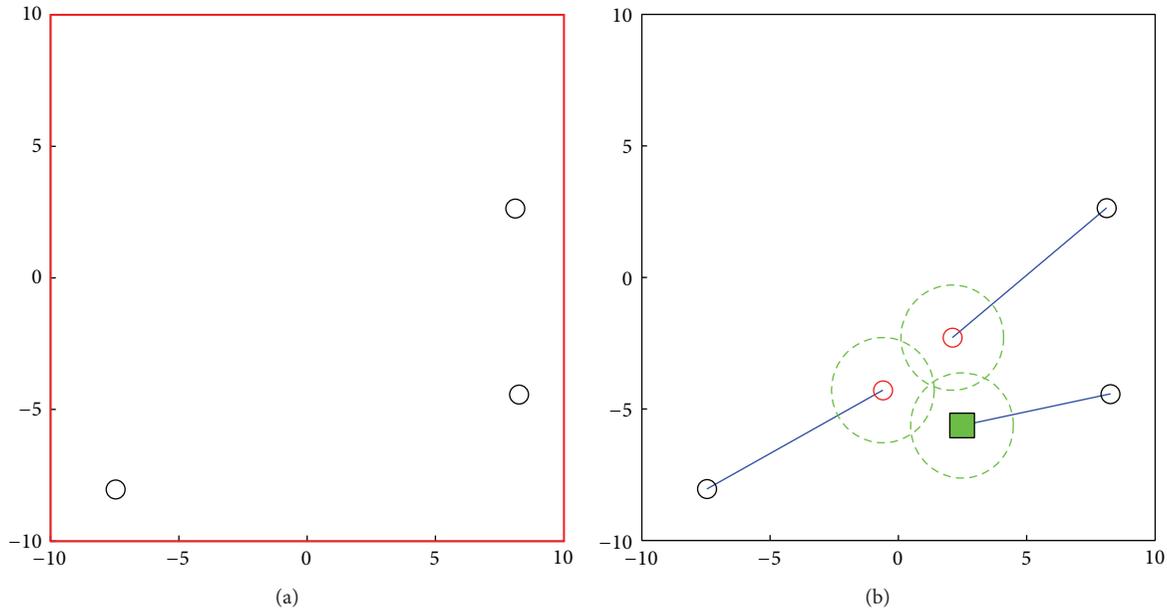


FIGURE 6: Example of 3 fishes forming a swarm.

- (d) The artificial fish simulate the gathering and chasing behavior of fish swarm, and the fish with big  $Y$  value is selected to conduct preying behavior.
- (e) After each action of various artificial fish, the food concentration of current location  $Y$  will be compared to the  $Y_T$  of artificial fish on the billboard, and if it is bigger than the value of  $Y_T$  on the Billboard, this artificial fish will replace the billboard fish and enter billboard.
- (f) Determine the end condition. If it has reached the maximum iteration times, the  $Y$  value of billboard will be output, that is, the optimum formation solution; otherwise,  $k = k + 1$ , go to (d).

3.3. *Example for Optimization Formation.* How to make the fish swarm search in a bigger scale? An easy approach is to make all fish in the fish swarm stay on the same plane, this plane is vertical to the direction which this fish swarm moves toward, and this could realize maximization of the cross-section area. In the meantime, efforts should be made to avoid any gap in the middle of fish swarm because the gap might cause blind spot during the search. Take a search fish swarm of 3 fishes for example, and after the 3 fishes reach the accident center, respectively, the formation process will successively begin. The initial positions of the 3 fishes could be random, they are optimized in accordance with the optimization algorithm in Section 3.2, they come closer to form a fish swarm, in the meantime, the distance between individuals is controlled to avoid any gap, and Figure 6 has shown this process.

The small circles on the left diagram represent the initial positions of 3 fishes, the red small circles on the right diagram represent the final positions after forming a swarm, the blue solid line refers to the trace of moving from the initial position

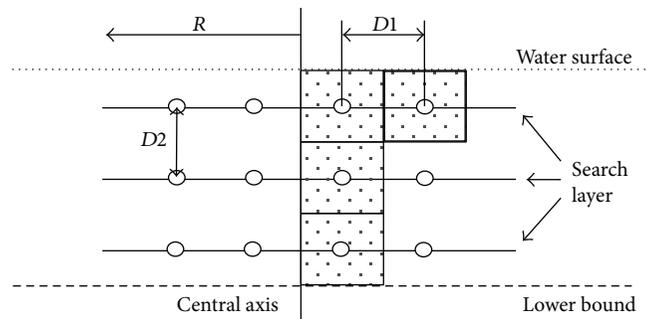


FIGURE 7: Diagrammatic sketch of search cross-section.

to the final position, and the green block represents the leader of fish swarm after formation (cluster head node).

Apparently, if the search is conducted in accordance with the formation in Figure 6, it might cause blind spot during search because it is very difficult to ensure that there is no gap between various search layers. If there is a fish swarm consisting of  $N$  fish, for the convenience of not leaving any gap between various search layers or between pitches, as shown in Figure 7, we need to calculate the maximum rectangular cross-sectional area for searching the fish swarm.

In Figure 7,  $R$  represents the radius of search area  $A$ ,  $D1$  refers to the space between helixes on the same search layer, and  $D2$  refers to the space between various search layers. The shadow block area in bold represents the search cross-section, and we can see the size of cross-section is  $D1 \times D2$ . When  $N = 1, 4, 9$ ,  $D1$  and  $D2$  are as shown in Figure 8.

#### 4. Simulation and Results

We consider mobile nodes to be operating in a three-dimensional underwater mission space. Assume that the

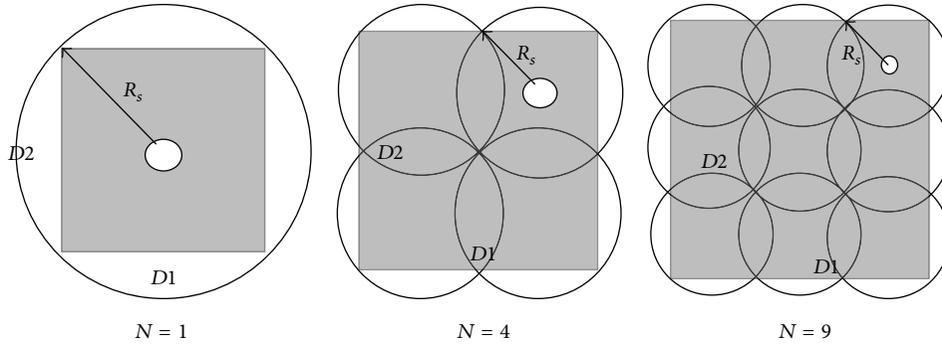


FIGURE 8: Diagrammatic sketches of  $D1$  and  $D2$  when  $N = 1, 4, 9$ .

mission is to find survivors from  $M$  targets using  $N$  nodes. Let set  $G$  denote the  $M$  targets  $G = \{m_j \mid j = 1, 2, \dots, M\}$  and let set  $B$  denote  $N$  nodes  $B = \{n_i \mid i = 1, 2, \dots, N\}$ . Note that a target may change its location during operation of the system, or new targets may show up; hence  $M$  and  $y_j$  may not be constant. At the same time, a node may malfunction, and thus  $N$  may also change in time. Associated with the  $j$ th target is a life value  $L_j$ , if  $L_j \leq 0$ , this means target  $j$ th fails, and in reality, it might mean that the survivor is dead. The mission's objective is to maximize the total life value collected by visiting target points in the set  $G$  within a given mission time  $T$ . Target life value may be time dependent, typically decreasing in time. The exact location of targets may not always be known in advance and there may be obstacles in the mission space, which constrain the feasible trajectories of nodes.

We model the mission space as a cylinder  $A \subset \mathbf{R}^3$ , over which there is an event density function  $\rho(x)$ ,  $x \in A$ , that captures the frequency or density that a specific event takes place (in Hz/m<sup>3</sup>).  $\rho(x)$  satisfies  $\rho(x) \geq 0$  for all  $x \in A$ . In this paper,  $\rho(x)$  may be the frequency that a survivor appears at target point. When an event occurs at point  $x$ , it emits a signal and this signal is observed by a sensor at that location nearby.

To distinguish the relative importance of targets at time  $t$ , each target has an associated life function denoted by  $L_j \Phi_j(t)$ , where  $L_j$  is the maximal life value and  $\Phi_j(t) \in [0, 1]$  is a discounting function which describes the life value change over time. When a deadline is associated with a particular target point, we can use

$$\Phi_j(t) = \begin{cases} 1 - \frac{\mu_j}{H_j}t, & t \leq H_j, \\ (1 - \varepsilon_j) e^{-\eta_j(t-H_j)}, & t > H_j, \end{cases} \quad (13)$$

where  $H_j$  is a deadline assigned to target point  $j$  and  $\varepsilon_j \in (0, 1]$ ,  $\eta_j > 0$  are parameters which may be target specific and are chosen to reflect different cases of interest.

The optimal searching problem can be formulated as an optimization problem to maximize the expected life value collected by the sensors over the mission space  $A$ :

$$\text{MAX} \int_A L_j \Phi_j(A) \rho(A) P_A(B) F(A) dA. \quad (14)$$

By referring to the above definition,  $L_j \Phi_j(A)$  in Formula (14) refers to the fact that the life value represented by the

target points in area  $A$  reduces with the discounting function,  $\rho(A)P_A(B)$  refers to the joint detection probability of the target points in area  $A$  by all nodes in set  $B$ , and  $F(A)$  refers to the node trajectory equation in area  $A$ . Therefore, the meaning of Formula (14) is to maximize the expected life value collected by the sensors over the mission space  $A$ .

Table 1 shows the parameters of simulation experiment, and in order to make the experiment as close to the actual situation as possible, we have set the experiment parameters.

**4.1. Total Search Time versus Nodes Number.** Figure 9 shows the relational diagram between the overall searching time and the node number under two schemes. In the first scheme, the fish swarm conducts cruise in accordance with a helix equation (blue bar graph), and in the second scheme, random walk is adopted. In accordance with the diagram, we can see that in both schemes, the overall searching time reduces with the increase of node number ( $N = 1, 4, 9$ ), but the random walk scheme takes significant more overall searching time than the helix scheme. When the node number is  $N = 1, 4, 9$ , the time taken by the helix scheme is, respectively, 21.9%, 22.2%, and 22.6% of the time taken by the random walk scheme. When the node number is  $N = 1$ , the time taken by the helix scheme is 2.99 times of that when  $N = 4$  and 8.96 times of that when  $N = 9$ . It can significantly reduce the searching time by increasing the number of fishes in the fish swarm, but of course, this will also increase the communication consumption and complexity. A feasible method is to increase the number of fish swarms, which can also achieve the effect of reducing the overall searching time.

**4.2. Average Rescue Time versus Nodes Number.** Figure 10 shows the relational diagram between the average rescuing time and the node number under two schemes. The average rescuing time refers to the average time to find 5 target points. In accordance with the diagram, we can see that in both schemes, the average rescuing time reduces with the increase of node number, but the average rescuing time taken by the random walk scheme is longer than the helix scheme. When the node number is  $N = 1, 4, 9$ , the time taken by the helix scheme is, respectively, 41.6%, 38.2%, and 55.4% of the time taken by the random walk scheme. In accordance with the set parameters in Table 1, when the node number is  $N = 1$ , the average rescuing time taken by the helix

TABLE I: Experiment parameters.

Network configuration	Node	Artificial fish	Others
Radius $R = 1$ km	$R_s = 50$ m	Swimming speed $V = 1.5$ m/s	Life value = 10 minutes
Depth $d = 200$ m	$R_c = 100$ m	$Visual = R_c$	
$c_{th} = 0.9$	$R_e = 0.5 R_s$	$Step = R_s$	
$N = \{1, 4, 9\}$	$\alpha = 0.2; \beta = 2.0$		
Targets: $M = 5$			

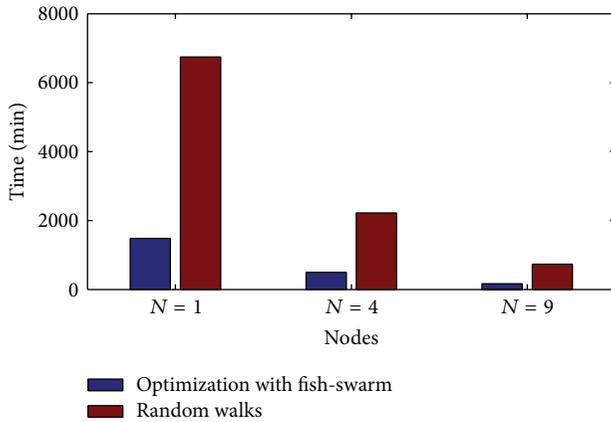


FIGURE 9: Overall searching time versus node number.

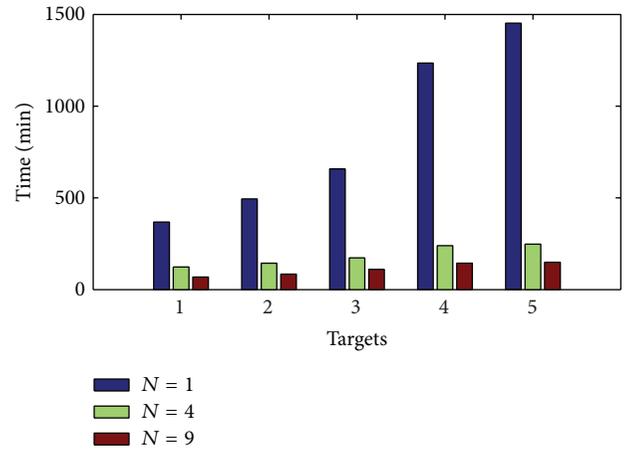


FIGURE 11: Target discovery time.

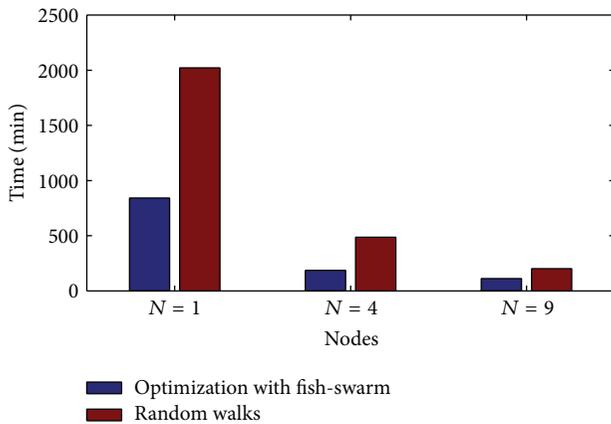


FIGURE 10: Average rescuing time versus node number.

scheme is 841.5 minutes, and when  $N = 4$  and  $N = 9$ , it takes 185.5 minutes and 111.2 minutes, respectively. This is far from the 10-minute threshold value of life value we expect, and the solutions include increasing the swimming speed of fish swarm, increasing the number of fish swarms, or increasing the number of fishes in the fish swarm. Increasing the swimming speed of fish requires consideration on the aspect of mechanics, which exceeds the discussion in this paper; increasing the number of fishes in the fish swarm is restricted because excessive nodes in the swarm will cause communication delay, which will also increase the control overhead. A feasible method is to increase the number of fish swarms while at the time ensuring that there will not

be excessive fishes in the fish swarm, which will significantly reduce the average rescuing time.

**4.3. Target Discovery Time.** Figure 11 shows the target discovery time by the fish swarm in accordance with the helix equation under the cruise scheme, and 5 total target points are found, which are randomly distributed in the whole task Area A. Record the 5 time points where the fish swarm find them in accordance with the cruise route. In accordance with the diagram, we can see that the discovery time of 5 targets presents gradual progressive increase, which is easy to understand. In reality, the sooner the target is discovered, the more beneficial it is to the rescue work, which requires the target discovery time to be as short as possible. In the diagram, we can see that it can significantly reduce the target discovery time by increasing the node number. For example, when  $N = 9$ , the time to discover the first target is 244.9 minutes shorter than that when  $N = 1$ , and the time to discover the fifth target is 1204.2 minutes shorter. In the meantime, in order to further reduce the target discovery time, the number of fish swarms can also be increased, so that the fish swarms can conduct searching at different search layers, which can significantly reduce the target discovery time and increase the rescuing efficiency.

**4.4. Optimized Formation versus No Optimization.** In accordance with the above result, we can see that the optimized formation can increase the search cross-sectional area of fish swarm, and Figure 12 shows the comparison between the optimized formation and the formation with no optimization

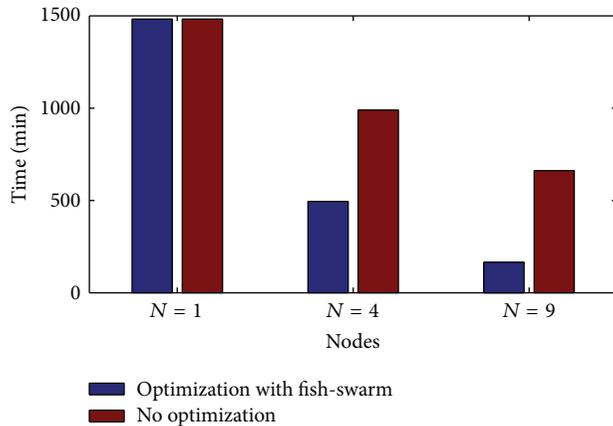


FIGURE 12: Comparison between the optimized formation and the formation with no optimization.

when the node number is  $N = 1, 4, 9$ . In accordance with the diagram, we can see that optimized formation to the fish swarm can significantly reduce the searching time. When  $N = 1$ , because there is only one fish, there is no so-called formation optimization; when  $N = 4$ , the efficiency has increased by 51.3% after optimization; when  $N = 9$ , the efficiency has increased by 74.6% after optimization. Apparently, the more nodes there are, the more significant the optimization result is. However, during the actual application, there should not be too many fishes in the fish swarm because this will cause a high cost of optimization.

## 5. Conclusions

Under the background of sea rescue, this paper proposes a search scheme based on the fish-swarm optimization algorithm, and the sensor nodes moving underwater are considered as the artificial fishes. In this paper, a trajectory model of fish swarm cruise is built in accordance with the helix equation, the fish-swarm optimization algorithm is used to optimize the formation of nodes, and the fish swarm behavior, such as gathering, chasing, and preying, is used to control the movement of nodes, so that maximum cross-sectional area of search by the fish swarm can be achieved. The experiment result shows that compared to the random walk model, it can significantly reduce the searching time in accordance with the preset trajectory cruise, which has reference value for practical application; in the meantime, it can also help reduce the searching time through formation optimization of fish swarm. Sea rescue is a task that has high requirement of saving time, the next step of work should be further in-depth research on the issues discussed in this paper, and how to adapt to underwater barriers will also be discussed.

## Conflict of Interests

The authors declare that they have no conflict of interests.

## Acknowledgment

This paper is a part of the work supported by the National Natural Science Foundation of China (no. 61301108).

## References

- [1] M. Kemp, B. Hobson, J. Meyer, R. Moody, H. Pinnix, and B. Schulz, "MASA: a multi-AUV underwater search and data acquisition system," in *Proceedings of the MTS/IEEE (OCEANS '02)*, vol. 1, pp. 311–315, October 2002.
- [2] B. Anderson and J. Crowell, "Workhorse AUV—a cost-sensible new autonomous underwater vehicle for surveys/soundings, search & rescue, and research," in *Proceedings of the MTS/IEEE (OCEANS '05)*, pp. 1–6, September 2005.
- [3] A. Jacoff, *Search and Rescue Robotics*, Springer, 2008.
- [4] E. P. Gels, *Electroactive Polymer (EAP) Actuators As Artificial Muscles: Reality, Potential, and Challenges*, 2004.
- [5] J. Yu, S. Wang, and M. Tan, "A simplified propulsive model of bio-mimetic robot fish and its realization," *Robotica*, vol. 23, no. 1, pp. 101–107, 2005.
- [6] Z. Wang, Y. Wang, J. Li, and G. Hang, "A micro biomimetic manta ray robot fish actuated by SMA," in *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO '09)*, pp. 1809–1813, December 2009.
- [7] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava, "Coverage problems in wireless ad-hoc sensor networks," in *Proceedings of the 20th IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM '01)*, vol. 3, pp. 1380–1387, April 2001.
- [8] J. Cortés, S. Martínez, T. Karataş, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243–255, 2004.
- [9] W. Li and C. G. Cassandras, "A minimum-power wireless sensor network self-deployment scheme," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '05)*, vol. 3, pp. 1897–1902, March 2005.
- [10] L. Mihaylova, T. Lefebvre, H. Bruyninckx, and K. Gadeyne, "Active sensing for robotics—a survey," in *Proceedings of the 5th International Conference on Numerical Methods and Applications*, pp. 316–324, Borovets, Bulgaria, 2002.
- [11] A. Howard, M. J. Mataric, and G. S. Sukhatme, "Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem," in *Distributed Autonomous Robotic Systems*, vol. 5, pp. 299–308, Springer, Tokyo, Japan, 2002.
- [12] D. O. Popa, H. E. Stephanou, C. Helm, and A. C. Sanderson, "Robotic deployment of sensor networks using potential fields," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '04)*, vol. 1, pp. 642–647, May 2004.
- [13] G. Tan, S. A. Jarvis, and A.-M. Kermarrec, "Connectivity-guaranteed and obstacle-adaptive deployment schemes for mobile sensor networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 836–848, 2009.
- [14] A. Casteigts, J. Albert, S. Chaumette, A. Nayak, and I. Stojmenovic, "Biconnecting a network of mobile robots using virtual angular forces," *Computer Communications*, vol. 35, no. 9, pp. 1038–1046, 2012.
- [15] Y. Song, B. Wang, Z. Shi, K. Pattipati, and S. Gupta, *Distributed Algorithms for Energy-Efficient Even Self-Deployment in Mobile Sensor Networks*, 2013.

- [16] C. Nowzari and J. Cortés, “Self-triggered coordination of robotic networks for optimal deployment,” *Automatica*, vol. 48, no. 6, pp. 1077–1087, 2012.
- [17] J. Cortes, “Coverage optimization and spatial load balancing by robotic sensor networks,” *IEEE Transactions on Automatic Control*, vol. 55, no. 3, pp. 749–754, 2010.
- [18] C. Yao, X. C. Ding, and C. G. Cassandras, “Cooperative receding horizon control for multi-agent rendezvous problems in uncertain environments,” in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC '10)*, pp. 4511–4516, December 2010.
- [19] J. Cortes, S. Martinez, T. Karatas, and F. Bullo, “Distributed coverage control and data with mobile sensor networks,” *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2445–2455, 2011.
- [20] X. C. Ding, C. Belta, and C. G. Cassandras, “Receding horizon surveillance with temporal logic specifications,” in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC '10)*, pp. 256–261, December 2010.
- [21] X. L. Li and J. X. Qian, “Studies on artificial fish swarm optimization algorithm based on decomposition and coordination techniques,” *Journal of Circuits and Systems*, vol. 1, pp. 1–6, 2003.
- [22] H.-C. Tsai and Y.-H. Lin, “Modification of the fish swarm algorithm with particle swarm optimization formulation and communication behavior,” *Applied Soft Computing Journal*, vol. 11, no. 8, pp. 5367–5374, 2011.
- [23] Y. Zou and K. Chakrabarty, “Sensor deployment and target localization based on virtual forces,” in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 2, pp. 1293–1303, April 2003.