

Complexity

Collective Behavior Analysis and Graph Mining in Social Networks 2021

Lead Guest Editor: Fei Xiong

Guest Editors: Shirui Pan, Sen Wang, Ziming Zhang, and Xuzhen Zhu





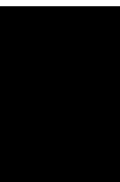
**Collective Behavior Analysis and Graph Mining
in Social Networks 2021**

Complexity

Collective Behavior Analysis and Graph Mining in Social Networks 2021

Lead Guest Editor: Fei Xiong

Guest Editors: Shirui Pan, Sen Wang, Ziming Zhang, and Xuzhen Zhu



Copyright © 2022 Hindawi Limited. All rights reserved.

This is a special issue published in "Complexity." 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.

Chief Editor

Hiroki Sayama, USA

Editorial Board

Oveis Abedinia, Kazakhstan
José Ángel Acosta, Spain
Andrew Adamatzky, United Kingdom
Marcus Aguiar, Brazil
Carlos Aguilar-Ibanez, Mexico
Mojtaba Ahmadiéh Khanesar, United Kingdom
Tarek Ahmed-Ali, France
Alex Alexandridis, Greece
Basil M. Al-Hadithi, Spain
Diego R. Amancio, Brazil
Maia Angelova, Australia
David Arroyo, Spain
Tomaso Aste, United Kingdom
George Bassel, United Kingdom
Abdellatif Ben Makhlouf, Saudi Arabia
Rosa M. Benito, Spain
Johan Bollen, USA
Erik M. Boltt, USA
Mohamed Boutayeb, France
Dirk Brockmann, Germany
Átila Bueno, Brazil
Seth Bullock, United Kingdom
Ning Cai, China
Eric Campos, Mexico
Anirban Chakraborti, India
Émile J. L. Chappin, The Netherlands
Chih-Chiang Chen, Taiwan
Yu-Wang Chen, United Kingdom
Diyi Chen, China
Siew Ann Cheong, Singapore
Hocine Cherifi, France
Matteo Chinazzi, USA
Giulio Cimini, Italy
Danilo Comminiello, Italy
Pierluigi Contucci, Italy
Roger Cremades, The Netherlands
Salvatore Cuomo, Italy
Sergey Dashkovskiy, Germany
Manlio De Domenico, Italy
Pietro De Lellis, Italy
Albert Diaz-Guilera, Spain
Thach Ngoc Dinh, France
James A. Dixon, USA

Alan Dorin, Australia
Sheng Du, China
Jordi Duch, Spain
Marcio Eisencraft, Brazil
Mondher Farza, France
Guilherme Ferraz de Arruda, Italy
Giacomo Fiumara, Italy
Thierry Floquet, France
José Manuel Galán, Spain
Lucia Valentina Gambuzza, Italy
Harish Garg, India
Georgi Yordanov Georgiev, USA
Carlos Gershenson, Mexico
Peter Giesl, United Kingdom
Sergio Gómez, Spain
Lingzhong Guo, United Kingdom
Xiangui Guo, China
Sigurdur F. Hafstein, Iceland
Zakia Hammouch, Morocco
Chittaranjan Hens, India
Giacomo Innocenti, Italy
Sarangapani Jagannathan, USA
Mahdi Jalili, Australia
Marco Javarone, United Kingdom
Peng Ji, China
Hang-Hyun Jo, Republic of Korea
Jeffrey H. Johnson, United Kingdom
Fariba Karimi, Austria
Mohammad Hassan Khooban, Denmark
Toshikazu Kuniya, Japan
Jurgen Kurths, Germany
C. H. Lai, Singapore
Guang Li, United Kingdom
Qingdu Li, China
Fredrik Liljeros, Sweden
May T. Lim, Philippines
Xinzhi Liu, Canada
Chongyang Liu, China
Xiaoping Liu, Canada
Joseph T. Lizier, Australia
Francesco Lo Iudice, Italy
Rosa M. Lopez Gutierrez, Mexico
Vittorio Loreto, Italy
Nishant Malik, USA

Rosario Nunzio Mantegna, Italy
Inés P. Mariño, Spain
Eulalia Martínez, Spain
André C. R. Martins, Brazil
Rossana Mastrandrea, Italy
Naoki Masuda, USA
Jose F. Mendes, Portugal
Ronaldo Parente De Menezes, United Kingdom
Fanlin Meng, United Kingdom
Marcelo Messias, Brazil
Ana Meštrović, Croatia
Ali Minai, USA
saleh mobayen, Taiwan, R.O.C., Iran
Osnat (Ossi) Mokryn, Israel
Christopher P. Monterola, Philippines
Marcin Mrugalski, Poland
Jesus Manuel Muñoz-Pacheco, Mexico
Roberto Natella, Italy
Chrystopher L. Nehaniv, Canada
Sing Kiong Nguang, New Zealand
Vincenzo Nicosia, United Kingdom
Irene Otero-Muras, Spain
Andreas Pape, USA
María Pereda, Spain
Nicola Perra, United Kingdom
Giovanni Petri, Italy
Cornelio Posadas-Castillo, Mexico
Mahardhika Pratama, Singapore
Karthikeyan Rajagopal, India
Gabriel Ramos-Fernández, Mexico
Andrea Rapisarda, Italy
Andrea Roli, Italy
Céline Rozenblat, Switzerland
Daniele Salvati, Italy
M. San Miguel, Spain
Matilde Santos, Spain
Fabio Saracco, Italy
Antonio Scala, Italy
Samuel V. Scarpino, USA
Enzo Pasquale Scilingo, Italy
Dan Selișteanu, Romania
Saray Shai, USA
Wen-Long Shang, China
Roberta Sinatra, Italy
Dimitrios Stamovlasis, Greece
Samir Suweis, Italy

Misako Takayasu, Japan
Ana Teixeira de Melo, Portugal
Vito Trianni, Italy
Shahadat Uddin, Australia
Jose C. Valverde, Spain
Sander E. Van Der Leeuw, USA
Alejandro F. Villaverde, Spain
Dimitri Volchenkov, USA
Christos Volos, Greece
Qingling Wang, China
Wenqin Wang, China
Zidong Wang, United Kingdom
Yong Xu, China
Honglei Xu, Australia
Xiao-An Yan, China
Xinggang Yan, United Kingdom
Zhile Yang, China
Baris Yuce, United Kingdom
Massimiliano Zanin, Spain
Hassan Zargazadeh, USA
Hector Zenil, United Kingdom
Xianming Zhang, Australia
Zhen Zhang, China
Fengtai Zhang, China
Rongqing Zhang, China
Xiaopeng Zhao, USA
Tao Zhou, China
Asim Zia, USA

Contents

Collective Behavior Analysis and Graph Mining in Social Networks 2021

Fei Xiong , Shirui Pan, and Xuzhen Zhu 

Editorial (2 pages), Article ID 9873569, Volume 2022 (2022)

Extraction of Psychological Effects of COVID-19 Pandemic through Topic-Level Sentiment Dynamics

Abdul Razzaq , Touqeer Abbas , Sarfraz Hashim , Salman Qadri , Imran Mumtaz , Najia Saher, Muzammil Ul-Rehman, Faisal Shahzad, and Syed Ali Nawaz 

Research Article (10 pages), Article ID 9914224, Volume 2022 (2022)

A Study of the Influence of Collaboration Networks and Knowledge Networks on the Citations of Papers in Sports Industry in China

Yu Zhang , Jianlan Ding , Hui Yan , Miao He , and Wei Wang 


Research Article (10 pages), Article ID 9236743, Volume 2022 (2022)

Evolutionary Game of Social Network for Emergency Mobilization (SNEM) of Magnitude Emergencies: Evidence from China

Rui Nan , Jingjie Wang , and Wenjun Zhu 




Research Article (13 pages), Article ID 3885934, Volume 2022 (2022)

Social Network Structure as a Moderator of the Relationship between Psychological Capital and Job Satisfaction: Evidence from China

Fan Gu  and Yuanyuan Xiao

Research Article (12 pages), Article ID 2550944, Volume 2021 (2021)

Research on the Structural Characteristics of Entertainment Industrial Correlation in China: Based on Dual Perspective of Input-Output and Network Analysis

Yang Xun , Wensheng Shi , and Tianyu Liu 

Research Article (11 pages), Article ID 6426123, Volume 2021 (2021)

A Resilience-Based Security Assessment Approach for CBTC Systems

Ruiming Lu , Huiyu Dong , Hongwei Wang , Dongliang Cui , Li Zhu , and Xi Wang 

Research Article (10 pages), Article ID 2175780, Volume 2021 (2021)

Research on the Relationship between Social Support and Employment Quality of Chinese Athletes from the Perspective of Social Network Structure

Meijuan Cao, Shuairan Li , Wenfei Yue, and Huanqing Wang 



Research Article (9 pages), Article ID 9916024, Volume 2021 (2021)

DWNet: Dual-Window Deep Neural Network for Time Series Prediction

Jin Fan, Yipan Huang , Ke Zhang, Sen Wang, Jinhua Chen, and Baiping Chen 

Research Article (10 pages), Article ID 1125630, Volume 2021 (2021)

Analysis on Quantified Self-Behavior of Customers in Food Consumption under the Perspective of Social Networks

Lei Lei , Yaling Zhu, and Qiang Liu 



Research Article (14 pages), Article ID 6001654, Volume 2021 (2021)

The Influence of Author Degree Centrality and L-Index on Scientific Performance of Physical Education and Training Papers in China Based on the Perspective of Social Network Analysis

Bin Zhang , Jian Wu , Qian Huang , Yujiao Tan , Lu Zhang , Qian Zheng , Yu Zhang , Miao He , and Wei Wang 



Research Article (14 pages), Article ID 3066602, Volume 2021 (2021)

The Influence of Individual Characteristics on Cultural Consumption from the Perspective of Complex Social Network

Hui Liu, Shuang Lu, Ximeng Wang , and Shaobo Long 

Research Article (14 pages), Article ID 7404690, Volume 2021 (2021)

Firms' Investment Behaviours in Temperature-Controlled Supply Chain Networks

Mengshuai Zhu, Hao Chen, Ximeng Wang, Yonghan Wang, Chen Shen , and Cong Zhu 






Research Article (11 pages), Article ID 5359819, Volume 2021 (2021)

EMM-CLODS: An Effective Microcluster and Minimal Pruning CLustering-Based Technique for Detecting Outliers in Data Streams

Mohamed Jaward Bah , Hongzhi Wang , Li-Hui Zhao , Ji Zhang , and Jie Xiao

Research Article (20 pages), Article ID 9178461, Volume 2021 (2021)

Layer Information Similarity Concerned Network Embedding

Ruili Lu , Pengfei Jiao , Yinghui Wang , Huaming Wu , and Xue Chen 



Research Article (10 pages), Article ID 2260488, Volume 2021 (2021)

The Comprehensive Contributions of Endpoint Degree and Coreness in Link Prediction

Yang Tian , Yanan Wang , Hui Tian , and Qimei Cui

Research Article (9 pages), Article ID 1544912, Volume 2021 (2021)

Link Prediction Based on the Derivation of Mapping Entropy

Hefei Hu , Yanan Wang, Zheng Li, Yang Tian , and Yuemei Ren

Research Article (7 pages), Article ID 4156832, Volume 2021 (2021)

Dual-Channel Reasoning Model for Complex Question Answering

Xing Cao, Yun Liu , Bo Hu, and Yu Zhang

Research Article (13 pages), Article ID 7367181, Volume 2021 (2021)

Privacy-Preserving Efficient Data Retrieval in IoMT Based on Low-Cost Fog Computing

Na Wang , Yuanyuan Cai , Junsong Fu, and Jie Xu

Research Article (13 pages), Article ID 6211475, Volume 2021 (2021)





The Sustainability of Knowledge-Sharing Behavior Based on the Theory of Planned Behavior in Q&A Social Network Community

Xin Feng , Lijie Wang , Yue Yan , Qi Zhang , Liming Sun , Jiangfei Chen , and Ye Wu 

Research Article (12 pages), Article ID 1526199, Volume 2021 (2021)

Contents

Topic Detection and Tracking Techniques on Twitter: A Systematic Review

Meysam Asgari-Chenaghlu , Mohammad-Reza Feizi-Derakhshi , Leili Farzinvasht , Mohammad-Ali Balafar , and Cina Motamed



Review Article (15 pages), Article ID 8833084, Volume 2021 (2021)

Credit Behaviors of Rural Households in the Perspective of Complex Social Networks

Qiang Zhao, Yue Shen , and Chaoqian Li








Research Article (13 pages), Article ID 9975856, Volume 2021 (2021)

Efficient Data Transmission for Community Detection Algorithm Based on Node Similarity in Opportunistic Social Networks

Aizimaiti Xiaokaiti, Yurong Qian , and Jia Wu 



Research Article (18 pages), Article ID 9928771, Volume 2021 (2021)

The Hotspots of Sports Science and the Effects of Knowledge Network on Scientific Performance Based on Bibliometrics and Social Network Analysis

Linxiao Ma , Yuzhu Wang , Yue Wang , Ning Li , Sai-Fu Fung , Lu Zhang , and Qian Zheng 

Research Article (12 pages), Article ID 9981202, Volume 2021 (2021)

Power Control Algorithm Based on a Cooperative Game in User-Centric Unmanned Aerial Vehicle Group

Yuexia Zhang  and Pengfei Zhang 

Research Article (6 pages), Article ID 7108198, Volume 2021 (2021)

Editorial

Collective Behavior Analysis and Graph Mining in Social Networks 2021

Fei Xiong ¹, Shirui Pan,² and Xuzhen Zhu ³

¹*School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China*

²*Faculty of Information Technology, Monash University, Clayton VIC 3800, Australia*

³*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Correspondence should be addressed to Fei Xiong; xiongf@bjtu.edu.cn

Received 8 March 2022; Accepted 8 March 2022; Published 16 April 2022

Copyright © 2022 Fei Xiong 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 the development of Internet technology, social networks have become a popular communication place in which people can interact with others conveniently and freely. Many sorts of social networks have emerged, such as online social networks, employment relationship networks, and scientific cooperation networks. Users in social networks may create new connections with other users, so they can interact through those links, or they may break existing connections. Therefore, the structure of social networks is evolving every day [1,2]. Users can contact their neighbors, send information to them, and exchange their opinions, leading to dynamics on social networks, including opinion formation [3], spreading dynamics [4], and collaborative behaviors [5]. Individual behaviors accumulate through local interactions and may result in complex collective phenomena, demonstrating the significant role of social networks in driving society development. Analyzing complex human behaviors and mining graph topology can help understand the essential mechanism of macroscopic phenomena [6]. These studies would help attract public interest and excavate useful resources. Therefore, social network mining has become a promising research area and attracts lots of attention.

Studies on social networks can be divided into two categories: theoretical modeling and data-driven methods. Theoretical methods characterize individual user behaviors and try to find the microscopic origin of collective phenomena by statistical physics, probability theory, and numerical stochastic processes [7]. Data-driven methods use machine learning, data mining, and natural language

processing to exploit hidden patterns from the data in social networks [8]. They are also used to estimate the future evolution of social behaviors. In recent years, with the growth of the massive amount of data, it becomes more difficult to mine social networks and analyze user behaviors. Therefore, advanced interdisciplinary data analysis and data mining methods should be suggested and developed to study social networks [9].

In this special issue, 23 papers have been published. All the papers must undergo a rigorous peer review process. The published papers can be divided into three categories, i.e., empirical social behavior analysis, social network modeling, and network data mining.

In the category of empirical social behavior analysis, the article by Liu et al. [10], “The Influence of Individual Characteristics on Cultural Consumption from the Perspective of Complex Social Network,” studied the social network effect on cultural consumption. The article by Gu and Xiao [11], “Social Network Structure as a Moderator of the Relationship between Psychological Capital and Job Satisfaction: Evidence from China,” primarily examined the role of social network structure in the relationship between psychological capital and employment satisfaction by adopting a two-wave data from undergraduate students. In the article titled “The Sustainability of Knowledge-Sharing Behavior Based on the Theory of Planned Behavior in Q&A Social Network Community,” by Feng et al. [12], the authors investigated the important factors which drive users to share knowledge and they verified their model by questionnaire data.

In the category of social network modeling, the article by Xiaokaiti et al. [13], “Efficient Data Transmission for Community Detection Algorithm Based on Node Similarity in Opportunistic Social Networks,” studied the problem of community detection and presented a new data transmission method which considers social attributes. In the article titled “Layer Information Similarity Concerned Network Embedding,” by Lu et al. [14], the authors explored the method of network embedding and they introduced the layer information similarity to enhance the representation of nodes.

In the category of network data mining, the article by Asgari-Chenaghlu et al. [15], “Topic Detection and Tracking Techniques on Twitter: A Systematic Review,” overviewed the technology of detecting popular topics on Twitter and analyzed some common problems in this area. The article “DWNNet: Dual-Window Deep Neural Network for Time Series Prediction,” by Fan et al. [16], exploited multi-granularity dependencies of time series and constructed a dual-window deep neural network to predict future data.

Conflicts of Interest

The editors declare that they have no conflicts of interest.

Acknowledgments

We would like to thank all the authors and reviewers in our special issue. We also appreciate the help of the Editor-in-Chief and staff members. This special issue acknowledges the help of guest editors.







Sen Wang
Ziming Zhang

References

- [1] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, “Popularity versus similarity in growing networks,” *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.
- [2] M. Szella, R. Lambiotte, and S. Thurner, “Multirelational organization of large-scale social networks in an online world,” *Proceedings of The National Academy of Sciences of The United States of America*, vol. 107, no. 31, pp. 13636–13641, 2010.
- [3] F. Xiong and Y. Liu, “Opinion formation on social media: an empirical approach,” *Chaos*, vol. 24, Article ID 013130, 2014.
- [4] F. Xiong, Y. Zheng, W. Ding, H. Wang, X. Wang, and H. Chen, “Selection strategy in graph-based spreading dynamics with limited capacity,” *Future Generation Computer Systems*, vol. 114, pp. 307–317, 2021.
- [5] D. Centola, “The spread of behavior in an online social network experiment,” *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.
- [6] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, “The physics of spreading processes in multilayer networks,” *Nature Physics*, vol. 12, no. 10, pp. 901–906, 2016.
- [7] N. Crokidakis and C. Anteneodo, “Role of conviction in nonequilibrium models of opinion formation,” *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 86, no. 6, Article ID 061127, 2012.
- [8] Q. Bao, W. K. Cheung, Y. Zhang, and J. Liu, “A component-based diffusion model with structural diversity for social networks,” *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 1078–1089, 2017.
- [9] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, “Modeling information diffusion over social networks for temporal dynamic prediction,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1–1997, 2017.
- [10] H. Liu, S. Lu, X. Wang, and S. Long, “The influence of individual characteristics on cultural consumption from the perspective of complex social network,” *Complexity*, vol. 2021, Article ID 7404690, 14 pages, 2021.
- [11] F. Gu and Y. Xiao, “Social network structure as a moderator of the relationship between psychological capital and job satisfaction: evidence from China,” *Complexity*, vol. 2021, Article ID 2550944, 12 pages, 2021.
- [12] X. Feng, L. Wang, Y. Yan et al., “The sustainability of knowledge-sharing behavior based on the theory of planned behavior in q&a social network community,” *Complexity*, vol. 2021, Article ID 1526199, 12 pages, 2021.
- [13] A. Xiaokaiti, Y. Qian, and J. Wu, “Efficient data transmission for community detection algorithm based on node similarity in opportunistic social networks,” *Complexity*, vol. 2021, Article ID 9928771, 18 pages, 2021.
- [14] R. Lu, P. Jiao, Y. Wang, H. Wu, and X. Chen, “Layer information similarity concerned network embedding,” *Complexity*, vol. 2021, pp. 1–10, Article ID 2260488, 2021.
- [15] M. Asgari-Chenaghlu, M. Feizi-Derakhshi, L. Farzinvas, M. Balafar, and C. Motamed, “Topic detection and tracking techniques on Twitter: a systematic review,” *Complexity*, vol. 2021, Article ID 8833084, 15 pages, 2021.
- [16] J. Fan, Y. Huang, K. Zhang, S. Wang, J. Chen, and B. Chen, “DWNNet: dual-window deep neural network for time series prediction,” *Complexity*, vol. 2021, Article ID 1125630, 10 pages, 2021.

Research Article

Extraction of Psychological Effects of COVID-19 Pandemic through Topic-Level Sentiment Dynamics

Abdul Razzaq ¹, Touqeer Abbas ¹, Sarfraz Hashim ², Salman Qadri ¹,
Imran Mumtaz ³, Najia Saheer,⁴ Muzammil Ul-Rehman,⁵ Faisal Shahzad,⁴
and Syed Ali Nawaz ⁴

¹Department of Computer Science, MNS University of Agriculture, Multan, Pakistan

²Department of Agricultural Engineering, MNS University of Agriculture, Multan, Pakistan

³Department of Computer Science, University of Agriculture, Faisalabad, Pakistan

⁴Department of Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

⁵Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

Correspondence should be addressed to Abdul Razzaq; abdul.razzaq@mnsuam.edu.pk and Sarfraz Hashim; sarfraz.hashim@mnsuam.edu.pk

Received 27 March 2021; Revised 19 April 2021; Accepted 31 January 2022; Published 18 March 2022

Academic Editor: Hassan Zargarzadeh

Copyright © 2022 Abdul Razzaq 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 rapid increase in COVID-19 cases has become the symbol of fear, anxiety, and panic among people around the globe. Mass media has played an active role in community education by addressing the health information of this pandemic. People interact by sharing their ideas and feelings through social media platforms. There is a considerable need to implement different measures and better perceive COVID-19 pertinent facts and information by demystifying public sentiments. In this study, the Quarantine Life dataset of thousand tweets is based on #Quarantine, #Quarantine Days, #Quarantine Life, #My Pandemic Plan, and #Quarantine and Chill from January to September 2020 has been collected from Twitter. The extracted data have been scrubbed through preprocessing techniques. The sentiments and topics extracted from tweets have been analyzed through the TEXT BLOB, VADER, and AFFIN approach. Results show that people were distressed and fearful due to the COVID-19 pandemic. However, most people enjoyed by playing games, watching movies, and reading books during the lockdown period. According to the present meta-analysis, physical activity interventions are beneficial for patients with dementia in terms of cognition. The proposed framework illustrates the insight impact of COVID-19 on human physiological and mainly focuses on the evaluation of sentiment dynamics at the topical level.

1. Introduction

COVID-19 is more than just an infectious disease spread by droplets emitted when people cough, sneeze, or talk; misinformation spread promoted on social media has made it a source of stress, depression, and anxiety. Fake information spreads quickly on social media, which negatively impacts mental health. During this period of social distancing and lockdown, individuals rely primarily on the Internet, and the most activity is reported on social media. Opinion mining and sentiment analysis are emerging Natural Language Processing application whose importance is becoming

progressively higher. It analyzes a textual dataset for people's opinions, assessments, sentiments, attitudes, and emotions using opinions mining and sentiment analysis. It is possible to determine people's sentiments by applying sentiment analysis to almost every social media platform, such as Twitter, Facebook, YouTube, and Tumblr.

It has been proven that comprehending the emotions expressed through certain resources, such as tweets, blogs, reports, documents, or segments from political speeches, is significant for humans [1]. However, enormous, large number of opinions are challenging task for human processing. The extraction of sentiments from multiple sources

that keep growing in volume, complexity, and diversity requires automated processes. Online social networks (OSNs) provide a medium (where different people can engage, demonstrate, and express their ideas). Microblogging is a fast way for communication such as Twitter, Tumblr, and Facebook are most popular microblogging platforms, in which millions of messages appear. By analyzing the content over social media, a notable transformation in the shaping of public perceptions has been made in a society that revealed dominant users. [2]. Recently, data are available on social media platform related to COVID-19 pandemic need to process and extract meaningful information to create awareness among people. Due to this pandemic situation, government has imposed lock down from January 2020 to September 2020 to save people's life. The period of lock down was very tough for people as they are bounded in their homes and used social media apps for exchange information and awareness. Twitter has become a popular social media platform, which people share their thoughts, views, audio, videos, and comments on various topics and ideas. Mostly tweets were viral on COVID-19 with hashtag symbol [3]. The hashtag is a set of keywords that are helpful to find useful information. Hashtag (#) symbol that indicates posted information, comment, and ideas, are important as 50 million information are organized with keyword hashtag on Twitter [4]. People from all over the world are affected badly due to this pandemic. In present study we target "#sentences" on COVID-19 pandemic from people all over the world in Twitter. Six important keywords or trends related to COVID-19 pandemic are targeted as follows:

- (1) #Quarantine
- (2) #Covid-19
- (3) #Quarantine Days
- (4) #Quarantine Life
- (5) #My Pandemic Plan
- (6) #Quarantine and Chill

Data related to these trends for the duration January–September 2020 was collected to find people's everyday life and their daily routines using sentiment analysis and topic modeling. In this proposed study, the following objectives were set as follows:

- (1) Collection of data from Twitter on COVID-19 pandemic
- (2) Analyzing people views using polarity tendency
- (3) Comparison between algorithms for better visualization of results
- (4) Which trend is mainly focused by people on social media users using LDA models?

2. Related Work

With the tremendous increase in COVID-19 all over the world, researchers applied sentiment analysis methods focused on social media to observe people's mental well-being.

This section contains the summary of work related to COVID-19 research based on social media data. A. Jenifer describes hybrid approaches to sentiment analysis based on Twitter data. A sentiment lexicon was created and enhanced by Senti-WordNet, along with semantic rules, unsupervised Machine Learning methods, and fuzzy sets [5]. TA hybrid standard classification was first carried out and was then upgraded to a hybrid advanced classification [6]. They built-in the hybrid advanced approach into the linguistic semantic polarity classification that was modeled using fuzzy sets. The new sentiment analysis methodology was used to compute the polarity of a given sentence for the movie review dataset.

Suresh et al. [5] described a fuzzy clustering model using real tweets collected over a one-year period for the purpose of analyzing the sentiments associated with a particular brand name. They conducted a comparative study with *K*-means clustering algorithm, expectation-maximization techniques, and accuracy, precision, recall, and time complexity were used. According to experimental analysis, the proposed method was proved to be effective in performing high-quality sentiment analysis on twitter. As compared to the other two methods, this model gave an accuracy of 76.4 and required less time to build. Supriya et al. [7] presented a three-step algorithm presented for analyzing the public sentiment in Twitter tweets. The algorithm steps consisted of cleaning, entity identification, and classification for sentiment analysis. The performance of the classifier was measured using precision, recall, and accuracy. Elaziz et al [8] proposed a novel approach to visual diagnosis of COVID-19 through machine learning by classifying the chest X-ray images into two classes, positive COVID-19 patient or negative COVID-19 person. They used new fractional multichannel exponent moments (FrMEMs) for features extraction from the chest X-ray images. They utilized a framework to accelerate the computational process. After that, they used a modified MRFO (Manta-Ray Foraging Optimization) that was based on differential evolution to extract the most important features. They performed this methodology on two COVID-19 datasets and got the accuracy of 96.09% and 98.09% from these two datasets. Jain and Sinha [2] proposed weighted correlated influence (WCI) approach in order to integrate the relative impact of trend-specific and timeline-based features of twitter users. They used the Twitter trend #Coronavirus Pandemic to quantify their proposed approach performance. The proposed WCI showed better performance than the existing methods. A Sharma et al. [9] gave insights of the foremost issues the firms are facing due to COVID-19 and how they are examining the strategic options. They took data from twitter about NASDAQ 100 firms and used text analytics tools to find out the issues that firms are facing and the strategies they are adopting. They also recommended some futuristic strategies for innovation of the supply chain. Samuel et al. [10] provided insight into COVID-19 pandemic fear sentiment progression. They also outlined interrelated methods, implications, opportunities, and limitations.

Their analysis was based on Covid-19 linked Tweets and R statistical tools along with text mining packages. They also established evidence that growth of fear-sentiments existed

from the beginning of COVID-19, as the outbreak reached its peak in the US, by applying descriptive text analytics.

Furthermore, they provided a methodological overview of two fundamental machine learning classification approaches (Naïve Bayes and logistic regression), as applied to textual analytics, and compared their efficiency when it came to categorizing coronavirus tweets. Both Naïve Bayes and logistic regression classification methods provided an accuracy of 91% and 74%, respectively, with short length tweets but both approaches showed relatively lower accuracy with lengthy tweets. Li et al. [11] examined the impact of COVID-19 on mental health. They used the method of online ecological recognition [12] based on several machine learning predictive models to evaluate Weibo (a Twitter-like microblogging framework in China) articles. They used the collected data to calculate the word frequency, scores of emotional indicators (depression, anxiety, indignation, and Oxford happiness), as well as cognitive indicators (life satisfaction and risk judgment). They performed sentiment analysis and sample *t*-test to examine the differences before and after the affirmation of COVID-19.

The results showed that the scores of negative emotions were increased as compared to positive ones. Cinelli et al. [13] used different social media platforms (Twitter, Instagram, Reddit) to analyze awareness and concern in the subject of COVID-19 and provided a differential assessment of the global discourse evolution of each platform and their users. They found similar spreading patterns from reliable and suspicious information sources. Zhou et al. [14] analyzed the sentiment dynamics of people of New South Wales (NSW) Australia during the COVID-19 period by exploiting the tweets on Twitter. They analyzed the sentiments at local government areas (LGAs) level that was based on more than 94 million tweets collected from Twitter for a 5-month period started from 1st January 2020. The results showed that the positive sentiments were decreased due to massive increase in COVID-19 confirmed cases. Han et al. [15] proposed a topic extraction and classification model to analyze media data in the early stage of COVID-19 in China. They generalized COVID-19 related microblogs into 7 topics and 13 more detailed subtopics. However, their study had some limitations. They used social media to analyze text only, but pictures and videos could also be informative.

3. Proposed Methodology

This research is divided into a series of steps as shown in Figure 1. The first step is to collect the data through Twitter API, after collecting the significant number of tweets, all these tweets are stored in a text file. In the second step, the classification accuracy was improved by performing some preprocessing techniques such as case folding, cleansing, word formalization, and stemming. These processes are conducted for Lexicon-based machine learning approaches. In the Lexicon-based approaches, Text Blob, Vader Sentiment, and AFINN have been used to determine the polarity of each Twitter user. Topic modeling has been used to find the useful information from the group of tweets. Furthermore, we adopt the *t*-distributed stochastic neighbor embedding

(t-SNE) technique that partly reduces the fact that humans cannot perceive vector spaces beyond 3D.

3.1. Data Set. Real time data were collected from Twitter using the scripting language Python for getting off data from Twitter. The data were collected from 1st January 2020 to 19th august 2020. For the collection and distribution of the datasets, Twitter API (Tweepy) was used.

API collects data real-time data of twitter on geographical regions of all countries illustrated in Figure 2. There should be a valid twitter account, and the application should be registered on Twitter to extract the tweets. The user sends the request to API for the twitter data, and it returns data according to the user-defined query. A sample of 16696 tweets was extracted. In this work, query was “Quarantine Life”, all data were extracted belonging to this keyword. The extracted data included Tweet ids, names, screen names, locations, descriptions, followers, and following counts of users.

3.2. Preprocessing. Text mining needs some primary section which in essence is preparing for the document to be converted to make it more structured. In this analysis, preprocessing steps are as follows:

- (1) *Case folding* is the first step that replaces whole text into lowercase, i.e., “Alan, self-isolation, Day 4” into “Alan, self-isolation, day 4.”
- (2) *Cleansing* is required to derive the figures used in this analysis. From this stage, we exclude grammar, symbols, abbreviations, specifying the client, and tweets. The only characters left in this phase are words.
- (3) *Formalization* on twitter results is limited to 160 letters only. Therefore, Twitter users tend to write unorthodox sentences. To resolve this, a formalization application is required in order to embed the word in its default form.
- (4) *Stemming* is the step process word using a tool that converts words holding a document into their fundamental forms using fixed rules.

3.3. Lexicon-Based Approaches. Machine learning algorithms were applied to check polarity in text. We used three algorithms AFINN, VADER (Valence Aware Dictionary for Sentiment Reasoning), and TEXTBLOB to check polarity in text and for semantics analysis.

3.3.1. VADER. VADER is a rule-based lexicon and analysis tool that is particularly used for sentiment analysis. It is used to extract sentiments being expressed in social media, and it performs exceptionally very well in this domain. VADER sentiment analysis [16] is primarily based on definite key factors such as punctuation, capitalization, conjunctions, degree modifiers, and preceding trigram. VADER classifies the sentiments into positive, neutral, and negative categories and secure complex scores which is determined by summing up

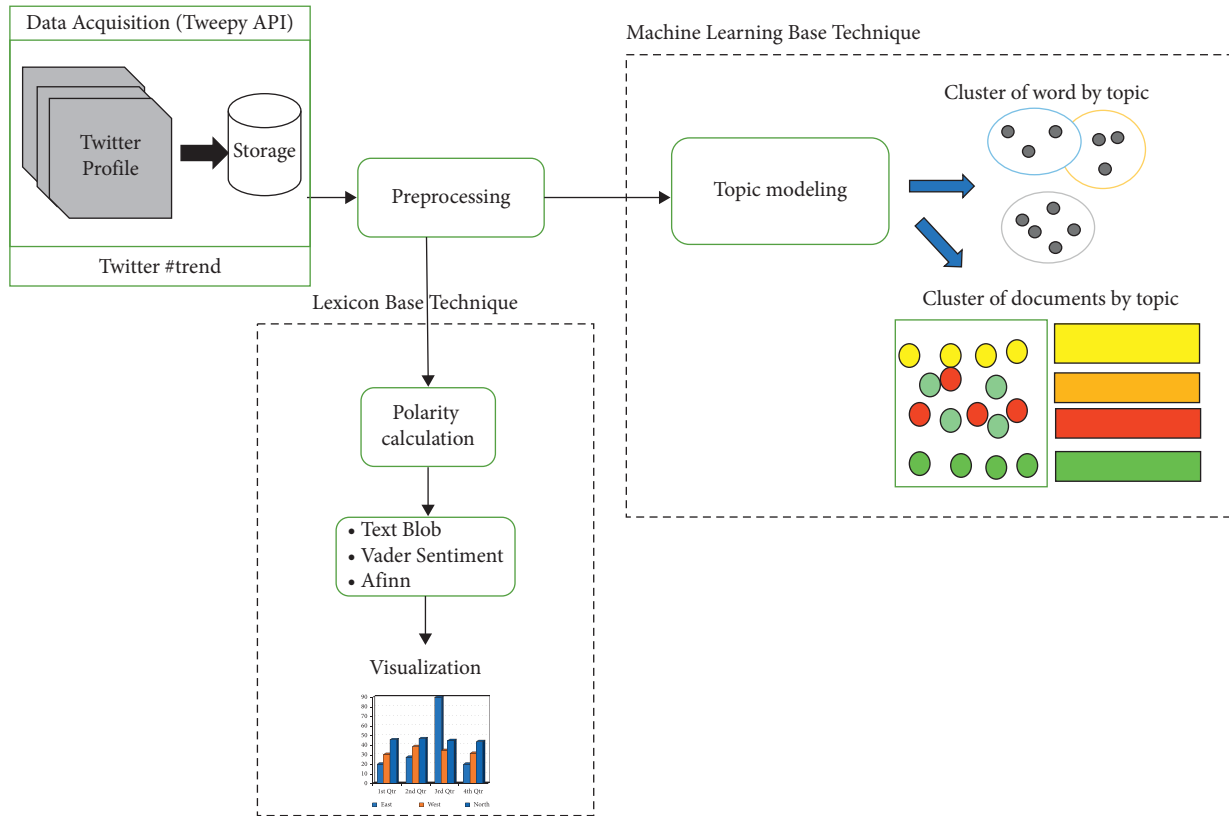


FIGURE 1: Proposed framework.

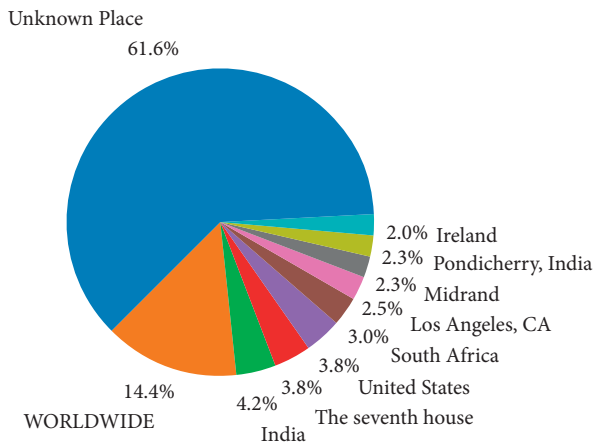


FIGURE 2: Percentage of tweets according to their location.

each word’s valence scores in the lexicon and normalized in the range $(-1, 1)$, the most extreme positive is “1,” and the most extreme negative is “-1.” If the compound score is less than -0.05 , the text will be considered negative; if the score is greater than 0.05 , the text will be considered positive; if the score is between 0.05 and -0.05 , the polarity of the text will be neutral. VADER has one major advantage that it does not entail the preprocessing of data and training of model can be utilized directly on the raw tweets to generate sentiment polarity. It also

supports emoji for sentiment classification and is fast enough to be used online without affecting speed-performance.

3.3.2. *AFINN*. Afinn is an English word list with an integer between 5 and -5 which has been significantly designed for microblogs such as tweets. It has the biggest advantage that it gets updated with new terms and phrases every year.

3.3.3. *TEXTBLOB*. Text blob is a Python library (just like a python string) which is used for processing the textual data. It aims in providing a consistent API to deal with common NLP (natural language processing) tasks such as part-of-speech tagging, noun phrase extraction, translation, text mining, text processing modules, text analysis, sentiment analysis, classification, and more. Text blob analyzes the text on sentence level [16]. Firstly, it takes input from the dataset, and then it splits the review into sentences. The polarity of the entire dataset can be determined by counting the number of negative and positive sentences and deciding whether the response is positive or negative based on the total number of negative and positive reviews. A *sentiment ()* function can be used to find the polarity and subjectivity of a given review. It returns a tuple with two parameters called polarity and subjectivity. The function returns a tuple consisting of polarity and subjectivity, where the polarity score ranges

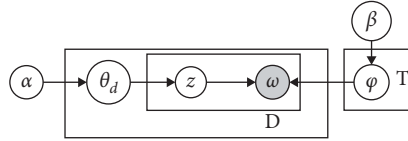
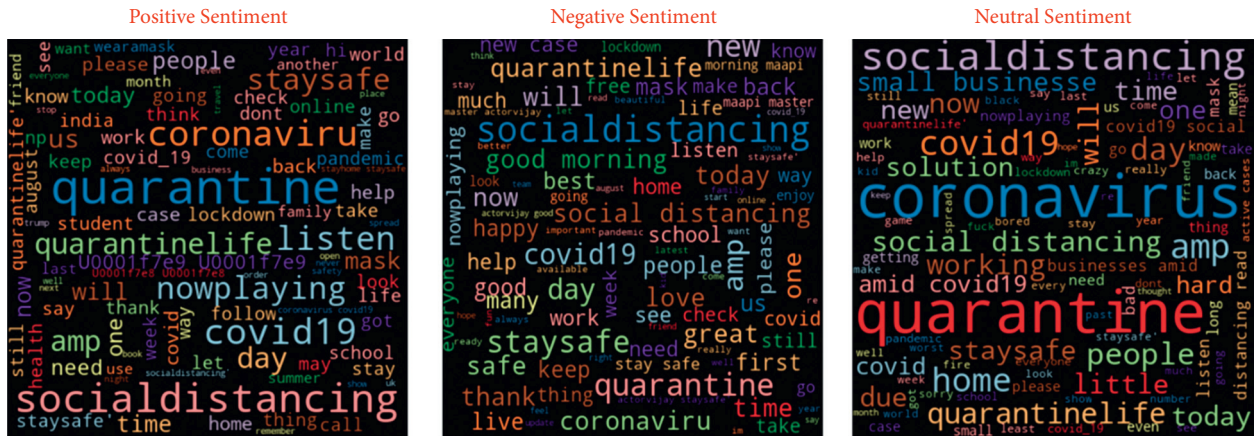


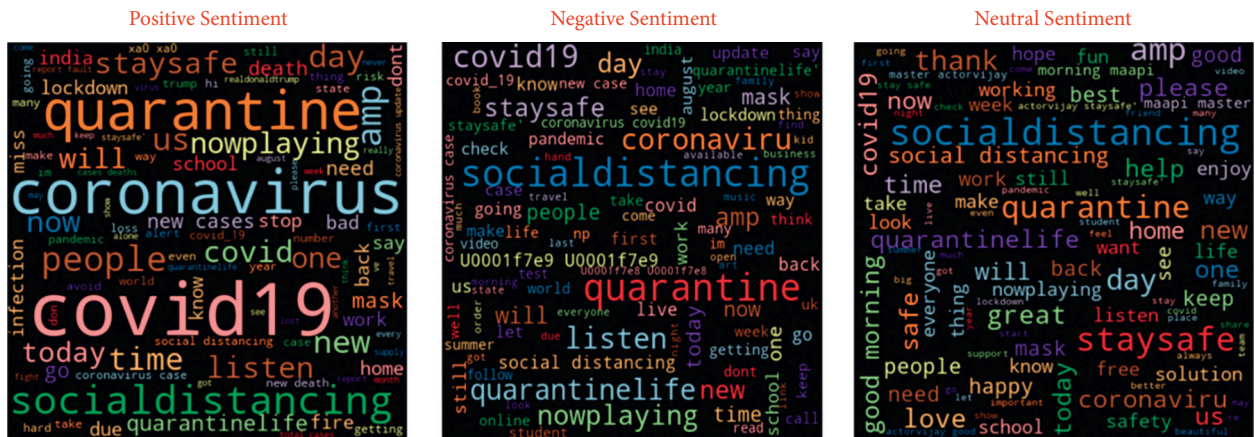
FIGURE 3: LDA modeling.



(a)



(b)



(c)

FIGURE 4: (a) Example of common words in positive, negative, and neutral sentiments have been represented by these three Word Clouds in Vader Semantics. (b) Example of common words in positive, negative and neutral sentiments has been represented by these three Word Clouds text blob. (c) Example of common words in positive, negative and neutral sentiments has been represented by these three Word Clouds in Affinn.

TABLE 1: Polarity calculation with sentiment analyzer percentage accuracy.

Tweets	TexBlob	Vader	Afinn
Is it just me or is everyone feeling like a dead body at home now lockdown quarantine	Negative	Negative	Negative
Full list of “safe” holiday countries right now amid fear more destinations will require quarantine. . .	Positive	Negative	Negative
Night six of my quarantine could not sleep a single minute	Negative	Negative	Negative
I love my cat what would i do without her ☐ quarantinelife queen quarantine roar rawr umbreon. . .	Positive	Positive	Positive
That feeling I have all the time I care about everyone’s health everyone is acting like things are gonna be okay	Neutral	Positive	Neutral
Now I understand my mom all this cooking and cleaning is crazy and I do not even have children quarantine quarantinelife	Negative	Negative	Negative
Watch quarantine hbo robertdeniro michellepfeiffer excellent timely film available in canada on crave tv	Positive	Positive	Positive
The pandemic has made seniors crazy quarantine cannot stop their adventurous souls from wandering beware of. . .	Negative	Negative	Negative
Lessons learned in quarantine	Neutral	Neutral	Neutral
Fearing coronavirus, a Michigan college tracks its students with a flawed app–techcrunch	Negative	Negative	Negative
The university of North Carolina is retreating to remote learning for undergrads because of coronavirus outbreaks. . .	Negative	Negative	Negative
Stress less and enjoy the best stay home insta pic of the day followme COVID-19 quarantine art coronavirse	Positive	Positive	Positive
From being in quarantine for 202 days has me so board	Negative	Negative	Negative
Homehaircut quarantine this is the ghetto	Neutral	Negative	Neutral
Distance learning at its finest melody first day of school 5th grade 2020 quarantine distance learning	Neutral	Neutral	Neutral

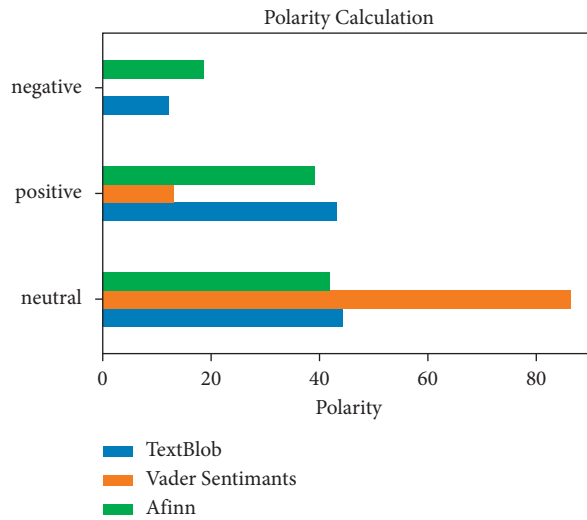


FIGURE 5: Semantics calculation.

TABLE 2: Tweets with target class.

Sentiment analyzer	Positive	Neutral	Negative
Text blob	7232 (43.31%)	7417 (44.42%)	2047 (12.26%)
Vader sentiment	6958 (41.67%)	398 (2.38%)	9340 (55.94%)
Afinn	6556 (39.26%)	7026 (42.08%)	3114 (18.65%)

TABLE 3: Description of top four dominant topics.

Topic no.	Topic precision contribution	Keywords	Representative tweet
0.0	0.8920	New, case, follow, death, confirm, safe, month, tell, recover, today	Beautiful, day, go, outside, keep, calm
1.0	0.9219	Walk, still, soon, reach, read, free, wait, back, pay, big	Way, stay, touch, find, way, stay, positive, challenging, time
2.0	0.9300	Walk, still, soon, reach, read, free, wait, back, pay, big	Update, new, case, confirm, new, recover, new, death, total, case
3.0	0.8725	Listen, amp, people, live, love, watch, order, keep, mask, play	State, add, visitor, state, must, quarantine, read

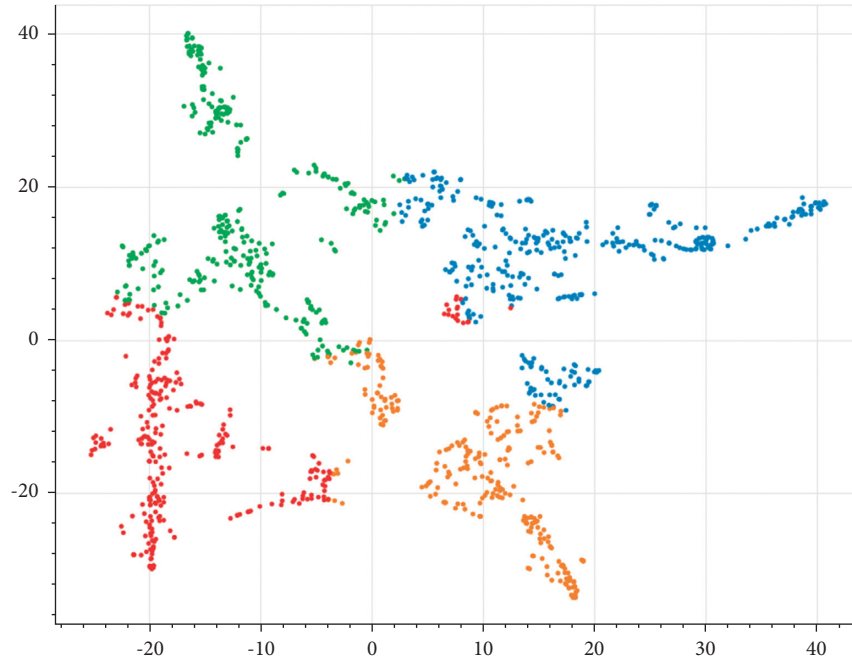


FIGURE 6: t-SNE cluster plot.

from -1 to 1 . The subjectivity range is 0 to 1 , where 1 is the most subjective and 0 is the most objective.

3.4. Topic Modeling. Topic Modeling [17] consists of finding the information contained in textual documents and presenting it in the form of themes (depending on the technique used, the relative importance of the themes can also be found). Topic modeling is therefore an unsupervised technique for classifying documents in multiple themes. From the perspective of the representation space, the TM is a reduction of dimensions in the vector representation. Instead of representing a document of a corpus by a vector in the space of the words, composing the

vocabulary of this corpus is represented by a vector. In the space of the themes of this corpus, each value of this vector corresponds to the relative importance of the theme in this document. Popular modern technique Latent Dirichlet allocation (LDA) was used in this study [18], and LDA is used for topic recognition in documents. It basically tells how many topics exist on similarity bases in each document. This model observes all words and produces topic distribution with P (P =probability) as shown in Figure 3. Researchers prefer LDA method for finding topics within context-based documents or text-based data [19, 20].

Mathematical representation of LDA:

$$P(\beta, \theta, z, w) = \left(\prod_{i=1}^k p(\beta_i | n) \right) \left(\prod_{d=1}^k P(\theta_d | \alpha) \prod_{N=1}^k P(Z_{d,n} | \theta_d) P(W_{d,n} | \beta_1 : k, Z_{d,n}) \right), \quad (1)$$

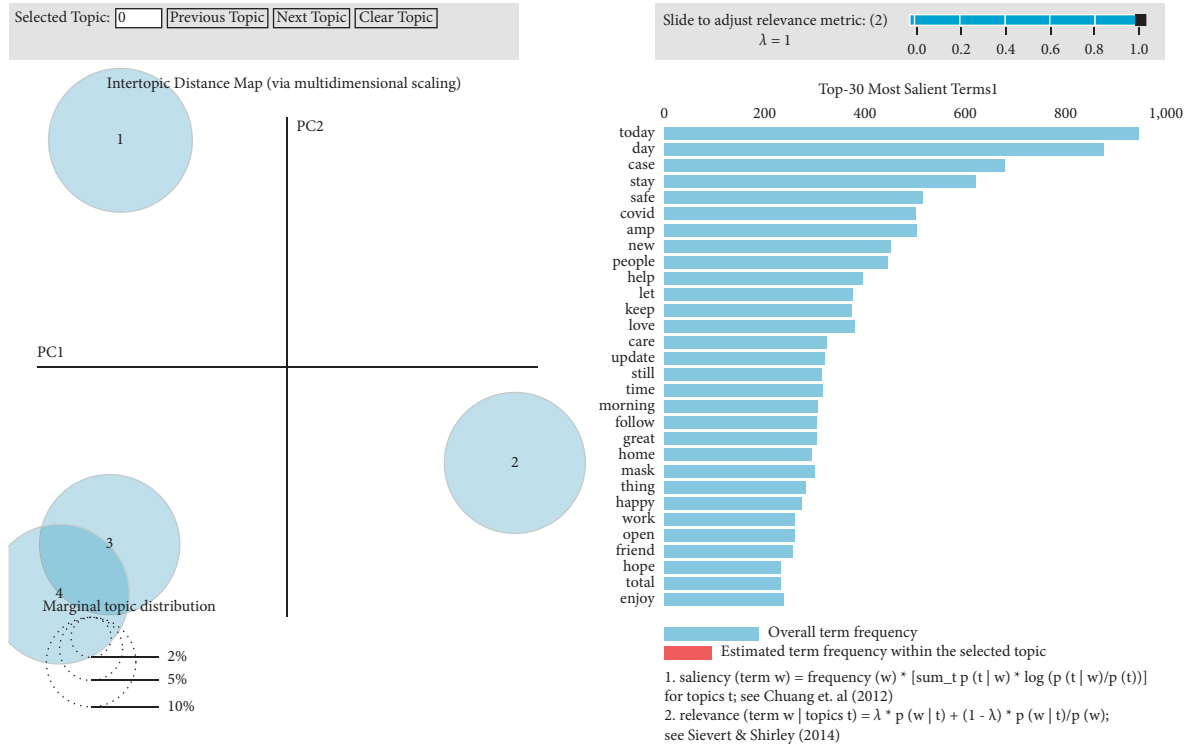


FIGURE 7: Intertopic distance map and Top 30 most relevant terms for topic 1 (25.1% of tokens).

where βK is the word distribution of the topic K , θ is the topic proportions of the document, and z is the topic assignment of the word in a document.

4. Results and Discussion

4.1. Polarity Calculation. A total of 16696 tweets were collected from twitter API. The collected records do not have a target group of people. To make the target group view, three lexicon algorithms has been used. Every analyzer describes whether the tweet is a positive, neutral, or negative. The claim of the authors that social media has been incapable to present the proper direction in which the netizens shall combat a pandemic like COVID-19, has been confirmed by the Word Cloud as shown in Figures 4(a)–4(c). Most of the words that have been described in each of the sentiments have been visualized using the WordCloud modules. These also present words that do not verify any efficiency in representing a possible solution during crises. Among the three sentiment analyzers we found that text blob had the highest rate of tweets with the neutral sentiment 44.42%. Vader sentiment gave the highest negative sentiment rate of 86.46%. However, AFINN gave the highest sentiment rate 42.08% as shown in Table 1 and Figure 5. Table 2 shows some random tweets with target class: positive, negative, and neutral. We perform experiments through Affin, Veder, and text blob, by taking random tweets, and then the algorithm analyzes which tweets contain positive, negative, and neutral sentiments.

4.2. Topic Modeling. After finding the sentiment from the data, the next step is to identify the topic. Topic modeling is the best way to discover how many abstract topics exist in the corpus. A document contains multiple topics in LDA models. The dominant topic is usually one of the topics. Table 3 shows extracted dominant topic for each sentence. Every keyword in LDA topic modeling contains weights and these weights show how much a specific keyword is important in topics. However, word counts represent the frequency of repeated words in specific topic.

Figure 5 shows the weight of the topic and the keywords. Our goal is to find words that are found across multiple topics and whose relative frequency outweighs their weight. In many cases, such words are not as important as they seem. From Figure 6, by using t-SNE plot high dimensional data into a lower dimension which is difficult for humans to understand such as word embedding. It categories words according to four topics and overlaid the word-level sentiment by color.

From Figures 7 and 8, we need to select four topics to analyze using Python 3.6.1 and LDAvis [16]. We set $\lambda = 1$ and set 4 topics and their keywords. Topics' names were generated according to their similar keywords to expatiate the topics. Bubbles are represented the topics and the size of the bubble is proportional to its prevalence in the corpus. Similar topics take shape close to each other; topics further apart are less similar. The topic distance is used to determine their centers [16].

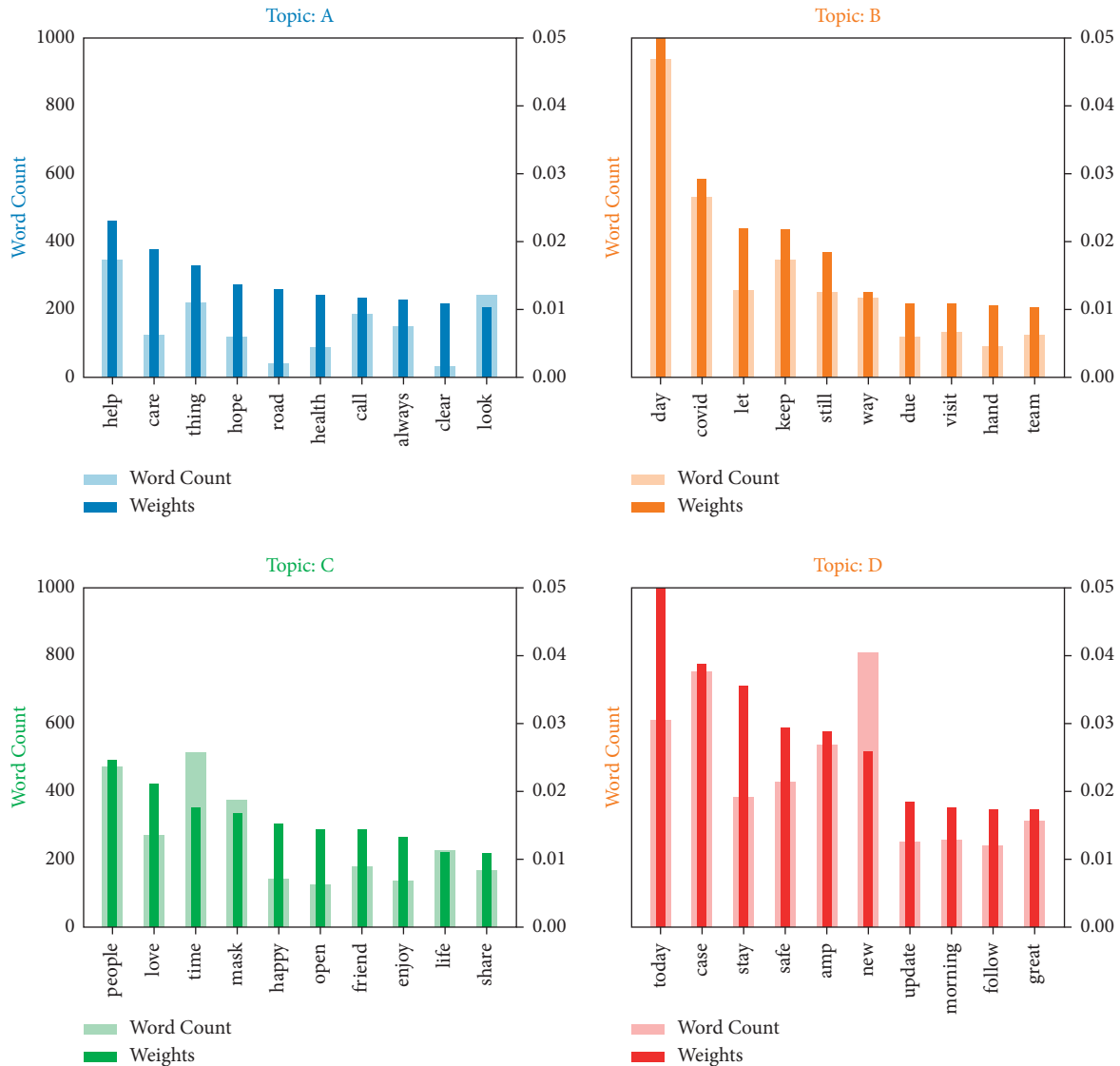


FIGURE 8: Extracted keywords and weight of their topic.

5. Conclusion

VADAR, AFINN, and TEXT BLOB polarity shows positive, negative, and neutral frequencies. Most results between these dictionaries-based algorithms are neutral. This study described that majority people were afraid of the COVID-19 pandemic situation and on the other side some people were enjoying their lockdown period such as they prefer to live at home, play, watch movies, and reading. The proposed study suggests that physical activities and exercise can refine cognition during the pandemic. The COVID-19 outbreak has been studied for its etiology, clinical features, transmission patterns, and management, but little has been done to explore its effects on mental health and ways to prevent stigmatization. People’s behaviors can significantly influence the dynamics of the pandemic by altering its severity, transmission, spread, and consequences. Raising public awareness can help deal with this calamity in the present situation. Despite the fact that high-frequency interventions had little concomitant effect on cognition functions, the threshold

remains to be worked out. In the future, we will extract various kinds of vaccination tweets datasets to study and analyze vaccine efficacy and effectiveness.

Data Availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank their colleagues from different institutes in Pakistan for sharing their insight and expertise during this research project.

References

- [1] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to sentiment analysis," in *Proceedings of the 2016 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, Vancouver, BC, Canada, 24 July 2016.
- [2] S. Jain and A. Sinha, "Identification of influential users on Twitter: a novel weighted correlated influence measure for Covid-19," *Chaos, Solitons & Fractals*, vol. 139, Article ID 110037, 2020.
- [3] T. A. Small, "What the hashtag? Information, Communication & Society," *Information, Communication & Society*, vol. 14, no. 6, pp. 872–895, 2011.
- [4] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, DBLP, Valletta, Malta, 17 May 2010.
- [5] H. Suresh, "An unsupervised fuzzy clustering method for twitter sentiment analysis," in *Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 6–8 October 2016.
- [6] J. Xiao, Y. Tian, L. Xie, X. Jiang, and J. Huang, "A hybrid classification framework based on clustering," vol. 16, no. 4, April 2020), 2019.
- [7] B. Supriya, V. Kallimani, S. Prakash, and C. B. Akki, "Twitter sentiment analysis using binary classification technique," in *Proceedings of the International Conference on Nature of Computation and Communication*, 17 March 2016.
- [8] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, "New machine learning method for image-based diagnosis of COVID-19," *PLoS One*, vol. 15, no. 6, Article ID e0235187, 2020.
- [9] A. Sharma, A. Adhikary, and S. B. Borah, "Covid-19's impact on supply chain decisions: strategic insights from NASDAQ 100 firms using Twitter data," *Journal of Business Research*, vol. 117, pp. 443–449, 2020.
- [10] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.
- [11] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users," *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, p. 2032, 2020.
- [12] M. Capelle, F. Fransincar, M. Moerland, and F. Hogenboom, "Semantics-based news recommendation," in *Proceedings of the 2nd international conference on web intelligence, mining and semantics*, 13 June 2012.
- [13] M. Cinelli, W. Quattrociocchi, A. Galeazzi et al., "The covid-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, pp. 16598–16610, 2020.
- [14] J. Zhou, S. Yang, C. Xiao, and F. Chen, "Examination of community sentiment dynamics due to Covid-19 pandemic: a case study from Australia," 2020, <https://arxiv.org/abs/2006.12185#:~:text=Based%20on%20the%20analysis%20of,polarity%20during%20the%20pandemic%20period>.
- [15] X. Han, J. Wang, M. Zhang, and X. Wang, "Using social media to mine and analyze public opinion related to COVID-19 in China," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, p. 2788, 2020.
- [16] R. D. Endsuy, "Sentiment analysis between VADER and EDA for the US presidential election 2020 on twitter datasets," *Journal of Applied Data Sciences*, vol. 2, no. 1, pp. 08–18, 2021.
- [17] F. Pakzad, M. Portmann, W. L. Tan, and J. Indulska, "Efficient topology discovery in software defined networks," in *Proceedings of the Signal Processing and Communication Systems (ICSPCS), 2014 8th International Conference on*, 15 December 2014.
- [18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [19] H. T. Vo, H. C. L. D. D. Nguyen, and N. H. Tuong, "Topic classification and sentiment analysis for Vietnamese education survey system," *Asian Journal of Computer Science & Information Technology*, pp. 27–34, 2016.
- [20] A. Razzaq, M. Asim, Z. Ali et al., "Text sentiment analysis using frequency-based vigorous features," *China Communications*, vol. 16, no. 12, pp. 145–153, 2019.
- [21] A. Z. Abbasi, N. Islam, and Z. A. Shaikh, "A review of wireless sensors and networks' applications in agriculture," *Computer Standards & Interfaces*, vol. 36, no. 2, pp. 263–270, 2014.

Research Article

A Study of the Influence of Collaboration Networks and Knowledge Networks on the Citations of Papers in Sports Industry in China

Yu Zhang ¹, Jianlan Ding ¹, Hui Yan ², Miao He ¹, and Wei Wang ¹

¹*Xi'an Physical Education University, Xi'an 710068, China*

²*School of Economics and Management, Shanghai University of Sport, Shanghai 200438, China*

Correspondence should be addressed to Jianlan Ding; 101014@tea.xaipe.edu.cn

Received 11 June 2021; Revised 15 December 2021; Accepted 3 January 2022; Published 9 February 2022

Academic Editor: Fei Xiong

Copyright © 2022 Yu 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 scientific paper's citation represents its influence, which is the most intuitive indicator to access the quality of papers. This paper mainly adopts the social network analysis method, using the authors and the keywords of sports industry papers in China to constitute the networks of collaboration and knowledge, to explore effects of the degree centrality of authors and keywords and the structural hole of authors and keywords on the citation of papers in the collaboration and knowledge networks and draw the following conclusions: (1) as for collaboration networks, the degree centrality at the paper level is positively correlated with citations; (2) in the collaboration network, the positive correlation between the structural hole at the paper level and citations does not exist; (3) within knowledge networks, an inverted-*U* shape was found between degree centrality and paper's citation; and (4) within knowledge networks, a positive correlation is in existence between the structural hole of papers and their citation. This study synthesizes the already widely used collaboration network with the knowledge network constructed through keywords, distinguishes from the previous network features focusing at the author level, and explores research projects of Chinese Sports Industry from the paper level, providing a new perspective for the research of sports industry in China, complementing the methods and ideas of sports industry research, as well as providing a reference for the research in other disciplinary fields.

1. Introduction

An article's citation manifests the times of the article will be used in subsequent research [1, 2], also the simplest and direct way to evaluate the paper, which represents the position of the paper in the context of academic research and the role it plays in scientific activities from the perspective of historical retrospection [3], and it plays an important role in talent evaluation, scientific research project establishment, and scientific research award. In addition, a series of indexes are derived from the citation amount of the paper. Therefore, the citations of a paper are often used to appraise the papers' impact [4, 5]. As the final manifestation of research achievements, the influence of papers can reflect the usefulness in some fields [6–8]; the higher the citations of a paper, the more recognized the paper's conclusion and the

greater the reference and significance for other scholars to conduct follow-up research. Previous studies have found that papers' citations vary greatly, with some papers receiving many citations, some reaching hundreds or even thousands of citations, while the majority of papers receive fewer citations, or even nearly 20% of papers are not cited at all [9, 10]. This may cause the consideration of academia: what exactly are the principal factors affecting the papers' citation?

Scholars studied several factors affecting the citations of a paper from different perspectives. For example, Bornmann et al. (2008) found that the emergence of citation relationships in academic papers may be due to academic aspects, or it may be due to some nonacademic considerations [11]. Tahamtan et al. (2016) outline the factors affecting the citations of a paper, most notably, followed by reasons such

as the abstract and other paper-related factors, the impact factor and other journal-related factors, number of authors, and other author-related factors [12].

With the development of social network research, Abbasi and Jaafari (2013) suggest that it is feasible to explore the number of citations of papers by constructing social networks [13]. Scientific collaboration networks are also a typical type of social network that is coming into the limelight [14, 15]. Authors have relatively different access to resources depending on their position in the scientific collaboration network, which can significantly affect the number of citations they receive [16, 17]. For instance, Li et al. (2013) found that the degree centrality of authors in a collaboration network was positively correlated with their citations [18]. Abbasi et al. (2011) investigated the relationship between authors' network characteristics in scientific collaboration networks and their citation performance and found that authors' degree centrality was positively associated with author citation performance, as was structural hole [19]. Based on this, since the knowledge network and collaboration network are also typical social networks, and they will affect the citation at the same time, this research attempts to further explore how the knowledge network affects the citation of papers, that is, quantity.

Compared with the collaboration network, another typical social network, that is the knowledge network, has not been attached importance in the research of paper citations. Knowledge networks are networks formed by the combination of scientific knowledge elements that can represent categories of knowledge domains. For example, patents are often divided into categories to distinguish them from technical features, and different patent categories represent different knowledge elements [20]. Similarly, after embodying knowledge elements that are keywords in scientific research papers, a collection of papers can also construct a complex knowledge network from a macro perspective. The current research also proves that the keywords of papers can be considered as knowledge elements. For instance, Su and Lee (2010), Assefa and Rorissa (2013), and Yang et al. (2016) built a knowledge structure map of a subject through keywords [21–23], Xie et al. (2008) used keywords to analyze trends in the evolution of research hotspots [24], and Chen (2006) used keywords to detect knowledge hotspots [25]. However, there are few research projects on the issue how knowledge networks affect the influence of papers. On the theoretical level, the collaboration network is decoupled from the knowledge network reciprocally, and the author's position does not commensurate with his knowledge elements' position. On the one hand, papers usually have different network structure characteristics in the embedded two networks. On the other hand, the formation mechanism of the two types of network structure characteristics is different. The formation of the collaboration networks structure characteristics mainly depends on the author of the paper. The formation of the structural characteristics of knowledge network depends on a large number of other past studies, so the two types of network should be treated differently. On the practical level, since the optimization of collaboration networks mainly

involves collaborators selection, the optimization of knowledge networks further involves the adjustment of the research theme and the way of expression during the whole research process, and the content and timing of the optimization of them are different. Therefore, a separate discussion between the two can also further enrich the solution to the problem of increasing paper citation and influence.

Based on this, this study uses keywords to represent the knowledge elements of the papers, and because knowledge elements are affiliated by the co-occurrence of keywords in predecessor's research, we construct a knowledge network through keyword co-occurrence. The size of the knowledge network will grow larger over time, while the categories of knowledge elements will become richer and new combinations of knowledge elements will be added [26]. The richer the variety of knowledge elements and the new combinations that are constantly being added ultimately also promote the evolution of knowledge networks. In this research, we argue that a knowledge element's position affects the opportunities of combining knowledge elements with others. For instance, a central knowledge element is more searchable because it has more element-coupled content and experience. Therefore, we believe that the location attribute of the knowledge element affects the citation of the paper. By introducing knowledge networks, we hope to enrich the research on the influencing factors of paper citations.

The data used in this study are publications on China's sports industry from 2000 to 2021. The 2018 Government Work Report made it clear that the sports industry entered the overall layout of the national economy for the first time, and the sports industry was also clearly defined by the National Development and Reform Commission as a "new wind outlet" for economic development. In 2019, the Outline for Building a Leading Sports Nation was issued by the General Office of the State Council, which proposed that by 2035, the sports industry will be developed into one of the pillar industries of the national economy. It can be predicted that the sports industry will usher in a golden period of development. The scientific research on the sports industry also provides a scientific guarantee for the sports industry to become the pillar industry of the national economy. Of course, the in-depth and promotion of the research cannot be separated from the reference and significance brought by the related citations. Previous publications on citations focus on the author level, institutional level, or journal level. However, considering that citations vary in different publications for the same author, and the average citation is not enough to reflect the influence level of each paper, this paper made a specific analysis of the citation amount from the paper level. Considering the average level of authors or keywords in each paper, this paper studies the influence of the average degree centrality and average structure hole of authors or keywords on the citation from the collaboration and knowledge networks, respectively. It concerned how authors' and keywords' location attributes in the network affect the citations of a paper, which gives a new dimension to study the influencing factors of citations in the sports industry and provides a reference for the improvement of the quality of papers in sports industry and even the field of

sports science. At the same time, this study can not only provide theoretical reference for subsequent scholars on how to improve scientific research quality and the amount of citation but also provide the theory basis for performance evaluation of the scientific research of scholars and the scientific guarantee for the sports industry to become a pillar industry of the national economy. This study has the following contributions: (1) we have constructed a knowledge network by using keywords from the paper, which fills a gap in previous research and will inspire further relevant research; (2) particularly, the mechanism about how degree centrality and structural holes in collaboration and knowledge networks impinge on citations is probed respectively. Specifically, the location attributes of nodes in those two networks are taken as the influencing factors on citations; and (3) this study places emphasis on article-level citations.

2. Theoretical Background and Research Hypothesis

Collaboration and knowledge networks are important contents of social network analysis and common methods in scientific research. The network characteristics of each node in the network vary, and the opportunity to acquire new information in the network is also different [27]. The feature analysis for nodes in collaboration and knowledge networks is the key to the application of social network analysis methods and high-quality achievement. In our study, degree centrality and structural hole are picked as two network attribute indexes to carry out related research. Degree centrality shows the amount of nodes who are in direct contact with node i in an N -nodes network. The higher the degree centrality of a node, the more nodes it contacts, and the more superior the node is in the network. Structural hole refers to that in the network, if a node has a direct connection with two nodes that are not in direct connection with each other, then this node occupies a position of the structural hole. Structural hole is a key attribute of nodes in a network. By occupying the position of the structural hole, nodes can obtain nonredundant information efficiently. Two network characteristic indexes, degree centrality and structural hole, were selected for the following two reasons: first, with the deepening of the scientific research, more scholars prefer to study local indicators of network attributes, and the degree of centrality and structure hole is not only a commonly used indicator in network analysis but also the network attribute local index [27]; second, if other independent variables, such as betweenness centrality and closeness centrality, are added, the inhibiting effect may be generated when the model is established and the combined effect of independent variables on dependent variables is analyzed, thus leading to the reverse β coefficient [19].

2.1. The Collaboration Network. The collaboration network in this research, in other words, is the researcher coauthorship network, where nodes and ties represent researchers' cooperative relationship in prior papers. The

collaboration network refers to scientific collaboration network, in which each node means an author, and the existing edges between nodes indicate that two authors have worked together previously.

2.1.1. The Influence of Degree Centrality on Citations in Collaboration Network. If a central location is occupied by an author in a collaboration network, it means that the author is likely to have numerous connections and access to the required information and resources [16]. The external information and fresh ideas provided by the resource can promote the research process. At the same time, by exchanging ideas with more different authors, their theoretical horizons can be broadened to a certain extent, which is of great benefits to improving the research quality. All these make them more likely to produce highly influential scientific research results [28]. Meanwhile, the higher the degree centrality of authors, the more frequently they cooperate with others in the collaboration network, which is bound to enhance their popularity, gain structural social capital, and get more attention and citation for their achievements. Some scholars have found that in the collaboration network at the author level, the degree centrality of authors is positively correlated with their citation performance. Abbasi et al. (2011) found that the degree centrality of authors was positively correlated with the g -index constructed based on the citation of articles [19]. Hao et al. (2020) studied 14,913 publications in SCIENCE journal from 2000 to 2018 and found that, in the scientific collaboration network at the author level, authors' degree centrality has a positive correlation with citation of their papers [29]. Thus,

H1a: for a paper, its authors' average degree centrality in the collaboration network is positively correlated to its citation count.

2.1.2. The Influence of Structural Holes on Citations in Collaboration Network. If an author is connected with those who are not connected to each other, then he crosses the structural hole. For instance, there are authors A, B, and C, author A is a structural loophole connecting author B and author C. When author A is in direct contact with two partners (author B and author C), there is no connection between the two partners. In other words, author A occupies the structural holes location. The structural hole illustrates the degree of interrelation among authors who have cooperated with an author in the cooperation network. The more structural holes the author occupies, the easier it is to gain control advantage [30], that is, authors occupying structural holes are more likely to have potential opportunities to control the flow of information between unconnected authors [31]. Moreover, structural holes are the hub of heterogeneous information flow in the network. The ties established by structural holes are nonredundant, and authors occupying the positions of structural holes can obtain a large amount of nonredundant information. With such nonredundant information, authors are more likely to improve their research quality and obtain more citations.

H1b: for a paper, its authors' structural holes in the collaboration network are positively correlated to its citation count.

2.2. The Knowledge Network. The knowledge network here is composed of the keywords in papers as knowledge elements, in which every node signifies a keyword, and the inter-connection between nodes means the co-occurrence of keywords in previous studies. The knowledge network refers to the network composed of the keywords of the paper as knowledge elements, in which each node means a keyword, and the existing edges between nodes indicate the co-occurrence of keywords in previous studies.

2.2.1. The Influence of Degree Centrality on Citations in Knowledge Network. Knowledge elements' degree centrality in the knowledge network indicates the degree of combination with other elements. Knowledge elements consist of parts that depend on each other to form a larger scale of knowledge system [32]. The combinatorial opportunities tend to rise with the increasing of knowledge element's centrality, and two main reasons have been issued here. Firstly, an element with higher degree centrality must have been integrated to more of them, that is to say, it is a knowledge element with a wider scope and better applicability, which will also prompt the authors to carry out a more in-depth discussion on this knowledge element in the subsequent research [33]. Secondly, higher degree centrality of knowledge elements can provide authors with more examples of this combination of knowledge elements and inspire them to carry out innovative research from different ideas and perspectives. Therefore, the citation count of papers related to this knowledge element will also be enhanced with the continuous development of related research at this knowledge element. However, when the degree centrality of knowledge elements reaches a certain degree, the combination opportunities of knowledge elements may decrease. That is to say, when knowledge elements are excessively combined, their combined value will reduce. Combining with this knowledge element may lead to insufficient innovation in the final scientific research results, and the citation count of papers related to this knowledge element in subsequent studies will also be reduced. Therefore,

H2a: for a paper, its keywords' average degree centrality in the knowledge network has an inverted- U shape on its citation count.

2.2.2. The Influence of Structural Holes on Citations in Knowledge Network. If a knowledge element is directly related to two nondirectly connected elements in the knowledge network, then the knowledge element occupies the position of structural holes. The search for knowledge is mostly internal search or related search [34]. Therefore, knowledge elements located in structural holes can provide more opportunities for combination of two knowledge elements that have no direct connection [35]. If a knowledge

element occupies more structural holes, then the author can find more nonredundant relevant knowledge elements through this knowledge element and can find more combinations of knowledge elements that have never appeared, and the paper containing this knowledge element will also receive more references. Thus,

H2b: for a paper, its keywords' average structural holes in the knowledge network are positively related to its citation count.

3. Research Methods

3.1. Data Collection. All data acquired are from the Chinese Social Science Citation Index (CSSCI) database and the General Contents of Chinese Core Journals (core of Peking University) database. We search "sports industry" in the retrieval field, which is limited to "subject," and the publication time is limited to 2000–2021. The final excerpt preserves the name, author, key words, citation count and other information of all publications, and 7,465 pieces of original data obtained.

3.2. Variables Selection and Measurement

3.2.1. Dependent Variables. The dependent variable is citations that have been normalized of each paper in our sample publications. Referring to the method proposed by Cannella and McFadyen in 2016 [36], the citation count of an article is first subtracted from the average citations of all sports industry articles published during the same year and then divided by the standard deviation of the citations for all sports industry articles published in the same year. Finally, the normalized citations of this paper are obtained. This method can eliminate the citation bias of articles published in different years. Its normalized citations calculation is as follows:

$$\text{normalized citations}_i = \frac{\text{citations}_i - \text{citationsmean}_{\text{all}}}{\text{citations standard deviation}_{\text{all}}} \quad (1)$$

3.2.2. Independent Variables. The disquisitive independent variables involve in degree centrality and structural holes in collaboration and knowledge networks constructed based on all the papers in the sample. The specific methods are as follows:

(1) *Construction of the Collaboration and Knowledge Networks.* The collaboration and knowledge networks here are typical social networks constructed based on the papers of sports industry journals in China from 2000 to 2021. Both in this study are constructed by Python.

(2) *Measurements of Degree Centrality and Structural Holes.* Two groups of independent variable, the degree centrality and structural holes, in both collaboration and knowledge networks, are discussed in this study at the level of papers. The specific node degree and structural hole are calculated by Python.

(a) Calculation of degree centrality

- (i) Firstly, to acquire the degree centrality, we reckon the number nodes that are directly relevant to the node. Then, we make standardized treatment. To obtain the standardized degree centrality, the value is divided by the amount of remaining nodes. The formula is as follows:

$$\text{normalized degree centrality}_i = \frac{\text{degree centrality}_i}{g - 1} \quad (2)$$

- (ii) In the formula, g represents the amount of nodes in the network.

(b) Calculation of structural holes

- (i) Burt's constraint method was first adopted [37, 38] to reckon the network constraint C_i , which indicates the strength i is constrained by its adjacent nodes. We use 2 minus the constraint metric C_i 2 to obtain the control advantage that i generates by crossing structural holes.

$$\text{Structural holes}_i = 2 - C_i = 2 - \sum_j \left(p_{ij} + \sum_{k, k_j, k_j^1} p_{ik} p_{kj} \right)^2, \quad (3)$$

where i indicates the target node and p_{ij} represents the ratio node j to the contact point of node i . For example, i is connected to five nodes including j , then p_{ij} equals 1/5. Node k has a connection with node i and j simultaneously. The more ties between i and other elements exist, the smaller value of p_{ij} and p_{ik} and less constraint node i have. Meanwhile, the more ties k has with other elements, the lower p_{kj} k has, thereby lower the constraint on i .

(3) *Degree centrality and structural holes at paper level.* The basic unit focuses on paper-level degree centrality and structural hole. Because of the nonuniqueness of authors and elements in a paper, both indicators need to be averaged to paper level. This research draws on the method of calculating paper-level degree centrality and structural holes proposed by Guan et al. (2017) [27]. For instance, there are three authors in a paper whose degrees are 1.2, 1.3, and 1.4, respectively, then the average degree centrality value of the paper's authors in the collaboration network is $(1.2 + 1.3 + 1.4)/3 = 1.3$. Structure holes in collaboration networks and degree centrality in knowledge networks both are equal to structure holes.

4. Regression Results and Analysis

4.1. Descriptive Statistics. In the collaboration network, if the author's degree centrality is 0, it indicates that the author has no cooperative relationship with other authors. Therefore, authors with degree centrality of 0 can be excluded from the collaboration network. After processing, the number of nodes

in the collaboration network is finally 4,851. The mean, median, standard deviation, minimum, and maximum values of variables in the collaboration network are depicted in Table 1. In the knowledge network, although there is no knowledge element with degree centrality of 0, some keywords need to be combined manually. For example, "2008 Olympic Games" and "2008 Beijing Olympic Games" both refer to the 2008 Olympic Games held in Beijing, so they can be combined into one. Through manual screening and merging of all the keywords, the final number of knowledge network nodes is 7379. And the mean value, median, standard deviation, minimum value, and maximum value of variables in the knowledge network are listed in Table 2.

In Tables 1 and 2, as for the collaboration network, degree centrality is significantly different from that in the knowledge network. The mean value of degree centrality in the collaboration network equals to 0.001, the standard deviation is 0.001, and the maximum value is 0.014. However, in the knowledge network, the mean value of degree centrality equals to 0.008, the standard deviation and the maximum value are 0.009 and 0.074, respectively. The difference between structural holes in the collaboration network and knowledge network is small. The mean value of the structural holes in the collaboration network equals to 1.335, the standard deviation is 0.295, the minimum and maximum values are 0.009 and 1.959, respectively. However, in the knowledge network, the mean value of structural holes is 1.783, the standard deviation is 0.21, the minimum value is 0.875, and the maximum value is 1.997. Because of the normalization, the average of the citations in both collaboration and knowledge networks is 0.

4.2. Regression Analysis

4.2.1. The Collaboration Network

(1) *Degree Centrality as the Independent Variable.* In order to verify whether there is a nonlinear relation between degree centrality of collaboration network and the citation count of the paper, a quadratic term was added into the regression, and the regression analysis was conducted by using Stata. The results are shown in Table 3.

Table 3 signifies that the quadratic regression is significant ($p \leq 0.05$), and the quadratic coefficient is $-6441.203 < 0$, preliminarily determined to be an inverted- U relationship. The UTEST test was conducted to further verify the relationship. The results are shown in Table 4.

As can be seen from Table 4, the extreme value is out of $[0.000, 0.014]$, so the null hypothesis cannot be rejected. Therefore, there is no U or inverted U relationship. In order to verify whether there is a linear relationship, unary linear regression is conducted. The results are shown in Table 5.

Table 5 signifies the significance ($p \leq 0.01$) of unary linear regression, and the coefficient is $135.674 > 0$, that indicates the degree is marked positive correlated with citation count, which verifies the hypothesis H1a, that is, the average degree centrality of a paper's author in the collaboration network is positively correlated to its citation count.

TABLE 1: Descriptive statistics of collaboration network variables.

	<i>N</i>	Mean	Std. Dev.	Min.	Max.
Citation	4851	0	0.998	-0.984	12.351
Degree	4851	0.001	0.001	0	0.014
Structure hole	4851	1.335	0.295	0.875	1.959

TABLE 2: Descriptive statistics of knowledge network variables.

	<i>N</i>	Mean	Std. Dev.	Min.	Max.
Citation	7379	0	0.999	-0.888	13.376
Degree	7379	0.008	0.009	0	0.074
Structure hole	7379	1.783	0.21	0.875	1.997

TABLE 3: Collaboration network degree centrality quadratic regression.

	Coef.	St. Err.	<i>t</i> -value	<i>p</i> -value	[95% Conf. Interval]		Sig.
Degree	182.095	21.43	8.50	≤0.001	140.082	224.108	***
Degree ²	-6441.203	2627.278	-2.45	0.014	-11591.858	-1290.547	**
Constant	-0.178	0.022	-8.22	≤0.001	-0.221	-0.136	***

*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

TABLE 4: Collaboration network degree centrality UTEST.

	Lower bound	Upper bound
Interval	0.000	0.014
Slope	179.471	3.683

Extreme point: 0.0141352
 Test: H1: inverse *U* shape vs. H0: monotone or *U* shape
 Extremum outside interval: trivial failure to reject H0

TABLE 5: Collaboration network degree centrality unary linear regression.

	Coef.	St. err.	<i>t</i> value	<i>p</i> value	95% conf. interval		Sig.
Degree	135.674	10.042	13.51	≤0.001	115.986	155.361	***
Constant	-0.148	0.018	-8.30	≤0.001	-0.183	-0.113	***

*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

(2) *Structural Holes as the Independent Variable*. In order to verify whether there is a nonlinear relationship between structural holes in collaboration network and citation count of a paper, a quadratic term was added into the regression, and the regression analysis was conducted by using Stata. The results are shown in Table 6.

Table 6 shows that the result of quadratic regression is significant ($p \leq 0.01$), and the quadratic coefficient is $0.981 > 0$, preliminarily determined to be a *U*-shaped relationship. The UTEST test was conducted to further verify the relationship. The results are shown in Table 7.

As can be seen from Table 7, the extreme point is in range of the data [0.875, 1.959], and the UTEST results are significant ($p \leq 0.05$), so the null hypothesis is rejected at the statistical level of 5%. Meanwhile, the slope interval in the result ranges from -0.361 to 1.767 , which is consistent with the preliminary *U* relationship determined by quadratic regression. Therefore, we can consider that there is a *U*

relationship. The regression results are inconsistent with the research hypothesis H1b.

4.2.2. The Knowledge Network

(1) *Degree Centrality as the Independent Variable*. In order to verify whether there is a nonlinear relationship between the moderate centrality of knowledge network and the citation amount of the paper, a quadratic term was added into the regression, and the regression analysis was conducted by using Stata. The results are shown in Table 8.

We can see from Table 8 that the quadratic regression is significant ($p \leq 0.05$), and the quadratic coefficient is $-204.433 < 0$, preliminarily determined to be an inverted-*U* relationship. The UTEST test was conducted to further verify the relationship. The results have been shown in Table 9.

As can be seen from Table 9, the extreme value is within [0, 0.074], and the UTEST results are significant ($p \leq 0.05$),

TABLE 6: Quadratic regression of collaboration network structure holes.

Citation	Coef.	St. err.	t value	p value	95% conf. interval		Sig.
Structure hole	-2.078	0.477	-4.36	≤ 0.001	-3.013	-1.143	***
Structure hole ²	0.981	0.175	5.62	≤ 0.001	0.639	1.323	***
Constant	0.939	0.314	2.99	0.003	0.324	1.555	***

*** $p < 0.01$, ** $p < 0.05$ and * $p < 0.1$.

TABLE 7: Collaboration network structure UTEST.

	Lower bound	Upper bound
Interval	0.875	1.959
Slope	-0.361	1.767
t value	-2.052	8.231
$P > t $	0.020	0.000
Extreme point: 1.058766		
Test: $H1$: UU shape vs. $H0$: monotone or inverse UU shape		
Overall test of presence of a U shape: t value = 2.05		
$P > t = 0.020$		

TABLE 8: Knowledge network degree central quadratic regression.

Citation	Coef.	St. err.	t value	p value	95% conf. interval		Sig.
Degree	10.171	3.068	3.32	0.001	4.157	16.184	***
Degree ²	-204.433	84.489	-2.42	0.016	-370.054	-38.811	**
Constant	-0.055	0.019	-2.84	0.004	-0.093	-0.017	***

*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

TABLE 9: Knowledge network degree centrality UTEST.

	Lower bound	Upper bound
Interval	0.000	0.074
Slope	10.138	-20.003
t -value	3.318	-2.039
$P > t $	0.000	0.021
Extreme point: 0.0248758		
Test: $H1$: UU shape vs. $H0$: monotone or inverse UU shape		
Overall test of presence of an inverse UU shape: t value = 2.04		
$P > t = 0.0208$		

so the null hypothesis is rejected at the statistical level of 5%. Meanwhile, the slope interval in the result ranges from 10.138 to -20.003, which is consistent with the preliminary result of quadratic regression that it is an inverted- U relationship. Therefore, it can be considered that there is an inverted- U relationship, which also verifies the hypothesis H2a.

(2) *Structural Holes as the Independent Variable*. In order to verify whether there is a nonlinear relationship between structural holes in the knowledge network and citation count of a paper, a quadratic term is added into the regression, and the regression analysis is conducted by using Stata. The results are shown in Table 10.

Table 10 signifies the significance ($p \leq 0.01$) of quadratic regression, and the quadratic coefficient is $0.444 > 0$, preliminarily determined to be a U relationship. The UTEST test was conducted to further verify the relationship. The results are shown in Table 11.

TABLE 10: Quadratic regression of knowledge network structural holes.

Citation	Coef.	St. err.	t value	p value	(95% conf. interval)		Sig.
Structure hole	-1.122	0.544	-2.06	0.039	-2.188	-0.057	**
Structure hole ²	0.444	0.169	2.63	0.009	0.113	0.776	***
Constant	0.569	0.431	1.32	0.186	-0.275	1.414	

*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 11 shows that the extreme value is in range of [0.875, 1.997], but the UTEST results are not significant ($p > 0.05$). Therefore, there is no U or inverted- U shape relation. In order to verify whether there is a linear relationship, unary linear regression is conducted. The results are shown in Table 12.

In Table 12, the unary linear regression result was significant ($p \leq 0.01$), and the coefficient was $0.298 > 0$,

TABLE 11: Knowledge network structural holes UTEST.

<!--Col Count:3F0E0	Lower bound	Upper bound
Interval	0.875	1.997
Slope	-0.345	0.652
<i>t</i> -value	-1.374	4.480
$P > t $	0.085	0.000
Extreme point: 1.263264		
Test: H_1 : UU shape vs. H_0 : monotone or inverse UU shape		
Overall test of presence of a inverse UU shape: t value = 1.37		
$P > t = 0.0847$		

TABLE 12: Knowledge network structural holes unitary linear regression.

Citation	Coef.	St. err.	<i>t</i> value	<i>p</i> value	(95% conf. interval)		Sig.
Structure hole	0.298	0.055	5.40	≤ 0.001	0.19	0.407	***
Constant	-0.532	0.099	-5.36	≤ 0.001	-0.726	-0.337	***

*** $p < 0.1$, ** $p < 0.05$, and * $p < 0.1$

indicating that the structural holes are significant positively correlated with its citation, which also verifies the hypothesis H2b, that is, for a paper, its keywords' average structural holes in the knowledge network are positively related to its citation count.

5. Conclusion

This paper, by using the 2000–2021 Chinese Social Science Citation Index (CSSCI) database and Chinese Core Journals Particular Overview (core) of Peking University database in sports industry data in China, constructs the keywords knowledge and authors collaboration network and explores the relationship between network structure attributes and citation count from two perspectives. Our above results can be concluded as the following findings.

First of all, as for the collaboration network, the average degree centrality of a paper's author is positively correlated with its citation count, that is, with the increase of the average degree centrality of the paper's author, the citation count also increases. Authors with higher degree centrality tend to cooperate with others with high degree centrality, which makes it easier to find more innovative ideas and acquires more opportunities to share resources, thus improving the quality of paper research. At the same time, authors with the higher degree centrality usually acquire relatively higher academic status in the field, and cooperating with them is more likely to gain the attention and support of peers, thus increasing the citation count of their papers.

Secondly, in the collaboration network, the average structural holes of a paper's author are not positively correlated with citation.

Thirdly, we confirmed that the average degree centrality of keywords in the knowledge network had an inverted- U shape impact on citation count, that is, with the increase of the average degree centrality of all keywords, the citation of a paper increases at first and then decreases when it reaches a certain altitude. The knowledge element tends to combine with other knowledge elements in pace of the increase of the

degree centrality of the knowledge element. It will improve the utilization rate of existing knowledge elements and provide more elements combination model, and the knowledge elements of related papers citation count will rise with knowledge elements' growth. When the degree centrality of knowledge element reaches a certain degree, the research on this knowledge element has been relatively sufficient, and the value of combining and studying this knowledge element is relatively low. Therefore, the citation amount of papers related to this knowledge element will also decrease due to the lack of innovation.

Finally, we found that the average structural holes of keywords in the knowledge network for a paper are positively correlated with citation count. The increase of the structure holes of all the keywords in a paper can lead to an increase in citation count for a paper. When the knowledge element occupies more of the structural holes, the knowledge element connects more with more nonredundant knowledge elements and may find more fresh knowledge element combination. Therefore, as the paper's richness of the average structural holes, perhaps it contains more novel knowledge element combinations, thus increasing the citation count.

6. Research Implications and Limitations

By analyzing the characteristics of the social networks at the paper level, this paper enriches the research on collaboration and knowledge networks of sports industry and paper citations and reveals the influence of the two networks on citation count in the sports industry. Through this study, we can find that the citation count of sports industry papers will be affected by the attributes of its own network structure. Therefore, it is necessary to seek cooperation with more scholars while writing papers and select collaborators with a high degree centrality as far as possible, which is conducive to increasing the citation count of papers and improving the influence of scientific research. At the same time, the study will make the knowledge elements of the paper highly cited and innovative, which is conducive to improving the citation

count of a paper. Moreover, the more structural holes knowledge elements occupy, the more conducive to improving the citation count of a paper.

At the same time, some contributions at theoretical level have been made. Firstly, this study not only involves the collaboration network, which has been widely used, but also constructs the knowledge network through keywords and applies the collaboration and knowledge network to the research of the sports industry, which provides new approaches and thoughts for the research of sports industry in China. Secondly, compared with previous studies that mostly focus on the network attributes at the author level, the basic research unit of this study focuses on the paper level, which furnishes a new perspective for following research projects of the sports industry and other disciplines. Finally, the study can not only provide theoretical reference for scholars in the sports industry on how to improve the quality of scientific research and increase the citation count but also provide a theoretical basis for predicting the citations count of papers and evaluating scholars' research performance through a more scientific comprehensive way.

Our study also has some limitations. Firstly, authors of collaboration and keywords of knowledge networks have not been given a certain weight, and the contribution degree for the author and the importance of keywords in each paper have not been distinguished, which have reached a common understanding in academia, and we need to further promote in the subsequent research projects. Secondly, the relationship between structural holes and citations in the collaboration network and the reasons for their generation need to be discussed and explained in the following research projects.

Data Availability

The original data used in this study are from the Chinese Social Sciences Citation Index (CSSCI) database and General Contents of Chinese Core Journals (core of Peking University) database. The original data used to support the findings of this study are available from <https://cssci.nju.edu.cn/>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] G. A. Lozano, V. Larivière, and Y. Gingras, "The weakening relationship between the impact factor and papers' citations in the digital age," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 11, pp. 2140–2145, 2012.
- [2] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.
- [3] J. Zhao, "Citation index and its influencing factors in the evaluation of academic journals," *Modern Publishing*, vol. 20, no. 4, pp. 67–70, 2013.
- [4] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, vol. 178, no. 4060, pp. 471–479, 1972.
- [5] L. Leydesdorff and T. Opthof, "Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on Fractional counting of citations," *Journal of the Association for Information Science & Technology*, vol. 61, no. 11, pp. 2365–2369, 2014.
- [6] E. Garfield, "Is citation analysis a legitimate evaluation tool?" *Scientometrics*, vol. 1, no. 4, pp. 359–375, 1979.
- [7] L. Leydesdorff and L. Bornmann, "Integrated impact indicators compared with impact factors: an alternative research design with policy implications," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 11, pp. 2133–2146, 2011.
- [8] R. Rousseau, C. García-Zorita, and E. Sanz-Casado, "The h-bubble," *Journal of Informetrics*, vol. 7, no. 2, pp. 294–300, 2013.
- [9] S. Redner, "How popular is your paper? An Empirical study of the citation Distribution," *Physics of Condensed Matter*, vol. 4, no. 2, pp. 131–134, 1998.
- [10] J. Mingers and Q. L. Burrell, "Modeling citation behavior in Management Science journals," *Information Processing & Management*, vol. 42, no. 6, pp. 1451–1464, 2006.
- [11] L. Bornmann and H. D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.
- [12] I. Tahamtan, A. Safipour Afshar, and K. Ahamdzadeh, "Factors affecting number of citations: a comprehensive review of the literature," *Scientometrics*, vol. 107, no. 3, pp. 1195–1225, 2016.
- [13] A. Abbasi and A. Jaafari, "Research impact and scholars' geographical diversity," *Journal of Informetrics*, vol. 7, no. 3, pp. 683–692, 2013.
- [14] M. A. McFadyen and A. A. Cannella, "Social capital and knowledge creation: diminishing returns of the number and strength of exchange relationships," *Academy of Management Journal*, vol. 47, no. 5, pp. 735–746, 2004.
- [15] J. Guan, K. Zuo, K. Chen, and R. C. M. Yam, "Does country-level R&D efficiency benefit from the collaboration network structure?" *Research Policy*, vol. 45, no. 4, pp. 770–784, 2016.
- [16] A. Abbasi, L. Hossain, and L. Leydesdorff, "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 403–412, 2012.
- [17] M. E. J. Newman, "Co-authorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, vol. 101, no. s1, pp. 5200–5205, 2004.
- [18] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: a social capital perspective," *Research Policy*, vol. 42, no. 9, pp. 1515–1530, 2013.
- [19] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures," *Journal of Informetrics*, vol. 5, no. 4, pp. 594–607, 2011.
- [20] G. Carnabuci and E. Operti, "Where do firms' recombinant capabilities come from? Intraorganizational networks, knowledge, and firms' ability to innovate through technological recombination," *Strategic Management Journal*, vol. 34, no. 13, pp. 1591–1613, 2013.
- [21] H.-N. Su and P.-C. Lee, "Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight," *Scientometrics*, vol. 85, no. 1, pp. 65–79, 2010.

- [22] S. G. Assefa and A. Rorissa, "A bibliometric mapping of the structure of STEM education using co-word analysis," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 12, pp. 2513–2536, 2013.
- [23] S. Yang, R. Han, D. Wolfram, and Y. Zhao, "Visualizing the intellectual structure of information science (2006–2015): introducing author keyword coupling analysis," *Journal of Informetrics*, vol. 10, no. 1, pp. 132–150, 2016.
- [24] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 359–377, 2006.
- [25] S. Xie, J. Zhang, and Y.-S. Ho, "Assessment of world aerosol research trends by bibliometric analysis," *Scientometrics*, vol. 77, no. 1, pp. 113–130, 2008.
- [26] R. Garud and A. Kumaraswamy, "Vicious and virtuous circles in the management of knowledge: the case of infosys technologies," *MIS Quarterly*, vol. 29, no. 1, pp. 9–33, 2005.
- [27] J. Guan, Y. Yan, and J. J. Zhang, "The impact of collaboration and knowledge networks on citations," *Journal of Informetrics*, vol. 11, no. 2, pp. 407–422, 2017.
- [28] M. K. Ahuja, D. F. Galletta, and K. M. Carley, "Individual centrality and performance in virtual R&D groups: an empirical study," *Management Science*, vol. 49, no. 1, pp. 21–38, 2003.
- [29] Z. H. Hao, Y. Chen, and P. Wang, "The relationship between scientific collaboration network centrality and research impact - using the journal science(2000–2018) as an example," *Library Tribune*, vol. 44, no. 4, pp. 79–88, 2020.
- [30] S. Rodan and C. Galunic, "More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness," *Strategic Management Journal*, vol. 25, no. 6, pp. 541–562, 2004.
- [31] E. Lazega and R. S. Burt, "Structural holes: the social structure of Competition," *Revue Française de Sociologie*, vol. 36, no. 4, p. 779, 1995.
- [32] J. Guan and N. Liu, "Exploitative and exploratory innovations in knowledge network and collaboration network: a patent analysis in the technological field of nano-energy," *Research Policy*, vol. 45, no. 1, pp. 97–112, 2016.
- [33] T. S. Kuhn, "The structure of scientific revolutions," *Physics Today*, vol. 16, no. 4, p. 69, 1962.
- [34] R. M. Cyert and J. G. March, *A Behavioral Theory of the Firm*, Social Science Electronic Publishing, Rochester, NY, USA, 1963.
- [35] C. Wang, S. Rodan, M. Fruin, and X. Xu, "Knowledge networks, collaboration networks, and exploratory innovation," *Academy of Management Journal*, vol. 57, no. 2, pp. 484–514, 2014.
- [36] A. A. Cannella and M. A. McFadyen, "Changing the exchange," *Journal of Management*, vol. 42, no. 4, pp. 1005–1029, 2016.
- [37] R. S. Burt, *Structural Holes: The Social Structure of Competition*, Harvard University Press, Cambridge, MA, USA, 1995.
- [38] R. S. Burt, "Structural holes and good ideas," *American Journal of Sociology*, vol. 110, no. 2, pp. 349–399, 2004.

Research Article

Evolutionary Game of Social Network for Emergency Mobilization (SNEM) of Magnitude Emergencies: Evidence from China

Rui Nan , Jingjie Wang , and Wenjun Zhu 

School of Law and Humanities, China University of Mining and Technology, Beijing 100083, China

Correspondence should be addressed to Rui Nan; nr19841018@163.com

Received 10 June 2021; Revised 9 October 2021; Accepted 5 January 2022; Published 29 January 2022

Academic Editor: Fei Xiong

Copyright © 2022 Rui Nan 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.

As a common social network, the SNEM plays an important role in emergency management. Magnitude emergencies are characterized by high complexity and uncertainty, and it is impossible to rely on the government for emergency management alone. We should absorb multiple subjects to build the SNEM and carry out extensive emergency mobilization in the whole society. The SNEM can integrate resources, gather consensus, promote participation, and reduce risks. The analysis of the types, generation mechanism, subject behavior, and strategy selection of the SNEM aid in adopting appropriate mobilization strategy based on magnitude emergencies, achieving the adaptation of the SNEM and emergency scenarios. By constructing the evolutionary game model of the SNEM for magnitude emergencies, taking China as an empirical sample, this paper explores the behavior evolution law and stable strategy of the government, social organizations, and the public. The results showed that the symbiotic SNEM with a positive response of social organizations and the public under the path of high-intensity mobilization by the government is the best strategy combination, and it is conducive to maximizing the emergency joint force.

1. Introduction

The social network is a collection of multiple points (social actors) and connections between points (relationships between actors). They coordinate resources, transmit information, provide services, and solve problems through conscious coordination and cooperation [1, 2]. The SNEM is also a type of social network formed by internal and intersubjects of emergency mobilization through the information exchange and mutual assistance of resources in the context of emergency mobilization for magnitude emergencies. The SNEM plays a pivotal role in emergency management.

The “9–11” attacks brought painful losses to the United States and were a call to guard against magnitude emergencies worldwide. The SARS virus greatly threatened the lives and property of people all over the world. The outbreak, spread, and continuous development of the COVID-19 have plunged the world into a protracted conflict and brought serious damage to the world. At present, the world faces volatility, uncertainty, complexity, ambiguity (VUCA), and

unpredictable events are becoming more frequently, which pose a potential but unpredictable serious threat to sustainable development. The highly complex governance scenarios, the uncertainty of magnitude emergencies, and the limitation of risk governance led to emergency failure of the government. It is necessary to mobilize social organizations and the public to participate in collective actions to improve the emergency quality and efficiency [3]. Countries all over the world are actively exploring the emergency management of magnitude emergencies and have formed different characteristics and models, and China is no exception. The large-scale SNEM formed through social mobilization is an essential feature of emergency management and a potent weapon, helping China to overcome many emergencies. The SNEM generated in this process is considered an effective form of organization to deal with magnitude emergencies. In the fight against the COVID-19, the Chinese government launched a wide range of social mobilization, calling on social organizations and the public quickly to reach substantive cooperation with the government, establish the broad SNEM, and work together to effect

major change through the network. The cooperation of multiple subjects prevented the spread and development of the epidemic in a short period.

Due to the attributes of collective activities, heterogeneity among organizations, and interest differences among subjects and limited rationality in emergency management, each subject in the SNEM has repeated game behavior; that is, each subject will constantly adjust their own strategy according to different emergency scenarios and the strategy changes of other subjects [4]. The different strategy combinations led to varying structures and functions of the SNEM, resulting in various emergency management results [5]. Analyzing different types of the SNEM and cracking the internal logic of the formation of the “black box” are conducive to better grasping the formation law, building the SNEM that highly matches the emergency scenarios, and realizing the coincidence of emergency demands and network functions. At present, there are few studies on the SNEM, especially on the generation mechanism, subject game, and strategy selection of the SNEM. On this basis, taking China as an empirical sample, aiming at the governance scenarios under magnitude emergencies, this paper uses the evolutionary game model to study the game process and strategy selection among the government, social organizations, and the public in the SNEM. The research attempted to answer the following questions: (1) Can the SNEM be divided into several types, and what is the basis for its division? (2) What are the scenarios and generation mechanism of cooperation among subjects in the SNEM? (3) What are the game processes in the SNEM, and how do emergency subjects make their selections? Through the investigation of these problems, this paper aims to explore the game law among subjects in the SNEM, analyze the strategy combination and evolutionary path of different emergency mobilization, and seek the optimal solution of strategy selection, so as to provide theoretical guidance and practical support for the strategy selection of emergency mobilization.

2. Literature Review

At present, the studies related to social network are mainly divided into two categories: social network as research methods or research objects; this paper is the latter. The studies on social network as the object are mainly concentrated in community governance, emergency management, and public services [6–10]. In emergency management, few people have directly focused on the SNEM, but the studies related to social network are relatively rich and mainly focus on the construction, types, functions, and organizational analysis of social network.

2.1. Emergence of Social Network. As regards the emergence, social networks are a form of governance that coordinate and cooperate with each other based on certain conditions under the double superposition of the nature of collective action and internal and external factors of the actors and their relationship attributes [11]. The essence of collective

action is the interdependence of tasks and resources between network organizations [5]. The willingness to cooperate between actors is affected by homogeneity, geographic proximity, and trust [12–15]. In public and nonprofit sectors, the generating conditions are usually in the form of controlling public resources, whereas, in the private sector, they are usually in the form of signing formal contracts, and the network is usually structured by a top-down approach [16].

2.2. Classification of Social Network. As regards the classification of social network under different standards, based on their stability, governance model, hierarchical structure, and construction procedures, social network can be divided into different types. According to the stability and the willingness of cooperation among their actors, social network can be divided into the incentive-compatible, stable, and comprehensive cooperation type or one-way pay, one-sided noncooperation type [4]. Social network can also be divided into the shared or participatory governance, governed leading organizations, and governed network management organizations based on the governance model, with each network performing different functions. Next, social network is also divided into the intraorganizational type, interorganizational type, and cross-level type based on network boundaries and the closeness of connections between different types of organizations [1]. Then, according to the construction procedure, social network is also divided into the top-down type and bottom-up type. The former usually exists in public and nonprofit sectors and is occasionally compulsory [8], whereas the latter is usually created informally by network members and is voluntary [17].

2.3. Role of Social Network. As regards the role, social network has the functions of reducing costs, creating social capital, promoting collective action, and enhancing public value. Social network helps in building trust, establishing reciprocal norms, and obtaining funds and other resources through information exchange and cooperation between organizations, thereby reducing transaction costs [18, 19]. As an integral part of social capital, the construction and development of social network strengthen the communication and assistance among different organizations and improve the stock of social capital by building trust [20]. Social network also promotes social capital while connecting actors with common interests, information, and skills necessary for organizational actions, promoting a lasting collective action [21–24]. However, several scholars have pointed out that social network also has disadvantages. Given the homogeneity and geographic proximity of social network, homogeneous clusters easily formed in the wider network, which is not conducive to cooperation and coordination between different organizations [5].

2.4. Organizations (Subjects) of Social Network. Organizations in the social network mainly include the government, social organizations, the public, and international institutions [9]. Each subject can act as the core

subject in the network, but usually, the government acts as the core subject of the network. Most of the studies on the role of organizations are measured and analyzed by structural indicators such as centrality and density [25]. When the organization is at the centre of the network, its structural position plays a more obvious role [26]. Organizations can achieve their specific goals by adjusting their own behavioral strategy (such as cooperation and non-cooperation) to change the network structures to achieve their specific goals [27], while the social network structure will, in turn, affect the subject behavior strategy [28]. This phenomenon occurs in emergency mobilization and is reflected in fields such as public participation [29].

Generally, the existing literature has two deficiencies. First, the existing studies mainly analyze the various effects of the network itself and the organizational structure embedded in the network from a static perspective and lack a dynamic perspective to analyze the construction mechanism of the network. Second, the lack of classification of the SNEM has led to the inability to choose effectively and reduced the effectiveness of emergency management. The possible innovations are embodied in two aspects. First, it uses game theory to analyze the strategy selection of different subjects in the network and discusses the operation mechanism of the SNEM. Second, it puts forward the classification standard, analyzes four types of the SNEM, and points out the best type.

3. Types and Generation Mechanism of the SNEM in Magnitude Emergencies

In the face of “black swan” incidents, the public’s emergency awareness is awakened, causing an emergency force to emerge with individuals or organizations as the unit and forming the SNEM under the corresponding emergency mobilization mechanism. However, given the differences in the collective action attribute of emergency management and the interest orientation of multiple subjects, different situations arise regarding the willingness to cooperate and strategy selections between the subjects and objects, thereby forming different types of the SNEM. Taking the path of social mobilization for emergency as the first-level indicator, and the types of participants and the willingness to cooperate among different subjects as the second-level indicators, the SNEM is divided into four categories: symbiotic type or conflict type under the top-down mobilization, and binary type or discrete type under the bottom-up mobilization. This paper only discusses the willingness to cooperate between the government and social organizations or the public, which is used as the basis for the classification of the SNEM, without considering the willingness to cooperate between the public and social organizations.

3.1. SNEM under the High-Intensity Mobilization Path. In the complex governance scenarios, the government conducts the high-intensity emergency social mobilization for the public and social organizations to form the high-density and integrated SNEM among the government, social

organizations, and the public. Under the high-intensity mobilization path, the SNEM is divided into two types according to the willingness of social organizations and the public: symbiotic type or conflict type.

3.1.1. Symbiotic SNEM. The symbiotic SNEM means that, under the high-intensity emergency mobilization by the government, social organizations and the public respond positively, mobilize their own, and cooperate positively with the government’s commands and coordination, forming the “one nuclear and multiple” SNEM with government-centric, social organization, and the public-participation. The network has the characteristic of symbiosis, which means that, in the SNEM, the development and conduct of activities of each subject depend on the assistance and support of the other subjects, and the subjects have a mutual dependence and joint promotion relationship (Figure 1). The role of the government in the symbiotic SNEM mainly includes the makers of emergency decision, the trustee of emergency services, and the commander and coordinator of emergency work. In emergencies, the dilemma of “government failure” and the complexity of the governance scenarios cause difficulty for the government to resist disasters on its own. However, the government can use its own advantages and power to make rapid and scientific emergency decisions. The government also entrusts the production and provision of emergency services to share its burden and save emergency costs and maintain the emergency order through the command and coordination of multiple emergency subjects [30]. The public’s response to the government’s mobilization in emergency management mainly includes independent participation and organized participation. The public’s independent participation is achieved mainly through emergency social mobilization to enable the public to consciously cooperate with the government and encourage the public to form or change certain values, attitudes, and expectations. The public’s organized participation occurs mainly through the community, which is a public self-government platform and the basic unit of public life, to respond to emergency mobilization and conduct emergency management in an organized manner under the leadership of the community’s party committee and with the help of social organizations.

The forms of social organizations responding to the government’s mobilization in emergency management mainly include assisting the government in decision-making, connecting the government and the public, collecting and distributing materials. As an intermediate force between the government and the public, social organizations help the government realize the matching of emergency information supply and public emergency information demand [31, 32]. Through the analysis of the responsibility and role division of the subjects in the symbiotic SNEM, it is not difficult to see that, in the SNEM, the subjects are interdependent and supported, showing the characteristics of high density and strong adhesion.

3.1.2. Conflict SNEM. The conflict SNEM means that, under the high-intensity emergency mobilization by the

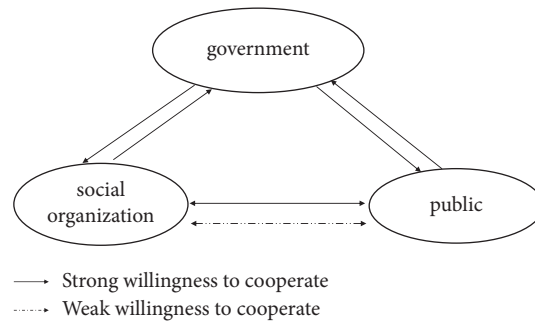


FIGURE 1: Schematic diagram of the symbiotic SNEM.

government, social organizations and the public, or one of them, have adopted the strategy of negative response. Under this path, the government strives to establish cooperative relations with social organizations for emergency cooperation through efforts such as empowerment and coordination and mobilizes the public to participate through media propaganda and offline calls. However, social organizations and the public are affected by different interest orientations and fail to respond to the government's mobilization positively. The network also has the characteristics of conflict, because the low willingness to cooperate leads to the goal conflict among subjects. Based on the degree of willingness of social organizations and the public to respond to the government's mobilization, the conflict SNEM can be divided into three types (Figures 2(a) to 2(c)).

- (1) When social organizations respond positively, the public's willingness to respond is not high. At this point, social organizations use their professional advantages to assist the government positively in the emergency decision-making, use their grassroot advantages to deepen the "last mile" for emergency communication and emergency information searching, and fully use the social organization resource system for emergency supplies. In the context of emergency management of magnitude emergencies, it is generally necessary for the public to change their previous life and work style to improve the emergency efficiency and quality quickly. However, the public often habitually wants to maintain the original life and work style, and the enthusiasm and initiative to make changes are not high, so it will also show a low willingness to respond to the government's mobilization (Figure 2(a)).
- (2) When the public responded positively, social organizations were not willing to respond. At this point, the public restrict their own behavior, cooperate positively with the government's emergency decisions, and participate in the emergency management in the form of individuals and organizations by means of donations and acting as volunteers. However, social organizations show a low willingness to cooperate or take negative actions to ignore the crisis or act according to their own will to carry out emergency work (Figure 2(b)).

- (3) When the public and social organizations' willingness to respond is low, at this point, the government helps social organizations and the public cooperate in emergency management by publishing information and providing resources quickly. However, due to the low willingness of social organizations and the public to respond to the government's mobilization, the social responsiveness is poor, easily leading to the dilemma of "fragmented" in emergency management. An information gap occurs between government and society, introducing difficulties in ensuring the quality of emergency response.

3.2. SNEM under the Low-Intensity Mobilization Path. Affected by various factors, such as the traditional government's "take on everything" emergency management and the weak awareness of emergency [8], the government also has a negative mobilization during magnitude emergencies. In this situation, the government has low responsiveness, and the public in emergency response is low. However, with the enhancement of emergency management capabilities in China, the establishment of a limited government, and the development and maturity of civil society, a bottom-up social automobilization model has gradually taken shape. Both play an increasingly important role in the risky society with complexity and turbulence. Therefore, under the low-intensity mobilization path, the SNEM can be divided into the binary type or discrete type according to the willingness to cooperate among the subjects.

3.2.1. Binary SNEM. The binary SNEM refers to the low willingness of the government, while social organizations and the public have a high willingness to cooperate, thereby forming the "government-society" binary SNEM (Figure 3). In this type, the government either responds negatively or follows the traditional administrative-controlled path to take care of all aspects of emergency work. Social organizations mobilize themselves positively, combine superior resources, and provide guidance and organizational guarantee for public participation in emergency work. Through active self-mobilization, the public form a high awareness of cogovernance and carry out orderly emergency work by assisting and participating in social organizations. Based on self-

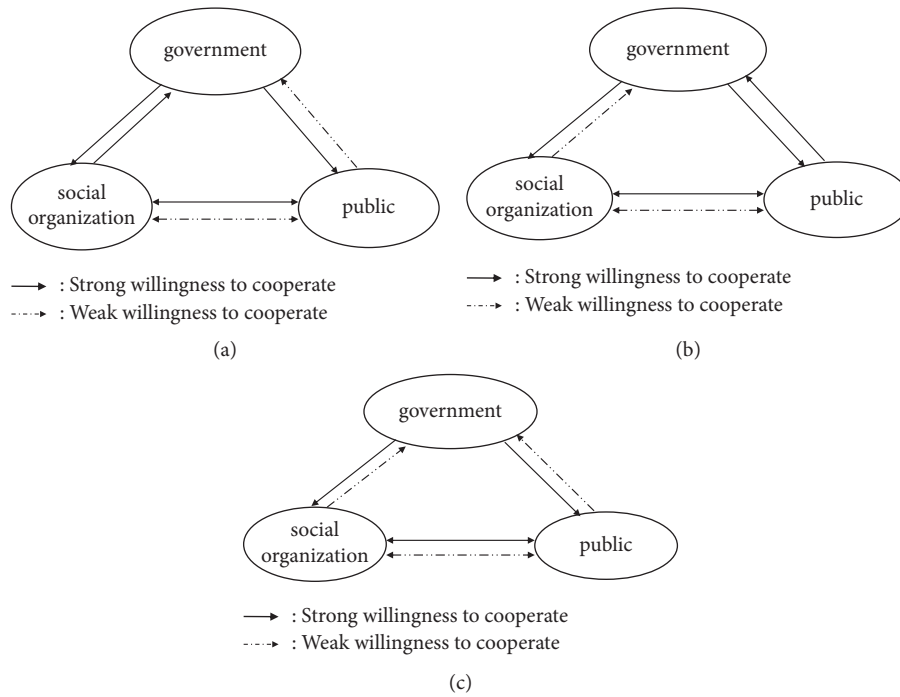


FIGURE 2: Schematic diagram of the conflict SNEM.

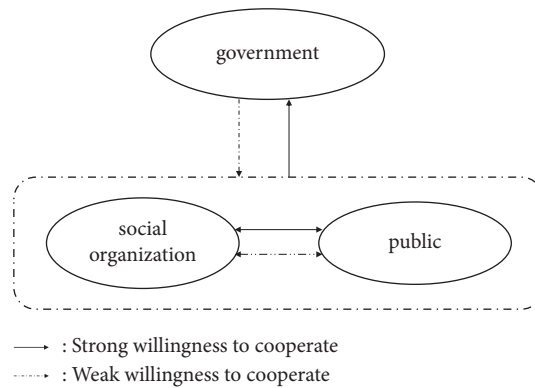


FIGURE 3: Schematic diagram of the binary SNEM.

mobilization, the public and social organizations carry out emergency work by establishing a cooperative partnership and assist and supervise the government’s emergency work.

3.2.2. *Discrete SNEM.* The discrete SNEM refers to the active self-mobilization by social organizations and the public, when the willingness of the government’s mobilize is low. In this case, the willingness to cooperate is low, and at most, a kind of subject seeks to cooperate with the government, leading to the loose cooperation among emergency subjects, forming the discrete SNEM (Figures 4(a) to 4(c)). This network presents discrete characteristics due to the high willingness to self-mobilize and low willingness to cooperate among subjects.

- (1) Social organizations seek to cooperate positively with the government, but the public’s willingness to cooperate is low. In this type, social organizations provide various resources for emergency positively by self-mobilization, but the government does not recognize it and shows low responsiveness. The public have low willingness to cooperate because of self-mobilization and are unwilling to sacrifice their own interests to cooperate with the government (Figure 4(a)).
- (2) The public seek to cooperate positively with the government, but the willingness of social organizations to cooperate is low. In this type, the public cooperate with the government positively based on self-mobilization and assist the government through

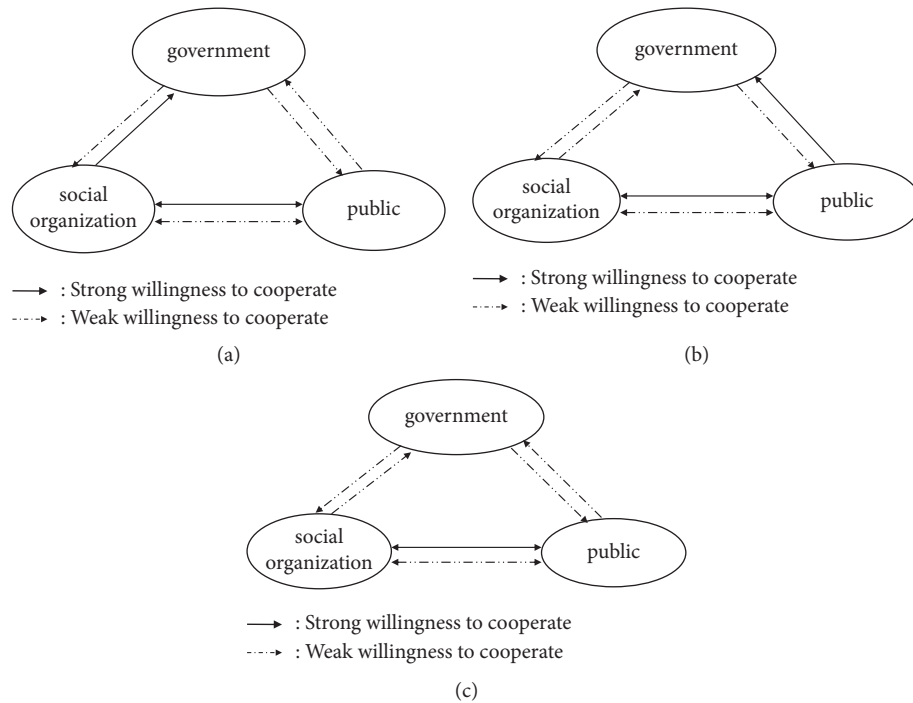


FIGURE 4: Schematic diagram of the discrete SNEM.

donations, advice, and suggestions. However, the government shows low responsiveness to these actions, whereas social organizations show distrust and refuse to cooperate with the government (Figure 4(b)).

- (3) The public, social organizations, and the government all show low willingness to cooperate. In this type, the government, social organizations, and the public failed to form a scientific and reasonable division of labor due to the lack of information exchange platform and mechanism. As a result, the government undertakes too many responsibilities and functions, social organizations cannot give full play to their professional advantages quickly, and the public cannot respond to emergency measures in time. So, there is no efficient cooperation pattern among subjects, which affects the efficiency and effectiveness of emergency management (Figure 4(c)).

3.3. Summary. In summary, according to the cooperation intensity among subjects, the SNEM is divided into four categories. In China, the SNEM will be affected by the interaction between the vertical network controlled by the central government and the horizontal network created by members, thus forming the multiple SNEM with the top-down or bottom-up. The purpose of constructing the SNEM is to promote multiple subjects to participate in the collective actions of emergency management. However, due to the heterogeneity of subjects and different interests, the cooperation in the SNEM presents the characteristics of limited rationality, and different subjects will make new strategy

selection according to the behavior changes of other subjects. Therefore, it is more realistic to use evolutionary game to analyze the cooperation of the government, social organizations, and the public in the SNEM (Table 1).

4. Evolution Game Analysis of the SNEM

4.1. Model Assumptions. In China, the social mobilization for magnitude emergencies is generally initiated by the government and responded or participated by social organizations and the public. In this process, the game subjects are the government, social organizations, and the public, which constitute the SNEM. It belongs to a typical social network. Given that the social mobilization for magnitude emergencies has a strong character with emergency and policy, the government must initiate and mobilize the participation of social organizations and the public. The function of social organizations determines that, in many cases, they need to respond positively to government and further mobilize other organizations and individuals. As the subject and object of mobilization, the public have realized the unity of subject and object. It is essential to participate in the social mobilization of magnitude emergencies. Under the premise that information is not completely symmetrical, all subjects in the SNEM are bounded rationality, and there are differences in interest pursuit, which constitutes the asymmetric game scenarios. Among them, the government aims to maximize social interests, while social organizations and the public pursue the maximization of their own interests. Therefore, how to optimize the strategy selection of subjects in asymmetric game scenarios to promote cooperation is essential to finding the optimal game path to overcome social dilemmas.

TABLE 1: Classification of the SNEM.

	Public	Social organizations			
		Response positively		Response negatively	
		Participate positively	Participate negatively	Participate positively	Participate negatively
Government	High mobilization intensity	Symbiosis	Conflict	Conflict	Conflict
	Low mobilization intensity	Binary	Discrete	Discrete	Discrete

Hypothesis 1. The three types of subjects in the SNEM adopt two strategies to participate in emergency mobilization. The government adopts the strategy with the high-intensity or low-intensity, the strategy set is {high-intensity, low-intensity}, which is denoted as (G_1, G_2) , and their probabilities are $x, 1 - x$ ($0 \leq x \leq 1$), respectively. The social organizations adopt the strategy with the positive response or negative response, the strategy set is {positive response, negative response}, which is denoted as (S_1, S_2) , and their probabilities are $y, 1 - y$ ($0 \leq y \leq 1$), respectively. The public adopt the strategy with the positive participation or negative participation, the strategy set is {positive participation, negative participation}, which is denoted as (P_1, P_2) , and their probabilities are $z, 1 - z$ ($0 \leq z \leq 1$), respectively.

Hypothesis 2. When the government adopts the high-intensity strategy, the costs paid are C_{11} (the labor cost caused by the high-intensity social mobilization and related subsidies or compensation used to mobilize other subjects), and the benefits obtained are R_{11} (the improvement of public satisfaction and the increase of income allocated by the superior government due to the positive mobilization effect.). It also includes spillover income (the motivation and enthusiasm for public participation caused by the positive mobilization effect have increased, the degree of cooperation with other government measures has increased significantly, the rewards given by the superior government, etc.). When the government adopts the low-intensity strategy, the costs paid are C_{12} (the increase of transaction cost caused by low-intensity mobilization, etc.), which also includes spillover costs (the public distrust and disapproval of the government caused by the failure to obtain a positive mobilization effect, resulting in a significant reduction in the degree of cooperation with other government measures, etc.), and the benefits obtained are R_{12} (general gains without positive mobilization effects, such as normal financial allocations). At this time, $C_{11} > C_{12}$, $R_{11} > R_{12}$.

Hypothesis 3. When the social organizations adopt the positive response strategy, the costs paid are C_{21} (the cost of mobilizing other organizations and individuals, such as publicity and personnel costs), and the benefits obtained are R_{21} (related subsidies given by the government; new

donations and other benefits attracted by social organizations after their reputation has increased, etc.). Given that social organizations respond positively and play a key role in responding to emergencies, the government will also provide certain rewards or subsidies V_{21} (if the government adopts the low-intensity strategy, it will prefer not to give rewards for various reasons). When the social organizations adopt the negative response strategy, the costs paid are C_{22} (the waste of resources caused by the failure of its own function and role due to the negative response, etc.), and the benefits obtained are R_{22} (normal general benefit). However, the social organizations adopt the negative strategy and will generally be punished by the government F_{21} (if the government adopts the low intensity strategy, they may turn a blind eye or take lighter punishments, such as formal persuasion and education, etc.). At this time, $C_{21} > C_{22}$, $R_{21} > R_{22}$.

Hypothesis 4. When the public adopt the positive participation strategy, the costs paid are C_{31} (the cost of positive participation in social mobilization, such as the opportunity cost caused by time and energy investment, etc.), and the benefits obtained are R_{31} (the psychological satisfaction brought by positive participation in social mobilization, the benefits brought by the restoration of normal production and life order, etc.). When the public adopt the negative participation strategy, the government will give certain subsidies or rewards W_{31} (if the government adopts the low intensity strategy, it will not grant subsidies for various reasons). When the public adopt the negative participation strategy, the costs paid are C_{32} (the losses caused by negative participation in social mobilization, etc.), and the benefits are R_{32} (saving time and other resources due to negative participation, etc.). At this time, $C_{31} > C_{32}$, $R_{31} > R_{32}$.

Table 2 shows the benefits of the government, social organizations, and the public because of these assumptions.

4.2. Evolutionary Game Analysis

4.2.1. Replicated Dynamics Equations of Game Subjects. Combined with Table 2, the expected benefits of each subject under different behavior strategies are as follows:

- (1) The expected benefits of the government adopting the high-intensity strategy:

TABLE 2: Game benefits matrix of the government, social organizations, and the public.

Public		Social organizations			
		Respond positively (y)		Respond negatively ($1 - y$)	
		Participate positively (z)	Participate negatively ($1 - z$)	Participate positively (z)	Participate negatively ($1 - z$)
Government	High-intensity (x)	$R_{11} - C_{11} - V_{21} - W_{31}$	$R_{11} - C_{11} - V_{21}$	$R_{11} - C_{11} - W_{31} + F_{21}$	$R_{11} - C_{11} + F_{21}$
		$R_{21} - C_{21} + V_{21}$	$R_{21} - C_{21} + V_{21}$	$R_{22} - C_{22} - F_{21}$	$R_{22} - C_{22} - F_{21}$
	Low-intensity ($1 - x$)	$R_{31} - C_{31} + W_{31}$	$R_{32} - C_{32}$	$R_{31} - C_{31} + W_{31}$	$R_{32} - C_{32}$
		$R_{12} - C_{12}$	$R_{12} - C_{12}$	$R_{12} - C_{12}$	$R_{12} - C_{12}$
		$R_{21} - C_{21}$	$R_{21} - C_{21}$	$R_{22} - C_{22}$	$R_{22} - C_{22}$
		$R_{31} - C_{31}$	$R_{32} - C_{32}$	$R_{31} - C_{31}$	$R_{32} - C_{32}$

$$\begin{aligned}
U_{g1} &= yz(R_{11} - C_{11} - V_{21} - W_{31}) + y(1 - z)(R_{11} - C_{11} - V_{21}) \\
&\quad + (1 - y)z(R_{11} - C_{11} - W_{31} + F_{21}) + (1 - y)(1 - z)(R_{11} - C_{11} + F_{21}) \\
&= -yV_{21} - zW_{31} + R_{11} - C_{11} - yF_{21} + F_{21}.
\end{aligned} \tag{1}$$

The expected benefits of the government adopting the low-intensity strategy:

$$\begin{aligned}
U_{g2} &= yz(R_{12} - C_{12}) + y(1 - z)(R_{12} - C_{12}) \\
&\quad + (1 - y)z(R_{12} - C_{12}) + (1 - y)(1 - z)(R_{12} - C_{12}) \\
&= R_{12} - C_{12}.
\end{aligned} \tag{2}$$

The average expected benefits of the government: $U_g = xU_{g1} + (1 - x)U_{g2}$.

According to the Malthusian model, the evolutionary game replicated dynamics equation of the government is as follows:

$$\begin{aligned}
F(x) &= \frac{dx}{dt} = x(U_{g1} - \bar{U}_g) \\
&= x(1 - x)[-y(F_{21} + V_{21}) - zW_{31} + R_{11} - C_{11} - R_{12} + C_{12} + F_{21}].
\end{aligned} \tag{3}$$

(2) The expected benefits of social organizations adopting the positive response strategy:

$$\begin{aligned}
U_{S1} &= xz(R_{21} - C_{21} + V_{21}) + x(1 - z)(R_{21} - C_{21} + V_{21}) \\
&\quad + (1 - x)z(R_{21} - C_{21}) + (1 - x)(1 - z)(R_{21} - C_{21}) \\
&= xV_{21} + R_{21} - C_{21}.
\end{aligned} \tag{4}$$

The expected benefits of social organizations adopting the negative response strategy

$$\begin{aligned}
U_{S2} &= xz(R_{22} - C_{22} - F_{21}) + x(1 - z)(R_{22} - C_{22} - F_{21}) \\
&\quad + (1 - x)z(R_{22} - C_{22}) + (1 - x)(1 - z)(R_{22} - C_{22}) \\
&= -xF_{21} + R_{22} - C_{22}.
\end{aligned} \tag{5}$$

The average expected benefits of social organizations:

$$\bar{U}_s = yU_{S1} + (1 - y)U_{S2}. \tag{6}$$

According to the Malthusian model, the evolutionary game replicated dynamics equation of social organizations is as follows:

$$\begin{aligned}
F(y) &= \frac{d_y}{d_t} = y(U_{s1} - \overline{U}_s) \\
&= y(1-y)[x(V_{21} + F_{21}) + R_{21} - C_{21} - R_{22} + C_{22}].
\end{aligned} \tag{7}$$

(3) The expected benefits of the public adopting the positive participation strategy:

$$\begin{aligned}
U_{p1} &= xy(R_{31} - C_{31} + W_{31}) + x(1-y)(R_{31} - C_{31} + W_{31}) \\
&\quad + (1-x)y(R_{31} - C_{31}) + (1-x)(1-y)(R_{31} - C_{31}) \\
&= xW_{31} + R_{31} - C_{31}.
\end{aligned} \tag{8}$$

The expected benefits of the public adopting the negative participation strategy:

$$\begin{aligned}
U_{p2} &= xy(R_{32} - C_{32}) + x(1-y)(R_{32} - C_{32}) \\
&\quad + (1-x)y(R_{32} - C_{32}) + (1-x)(1-y)(R_{32} - C_{32}) \\
&= R_{32} - C_{32}.
\end{aligned} \tag{9}$$

The average expected benefits of the public:

$$\overline{U}_p = zU_{p1} + (1-z)U_{p2}. \tag{10}$$

According to the Malthusian model, the evolutionary game replicated dynamics equation of the public is as follows:

$$\begin{aligned}
F(z) &= \frac{d_z}{d_t} = z(U_{p1} - \overline{U}_p) = z(1-z)(U_{p1} - U_{p2}) \\
&= z(1-z)(xW_{31} + R_{31} - C_{31} - R_{32} + C_{32}).
\end{aligned} \tag{11}$$

4.2.2. Evolution Stability Analysis of Game Subject Strategy. According to equations (1)–(11), the power system of the subject consisting of the government, social organizations, and the public is as follows:

$$\begin{cases}
\frac{d_x}{d_t} = x(U_{g1} - \overline{U}_g) = x(1-x)[-y(F_{21} + V_{21}) - zW_{31} + R_{11} - C_{11} - R_{12} + C_{12} + F_{21}], \\
\frac{d_y}{d_t} = y(U_{s1} - \overline{U}_s) = y(1-y)[x(V_{21} + F_{21}) + R_{21} - C_{21} - R_{22} + C_{22}], \\
\frac{d_z}{d_t} = z(U_{p1} - \overline{U}_p) = z(1-z)(xW_{31} + R_{31} - C_{31} - R_{32} + C_{32}).
\end{cases} \tag{12}$$

As shown in equation (12), the equilibrium point of the system is obtained and set as follows: $d_x/d_t = 0$, $d_y/d_t = 0$, $d_z/d_t = 0$. After solving, eight special equilibrium points can be obtained: (0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1).

The equilibrium point obtained by the replicated dynamics equation is not necessarily the stable strategy of system evolution. The stability of the equilibrium point must be analyzed by Lyapunov's stability theory, that is, judged by the eigenvalue of Jacobian matrix.

The derivation of x , y and z is for d_x/d_t , d_y/d_t and d_z/d_t , respectively, and the Jacobian matrices are as follows, to obtain the system equilibrium points and eigenvalues (Table 3).

According to the nature of evolutionarily stability strategy (ESS), the necessary condition for government to reach the evolutionary stability is $dF(x)/dx < 0$. According to the replicated dynamic equation (1), when $z = \{-(F_{21} + V_{21})y + R_{11} - R_{12} + C_{12} + F_{21} - C_{11}\}/W_{31}$, then $F(x) = 0$, so it is stable for all x ; when $z > \{-(F_{21} + V_{21})y + R_{11} - R_{12} + C_{12} + F_{21} - C_{11}\}/W_{31}$, then $dF(x)/d_x|_{x=0} < 0$, $dF(x)/d_x|_{x=1} > 0$, so $x = 0$ satisfies the necessary conditions and is the evolutionary stable point; when $z < \{-(F_{21} + V_{21})y + R_{11} - R_{12} + C_{12} + F_{21} - C_{11}\}/W_{31}$, then $dF(x)/d_x|_{x=0} > 0$, $dF(x)/d_x|_{x=1} < 0$, so $x = 1$ satisfies the necessary conditions and is the evolutionary stable point. According to these three cases, Figure 5(a) can be drawn briefly. The point in the plane is stable in the x -axis direction, the point above the plane will tend

TABLE 3: System equilibrium points and eigenvalues.

Equilibrium points	Eigenvalues		
	λ_1	λ_2	λ_3
$E_1: (0, 0, 0)$	$R_{11} - C_{11} - R_{12} + C_{12} + F_{21}$	$R_{21} - C_{21} - R_{22} + C_{22}$	$R_{31} - C_{31} - R_{32} + C_{32}$
$E_2: (0, 0, 1)$	$-W_{31} + R_{11} - C_{11} - R_{12} + C_{12} + F_{21}$	$R_{21} - C_{21} - R_{22} + C_{22}$	$-(R_{31} - C_{31} - R_{32} + C_{32})$
$E_3: (0, 1, 0)$	$R_{11} - V_{21} - C_{11} - R_{12} + C_{12}$	$-(R_{21} - C_{21} - R_{22} + C_{22})$	$R_{31} - C_{31} - R_{32} + C_{32}$
$E_4: (0, 1, 1)$	$-V_{21} - W_{31} + R_{11} - C_{11} - R_{12} + C_{12}$	$-(R_{21} - C_{21} - R_{22} + C_{22})$	$-(R_{31} - C_{31} - R_{32} + C_{32})$
$E_5: (1, 0, 0)$	$-(R_{11} - C_{11} - R_{12} + C_{12} + F_{21})$	$V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21}$	$W_{31} + R_{31} - C_{31} - R_{32} + C_{32}$
$E_6: (1, 0, 1)$	$W_{31} - R_{11} + C_{11} + R_{12} - C_{12} - F_{21}$	$V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21}$	$-(W_{31} + R_{31} - C_{31} - R_{32} + C_{32})$
$E_7: (1, 1, 0)$	$V_{21} - R_{11} + C_{11} + R_{12} - C_{12}$	$-(V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21})$	$W_{31} + R_{31} - C_{31} - R_{32} + C_{32}$
$E_8: (1, 1, 1)$	$V_{21} + W_{31} - R_{11} + C_{11} + R_{12} - C_{12}$	$-(V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21})$	$-(W_{31} + R_{31} - C_{31} - R_{32} + C_{32})$

to $x = 0$, and the point under the plane will tend to $x = 1$. Similarly, for the social organizations, in plane V , that is, the point in $x = C_{21} + R_{22} - C_{22} - R_{21}/V_{21} + F_{21}$ is stable in the y -axis direction, the point on the left side of plane V will tend to $y = 0$, and the point on the right side of plane V will tend to $y = 1$, as shown in Figure 5(b). For the public, in plane W , that is, the point in $x = C_{31} - C_{32} - R_{31} + R_{32}/W_{31}$ is stable in the z -axis direction, the point on the left side of plane W will evolve towards $z = 0$, and the point on the right side of plane W will evolve towards $z = 1$, as shown in Figure 5(c).

4.3. Analysis on the Behavior Strategy of the Tripartite Game.

$E_1: (0, 0, 0)$ when $\lambda_1 = R_{11} - C_{11} - R_{12} + C_{12} + F_{21} < 0$, $\lambda_2 = R_{21} - C_{21} - R_{22} + C_{22} < 0$ and $\lambda_3 = R_{31} - C_{31} - R_{32} + C_{32} < 0$, so $(0, 0, 0)$ is balanced and stable. At this time, the government, social organizations, and the public adopt the low intensity, negative response, and low intensity participation strategy, respectively, which is the most unsatisfactory behavior strategy, corresponding to the discrete SNEM (Figure 4(c)). In the game process, the government fails to adopt the effective social mobilization due to various scruples or limited energy in dealing with magnitude emergencies and could not effectively mobilize other subjects. This will directly lead to the termination of negotiation and maintain the status quo and then lead to the effect of social mobilization being not obvious. Thus, the government needs to reformulate the rules of the game and play again to prompt other subjects to change their strategies. If magnitude emergencies worsen, social organizations and the public may also switch strategies to form the bottom-up type social mobilization, which is different from the type of government-led mobilization, that is, strategies E_2 , E_3 and E_4 . However, due to the low intensity strategy adopted by the government at this time, under the constraints of economic rationality, the government will also choose not to mobilize positively or even resist. This type of strategy transformation is the most difficult and depends on a series of institutional conditions.

$E_2: (0, 0, 1)$: when $\lambda_1 = -W_{31} + R_{11} - C_{11} - R_{12} + C_{12} + F_{21} < 0$, $\lambda_2 = R_{21} - C_{21} - R_{22} + C_{22} < 0$ and $\lambda_3 = -(R_{31} - C_{31} - R_{32} + C_{32}) < 0$, so $(0, 0, 1)$ is balanced and stable. At this time, the public, social organizations, and the government adopt the positive participation, negative response, and

low-intensity strategy, respectively, which correspond to the discrete SNEM (Figure 4(b)). This type of social network corresponds to the subject strategy, and the public is in a relatively passive position; they cannot obtain the government support nor the favor of social organizations. If effective measures are implemented in time, the public will quickly turn to strategy E_1 , that is, the worst state. If the community where the public is located or the private group formed can select opinion leaders to participate in lobbying based on full consideration of public opinions, the public may also turn to strategy E_4 , but turning to strategies E_8 or E_6 is more difficult.

$E_3: (0, 1, 0)$: when $\lambda_1 = R_{11} - V_{21} - C_{11} - R_{12} + C_{12} < 0$, $\lambda_2 = -(R_{21} - C_{21} - R_{22} + C_{22}) < 0$ and $\lambda_3 = R_{31} - C_{31} - R_{32} + C_{32} < 0$, so $(0, 1, 0)$ is balanced and stable. At this time, social organizations, the public, and the government adopt the positive response, negative participation, and low-intensity strategy, respectively, which corresponds to the discrete SNEM (Figure 4(a)). This type of social network corresponds to the subject strategy, where social organizations are in a relatively passive position; that is, they cannot obtain the support by the government nor can they gain public trust. If effective measures are implemented in time, they will quickly turn to strategy E_1 , which is the worst state. However, if social organizations use their professional advantages to lobby through efforts such as propaganda and display to obtain support from the government, they will turn to the more favorable strategy E_7 . If social organizations can continue to persuade the public with the government through practices such as preaching and persuasion, warning them about the importance and necessity of participating in magnitude emergencies, and mobilize them to switch strategy and participate positively, they will turn to the most ideal strategy E_8 to maximize the benefits of cooperation.

$E_4: (0, 1, 1)$: when $\lambda_1 = -V_{21} - W_{31} + R_{11} - C_{11} - R_{12} + C_{12} < 0$, $\lambda_2 = -(R_{21} - C_{21} - R_{22} + C_{22}) < 0$ and $\lambda_3 = -(R_{31} - C_{31} - R_{32} + C_{32}) < 0$, so $(0, 1, 1)$ is balanced and stable. At this time, social organizations, the public, and the government adopt the positive response, positive participation, and low-intensity strategy, respectively, which correspond to the dual-type SNEM (Figure 3). This type of social network corresponds to the subject strategy, where social organizations and the public reach a consensus and adopt a positive attitude to participate in social mobilization. This mobilization can often achieve remarkable success in

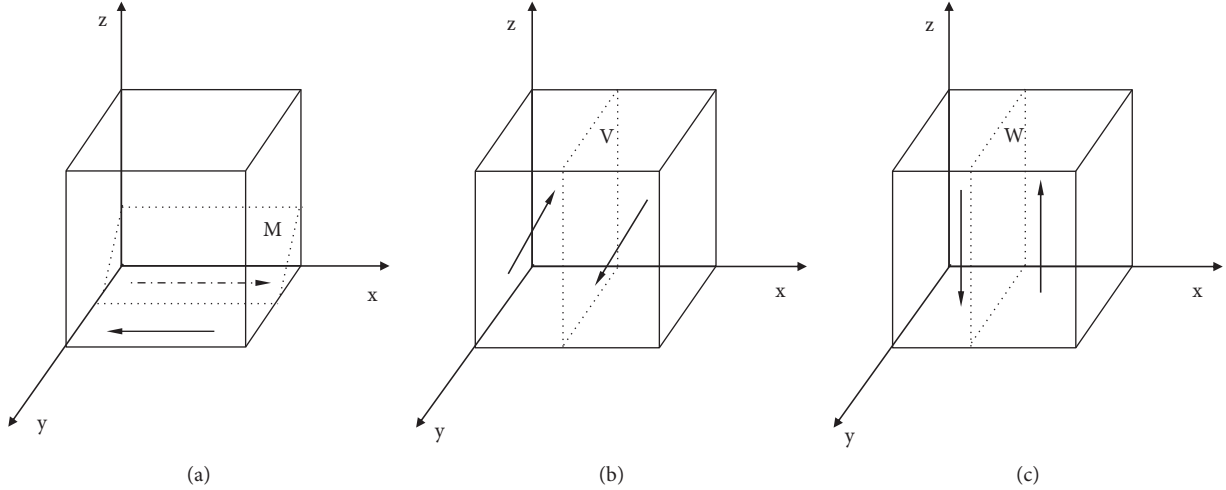


FIGURE 5: Evolutionary process of government, social organizations, and the public. (a) Evolution process of government strategy. (b) Evolution process of social organizations strategy. (c) Evolution process of the public strategy.

communities with a high degree of autonomy. If the government adopts the low-intensity strategy, the emergency social mobilization is often interrupted or stopped due to the lack of government support and promotion, which is generally difficult to continue. However, social organizations and the public will also negotiate with the government, aiming to persuade the government through various means to achieve the transition to the most ideal strategy E_8 .

E_5 : (1, 0, 0): when $\lambda_1 = -(R_{11} - C_{11} - R_{12} + C_{12} + F_{21}) < 0$, $\lambda_2 = V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21} < 0$ and $\lambda_3 = W_{31} + R_{31} - C_{31} - R_{32} + C_{32} < 0$, so (1, 0, 0) is balanced and stable. At this time, the government, social organizations, and the public adopt the high-intensity, negative response, and negative participation strategy, respectively, which correspond to the conflict SNEM (Figure 2(c)). At this time, the government promotes positively, but the attitude of social organizations and the public is extremely negative and may result in their nonparticipation. The possible reasons are as follows: on the one hand, the government does not publicize the importance and necessity of emergency social mobilization, participation in the mobilization requires considerable public energy, and the public assumes a negative attitude; on the other hand, social organizations perceive the noncooperation of the public or their negative attitudes and may have trouble in participating in social mobilization continually. Thus, they may also decide not to participate in the strategy. In this game process, given that both parties adopt a negative attitude of participation, this type of cooperation is often difficult to achieve without the help of administrative authority.

E_6 : (1, 0, 1): when $\lambda_1 = W_{31} - R_{11} + C_{11} + R_{12} - C_{12} - F_{21} < 0$, $\lambda_2 = V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21} < 0$ and $\lambda_3 = -(W_{31} + R_{31} - C_{31} - R_{32} + C_{32}) < 0$, so (1, 0, 1) is balanced and stable. At this time, the government, the public, and social organizations adopt the high-intensity, positive participation, and negative response strategy, respectively, which correspond to the conflict SNEM (Figure 2(b)). This type of social network corresponds to the subject strategy,

where the government and the public reach a consensus on emergency social mobilization, which is generally manifested as movement mobilization. However, at this point, social organizations have low enthusiasm for participation; they even refuse to participate. Organizational participation costs are extremely high, no other capital injection is available, economic pressure is high, or certain behaviors are inconsistent with the organizations' tasks and missions, and they cannot be recognized by members of social organizations.

E_7 : (1, 1, 0): when $\lambda_1 = V_{21} - R_{11} + C_{11} + R_{12} - C_{12} < 0$, $\lambda_2 = -(V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21}) < 0$ and $\lambda_3 = W_{31} + R_{31} - C_{31} - R_{32} + C_{32} < 0$, so (1, 1, 0) is balanced and stable. At this time, the government, social organizations, and the public adopt the high-intensity, positive response, and negative participation strategy, respectively, which correspond to the conflict SNEM (Figure 2(a)). This type of social network corresponds to the subject game strategy, where the government and social organizations reach a consensus to adopt a wide range of mobilization in response to magnitude emergencies, and the public's negative participation strategy may affect the mobilization effect to a certain extent. However, given the authority and organizational advantages of the government and social organizations, some community members could be potentially mobilized to engage in public participation in a relatively short period, prompting them to shift quickly and turn to the most ideal strategy E_8 .

E_8 : (1, 1, 1): when $\lambda_1 = V_{21} + W_{31} - R_{11} + C_{11} + R_{12} - C_{12} < 0$, $\lambda_2 = -(V_{21} + R_{21} - C_{21} - R_{22} + C_{22} + F_{21}) < 0$ and $\lambda_3 = -(W_{31} + R_{31} - C_{31} - R_{32} + C_{32}) < 0$, so (1, 1, 1) is balanced and stable. At this time, the government, social organizations, and the public adopt the high-intensity, positive response, and positive participation strategy, respectively, which correspond to the symbiotic SNEM (Figure 1). This type of social network corresponds to the subject strategy, where the three subjects reach a consensus. In response to magnitude emergencies, all subjects are mobilized positively,

forming a situation of widespread mobilization of the entire society. Generally, it can quickly release the positive effect of emergency social mobilization in a short time. With the COVID-19 epidemic as an example, China has contributed Chinese experience and solutions to the world. The important aspects of such contributions are adopting a wide range of emergency social mobilization, selecting strategy E_8 , and achieving an ideal state.

5. Conclusion

Taking China as a sample, this paper analyzed the evolutionary game of the SNEM and explored the internal generation mechanism, subject game, and their strategy selections. In recent years, the outbreak of magnitude emergencies, such as the COVID-19, Ebola epidemic, and Zika virus, has caused severe damage to countries worldwide. In the face of frequent magnitude emergencies in the context of a risky society, countries all over the world are actively exploring new programs for efficient and timely emergency management. In this context, combined with Chinese practice, this paper analyzed the SNEM and its game and drew the following conclusion: (1) the SNEM, where the government, social organizations, and the public adopt the high-intensity, positive response, and positive participation strategy, respectively, is the most ideal network; (2) the SNEM, where the government, social organizations, and the public adopt the low intensity, negative response, and negative participation strategy, respectively, is the worst network; (3) however, different types of the SNEM are not fixed, and dynamic transformation can be realized by guiding the subject behavior strategy. Through the empirical analysis of the types, generation mechanism, subject behavior, and strategy selection of the SNEM in China, this paper seeks the optimal strategy of social mobilization matching the emergency scenarios based on utility maximization, in order to provide useful reference for emergency social mobilization in magnitude emergencies all over the world and contribute Chinese wisdom and China's plans to the world.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are extremely grateful to Ouyang Fan for her many helpful suggestions. This study was supported by the Fundamental Research Funds for the Central Universities (Grant no. 2021SKWF04) and Beijing Social Science Foundation (Grant no. 20GLC044).

References

- [1] P. Kenis and K. G. Provan, "Towards an exogenous theory of public network performance," *Public Administration*, vol. 87, no. 3, pp. 440–456, 2009.
- [2] N. Rui and O. Y. Fan, "Influence of social networks on citizens' willingness to participate in social governance: evidence from China," *Complexity*, vol. 2020, Article ID 3819514, 16 pages, 2020.
- [3] D. Nohrstedt, "Explaining mobilization and performance of collaborations in routine emergency management," *Administration & Society*, vol. 48, no. 2, pp. 135–162, 2016.
- [4] V. Bala and S. Goyal, "A noncooperative model of network formation," *Econometrica*, vol. 68, no. 5, pp. 1181–1229, 2000.
- [5] A. Sapat, A.-M. Esnard, and A. Kolpakov, "Understanding collaboration in disaster assistance networks: organizational homophily or resource dependency?" *The American Review of Public Administration*, vol. 49, no. 8, pp. 957–972, 2019.
- [6] J. S. Coleman, "Social capital in the creation of human capital," *American Journal of Sociology*, vol. 94, pp. 95–120, 1988.
- [7] E. W. Lee and H. K. Liu, "Factors influencing network formation among social service nonprofit organizations in Hong Kong and implications for comparative and China studies," *International Public Management Journal*, vol. 15, no. 4, pp. 454–478, 2012.
- [8] K. G. Provan and K. Huang, "Resource tangibility and the evolution of a publicly funded health and human services network," *Public Administration Review*, vol. 72, no. 3, pp. 366–375, 2012.
- [9] K. G. Provan, K. Huang, and H. B. Milward, "The evolution of structural embeddedness and organizational social outcomes in a centrally governed health and human services network," *Journal of Public Administration Research and Theory*, vol. 19, no. 4, pp. 873–893, 2009.
- [10] J. P. Breimo, H. Turba, O. Firbank, I. Bode, and J. T. Sandvin, "Networking enforced-comparing social Services' collaborative rationales across different welfare regimes," *Social Policy and Administration*, vol. 51, no. 7, pp. 1348–1366, 2017.
- [11] D. Nohrstedt and Ö. Bodin, "Collective action problem characteristics and partner uncertainty as drivers of social tie formation in collaborative networks," *Policy Studies Journal*, vol. 48, no. 4, pp. 1082–1108, 2020.
- [12] Y. C. Atouba and M. Shumate, "International nonprofit collaboration," *Nonprofit and Voluntary Sector Quarterly*, vol. 44, no. 3, pp. 587–608, 2015.
- [13] J. A. Musso and C. Weare, "From participatory reform to social capital: micro-motives and the macro-structure of civil society networks," *Public Administration Review*, vol. 75, no. 1, pp. 150–164, 2015.
- [14] E. R. Gerber, A. D. Henry, and M. Lubell, "Political homophily and collaboration in regional planning networks," *American Journal of Political Science*, vol. 57, no. 3, pp. 598–610, 2013.
- [15] M. Granovetter, "Economic Action and Social Structure," *The Problem of Embeddedness*, Routledge, London, UK, 2018.
- [16] K. G. Provan and P. Kenis, "Modes of network governance: structure, management, and effectiveness," *Journal of Public Administration Research and Theory*, vol. 18, no. 2, pp. 229–252, 2008.
- [17] N. Lin, Y. C. Fu, and R. M. Hsung, "The Position Generator," *Measurement Techniques for Investigations of Social Capital*, Routledge, London, UK, 2001.
- [18] N. Lin, "A network theory of social capital," *The Handbook of Social Capital*, Oxford University Press, Oxford, UK, 2008.

- [19] R. D. Putnam, R. Leonardi, and R. Y. Nanetti, *Making Democracy Work*, Princeton University Press, Princeton, NJ, USA, 1994.
- [20] A. Rosenberg, K. Hartwig, and M. Merson, "Government-NGO collaboration and sustainability of orphans and vulnerable children projects in southern Africa," *Evaluation and Program Planning*, vol. 31, no. 1, pp. 51–60, 2008.
- [21] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.
- [22] P. E. Oliver and G. Marwell, "The paradox of group size in collective action: a theory of the critical mass. II," *American Sociological Review*, vol. 53, no. 1, pp. 1–8, 1988.
- [23] R. D. Putnam, *Bowling Alone: The Collapse and Revival of American Community*, Simon & Schuster, New York, NY, USA, 2000.
- [24] R. D. Putnam, L. Feldstein, and D. J. Cohen, *Better Together: Restoring the American Community*, Simon & Schuster, New York, NY, USA, 2004.
- [25] M. D. Siciliano and J. R. Thompson, "If you are committed, then so am I: the role of social networks and social influence on organizational commitment," *Administration and Society*, vol. 50, no. 7, pp. 916–946, 2018.
- [26] I. Luxton and J. Sbicca, "Mapping movements: a call for qualitative social network analysis," *Qualitative Research*, vol. 21, no. 2, pp. 161–180, 2021.
- [27] N. Holman, "Community participation: using social network analysis to improve developmental benefits," *Environment and Planning C: Government and Policy*, vol. 26, no. 3, pp. 525–543, 2008.
- [28] C. Wukich, M. D. Siciliano, J. Enia, and B. Boylan, "The formation of transnational knowledge networks on social media," *International Public Management Journal*, vol. 20, no. 3, pp. 381–408, 2017.
- [29] S. Schnell and J. Saxby, "Mobilizing against hunger and poverty: capacity and change in a Brazilian social mobilization network," *Public Administration and Development*, vol. 30, no. 1, pp. 38–48, 2010.
- [30] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2018.
- [31] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [32] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2019.

Research Article

Social Network Structure as a Moderator of the Relationship between Psychological Capital and Job Satisfaction: Evidence from China

Fan Gu  and Yuanyuan Xiao

Business School, China University of Political Science and Law, Beijing 100088, China

Correspondence should be addressed to Fan Gu; gufan@cupl.edu.cn

Received 10 June 2021; Accepted 8 October 2021; Published 2 November 2021

Academic Editor: Fei Xiong

Copyright © 2021 Fan Gu and Yuanyuan Xiao. 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.

Although networking is reported to be a job search strategy in the literature, research on the interaction between social networking and other personal resources and its effect on job satisfaction is scarce. In the perspective of social networks, the present study explored whether the social network structure, which consists of network size and tie strength, moderates the relationship between psychological capital and job satisfaction. By using a two-wave longitudinal design, we collected the quantitative data (survey of 344 undergraduate students who were about to graduate soon) from 19 universities in Beijing city, Shandong Province, and Jiangsu Province in Eastern China. Factor analysis and hierarchical regression analysis were adopted to analyze the data of the survey. We found that psychological capital has a positive impact on job seekers' job satisfaction. Furthermore, smaller networks and weaker ties in social networks both render the positive effect of psychological capital on job satisfaction even stronger.

1. Introduction

Social networks, an essential component of social capital, refer to interpersonal relationships, which are built, developed, and maintained through stable ties and frequent interactions [1]. The previous research had divided social networks mainly into two typical categories: formal and informal networks [2]. Although social networks are reported to be universally valid, it is still found that social networks in China have unique characteristics compared to their counterparts in Western countries. Social networks in East Asia are influenced by local culture [3] and become a dominant norm in Chinese society [4]. Thus, due to the remarkable intraregional differences, scholars even adopted *guanxi* to describe the informal social network in Chinese society. Contrary to social networks in Western countries, weak ties are found neither effective nor preferable in Eastern Asia, while strong ties are more accepted and powerful [5]. Hence, the yawning divergence between the characteristics of social networks in China and the West is obvious.

Whether in China or in the West, a large number of fresh graduates are confronted with difficulty in seeking employment in the last year of their student lives. Chinese national government pays great attention to this issue and publishes reports to announce the rate of unemployment, which partly represents the economic prosperity of a country, every year. From the last year to the present, the coronavirus (COVID-19) pandemic leads to an increase in the rate of unemployment all over the world. As a matter of fact, a large number of studies showed that unemployment is harmful to individuals' mental and physical health and decreases satisfaction with their lives [6]. Therefore, the research had concentrated on the exploration of the relationship between individuals' job search behaviors and employment quality, such as job-organization fit, job satisfaction, and turnover intention [7–9], in order to clarify what factors in the job search process could eventually increase the individuals' opportunities of obtaining job offers and then enhance their performance [10].

Job satisfaction is defined by Locke [11] as "a pleasurable or positive emotional state resulting from the appraisal of

one's job or job experiences." Obviously, Locke's definition consists of two important individuals' reactions: affects and cognition. As such, when individuals evaluate their jobs, both thinking and feeling are involved in the whole process. It is widely discussed that the determinants of job satisfaction primarily consist of two types: job-specific characteristics, such as wages, labor market status, and observable attributes of the current job, and worker-specific characteristics, such as workers' education background, health, and worker's perceptions of the match between education and employment [12].

The literature regarding job search behaviors and process also paid great attention to job satisfaction, which refers to job hunters' satisfaction with the outcome of the job-seeking behaviors, i.e., the job they finally obtained [13]. Scholars found that both of the job seekers' internal factors, such as psychological capital and social capital and human capital, and external factors, such as labor market and national employment policy, can influence their job satisfaction. Among all these factors, it is reported that the subjective factors, which are considered as job seekers' internal motivation to promote them to make great endeavors in the job-hunting process, influence their attitudes and behaviors to a large extent [14]. It is reported that a large amount of job seekers in China always experiences a long job-hunting process, which lasts from the beginning of the autumn to the end of the spring [15]. Moreover, the difficulties for all job seekers in finding a satisfactory job had been increased, due to the huge supply of fresh graduates in the Chinese labor market in recent years [15]. Thus, to help job seekers enhance their employment opportunities, it is essential to pay attention to their behaviors in the Chinese context.

Networking is an important source for job seekers to gather information regarding employment opportunities. A popular book, which introduces the techniques of job hunting, also suggests that individuals could obtain useful information on job vacancies by contacting the important persons who can help in their social networks [16]. This echoes the research on recruitment, which claimed that knowing the job opportunities through one's social networks always has a positive impact on employees' behaviors [17, 18]. However, the previous research on the role of networking in the job search process primarily focused on individual difference determinants of networking [19] and the impact of networking on employment quality and outcomes [20]. It is obvious that the literature ignores the moderation effect of social networks on the relationship between other important internal factors (e.g., psychological capital) in the job search process and the quality of employment (e.g., job satisfaction). Thus, this study examines the moderation effect of social networks on the relationship between psychological capital and job satisfaction by responding to the following question:

How do social networks moderate the impact of psychological capital on job satisfaction?

In fact, this study takes an interactionist approach by bringing together individual variables (psychological capital) and contextual variables, that is, social network structure, to predict job satisfaction in a single frame for the first

time. To answer the question above, the dependent variable in this study is "job satisfaction," while the independent variable is "psychological capital." "Social network structure" involving network size and tie strength is examined as the moderator of the relationship between psychological capital and job satisfaction. The exploratory factor analysis is used to assess common method variance among variables while the hierarchical regression model is adopted to test the influence of psychological capital on job satisfaction and the moderative effect of network size and tie strength on the relationship between psychological capital and job satisfaction (Figure 1).

As both social networks and job seekers' behaviors have a uniqueness in Chinese society, this study extends previous relevant research by examining the moderation effect of social network structure on the relationship between psychological capital and job satisfaction in the Chinese context. Thus, we could better realize the important role of social networks in Chinese daily life. This is the main academic contribution of this study.

2. Literature Review and Hypotheses

2.1. Networking Behaviors in Job Search Process. Networking behaviors refer to "individuals' attempts to develop and maintain relationships with others who have the potential to assist them in their work or career" [21]. Networking behaviors are widely studied in various contexts in the management research field, because both organizations or individuals adopt social networks as a strategy on how to gain resources. In the context of the job search process, networking behaviors refer to individuals' proactive actions aiming at acquiring information, leads, or suggestions on obtaining a job through friends or other people in their social networks [22].

In the past twenty years, it is very popular to examine the characteristics of networking behaviors in the job search process. Zottoli and Wanous [23] defined networking as a typical informal job search behavior, in the perspective of formal-informal classification. That is to say, in contrast to formal job search behavior, which relies on formal intermediaries such as campus recruitment or employment agencies, which are built only for recruitment purposes, informal job search behaviors do not need the help from those formal intermediaries but contact the persons they are acquainted with for new opportunities. Saks and Ashforth [24] also made a great contribution to the characteristics of networking behaviors in the job search process, by stating that networking is a typical preparatory job search behavior, according to the preparatory-active classification. However, Hoyer et al. [20] pointed out that networking could be classified as both preparatory behaviors and active behaviors because the job search process could be divided into two stages. As such, at the beginning of the job search process, individuals experience a preparatory job search stage, in which they collect relevant information regarding potential jobs through networking. Following this, individuals will be actively contacting prospective employers in order to get a new job, which refers to active behaviors.

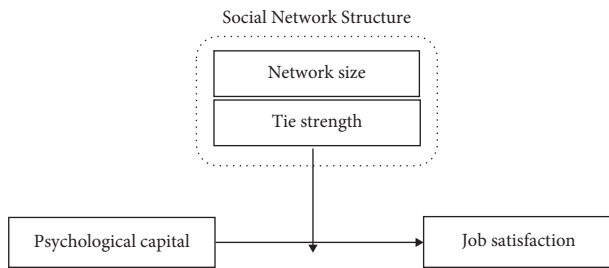


FIGURE 1: Model logic diagram.

Besides the research on characteristics of networking behaviors, scholars also paid great attention to the frequency of networking behaviors in the job search process. Kanfer et al. [10] used general job search intensity to describe the frequency of combined job search behaviors adopted by individuals in the process. The composition of general job search behaviors involves not only networking behaviors (e.g., turning to friends for help) but also other behaviors (e.g., through advertisements on newspapers, TV, or the Internet). The frequency and intensity of networking behaviors are reported to be positively related to job search outcomes [10]. Thus, the prior research considers networking as a specific job search behavior, differing from others, and this provides the theoretical insight for our current study and a sufficient approach to support our following research.

2.2. Psychological Capital and Job Satisfaction.

Psychological capital, which refers to “an individual’s positive psychological state of development [25],” consists of four psychological resources, i.e., “(1) having confidence (self-efficacy) to take on and put in the necessary effort to succeed at challenging tasks; (2) making a positive attribution (optimism) about succeeding now and in the future; (3) persevering toward goals and, when necessary, redirecting paths to goals (hope) in order to succeed; and (4) when beset by problems and adversity, sustaining and bouncing back and even beyond (resilience) to attain success [26].” As psychological capital is “state-like,” it is also malleable and open to development [27, 28]. A large number of studies had proved that psychological capital has positive impacts on various individuals’ attitudes or behaviors, such as performance, engagement, satisfaction, well-being, and citizenship behavior [29–31]. In the context of job search behaviors, the individuals with higher psychological capital are likely to be more confident with themselves in the job-hunting process and ultimately obtain more job offers, compared to the individuals with lower psychological capital. Subsequently, those individuals with higher psychological capital will be likely to be satisfied with their jobs. Thus, we predict the following.

Hypothesis 1. Psychological capital is positively related to job satisfaction.

2.3. Social Network Structure as a Moderator. In the literature, social network theorists emphasized the importance of the social network structure, which was defined as the formal

structure of the ties or relationships making up the social network [20], and proposed it as an essential source of social capital [32, 33]. On the basis of this point, we expect that social network structure will influence the relationship between psychological capital and job satisfaction. According to the prior studies, social network structure consists of two important components, i.e., network size and tie strength. First of all, network size refers to the total number of persons to whom an individual is tied [34]. According to Hoye et al. [20], network size is positively associated with time spent networking. The individuals, whose social network size is large, always need to spend more time on networking behaviors, because they need to contact more people in their networks. Thus, the job seekers with larger networks are likely to ask more persons for help and collect job-related information from them. The more people job seekers contact, the more job opportunities they could obtain. The more people job seekers ask for help, the greater progress they will make in the job search process.

Thus, in the context of large network size, job seekers are likely to rely heavily on the information or opportunities networks bring to them, which will weaken the effects of psychological capital on the job search process. In contrast, job seekers, who just have a small social network, can only have limited friends or acquaintances who can help and then need mainly to seek a job on their own. Therefore, in the context of a small network size, job seekers’ internal psychological capitals play a more significant role in the acquisition of job-related information and job offers. Thus, we propose the following.

Hypothesis 2. Network size moderates the relationship between psychological capital and job satisfaction, such that the relationship is more positive for job seekers with smaller social networks.

Besides network size, the other important element of social network structure is the strength of the ties in networks [2, 35, 36]. Tie strength refers to the closeness of the interpersonal relationship between the individual and the other persons in social networks [2]. For instance, parents, relatives, or close friends represent strong ties in social networks, while seldom-contacted friends could be an example of weak ties [20]. Reingen and Kernan [37] stated that individuals are always likely to contact those people with whom they are much more familiar to gather information regarding job vacancies in the job-hunting process. In contrast, the individuals have fewer interactions with those people they are not familiar with and could obtain little help from them. The previous research on marketing also claimed that individuals are more inclined to gather information from strong ties rather than weak ties [37, 38]. Hoyer et al. [20] collected empirical evidence to find out that individuals with strong ties in social networks always spend more time on networking behaviors in order to acquire adequate information when they are looking for a new job.

Based on these, in the context of strong ties, the job seekers will do endeavor to contact those close friends for help, which means that they are likely to rely heavily on networking behaviors and ignore the role of psychological

capital in the job search process. In contrast, job seekers with just weak ties could not obtain adequate information and enough help from their friends; thus, they are more likely to get job-related information from other sources (e.g., via employment advertising or job sites.) by themselves. In this situation, psychological capitals help them to have a positive and optimistic state to be confronted with difficulties and fierce competition in the job search process. Taken together, we predict the following.

Hypothesis 3. Tie strength moderates the relationship between psychological capitals and job satisfaction, such that the relationship is more positive for job seekers with weaker ties in their social networks.

3. Methods

3.1. Sample and Procedure. Our participants were university undergraduate students who soon graduate from 19 universities in Beijing city, Shandong Province, and Jiangsu Province, which are all located in Eastern China. Selection criteria for inclusion in the survey were that students were in the fourth year of their undergraduate studies and decided to work after graduation and had obtained at least one job offer till now. We then contacted the potential respondents and invited them to participate in the survey as a volunteer with the guarantee of confidentiality and anonymity. If the respondents agreed to participate, we asked them to choose to answer the questions via e-mail or we-chat (i.e., a free communication application in China), and the questionnaires were sent to them via the platform they chose.

To alleviate potential common method biases [39], survey data was collected in a two-wave longitudinal design. The literature showed the advantages of collecting data for more than one round of data within a long period [40]. As Gollob and Reichardt [41] stated, “no one time lag by itself can give a complete understanding of a variable’s effects.” Thus, the data was collected at two different time points within one month. At time 1, respondents completed measures of psychological capital, network size, and tie strength. One month later, at time 2, students were invited to finish the measurement of job satisfaction. In total, 403 students were invited to participate in the study, and accordingly, in time 2, a total of 378 questionnaires was acquired, yielding a response rate of 93.7%. Finally, 344 valid types of data were selected and adopted in this study.

3.2. Variable Designing and Measurement. All variables in this study were measured by adopting a five-point Likert scale, which ranges from 1 (strongly disagree) to 5 (strongly agree). Brislin’s [42] back-translation method was used to translate English items into Chinese. Some items are adapted to the uniqueness of Chinese characteristics and culture.

3.2.1. Independent Variable. The dependent variable is psychological capital. According to Luthans et al.’s [26] definition, psychological capital is categorized into four elements, self-efficacy, optimism, hope, and resilience. The

measurement developed by Luthans et al. [26] was adapted to fit the Chinese context and includes 24 items. Some sample items for psychological capital consist of the following: “I feel confident helping to set targets/goals in my work area” (self-efficacy); “I always look on the bright side of things regarding my job” (optimism); “If I should find myself in a jam at work, I could think of many ways to get out of it” (hope); “I usually manage difficulties one way or another at work” (resilience).

3.2.2. Dependent Variable. The dependent variable is job satisfaction. Job satisfaction was measured by three items, which were adapted from Bharati and Chaudary [43]. The first item was “the current job offers I got are what I expect.” The second item was “I hope that I can have a bright future in my job, which I choose from all my job offers.” The third item was “I believe that I make a right decision in choice of the job offers.”

3.2.3. Moderator. As discussed above, social network structure was chosen as a moderator in this study. According to Hoye et al. [20], social network structure consists of two components: network size and tie strength. As for the measurement of network size, 3 items adapted from Hoye et al. [20] were used. The first item is “I know a lot of people who might help me find a job.” The second item is “I can count on many relatives, friends, or acquaintances for information about jobs.” The third item is “I have connections I can talk to help me find a job.” On the other hand, three items, adapted from Hoye et al. [20], were also chosen for the measurement of tie strength. The first item is “most people who might help me find a job are people I know very well, such as family or friends.” The second item is “most people who might help me find a job are people I often talk to.” The third item is “most people who might help me find a job are people I feel comfortable talking to, even about touchy subjects”.

3.2.4. Control Variables. The control variables consist of gender, single child, student leader, CPC member, university classification, and major classification. We do not include age and education, as the participants of this study were all in their final year of undergraduate period, and have similar ages and same educational background. For education background, we also use university classification, which represents the level of the university and major classification, which represents different categories of major to distinct these respondents. The variable’ scaling and statistical description are as shown in Table 1.

3.3. Model. In the process of data analysis and hypotheses test, factor analysis and hierarchical regression model were used.

3.3.1. Factor Analysis. Factor analysis is a widely used explorative methodology of statistics, aiming at grouping similar variables into the same factor. It uses the data

TABLE 1: Control variable scaling and statistics.

Category	Scaling	No.	Ratio (%)
Gender	0 = male	131	38.1
	1 = female	213	61.9
Single child	1 = yes	211	61.3
	0 = no	133	38.7
Student leader	1 = yes	165	48
	0 = no	179	52
CPC member	1 = yes	141	41
	0 = no	203	59
University classification	1 = institute	4	1.2
	2 = 985 projects	62	18
	3 = 211 projects	131	38.1
	4 = normal university	147	42.7
Major classification	1 = humanity and social science	115	33.4
	2 = natural science	24	7
	3 = technology	62	18
	4 = management	129	37.5
	5 = others	14	4.1

correlation matrix to identify the latent variables. When conducting the analysis of data, factor analysis is always used to reduce the dimension of the dataset at the beginning. Normally, factors will be rotated after it has been extracted, in order to decrease the number of the variables in the regression model. Moreover, factor analysis is always adopted to cope with the problems of multicollinearity in regression analysis. In the following section, *Varimax*, as a popular rotation method, will be used to simplify the interpretation of each factor in the model. According to Pereira et al. [44], the steps conducted in exploratory factor analysis were stated as follows: “(1) collect data: choose relevant variables used in the regression model; (2) extract initial factors; (3) choose the number of factors to retain; (4) choose estimation method and estimate model; (5) rotate and interpret the results” [44].

According to [45, 46], the fundamental model for factor analysis is stated as follows:

$$\begin{aligned}
 X &= AF + B, \\
 x_1 &= a_{11}F_1 + a_{12}F_2 + a_{1m}F_m + \varepsilon_1, \\
 x_1 &= a_{21}F_1 + a_{22}F_2 + a_{2m}F_m + \varepsilon_2, \\
 x_1 &= a_{p1}F_1 + a_{p2}F_2 + a_{pm}F_m + \varepsilon_p, \\
 x_i &= a_{i1}F_1 + a_{i2}F_2 + a_{im}F_m + \varepsilon_i.
 \end{aligned} \tag{1}$$

In the model above [45, 46], X refers to the observable random vector while F refers to the common factor of X . A refers to the coefficient of F while B refers to the special factor of X . a_{ij} refers to the correlation coefficient while ε_i refers to the error factor.

To conduct factor analysis [45, 46], we need to normalize the X matrix, in order to ensure that the mean value equals 0 and the variance equals 1. Subsequently, F and ε_i should be assumed to be independent. Thirdly, we need to calculate the correlation coefficient matrix (i.e., $R = (rij)p * p$) and its latent root. Ultimately, the number of the factors should be

decided, and the common factors will be acquired by rotating the loading matrix.

3.3.2. Moderation Effect on Hierarchical Regression Analysis. Hierarchical regression analysis is a commonly used statistical methodology, which helps to quantify the relationships between different variables. In order to testify the relationship between the variables in the hypotheses above, it is necessary to conduct a hierarchical regression analysis to make sure whether the hypotheses are supported. In the process of a hierarchical regression model, we can decide what predictors should be entered into the model, and we also need to determine in what step we enter each variable. Normally, the logical and theoretical considerations will influence the order we enter each predictor.

The most important aim of this study is to decide the interaction of social network structure and psychological capital and its effect on job satisfaction. As such, we need to examine the moderation effect of social network structure via a hierarchical regression model. Hence, the model equation of the moderation effect is stated as follows:

$$Y = b_0 + b_1X + b_2W + b_3XW. \tag{2}$$

In the model, Y refers to the dependent variable while X refers to the independent variable. W refers to the moderator variable. Here, b_0 is the intercept; b_1 refers to the regression coefficient of X on Y ; b_2 refers to the regression coefficient of W on Y ; b_3 refers to the regression coefficient of the interaction between X and W on Y . The relationship between these variables is shown in Figure 2.

In order to calculate the conditional effects of the moderator, the model equation should be written as $Y = a + bX$. Hence, after grouping the terms to form $Y = a + bX$, we got the following equation model:

$$Y = (b_0 + b_2W) + (b_1 + b_3W)X. \tag{3}$$

Here, $b_1 + b_3W$ represents the direct effect of X on Y , conditional on W . In the following part of the hypotheses test, we conducted the hierarchical regression analysis by using SPSS 24.0. We will enter control variables, independent variable (i.e., psychological capital), dependent variable (i.e., job satisfaction), and moderation variables (i.e., network size and tie strength in the social network) into different steps in the model, in order to examine the moderation effect of network size and tie strength on the relationship between psychological capital and job satisfaction.

4. Results and Analysis

4.1. Analytical Strategy and Descriptive Statistics. In order to test the effect of psychological capital on job satisfaction and the moderation effect of social network structure on this relationship, the hierarchical regression model was adopted to test the hypotheses. In order to avoid multicollinearity, all variables in the study were grand-mean-centered. To make sure reliability and validity, relevant tests were also conducted. Firstly, the test results told us that the Cronbach

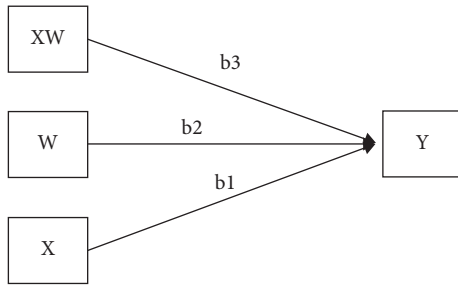


FIGURE 2: Model of moderation effect in hierarchical regression analysis.

Alpha Coefficient is 0.886, which shows that the internal consistency is great. Secondly, it is found that the Kaiser-Meyer-Olkin is 0.869, while the significance level of Bartlett's test of sphericity is 0.000, which shows a high level of validity (Tables 2 and 3). Means, standard deviations, and correlations among all variables are presented in Table 4.

4.2. Exploratory Factor Analysis. In this study, we collected the data for four major variables (i.e., psychological capital, job satisfaction, network size, and tie strength) from the same source, although it has been done two times within one month. This could lead to common method problems, which exist if only one factor emerged or if a single factor accounted for the majority of the variance among the variables. Thus, it is necessary to assess common method variance among these four variables. We conducted a principal component exploratory factor analysis, which is suggested by Podsakoff and Organ [47]. The results showed that four factors had eigenvalues greater than one and explained 79.6% of the variance (Table 5). According to Table 5, it is obvious that the factor loadings are all higher than 0.75, which indicated good internal reliability. Thus, these four variables will be all used in the following hypotheses tests.

4.3. Results of Hypotheses Test

4.3.1. Results of Test of Hypothesis 1. To test the three hypotheses proposed above, a hierarchical regression analysis was performed. To test Hypothesis 1, we created a two-step model by SPSS 24.0. We entered control variables and job satisfaction (dependent variable) in the first step. Subsequently, we added psychological capital (independent variable) into the equation in the second step. The results showed that R^2 significantly increased in step 2 (regression coefficient = 0.374, $p < 0.001$), compared to the counterpart in step 1, which represents the positive impact of psychological capital on job satisfaction and thus supports Hypothesis 1 (Table 6).

4.3.2. Results of Test of Hypothesis 2. To test the moderating effect of social network structure on the relationship between psychological capital and job satisfaction, hierarchical regression analysis was, respectively, performed for network

TABLE 2: The reliability of the survey.

Number of the items	Cronbach's α coefficient
33	0.886

size and tie strength. To begin with the test of Hypothesis 2, the control variables were entered as well as job satisfaction (dependent variable). Then, psychological capital (independent variable) was added to the equation in the second step. Finally, we added the interaction between psychological capital and social network size in the third step. Meanwhile, in order to cope with multicollinearity problems, we followed Aiken and West's [48] suggestion in the moderated regression model that the dichotomous variables were effects coded and continuous variables were centered.

As indicated in step 3 of Table 7, the results showed that R^2 increased from 0.173 to 0.258, which slightly increased the variance explained. This supports Hypothesis 2 stating that social network size moderates the relationship between psychological capital and job satisfaction (regression coefficient = -0.306 , $p < 0.001$). Besides this, Figure 3 also shows that the positive impact of psychological capital on job satisfaction was stronger when job seekers have small size networks than when job seekers have large size networks. Hence, Hypothesis 2 was supported.

4.3.3. Results of Test of Hypothesis 3. To test the moderating effect of social network tie strength on the relationship between psychological capital and job satisfaction, we followed the same procedure conducted above. We entered control variables and job satisfaction (dependent variable) in the first step, psychological capital (independent variable) in the second step, and the interaction between psychological capital and tie strength in social networks in the third step. The results (Table 8) indicate that R^2 increased from 0.182 to 0.258, which supports Hypothesis 3 (regression coefficient = -0.291 , $p < 0.001$). Furthermore, Figure 4 shows that the positive impact of psychological capital on job satisfaction was stronger when job seekers have weak ties than when job seekers have strong ties. Thus, Hypothesis 3 was supported.

5. Discussion

Extending the prior job search research, we applied a social network perspective to explore whether social network structure moderates the impact of psychological capital on job satisfaction. By using the data from a two-wave longitudinal design, the results show that network size moderates the relationship between psychological capital and job satisfaction, such that the relationship is more positive for job seekers with small social networks. Moreover, tie strength also moderates the relationship between psychological capital and job satisfaction, such that the relationship is more positive for job seekers with weaker ties in their social networks.

TABLE 3: The reliability of each item in the survey.

Variable	Item	CITC	α coefficient after deleting this item	α coefficient
Self-efficacy (one dimension of psychological capital)	SE1	0.755	0.903	0.917
	SE2	0.756	0.903	
	SE3	0.776	0.900	
	SE4	0.772	0.901	
	SE5	0.777	0.900	
	SE6	0.752	0.904	
Hope (one dimension of psychological capital)	H1	0.774	0.922	0.931
	H2	0.818	0.916	
	H3	0.806	0.918	
	H4	0.739	0.926	
	H5	0.846	0.913	
	H6	0.813	0.917	
Resilience (one dimension of psychological capital)	R1	0.886	0.954	0.962
	R2	0.889	0.954	
	R3	0.909	0.952	
	R4	0.866	0.956	
	R5	0.870	0.956	
	R6	0.854	0.958	
Optimism (one dimension of psychological capital)	O1	0.822	0.956	0.958
	O2	0.825	0.955	
	O3	0.914	0.945	
	O4	0.906	0.946	
	O5	0.861	0.951	
	O6	0.884	0.949	
Job satisfaction	JS1	0.636	0.723	0.794
	JS2	0.599	0.758	
	JS3	0.680	0.677	
Network size	NS1	0.805	0.906	0.918
	NS2	0.896	0.832	
	NS3	0.805	0.907	
Tie strength	TS1	0.863	0.919	0.939
	TS2	0.897	0.893	
	TS3	0.860	0.921	

TABLE 4: Means, SDs, and correlations of study variables.

Factor	Mean	SD	1	2	3	4	5	6	7	8	9	10
(1) Gender	0.616	0.487	1									
(2) Single child	0.384	0.487	0.168**	—								
(3) Student leader	0.517	0.500	0.063	0.068	—							
(4) CPC member	0.579	0.495	0.041	0.020	0.330**	—						
(5) University classification	3.201	0.803	0.130*	0.064	0.017	0.147**	—					
(6) Major classification	2.683	1.373	-0.078	-0.005	0.023	0.133*	0.058	—				
(7) Psychological capital	3.427	0.631	-0.026	-0.047	-0.045	0.104	-0.019	0.097	—			
(8) Job satisfaction	3.363	0.693	-0.144*	-0.023	-0.025	0.046	-0.023	0.079	0.381**	—		
(9) Network size	3.247	1.145	-0.068	0.068	0.046	0.042	-0.001	0.094	0.203**	0.179**	—	
(10) Tie	2.949	1.148	-0.072	-0.048	0.043	-0.040	0.012	0.011	0.063	-0.102	0.159**	—

Note: $N = 344$. * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

This study results in several essential findings which contribute to the literature regarding social networks and job search. First, psychological capital is a type of invisible resource, and it can help job seekers to have better job search behaviors and then good employment outcomes, e.g., acquirement of satisfactory job offers. It is well known that China has a large population and the competition in the job market is very fierce. Thus, job seekers with higher psychological capital are more likely to succeed in the

competition. This finding is consistent with previous research on the positive effect of psychological capital on employees' attitudes or behaviors (e.g., satisfaction, performance, and well-being) [29–31].

Second, social network structure, i.e., network size and tie strength, is found to be a moderator of the relationship between psychological capital and job satisfaction. Specifically, the impact of psychological capital on job satisfaction will be stronger, when job seekers have a smaller network or

TABLE 5: Results of exploratory factor analysis (EFA) in the study.

Variables	Items	Factors			
		1	2	3	4
Psychological capital	PC1	0.816			
	PC2	0.832			
	PC3	0.844			
	PC4	0.832			
	PC5	0.846			
	PC6	0.820			
	PC7	0.837			
	PC8	0.859			
	PC9	0.860			
	PC10	0.808			
	PC11	0.883			
	PC12	0.860			
	PC13	0.906			
	PC14	0.913			
	PC15	0.933			
	PC16	0.898			
	PC17	0.903			
	PC18	0.890			
	PC19	0.859			
	PC20	0.882			
	PC21	0.931			
	PC22	0.922			
	PC23	0.886			
	PC24	0.908			
Job satisfaction	JS1		0.792		
	JS2		0.788		
	JS3		0.842		
Network size	NS1			0.891	
	NS2			0.945	
	NS3			0.893	
Tie strength	T1				0.932
	T2				0.944
	T3				0.933

TABLE 6: Hierarchical regression of the influence of psychological capital on job satisfaction.

Predictor	Job satisfaction	
	Step 1	Step 2
<i>Control variables</i>		
Gender	-0.137*	-0.135**
Single child	0.003	0.018
Student leader	-0.037	-0.005
CPC member	0.058	0.010
University classification	-0.017	-0.003
Major classification	0.062	0.031
<i>Independent variable</i>		
Psychological capital		0.374***
R^2	0.029	0.164
ΔR^2	0.011	0.147
F	1.652	9.430***

Note: $N=344$. The values in the table are standardized beta weights (b). * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

when job seekers have weaker ties in social networks. This is an interesting and important finding, which indicates the interaction between social network structure and

TABLE 7: Hierarchical regression of moderating effect of network size on the relationship between psychological capital and job satisfaction.

Variables	Job satisfaction		
	Step 1	Step 2	Step 3
<i>Control variables</i>			
Gender	-0.137*	-0.127*	-0.104*
Single child	0.003	0.009	-0.007
Student leader	-0.037	-0.010	-0.023
CPC member	0.058	0.011	0.005
University classification	-0.017	-0.003	-0.005
Major classification	0.062	0.024	0.015
<i>Independent variable</i>			
Psychological capital		0.354***	0.264***
<i>Moderator</i>			
Network size		0.096	0.124*
<i>Interaction</i>			
Psychological capital \times network size			-0.306***
R^2	0.029	0.173	0.258
ΔR^2	0.011	0.153	0.238
F	1.652	8.748***	12.894***

Note: $N=344$. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

psychological capital. Prior research showed the important role of social networks in job search in Chinese society. It is reported that Chinese *guanxi* (i.e., interpersonal relationship between two individuals) [49], which is a typical social network in China, could help individuals to get job offers from a company and obtain advantages and benefits through *guanxi* HRM practices (i.e., HRM practices based on *guanxi*) [50, 51]. The findings in this study proved the formidable influence of social networks in Chinese society again, which echoes the findings in prior research [49–51]. That is to say, if job seekers have strong ties or large network sizes, the influence of psychological capital on job search behaviors and outcomes will be weakened.

Third, the findings in this study also showed that different personal resources (e.g., psychological capital and social capital) are helpful for job seekers to obtain satisfactory job offers in the job search process. In the last century, social networks, which are considered the most important personal resource, significantly influence almost every aspect of Chinese people's life. As Xiong et al. pointed out, one's decision will be always affected by their neighbors in social networks [52, 53]; e.g., users' decisions are often affected by the existing ratings on social media [54]. However, the economic reform and society development lead to the perfection of legal regulation and policies and transparency of the procedures. This results in the declining significance of social networks in Chinese society. That is to say, social networks as a part of social capital are no longer the only personal resource that can determine whether Chinese people could succeed in the competition, and this represents the development of the society. On the other hand, it is notable that job seekers who have large network sizes or strong ties still rely heavily on the help of social

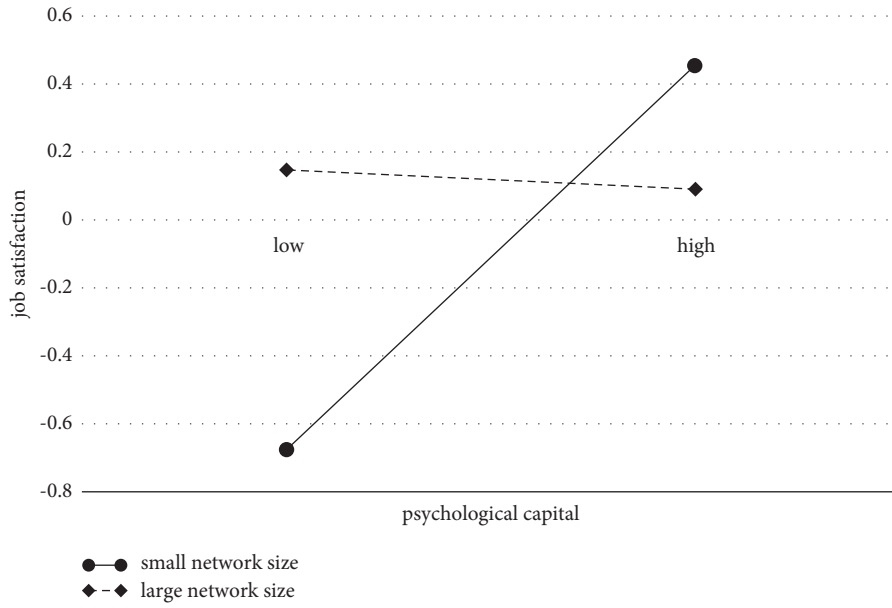


FIGURE 3: Interaction of psychological capital and network size on job satisfaction.

TABLE 8: Hierarchical regression of moderating effect of tie strength on the relationship between psychological capital and job satisfaction.

Variables	Job satisfaction		
	Step 1	Step 2	Step 3
<i>Control variables</i>			
Gender	-0.137*	-0.144**	-0.144**
Single child	0.003	0.012	0.004
Student leader	-0.037	0.005	0.009
CPC member	0.058	0.000	-0.020
University classification	-0.017	0.002	-0.026
Major classification	0.062	0.032	0.040
<i>Independent variable</i>			
Psychological capital		0.383***	0.299***
<i>Moderator</i>			
Tie strength		-0.136**	-0.133**
<i>Interaction</i>			
Psychological capital × tie strength			-0.291***
R^2	0.029	0.182	0.258
ΔR^2	0.011	0.163	0.238
F	1.652	9.342***	12.925***

Note: $N=344$. * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

networks in the job search process. Every coin has two sides; everything has both good and bad aspects. For instance, it is proper for job seekers to collect more information through social networks. However, it is inappropriate for those job seekers to join a company through informal social networks by violating procedural justice because this kind of practice based on informal social networks leads to the competitors' perceptions of unfairness [50, 51]. Therefore, it is important for Chinese enterprises and all other organizations to take action to ensure justice in the recruitment process. For instance, managers could ensure a fair procedure in recruitment and

more transparent decision making, in order to reduce the factors, which lead to unfairness [55, 56].

The study is not without its limitations. Firstly, this study adopted a convenience sample, including only undergraduates who are about to graduate. Future studies may test the hypothesized model with various respondents, such as the employees who have been working for a period. Secondly, due to the limited time for all empirical studies, we collected all the data for the survey within one month. Future studies could consider collecting the data within a longer period, e.g., collecting the second time data when the graduates start their work for a period.

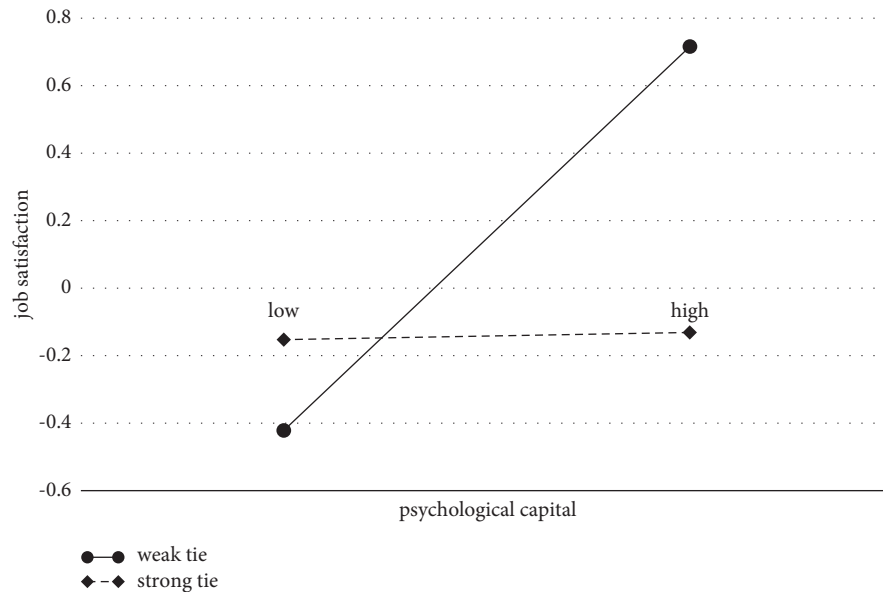


FIGURE 4: Interaction of psychological capital and tie strength on job satisfaction.

Data Availability

The data adopted to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities (CUPL: 20ZFQ63001), the Social Science Foundation from the China University of Political Science and Law (10820366 and 20ZFQ63001), the National Science Foundation of China (71874205), and the National Social Science Foundation of China (20AZD071).

References

- [1] R. Putnam, R. Leonardi, and R. Nanetti, *Making Democracy Work: Civic Tradition in Modern Italy*, Princeton University Press, New York, NY, USA, 1993.
- [2] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.
- [3] Y. Sato, "Are Asian sociologies possible?: universalism versus particularism," in *Facing an Unequal World: Challenges for a Global Sociology*, M. Burawoy, M. Chang, and M. F. Hsieh, Eds., vol. 2, pp. 192–200, Institute of Sociology, Academia Sinica and Council of National Associations of International Sociological Association, Taipei, Taiwan, 2010.
- [4] P. P. Li, "Social tie, social capital, and social behavior: toward an integrative model of informal exchange," *Asia Pacific Journal of Management*, vol. 24, no. 2, pp. 227–246, 2007.
- [5] S. Horak, M. Taube, I. Yang, and K. Restel, "Two not of a kind: social network theory and informal social networks in east Asia," *Asia Pacific Journal of Management*, vol. 36, no. 2, pp. 349–372, 2019.
- [6] F. McKee-Ryan, Z. Song, C. R. Wanberg, and A. J. Kinicki, "Psychological and physical well-being during unemployment: a meta-analytic study," *Journal of Applied Psychology*, vol. 90, no. 1, pp. 53–76, 2005.
- [7] E. E. Brasher and P. Y. Chen, "Evaluation of success criteria in job search: a process perspective," *Journal of Occupational and Organizational Psychology*, vol. 72, no. 1, pp. 57–70, 1999.
- [8] A. M. Saks and B. E. Ashforth, "Is job search related to employment quality?: it all depends on the fit," *Journal of Applied Psychology*, vol. 87, no. 4, pp. 646–654, 2002.
- [9] C. R. Wanberg, L. M. Hough, and Z. Song, "Predictive validity of a multidisciplinary model of reemployment success," *Journal of Applied Psychology*, vol. 87, no. 6, pp. 1100–1120, 2002.
- [10] R. Kanfer, C. R. Wanberg, and T. M. Kantrowitz, "Job search and employment: a personality-motivational analysis and meta-analytic review," *Journal of Applied Psychology*, vol. 86, no. 5, pp. 837–855, 2001.
- [11] E. A. Locke, "The nature and causes of job satisfaction," in *Hand Book of Industrial and Organizational Psychology*, M. D. Dunnette, Ed., pp. 1297–1349, Rand McNally, Chicago, Illinois, 1976.
- [12] E. Fabra Florit and L. E. Vila Lladosa, "Evaluation of the effects of education on job satisfaction: independent single-equation vs. structural equation models," *International Advances in Economic Research*, vol. 13, no. 2, pp. 157–170, 2007.
- [13] G. Blau, "Testing a two-dimensional measure of job search behavior," *Organizational Behavior and Human Decision Processes*, vol. 59, no. 2, pp. 288–312, 1994.
- [14] Z. H. Qiao, "The construct of graduates' employability and its effect on graduates' employment results," *Psychological Development and Education*, vol. 27, pp. 274–281, 2011, in Chinese.
- [15] L. Y. Sun, "Analysis of college students' job-seeking situation in the new era and countermeasures to improve employment services: a case study of Nanjing university of science and technology," *Employment Guidance*, vol. 12, pp. 54–59, 2020.
- [16] R. N. Bolles, *What Color is Your Parachute?: A Practical Manual for Job-Hunters and Career-Changers*, Ten Speed Press, Berkeley, CA, USA, 2006.

- [17] C. J. Collins and C. K. Stevens, "The relationship between early recruitment-related activities and the application decisions of new labor-market entrants: a brand equity approach to recruitment," *Journal of Applied Psychology*, vol. 87, pp. 1121–1133, 2002.
- [18] G. V. Hoye and F. Lievens, "Recruitment-related information sources and organizational attractiveness: can something be done about negative publicity?" *International Journal of Selection and Assessment*, vol. 13, pp. 179–187, 2005.
- [19] A. Tziner, E. Vered, and L. Ophir, "Predictors of job search intensity among college graduates," *Journal of Career Assessment*, vol. 12, no. 3, pp. 332–344, 2004.
- [20] G. V. Hoye, E. A. J. Hoof, and F. Lievens, "Networking as a job search behaviour: a social network perspective," *Journal of Occupational and Organizational Psychology*, vol. 82, pp. 661–682, 2009.
- [21] M. L. Forret and T. W. Dougherty, "Correlates of networking behavior for managerial and professional employees," *Group & Organization Management*, vol. 26, no. 3, pp. 283–311, 2001.
- [22] C. R. Wanberg, R. Kanfer, and J. T. Banas, "Predictors and outcomes of networking intensity among unemployed job seekers," *Journal of Applied Psychology*, vol. 85, no. 4, pp. 491–503, 2000.
- [23] M. A. Zottoli and J. P. Wanous, "Recruitment source research: current status and future directions," *Human Resource Management Review*, vol. 10, no. 4, pp. 353–382, 2000.
- [24] A. M. Saks and B. E. Ashforth, "Change in job search behaviors and employment outcomes," *Journal of Vocational Behavior*, vol. 56, no. 2, pp. 277–287, 2000.
- [25] A. Rego, F. Sousa, C. Marques, and M. P. E. Cunha, "Authentic leadership promoting employees' psychological capital and creativity," *Journal of Business Research*, vol. 65, no. 3, pp. 429–437, 2012.
- [26] F. Luthans, C. M. Youssef, and B. J. Avolio, *Psychological Capital*, Oxford University Press, Oxford, UK, 2007.
- [27] F. Luthans, J. B. Avey, B. J. Avolio, S. M. Norman, and G. M. Combs, "Psychological capital development: toward a micro-intervention," *Journal of Organizational Behavior*, vol. 27, no. 3, pp. 387–393, 2006.
- [28] F. Luthans, B. J. Avolio, J. B. Avey, and S. M. Norman, "Positive psychological capital: measurement and relationship with performance and satisfaction," *Personnel Psychology*, vol. 60, no. 3, pp. 541–572, 2007.
- [29] F. Luthans, S. M. Norman, B. J. Avolio, and J. B. Avey, "The mediating role of psychological capital in the supportive organizational climate—employee performance relationship," *Journal of Organizational Behavior*, vol. 29, no. 2, pp. 219–238, 2008.
- [30] J. B. Avey, T. S. Wernsing, and F. Luthans, "Can positive employees help positive organizational change?: impact of psychological capital and emotions on relevant attitudes and behaviors," *The Journal of Applied Behavioral Science*, vol. 44, no. 1, pp. 48–70, 2008.
- [31] J. B. Avey, R. J. Reichard, F. Luthans, and K. H. Mhatre, "Meta-analysis of the impact of positive psychological capital on employee attitudes, behaviors, and performance," *Human Resource Development Quarterly*, vol. 22, no. 2, pp. 127–152, 2011.
- [32] S. P. Borgatti and P. C. Foster, "The network paradigm in organizational research: a review and typology," *Journal of Management*, vol. 29, no. 6, pp. 991–1013, 2003.
- [33] P. S. Adler and S.-W. Kwon, "Social capital: prospects for a new concept," *Academy of Management Review*, vol. 27, no. 1, pp. 17–40, 2002.
- [34] S. E. Seibert, M. L. Kraimer, and R. C. Liden, "A social capital theory of career success," *Academy of Management Journal*, vol. 44, no. 2, pp. 219–237, 2001.
- [35] M. S. Granovetter, *Getting a Job: A Study of Contacts and Careers*, University of Chicago Press, Chicago, Illinois, 2nd edition, 1995.
- [36] D. W. Brown and A. M. Konrad, "Granovetter was right: the importance of weak ties to a contemporary job search," *Group & Organization Management*, vol. 26, no. 4, pp. 434–462, 2001.
- [37] P. H. Reingen and J. B. Kernan, "Analysis of referral networks in marketing: methods and illustration," *Journal of Marketing Research*, vol. 23, no. 4, pp. 370–378, 1986.
- [38] M. C. Gilly, J. L. Graham, M. F. Wolfinbarger, and L. J. Yale, "A dyadic study of interpersonal information search," *Journal of the Academy of Marketing Science*, vol. 26, no. 2, pp. 83–100, 1998.
- [39] P. M. Podsakoff, S. B. MacKenzie, and N. P. Podsakoff, "Sources of method bias in social science research and recommendations on how to control it," *Annual Review of Psychology*, vol. 63, no. 1, pp. 539–569, 2012.
- [40] J. Hu, B. Erdogan, T. N. Bauer, K. Jiang, S. Liu, and Y. Li, "There are lots of big fish in this pond: the role of peer overqualification on task significance, perceived fit, and performance for overqualified employees," *Journal of Applied Psychology*, vol. 100, no. 4, pp. 1228–1238, 2015.
- [41] H. F. Gollob and C. S. Reichardt, "Taking account of time lags in causal models," *Child Development*, vol. 58, no. 1, pp. 80–92, 1987.
- [42] R. W. Brislin, "The wording and translation of research instrument," in *Field Methods in Cross-Cultural Research*, W. J. Lonner and J. W. Berry, Eds., pp. 137–164, Sage, Beverly Hills, CA, USA, 1986.
- [43] P. Bharati and A. Chaudhury, "An empirical investigation of decision-making satisfaction in web-based decision support systems," *Decision Support Systems*, vol. 37, no. 2, pp. 187–197, 2004.
- [44] V. Pereira, F. Tavares, P. Mihaylova, V. Mladenov, and P. Georgieva, "Factor analysis for finding invariant neural descriptors of human emotions," *Complexity*, vol. 2018, Article ID 6740846, 8 pages, 2018.
- [45] K. Zhang, M. Hassan, M. Yahaya, and S. Yang, "Analysis of work-zone crashes using the ordered probit model with factor analysis in Egypt," *Journal of Advanced Transportation*, vol. 2018, Article ID 8570207, 10 pages, 2018.
- [46] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, NY, USA, 1984.
- [47] P. M. Podsakoff and D. W. Organ, "Self-reports in organizational research: problems and prospects," *Journal of Management*, vol. 12, no. 4, pp. 531–544, 1986.
- [48] L. S. Aiken and S. G. West, *Multiple Regression: Testing and Interpreting Interactions*, Sage, Newbury Park, CA, USA, 1991.
- [49] C. C. Chen, X.-P. Chen, and S. Huang, "Chinese Guanxi: an integrative review and new directions for future research," *Management and Organization Review*, vol. 9, no. 1, pp. 167–207, 2013.
- [50] D. Guthrie, "The declining significance of guanxi in China's economic transition," *The China Quarterly*, vol. 154, pp. 254–282, 1998.
- [51] C. Chen, Y. Chen, and K. Xin, "Guanxi practices and trust in management: a procedural justice perspective," *Organization Science*, vol. 15, no. 2, pp. 200–209, 2004.
- [52] F. Xiong, X. M. Wang, S. R. Pan, H. Yang, H. S. Wang, and C. Q. Zhang, "Social recommendation with evolutionary

- opinion dynamics,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [53] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, “Bayesian personalized ranking based on multiple-layer neighborhoods,” *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [54] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, “Exploiting implicit influence from information propagation for social recommendation,” *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [55] C. C. Chen and X.-P. Chen, “Negative externalities of close guanxi within organizations,” *Asia Pacific Journal of Management*, vol. 26, no. 1, pp. 37–53, 2009.
- [56] R. Nan and F. Ouyang, “Influence of social networks on citizens’ willingness to participate in social governance: evidence from China,” *Complexity*, vol. 2020, Article ID 3819514, 16 pages, 2020.

Research Article

Research on the Structural Characteristics of Entertainment Industrial Correlation in China: Based on Dual Perspective of Input-Output and Network Analysis

Yang Xun ¹, Wensheng Shi ², and Tianyu Liu ³

¹School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710049, China

²Yulin University, Yulin 719000, China

³School of Leisure Sport, Shanghai University of Sport, Shanghai 200444, China

Correspondence should be addressed to Wensheng Shi; 502420007@qq.com

Received 11 June 2021; Accepted 28 September 2021; Published 31 October 2021

Academic Editor: Fei Xiong

Copyright © 2021 Yang Xun 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.

Based on the dual perspective of input-output and network analysis, this study takes typical industrial sectors of China's entertainment industry as representatives. Through the input-output analysis of industrial correlation characteristic indicators and construction of an industrial correlation network, we conduct a systematic and quantitative study on the entertainment industrial correlation characteristics and structural characteristics of the industrial correlation network in China. Furthermore, we clarify the role of the entertainment industry in China's industrial development and its positioning in China's whole industrial correlation network. We have the following key findings: China's entertainment industry as a whole shows the characteristics of final demand-oriented industries, whose rapid development plays a certain positive role in boosting consumption and driving economic growth. Within the whole industrial correlation network in China, there is frequent interaction between the entertainment industry and other industry sectors within the directly related network; it will especially exert obvious radiation and driving effect on the upstream industry. However, limited by the scale of the direct industrial correlation network, such influence is still difficult to achieve the common development of most industries in China.

1. Research Background

In recent years, with the rapid progress of productivity and economic and social undertakings, the overall income level of Chinese residents has been greatly improved, and entertainment and leisure have become increasingly important in people's daily life. As people's lifestyles show the characteristics of leisure and entertainment, more and more people are willing to devote energy, money, and time to participate in various forms of entertainment. Entertainment consumption has gradually become an important part of residents' daily consumption [1]. In this process, China's commercial entertainment industry has also ushered in a stage of rapid development. As a representative industry of the modern service industry, although China's entertainment industry started late, due to the acceleration of the

domestic industrial structure transformation and upgrading and the rapid growth of entertainment consumption demand, the overall growth momentum of China's entertainment industry in recent years has been strong. Relevant statistics show that the scale of China's pan-entertainment core industry has maintained an average annual growth rate of more than 15% from 2015 to 2019, which is significantly higher than the GDP growth rate during the same period [2]. It can be seen that the rapid development of the entertainment industry has gradually begun to play a role in boosting China's overall national economic growth.

Existing theoretical studies have shown that there is a natural direct connection between the production activities of the entertainment industry and various industries such as media, tourism, real estate, and catering. This connection will be further passed on to other industries through the

exchange of intermediate products, which makes the entertainment industry show obvious industry linkage effects [3]. Therefore, a mature entertainment industry will have a greater radiation effect on the development of many industrial sectors in a country or region. At present, although China's entertainment industry has made considerable progress in the process of economic development in recent years, its extensive relevance with other industries and economic driving ability have also received general attention from the industry and academia, but existing researches show that the theoretical research on the economic effects of China's entertainment industry is still mostly at the qualitative level, and the positioning of the entertainment industry in the overall industrial structure of the national economy and the quantitative analysis of the characteristics of its industrial correlation structure are still relatively lacking [4]. Therefore, in order to deepen the understanding of the role of the entertainment industry in the development of China's industry and to further clarify the industry correlation and interaction effects between the entertainment industry and other industries, this study uses the latest input-output data released by the National Bureau of Statistics of China, combined with input-output analysis and network analysis to conduct a systematic and quantitative study on the characteristics of China's entertainment characteristics of industrial correlation and industrial correlation network structure, so as to provide theoretical support for further exerting the extensive driving role of the entertainment industry in the development of China's industry and accelerating the pace of industrial transformation and upgrading.

2. Analytical Framework for the Characteristics of Industrial Correlation Structure of China's Entertainment Industry

The traditional industry correlation analysis mainly uses the input-output table to reveal the interdependence between the sources of production input and the destination of output use among the various industrial sectors of the national economy. By converting the flow data in the input-output table into different coefficient matrices, various coefficients that reflect the direct and indirect economic links between industrial sectors can be calculated, so as to further carry out a more in-depth quantitative analysis of the relationship between industries [5]. However, although the classic input-output model can give a clear quantitative description of the relationship between the two industries, the information obtained through input-output analysis is too scattered, and it is difficult to fully reflect the positioning of the entertainment industry in China's complete industrial structure system and structural features. Therefore, if we want to systematically show the role of the entertainment industry in China's overall industrial layout, we need to find a basis from the perspective of network analysis.

As a new quantitative analysis method, network analysis integrates theories and research methods of statistics, topology, system dynamics, and other disciplines. By

abstracting each action subject and their interrelationships in a complex network system into nodes and ties, network analysis can more comprehensively describe the whole structural characteristics of the network and the relationship between the subjects in the network [6]. At present, the research methods of network analysis have been widely used in most practical network research fields such as sociology, economics, and communication [7–9], especially in the field of industrial economics. The analysis of the industrial correlation network based on the input-output relationship has basically formed a relatively mature research framework, which provides new research tools for revealing the characteristics of the interconnected network between industrial sectors and grasping the direction of macroindustrial structure optimization. In the current related research on industrial correlation networks, the main research idea is to abstract each industrial sector as a node in network analysis and use the 0-1 adjacency matrix formed by filtering various coefficients in the input-output model as the directed edge basis to build an industrial correlation network and quantitatively analyze the overall structural characteristics and node characteristics of the industrial correlation network such as network density, clustering coefficient, symmetry, and centrality degree, to systematically study the positioning of each industry sector and the characteristics of the overall industrial correlation structure in the industrial correlation network [10].

3. Analytical Framework for the Characteristics of Industrial Correlation of the Entertainment Industry Based on Input-Output Analysis

In the input-output analysis stage, this study will use the input-output table to calculate the direct consumption coefficient, direct demand coefficient, intermediate consumption rate, intermediate demand rate, diffusion coefficient, and inducing coefficient of China's entertainment industry. Through the analysis of the above-mentioned characteristic indicators, the research will clarify the quantitative relationship between the input and output of the typical entertainment industry sector from production to final use and other industry sectors and provide an overall analysis of the industry type and characteristics of China's entertainment industry. The analysis of the role played in the national economic system provides a basis for quantitative research on the structural characteristics of the industry's correlation network. The calculation methods and descriptions of specific characteristic indicators are as follows.

3.1. Direct Consumption Coefficient and Direct Distribution Coefficient. In the input-output analysis, the direct consumption coefficient refers to the value of the product or service of the i sector that will be directly consumed for each additional unit of total output in the production and operation process of the j industry sector, recorded as a_{ij} ($i, j = 1, 2, \dots, n$). The calculation method is

$$a_{ij} = \frac{x_{ij}}{X_j}, \quad i, j = 1, 2, \dots, n. \quad (1)$$

Among them, x_{ij} is the value of the products or services of the i industrial sector directly consumed during the production and operation of the j industrial sector and X_j is the total input of the j sector.

The direct distribution coefficient refers to the amount of value in which the total output of each unit of the i sector is allocated to the j sector for direct use as factors of production, recorded as h_{ij} ($i, j = 1, 2, \dots, n$). The calculation method is

$$h_{ij} = \frac{x_{ij}}{X_i}, \quad i, j = 1, 2, \dots, n. \quad (2)$$

Among them, x_{ij} is the value of the products or services used by the i sector allocated to the j sector and X_i is the total output of the i sector.

The above two indicators reflect the direct technical and economic connection between a certain industrial sector and another industrial sector in the process of production and operation from the perspective of input and output, respectively. When the direct consumption coefficient of a certain industry on another industry is larger, it means that the direct dependence of the industry on another industry sector in the production process is more obvious. When the direct distribution coefficient of a certain industry to another industry is larger, it means that the direct support effect of the industry on another industry sector is greater in the production process.

3.2. Intermediate Demand Rate and Intermediate Input Rate.

The intermediate demand rate refers to the ratio of the intermediate demand for products of the i industrial sector in other sectors of the national economy to the total demand for the products of the industrial sector. It reflects the proportion of the product or service produced by the industrial sector as a means of production and consumption. The calculation method is

$$w_i = \frac{\sum_{j=1}^n x_{ij}}{X_i}, \quad i = 1, 2, \dots, n. \quad (3)$$

Among them, the numerator is the sum of the intermediate demand for products or services of the i industrial sector by other industrial sectors, and the denominator X_i is the total demand for the products or services of the i industrial sector.

Intermediate input rate refers to the ratio of the amount of intermediate input obtained from other industrial sectors during the production and operation of the j industrial sector to its total input, which reflects the proportion of intermediate products obtained by the industrial sector from other industrial sectors in the process of production. The calculation method is

$$u_j = \frac{\sum_{i=1}^n x_{ij}}{X_j}, \quad j = 1, 2, \dots, n, \quad (4)$$

where the numerator is the sum of the intermediate input of the j industrial sector from other industrial sectors in the production and operation process, and the denominator X_j is the total input of the j industrial sector.

The above two characteristic indicators can reflect the overall position of a certain industrial sector in the national economy and its relationship with upstream and downstream industrial sectors. When the intermediate demand rate of a certain industry is higher, it means that the products produced by the industry are more of the nature of the means of production and play the role of the basic industry in the process of correlating with other industries; otherwise, it means that the products produced by the industry are more oriented toward final demand and have a stronger effect on expanding domestic demand. When the intermediate input rate of a certain industry is high, it means that the industry needs to obtain more intermediate products from other industrial sectors in the process of production and operation; that is, the industry sector has a more obvious leading role in its upstream industry; on the contrary, it shows the higher industrial added value of the products of the industry sector.

3.3. The Diffusion Coefficient and Inducing Coefficient.

The diffusion coefficient is the ratio of the influence of a certain industrial sector to the average level of the influence of various industrial sectors in the national economy, reflecting the extent to which changes in the final product of the industry affect the production demand of various industrial sectors of the national economy. The calculation method is

$$F_j = \frac{\sum_{i=1}^n b_{ij}}{(1/n) \sum_{j=1}^n \sum_{i=1}^n b_{ij}}, \quad j = 1, 2, \dots, n, \quad (5)$$

where the numerator is the sum of the j column of the Leontief inverse matrix (Leontief inverse matrix A is the direct consumption coefficient matrix composed of the direct consumption coefficient a_{ij}), which means that every unit of final product produced by the j industrial sector is the complete demand for products in all sectors of the national economy. The denominator is the average of the column sums of the Leontief inverse matrix.

The inducing coefficient is the ratio of the sensitivity of a certain industry sector to the average level of the sensitivity of each industry sector in the national economy and reflects the degree of demand sensitivity experienced by the industry when the final product of each industry sector in the national economy changes. The formula is as follows:

$$E_i = \frac{\sum_{j=1}^n b_{ij}}{(1/n) \sum_{i=1}^n (\sum_{j=1}^n b_{ij})}, \quad i = 1, 2, \dots, n, \quad (6)$$

where the numerator is the sum of the i row of the Leontief inverse matrix, which represents the total demand for the i industrial sector when each industry sector produces a unit of the final product. The denominator is the average of the row sums of the Leontief inverse matrix.

The above two coefficients both reflect the role or impact of a certain industrial sector in the national economic system. Among them, the diffusion coefficient represents the degree of impact on the supply sector when the input-output relationship of a certain industrial sector changes. The inducing coefficient indicates the degree to which the input-output relationship of a certain industrial sector is affected by the demand sector. The larger the diffusion coefficient, the stronger the demand rate of the industry for other industrial sectors, and the more obvious its pulling function on other industrial sectors. When the inducing coefficient is larger, it means that the relative demand of other industry sectors for the industry is greater; that is, the promoting effect of this industry on other industry sectors will be more obvious.

4. An Analysis Framework of the Structural Characteristics of China's Entertainment Industry Correlation Network

In the network analysis stage, the research first converts the direct consumption coefficient matrix obtained from the input-output analysis into a binary adjacency matrix and uses the binary data to construct an industry-correlated network covering China's 149 industrial sectors, so as to fully display the structure of the linkages between the various industrial sectors of the national economy. Then, on this basis, the research will build a self-centered network centered on the typical sectors of the entertainment industry and more specifically show the correlated network structure between the entertainment industry and each directly correlated industry. The calculation methods and descriptions of the characteristic indicators of the industry-correlated networks involved in the research at this stage are as follows.

4.1. Indicators of the Overall Structural Characteristics of Industrial-Related Networks

4.1.1. The Structural Characteristics of the Small-World Network. In a network, if most of the nodes in the network are not adjacent to each other, but any node can access other nodes through its adjacent nodes with fewer jumps, then this network will have the characteristics of the small-world structure. The characteristics of the small-world structure are a key index to reflect the network of the overall structure, and they are also the premise to analyze the network structure. Generally, describing mainly the small-world structure characteristics of the network is through two indexes: the average aggregation coefficient and the average shortest path length.

Cluster coefficient is the ratio of the number of connections between adjacent nodes of a node in the network to the number of possible connections, which reflects the degree of the close connection between a node in the network and its neighbors. And the average agglomeration coefficient is the average of the agglomeration coefficient of each node in the network, reflecting the degree of agglomeration of the whole network. In industrial related networks, the higher the agglomeration coefficient is, the

closer the technological and economic connection between each industrial node in the network and its neighboring nodes is. The calculation formula of the average clustering coefficient in the directed network is as follows:

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n \frac{A_i}{k_i(k_i - 1)} \quad (7)$$

Among them, the numerator A_i is the number of edges that actually exist between neighboring nodes of node v_i and the denominator is the maximum number of edges that may exist between neighboring nodes of node v_i .

The shortest path length is the distance of the shortest path among the multiple paths connected between two nodes, while the average shortest path is the average of the shortest distance between any two points in the network. In industrial related networks, the average shortest path length reflects the transmission efficiency of resources and information among various industrial departments. The smaller the average shortest path length is, the closer the relationship between various industries in the network is, and the more efficient the circulation of resources and information among industrial departments is. The calculation method is

$$d = \frac{1}{N(N-1)} \sum_{i,j \in N, i \neq j} d_{ij} \quad (8)$$

Among them, d_{ij} is the shortest path length between nodes v_i and v_j and N is the number of nodes in the network.

In network analysis, if a network has both a higher average agglomeration coefficient and a lower average shortest path length, it indicates that the network has the structural characteristics of a small-world network [11]. For the industrial related networks, the more obvious the small-world feature is, the closer the connection between each industrial sector and its neighboring industrial sector is, and the higher the quality and efficiency of the industrial sector in the exchange of resources or information are.

4.1.2. Network Density. Network density refers to the ratio between the actual number of existing edges and the maximum number of possible existing edges among all nodes in a certain network. It reflects the tightness of connections among all nodes in the network from the overall level. The calculation method of network density in a directed network is as follows:

$$\Delta = \frac{L}{N(N-1)} \quad (9)$$

Among them, L is the number of edges that actually exist in the network and N is the number of nodes in the network.

In the industrial related network, the level of network density reflects the close degree of economic interaction between various industrial departments and also reflects the close degree of connection among the members of the entire industrial related network, which is an important indicator to understand its structure.

4.2. Industrial Node Characteristic Indexes of Industrial-Related Network

4.2.1. Node Degree. Node degree is the number of edges connected between a node and other nodes in the network. In a directed network, node degree is divided into point-out degree and point-in degree according to edge pointing, in which point-out degree refers to the number of edges where the node points to other nodes, while point-in degree refers to the number of edges where other nodes point to the node.

On the whole, node degree in the industrial related network reflects the position of a particular industrial department in the overall industrial association network. The higher the node degree of an industrial department, the more other industrial departments it is associated with, and the greater its influence in the industrial related network. As the industrial related network constructed by direct consumption coefficient has directivity, the node degree in the industrial association network is also divided into point-in degree and point-out degree. The point-in degree indicates how many final products of the industrial sector need to be consumed in the production process, reflecting the forward relevance of the industrial sector, and the point-out degree indicates how many production demands of industrial sectors will be met by the final products of the industry sector, reflecting the backward correlation of the industry sector.

4.2.2. Betweenness Centrality. In the industry-related network, betweenness centrality refers to the number of times where a node in a network acts as the shortest path bridge between other two nodes, reflecting the intermediary role and importance of this node in the network. The higher intermediate centrality is, the more obvious the industrial sector is as an intermediary to connect other industrial sectors, and the more it is at the center of the industrial related network. The calculation method is as follows:

$$C_B(v_i) = \sum_{j < k} \frac{g_{jk}(v_i)}{g_{jk}} \quad (10)$$

Among them, g_{jk} represents the number of shortest paths between node v_j and v_k and $g_{jk}(v_i)$ represents the number of shortest paths between node v_j and v_k that pass through node v_i .

4.3. Data Sources, Industrial Sector Selection, and Data Processing Methods. In this study, the original data were used as “China’s 2017 Input-Output Table” released by the National Bureau of Statistics of China in 2019. Different from the classifying way in previous years of 42 industrial categories, this input-output table divides the production sectors of China’s national economy into 149 specific industrial sectors. According to the classification interpretation of each industrial sector, among all industrial sectors, the “entertainment” sector includes indoor entertainment activities, amusement parks, leisure and sightseeing activities, lottery activities, cultural and sports entertainment activities, and

other entertainment industries. “Radio, television, film, and film recording production” departments include radio, television, film, and television program production, integrated broadcasting control of radio and television, film, and radio and television program distribution, film screening, recording production, and other industries [12]. The above two industrial sectors basically cover the core industries of China’s entertainment industry. Therefore, this study will select “entertainment” and “radio, television, film, and film recording production” as the representatives of the entertainment industry to conduct research and analysis.

In data processing, Excel and RStudio were first used to conduct input-output analysis on the original data to calculate the industrial related characteristic indexes of the above two industrial sectors. Then, Gephi and the direct consumption coefficient matrix obtained from the previous analysis were used to construct the overall industrial association network and the entertainment industry direct association network. Furthermore, this paper analyzes the structural characteristics of the association network of the core departments of the entertainment industry.

5. Input-Output Analysis of the Related Characteristics in China’s Entertainment Industry

5.1. Direct Consumption Coefficient and Direct Distribution Coefficient of the Entertainment Industry. According to equations (1) and (2), the direct consumption coefficient and direct distribution coefficient of entertainment, radio, television, film, and film recording production are measured by using the basic flow table in the input-output table.

According to the calculation results of the direct consumption coefficient, “entertainment” and “radio, television, film, and video recording production” will have a direct consumption relationship with other 111 and 99 industrial sectors, respectively, in the production process. It shows that the core sector of China’s entertainment industry will produce direct demand for most other industrial sectors in the production process and has a wide range of pulling capacity for the development of upstream industries. After ranking, the 10 industrial sectors that are most closely related to the direct consumption of the entertainment industry are shown in Table 1.

From the above statistical results, it can be seen that the three industry sectors with the highest direct consumption coefficients, which represent the indoor exterior entertainment industry, are the real estate, beverage, and alcohol and liquor industries. The entertainment industry sector’s direct consumption coefficients for the above three industry sectors all exceed 0.04. It shows that every 10,000 yuan of indoor and outdoor leisure and entertainment products produced by the entertainment industry directly consume more than 400 yuan of the products of these sectors, and the direct dependence and traction effect on them is the largest compared with other industrial sectors. Representing the film and television entertainment industry, the radio, television, film, and video recording production industries have

TABLE 1: Statistical table of direct consumption coefficient of the Chinese entertainment industry on other industrial sectors.

Entertainment		Radio, film, television, and (other) audiovisual media	
Industry sector	Direct consumption factor	Industry sector	Direct consumption factor
Real estate	0.0611	Manufacture of textiles, clothing; apparel industry	0.0602
Beverages	0.0497	Commercial services	0.0598
Manufacture of alcohol	0.0433	Catering	0.0427
Money, finance, and other financial services	0.0367	Radio, film, television, and (other) audiovisual media	0.0403
Commercial services	0.0277	Accommodation	0.0259
Entertainment	0.0270	Special chemical products and explosives, pyrotechnics, fireworks products	0.0255
Manufacture of tobacco	0.0204	Other services	0.0156
Retail trade	0.0191	Aerospace	0.0154
Refined tea	0.0141	Real estate	0.0140
Wholesale trade	0.0138	Printed and recorded media reproductions	0.0139

the highest direct consumption coefficients in three industrial sectors: textile, clothing, business services, and catering industries. Among them, the direct consumption coefficient of the broadcasting, television, film, and film and television recording production industries on the textile, clothing, and business services is the same, which is about 0.06. It shows that for every 10,000 yuan of film and television entertainment products produced by the radio, television, film, and film and television recording production industries, the direct consumption of these two departments' products is about 600 yuan, which has a strong backward correlation.

Comparing the 10 industry sectors with the highest direct consumption coefficients of the two typical industry sectors in China's entertainment industry, we can find that the only industries in the top ten of the two typical industry sectors in the entertainment industry are real estate and business services. This shows that China's entertainment industry covers a wide range of backward correlations, and it has a certain degree of dependence on most other industries and will also have a certain driving effect on its development.

From the calculation results of the direct distribution coefficient, it can be seen that there is more or less a direct distribution relationship between the entertainment industry and the radio, television, film, and film recording production industry and all other industrial sectors. After ranking, the 10 industrial sectors with the highest direct distribution coefficients of the two industrial sectors are shown in Table 2.

From the above statistical results, although there is a certain direct distribution relationship between the two typical Chinese entertainment industries and other industries, it can be seen from the numerical point of view that the direct distribution coefficient is the highest industry itself, and of the other sectors under the ranking position with direct distribution coefficient significantly less than the direct consumption coefficients, it is shown that, in addition to the final demand oriented, the products or services produced by China's entertainment industry are mainly to meet the needs of its own industry, and it is difficult to promote the

development of other industrial sectors through direct supply.

5.2. Intermediate Demand Rate and Intermediate Input Rate of the Entertainment Industry. According to equations (3) and (4), the research has measured the intermediate demand rate and intermediate input rate of two typical industrial sectors in China's entertainment industry. The results are shown in Table 3.

It can be seen from Table 3 that the intermediate demand rates of the entertainment industry sector representing the indoor exterior light entertainment industry and the broadcasting, television, film, and film and television recording production sector representing the film and television entertainment industry are 55.40% and 37.51%, respectively, in 149 industries across the country. The rankings in the departments are relatively low. From the perspective of the intermediate investment rate, the value and ranking of the two typical sectors of China's entertainment industry are basically the same, with an intermediate investment rate lower than 50%, ranking basically at the bottom of all industrial sectors in China.

A comprehensive comparison of the intermediate demand rate and intermediate input rate of the two industrial sectors shows that although the intermediate demand rate of the entertainment industry, which represents the indoor exterior light entertainment industry, is slightly higher than 50%, it is still relatively low in terms of ranking. Therefore, the two typical sectors of the entertainment industry both show the characteristics of low intermediate demand and low intermediate input. According to the industry classification method of input-output analysis, this means that China's entertainment industry as a whole shows the characteristics of the basic industry of final demand, and the products and services it produces are more directly meeting final demand, which is positive for expanding consumption. Because of the low intermediate input rate, the industry has a high industrial added value, which can create economic value and stimulate economic growth.

TABLE 2: Statistical table of direct distribution coefficients of the Chinese entertainment industry to other industrial sectors.

Entertainment		Radio, film, television, and (other) audiovisual media	
Industry sector	Direct consumption factor	Industry sector	Direct consumption factor
Entertainment	0.0270	Radio, film, television, and (other) audiovisual media	0.0403
Insurance	0.0121	Radio, television, and satellite transmission services	0.0206
Monetary and financial services and other financial activities	0.0111	Water transport	0.0070
Culture and arts	0.0066	Radio and television equipment and radar and ancillary equipment	0.0060
Sports	0.0061	Entertainment	0.0055
Resident services	0.0038	Accommodation	0.0032
Internet and related services	0.0032	Production and distribution of tap	0.0026
Loading/unloading, removal, and storage	0.0031	Sports	0.0023
Radio, television, and satellite transmission services	0.0029	Social work	0.0020
Entertainment	0.0270	Other electrical machinery and equipment	0.0020

TABLE 3: Statistical table of intermediate demand rate and intermediate investment rate of the Chinese entertainment industry.

Industry sector	Intermediate demand rate	Rank	Intermediate investment rate	Rank
Entertainment	0.5540	97	0.4455	130
Radio, film, television, and (other) audiovisual media	0.3751	116	0.4571	129

5.3. *The Diffusion Coefficient and Inducing Coefficient of the Entertainment Industry.* According to equations (5) and (6), the diffusion coefficient and inducing coefficient of two typical industry sectors in China's entertainment industry are further measured by using the calculated direct consumption coefficient. The results are shown in Table 4.

From the diffusion coefficient calculated result on behalf of the film and television entertainment industry of radio, television, film, and video recording production sector, the diffusion coefficient is slightly higher than the representative of the indoor and outdoor sightseeing leisure entertainment industry sector in the entertainment business, but it is obviously lower than the average level of other industries in China, among all sectors. From the calculation results of the reaction coefficient, the entertainment industry sector and radio, television, film, and film recording production industry sector of the reaction coefficient are basically the same but are lower than the national average level.

A comprehensive comparison of the diffusion coefficient and the inducing coefficient of two typical industry sectors in China's entertainment industry shows that the diffusion coefficient of both is higher than the inducing coefficient. This indicates that the above industries play a more obvious role as users in China's national economy, and their demand for other industrial sectors is higher than that of other industrial sectors in the process of production and operation, so their pulling effect is more obvious; that is, their backward correlation with other industrial sectors is more close. However, from the numerical point of view, the diffusion coefficient of these two industrial sectors is higher than the inducing coefficient, but compared with other industrial sectors in the country, their influence level is relatively low, and their economic driving effect is still limited.

6. Analysis on the Characteristics of Industrial-Related Network Structure of the Chinese Entertainment Industry

Through the input-output analysis of the industrial relevance characteristics of China's entertainment industry, the research has sorted out the general industrial relevance of China's entertainment industry. At this stage, the research will build a more systematic description of the entertainment industry's industry-related network structure characteristics by constructing an industry-related network based on the data obtained from the input-output analysis.

6.1. *Construction of Industrial-Related Network.* According to research needs, this research will first build the overall industry-related network of China's industry sectors based on the direct consumption coefficients obtained from the previous research, and then on this basis, we will build China's entertainment industry with two typical industry sectors as the center connected directly to the network.

The industrial relational network constructed by this research is an unauthorized directed network, and the direction of the edges is the direction of economic and technological connections between various industrial departments. Whether there is a connection between various industrial departments requires the binary value obtained by the direct consumption coefficient matrix conversion adjacency matrix for judgment. In order to show the network structure more clearly, it is necessary to select an appropriate threshold to filter the weak connections between nodes in the network. Referring to the practice of Wang [13], Li [10], and others, this study selected the mean value of direct

TABLE 4: Statistics of diffusion coefficient and inducing coefficient of the Chinese entertainment industry.

Industry sector	Diffusion coefficient	Rank	Inducing coefficient	Rank
Entertainment	0.7275	133	0.4881	109
Radio, film, television, and (other) audiovisual media	0.7857	126	0.4612	123

consumption coefficient matrix 0.004454 as the threshold value. If the direct consumption coefficient of the industrial sector I to industrial sector j is greater than the threshold value, the corresponding adjacency matrix is 1. This indicates that there is an edge pointing to I from j between industrial sector i and industrial sector j . In addition, since the research mainly wants to show the correlation between the X industry and other industries, the self-loop of industrial nodes is removed during the network construction [14]. The overall industrial related network constructed according to this matrix is shown in Figure 1 (because there are too many nodes, the industrial node names are not marked in the figure).

In the graph of the overall industry association network, the research selects entertainment, radio, television, film, and film and television recording production as the center, determines the depth of 1, filters the overall association network, and constructs the corresponding Chinese entertainment industry direct association network. The resulting network diagram is shown in Figures 2 and 3.

6.2. Analysis of the Overall Structural Characteristics of the Entertainment Industry-Related Network. According to equations (7)–(9), the average agglomeration coefficient, average shortest path length, and network density of the above three industrial association networks are, respectively, measured, and the basic information of the three network graphs is statistically analyzed. The results are shown in Table 5.

First of all, from the perspective of the basic information of the overall industrial related network, 149 industrial departments in China are all covered in the overall industrial association network, and the number of directed edges between industrial nodes is 2933, indicating that all industrial departments in China will still be related to other industrial departments through direct or indirect connections after filtering out weak connections. And in the direct industrial related network centered on two typical Chinese entertainment industry sectors, although the broadcast, television, film, and video recording industry is a directly close network within the industry sector nodes and directed edge quantity is higher than that of the entertainment industry department, compared with the overall industry association network, the number of industry sectors in these two direct industry association networks is relatively small. This shows that although China's entertainment industry has experienced a period of rapid development, the industrial sectors that have a strong direct connection with China's entertainment industry still account for a small proportion in all industrial sectors.

Secondly, from the perspective of the average agglomeration coefficient, the average agglomeration coefficients of

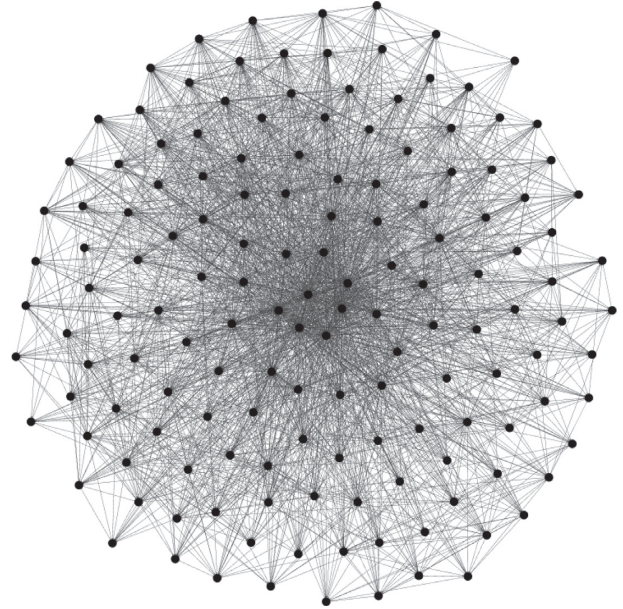


FIGURE 1: China's overall industrial related network.

China's overall industrial association network and the network directly associated with typical industrial sectors of China's entertainment industry are 0.34, 0.51, and 0.50, respectively. On the whole, the average agglomeration coefficient of these three networks is relatively high, indicating that, in the above association network, the association between industrial nodes and their neighboring industries is relatively close. Compared to the overall industry-related network, two direct industrial networks can be found where the Chinese entertainment industry departments of the two typical industries directly affiliate network average concentration coefficient, the related network was greater than the overall industry, and the entertainment industry of China is directly related to individual industry between nodes in the network economy as a whole technical contact more closely.

Thirdly, from the perspective of average shortest path length, the average shortest path length of the three industrial related networks is relatively close, and all of them are relatively short, indicating that no matter the overall network or the direct network, any industrial nodes in the network can generate contact through a shorter path, and the overall circulation speed and efficiency of resources are relatively high. Combined with the average agglomeration coefficient of the three networks, the three industrial related networks constructed in this study all have high average agglomeration coefficient and short average shortest path length, indicating that the above networks all have the structural characteristics of small-world networks.

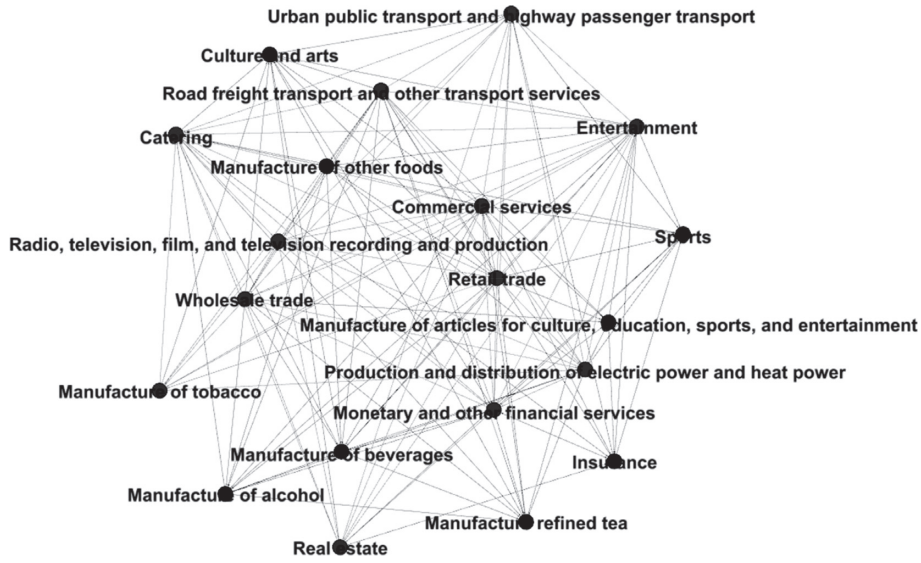


FIGURE 2: Diagram of the direct industrial related network of the entertainment industry.

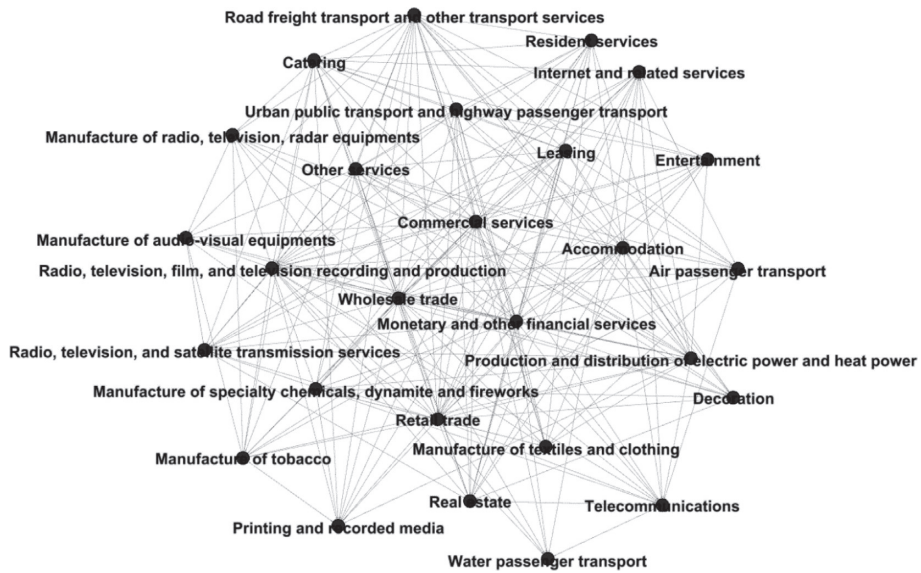


FIGURE 3: Diagram of direct industry connected network of broadcasting, television, film, and video recording production industry.

TABLE 5: Statistical table of the overall structural characteristics of the industrial related network.

Network type	Node number	Edges	Average agglomeration coefficient	Average shortest path length	Network density
The whole industry-related network	149	2933	0.34	2.20	0.13
The entertainment industry is directly connected to the Internet	20	158	0.51	1.70	0.42
The radio, television, film, and video recording industries are directly connected	27	261	0.50	1.70	0.37

Finally, from the perspective of network density, among the three industrial related networks, China’s overall industrial related network has the lowest network density, which is 0.13. The network densities of the other two direct industrial related networks were close, 0.42 and 0.37,

respectively, which were significantly higher than the overall industrial related network. It can be seen from this that, within the direct industrial related network of China’s core industry sectors, strong economic interaction has been established among various industrial sectors, and there is a

TABLE 6: Statistical table of node characteristics of the entertainment industry.

Industry nodes	Node degree		Point-in degree		Point-out degree		Betweenness centrality	
	Value	Rank	Value	Rank	Value	Rank	Value	Rank
Entertainment	20	126	16	97	4	109	283.08	18
Radio, film, television, and (other) audiovisual media	26	101	22	61	4	109	36.71	98

high synergistic effect among various industrial sectors. Its industrial related channels and cooperative behaviors are higher than the overall level of China's industrial related network.

Through the above analysis, it can be found that the directly related networks related to China's entertainment industry generally have the characteristics of high agglomeration coefficient, short average path, and high network density, but few nodes. This feature indicates that although the number of industrial sectors with a strong correlation with China's entertainment industry is obviously rare, the established industrial sectors have frequent interactions and close relationships within the directly related network [15]. Although such network structures will expand the driving ability of each other in the same direction by forming a small industrial group, its influence will be limited by the number of nodes, and it is difficult to drive the common development of most industries.

6.3. Analysis of Node Characteristics of the Entertainment Industry's Related Network. After sorting the overall characteristics of the industry-related network, we study further using Gephi software and put two typical Chinese entertainment industry sectors in China's overall industrial connection within the network. And the node characteristics were more specific; thus, there is a more comprehensive display of the entertainment industry in the positioning and characteristics of the industrial structure of China. The measurement results of network node characteristics of specific industrial sectors are shown in Table 6.

First of all, from the point of view of the node degree of industrial nodes. The nodal degrees of the entertainment industry, which represents the indoor and outdoor tourism and leisure and entertainment industry, and the radio, television, film, and film recording production industry, which represents the film and television entertainment industry, are 20 and 26, respectively, ranking the lowest among all industrial sectors in China. It can be seen from this that China's entertainment industry is still in a marginal position in the national overall industrial network, and the number of industrial sectors with a strong correlation with it is still relatively small.

Secondly, the specific point-in degree and point-out degree are analyzed. The point-in degree and point-out degree of the entertainment industry, which represents the indoor and outdoor sightseeing and leisure and entertainment industry, rank relatively low among all the industry sectors. This indicates that the forward and backward radiating driving ability of the industrial sector is relatively limited after filtering out the weak connection, but in

comparison, the backward correlation is higher than the forward correlation in the industrial related network. However, in the radio, television, film, and film and television recording production industry, which represents the film and television entertainment industry, the point-in degree is 22, ranking relatively higher among 149 industrial sectors in China. This shows that the industrial sector is supported a lot by other industrial sectors in the production process, and it will have a certain radiating and driving effect on the upstream industry during its development.

Finally, from the calculation results of the betweenness centrality of two industrial sector nodes. The betweenness centrality of the entertainment industry and the broadcasting, television, film, and film recording production industry are, respectively, 283.08 and 36.71, which have a great difference in the rank among all Chinese industrial sectors. This indicates that different industrial sectors in China's entertainment industry have different roles in China's national industrial structure. Though the entertainment industry sector, representing indoor and outdoor tourism and leisure and entertainment industry, has a less strong correlation in the industrial related network, it plays a relatively obvious intermediary role in the overall related network. Compared with other industrial sectors, the ability of resource control and regulation is more prominent, and it has a special position in the overall industrial related network.

7. Conclusion

In order to better clarify the characteristics of economic links between China's entertainment industry and other industrial sectors and further clarify the role of the entertainment industry in China's industrial development and its positioning in China's overall industrial related network, this study is based on the dual perspectives: input-output analysis and network analysis. Taking the typical industry sectors of the Chinese entertainment industry—entertainment industry sector and radio, television, film, and film recording production industry sector—as representatives, this paper systematically and quantitatively studies the characteristics of the industrial related and industrial related network structure of Chinese entertainment industry. Through the two-stage research, the following conclusions are drawn.

Firstly, through the characteristics of the Chinese entertainment industry-related network of input-output analysis, the study found that the Chinese entertainment industry as a whole showed the characteristics of the final demand basic industries, in the process of production in the industry, for most other sectors have certain dependence,

and its products and services more directly meet the final demand. This characteristic shows that the development of China's entertainment industry plays a positive role in expanding consumption to a certain extent. Moreover, due to the low intermediate investment rate, the industry has a high added value, which can better create economic value and drive economic growth. It can be seen that, in order to play the role of China's entertainment industry in expanding consumption, it needs the development of other related industries as a foundation. Therefore, in the current economic development environment of China, in addition to the acceleration of the entertainment industry itself, the development of other related industries also needs to be promoted simultaneously.

Secondly, through the analysis of the overall structural characteristics of China's overall industrial association network and China's entertainment industry's direct industrial association network, the research finds that the direct association network related to China's entertainment industry generally has the characteristics of a high agglomeration coefficient, short average path, and high network density, but few nodes. This characteristic indicates that although the number of industrial sectors with strong correlation with China's entertainment industry is obviously rare, the established industrial sectors have frequent interaction and close relationships within the direct correlation network. Although such network structures will expand the driving ability of each other in the same direction by forming a small industrial group, its influence will be limited by the number of nodes, and it is difficult to drive the common development of most industries. Therefore, the research believes that it is necessary to further expand the downstream industrial chain related to the entertainment industry and expand the overall scale of the direct industrial association network of the entertainment industry. This will extend its original industry-driven capacity to more industries.

Third, through the analysis of the overall industry-related network in China to Chinese entertainment typical industry departments within the network node characteristics, the study found that the backward correlation of typical industry sectors of Chinese entertainment is obviously higher than the forward correlation. This shows that in the development process of the Chinese entertainment industry, there will be a certain amount of radiation on the upstream industry. In addition, although Chinese entertainment industry has few strong correlations within the industrial connection network, its core industrial sector has a clear intermediary role in the overall connection network. This shows that Chinese entertainment industry's resource control and adjustment capacity are more prominent than other industrial sectors.

Data Availability

The original data used in this study are from the input-output tables of China released by the National Bureau of Statistics of China in 2019. The original data used to support the findings of this study are available from http://www.stats.gov.cn/zjtj/tjzdgg/trccxh/zlxz/trccb/201701/t20170113_1453448.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Social Sciences Foundation of Shaanxi Province, China (Grant no. 2019R007).

References

- [1] J. He, "Study on development orientation of urban leisure industry and its products—based on an analysis of leisure demand structure and behaviors," *Tourism Tribune*, vol. 23, no. 7, pp. 13–17, 2008.
- [2] J. Xia and Y. Xiao, "Development trend and future orientation of digital entertainment consumption," *Reform*, vol. 32, no. 12, pp. 56–64, 2019.
- [3] C. Tao, M. Xu, and J. Yu, "Government innovation investment, inter-regional correlation and industrial structure upgrading: an empirical analysis of 283 cities in China," *Statistics and Information Forum*, vol. 35, no. 7, pp. 89–100, 2020.
- [4] Q. Qing and Y. Hu, "Leisure industry: review of the related in China," *Economist*, vol. 18, no. 4, pp. 40–46, 2006.
- [5] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [6] Q. Zhao and Q. Yan, "Research on the structural characteristics of China's industrial correlation network," *Statistics and Decision*, vol. 32, no. 15, pp. 104–108, 2017.
- [7] L. Jing, C. Shu, G. Wan, and C. Fu, "Study on the spatial correlation and explanation of regional economic growth in China—based on analytic network process," *Economic Research Journal*, vol. 49, no. 11, pp. 4–16, 2014.
- [8] H. Qian, X. Yang, J. Ding, and L. Yang, "Whole social network for professional athletes in China," *Journal of Wuhan Institute of Physical Education*, vol. 50, no. 7, pp. 77–83+88, 2016.
- [9] Z. Wang, S. Liu, W. Wei, and Y. Yao, "Spatial characteristics of China's local government debt risk correlation network and the influencing factors," *Statistics and Information Forum*, vol. 34, no. 12, pp. 22–31, 2019.
- [10] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [11] F. Wasserman, *Social Network Analysis: Methods and Applications*, China Renmin University Press, Beijing, China, 2011.
- [12] National Economic Accounting Department, National Bureau of Statistics, *Input-Output Tables of China*, China Statistics Press, Beijing, China, 2019.
- [13] T. Wang, "Research on the general feature of Chinese industrial structure based on social network analysis," *Science Research Management*, vol. 35, no. 7, pp. 124–129, 2014.
- [14] D. Zhai and J. Han, "IUR knowledge collaboration research based on networked evolutionary game," *Statistics and Information Forum*, vol. 34, no. 2, pp. 64–70, 2019.
- [15] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.

Research Article

A Resilience-Based Security Assessment Approach for CBTC Systems

Ruiming Lu ¹, Huiyu Dong ¹, Hongwei Wang ², Dongliang Cui ³, Li Zhu ⁴,
and Xi Wang ⁴

¹National Research Center of Railway Safety Assessment, Beijing Jiaotong University, Beijing, China

²State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

³State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China

⁴State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Correspondence should be addressed to Hongwei Wang; hwwang@bjtu.edu.cn

Received 29 June 2021; Accepted 28 September 2021; Published 21 October 2021

Academic Editor: Xuzhen Zhu

Copyright © 2021 Ruiming Lu 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 the rapid development of urban rail transit systems, large amounts of information technologies are applied to increase efficiency of train control systems, such as general computers, communication protocols, and operation systems. With the continuous exposure of information technology vulnerabilities, security risks are increasing, and information is easy to use by malicious attackers, which can bring huge property and economic losses. The communication-based train control (CBTC) system is the most important subsystem of urban rail transit. The CBTC system ensures safe and efficient operation of trains, so the quantitative assessment of cyber security is quite necessary. In this paper, a resilience-based assessment method is proposed to analyze the security level of CBTC systems based on indicators of both the cyber domain and the physical domain. The proposed method can demonstrate the robustness and recovery ability of CBTC systems under different security attacks. Based on the structural information entropy, the fusion of different indicators is achieved. Two typical attacking scenarios are analyzed, and the simulation results illustrate the effectiveness of the proposed assessment approach.

1. Introduction

At present, railway is developing rapidly around the world, especially in China, where the high-speed railway (HSR) has a total length of 35,000 kilometers, accounting for approximately 66.7% of the world's high-speed railways [1]. China has also made significant progress in urban rail transit; there are more than 200 lines, and the total operation length is more than 6000 km [2]. Ensuring the punctuality of trains is the most significant goal of railways, and it can promote the sustainable development and bring the maintenance of social stabilization.

Communication-based train control (CBTC) is the key technology of urban rail transit to keep trains operation safe and efficient, which can provide real-time operation information for trains and generate control and dispatch strategies. In order to increase the automation and

informatization level of CBTC systems, communication, computer, and control technologies have been widely applied [3]. Additionally, security risks are introduced in CBTC systems and can cause the destruction of railway transportation organization, which is the same as the other industrial control systems [4, 5].

Generally, a CBTC system can be taken as a typical cyber-physical system [6], where the computer network is working at the cyber domain while trains are running at the physical domain. Cyber attacks are usually carried out on computer nodes or communication links, which will cause information delay and tampering. Considering the principles of CBTC systems, the normal operation of trains could be disturbed, such as emergency braking. For example, wireless local area networks (WLANs) are adopted as the main method of bidirectional train-ground communications of train control systems [7, 8], which could be easily

interfered and attacked [9] as WLANs work at the public frequency and the authentication mechanism is unidirectional. Once wireless links are cut off under denial of service (DoS) attacks, trains cannot receive the movement authority (MA) from the control center, and emergency braking must be implemented in order to keep trains safe. Obviously, the operation efficiency is seriously reduced.

As urban railways are designed to deliver passengers, CBTC systems are safety-critical, and the fail-safe mechanisms are applied in order to achieve the demanded performance including reliability, availability, maintainability, and safety (RAMS) [10, 11]. In the traditional assessment approach to CBTC systems, RAMS is the significant statistical indicator system [12, 13] according to IEC 62278 [14], where qualitative measures include failure probabilities, mean time to failure (MTTF), mean time between failures (MTBF), and two-dimensional risk matrixes (risk probabilities and risk consequences). Therefore, the existing assessment approach focuses on the large time scale, which cannot determine in real time the effects caused by the temporal or sudden disruption. However, security events are often unexpected, and malicious attacks are implemented depending on the subjective will of attackers, being random. As a result, it is not appropriate to adopt traditional statistics indicators to evaluate performance of train control systems under attack.

As mentioned above, CBTC systems are designed to provide transportation service, and the robustness and recovery capability are critical when cyber attacks are performed. The Department of Homeland Security developed a plan to achieve critical infrastructure security and resilience in 2013 [15]. The transportation systems sector-specific plan [16] was also proposed. It identifies the transportation system's security and resilience priorities and describes the approach to managing critical infrastructure risks, where the railway system is included. Therefore, a novel assessment approach based on resilience is proposed in this paper.

The resilience of a CBTC system could be illustrated as Figure 1 according to [17]. Generally, a CBTC system keeps at the normal operation level, and trains are running according to the predesigned timetable. At t_i^o , the cyber attacks are implemented, and system performance is still kept at the same level. From t_i^o to t_i^d , attackers search the target and inject malware to affect the normal operation. Therefore, system performance begins to go down at t_i^d , and it reaches the lowest point at t_i^m . Meanwhile, some protection mechanisms are triggered to mitigate effects of attacks. From t_i^s , the system begins to recover and reaches a new stable level at t_i^r . Therefore, when a security assessment approach is applied, there are three stages which should be considered: the preevent stage ($t < t_i^o$), the during-event stage ($t_i^o < t < t_i^m$, $t_i^m < t_i^s$), and the postevent stage ($t_i^s < t < t_i^r$, $t_i^m < t_i^s$). Considering the characteristics of cyber-physical systems, performance of a train control system should contain both indicators of network from cyber domain and those of train operation performance from physical domain. Generally, the cyber domain is discrete while the physical domain is continuous. Therefore, the structure information entropy is applied to fuse different indicators [18], which can measure consequences of cyber

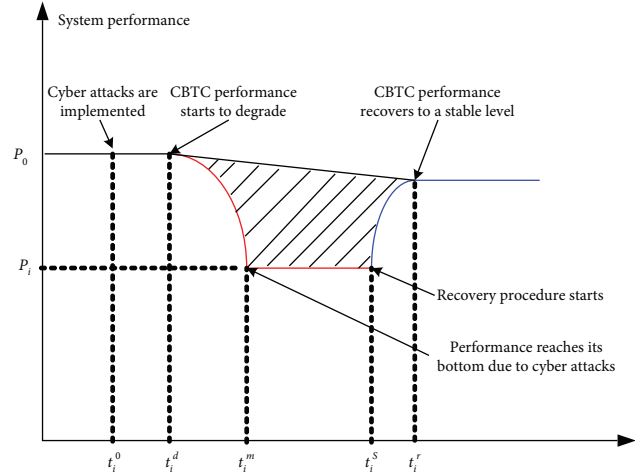


FIGURE 1: The resilience of a CBTC system.

attacks and demonstrate overall performance changes of CBTC systems versus the whole process of security events. In this paper, the resilience of CBTC system is assessed by the structure information entropy.

The rest of this paper is organized as follows. A typical CBTC system is shown in Section 2. Section 3 describes the assessment model based on structural information entropy. Section 4 presents simulation results and some discussions. Finally, we conclude the study in Section 5.

2. Overview of CBTC Systems

Figure 2 demonstrates a typical CBTC system for urban rail transit, which includes some critical equipment, e.g., automatic train supervision (ATS), data storage unit (DSU), computer interlock (CI), zone controller (ZC), and vehicle on-board controller (VOBC). VOBC receives the control command from ZC and transmits the train status through wireless communications, where WLANs and long-term evolution for metro (LTE-M) are usually applied. WLANs-based train-ground communication systems consist of wayside access points (APs) and on-board mobile stations (MSs).

Generally, trains are running at a high speed and sending the corresponding information including velocity, position, and direction to the ZC. ZC generates movement authorities (MAs) to trains to inform the train about the location of the nearest obstacle, which could be a running train, a station, or a turnout. The train obtaining the MA should calculate the permitted maximum velocity to keep a safe distance to the nearest obstacle. During the process, messages between trains and ZCs are transmitted through WLANs or LTE-M. Obviously, the reliability and dependability of wireless communications are significant to CBTC systems.

As mentioned above, the fail-safe mechanisms are embedded in the operating principle of the CBTC system so that when a specific type of failure occurs, it will not cause harm to other equipment, the environment, or the personnel or cause minimal harm. Therefore, redundant and fault tolerance architectures are applied, such as double 2-vote-2 architecture for ZC, DSU, and CI. On the left part of

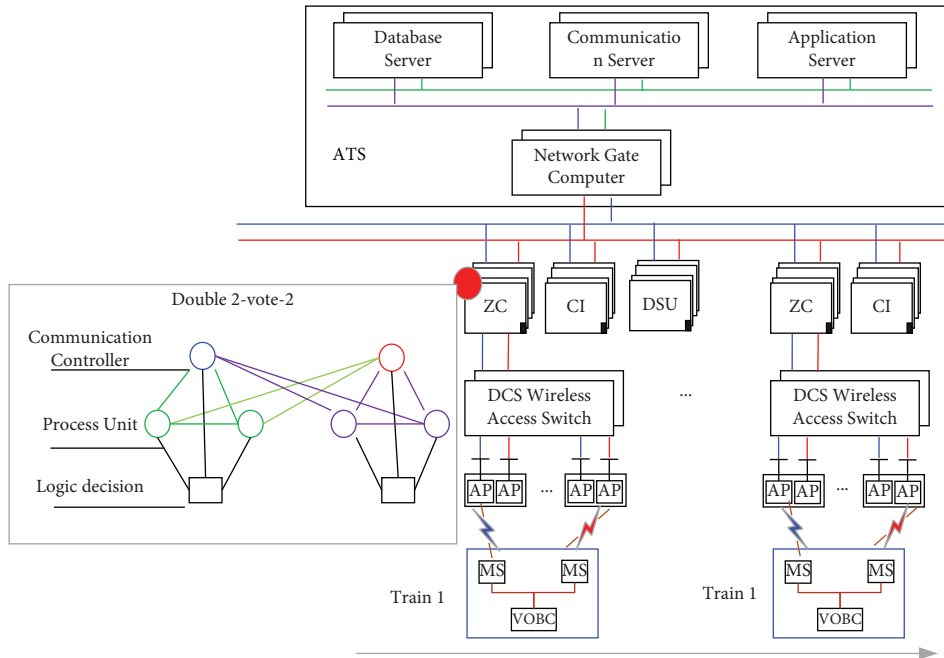


FIGURE 2: A typical CBTC system.

Figure 2, the double 2-vote-2 architecture is demonstrated, where there are two communication controllers (CCs), four processing units (PUs), and two logic decision makers. In the architecture, one CC, two PUs, and one logic decision maker make up the main system while the others are the standby system. Generally, when the main system does not work well, the standby system switches to the main role. Therefore, ZC, DSU, CI, and ATS are not standalone devices but subsystems. For example, ATS includes database servers, communication servers, application servers, and network gate computers. Some dedicated protocols are developed to keep the confidentiality, integrity, and availability of information, such as railway signalling safe protocols (RSSP) derived from EN 50159.

Conversely, applications of general information technologies could bring security risks, such as server message block (SMB) protocol vulnerabilities, remote code execution vulnerabilities, authentication vulnerabilities, DoS threats on wireless communications, and false data injection threats. The combined effects of security threats and vulnerabilities can generally bring changes of CBTC network topology, such as the downtime of one server due to virus, which can lead to interruptions of communications from the server to any other equipment. For some specific scenarios, under protection of fail-safe mechanisms, changes of CBTC topology cannot affect the normal train operation. With dual-network redundancy of wireless communications, although one wireless link between a train and ZC is blocked due to jamming attacks or DoS attacks, the train could still keep the preset running trajectory as the other wireless link can provide one channel to transmit the control command. Therefore, a security assessment approach should consider effects of the existing fail-safety mechanisms, which can

precisely evaluate the practical robustness and the recovery capability of train control systems.

3. The Resilience Assessment Model of CBTC Systems

As mentioned above, cyber domain of CBTC system is a computer network with different computer nodes and communication links. The physical domain consists of trains with effects of traction and braking according to commands from the cyber domain. Obviously, abnormal performance of cyber domain could affect the operation of trains and bring on disturbance to the transportation service of urban rail transit.

According to the definition of resilience, system performance indicators should be determined based on the characteristics of CBTC. As a cyber-physical system, there are amounts of performance indicators of cyber domain and physical domain. Therefore, the performance variance due to cyber attacks should be described based on difference indicators. In this section, we develop a novel method based on the structural information entropy to demonstrate the real-time system performance of both the cyber domain and the physical domain.

3.1. Cyber Domain. As a CBTC system could be treated as a computer network, we built a graph model $G(V, E)$, where $v_i \in V$ is the device of CBTC systems and $e_i \in E$ is the communication link among devices. Two-dimensional structural information of graphs is proposed to quantitatively measure the force of the network to resist cascading failures caused by intentional virus attacks, as the general Shannon's information entropy failed to support

communication network. The definition of two-dimensional structural information entropy is shown as follows:

$$\begin{aligned} H^s(G) &= \sum_{j=1}^L \frac{V_j}{2m} \cdot H \left\{ \frac{d_1^{(j)}}{V_j}, \dots, \frac{d_n^{(j)}}{V_j} \right\} - \sum_{j=1}^L \frac{g_j}{2m} \log_2 \frac{V_j}{2m}, \\ &= - \sum_{j=1}^L \frac{V_j}{2m} \sum_{i=1}^{n_j} \frac{d_i^j}{V_j} \log_2 \frac{d_i^j}{V_j} - \sum_{j=1}^L \frac{g_j}{2m} \log_2 \frac{V_j}{2m}, \end{aligned} \quad (1)$$

where L is the number of modules in partition \mathbb{P} which could be the subsystem of a CBTC system, n_j is the number of nodes in module X_j , $d_i^{(j)}$ is the degree of the i th node in X_j , V_j is the volume of module X_j (i.e., the sum of degrees of all the nodes in X_j), g_j is the number of edges with exactly one endpoint in module j , m is the number of edges in G , and $2m$ is the volume of G .

Equation (1) assumes that each vertex and each edge is completely the same. However, in CBTC systems, different operation systems and hardware platforms are adopted based on functional attributes of devices. Meanwhile, according to safety-critical requirements, RSSP-1 and RSSP-2 are individually applied to the closed network and the open network. As a matter of fact, some private protocols (PPs) are also developed due to specific requirements. For some unsafe communication links, information is transmitted in clear text. Therefore, there are a few types of vertexes and edges, which means every element of a CBTC graph model should be described with specific parameters according to its inherent features.

Based on the password strength, the security protection policies, and the number and level of vulnerabilities, a security factor of a node could be designed. Vulnerabilities could be classified into five levels according to the common vulnerability scoring system (CVSS), where the corresponding weight of a node can be determined.

$$\begin{aligned} VN_i &= \chi \times \frac{\nu_\alpha}{\max(\nu_\alpha)} \times \frac{\nu_\beta}{\max(\nu_\beta)}, \\ \chi &= \frac{n_{pp}}{n_{dp}}, \\ \nu_\alpha &= \sum_{k=1}^N n_k * w_k, \\ \nu_\beta &= -\log_2 N_{op}^L, \end{aligned} \quad (2)$$

where χ denotes the security protection situation of a device, ν_α is the index which demonstrates the overall state of vulnerabilities, ν_β is the measure of the password strength, n_{pp} is the number of practical security protections, n_{dp} is the number of desirable security protections, N is the number of vulnerability classifications, n_k is the number of the k th vulnerabilities, w_k is the weight of the k th vulnerability, N_{op} is the number of the password character sets, and L is the length of the password.

Similarly, for an edge, based on protocols of communication links, the weight of each edge could be determined as follows:

$$\omega_e = S_j \times R_j \times \frac{N_s}{N_d}, \quad (3)$$

where S_j is the security level of the protocol adopted by the communication link j , R_j is the reliability level of the protocol, N_s is the number of protection methods practically adopted by the protocol, and N_d is the number of protection methods which should be adopted by the protocol. Therefore, S_j is determined by the openness of the standard protocol, where the private protocol can be assigned to the maximum value. R_j depends on whether the communication link is wireless, where the value of a wireless link is obviously smaller than that of a wired link. N_d is the maximum value of N_s in the system.

Therefore, the structural entropy of a CBTC system can be formulated as follows:

$$H^s(G(t)) = - \sum_{j=1}^L \frac{V_j(\omega_e)}{2m} \sum_{i=1}^{n_j} \frac{d_i^j(\omega_e, VN_i)}{V_j(\omega_e)} \log_2 \frac{d_i^j(\omega_e, VN_i)}{V_j(\omega_e)} - \sum_{j=1}^L \frac{g_j(\omega_e)}{2m} \log_2 \frac{V_j(\omega_e)}{2m}, \quad (4)$$

where $V_j(\omega_e)$ is the weighted V_j in (1), $d_i^j(\omega_e, VN_i) = VN_i \times d_i^j(\omega_e)$, $d_i^j(\omega_e)$ is the weighted d_i^j in (1), and $g_j(\omega_e)$ is the weighted g_j in (1).

3.2. Physical Domain. The structural entropy in (4) can demonstrate changes of network typologies due to node failures and interruptions of communication links caused by security issues, which is the performance variance of cyber space. However, due to cyber-physical characteristics of CBTC systems, performance variances of physical space should also be considered. Based on the transportation service attribute of CBTC systems, the achievement rate of timetables can be used to describe effects caused by security attacks on train operation. Firstly, the normalized value of the performance loss of a train is expressed as follows, where the min-max principle is applied.

$$\begin{aligned} \Delta p_{\text{norm}} &= \alpha \Delta s_{\text{norm}} + \beta \Delta v_{\text{norm}} + \gamma \Delta t_{\text{arr-norm}}, \\ \Delta s_{\text{norm}} &= \frac{\Delta s - \Delta s_{\text{min}}}{\Delta s_{\text{max}} - \Delta s_{\text{min}}}, \\ \Delta v_{\text{norm}} &= \frac{\Delta v - \Delta v_{\text{min}}}{\Delta v_{\text{max}} - \Delta v_{\text{min}}}, \\ \Delta t_{\text{norm}} &= \frac{\Delta t_{\text{arr}} - \Delta t_{\text{arr-min}}}{\Delta t_{\text{arr-max}} - \Delta t_{\text{arr-min}}}, \end{aligned} \quad (5)$$

where Δs_{min} , Δv_{min} , $\Delta t_{\text{arr-min}}$, Δs_{max} , Δv_{max} , and $\Delta t_{\text{arr-max}}$ are the minimum and maximum value of the variation of the displacement, velocity, and arriving time, and α , β , and γ are the weight of three parameters.

Therefore, the performance of a whole subway line under attack can be formulated as follows, which is the y axis of Figure 1.

$$AR(t) = \frac{p_p(t) - p_l(t)}{p_p(t)} = 1 - \sum_{i=1}^N \frac{\Delta p_{\text{norm}}^i}{N}, \quad (6)$$

where $p_p(t)$ is the train operation performance of the entire line under normal conditions and $p_l(t)$ represents the performance loss of the whole line under attack.

3.3. Resilience of CBTC Systems. Equation (6) demonstrates the overall performance of CBTC systems under attack. With the attacking process being implemented, states of nodes and edges are changing. Therefore, $AR(t)$ and $H^s(G(t))$ are time-varied functions. By combining AR and $H^s(G(t))$, the performance of cyber space and physical space can be monitored, which can demonstrate effects of security attacks on CBTC systems shown as follows:

$$H(t) = AR(t) \times H^s(G(t)). \quad (7)$$

According to the metric proposed in [19], there are three attributes to measure resilience: absorptive capacity, adaptive capacity, and restorative capacity, and the corresponding expression is shown as follows:

$$\rho_j(S_p, H_r, H_d, H_o) = S_p \frac{H_r}{H_o} \frac{H_d}{H_o}, \quad (8)$$

where j is the j th cyber attack, S_p is the recovery speed factor, H_r is the stable level after the system recovers from cyber attacks, H_d is lowest performance level of the system due to attacks, and H_o is the normal performance level of the systems. Obviously, H_r/H_o describes the adaptive capacity while H_d/H_o presents the absorptive capacity.

In addition, the recovery speed factor is determined according to some key timing.

$$S_p = \begin{cases} \frac{t_\delta}{t_{r^*}} \exp^{-a(t_r - t_{r^*})}, & t_r > t_{r^*}, \\ \frac{t_\delta}{t_{r^*}}, & \text{else,} \end{cases} \quad (9)$$

where t_{delta} is the tolerable time before the recovery measures are implemented, t_{r^*} is the time when some initial measures are performed to decrease the effects of attacks, t_r is the time when CBTC system recovers to a stable operation level, and a is a decay factor.

Considering operation principles of CBTC systems, when attacks are performed and cause failures of critical equipment such as ZC, trains will implement emergency braking to keep safe based on fail-safe mechanisms. Obviously, the performance of the whole subway line will fall down to a lowest level H_d , and the corresponding time is t_δ . Due to the existence of backup operation mode, CBTC system can still operate with ZCs and trains will recover from the emergency braking state, which is the initial measure to keep the continuous service. Therefore, t_{r^*} is determined. Finally, after cyber attacks finish, ZCs being attacked can run at a normal state and CBTC systems can return to a stable level H_r which is generally smaller than the normal level H_o . Hence, t_r can also be obtained.

4. Simulation Results and Discussions

4.1. Simulation Description. Take Beijing Subway Yizhuang Line, for example, where there are 13 stations, 6 ZCs, and 6 CIs and the length is 23.3 km. Based on the structure of CBTC systems and the architectures of ZCs, CIs, and ATS, a computer network of CBTC is demonstrated in Figure 3, where the double 2-vote-2 architecture is applied in ZC and CI subsystems.

The normal timetable of Beijing Subway Yizhuang Line is taken as the input of simulations as shown in Figure 4. The typical jamming attack is implemented on train-ground wireless communications. There are two scenarios:

Scenario 1 took ZCs as attacking targets. Generally, ZC failures could cause serious disturbances to train operations, as trains have to perform the emergency braking when they cannot receive MAs from ZCs. Therefore, operators must try their best to repair failures or implement some other emergency response

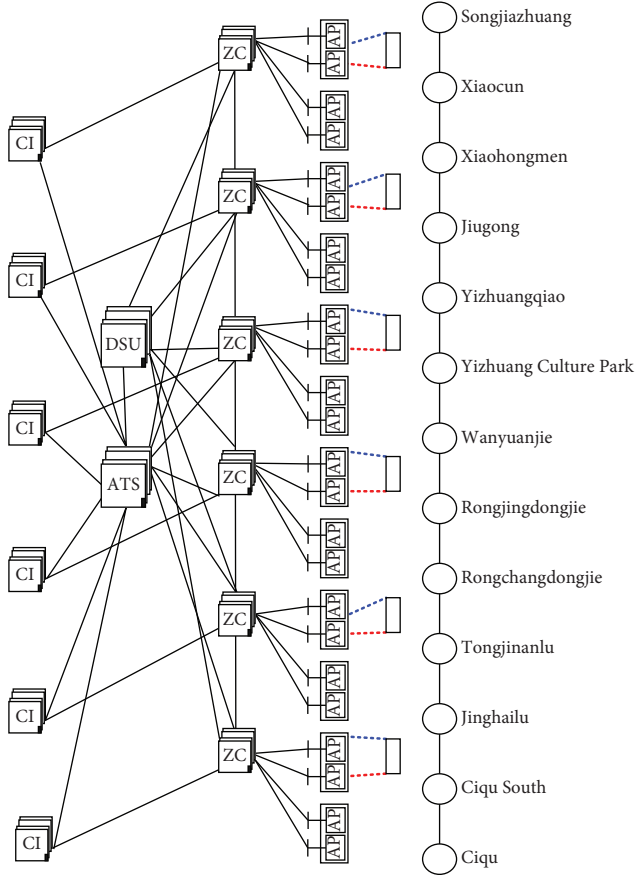


FIGURE 3: The computer network of Beijing Subway Yizhuang Line.

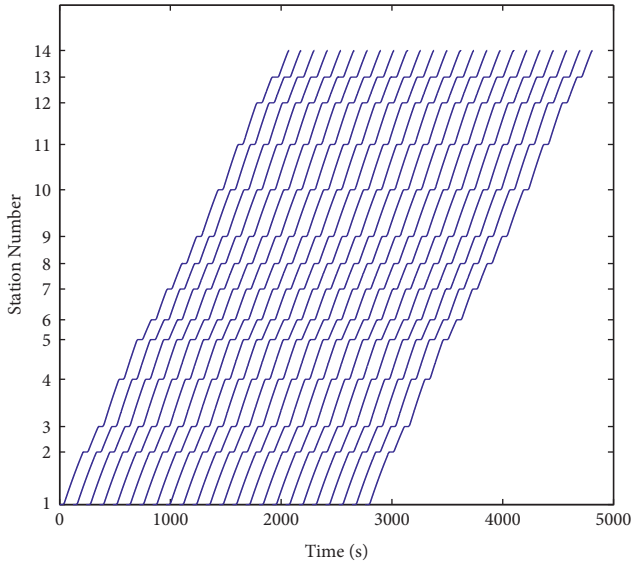


FIGURE 4: The timetable of Beijing Subway Yizhuang Line.

measures. We assume that operators should take several minutes to make the system recover from ZC failures. According to the architecture in Figure 2, the attacking path is CC1 (300 s) \rightarrow PU1 and PU2 (400 s) \rightarrow CC2 (600 s) \rightarrow PU3 and PU4 (700 s). In the scenario, there were three ZC systems being attacked

and crashing. After several minutes, the three ZC systems successively recovered at 2500 s, 2900 s, and 3300 s, respectively.

Scenario 2 took trains as attacking targets, where DoS attacks were implemented on wireless communications between ZCs and VOBCs. Through sending a large number of data packets to exhaust bandwidth resources, communication interruptions could be caused, and trains have to perform the emergency braking to keep safe based on “fail-safe” mechanisms. Therefore, trains worked under the degraded mode depending on operation of drivers until wireless communications recovered to normal. In the scenario, we attacked the 5th, 10th, and 15th trains, respectively, at $t = 1000$ s, $t = 2000$ s, and $t = 2500$ s. Successively, trains ran under the normal mode at $t = 1500$ s, $t = 2500$ s, and $t = 3000$ s.

4.2. Simulation Results

4.2.1. Performance of Cyber Space. Figure 5 shows the network performance based on the two-dimensional structure information entropy under scenario 1, where A, B, C, and D, respectively represent failures of CC1, PU1 and PU2, CC2, and PU3 and PU4 of ZC2. The initial network performance under the normal mode was 7.6754. During the attacking process, the main system and the standby system of ZC2 crashed, and the network performance fluctuated. When ZC3 and ZC4 successively crashed, the network performance reached 7.4068. Then, some measures were implemented to make ZCs recover to normal. Therefore, ZC2, ZC3, and ZC4 started to work normally in sequence, and the network performance quickly returned to the original value before the attack.

Figure 6 demonstrates the network performance under scenario 2, where wireless communications between trains and ZC were blocked by DoS attacks. The network performance had little influence, which means attacks on single or several wireless links could hardly bring obvious changes of the network topology. However, communication interruptions could lead to the emergency braking of trains, which obviously affected the operation of a subway line. Therefore, gentle changes of network performance cannot describe effects of DoS attacks on CBTC systems.

4.2.2. Performance of Physical Space. Figures 7 and 8 present practical timetables under two different scenarios. It can be seen from the timetable that, in the two attack scenarios, the normal operation of the train was greatly affected. Figure 7 indicates that after the first ZC system was compromised, the timetable began to be delayed under scenario 1. With the restoration of the ZC system one by one, the timetable began to recover, but it still had an impact on subsequent train operations. Figure 8 shows that even if there was no impact on the network domain, due to the DoS attack on the wireless communication between ZCs and VOBCs, the train could not obtain MAs, so it led to emergency braking, which still had an impact on the timetable. Therefore, cyber attacks

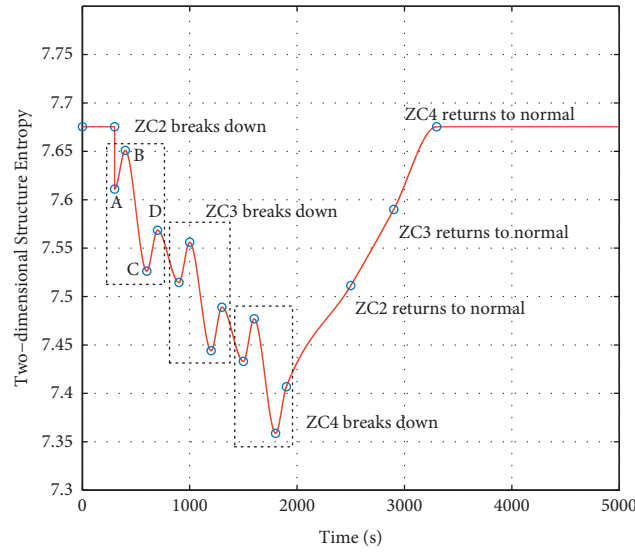


FIGURE 5: Network performance changes of the CBTC system under scenario 1.

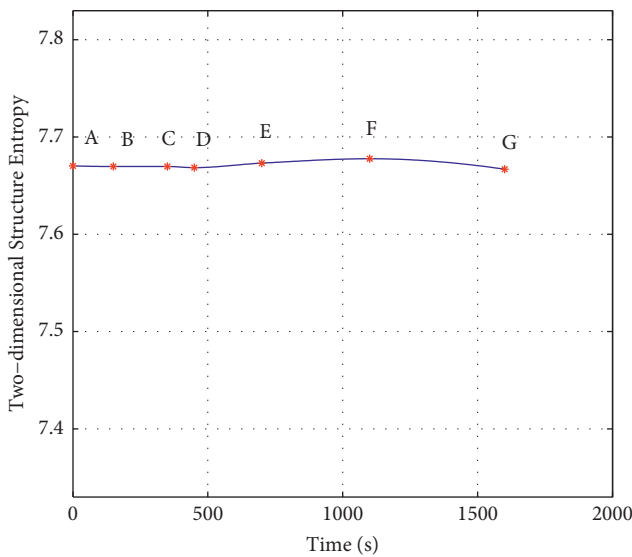


FIGURE 6: Network performance changes of the CBTC system under scenario 2.

can badly affect the normal operation of a whole subway line. It is necessary to evaluate the train performance loss of a whole subway line under cyber attacks.

As shown in Figure 9, the train operation performance (defined in (6)) of the subway line decreased as several ZCs broke down (A ~ B) and then increased with the recovery of ZCs (B ~ C). With the recovery of ZCs, each train returned to the normal operation state. However, the performance loss is irreparable, and the curve could not reach the normal operation level as shown in area C.

The train operation performance of the subway line under DoS attacks on wireless communications is shown in Figure 10. It began to decrease (area A) and fell to the lowest point (area B) at $t = 2300$ s. In order to keep the continuity of transportation service, trains had to recover to the normal operation mode, and

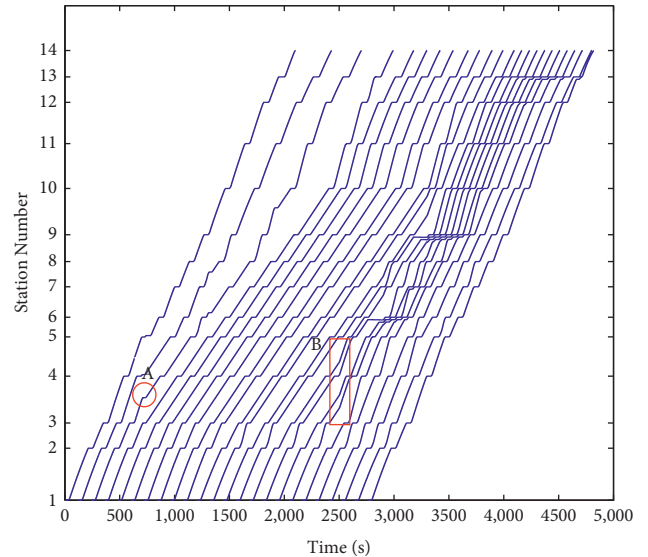


FIGURE 7: Practical timetable under scenario 1.

then the curve started rising. However, trains performing the emergency braking could affect normal operation of following trains in a certain area, and those far from the attacked ones could keep the normal model.

4.2.3. Resilience Assessment of CBTC Systems. As shown in Figures 11 and 12, the performance of the cyber domain and that of the physical domain were integrated, which could demonstrate the security state of the whole subway line under attack, and the corresponding polynomial fitting results were also included.

According to fitting results, the key parameters of (8) and (9) were determined as shown in Table 1. The lowest value of the security level under the two scenarios was close. In scenario 1, failures of one single ZC could affect all the trains within its control. Meanwhile, with the longer attacking

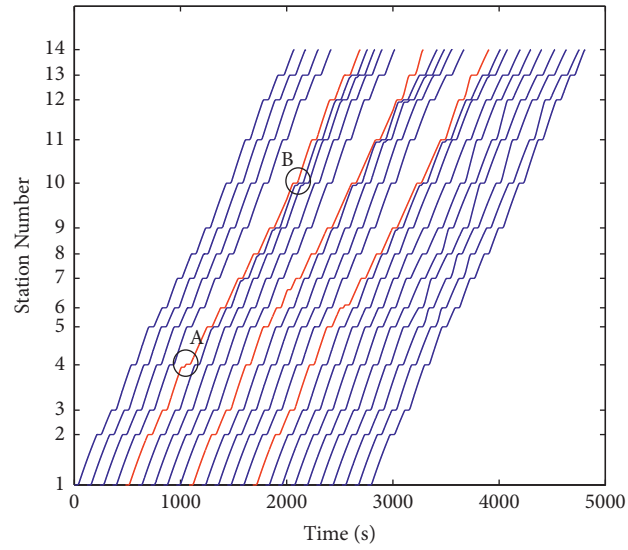


FIGURE 8: Practical timetable under Scenario 2.

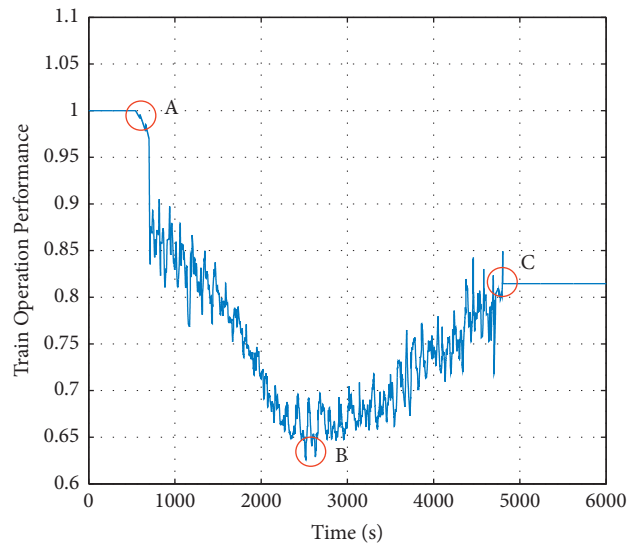


FIGURE 9: The train operation performance of the subway line under Scenario 1.

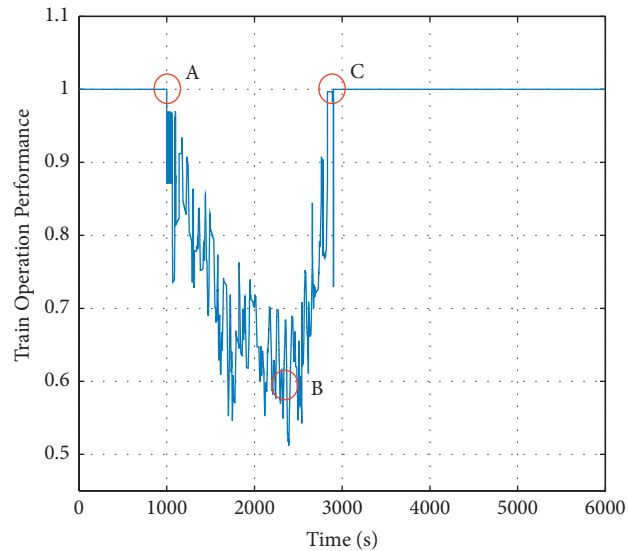


FIGURE 10: The train operation performance of the subway line under scenario 2.

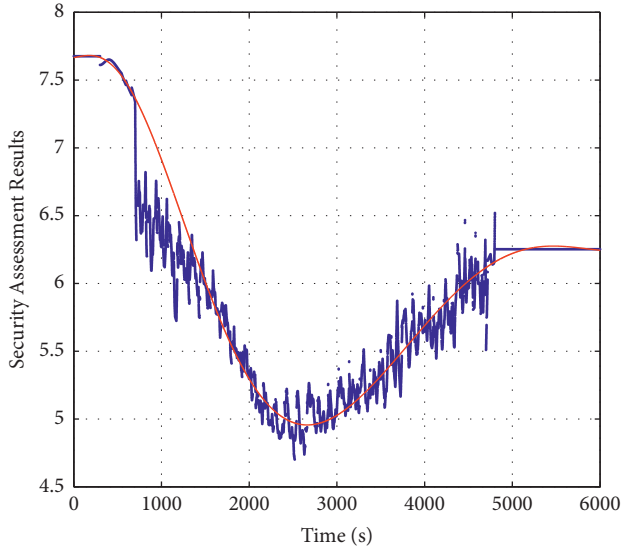


FIGURE 11: The overall performance of the subway line under scenario 1.

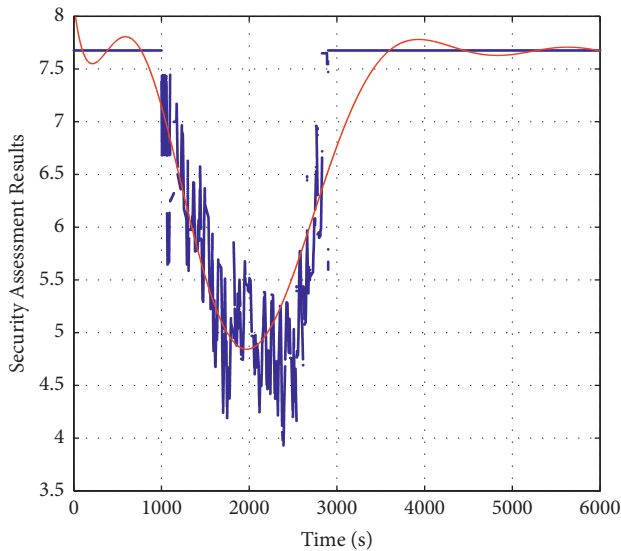


FIGURE 12: The overall performance of the subway line under scenario 2.

TABLE 1: Resilience parameters of CBTC system under two different scenarios.

Resilience parameters	Scenario 1	Scenario 2
H_o	7.6754	7.6754
H_d	4.9438	4.8213
H_r	6.2394	7.6754
t_δ (s)	3400	2500
t_{r^*} (s)	2119	3672
t_r (s)	4821	3672

time, the affected area was wider. Hence, it should take more time to recover to the normal level compared with scenario 2. In addition, interruptions of wireless communication could directly affect performance of trains. Therefore, the

TABLE 2: Resilience assessment results of CBTC systems under two scenarios.

Assessment metrics	Scenario 1	Scenario 2
Absorptive capacity	0.6441	0.6281
Adaptability	0.8129	1
Recovery capacity	1.6049	1.2310
Resilience index	0.6446	1.7734

performance fading rate of scenario 2 was larger. In scenario 2, due to the DoS attack on the wireless communication between ZCs and VOBCs, although it still causes train delays, system performance will return to normal levels after the attack ends.

We could calculate three attributes of resilience as shown in Table 2. The absorptive capacities under the two scenarios were almost the same, which indicated that the CBTC system had similar robustness. As one ZC can control several trains, adaptability and recovery capacity of CBTC systems were weaker under scenario 1. Therefore, resilience can be quantitatively assessed according to the process of attacks.

5. Conclusion

In this paper, a resilience-based assessment approach is proposed to measure the security level of CBTC systems. The two-dimensional structure entropy is adopted to describe the performance of the cyber domain, and that of physical space is calculated according to the practical timetable and running states of trains. Based on stages of attacks, resilience metrics are utilized to analyze the security level of the whole subway line, where both cyber space and physical space are considered. Two typical attacking scenarios were built, and a practical subway line was taken as an example. Simulation results show that the resilience-based approach can efficiently evaluate the security level of CBTC systems under different attacks.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by grants from the Fundamental Research Funds for the Central Universities (No. 2021QY007), National Natural Science Foundation of China under Grant (U18341211, 61925302, 61971030, 61973026), the Railway Traffic joint fund of Beijing Natural Science Foundation and TCT Technology (L181004), Traffic Control Technology (TCT) Innovation Funding under Grant 9907006509, the open project of State Key Laboratory of Synthetical Automation for Process Industries, Beijing Natural Science Foundation: L201002, and Natural Science Foundation of China under Grants: 61973026.

References

- [1] S. Peng, X. Yang, H. Wang et al., "Dispatching high-speed rail trains via utilizing the reverse direction track: adaptive rescheduling strategies and application," *Sustainability*, vol. 11, no. 8, p. 2351, 2019.
- [2] X. Yang, H. Yin, J. Wu, Y. Qu, Z. Gao, and T. Tang, "Recognizing the critical stations in urban rail networks: an analysis method based on the smart-card data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 1, pp. 29–35, 2019.
- [3] R. Pascoe and T. Eichorn, "What is communication-based train control?" *IEEE Vehicular Technology Magazine*, vol. 4, no. 4, pp. 16–21, 2009.
- [4] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "Real time security assessment of the power system using a hybrid support vector machine and multilayer perceptron neural network algorithms," *Sustainability*, vol. 11, pp. 1–18, 2019.
- [5] S. M. Wu, D. Guo, Y. J. Wu, and Y. C. Wu, "Future development of taiwans smart cities from an information security perspective," *Sustainability*, vol. 10, pp. 1–18, 2018.
- [6] L. Bu, D. Xie, X. Chen, L. Wang, and X. Li, "Demo abstract: bachol- modeling and verification of cyber-physical systems online," in *Proceedings of the 2012 IEEE/ACM Third International Conference on Cyber-Physical Systems*, p. 222, Washington, DC; USA, April 2012.
- [7] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 719–728, 2012.
- [8] H. Wang, F. R. Yu, L. Zhu, T. Tang, and B. Ning, "A cognitive control approach to communication-based train control systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1676–1689, 2015.
- [9] H. Wang, F. R. Yu, and H. Wang, "A cognitive control approach to interference mitigation in communications-based train control (cbtc) coexisting with passenger information systems (piss)," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, Article ID 17-0959-3, 13 pages, 2017.
- [10] Y. Cao, H. Lu, and T. Wen, "A safety computer system based on multi-sensor data processing," *Sensors*, vol. 19, no. 4, p. 818, 2019.
- [11] Y. Cao, Y. Zhang, T. Wen, and P. Li, "Research on dynamic nonlinear input prediction of fault diagnosis based on fractional differential operator equation in high-speed train control system," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 1, Article ID 013130, 2019.
- [12] S. Hiraguri, K. Iwata, and I. Watanabe, *A Method of Evaluating Railway Signalling System Based on Rams Concept*, Springer, New York, NY, USA, pp. 97–105, 2011.
- [13] F. Yan, C. Gao, T. Tang, and Y. Zhou, "A safety management and signaling system integration method for communication-based train control system," *Urban Rail Transit*, vol. 3, no. 2, pp. 90–99, 2017.
- [14] E. CENELEC, *Railway Applications the Specification and Demonstration of Reliability, Availability, Maintainability and Safety (Rams)*, The European Standard, Brussels, Belgium, 1999.
- [15] S. O. Johnsen and M. Veen, "Risk assessment and resilience of critical communication infrastructure in railways," *Cognition, Technology & Work*, vol. 15, no. 1, pp. 95–107, 2013.
- [16] U. DHS, *Nipp 2013: Partnering for Critical Infrastructure Security and Resilience*, CreateSpace, Scotts Valley, CA, USA, 2013.
- [17] Q. Zhu, D. Wei, and K. Ji, *Hierarchical Architectures of Resilient Control Systems: Concepts, Metrics and Design Principles*, CRC Press, Boca Raton, FL, USA, 2015.
- [18] A. Li, Q. Hu, J. Liu, and Y. Pan, "Resistance and security index of networks: structural information perspective of network security," *Scientific Reports*, vol. 6, no. 1, p. 26810, 2016.
- [19] R. Francis and B. Bekera, "A metric and frameworks for resilience analysis of engineered and infrastructure systems," *Reliability Engineering & System Safety*, vol. 121, pp. 90–103, 2014.

Research Article

Research on the Relationship between Social Support and Employment Quality of Chinese Athletes from the Perspective of Social Network Structure

Meijuan Cao,¹ Shuairan Li ¹, Wenfei Yue,² and Huanqing Wang ²

¹*Xi'an Physical Education University, Xi'an 710065, China*

²*Sports College, Xi'an University of Architecture and Technology, Xi'an 710311, China*

Correspondence should be addressed to Shuairan Li; 1929939308@qq.com

Received 11 June 2021; Accepted 28 September 2021; Published 14 October 2021

Academic Editor: Shirui Pan

Copyright © 2021 Meijuan Cao 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.

Based on the theories of social network, social support, and retirement process, this study analyzes the source and composition of social support for Chinese athletes on the basis of constructing the social support network. Subsequently, we analyze the impact of social support on employment quality of Chinese athletes from different dimensions and further explore the mechanism of social support on the employment quality of athletes from the moderating role of athletes' self-employment cognition. The study found that the social support network of athletes showed a clear tendency toward "strong ties," and the social support they received mainly came from family members, teammates, and sports team managers. These kinds of social support will directly promote the employment quality of athletes after retirement. When athletes have full knowledge of their future employment status, the effect of social support in promoting employment quality will be further expanded.

1. Introduction

In recent years, with the continuous development of China's sports industry, the building of the leading sports nation has become an important strategic move in China's modernization. Competitive sports have always played a leading role in sports undertakings [1]. As the main body of competitive sports, whether athletes can achieve reasonable arrangements at the three key nodes of athlete selection, training, and transfer has become the core of the sustainable development of China's competitive sports. In order to solve the problem of reemployment and placement of athletes in a better way, General Administration of Sport of China has successively introduced a series of important measures. Education, personnel, labor security, and sports departments at all levels have also actively created conditions to broaden the employment channels for retired athletes and guide athletes to adapt to the social needs. Constructing and perfecting the support system for athletes' transformation have been an indispensable part of building the leading sports nation.

Although China has introduced relevant support policies for the professional transformation of retired athletes, the appearance of low employment rate, narrow employment scope, and low-quality employment still embarrasses them. In order to improve the reemployment quality of retired athletes, we must help athletes establish a correct self-awareness of employment first, and various social resources should be given full play. Studies have shown that, among the many factors that affect the personal employment quality, social support has played a very important role in the conversion of personal careers [2]. Similarly, in the career transition process of athletes, the social support system will also affect their smooth employment in terms of psychological adaptation needs, social survival needs, and individual career development needs. Whether athletes succeed in finding high-quality employment opportunities depends on whether they can make good use of social support. Among them, psychological social support will increase athletes' self-evaluation and the possibility of adopting positive coping styles, thereby obtaining better

reemployment opportunities [3]. It can be seen that social support is a key factor in promoting athletes' employment opportunities.

At present, more and more academic research on the career transformation of athletes is published, but most of it is qualitative research on the transformation and employment status of athletes. Quantitative research on the complex influencing mechanism of athletes' reemployment quality and the sources of influencing factors is less common, especially the empirical research on the impact of social support in athletes' social networks on their reemployment quality. Therefore, in order to explore the influence of mechanism of social support on the reemployment quality of Chinese retired athletes, this study will build a social support network for athletes and clarify the source and composition of social support. The theories of social support and retirement process are the theoretical basis. The impact of social support on the reemployment quality of Chinese athletes is analyzed from different dimensions, and on this basis, the modulating effect of athletes' self-employment cognition is further shown to demonstrate the impact of social support on athletes' reemployment quality. The mechanism of employment quality provides theoretical support and assistance for the proper placement and smooth transition of Chinese athletes after retirement.

2. An Analysis of Social Support Sources of Chinese Athletes from the Perspective of Social Networks

In the early 1970s, psychiatric literature introduced the concept of social support, sociology, and medicine using quantitative assessment methods to conduct a large number of studies on the relationship between social support and physical and mental health, and coping with special stress events. However, the connotation has not been uniformly defined within the various disciplines. At present, the academia mainly defines social support as follows: First, it is defined from the perspective of social interactivity. Sarason (1985) deemed that "social support is an objective existence between people or an interactive relationship that an individual can perceive" [4]. Second, from the perspective of social resources, it is believed that social support is "a form of resource exchange, which stems from the help generated by social relations, which is people's contact information or support network members" [5]. Third, from the perspective of social behavior, it is believed that "social support is the response of individuals or groups to the social needs of others. It refers to both the support and help behavior between people and the occurrence process of such mutual support and help behavior." [6]. Comparing the above three viewpoints, scholars define it from different angles, showing that the concept of social support varies according to the research object and purpose. Among them, the social resources view pays more attention to the types of social resources provided by each member of the social network when researching for social support and the role of various resources in the exchange and interaction between social

network members, such as emotional support, information support, and tool support. The focus of this research is to analyze the impact of the social support received by Chinese athletes on obtaining high-quality employment opportunities and improving the quality of employment, and existing studies have shown that the social resources obtained by individuals through social networks can make it easier to make connections with higher-ranking helpers. Therefore, in order to better reflect these sources of social support that can help athletes improve the quality of employment, this article defines social support from the perspective of social resources; that is, social support is a form of resource exchange, which is generated in the process of interaction between athletes in a support network.

From the above definition of social support, it can be seen that the source of social support must be in the social network where the individual is located. Chinese athletes have received closed training and management since childhood, so their social network structure is of unique characteristics. Therefore, if we want to correctly analyze the impact of social support on the reemployment quality of this group, we must first sort out the source and composition of social support by constructing individual social support networks for athletes.

2.1. The Measurement Index of the Social Support Network of Chinese Athletes. Although individual social network research has some common concepts when describing the network structure, the Chinese athlete social support network is unique, and the concepts of some indicators that need to be measured are different [7]. This study mainly focuses on the following measurement indicators in the social support network:

- (1) Network size: it is mainly used to measure the total number of people in the network. The larger the number, the richer the social support an individual has.
- (2) Relationship constitution: it mainly refers to the proportion of athletes' various relationships in the network. This study mainly involves five types: family members, relatives, classmates or friends, teammates, and team managers. Among them, family members specifically refer to the athlete's parents, spouses, and children; relatives are those other than the aforementioned ones; teammates are the athletes' companions in the same training team; team managers are the athlete's coaches, sparring partners, team leaders, and other daily contact training teams managers; and classmates or friends are friends other than teammates and managers in the team. The social relationships that constitute the athlete's social network do not overlap with each other.
- (3) Ties strength: tie strength can be divided into two types: "strong ties" and "weak ties." There are "interactive method" and "role method" [8] to measure the strong and weak ties. This research combines the

reality of Chinese athletes and adopts the “role method.” In the measurement of the “role method” in the past, “family members and relatives” were often defined as strong ties, and other relationships were defined as weak ties. Through our interviews with athletes, combined with their professional characteristics and closed group training, we here define “family members,” “teammates,” and “team managers” as strong ties and others as weak ties.

2.2. Data Sources of the Social Support Network for Chinese Athletes. The connection time before and after the retirement of Chinese athletes is very short, and the social support they obtain when researching for a job comes from the social network built during their service. Therefore, in order to better reflect the social network characteristics of athletes during this special period, this study used high-level Chinese professional athletes in service as samples when constructing the network. The athlete training management center selects athletes who have participated in large-scale sports events at or above the national level in the past as the survey objects and collects the data required for the construction of the social support network through a combination of on-site and online questionnaires distribution. 500 paper questionnaires and online questionnaires were distributed, 463 valid questionnaires with complete data content were received, and the effective rate of the questionnaire was 92.6%.

2.3. Analysis of Social Support Network of Chinese Athletes. In the survey, this study refers to van der Poel’s social support questionnaire and divides it into practical support and emotional support. Each type of subnetwork is described with a question. Among them, the question of emotional support is, when you are depressed because of certain problems, such as quarreling with people around you or dissatisfaction with training, life, etc., who do you usually talk to? The practical support question is, when you encounter difficulties in training, living, or finding a job, who do you think will help you? Through the research on the above issues, the number of helpers or associates who provide this type of social support to athletes will be obtained, so as to clarify the main source and composition of social support for professional athletes in China. The analysis results are shown in Table 1.

Firstly, with regard to the network size, the sizes of athletes’ practical and emotional support network are 4.12 and 5.13, respectively, which are relatively small. Among them, the emotional support network size is slightly larger than that of the practical support network, which indicates that the practical social support is not so important in the training and daily life, which may be related to their life pattern. Since Chinese athletes’ training centers are managed in a closed way and all foods and accommodation are provided uniformly, this may lead to low demands in this aspect.

Secondly, concerning the relation composition, in the two specific social support networks, the proportion of “teammates” is the highest, which is 40.06 and 59.49,

respectively, indicating that in daily life no matter the type of support is, practical things or emotional psychological help, most of it comes from their teammates in the sports team. In addition, family and friends also offer more social support to athletes. In terms of practical support, family support is slightly more than that of classmates and friends, while in terms of emotional support, athletes rely more on classmates and friends. In addition, it should also be noted that athletes are destined to contact more the coaches and other personnel due to their daily life and training, so they offer more practical support to them.

Finally, in respect of the strong and weak ties, the strong ties between the two kinds of social support networks account for more than 70%, showing a strong tendency toward “strong ties.” Because it is difficult for Chinese athletes to get in touch with other groups in the society in the training process, their social objects are mostly concentrated in family members, teammates, and team managers, which leads to their narrow social circle and limited social support from weak ties.

Through the network analysis, it can be seen that athletes are limited by the closed social environment, and most of their social support comes from strong ties networks such as family members, teammates, and sports team managers. Among these social support sources, teammates and family members account for a higher proportion.

3. Theoretical Basis and Research Hypothesis

3.1. Social Support and Employment Quality of Professional Athletes. The employment quality originated from the labor concept advocated by the international labor field in the 1990s. With the gradual deepening of research in recent years, its connotation is constantly evolving and expanding, but in general the employment quality is the synthesis evaluation of the results of individual employment behaviors, including aspects like labor compensation and benefits, employment safety, career development, and work and life satisfaction [9]. Judging from the performance of the employment problems faced by retired athletes, they mainly concentrated in the low level of labor remuneration and social security, lack of stability in employment, lack of clear career promotion channels, and low overall job satisfaction. Therefore, it is crucial to find out the factors that promote the employment quality of athletes after retirement for improving the employment placement situation.

Relevant studies have shown that social support is a particularly important one among the many factors that affect individual employment quality, and this also applies to athlete groups. After athletes retirement, the social support provided by their social network not only gives them emotional accompaniment to ensure that they can conduct effective emotional communication and information exchange, but also provides them with more high-quality and reliable job opportunities to help them establish a new lifestyle in a better way [10]. Other studies found that frequent communication with parents, coaches, and peers and support from them can enhance athletes’ confidence in dealing with issues related to career development, which not

TABLE 1: Analysis results of social support network for Chinese athletes.

Social support network	Network size	Relation composition (%)					Strong/weak ties (%)	
		Family members	Relatives	Classmates or friends	Teammates	Team managers	Strong ties	Weak ties
Practical support	4.12	25.99	3.29	24.19	40.06	6.48	72.52	27.48
Emotional support	5.13	12.47	2.48	22.84	59.49	2.72	74.68	25.32

only enables athletes to correctly accept the new living environment, but also helps them transfer their related athlete qualities to improve their evaluation of their work and life while entering a new position [11]. These forms of tangible and intangible social support have greatly enhanced the overall quality of employment of athletes after retirement. Therefore, paying attention to the social support received by athletes is of great significance to the improvement of the quality of employment of athletes after retirement.

Based on the above analysis, this research proposes the following research hypothesis:

H1: social support has a significant positive impact on the employment quality of Chinese athletes

From the source and composition of the social support of athletes, it can be seen that most of the social support comes from “strong ties” groups such as family members, teammates, and sports team managers. Therefore, in this study, the social support of can be subdivided according to the source into family support, teammate support, and sports team support. Therefore, research hypothesis H1 can be further subdivided into the following:

H1a: family support has a significant positive impact on the employment quality of Chinese athletes

H1b: teammate support has a significant positive impact on the employment quality of Chinese athletes

H1c: sports team support has a significant positive impact on the employment quality of Chinese athletes

3.2. The Relationship between Social Support, Employment Perception, and Employment Quality of Professional Athletes. Social support has an impact on the employment quality of retired athletes by giving professional athletes emotional companionship, helping athletes establish career choices, and providing high-quality employment opportunities. This impact mainly refers to the positive benefits produced by using the main force of social support to work after retirement [12]. In the process of reemployment, some athletes will develop advance understanding of their role positioning, ability evaluation, employment environment, career development willingness, etc. before employment, and make corresponding preparations to better grasp the employment opportunities of social support and obtain a better quality of employment. This kind of cognitive process in which individuals take the initiative to understand and prepare for the employment process is employment cognition [13].

Studies show that employment cognition is one of the important factors related to the effect of social support on the employment quality of athletes after retirement.

Retirement process theory believes that the retirement of athletes is a natural process of sports career development, which means that athletes enter another social culture from one social culture, and the renewal of their understanding of careers and future choices is also a natural process of change. When athletes retire and transitions without clear employment cognition, their career choice behavior will be in a passive state; on the other hand, athletes with clear employment cognition will actively strengthen their self-transformation cognition, improve work cognition ability, and be active in understanding the status of employment and actively constructing a career vision, which enables them to maximize the use of opportunities provided by social support they receive, thereby further enhancing the impact of social support on the employment quality [14]. Therefore, through theoretical analysis, the study believes that employment cognition may play a positive moderating role in the relationship between athletes’ social support and employment quality. Therefore, the study puts forward the following hypotheses:

H2: employment cognition has a positive moderating role in the influence of social support on the employment quality of Chinese athletes

H2a: employment cognition has a positive moderating role in the influence of family support on the employment quality of Chinese athletes

H2b: employment cognition plays a positive moderating role in the influence of teammate support on the employment quality of Chinese athletes

H2c: employment cognition plays a positive moderating role in the influence of sports team support on the employment quality of Chinese athletes

Based on above hypotheses, the theoretical model framework constructed in this study is shown in Figure 1.

4. Research Design

4.1. Research Objects and Data Sources. In the research process of this study, Chinese high-level athletes who are reemployed after retirement and who have participated in national level and above large games were selected as the research object. They are from athletes training management centers in Shaanxi, Shanxi, Tianjin, and Fujian. Through the questionnaires issued and distributed on-site and online, a combination of related research data was collected. A total of 500 questionnaires were distributed, among which 437 were valid questionnaires with complete data content, with an effective rate of 87.4%. The descriptive statistics for the sample data are shown in Table 2.

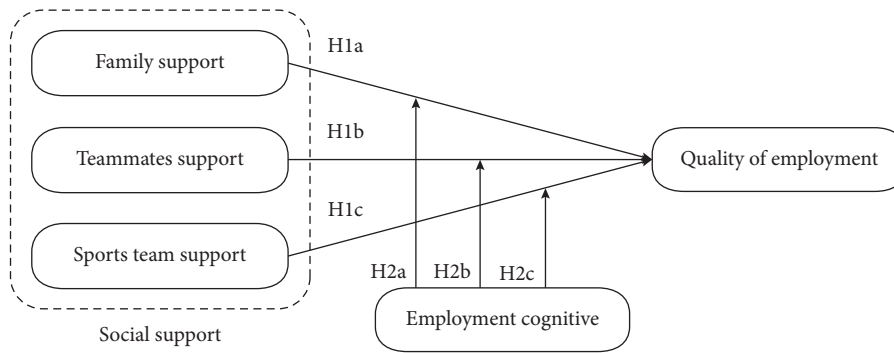


FIGURE 1: Schematic diagram of theoretical model.

TABLE 2: Analysis results of descriptive statistics.

Variables	N	Statistics ($X \pm S$)	Variables	N	Statistics (%)	
Quality of employment	437	3.68 ± 0.98	Gender	Female	204	46.7
Family support	437	3.44 ± 1.21		Male	233	53.3
Teammates support	437	3.47 ± 1.09	Education level	Bachelor's degree	358	81.9
Sports team support	437	3.53 ± 1.14		Master's degree or above	79	18.1
Employment perception	437	4.08 ± 0.51	Training years	<9 years	43	9.8
Age	437	29.63 ± 3.48		≥ 9 years	394	90.2

4.2. Variable Selection and Measurement Method. The core variables involved in this study include athletes' employment quality, employment perception, family support, teammate support, and sports team support. Among them, the explained variable is the employment quality of retired athletes; the core explanatory variables reflected three different sources of social support: family support, teammate support, and sports team support. The moderating variable was the athletes' employment perception near retirement. In addition to the above core variables, this paper also selects four variables that can reflect the individual characteristics of athletes, namely, gender, age, education level, and years of sports training, as the control variables.

In the measurement of core variables, the research used Likert 5-level scale to collect relevant information of the research objects. The options of each variable are "completely disagree = 1," "basically disagree = 2," "not sure = 3," "basically agree = 4," and "completely agree = 5." After calculating the scores of each scale, the measurement index of the sample was constructed by means of mean processing. Among them, social support was measured using the revised SSQA scale compiled by Du et al. (2020) [15], which was divided into three dimensions: family support, teammate support, and sports team support. Employment cognition refers to the quantity of career development and career attitude compiled by Thomas et al. (2010) [16], which is mainly measured from the four aspects of cognition of employment purpose, cognition of self-employability, cognition of employment situation, and intention of career development. Since there is no systematic measurement scale for employment quality, this study, on the basis of the research results of Erhel (2015) [17] and Bedeian (1991) et al. [18], measured athletes' subjective feelings on their relative

salary level, job stability, social security, career development, and job satisfaction.

4.3. Data Analysis Methods. In order to examine the influence of mechanism of social support and employment perception on the employment quality of Chinese athletes, the data analysis of this study is mainly divided into two stages. In the first stage, the study will build a multiple linear regression model of the impact of social support on the employment quality of Chinese athletes and analyze the main effects of family support, teammate support, and sports team support on the employment quality of athletes. In the second stage, the study will introduce employment cognition to construct interaction terms, and test the moderating effect of employment cognition through hierarchical regression model, so as to further clarify the influence mechanism between social support, employment cognition, and athletes' employment quality, so as to verify the theoretical hypotheses proposed by the study and finally draw research conclusions.

4.4. Reliability and Validity Tests. Since the scale used in this study was processed and modified on the basis of the original scale, in order to ensure the validity of the obtained data, it is necessary to test its reliability and validity first.

Questionnaire reliability method is as follows: through the internal consistency method, Cronbach's α coefficient was used to test the reliability of the recovered sample data. It was found that the α coefficient of the variables involved in this study, such as family support, teammate support, sports team support, employment perception, and employment quality, was between 0.838 and 0.913, all greater than 0.7,

showing high reliability. This indicates that the data measured by the 5 variable scales in this study are consistent, and the results obtained can be regarded as ones derived from a consistent measurement method and treated as an aggregate average.

Questionnaire validity is as follows: the KMO and Bartlett sphere test results show that the KMO value of the scale involved in the research is $0.740 > 0.5$ and the significance of the Chi-square statistic of the Bartlett sphere test is $0.000 < 0.1$. Through the significance test, further factor analysis can be performed. Through exploratory factor analysis of 22 items, a total of 5 common factors with characteristic roots greater than 1 were extracted. Among them, factors 1 to 5 are employment cognitive, family support, sports team support, teammates support, and quality of employment. The analysis results showed that the question factor loads of the five factors were all greater than 0.5, and there was no cross factor phenomenon, indicating that the measurement questions of the questionnaire had good convergent validity and met the research requirements.

5. Empirical Analysis of the Influence of Social Support and Employment Perception on Athletes' Employment Quality

5.1. Analysis of the Influence of Social Support on the Employment Quality of Chinese Athletes. In order to verify the influence of the three different dimensions of social support on the employment quality of Chinese athletes and at the same time screen out the influence of other noncore variables, this study adopts the regression method and divides it into two models to verify it. Only four control variables were added to Model 1, which were gender, age, training years, and education level. The purpose of Model 1 was to investigate whether the control variables had an impact on the employment quality of athletes. On the basis of Model 1, Model 2 added core explanatory variables reflecting the three dimensions of social support to observe whether the influence of social support and control variables on athletes' employment quality changed. As the explained variable employment quality is a continuous value, OLS regression model is adopted for statistical analysis. The results of empirical analysis are shown in Table 3.

The results showed that the fitting degree of Model 1 and the significance of the overall model were poor ($R^2 = 0.018$, $F = 1.960$, $P < 0.100$), indicating that the control variables had almost no impact on the employment quality of Chinese athletes. Therefore, it can be judged that gender, age, training years, and education level have no influence on the employment quality of athletes. The fitting degree of Model 2 and the significance of the overall model were high ($R^2 = 0.789$, $F = 228.926$, $P = 0.000 < 0.01$), and the three core explanatory variables all passed the significance level test of 1% ($P < 0.01$).

In previous related studies, personal factors such as the gender and education level of job seekers often have an impact on the quality of their employment. However, the analysis results of Model 1 show that these personal factors

TABLE 3: Analysis results of the influence of social support on the employment quality of Chinese athletes.

Variable	Model 1	Model 2
	Standardization factor	Standardization factor
Gender	-0.068 (-1.405)	0.003 (0.150)
Age	0.022 (0.450)	0.009 (0.386)
Training years	0.066 (1.385)	-0.004 (-0.169)
Education level	-0.095 (-1.975)	-0.044 (-1.959)
Family support	—	0.348*** (8.484)
Teammate support	—	0.316*** (7.246)
Sports team support	—	0.287*** (7.106)
R^2	0.018	0.789
F	1.960	228.926***

Note. * $P < 0.1$, ** $P < 0.05$, and *** $P < 0.01$.

in the athlete group do not have a significant impact on the quality of their employment. Research believes that the main reason for this phenomenon is that the reemployment of Chinese athletes is unique. Most professional athletes will engage in work related to their sports after retirement, such as coaches and sports management staff. In these areas of work, compared to the education level, recruiters value the athlete's professional ability and understanding of sports management more. Such employment environment and work content have also reduced the unequal treatment of athletes due to gender to a certain extent, which is different from previous studies.

According to the standard regression coefficient of the three core explanatory variables in Model 2, the coefficient of family support is the largest, which is 0.348, indicating that family support has the greatest influence on the employment quality of Chinese athletes. Every 1% increase in the score of family support will increase the score of the employment quality of athletes by 0.36%. Although the influence of teammate support and sports team support on athletes' employment quality is slightly lower than that of family support, the coefficient is still around 0.3, and the positive influence on Chinese athletes' employment quality is also very significant. Accordingly, it can be concluded that social support has a significant positive impact on the employment quality of Chinese athletes, which indicates that the higher the family support, teammate support, and sports team support are, the higher the employment quality of retired athletes will be. Therefore, hypothesis 1 of the study is verified.

Research has shown that there is a strong relationship between the support of family, teammates, and sports teams and the transition to retirement. When athletes are faced with transition decisions after retirement, the support of parents, peers, coaches, or team leaders will have a positive impact on athletes' career exploration and choice [19]. Existing literature shows more that athletes receive help from peers, sports teams, and families during the transition to retirement [20]. On this basis, this study further found that the information, resources, and actions from the social network will directly promote the employment quality of athletes after retirement. Because Chinese athletes training

in a closed system, the sports career during difficult to access to other kinds of work, so after the athlete retires, facing the sudden arrival of career transition, on the one hand they will produce in the psychological anxiety and tension, on the other hand they completely on your own is difficult to find suitable and quality work. At this time, family, the team teammates, as well as the long-term exposure to team management support for the athletes, which can give great extent to alleviate the employment pressure of athletes, and retired after the transformation of career development to provide substantial material support, help them get higher pay and better benefits, personal more satisfying jobs, and it has brought substantial promoting effect to the improvement of employment quality.

5.2. Analysis of Employment Perception Moderating Effect on the Relationship between Social Support and Employment Quality of Chinese Athletes. In order to further explore the influence of mechanism of social support on athletes' employment quality, the study further tested the moderating effect of employment cognition on the basis of the main effect analysis of social support. In this study, employment quality was taken as the dependent variable; family support, teammate support, and sports team support as the independent variable; and employment perception as the moderating variable to build the following hierarchical regression model:

$$\text{Model 3: } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \varepsilon$$

$$\text{Model 3': } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \alpha_6 \text{FS} * \text{EC} + \varepsilon$$

$$\text{Model 4: } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \varepsilon$$

$$\text{Model 4': } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \alpha_6 \text{TS} * \text{EC} + \varepsilon$$

$$\text{Model 5: } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \varepsilon$$

$$\text{Model 5': } QE = \alpha_0 + \alpha_1 \text{CONTROLS} + \alpha_2 \text{FS} + \alpha_3 \text{TS} + \alpha_4 \text{STS} + \alpha_5 \text{EC} + \alpha_6 \text{STS} * \text{EC} + \varepsilon$$

In order to eliminate the influence of multicollinearity, the variables FS, TS, STS, and EC were obtained by centralized processing on the basis of the aggregate mean of the original data. FS * EC, TS * EC, and STS * EC are the interaction terms of family support, teammate support, sports team support, and employment perception, respectively, representing the moderating effect of employment perception on the relationship between social support and athletes' employment quality. When the coefficient α_6 of Models 4', 5', and 6' is significant, it indicates that employment perception has a moderating effect. When the coefficient is positive, it has positive regulation effect; otherwise, it has negative regulation effect. Due to space constraints, only the analysis results of Models 4', 5', and 6' are presented here, as shown in Table 4.

The results show that the fitting effects of the three models are all good (R^2 [Model 6'] = 0.799; R^2 [Model 7'] =

0.803; R^2 [Model 8'] = 0.794). The results of Model 6' ($F = 189.169$, $P = 0.000 < 0.01$) show that the coefficient of interaction between employment cognition and family support is 0.095. The significance level test of 1% indicates that the establishment of good employment cognition in advance can guide athletes to better understand themselves and clarify their future career development direction, so that they have the initiative to choose in the process of employment, which effectively protects the effectiveness of the career development of retired athletes. After retirement, the most direct practical problem athletes face is the poor employment situation caused by single skills, little work experience, and lack of accurate employment service platform.

Family, as a harbor for athletes, provides financial, material, emotional, and other support for the entire life course of athletes. The accumulation and help of family capital can enhance the labor value of retired athletes and improve their employment quality. When athletes accurately assess their own positioning in advance and understand the employment environment, they can better grasp their personal abilities and specialties, determine the best career development direction, and more easily choose a career path that meets their interests and specialties during the accumulation of family capital, which enhances the employment quality of retired athletes. Therefore, employment perception has a significant positive moderating effect on the impact of family support on the employment quality of Chinese athletes, which supports hypothesis 2a.

The results of Model 7' ($F = 193.747$, $P = 0.000 < 0.01$) show that the interaction coefficient of employment cognition and teammate support is 0.119, and through the significance level test of 1%, it is shown that the influence of teammate support on athletes' employment quality will be positively adjusted by athletes' own employment cognition. After Chinese athletes enter professional training, they spend most of their time with their teammates day and night. In their spare time, more spiritual support comes from the communication with their teammates. A harmonious relationship can bring great psychological comfort to athletes, make them feel a sense of belonging, and create a comfortable and stable psychological environment. After retirement, athletes often fall into the "social anxiety," inferiority, and frustration syndrome due to the difference in social status and confusion about their future development, which will reduce their satisfaction with the new job and life style and then affect the quality of employment. However, employment cognition can often improve their sense of self-control and psychological quality, which undoubtedly enhances the ability of athletes to resist transition and emotions, so that the psychological comfort given by teammates can play a better role, help them enhance their confidence in career transition, and better relieve the negative psychology in the face of the new work environment. Furthermore, the positive influence of peer support on employment quality was expanded. Therefore, research hypothesis 2b holds.

The results of Model 8' ($F = 182.695$, $P = 0.000 < 0.01$) showed that the coefficient of interaction between employment perception and sports team support was 0.055, which

TABLE 4: The moderating effect of employment perception on the relationship between social support and athletes' employment quality.

Variable	Model 3' Standardization factor	Model 4' Standardization factor	Model 5' Standardization factor
Gender	-0.011 (-0.504)	-0.018 (-0.802)	-0.003 (-0.152)
Age	-0.002 (-0.075)	-0.001 (-0.036)	-0.005 (-0.210)
Training years	-0.007 (-0.313)	-0.007 (-0.303)	-0.004 (-0.185)
Education level	-0.044 (-1.975)	-0.034 (-1.543)	-0.043 (-1.904)
Family support	0.349*** (8.706)	0.346*** (8.702)	0.348*** (8.560)
Teammate support	0.329*** (7.631)	0.339*** (7.932)	0.322*** (7.338)
Sports team support	0.283*** (7.166)	0.271*** (6.903)	0.293*** (7.312)
Employment perception (centralized)	-0.034 (-1.444)	-0.015 (-0.642)	-0.039 (-1.635)
Family support * employment perception	0.095*** (4.251)	—	—
Teammate support * employment perception	—	0.119*** (5.165)	—
Sports team support * employment perception	—	—	0.055** (2.428)
R2	0.799	0.803	0.794
F	189.169***	193.747***	182.695***

Note. *** $P < 0.01$, ** $P < 0.05$, and * $P < 0.1$.

passed the significance level test of 5%, indicating that the influence of sports team support on athletes' employment quality would be positively moderated by athletes' employment perception. Judging from the current employment choices of Chinese athletes, most of them hope to be engaged in jobs related to professional sports after retirement, such as coaches and team leaders. In addition, the former sports teams' support of management personnel plays an important role in meeting their career development needs. Under this condition, if the athletes have a clear cognition of purpose and personal ability to engage in professional sports-related work in the future, they will better grasp the job opportunities offered by the sports teams, thus improving the quality of future employment. Therefore, employment perception has a significant positive moderating effect on the influence of sports team support on the employment quality of Chinese athletes. This conclusion supports research hypothesis 2c.

6. Conclusions

In order to explore the influence of mechanism of social support on the employment quality of Chinese athletes after retirement transition, this study uses the social support network of Chinese athletes to clarify the source and composition of the social support of this group and takes social network theory, social support theory, and retirement process theory as the theoretical basis. This paper analyzes the influence of social support on the employment quality of Chinese athletes from different dimensions and further explores the mechanism of social support on the employment quality of athletes from the regulating effect of athletes' self-employment cognition. Through network analysis and empirical test, the research mainly draws the following conclusions.

Firstly, through the construction of Chinese athletes' social support network, the research finds that, because of the closed management mode and social environment, the social support network showed a strong tendency to "strong ties." The group gained emotional and practical support mostly from family, teammates, and sports management and their training daily life related group, among which family

and social support from their teammates are higher. Therefore, in order to provide more social support for the future work transformation of Chinese athletes, the study believes that professional athletes should be given a more relaxed social environment and social opportunities in daily management to help them establish a wider social network.

Secondly, social support will have a significant positive impact on the employment quality of Chinese athletes after retirement transition from three dimensions: family support, teammate support, and sports team support. The more the emotional care and practical help received from their families, teammates, and sports teams in their social networks are, the higher the quality of their postretirement transition employment will be.

Thirdly, employment perception has a significant positive moderating effect on the influence of social support on the employment quality of Chinese athletes. When the athletes have a full understanding and preparation of their role positioning, employability, employment environment, and career development intention, they can better grasp the social support and employment help from their families, teammates, and sports teams, thus further improving the promotion effect of social support on their employment quality. Therefore, the study believes that, in addition to paying attention to the training of athletes' professional capabilities, Chinese athletes' management departments should also establish a more complete career development plan to help them envision their future transformation and development directions, so as to make better use of the high-quality job opportunities brought by social resources.

Data Availability

The original data used in this study are the questionnaire data obtained from the survey, which are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] B. Ren and X. Zhu, "Research on the measurement and path of China's high quality openness in the new era," *Statistics & Information Forum*, vol. 35, no. 09, pp. 26–33, 2020.
- [2] X. Zhang and J. Yan, "Comparison of the influences of human capital and social capital on the income of retired athletes in China," *Journal of Shanghai University of Sport*, vol. 44, no. 4, pp. 31–40, 2020.
- [3] J. Wang, "A quantitative analysis on the retirement awareness and psychological state of Chinese athletes," *Acta Psychologica Sinica*, vol. 4, pp. 496–506, 2008.
- [4] I. G. Sarason, B. R. Sarason, E. H. Potter, and M. H. Antoni, "Life events, social support, and illness," *Psychosomatic Medicine*, vol. 47, no. 2, pp. 156–163, 1985.
- [5] B. H. Gottlieb, "Assessing and strengthening the impact of social support on mental health," *Social Work*, vol. 30, no. 4, pp. 293–300, 1985.
- [6] C. D. Ryff and B. Singer, "Interpersonal flourishing: a positive health agenda for the new millennium," *Personality and Social Psychology Review*, vol. 4, no. 1, pp. 30–44, 2000.
- [7] H. He and G. Deng, "Multi-dimensional ordered clustering method based on common trend extraction," *Statistics & Information Forum*, vol. 35, no. 12, pp. 15–20, 2020.
- [8] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [9] Z. Peng, G. Lu, and L. Li, "Research on graduates' employment quality: influence factor and path analysis," *China Higher Education Research*, no. 1, pp. 57–64, 2020.
- [10] G. Cheng, C. Rong, and L. Wang, "The effect of social capital in virtual community on customer citizenship behavior: from the perspective of psychological ownership," *Statistics & Information Forum*, vol. 35, no. 11, pp. 121–128, 2020.
- [11] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [12] X. Zhang and H. Qian, "Structural characteristics of athletes' employment network: a survey of 130 professional athletes in Shaanxi province," *Journal of Wuhan Institute of Physical Education*, vol. 52, no. 3, pp. 17–23, 2018.
- [13] D. Wang, "The relationship among cognitive appraisal, psychological control, social support and employment stress in university studies," *China Journal of Health Psychology*, vol. 12, pp. 1142–1145, 2007.
- [14] Y. Zou and F. Xiao, "Difference analysis of salary evaluation of athletes under different contingency factors," *Journal of Xi'an Physical Education University*, vol. 37, no. 6, pp. 712–720, 2020.
- [15] W. Du, J. Qiu, F. Zhang, J. Zhong, B. Zhang, and Y. Shen, "Development and validation of social support questionnaire for athletes," *Journal of Wuhan Institute of Physical Education*, vol. 54, no. 11, pp. 94–100, 2020.
- [16] L. K. M. Thomas and W. F. Russell, *Applications of Rasch Measurement in Education*, Nova Science Publishers, New York, NY, USA, 2010.
- [17] C. Erhel and M. Guergoatlariviere, "Trends in job quality during the great recession and the debt crisis(2007–2012): a comparative approach for the EU," *Psychopharmacology*, vol. 232, no. 19, pp. 3563–3572, 2015.
- [18] A. G. Bedeian, E. R. Kemery, and A. B. Pizzolatto, "Career commitment and expected utility of present job as predictors of turnover intentions and turnover behavior," *Journal of Vocational Behavior*, vol. 39, pp. 331–343, 1991.
- [19] J. Wang, "Investigation on retirement consciousness, mental and strategy of our athletes," *China Sport Science*, vol. 7, pp. 47–59, 2006.
- [20] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542.

Research Article

DWNet: Dual-Window Deep Neural Network for Time Series Prediction

Jin Fan, Yipan Huang , Ke Zhang, Sen Wang, Jinhua Chen, and Baiping Chen 

Hangzhou Dianzi University, Hangzhou, China

Correspondence should be addressed to Baiping Chen; chenbp@hdu.edu.cn

Received 22 July 2021; Accepted 25 September 2021; Published 13 October 2021

Academic Editor: Fei Xiong

Copyright © 2021 Jin Fan 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.

Multivariate time series prediction is a very important task, which plays a huge role in climate, economy, and other fields. We usually use an Attention-based Encoder-Decoder network to deal with multivariate time series prediction because the attention mechanism makes it easier for the model to focus on the really important attributes. However, the Encoder-Decoder network has the problem that the longer the length of the sequence is, the worse the prediction accuracy is, which means that the Encoder-Decoder network cannot process long series and therefore cannot obtain detailed historical information. In this paper, we propose a dual-window deep neural network (DWNet) to predict time series. The dual-window mechanism allows the model to mine multigranularity dependencies of time series, such as local information obtained from a short sequence and global information obtained from a long sequence. Our model outperforms nine baseline methods in four different datasets.

1. Introduction

In the age of big data, sequence data is everywhere in life [1, 2]. Time series prediction algorithms are becoming more and more important in many areas, such as financial market prediction [3], passenger demand forecasting [4], and heart signal prediction [5]. In most cases, time series data is multivariate. The key to multivariate time series prediction is to obtain the spatial and temporal relationships between different attributes at different times [6]. As a widely used traditional time series prediction algorithm, ARIMA [7] has shown its effectiveness in many areas. However, ARIMA cannot model nonlinear relationships and can only be applied to stationary time series [8–10]. Recurrent neural network (RNN) [11] has achieved great success in sequence modeling. But RNN has the problem of vanishing gradients, and it is difficult to capture the long-term dependence of time series [12]. Long Short-Term memory (LSTM) [13] and gated recurrent unit (GRU) [14, 15] alleviate the problem of RNN's vanishing gradients and have developed many models for time series prediction, such as Encoder-Decoder networks [15, 16]. Encoder-Decoder networks are excellent in time series prediction tasks, especially Attention-based

Encoder-Decoder networks [17]. Attention-based Encoder-Decoder network can not only find the spatial-temporal correlation between different series but also find important information in raw data and increase its weight [17]. Among them, dual-stage attention-based recurrent neural network (DARNN) is one of the state-of-the-art methods, creatively using a two-stage attention mechanism [18].

Although DARNN can capture spatial correlations between different attributes at the same time and the temporal correlations between different times in the same attribute, when the length of the sequence is too long, the prediction effect will be worse [18]. This problem is common to all Encoder-Decoder networks. A long sequence means more historical information, so better results should be obtained. However, due to the limitations of Encoder-Decoders, the information of the long sequence is not effectively used, even interfere with the prediction results. This is because LSTM does not solve the problem of vanishing gradient, and when the length of the time series is too long, the previous information will be covered by the latter. Therefore, Encoder-Decoders generally use a small window size to ensure the accuracy of prediction. Dual-stage two-phase attention-based recurrent neural network (DSTP) [19] has made

improvements to this problem of DARNN and optimized the prediction effect of long sequences. However, DSTP still does not make effective use of long sequences.

When the time window size is small, the series is very close to the prediction point. Such data has the closest relationship with the prediction point. For instance, if the values before the prediction point are gradually increasing, then the value at the prediction point is also likely to increase. When the time window size is large, series contain more time steps. It is difficult for other models to extract recent information, such as trends, in such a long series, so it cannot get good prediction results. However, more information brought by more time steps is very important for time series prediction. It is key to how to make good use of the different characteristics of short sequence and long sequence.

To solve this problem, we propose a dual-window deep neural network (DWNNet). DWNNet consists of two parts. The first part captures spatial and temporal correlations from the short sequence and is responsible for providing recent details, based on Encoder-Decoder [15]. The second part obtains long-term dependencies such as periodicity and seasonality from the long sequence, based on TCN. Temporal convolutional network (TCN) [20] is an emerging CNN-based model. With the parallelism of convolution operation and large receptive field, it has gained everyone's expectations in the areas of sequence modeling. Short-term time series generally contain only one or two periods. However, long-term time series are the opposite, including enough time steps. The setting of two different time window sizes for long sequence and short sequence makes it possible to mine multigranularity dependencies.

The main contributions of our work are as follows:

- (i) We propose a dual-window mechanism that can extract multigranularity information from sequences of different lengths.
- (ii) We propose the DWNNet approach, which includes the advantages of Encoder-Decoder networks and TCN at the same time. Encoder-Decoder networks have a strong ability to mine dependence from the short sequence. Meanwhile, TCN's large receptive field and fast training speed are more suitable for long sequences.
- (iii) DWNNet can be applied to time series prediction tasks in many domains, and there is no requirement for input data. To justify the effectiveness of the DWNNet, we compare it with nine baseline methods using the Human Sports dataset, SML 2010 dataset, Appliances Energy dataset, and EEG dataset. The experiment showed the effectiveness and robustness of DWNNet.

2. Related Work

For the time series prediction task, there are various approaches from traditional methods to deep learning methods. As the most famous traditional method, ARIMA can effectively obtain the long-term dependence of target

series [7]. However, ARIMA does not consider the spatial correlation between exogenous series [18], can only be used to deal with stationary series [7], and cannot model nonlinear relationships [8]. ARIMA is not suitable for the increasingly complex time series data analysis. As a deep neural network dedicated to machine learning and data mining applications [21–23], RNN can model nonlinear relationships [24] and has achieved great success in time series prediction. However, the gradient vanishing of RNN makes it difficult to obtain long-term dependence from time series. LSTM [13] and GRU [15] add a gating mechanism based on RNN and process the addition and deletion of timing information through the gating mechanism, which alleviates gradient vanishing of RNN. Based on LSTM and GRU, many influential deep neural networks have been proposed, such as the Encoder-Decoder network that has received great attention in the area of natural language processing [17]. Encoder-Decoder networks convert input series into context vector through Encoder and then convert context vector into output through Decoder. Encoder-Decoder networks have a problem. When the length of the sequence increases, the performance of Encoder-Decoder will first become better and then worse [17]. Attention-based Encoder-Decoder network can automatically select important information, thereby effectively alleviating the shortcoming of performance degradation when the length of the sequence increases.

Many attention-based models emerge endlessly. And DARNN [18], GeoMAN [25], and DSTP [19] are models that are improved based on the Attention-based Encoder-Decoder and used for time series prediction. Inspired by some theories of human attention [26], DARNN uses a dual-stage attention mechanism. The first stage uses a spatial attention mechanism to assign different weights to exogenous series to the hidden state of Encoder at the previous time step. The second stage uses a temporal attention mechanism to select the most relevant Decoder hidden states in all time steps. After DARNN was proposed, it has always been one of the state-of-the-art methods in time series prediction. Multilevel Attention Network (GeoMAN) is specially used to predict geo-sensor time series data. Many time series data are collected by sensors distributed in many locations. Such data is called geo-sensor time series data. If each series in the geo-sensor time series is simply treated as a normal attribute, it will lose the connection between different locations. GeoMAN adds local spatial attention and global attention mechanisms to Encoder and adds external factor information to Decoder to solve this problem. DSTP adds a new spatial attention mechanism to Encoder to obtain a spatial correlation between target series and exogenous series so that DSTP achieved better results in the long time series prediction.

While the Attention-based Encoder-Decoder network has attracted much attention, TCN has also shown strong sequence modeling ability [20]. TCN is based on CNN and includes causal convolution, dilated convolution [27, 28], and residual block [29]. To apply to series data, TCN is specially adjusted for different data formats of series and image. TCN has advantages that RNNs do not have. (1) TCN

can process series in parallel and does not need to be processed sequentially like RNN or LSTM. This means that there is no possibility that the information of the previous time step will be overwritten and it also means that there is a faster training speed. (2) TCN's receptive field varies with the number of layers, kernel size, and dilation rate and can be flexibly changed according to a different situation. (3) Compared with LSTM, TCN rarely has the problem of gradient vanishing. Due to the flexible receptive field, fewer parameters than LSTM, and parallel processing, TCN can not only reduce the training time of long sequence but also ensure that the information of the previous time step will not be covered. Therefore, TCN has a strong ability to obtain information from long sequences and is suitable for long sequence modeling.

Long- and short-term time series network (LSTNet) [30] is based on CNN and RNN and realizes that time series have two different dependencies, short-term and long-term. Therefore, LSTNet uses a recurrent-skip mechanism to obtain short-term dependence and then uses RNN to obtain long-term dependence from previous results. But it does not consider that the closer to the prediction point, the more important the information. Therefore, LSTNet will lose some recent information in the time series prediction.

3. Dual-Window Deep Neural Network

3.1. Notation and Problem Statement. In our work, there are two different window sizes, T_l and T_s . Given n exogenous series, that is, $\mathbf{X} = \mathbf{X}_1 = (\mathbf{x}_1^1, \mathbf{x}_1^2, \dots, \mathbf{x}_1^n)^\top = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T_l}) \in \mathbb{R}^{n \times T_l}$, we segmented a short series like this $\mathbf{X}_2 = (\mathbf{x}_s^1, \mathbf{x}_s^2, \dots, \mathbf{x}_s^n)^\top = (\mathbf{x}_{T_l-T_s+1}, \mathbf{x}_{T_l-T_s+2}, \dots, \mathbf{x}_{T_l}) \in \mathbb{R}^{n \times T_s}$. We use $\mathbf{x}_l^i = (x_{t_1}^i, x_{t_2}^i, \dots, x_{t_{T_l}}^i)^\top \in \mathbb{R}^{T_l}$ to represent the i -th long exogenous series, use $\mathbf{x}_s^i = (x_{t_1-T_s+1}^i, x_{t_1-T_s+2}^i, \dots, x_{t_1}^i)^\top \in \mathbb{R}^{T_s}$ to represent the i -th short exogenous series, and use $\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^n)^\top \in \mathbb{R}^n$ to denote a vector of n exogenous series at time t . We use $\mathbf{Y} = (y_1, y_2, \dots, y_{T_l})^\top \in \mathbb{R}^{T_l}$ to represent target series, which has the long window size T_l .

Given previous values of target series and exogenous series, that is, $(y_1, y_2, \dots, y_{T_l})$ with $y_t \in \mathbb{R}$ and $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T_l})$ with $\mathbf{x}_t \in \mathbb{R}^n$, we aim to predict the next time step value of target series y_{T_l+1} :

$$\hat{y}_{T_l+1} = F(y_1, \dots, y_{T_l}, \mathbf{x}_1, \dots, \mathbf{x}_{T_l}), \quad (1)$$

where $F(\cdot)$ is a nonlinear mapping function we aim to learn.

3.2. Model. Figure 1 presents the framework of our method. The input of DWNNet is divided into two parts, long series with window size T_l and short series with window size T_s . Short series is a part of long series and is located at the end of the long series (Figure 1 shows the relationship between the two series). Long series is processed by TCN [20] and used to obtain more detailed historical information than short series. The short series is processed by Encoder-Decoder to capture local information. Finally, the output of the two

parts is combined to get the predicted value of the target series at time T_{l+1} .

3.2.1. Capture Short-Term Dependence. First of all, we introduce the short series processing module. This part is based on Encoder-Decoder and uses spatial attention and temporal attention mechanism [18] to emphasize key information in short series. Encoder is based on LSTM, the input data of Encoder is short series $\mathbf{X}_2 = (\mathbf{x}_{T_l-T_s+1}, \mathbf{x}_{T_l-T_s+2}, \dots, \mathbf{x}_{T_l}) \in \mathbb{R}^{n \times T_s}$. Given i -th short exogenous series $\mathbf{x}_s^i = (x_{T_l-T_s+1}^i, x_{T_l-T_s+2}^i, \dots, x_{T_l}^i)^\top \in \mathbb{R}^{T_s}$, we use the spatial attention module to adaptively obtain the spatial correlation between exogenous series:

$$e_t^i = \mathbf{v}_e^\top \tanh(\mathbf{W}_e [\mathbf{h}_{t-1}; \mathbf{s}_{t-1}] + \mathbf{U}_e \mathbf{x}_s^i + \mathbf{b}_e), \quad (2)$$

$$\alpha_t^i = \frac{\exp(e_t^i)}{\sum_{j=1}^n \exp(e_t^j)}, \quad (3)$$

where $\mathbf{v}_e \in \mathbb{R}^{T_s}$, $\mathbf{W}_e \in \mathbb{R}^{T_s \times 2p}$, $\mathbf{U}_e \in \mathbb{R}^{T_s \times T_s}$, and $\mathbf{b}_e \in \mathbb{R}^{T_s}$ are parameters to learn. Here, p is the hidden size of Encoder and $\mathbf{h}_{t-1} \in \mathbb{R}^p$ and $\mathbf{s}_{t-1} \in \mathbb{R}^p$ are the hidden state and cell state of LSTM unit in the Encoder at time $t-1$. α_t^i is the attention weight measuring the importance of i -th exogenous series at time t . After we get the attention weight, we can adaptively extract exogenous series with

$$\tilde{\mathbf{x}}_t = (\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \dots, \alpha_t^n x_t^n)^\top. \quad (4)$$

Thus, the hidden state at time t can be updated as

$$\mathbf{h}_t = f_e(\mathbf{h}_{t-1}, \tilde{\mathbf{x}}_t), \quad (5)$$

where f_e is an LSTM unit in the Encoder. The spatial attention module calculates the weight of each exogenous series through equations (2) and (3) at time t and uses $\tilde{\mathbf{x}}_t$ to adjust the hidden state at time t .

The input of Decoder is the previous target series and the output of the Encoder, which is the hidden state of Encoder. Decoder aims to predict \hat{y}_{T_l+1} . To get accurate prediction results, we need to capture the temporal correlation between each series. So, we add a temporal attention module to the Decoder. The same as Encoder, the attention weight of Encoder hidden state at time t is calculated based upon the previous Decoder hidden state and cell state of LSTM unit with

$$d_t^i = \mathbf{v}_d^\top \tanh(\mathbf{W}_d [\mathbf{h}'_{t-1}; \mathbf{s}'_{t-1}] + \mathbf{U}_d \mathbf{h}_t + \mathbf{b}_d) \quad (6)$$

$$\beta_t^i = \frac{\exp(d_t^i)}{\sum_{j=1}^{T_s} \exp(d_t^j)}, \quad (7)$$

where $\mathbf{v}_d^\top \in \mathbb{R}^p$, $\mathbf{W}_d \in \mathbb{R}^{p \times 2q}$, $\mathbf{U}_d \in \mathbb{R}^{p \times p}$, and $\mathbf{b}_d \in \mathbb{R}^p$ are parameters to learn. q is the hidden size of Decoder, and $\mathbf{h}'_{t-1} \in \mathbb{R}^n$ and $\mathbf{s}'_{t-1} \in \mathbb{R}^n$ are the hidden state and cell state of LSTM unit in the Decoder at time $t-1$. β_t^i is the attention weight and can show the importance of i -th Decoder hidden state at time $t-1$. And, we can get context vector with

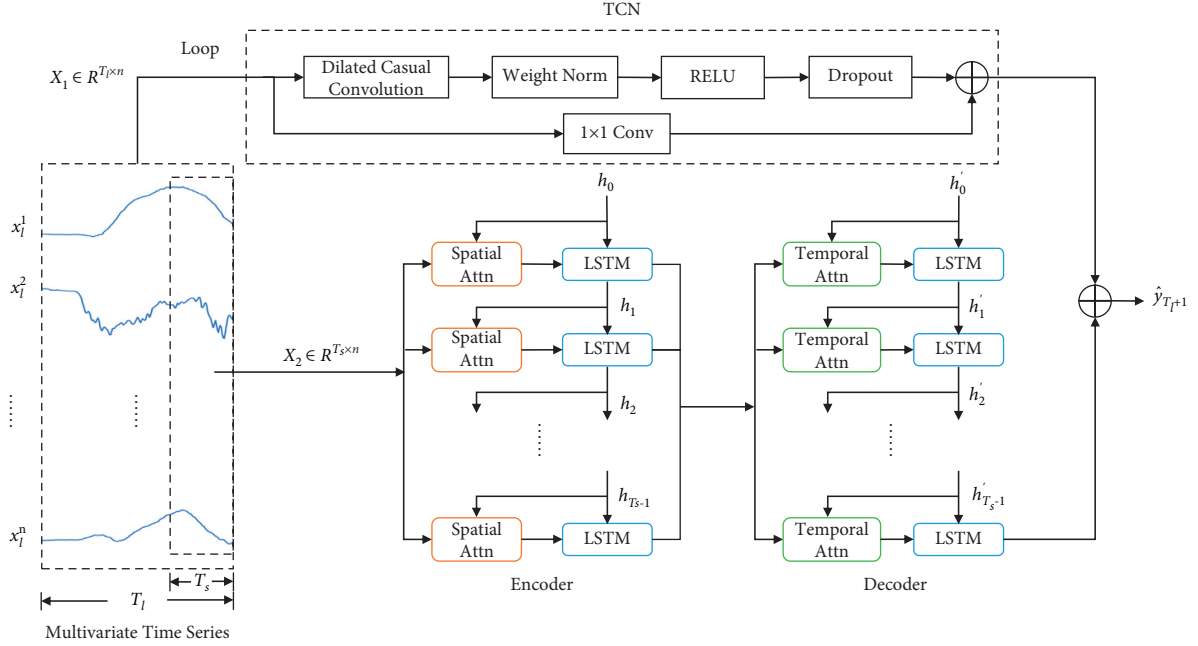


FIGURE 1: Framework of our method. T_l : the window size of long series. T_s : the window size of short series. n : the number of exogenous series. x_i^j : i -th long exogenous series. \mathbf{h}_t : hidden state in the encoder at time t . \mathbf{h}'_t : hidden state in the decoder at time t . Spatial Attn: spatial attention module. Temporal Attn: temporal attention module. \hat{y}_{T_l+1} : the predicting value at time $T_l + 1$.

$$\mathbf{c}_t = \sum_{i=1}^{T_s} \beta_i^t \mathbf{h}_i. \quad (8)$$

Context vector \mathbf{c}_t is the sum of all weighted encoder hidden states at time t . Then, we combine context vector \mathbf{c}_t and target series to update the Decoder hidden state \mathbf{h}'_t :

$$\mathbf{h}'_t = f_d(\mathbf{h}'_{t-1}, [\mathbf{c}_t; y_t]), \quad (9)$$

where f_d is an LSTM unit in the Decoder.

3.2.2. Capture Long-Term Dependence. We obtain long-term dependence through TCN [20], because TCN can process time series data in parallel and have much fewer parameters than LSTM. Therefore, TCN can quickly handle long series and improve time efficiency. And TCN does not have the problem of the previous information being covered. When window sizes are too large, the integrity of the information can be guaranteed. In our model, the input of the TCN part is long series from time 1 to T_l . In time series analysis, we cannot allow leakage from the future into the past. A high layer element at time t is obtained by convolution of elements from time t and earlier in the previous layer. To avoid information leakage, TCN uses casual convolution. To expand the receptive field, TCN uses dilated convolution [27, 28]. For long exogenous series $\mathbf{X}_1 = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{T_l}) \in \mathbb{R}^{n \times T_l}$ and filter \mathbf{g} : $(\mathbf{g}_0, \mathbf{g}_1, \dots, \mathbf{g}_{k-1})$, the element at time t is

$$O_t = (\mathbf{x} * d_g)(t) = \sum_{i=0}^{k-1} \mathbf{g}_i \mathbf{x}_{t-d,i}, \quad (10)$$

where d is the dilation factor, k is the filter size, and O is dilated convolution operation. d will increase exponentially with the number of layers to expand the receptive field. A deep neural network is so easy to have the problem of gradient exploding and gradient vanishing, so TCN uses residual block [29]. The residual connection enables the network to transfer information in a cross layer and improve the efficiency of feature extraction.

3.2.3. Training. Figure 1 shows that the predicted value is determined by two parts. We combine the output of Decoder \mathbf{h}'_{T_s} and TCN O_{T_l} to predict \hat{y}_{T_l+1} :

$$\begin{aligned} \hat{y}_{T_l+1} &= F(y_1, \dots, y_{T_l}, \mathbf{x}_1, \dots, \mathbf{x}_{T_l}) \\ &= \mathbf{v}_y^\top (\mathbf{W}_y [\mathbf{h}'_{T_s}; O_{T_l}] + \mathbf{b}_w) + b_v, \end{aligned} \quad (11)$$

where $\mathbf{v}_y \in \mathbb{R}^q$, $\mathbf{W}_y \in \mathbb{R}^{q \times (q+m)}$, $\mathbf{b}_w \in \mathbb{R}^q$, and $b_v \in \mathbb{R}$ are parameters to learn. Here, m is the number of hidden units per layer, and $[\mathbf{h}'_{T_s}; O_{T_l}] \in \mathbb{R}^{q+m}$. We use the back-propagation algorithm to train DWNet. We use the Adam optimizer [31] to minimize the mean squared error (MSE) between the predicted value \hat{y}_{T_l+1} and the ground truth y_{T_l+1} :

$$L(\theta) = \|\hat{y}_{T_l+1} - y_{T_l+1}\|_2^2, \quad (12)$$

where θ are all parameters to learn in DWNet.

4. Experiment

Our model and all baseline methods are implemented on the PyTorch framework [32]. In this section, we first introduce four different datasets used in the experiment. Then, we

introduce nine baseline methods. Next, we introduce the model evaluation methods and parameters. Finally, experiment results show the effectiveness of DWNet.

4.1. Datasets. We use four datasets to verify the effect of our model. They are in the field of sports, energy, climate, and medicine. We divide datasets into training sets and testing sets according to the ratio of 4:1.

4.1.1. Human Sports [33]. Human Sports data is collected by 10 volunteers of different genders, heights, and weights who performed sports including squat, walking, jumping jacks, and high knee. Four sensors worn on the arms and thighs record data every 50 milliseconds, including acceleration and angular acceleration of the x -axis, y -axis, and z -axis. In our experiment, we take the resultant acceleration as the target series and others as exogenous series. We only use the squat data of one volunteer and use the first 8796 data points as the training set and the remaining 2197 data points as the testing set.

4.1.2. SML 2010 [34]. SML 2010 is a public dataset for indoor temperature prediction. SML 2010 contains nearly 40 days of data, which is collected by the monitoring system. The data were sampled every minute, computing and uploading it smoothed with 15-minute means. In our experiment, we take the weather temperature as target series and select fifteen exogenous series. We use the first 1971 data points as the training set and the remaining 493 data points as testing set.

4.1.3. Appliances Energy [35]. Appliances energy is a public dataset used for home appliance energy consumption prediction. This dataset is at 10 minutes for about 4.5 months. Room temperature and humidity conditions were monitored with a wireless sensor network. The energy data is recorded with m-bus energy meters every 10 minutes. Weather data was downloaded from the nearest airport weather station. In our experiment, we take energy use as target series and others as exogenous series. We use the first 15548 data points as a training set and the remaining 3887 as a testing set.

4.1.4. EEG [36]. EEG is a public dataset for classification and regression. This database consists of 30 subjects performing Brain-Computer Interface for Steady-State Visual Evoked Potentials. In our experiment, we only use the data from one of those subjects. We take the electrode O1 attribute as the target series and others as exogenous series. We use the first 7542 data points as a training set and the remaining 1886 as a testing set.

4.2. Baseline

4.2.1. ARIMA [8]. It is one of the well-known statistical algorithms for time series prediction.

4.2.2. LSTM [13]. LSTM is improved by RNN, through the gating mechanism to control the adding and deletion of information, alleviating the gradient vanishing.

4.2.3. Encoder-Decoder [16]. It is widely used in machine translation. However, Encoder-Decoder has the disadvantage of losing information.

4.2.4. Input-Attn-RNN [18]. It adds a spatial attention module on the basis of Encoder-Decoder to the Encoder to obtain the spatial correlation of raw data.

4.2.5. Temp-Attn-RNN [19]. It adds a temporal attention module on the basis of Encoder-Decoder to the Decoder to obtain the temporal correlation of Encoder hidden state.

4.2.6. TCN [20]. It is an emerging sequence modeling model that has attracted much attention, including casual convolution, dilated convolution, and residual blocks.

4.2.7. LSTNet [30]. It combines CNN and RNN to obtain short-term and long-term dependencies in sequence.

4.2.8. DARNN [18]. As one of the state-of-the-art methods, inspired by the human attention system, DARNN uses both spatial attention and temporal attention to extract spatial-temporal correlation.

4.2.9. DSTP-RNN [19]. It improves DARNN and adds an attention module to Encoder. In the Encoder, more stationary weights can be obtained. DSTP-RNN is good at long time series prediction.

4.3. Evaluation Metrics. We employ root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (SMAPE) to evaluate our model and baseline methods. These four evaluation metrics are scale-independent and widely used in time series prediction. RMSE has a strong feedback ability for predicted results that deviate too much from the ground truth. MAE treats all results equally. MAPE is able to compare forecast accuracy among differently scaled time series data because relative errors do not depend on the scale of the dependent variable. However, when truth value y_t is small, different \hat{y}_t will have a huge difference in MAPE value. And SMAPE can solve this problem. Assuming \hat{y}_t is predicted value at time t and y_t is the ground truth, RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{y}_t - y_t)^2}. \quad (13)$$

MAE is defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^N |\hat{y}_t - y_t|. \quad (14)$$

MAPE is defined as follows:

$$\text{MAPE} = \frac{100\%}{N} \sum_{t=1}^N \left| \frac{\hat{y}_t - y_t}{y_t} \right|. \quad (15)$$

SMAPE is defined as follows:

$$\text{SMAPE} = \frac{100\%}{N} \sum_{t=1}^N \frac{|\hat{y}_t - y_t|}{(|\hat{y}_t| + |y_t|)/2}. \quad (16)$$

4.4. Parameters Settings. Most time series prediction models have chosen a small window size in their experiment. For example, DARNN set the window size to 10 [18], and GeoMAN set the window size to 6 [25]. To show the influence of window size on prediction, we select the window size $T = \{2, 4, 8, 16, 32, 128\}$. For DWNet, we set $T_l = 128$ and $T_s = 16$. For baseline methods, we conducted experiments on $T = 16$ and $T = 128$, respectively. In training, we set the batch size to 128 and learning rate to 0.001. In our model, there are also parameters such as the hidden size of Encoder p , the hidden size of Decoder q , kernel size, and levels of TCN. For simplicity, we use the same hidden size at Encoder and Decoder, that is, $p = q$, and conducted a grid search over $\{16, 32, 64, 128, 256\}$. For TCN level and kernel size, we also conducted a grid search. The setting in which $p = q = 128$, level = 8, kernel size = 7 outperforms the others in the testing set. And we fixed these parameters in all experiments.

5. Results and Discussion

In this section, we first compare our model with baseline methods on the four datasets. Then, we conduct a grid search to show the performance of our model in different long time steps and short time steps combinations. Next, we investigate ablation experiments and study the time efficiency of our model.

5.1. Model Comparison. To show the effectiveness of DWNet, we compare DWNet with 9 different methods, including the state-of-the-art methods and emerging methods. For the sake of fairness, we use two different window sizes for baseline methods so that we can compare the baselines' results of long window size and short window size with DWNet. The prediction results of DWNet and baseline methods are shown in Tables 1 and 2.

Table 1 shows that DWNet achieves the best RMSE and MAE across four datasets. Table 2 shows that DWNet also achieves the best MAPE and SMAPE in four datasets. This is because DWNet obtains not only the short-term dependency in the short sequence but also the long-term dependency in the long sequence. ARIMA performs worse than other models for ARIMA cannot capture linear relationships and does not consider the

spatial correlation between exogenous series [7]. Encoder-Decoder network performs better than normal LSTM in four datasets, which means Encoder-Decoder is easier to obtain dependency from raw data [16]. Attention-based Encoder-Decoder networks, that is, Input-Attn-RNN and Temp-Attn-RNN, are better than normal Encoder-Decoder networks in four datasets because the attention mechanism pays more attention to more important features in raw data. DARNN and DSTP combine spatial attention and temporal attention mechanism and have good performance in four datasets. The performance of TCN is very unstable, and its performance in Human Sports is better than DSTP, but it is far worse than DARNN and DSTP in other datasets, especially EEG. LSTMNet's performance is also unstable. And it performs very well in Human Sports, but it performs poorly in the other three datasets. Meanwhile, we can also find that LSTM-based networks perform better than long sequences in short sequences.

5.2. Time Step Study. In this section, we study the impact of long window size T_l and short window size T_s on prediction. When we vary T_l and T_s , we keep other parameters fixed. We plot the RMSE versus different long window size ($T_l \in \{64, 128, 256, 512\}$) and short window size ($T_s \in \{4, 8, 16, 32\}$) in Figure 2.

It is easily observed that the performance of DWNet is simultaneously affected by two parameters T_l and T_s . When T_l is fixed, the performance of DWNet will be worse when T_s is too large or too small and vice versa. And we notice that DWNet achieves the best performance when $T_l = 128$ and $T_s = 16$.

5.3. Ablation Experiment. To further investigate the effectiveness of each model component, we compare DWNet with Input-Attn-RNN, Temp-Attn-RNN, DARNN, and other variants in Human Sports and EEG datasets. In this experiment, we set window size T of Input-Attn-RNN, Temp-Attn-RNN, and DARNN to 16 and set $T_l = 128$ and $T_s = 16$. The variants of DWNet are as follows:

- (i) DWNet-ni: there is no spatial attention module in the Encoder part.
- (ii) DWNet-nt: there is no temporal attention module in the Decoder part.

The experiment results are shown in Figure 3. Input-Attn-RNN performs better than Temp-Attn-RNN in the EEG dataset but performs worse than Temp-Attn-RNN in the Human Sports dataset. However, DARNN achieves better RMSE and MAE than Input-Attn-RNN and Temp-Attn-RNN in both two datasets. Apparently, the model based on a two-stage attention mechanism is better than the single attention model. And that is why DWNet is superior to DWNet-ni and DWNet-nt. It is easily observed that DWNet achieves the best RMSE in Human Sports and EEG, which shows that the information in the long sequence is valuable for the time prediction task. Without the long

TABLE 1: RMSE and MAE performance comparison among different methods and datasets (best result is displayed in **boldface**).

Models	SML 2010		Human Sports		EEG		Energy	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
ARIMA (16)	0.2786	0.2219	0.1371	0.0617	0.5694	0.4724	0.8640	0.5740
LSTM (16)	0.1905	0.1489	0.0831	0.0325	0.2244	0.1724	0.6907	0.3663
LSTM (128)	0.2099	0.1671	0.0983	0.0437	0.3033	0.2283	0.8017	0.4376
Encoder-Decoder (16)	0.1438	0.0907	0.0774	0.0296	0.2499	0.1401	0.5983	0.2839
Encoder-Decoder (128)	0.1648	0.1012	0.0831	0.0303	0.4650	0.3036	0.6524	0.3117
Input-Attn-RNN (16)	0.1296	0.0762	0.0680	0.0282	0.2055	0.1447	0.5452	0.2564
Input-Attn-RNN (128)	0.1008	0.0897	0.0766	0.0362	0.4217	0.2881	0.5782	0.2502
Temp-Attn-RNN (16)	0.1097	0.0692	0.0646	0.0311	0.2220	0.1500	0.5414	0.2507
Temp-Attn-RNN (128)	0.1105	0.0770	0.0740	0.0334	0.3943	0.2998	0.5488	0.2563
TCN (16)	0.1156	0.0817	0.0628	0.0270	1.1845	0.9545	0.8279	0.5186
TCN (128)	0.1473	0.1136	0.0727	0.0329	1.1050	0.8696	0.8126	0.4567
LSTNet (16)	0.1277	0.0957	0.0582	0.0269	0.2322	0.1807	0.5733	0.2762
LSTNet (128)	0.1352	0.1020	0.0642	0.0312	0.2384	0.1868	0.6078	0.3296
DARNN (16)	0.0977	0.0644	0.0643	0.0232	0.1804	0.1442	0.5270	0.2439
DARNN (128)	0.1093	0.0778	0.0733	0.0435	0.3483	0.3250	0.5556	0.2525
DSTP (16)	0.0932	0.0614	0.0641	0.0227	0.1805	0.1414	0.5320	0.2459
DSTP (128)	0.0954	0.0670	0.0641	0.0235	0.1754	0.1384	0.5456	0.2525
DWNet	0.0891	0.0565	0.0575	0.0217	0.1702	0.1371	0.5015	0.2362

The window size of baseline methods is set to 16 and 128, and the short window size and long window size of DWNet are set to 16 and 128, respectively.

TABLE 2: MAPE and SMAPE performance comparison among different methods and datasets (best result is displayed in **boldface**).

Models	SML 2010		Human Sports		EEG		Energy	
	MAPE (%)	SMAPE (%)	MAPE (%)	SMAPE (%)	MAPE (%)	SMAPE (%)	MAPE (%)	SMAPE (%)
ARIMA (16)	123.0993	62.0098	22.3507	18.4392	159.6348	83.7905	178.4365	77.4287
LSTM (16)	78.9095	45.0082	17.4439	13.9803	80.3427	56.9819	163.9849	65.8066
LSTM (128)	83.1021	49.5439	17.9035	12.0033	87.5609	63.5271	176.3415	69.5442
Encoder-Decoder (16)	70.7142	43.0760	13.3326	9.5610	66.4635	38.5606	170.0863	69.3110
Encoder-Decoder (128)	78.9981	50.5022	15.7987	10.0923	79.3445	40.2327	181.7583	76.1764
Input-Attn-RNN (16)	61.1121	30.0997	11.7831	7.9462	41.8856	32.4032	145.8688	67.5386
Input-Attn-RNN (128)	68.9089	35.3459	12.0034	7.9897	41.4628	29.7608	152.3287	64.7685
Temp-Attn-RNN (16)	54.3435	32.8703	11.2627	7.2110	40.8683	29.0871	140.6838	56.0774
Temp-Attn-RNN (128)	57.8065	31.9911	11.4980	7.1153	45.8705	30.0085	128.0644	59.3527
TCN (16)	83.2350	49.5797	18.5920	11.0097	675.9030	131.4202	258.1707	93.9882
TCN (128)	85.4479	67.3689	14.6141	10.7326	995.0083	133.4580	265.3023	100.9782
LSTNet (16)	50.5956	29.6186	13.0975	8.6524	46.9208	34.3753	135.8396	68.8810
LSTNet (128)	83.2999	46.3060	13.3192	9.3698	50.0573	41.5482	140.2021	72.9974
DARNN (16)	43.1275	28.9558	11.9568	8.0686	36.4658	26.6514	123.0556	59.8798
DARNN (128)	45.2110	33.1612	11.8951	7.4177	33.8550	27.1255	139.0686	64.3326
DSTP (16)	40.6946	24.5261	11.7359	7.1343	34.5063	22.7594	138.8959	59.8884
DSTP (128)	36.2600	24.8048	11.7928	7.3406	35.1179	24.0093	142.8744	56.9903
DWNet	31.5764	23.0888	10.3833	7.0483	31.3287	20.6706	82.1119	52.5880

The window size of baseline methods is set to 16 and 128, and the short window size and long window size of DWNet are set to 16 and 128, respectively.

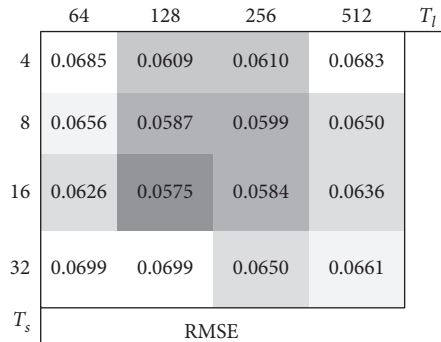


FIGURE 2: Performance of DWNet in Human Sports based on different short window size T_s and long window size T_l . We use different colors to indicate the prediction effect. The better the prediction, the darker the color.

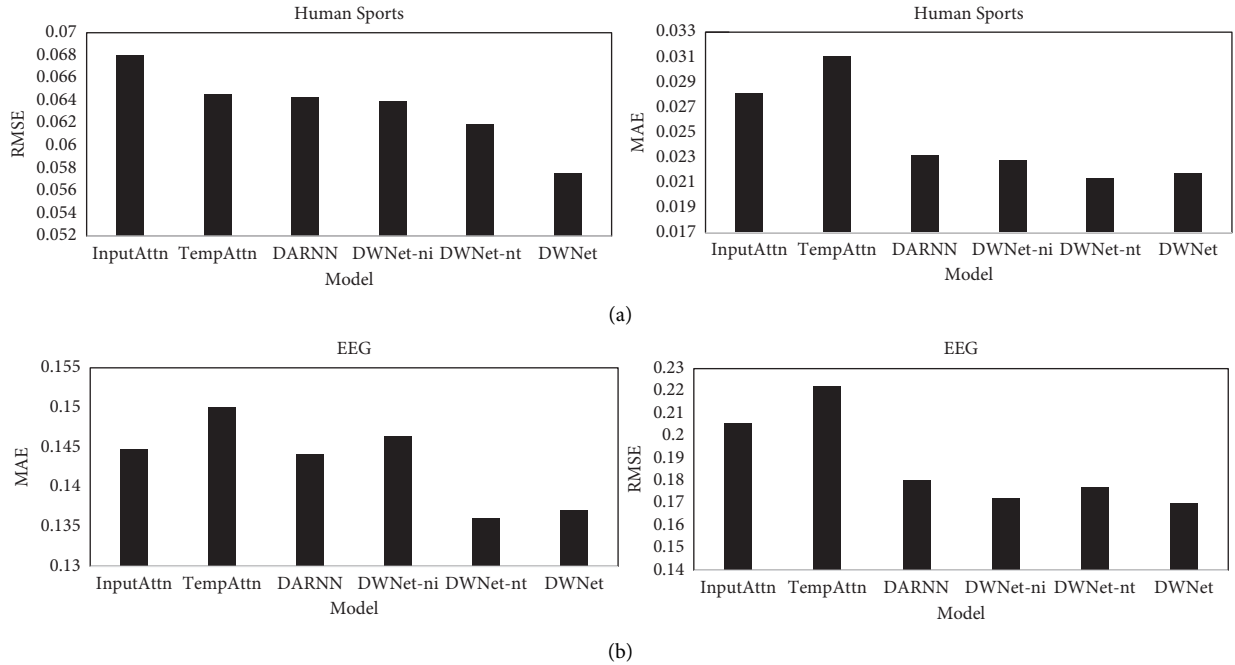


FIGURE 3: Performance of different methods in Human Sports and EEG. (a) RMSE and MAE versus different methods over Human Sports. (b) RMSE and MAE versus different methods over EEG.

sequence processing module, it is impossible to outperform the state-of-the-art methods in time series prediction.

5.4. Time Complexity. The time efficiency of deep neural networks is also an evaluation metric that needs to be considered. In this section, we compare the time efficiency of DWNet and baseline methods. In this experiment, we set T to 16, T_l to 128, T_s to 16, and fixed other parameters. We experimented on Human Sports and EEG datasets and recorded the time (in seconds) spent in 10 epochs. The results are shown in Figure 4. We can observe that, with more attention modules, the time spent by the model gradually increases. Input-Attn-RNN and Temp-Attn-RNN have only one attention module: one is spatial attention and the other is temporal attention, but the amount of computation is essentially the same. Temp-Attn-RNN’s training time is slightly longer than Input-Attn-RNN, but it is far less than the DARNN that both attention modules have. DSTP has two attention modules in the Encoder part and one attention module in the Decoder part, so the training time spent is longer than DARNN. TCN is superior to fewer parameters and the characteristics of parallel processing and has a very large advantage in time spent. It takes the least time in both two datasets. In DWNet, there are two attention modules and a long sequence processing module (implemented by TCN). Therefore, DWNet is inferior to DARNN in terms of time efficiency and even worse than TCN. However, DWNet has stronger time series forecasting capabilities than DARNN and TCN and is more suitable for situations that require high accuracy rather than low time consumption.

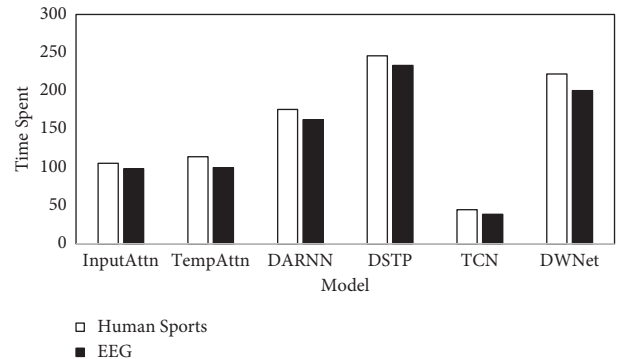


FIGURE 4: Time spent per 10 epoch in different models and datasets.

6. Conclusion

In this paper, we propose a dual-window deep neural network (DWNet) to make good use of the long sequence for time series prediction. The dual-window mechanism splits the end of a sequence as a short sequence and treats this sequence as a long sequence. The long sequence processing module in DWNet can extract historical information from long time series, and the short sequence processing module obtains recent information from short time series. These allow the model to learn both long-term dependence and short-term of the sequence. Our model outperforms the state-of-the-art methods in four datasets. In the future, we are going to perform model compression and reduce the model running time. Moreover, we will improve the long sequence processing module and enhance its stability, thereby enhancing the performance of DWNet.

Data Availability

The Human Sports dataset is available from Hangzhou Dianzi University's fitness club. Due to personal privacy, data cannot be made publicly available. The remaining datasets analyzed during the current study were derived from the following public domain resources: <https://archive.ics.uci.edu/ml/datasets/SML2010> <https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction> <https://archive.ics.uci.edu/ml/datasets/EEG+Steady-State+Visual+Evoked+Potential+Signals>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (no. U1609211) and National Key Research and Development Project (2019YFB1705102).

References

- [1] Q. Zhang, J. Wu, H. Yang, Y. Tian, and C. Zhang, "Unsupervised feature learning from time series," in *Proceedings of the IJCAI*, pp. 2322–2328, New York, NY, USA, July 2016.
- [2] H. Wang, Q. Zhang, J. Wu, S. Pan, and Y. Chen, "Time series feature learning with labeled and unlabeled data," *Pattern Recognition*, vol. 89, pp. 55–66, 2019.
- [3] B. Moews, J. M. Herrmann, and G. Ibikunle, "Lagged correlation-based deep learning for directional trend change prediction in financial time series," *Expert Systems with Applications*, vol. 120, pp. 197–206, 2019.
- [4] L. Bai, L. Yao, S. Kanhere, X. Wang, and Q. Z. Sheng, "Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting," 2019, <https://arxiv.org/pdf/2108.05940.pdf>.
- [5] S. Fraga, M. A. Aceves-Fernandez, J. C. Pedraza-Ortega, and J. M. Ramos-Arreguin, "Screen task experiments for eeg signals based on ssvep brain computer interface," *International Journal of Advanced Research*, vol. 6, no. 2, pp. 1718–1732, 2018.
- [6] Q. Zhang, J. Wu, P. Zhang, G. Long, and C. Zhang, "Salient subsequence learning for time series clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2193–2207, 2018.
- [7] D. Asteriou and S. G. Hall, "Arma models and the box-jenkins methodology," *Applied Econometrics*, vol. 2, no. 2, pp. 265–286, 2011.
- [8] A. Geetha and G. M. Nasira, "Time-series modelling and forecasting: modelling of rainfall prediction using arima model," *International Journal of Society Systems Science*, vol. 8, no. 4, pp. 361–372, 2016.
- [9] L. Yan, A. Elgamal, and G. W. Cottrell, "Substructure vibration narx neural network approach for statistical damage inference," *Journal of Engineering Mechanics*, vol. 139, no. 6, pp. 737–747, 2013.
- [10] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time Series: Theory and Methods: Theory and Methods*, Springer Science & Business Media, Berlin, Germany, 1991.
- [11] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Machine Learning*, vol. 7, no. 2-3, pp. 195–225, 1991.
- [12] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [14] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, <https://arxiv.org/abs/1412.3555>.
- [15] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014, <https://arxiv.org/abs/1406.1078>.
- [16] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: encoder-decoder approaches," 2014, <https://arxiv.org/abs/1409.1259>.
- [17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, <https://arxiv.org/abs/1409.0473>.
- [18] Q. Yao, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," 2017, <https://arxiv.org/abs/1704.02971>.
- [19] Y. Liu, C. Gong, L. Yang, and Y. Chen, "Dstp-rnn: a dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction," *Expert Systems with Applications*, vol. 143, p. 113082, 2020.
- [20] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, <https://arxiv.org/abs/1803.01271>.
- [21] X. Su, S. Xue, F. Liu et al., "A comprehensive survey on community detection with deep learning," 2021, <https://arxiv.org/abs/2105.12584>.
- [22] F. Liu, S. Xue, J. Wu et al., "Deep learning for community detection: progress, challenges and opportunities," 2020, <https://arxiv.org/abs/2005.08225>.
- [23] X. Ma, J. Wu, S. Xue et al., "A comprehensive survey on graph anomaly detection with deep learning," 2021, <https://arxiv.org/abs/2106.07178>.
- [24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [25] Y. Liang, K. Songyu, J. Zhang, X. Yi, and Y. Zheng, "Geoman: multi-level attention networks for geo-sensory time series prediction," in *Proceedings of the IJCAI*, pp. 3428–3434, Stockholm, Sweden, July 2018.
- [26] R. Hübner, M. Steinhauser, and C. Lehle, "A dual-stage two-phase model of selective attention," *Psychological Review*, vol. 117, no. 3, pp. 759–784, 2010.
- [27] A. Van Den Oord, D. Sander, H. Zen et al., "Wavenet: a generative model for raw audio," 2016, <https://arxiv.org/abs/1609.03499>.
- [28] Y. Fisher and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, <https://arxiv.org/abs/1511.07122>.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [30] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," in *Proceedings of the 41st International ACM SIGIR*

- Conference on Research & Development in Information Retrieval*, pp. 95–104, Ann Arbor, MI, USA, July 2018.
- [31] D. P. Kingma and B. Jimmy, “Adam: a method for stochastic optimization,” 2014, <https://arxiv.org/abs/1412.6980>.
- [32] P. Adam, S. Gross, S. Chintala et al., Automatic differentiation in pytorch, 2017.
- [33] J. Fan, H. Wang, Y. Huang, K. Zhang, and B. Zhao, “Aedmts: an attention-based encoder-decoder framework for multi-sensory time series analytic,” *IEEE Access*, vol. 8, pp. 37406–37415, 2020.
- [34] F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora, and J. Pardo, “On-line learning of indoor temperature forecasting models towards energy efficiency,” *Energy and Buildings*, vol. 83, pp. 162–172, 2014.
- [35] L. M. Candanedo, V. Feldheim, and D. Deramaix, “Data driven prediction models of energy use of appliances in a low-energy house,” *Energy and Buildings*, vol. 140, pp. 81–97, 2017.
- [36] S. M. Fernandez-Fraga, M. A. Aceves-Fernandez, J. C. Pedraza-Ortega, and S. Tovar-Arriaga, “Feature extraction of eeg signal upon bci systems based on steady-state visual evoked potentials using the ant colony optimization algorithm,” *Discrete Dynamics in Nature and Society*, vol. 2018, Article ID 2143873, 2018.

Research Article

Analysis on Quantified Self-Behavior of Customers in Food Consumption under the Perspective of Social Networks

Lei Lei ^{1,2}, Yaling Zhu,³ and Qiang Liu ⁴

¹School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710064, China

²Sports Department, Northwest A&F University, Yangling 712100, China

³International Business School, Shaanxi Normal University, Xi'an 710119, China

⁴Sports Department, Central South University, Changsha 410083, China

Correspondence should be addressed to Qiang Liu; answerxiaoliu@163.com

Received 19 May 2021; Revised 23 August 2021; Accepted 9 September 2021; Published 6 October 2021

Academic Editor: Fei Xiong

Copyright © 2021 Lei Lei 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.

China is still facing the double challenges of over nutrition and malnutrition. One of the main reasons is the lack of residents' understanding of the nutritional value of food. Quantified self, as a measure of consumer self-activity, has been used to analyze food consumption behavior recently. Although the research results are increasing, the conclusions are not consistent. What's more, previous literatures did not consider food consumption behavior based on the theory of information perception and the risk perception theory. In addition to obtaining information through their own human capital for quantitative activities, consumers will also obtain information through social networks. In view of the above understanding, this study uses experimental design and field survey to obtain data, uses Heckman two-step method and PLS path modeling method to analyze the impact of consumers' quantified self-behavior on their health food consumption, and discusses the moderating role of social networks based on the perspective of complex network. The results show that (1) consumers' health awareness can promote their choice of quantified self-behavior, (2) consumers' quantified self-behavior is helpful to promote their purchase intention and purchase scale of healthy food, and (3) social networks play a positive moderating role in consumers' quantified self-influence on their healthy food consumption. Both emotional networks and instrumental networks have significant moderating effect, but the formal is stronger. This article not only considers the relationship between food consumption behavior and social network but also the enhances literature based on the theory of information perception and the risk perception theory.

1. Introduction

A state relies on people and people relies on food. With the improvement of living standards and the growth of residents' income, the total dietary intake of residents in China has shown a gradual improvement trend in the past decade. From the data of Research Group of China's Health and Nutrition (2019), the malnutrition rate of residents aged 18 years and above as well as that of children younger than five years have decreased from 8% to 22.16% in 2000 to 2.2% and 9.3% in 2019, respectively. However, at the same time, the dietary intake structure of residents is unbalanced, such as excessive consumption of livestock meat and fat, low consumption of cereal food, and general lack of vitamins [1].

One of the main reasons for the coexistence of malnutrition and over nutrition is that the residents are lacking the understanding of the nutritional value of food and the rational judgment of nutrition. During the epidemic of COVID-19 in 2020, residents' consumption behavior also reflected the same problem. On one hand, the residents bought the food not according to their needs but to panic buying and hoarding. On the other hand, the residents concentrated on purchasing food like rice, flour, oil, meat, and so on, although the willingness to buy dairy products, coarse grains, and frozen vegetables was not high [2]. Therefore, the key problem to be solved in practice is to make the residents realize the nutritional value of food, to form the concept of healthy nutrition, to optimize the food

consumption structure, and improve the status of malnutrition and over nutrition of Chinese residents.

Quantified self as a consumer tracking measure of self-activity [3, 4] and the process of forming self-knowledge and conventional habits based on it, brings about the change in consumers' behavior. Recently, quantified self is also used in the research of analyzing food consumption behavior of consumers [4]. Although the results of quantified self-analysis in residents' consumption behavior are increasing day by day, the conclusion shows a dilemma of quantified self in consumption, which was like chicken ribs. According to the theory of information perception, quantification is the way to transform the professional and theoretical food safety assessment information into the data information that consumers can understand [5]. Consumers choose the appropriate food according to the food composition, safety level, and other information. This kind of quantified self-behavior is a short-term choice behavior made by consumers based on the food information provided by producers; it has little effect on consumers' food nutrition concept and health behavior [6, 7]. According to the risk perception theory, consumers will improve the collection, management, and reflection of quantitative information of self-health and demand status (such as blood pressure and blood glucose level) under risk perception and then accurately intervene and control self-behavior decision-making [4]. This quantified self-behavior is to select appropriate food according to consumers' health status; it has a positive role in promoting the formation of nutrition concept and health behavior of consumers [8, 9]. In addition, many scholars also point out that consumers' quantified self-behavior has the problem of short-term participation. Different conclusions make it necessary to further study the quantified self in food consumption: is the quantified self in food consumption effective for residents' health? Is quantified self-behavior a "chicken rib" to the formation of residents' nutrition concept? These problems need further analysis and verification in theory.

In the implementation process of quantified self-behavior, consumers obtain information not only through their own human capital but also through social networks for quantitative activities. If consumers want to eat safe and nutritious food, firstly, they need to obtain information about food quality, safety, and ingredients. Due to the limitation of cost and the profit-making purpose of food manufacturers, food manufacturers may adopt opportunistic behavior in the disclosure of information on food quality, safety, and nutrition [10]. At this time, consumers can only obtain the related information through their own efforts. To obtain more information about food quality, consumers will pay a high cost for searching information if they only make efforts to obtain it by themselves. At this time, social networks provide opportunities for consumers. The so-called social network refers to "a relatively stable association system formed between social individuals because of interaction." In this system, network members can exchange, interact, and share information. Through social networks, those consumers who have less information can establish effective contact with those who have more

information. Consumers who have less information will get the information they need from consumers who have more information, which makes food information shared between two kinds of consumers and reduces a lot of information searching costs [11]. Meanwhile, the mutual sharing of network members could help to promote consumers' quantitative behavior, so as to improve consumers' concept for food safety and nutrition.

Based on the complex social network theory, this study analyzes the impact of consumers' quantified self-behavior on their healthy food consumption and explores the role of social networks. The main contents include the following: firstly, whether or not to quantify self in residents' daily food consumption; secondly, what is the impact of quantified self-behavior on residents' health food consumption intention and scale; thirdly, whether social network has a moderating effect in the process of quantified self-behavior on residents' health food consumption. The article is divided into following parts: Section 2 is theoretical analysis and hypothesis, which theoretically analyses the relationship between quantified self and food consumption based on social network. Section 3 is materials and methods, which introduces the data sources and main methods. Section 4 is results and discussion, which shows the main empirical results and discusses them. The final part is conclusion, which summarizes the full and puts forward the shortcomings of the article.

2. Theoretical Analysis and Hypothesis

2.1. Quantified Self-Behavior of Healthy Food Consumers. With the improvement of people's living standards, food consumption of residents has gradually changed from "full" to "healthy." In recent years, food poisoning, overdue resale, and other problems have further strengthened consumers' preference for healthy food. Compared with consumers who buy ordinary food, those who prefer healthy food are more likely to pay attention to the origin, composition, and nutritional structure of food, and these information often need to be screened and compared. Relevant theories point out that in order to accurately predict the extent of food safety risks to their own health, consumers should understand the sources of risk variability and risk magnitude, acquire self-knowledge through quantitative information, and scientifically assess food safety risks through the levels of harmful and nutritional components. It is necessary and important to develop a suitable diet to ensure the safety and quality of individualized food [12]. Therefore, we propose Hypothesis 1.

Hypothesis 1. Compared with other consumers, consumers with healthy food attitude are more willing to take quantified self-behavior.

2.2. Influence of Quantified Self-Behavior on Healthy Food Consumption. The application of quantified self in food consumption is mainly reflected in two aspects: first, the residents' attention to food quantitative data information will improve their sensitivity and alertness to the

information of food harmful ingredients and their harmful amount, so as to enhance the accurate identification of food safety, risk, and efficiency [8, 12]. The preference for quantitative information will arouse consumers' quantitative consumption intention, that is, through tracking and observing self-related data (such as their maximum tolerance of specific food additives and daily consumption) and product efficiency data (such as the content of specific food additives and the magnitude of possible harm to the body), the self-behavior state and product efficiency knowledge will be formed. Based on this, the consumption decision is optimized [13].

Secondly, under the health risk perception, residents will improve the collection, management, and reflection of quantitative information of self-health and demand status (such as blood pressure and blood glucose level), so as to accurately intervene and control self-behavior decision-making [4]. Under the residents' own health perception, the residents' food consumption decisions will show a precise preference, which not only pay more attention to the quantitative data information of food but also purchase food according to the quantitative information of their own needs and make food decisions accurately and rationally based on objective quantitative data rather than subjective assumptions [14]. Quantitative information will enable residents to rethink their eating habits in a way driven by data rather than experience, so as to rationally decide food purchase and scientifically plan food intake. Quantified self will identify residents' self-health level and personality needs, further more intuitively understand self-status [15], judges product effectiveness based on quantitative data information, establish the association between product data indicators and consumers themselves, and realize accurate and rational consumption decision [16]. Therefore, we propose Hypotheses 2 and 3.

Hypothesis 2. Quantified self-behavior is helpful to strengthen consumers' willingness to buy healthy food.

Hypothesis 3. Quantified self-behavior is helpful to increase purchasing amount of healthy food consumption.

2.3. The Effect of Social Networks on Quantified Self-Behavior and Healthy Food Consumption. Consumers' searching cost is high because they usually rely on themselves to collect information about food safety, nutrition, and so on. Social networks provide consumers with information channels [17]. In social networks, information sharing behavior is a common behavior, which network members share information accidentally found or needed by others [18]. Information sharing behavior is a very important social behavior, which often occurs in the network or social groups, and is not a special behavior of individual [19]. Instead, it is a process of cooperation among network members under the condition of social networks, in which information providers transfer information to information searchers [20]. Also, social media influencers can shape corporate brand reputation through online followers' trust,

value creation, and purchase intentions [21]. In the sharing economy platforms, such as social networks, digital personal reputation and feedback systems facilitate interaction and trust between strangers and further form customer loyalty [22]. This means that the trust formed during the interaction of social networks further produce an effect on the decision of buying [23]. Another aspect, when customers develop a sense of trust in each other in interaction of social networks, consumer cognition will also affect the decision-making behavior. Drugău-Constantin and Mirică pointed out that consumer cognition could be reducible to neurophysiological functioning, and this would influence consumers' choice [24, 25]. That is to say, social networks provide information channels for members to share information, which is related to individuals' health or food safety. This information further help individuals to improve their quantified self-behavior. On other way, social networks can influence quantified self-behavior through information sharing among members and then affect consumers' healthy food consumption.

Social networks are composed of the relationship between different members. The more the members, the more complex the connection and the more complex the social network. Social networks with different complexity may affect both quantified self-behavior and food consumption behavior. Assuming that the network is only a star-style network with four consumers (Figure 1(a)), the core consumer S_1 can adopt information sharing strategy after obtaining the information, whether it is quantitative information or food health information, the rest of consumers can get it for free, while they can share the information again or choose not to share it. In Figure 1(b) star network with six consumers, consumers choose the sharing strategy as (a), but because it has more network connections, the speed of information transmission and sharing is wider. In the type of Figure 1(c) network connection, in addition to the core consumers and other consumers, there are also connections between other consumers such as S_{21} and S_{22} . In this case, the probability of other consumers choosing information sharing strategy will increase, that is, the multiconnection relationship between different members in the same network will strengthen the information sharing and transmission. In addition, social networks can also be reconnected through members. In the two star networks (a) and (b) of consumers, there are two kinds of connection choices: one is to form a new network structure as Figure 1(d) for the connection between the core consumers of star network (a) and the other consumers of star network (b); the other is to form a new network structure as Figure 1(e) for the increase in connection between the core consumers of star network (a) and the core consumers of star network (b). However, compared with social network (d), social network (e) breaks the original equilibrium and forms a new network because of the connection between core consumers. The network scope is wider, and core consumers will choose information sharing, whereas other consumers may choose not to share information because of free-riding behavior. Thus, the sharing behavior in network (e) is more wide. It means the more complex the social networks are, the stronger the

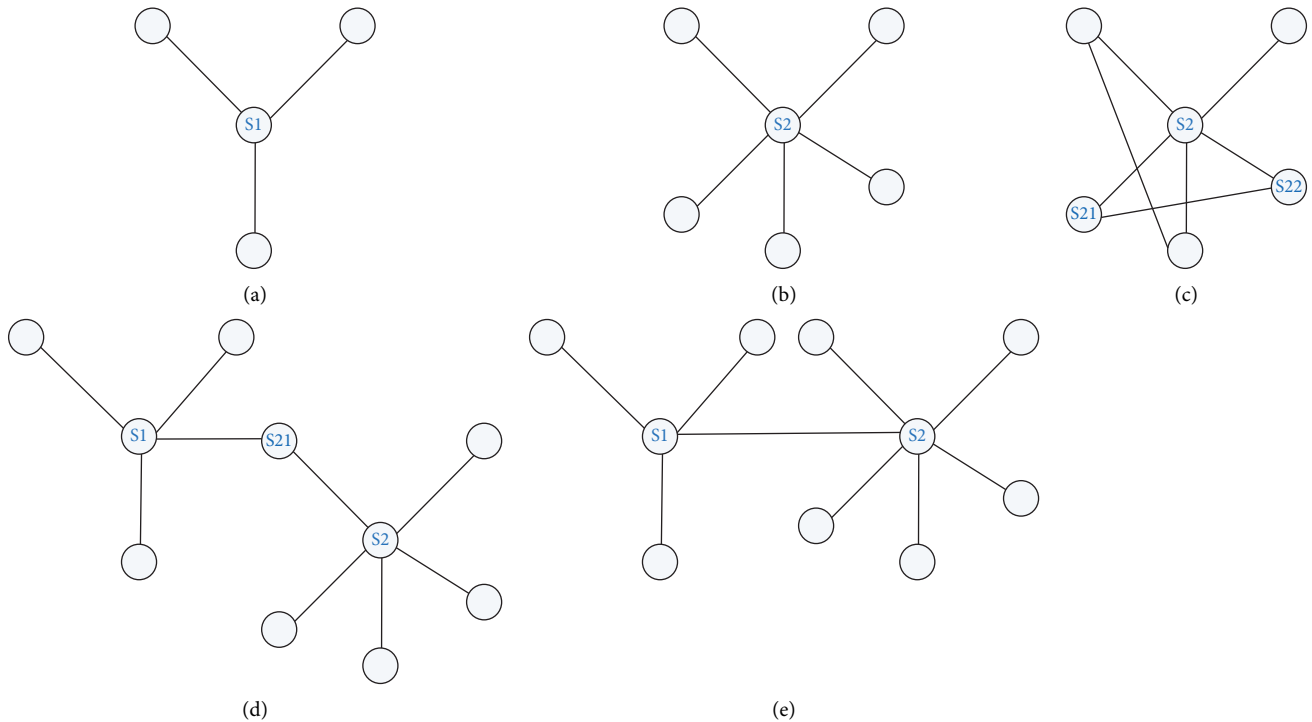


FIGURE 1: Social networks of different complexity.

impact of quantified self-behavior on consumers' healthy food consumption is. Therefore, we propose Hypothesis 4.

Hypothesis 4. Social networks will moderate the relationship between quantified self and consumers' healthy food consumption.

With the continuous development of China's social economy and the increase in population flow, the traditional concept of family has been affected. Individuals begin to pay more attention to individual value and interests and change from survival rationality to social rationality [26]. With the change, the coverage and strength of social relationship have changed from mainly emotional network to instrumental network mixed with emotional network. Emotional network is based on family concept and blood relationship, which has strong stability and nonselectivity [26]. Moreover, social network based on emotion and blood relationship pays more attention to individual health than individual achievement. Just like parents are more concerned about their children's health instead of their children's academic and career achievements. The instrumental network is mostly based on the relationship of career or classmates. Network members will establish more externalized social relations according to their own purposes and needs and network connection is mainly based on mutual interests. More attention is paid to individual achievement rather than physical health. For example, instrumental network members are more likely to start a business and share employment experience together rather than focus on health together [27]. In instrumental network, the information shared are

more career related than health related. Therefore, we propose Hypothesis 5.

Hypothesis 5. Compared with instrumental network, emotional network has a stronger moderating effect.

3. Materials and Methods

3.1. Subjects and Methods

3.1.1. Experimental Operation. The purpose of this experiment is to examine the impact of quantified self on residents' willingness and scale of healthy food consumption. The subjects were randomly divided into control group (CQG), active quantitative group (AQG), and passive quantitative group (PQG). The control group did not provide any information. The passive quantitative experimental group was offered food information including protein, carbohydrate, calorie, and so on, and provided the minimum nutrients needed by the human body every day. The food composition information of the active quantitative experimental group was hidden, and the subjects could actively view the food composition information or directly select without viewing it.

90 students from a university in Shaanxi participated in the experiment, and the subjects were arranged to participate in the experiment within a certain period. Before the experiment, the basic statistical characteristics, such as gender, age, food consumption preferences, and habits of the subjects were investigated with a short questionnaire. The subjects were randomly divided into control group (CQG), active quantization group (AQG), and passive quantization

group (AQG). The demographic characteristics of the whole sample are shown in Figure 2.

From Figure 2, in terms of gender, the distribution of the control group and the experimental group is similar, and the number of male subjects is equal to or more than that of female subjects. In the control group, male and female subjects accounted for 50%, respectively. In the active quantitative group and the passive quantitative group, male subjects accounted for 53.33% and 63.3%, respectively. In terms of age distribution, the minimum age of the subjects is 17 years, the maximum age is 25 years, and the overall age distribution is relatively uniform. For example, in the control group, 43.3% of the subjects were younger than 20 years, 36.7% of the subjects were between 21 and 23 years, and 20% of the subjects were older than 24 years. It can be seen from Figure 2 that the distribution of gender and age in different groups are similar. Because the samples are randomly assigned to different groups, the results show that the comparison between different groups is reliable.

3.1.2. Investigation Research. Based on the in-depth understanding of the existing research maturity scale and interviews with experts, we developed the related variable measurement scale w combined with the research situation. After the prediction test of 5 scholars in related fields and 30 ordinary consumers, this study finally determined the formal questionnaire. In order to ensure the efficient development of the survey, in addition to using the network questionnaire survey, this study also took some urban and rural residents and college students in Shaanxi Province and Zhejiang Province as the survey subjects. A total of 1000 questionnaires were distributed, and 987 questionnaires were collected. After screening out 45 invalid questionnaires, such as missing answers and consistent selection of test items, 942 valid questionnaires were finally obtained (the effective recovery rate was 94.2%). The demographic characteristics of the respondents are shown in Table 1.

3.2. Variable Selection and Measurement

3.2.1. Quantified Self. Combined with the research viewpoints of Zhang et al. [6] and Zhou et al [28], 10 items of consumers' quantitative information preference and quantitative willingness to participate in consumption were determined, respectively, including "I am very concerned about food nutrition," "I prefer healthy food," "I am willing to rationally consume food by evaluating food nutrition," "I am willing to participate in food consumption by evaluating food nutrition," "I am willing to choose safe alternative food with the same nutritional efficiency according to the dietary needs after quantified self." In this study, the confirmatory factor analysis of the questionnaire fitted well: $\chi^2/df = 3.26$, RMSEA = 0.06, NFI = 0.998, GFI = 0.999, and CFI = 0.999. Internal consistency coefficient of questionnaire is 0.742. Analyzing dimension reduction by SPSS, according to the standard of eigenvalue greater than 1, two variables were obtained: consumer quantitative information preference and quantitative willingness and five items belong to the former

and five items belong to the latter. The results are consistent with the expectation of the questionnaire design.

3.2.2. Healthy Food Consumption. Referring to the measurement scale developed by Penning and Wansink, the scale of residents' healthy food consumption is compiled, including 10 items of healthy food consumption willingness and scale, such as "I prefer organic food to ordinary food," "I will pay attention to the nutritional components of food when I buy it," "I will pay attention to whether food contains harmful ingredients such as additives when I buy it," "I will pay attention to the food quality when I buy it," "The amount of pork purchased per week," "the amount of organic pork purchased per week," "the cost of fruit purchased per week," and "the frequency of eating instant noodles and other fast food products per week." In this study, the confirmatory factor analysis of the questionnaire fitted well: $\chi^2/df = 2.84$, RMSEA = 0.07, NFI = 0.999, GFI = 0.998, and CFI = 0.999. Internal consistency coefficient of questionnaire is 0.728. Through factor analysis and dimensionality reduction, two variables were obtained according to the standard of eigenvalue greater than 1: consumer's food consumption intention and food consumption scale, with five items belonging to the former, four items belonging to the latter, and one item was deleted due to low load value. The results were basically consistent with the expectation of the questionnaire design.

3.2.3. Social Networks. The social network questionnaire compiled by Fang [29] adopted 13 items, such as the number of my brothers and sisters, the number of my friends in wechat group, the number of communities I join in on the Internet, and so on. Instrumental network contains 6 items, such as "I keep close contact with many classmates" and the like; emotional network contains 7 items, such as "I have close relationship with relatives of the same age" and the like (Kim and Lee). In this study, the confirmatory factor analysis of the questionnaire fitted well: $\chi^2/df = 2.69$, RMSEA = 0.08, NFI = 0.999, GFI = 0.999, and CFI = 0.98. Internal consistency coefficient of questionnaire is 0.676. According to the criterion of eigenvalue greater than 1, two variables were obtained: emotional network and instrumental network, with 7 items belonging to the former and 6 items belonging to the latter. The results were consistent with the expectation of the questionnaire design.

3.3. Empirical Methods and Models

3.3.1. Heckman Two-Step Method. Heckman two-step method is mainly used to deal with sample bias and self-selection problems, and it can also solve endogenous problems in self-selection behavior. In this article, the residents also have the problems of self-selection and sample bias. To study the influence of quantified self on the willingness and scale of healthy food consumption, the equation of quantified self and scale of food consumption is set as follows:

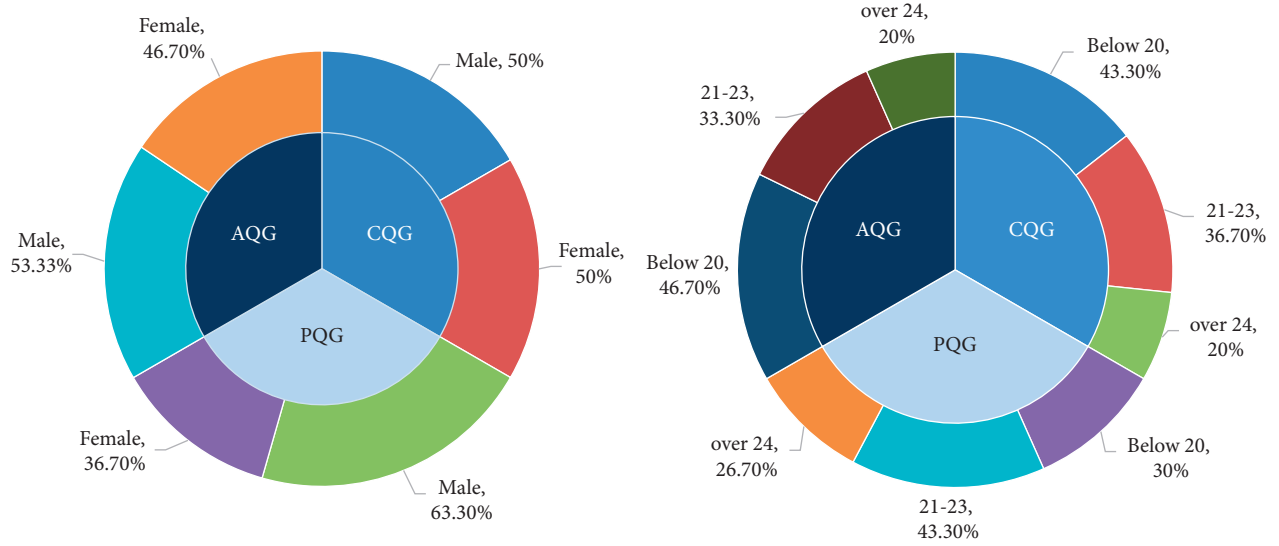


FIGURE 2: Distribution of gender and age in different groups.

TABLE 1: Demographic characteristics of respondents.

Category	Number	Percentage	Category	Number	Percentage		
Gender	Male	469	49.79	Average net monthly income	Below \$300	249	26.43
	Female	473	50.21		\$300-\$800	225	23.89
Age	Below 25	43	4.56		\$800-\$1500	334	35.46
	26-35	298	31.63	Above \$1500	134	14.23	
	36-45	291	30.89	Residence	Urban	445	47.24
	Above 46	310	32.91		Rural	497	52.76
Education	Below bachelor degree	331	35.14	High	598	63.48	
	Above bachelor degree	611	64.86	Health awareness	Medium	210	22.29
Minimum net monthly income	Below \$150	221	23.46		Low	134	14.23
	\$150-\$300	247	26.22	High	250	26.54	
	\$300-\$800	346	36.73	Health knowledge	Medium	289	30.68
	Above \$800	128	13.59		Low	403	42.78

$$y_{1i} = \chi_{1i}\beta_1 + \varepsilon_{1i}. \quad (1)$$

Here, y_{1i} is scale of healthy food consumption and x_{1i} is quantified self. The quantified self choice equation was as follows:

$$y_{2i} = \chi_{2i}\beta_2 + \varepsilon_{2i}. \quad (2)$$

Here, y_{2i} is whether to choose quantified self and x_{2i} is factors influencing self quantification, such as age, gender, education, and so on. If equation (1) is direct estimated, we will get the estimator with selective bias.

It can be seen from the model

$$E(\varepsilon_{1i} | y_{2i}^* \geq 0) = E(\varepsilon_{1i} | \varepsilon_{2i} \geq -\chi_{2i}\beta_2), \quad (3)$$

and

$$E(y_{1i} | \chi_{1i}, y_{2i}^* \geq 0) = \chi_{1i}\beta_1 + E(\varepsilon_{1i} | \varepsilon_{2i} \geq -\chi_{2i}\beta_2). \quad (4)$$

This indicates if we directly use equation (1) to estimate β_1^{\wedge} , we ignore the conditional mean value ε_{1i} . Furthermore, equation (4) can be written as:

$$E(y_{1i} | \chi_{1i}, y_{2i}^* \geq 0) = \chi_{1i}\beta_1 + \rho\sigma_1\lambda_i. \quad (5)$$

Then, we get,

$$y_{1i} = \chi_{1i}\beta_1 + \rho\sigma_1\lambda_i + \mu_i, \quad (6)$$

where ρ is the correlation coefficient of ε_{1i} , ε_{2i} ; σ_1 is the standard deviation of ε_{1i} , and σ_2 is the standard deviation of ε_{2i} .

Firstly, using Probit model to estimate equation (2) based on all samples, the tendency of all samples to choose quantified self-behavior was obtained. Using the estimated results, we can calculate λ_i .

Secondly, regard $\rho\sigma_1$ as a parameter to be estimated using the sample of choosing quantified self and estimate equation (6) to get β_1^{\wedge} .

3.3.2. *PLS Path Modeling.* PLS path model mainly includes two parts: measurement model is called external model, which describes the relationship between latent variables and measurable variables. Structural model, also known as internal model, describes the relationship between endogenous latent variables and exogenous latent variables, as well as the relationship between exogenous latent variables. This study mainly uses PLS modeling to analyze the role of social networks in quantified self influencing on residents' health food consumption.

Let ξ be a latent variable and X_h be a measurable variable, then, the relationship between ξ and X_h can be expressed in three forms: reflection type, constitutive type, and polygenetic type.

(1) *Reflection Type.* In reflection type, latent variable (LV) reflects every measurable variable (MV), and every measurable variable can be expressed as a simple regression equation about its latent variable:

$$X_h = \pi_{h_0} + \pi_h \xi + \varepsilon_h. \quad (7)$$

The mean value ξ is 0, and the standard deviation is 1, which satisfies the forecast specification conditions:

$$E\left(\frac{X_h}{\xi}\right) = \pi_{h_0} + \pi_h \xi. \quad (8)$$

When the reflection type appears in the model, the first step is to test the unique dimension using principal component analysis, Cronbach's α coefficient, and Dillon Goldstein's ρ coefficient.

Principal component analysis shows that if the first eigenvalue root is greater than 1, the second eigenvalue root is less than 1, or the second eigenvalue root is less than the first eigenvalue root, then the group of measurable variables is unique [30]. Therefore, it can be determined that the first principal component positively correlated with all or at least most of the measurable variables. If there is a negatively correlated measurable variable, it can be considered that the

variable cannot fully reveal its potential variables and should be removed from the model summary.

The premise of Cronbach's coefficient as the test standard of the only dimension is that a group of p -dimensional measurable variables positively correlated with each other. Firstly, the group of variables should be standardized. Write the variance as follows:

$$\text{Var}\left(\sum_{h=1}^p X_h\right) = p + \sum_{h \neq h'} \text{cor}(X_h, X_{h'}), \quad (9)$$

where the larger the $\sum_{h \neq h'} \text{cor}(X_h, X_{h'})$, the more the measurable variables meet the requirement of unique dimension, thus introducing α' ,

$$\alpha' = \frac{\sum_{h \neq h'} \text{cor}(X_h, X_{h'})}{p + \sum_{h \neq h'} \text{cor}(X_h, X_{h'})}. \quad (10)$$

It is found that Cronbach's coefficient reaches the maximum value $(P-1)/P$ when $\text{cor}(X_h, X_{h'}) = 1$

$$\alpha = \frac{\sum_{h \neq h'} \text{cor}(X_h, X_{h'})}{p + \sum_{h \neq h'} \text{cor}(X_h, X_{h'})} \times \frac{p}{p-1}. \quad (11)$$

The equation (11) can be transformed as

$$\alpha = \frac{\sum_{h \neq h'} \text{cor}(X_h, X_{h'})}{\text{Var}\left(\sum_{h=1}^p X_h\right)} \times \frac{p}{p-1}. \quad (12)$$

If Cronbach's is greater than 0.7, it indicates that the more the ratio of autocorrelation coefficient to variance approaches its maximum, the group of measurable variables meet the unique dimension.

Dillon Goldstein's ρ coefficient is slightly better than Cronbach's coefficient in the evaluation of uniqueness. The coefficient is mainly set based on the simple regression model between latent variables and measurable variables. Firstly, the variance $\sum_{h=1}^p X_h$ is calculated according to equation (7) and the residual ε_h is assumed to be independent.

$$\text{Var}\left(\sum_{h=1}^p X_h\right) = \text{Var}\left(\sum_{h=1}^p (\pi_{h_0} + \pi_h \xi + \varepsilon_h)\right) = \left(\sum_{h=1}^p \pi_h\right)^2 \text{Var}(\xi) + \sum_{h=1}^p \text{Var}(\varepsilon_h). \quad (13)$$

The larger of $(\sum_{h=1}^p \pi_h)^2$, the more the group of variables meet the requirement of unique dimension, which defined ρ as follows:

$$\rho = \frac{(\sum_{h=1}^p \pi_h)^2 \text{Var}(\xi)}{(\sum_{h=1}^p \pi_h)^2 \text{Var}(\xi) + \sum_{h=1}^p \text{Var}(\varepsilon_h)}. \quad (14)$$

Assuming that both the measurable variable X_h and the latent variable ξ are standardized, the latent variable ξ can be estimated by the first principal component t_1 of the measurable variable, and π_h can be estimated by the similarity coefficient $\text{cor}(X_h, t_1)$ and the first principal component t_1 .

$\text{Var}(\varepsilon_h)$ can be estimated by $1 - \text{cor}^2(X_h, t_1)$. Therefore, the estimate of Dillon-Goldstein's ρ is given as

$$\hat{\rho} = \frac{(\sum_{h=1}^p \text{cor}(X_h, t_1))^2}{(\sum_{h=1}^p \text{cor}(X_h, t_1))^2 + \sum_{h=1}^p (1 - \text{cor}^2(X_h, t_1))}. \quad (15)$$

If $\hat{\rho}$ is greater than 0.7, The group of measurable variables is considered to be unique.

(2) *Constructive Type.* In the constructive form, the latent variable ξ is generated by a group of measurable variables, which can be expressed as the sum of the weighted residuals:

$$\xi = \sum_h \bar{w}_h X_h + \delta. \quad (16)$$

In constructive form, measurable variables can belong to multiple latent variables. At the same time, it meets the prediction criteria:

$$E\left(\frac{\xi}{X_1, \dots, X_{pj}}\right) = \sum_h \bar{w}_h X_h. \quad (17)$$

Suppose that the mean value of the residual vector is 0, and it is not related to the measurable variable X_h . In the PLS algorithm parameter estimation, if the symbol is related, the variable should be deleted.

4. Results and Discussion

4.1. Results

4.1.1. The Impact of Quantified Self on Consumers' Healthy Food Consumption

(1) *Experimental Results.* Firstly, it analyzes whether different consumers engaged in quantified self-behavior. From the experimental subjects, consumers' quantified self-awareness and actual quantified self-behavior are different in different gender and age groups. From the perspective of gender (Figure 3), male's overall quantitative awareness is low, only 30% of all male subjects have high quantitative awareness while 48% of female subjects have high quantitative awareness. In the actual quantified self-behavior, 68.75% of the male subjects took the quantitative behavior in the process of choosing food, and 85.71% of the female subjects took the actual quantitative behavior. Therefore, from the perspective of gender, female subjects have higher quantitative awareness and behavior than male subjects.

From the perspective of age (Figure 4), the older the subjects are, the stronger their quantitative awareness is; 69.44% of the subjects younger than 20 years, 67.65% of the subjects between 21 and 23 years, and 75% of the subjects older than 24 years have medium or high quantitative awareness. But the results of quantified self-behavior show different trends: in the group of 21- 23-year-old subjects, 90% adopted quantified self-behavior, whereas in the 20-year-old and 24-year-old subjects, about 70% chose quantified self-behavior. It can be seen that age has no significant difference in the choice of quantified self-behavior. Combined with the quantitative consciousness and quantitative behavior of the subjects, it can be seen that more than 65% of the subjects will choose quantified self-behavior.

In addition, by observing the relationship between healthy food awareness and quantified self-behavior (Figure 5(a)), we can find that the subjects with healthy food awareness are more likely to choose quantified self. Only 70% of the subjects with low healthy food awareness chose quantified self, while 75% and 87.5% of the subjects with moderate or high health food awareness chose quantified self, respectively. With the improvement of healthy food awareness, the probability of consumers choosing to quantify increases, and Hypothesis 1 is verified.

The results of independent sample T test shows that there is no significant difference in consumption willingness of healthy food between CQG and PQG (Figure 5(b)). Compared with CQG, subjects of PQG report slightly higher willingness to consume healthy food ($M = 1.83$, $SD = 0.83$ vs $M = 2.1$, $SD = 0.80$, $F(1, 58) = 1.59$, $p = 0.2121$). However, there are significant differences in healthy food consumption intention between COQ and AQG, as well as PGQ and AQG. Compared with CQG, AQG has higher healthy food consumption intention ($M = 2.47$, $SD = 0.62$ vs $M = 1.83$, $SD = 0.83$, $F(1, 58) = 11.03$, $p = 0.0016$), indicating that the subjects who have the initiative to take quantified self-behavior have higher healthy food consumption intention. AQG has higher health food consumption intention than PQG ($M = 2.47$, $SD = 0.62$ vs $M = 2.1$, $SD = 0.80$, $F(1, 58) = 3.87$, $p = 0.0537$), which indicates that the subjects with active quantified self-behavior have higher healthy food consumption intention than those with passive quantified self.

As for the influence of quantified self on the consumption scale of healthy food, the result of independent sample T test shows that there is no significant difference between CQG and PQG (Figure 5(c)). Compared with CQG, PQG report slightly higher consumption scale of health food ($M = 23.43$, $SD = 3.97$ vs $M = 21.78$, $SD = 5.65$, $F(1, 58) = 1.70$, $p = 0.1968$). However, there are significant differences in the scale of healthy food consumption between COQ and AQG, as well as PQG and AQG. The subjects of AQG have higher scale of healthy food consumption than CQG ($M = 27.52$, $SD = 5.52$ vs $M = 21.78$, $SD = 5.65$, $F(1, 58) = 15.79$, $p = 0.0002$). Compared with PQG, AQG has higher consumption scale of health food ($M = 27.52$, $SD = 5.52$ vs $M = 23.43$, $SD = 3.97$, $F(1, 58) = 10.82$, $p = 0.0017$), which indicates that the subjects with active self-quantification have higher consumption scale of health food than those with passive self-quantification.

Through the analysis of the influence of CQG, PQG, and AQG on consumers' willingness and scale of healthy food consumption, it can be found that the subjects with active quantification have higher willingness and scale of food consumption, and quantified self-behavior can promote consumers' willingness and scale of healthy food consumption. Hypotheses 2 and 3 are verified.

(2) *Empirical Results.* Heckman two-step method is used to test whether consumers choose quantified self-behavior and the relationship between quantitative behavior and consumers' healthy food consumption intention and scale. The results are shown in Table 2. In that, column (1) is the Probit regression of consumers' quantitative behavior, and column (2) and column (3) indicate the influence of quantified self-behavior on consumers' willingness and scale of healthy food consumption under the control of self-selection bias.

From the results of column (1), the constant coefficient is 0.324, which means that the ratio of quantitative to non-quantified is 0.324 without considering other influences. The ratio of consumer selection quantification to nonquantified is 1.383. It shows that in the survey sample, the individual who chooses quantification is 38.3% higher than the

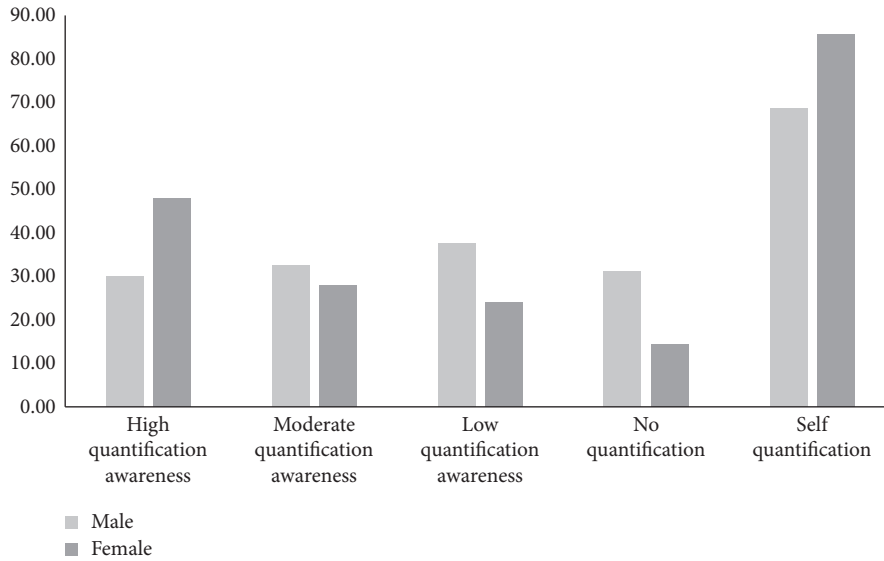


FIGURE 3: Gender differences in consumers' quantitative self-awareness and behavior.

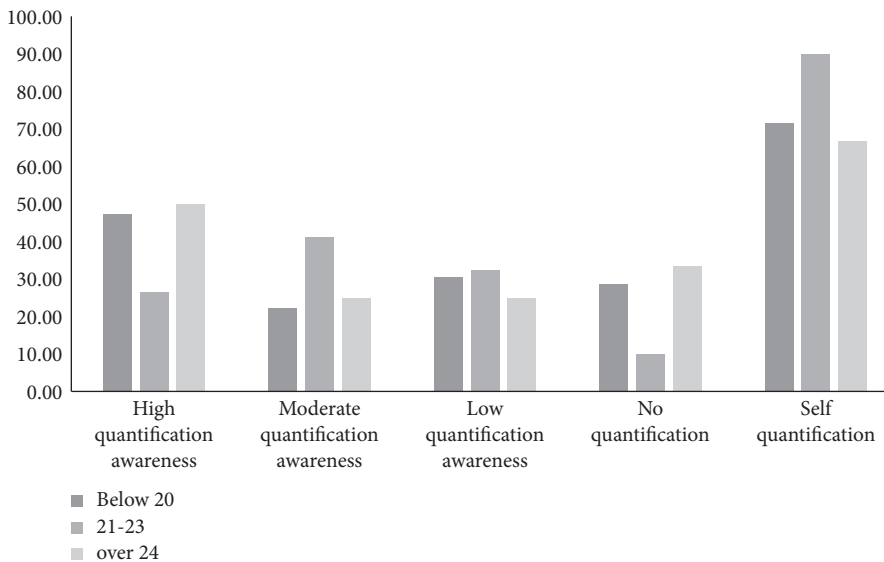


FIGURE 4: Gender differences in consumers' quantitative self-awareness and behavior.

individual who has not selected quantification. In addition, health awareness has a significant positive effect on consumers' quantified self choice. The coefficient after logarithm transformation is 1.035, which shows that the consumers with high health awareness are 3.5% higher than those with low health awareness. Hypothesis 1 is verified again. From the results of column (2), the consumer quantified self has a significant positive effect on the consumption willingness of healthy food, with a coefficient of 0.292 and significant at the level of 1%. It shows that under the condition, the consumer's quantified self-behavior is improved, and the consumers' willingness to consume healthy food products is increased by 29.2%, Hypothesis 2 is verified. The results of column (3) show that the influence of consumer quantified self on consumption scale of healthy food is 0.351, which shows that with the improvement in consumers' quantified

self-behavior, the consumption scale of healthy food of consumers increases by 35.1%, and Hypothesis 3 is verified.

4.1.2. The Moderating Role of Social Networks. Furthermore, PLS path modeling is used to test the moderating role of social networks in the process of quantified self impact on healthy food consumption. The results are shown in Table 3. In that, columns (1)–(3) are the moderating results of social networks in the process of quantified self-influence on consumers' healthy food consumption intention, and columns (4)–(6) are the moderating results of social networks in the process of quantified self-influence on consumers' healthy food consumption scale.

From the results of column (1), the quantified self-behavior and social networks have positive and significant effects on

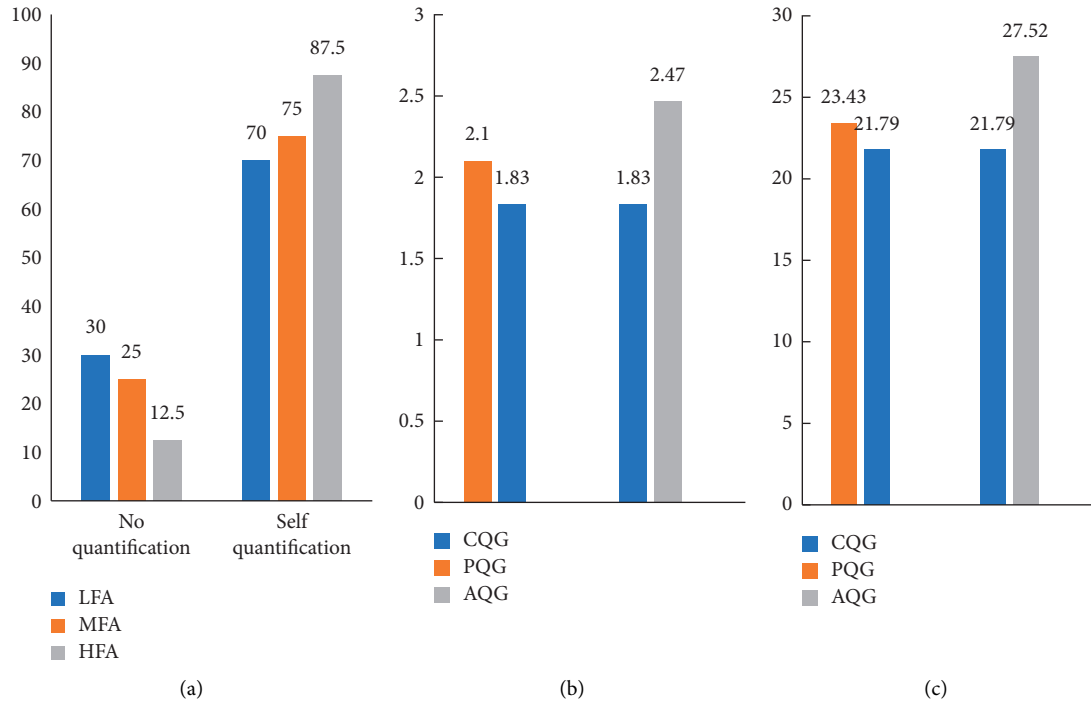


FIGURE 5: The effect of quantified self on consumers' willingness and scale of healthy food consumption. (a) Quantitative behavior choice. (b) Consumption awareness. (c) Consumption scale.

TABLE 2: The influence of quantified self on consumers' willingness and scale of healthy food consumption.

	(1) Quantified self	(3) Consumption intention	(4) Consumption scale
Constant	0.324*** (0.031)	-0.963*** (0.224)	-0.383 (0.228)
Quantified self		0.292*** (0.032)	0.351*** (0.033)
Register	0.118 (0.104)	0.153* (0.068)	0.212** (0.070)
Gender	-0.030 (0.089)	0.023 (0.059)	0.004 (0.060)
Age	0.051*** (0.006)	0.030*** (0.004)	0.004 (0.004)
Education	0.102 (0.094)	0.132* (0.062)	0.137* (0.063)
Income	0.122** (0.046)	0.002 (0.031)	0.052 (0.031)
Health knowledge	0.045 (0.056)	0.058*** (0.017)	0.060** (0.028)
Health awareness	0.034*** (0.011)	0.015 (0.040)	0.046*** (0.010)
<i>N</i>	942	942	942
Adj. <i>R</i> ²		0.233	0.206

Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

consumers' willingness to purchase healthy food with coefficients of 0.190 and 0.273, respectively, and both are significantly at the level of 1%. At the same time, the intersection of quantified self and social networks has a positive impact on the consumption intention, which means that social networks will

moderate the impact of quantified self-behavior on the consumption willingness. That is, the more complex the social networks are, the stronger the impact of quantified self-behavior on consumers' consumption of healthy food is. Column (4) is the moderating effect of social networks on the

TABLE 3: The moderating role of social networks.

	(1)	(2)	(3)	(4)	(5)	(6)
	Consumption intention			Consumption scale		
Constant	0.060 (0.035)	0.060 (0.035)	0.021 (0.029)	0.051 (0.035)	0.047 (0.034)	0.013 (0.029)
Quantified self	0.190*** (0.043)	0.143*** (0.043)	0.371*** (0.030)	0.197*** (0.043)	0.120** (0.042)	0.373*** (0.029)
Social network	0.273*** (0.044)			0.279*** (0.044)		
Quantified self * social network	0.082** (0.026)			0.069** (0.026)		
Emotional network		0.171*** (0.030)			0.228*** (0.030)	
Quantified self * emotional network		0.116*** (0.028)			0.073** (0.028)	
Instrumental network			0.334*** (0.044)			0.382*** (0.043)
Quantified self * instrumental network			0.080** (0.027)			0.063* (0.026)
Control variables	Control	Control	Control	Control	Control	Control
N	942	942	942	942	942	942
Adj. R^2	0.219	0.234	0.205	0.223	0.252	0.226

Standard errors in parentheses, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

consumption scale of healthy food, and its moderating coefficient is 0.069. Hypothesis 4 is verified.

Columns (2) and (5) show the moderating effects of emotional social networks on the relationship between quantified self and healthy food consumption. The results show that whether it is consumers' willingness to consume healthy food or consumption scale, emotional social networks have positive moderating effects with coefficients of 0.116 and 0.073 respectively. The results of column (3) and (6) show that the instrumental networks also have positive moderating effects with coefficients of 0.080 and 0.063, and the coefficients are significantly at the level of 5% and 10%. The results of emotional networks and instrumental networks further verify Hypothesis 4. At the same time, from the coefficients, whether it is healthy food consumption intention or consumption scale, the cross-term coefficient of quantified self and emotional networks is greater than that of quantified self and instrumental networks, indicating that the moderating effect of emotional networks is stronger, and Hypothesis 5 is verified.

4.2. Discussion. The results of experimental and empirical tests are consistent with the conclusion of Ortega et al. [31] and Pocol et al. [32] but has differences. The results of Ortega et al. [31] showed that the consumers' selection of food safety was affected by consumer preferences; they measured consumer preferences for select food safety attributes in pork and took food safety risk perceptions into account. Several choice experiment models, including latent class and random parameters logit, were constructed to capture heterogeneity in consumer preferences. Their results suggest that Chinese consumers have the highest willingness to pay for a government certification program, followed by third-

party certification, a traceability system, and a product-specific information. But our results show that consumers' willingness and scale of healthy food consumption could be also affected by quantified self-behavior. Pocol et al. [32] explored Generation Z university students' clusters based on the consumption of daily fruits and vegetables in an emerging market economy and found most cluster members are aware of the value of regular fresh fruit and vegetable consumption in order to maintain health and overall well-being [33]. Our results complement the conclusion that customers with quantified self-behavior will have higher willingness to buy safety food.

The results of moderating role of social networks show the similar situation of Kim and Lee [34]. They pointed out that the more friends in Facebook, the more likely individuals chose self-presentation, then individuals would make similar decisions. Hu et al. [35] used Bayesian personalized ranking based on multiple-layer neighborhoods and pointed out that social networks would affect individuals' behavior. Though Marcel et al. [36] stated that the information sources of consumers' selection of food safety was relatives and friends, our results shows that emotional social network plays a stronger role in quantifying self-influence on healthy food consumption behavior.

In the context of big data, the digital divide encourages opportunistic enterprises to exploit the rights and interests of consumers, and the emergence of quantitative self indicates the beginning of comprehensive digitization in the field of consumer cognition. Through the tracking and measurement of self-behavior state, consumers' health status and life process are becoming more and more visual and predictable. The implementation and penetration of quantitative self makes consumers' demand judgment and behavior choice more and more accurate and rational. This is very

important for consumers to pay attention to healthy food consumption or nutrition.

Based on the theory of information perception and the risk perception theory, combined with social cognition, this study explores the internal mechanism of self quantify on healthy food consumption, which brings new insights for analyzing quantitative behavior and understanding the formation of consumers' healthy food consumption participation intention. In terms of research perspective, due to the limitation of perspective, relevant studies lack holistic thinking on the internal mechanism of the formation of consumers' willingness to buy healthy food. Breaking through the limitations of existing studies focusing on the individual level to explore relevant issues, combined with the reality of the increasing socialization of quantitative self under the support of technology, this study more accurately explains the connotation of quantitative self-concept from the perspective of the combination of individual and community, and comprehensively understands the formation mechanism of consumers' willingness to quantitative self in healthy food consumption. In terms of research content, most of the existing studies regard constraints such as limited time and energy, insufficient operating skills, and unhealthy habits as the main reasons for consumers' quantifying themselves in healthy food consumption and do not consider the impact of social networks. From the perspective of individual differences, this study constructs a theoretical framework for analyzing consumers' healthy food consumption based on information perception and the risk perception theory, which makes a reasonable supplement to the research on the consumers' quantitative self-behavior and healthy food consumption behavior. Focusing on the essence of quantitative self as a socialized practice and relying on social cognitive theory, this article explains the differentiated choice of consumers on whether to participate in quantitative self in the face of social networks. From the perspective of relationship, the internal mechanism of different social networks on consumers' healthy food consumption is introduced to further clarify the applicable boundary of the theory. In terms of research methods, although the existing studies recognize the necessity of participation in the realization of quantitative self positive utility and emphasize the importance of clarifying the internal mechanism of consumers' quantitative self participation behavior on healthy food consumption, the understanding of relevant issues is still in the stage of descriptive exploration. Based on the literature review, this study forms relevant hypotheses, which are verified by a series of experiments and questionnaires, and defines and confirms the internal relationship of various elements in the relationship of consumers' quantitative self and healthy food consumption.

5. Conclusions

This study analyzed consumers' quantified self-behavior and healthy food consumption through laboratory experiments and field investigation and explored the role of social networks in quantified self-behavior and healthy food consumption from the perspective of complex network. Through theoretical analysis and empirical research, this study obtains three main

results: (1) Health awareness can promote consumers to choose quantified self-behavior; (2) consumers' quantified self-behavior is helpful to promote their purchase intention and purchase scale of healthy food; (3) social networks play a positive moderating role in the relationship of quantified self-behavior and healthy food consumption. Both emotional networks and instrumental networks have significant moderating effects, but the former is stronger.

This article discusses the impact of quantitative self on consumers' healthy food consumption from the perspective of social network and deeply explores how different social networks affect the relationship between the two. Also, this article expands the theory of information perception and the risk perception theory, and it strengthens the role of information cognition, information transmission, and risk cognition in the connection of different social networks, which will deeply affect the impact of quantifying self on food consumption. In addition, in terms of practical significance, the conclusions of this article can provide practical guidance for food enterprises to optimize product packaging and quantify innovative product design. In addition, the relevant conclusions can also provide some support for the government to guide residents to establish food nutrition concept, form quantitative self habits, make food consumption decisions, and buy food reasonably.

Compared with previous studies, this study has some innovations, such as using experimental data and survey data at the same time to strengthen the diversity of data sources and the reliability of empirical results. This article discusses the quantified self behavior and healthy food consumption behavior of consumers, which is helpful to expand the research of consumer behavior from the cross perspective of behavior and economics. However, there are also some shortcomings in this study. For example, although the scale of this study cites the mature scale of previous studies as reference, due to the diversity and complexity of consumers' behavior and social networks, the questionnaire items may not fully reflect the relevant variables. In addition, this study mainly focuses on the moderating role of social networks, but whether there are other roles needs to be further explored in the follow-up research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by National Natural Science Foundation of China (72003113), Humanities and Social Sciences Research Fund of the Ministry of Education (18YJC890012), and Soft Science General Project of Shaanxi Innovation Capability Support Plan (2020KRM118).




References

- [1] H.-Y. Lin and M.-H. Hsu, "Using social cognitive theory to investigate green consumer behavior," *Business Strategy and the Environment*, vol. 24, no. 5, pp. 326–343, 2015.
- [2] M. A. Mccrory, P. J. Fuss, N. P. Hays, A. G. Vinken, A. S. Greenberg, and S. B. Roberts, "Overeating in America: association between restaurant food consumption and body fatness in healthy adult men and women ages 19 to 80," *Obesity Research*, vol. 7, no. 6, 2012.
- [3] L. Robertsson, "Quantified self: an overview and the development of a universal tracking application," Unpublished Doctoral Dissertation, Umea University, Västerbotten, 2014.
- [4] M. Almalki, K. Gray, and F. Martin-Sanchez, "Activity theory as a theoretical framework for health self-quantification: a systematic review of empirical studies," *Journal of Medical Internet Research*, vol. 18, no. 5, pp. 131–148, 2016.
- [5] B. K. Choi, H. K. Moon, and E. Y. Nae, "Cognition- and affect-based trust and feedback-seeking behavior: the roles of value, cost, and goal orientations," *Journal of Psychology*, vol. 148, no. 5, pp. 603–620, 2014.
- [6] Y. Zhang, D. Li, and H. Jin, "Security risk perception, quantitative information preference and willingness to participate in consumption: decoding the decision logic of food consumers," *Modern Finance-Journal of Tianjin University of Finance and Economics*, vol. 39, no. 1, pp. 86–98, 2019.
- [7] M. Ruckenstein and M. Pantzar, "Beyond the quantified self: thematic exploration of a dataistic paradigm," *New Media and Society*, vol. 19, no. 3, pp. 401–418, 2017.
- [8] R. L. Trivers, "The evolution of reciprocal altruism," *The Quarterly Review of Biology*, vol. 46, no. 1, pp. 35–57, 1971.
- [9] S. Quan and Y. Zeng, "Research on consumers' search behavior of food safety information—based on the survey of consumers in Beijing," *Agricultural Technology Economics*, no. 4, pp. 45–54, 2013.
- [10] R. An, "Effectiveness of subsidies in promoting healthy food purchases and consumption: a review of field experiments," *Public Health Nutrition*, vol. 16, no. 7, pp. 1215–1228, 2013.
- [11] A. Booth, A. Barnes, A. Laar et al., "Policy action within urban African food systems to promote healthy food consumption: a realist synthesis in Ghana and Kenya," *International Journal of Health Policy and Management*, 2020.
- [12] D. Li and Y. Zhang, "The effect of quantifying self and its influence mechanism on consumer participation behavior," *Management Science*, vol. 31, no. 3, pp. 112–124, 2018.
- [13] D. C. Aryani, *Impact of Microbial Variability on Food Safety and Quality*, Wageningen University, Wageningen, Netherlands, 2016.
- [14] T. Zhang and J. Lu, "Research on the influence of food labeling information on consumers' purchasing decisions—taking infant food as an example," *Forum on Statistics and Information*, no. 9, pp. 107–113, 2012.
- [15] T. C. Schroeder, G. T. Tonsor, J. M. E. Pennings, and J. Mintert, "Consumer food safety risk perceptions and attitudes: impacts on beef consumption across countries," *The B.E. Journal of Economic Analysis & Policy*, vol. 7, no. 1, p. 1848, 2007.
- [16] T. Meyvis and C. Janiszewski, "Consumers' beliefs about product benefits: the effect of obviously irrelevant product information," *Journal of Consumer Research*, vol. 28, no. 4, pp. 618–635, 2002.
- [17] E. K. Choe, N. B. Lee, and B. Lee, "Understanding quantified-selfers' practices in collecting and exploring personal data," in *Proceedings of the Sigchi Conference on Human Factors in Computing System*, pp. 1143–1152, ACM, Toronto, ON, Canada, April 2014.
- [18] S. Erdelez and K. Rioux, "Sharing information encountered for others on the web," *New Review of Information Behaviour Research*, vol. 1, no. 1, pp. 219–233, 2000.
- [19] K. Rioux, "Information acquiring-and-sharing," *Theories of Information Behavior*, vol. 24, no. 2, pp. 169–173, 2005.
- [20] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [21] S. Bratu, "Can social media influencers shape corporate brand reputation? Online followers' trust, value creation, and purchase intentions," *Review of Contemporary Philosophy*, vol. 18, pp. 157–163, 2018.
- [22] J. C. Hollowell, Z. Rowland, T. Kliestik, J. Kliestikova, and V. V. Dengov, "Customer loyalty in the sharing economy platforms: how digital personal reputation and feedback systems facilitate interaction and trust between strangers," *Journal of Self-Governance and Management Economics*, vol. 7, no. 1, pp. 13–18, 2019.
- [23] R.-A. Pop, Z. Săplăcan, D.-C. Dabija, and M.-A. Alt, "The impact of social media influencers on travel decisions: the role of trust in consumer decision journey," *Current Issues in Tourism*, pp. 1–21, 2021.
- [24] A. Drugău-Constantin, "Is consumer cognition reducible to neurophysiological functioning?" *Economics, Management, and Financial Markets*, vol. 14, no. 1, pp. 9–14, 2019.
- [25] C.-O. Mirică Dumitrescu, "The behavioral economics of decision making: explaining consumer choice in terms of neural events," *Economics, Management, and Financial Markets*, vol. 14, no. 1, pp. 16–20, 2019.
- [26] J. Jin, Y. Li, X. Zhong, and L. Zhai, "Why users contribute knowledge to online communities: an empirical study of an online social Q&A community," *Information and Management*, vol. 52, no. 7, pp. 840–849, 2015.
- [27] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [28] Y. Zhou, L. Huo, and X. Peng, "Food safety: consumer attitudes, purchasing intentions, and the influence of information—a survey and analysis of supermarket consumers in Nanjing," *Journal of Consumer Research*, vol. 42, p. 6, 2016.
- [29] Z. W. Fang, "The relationship between Internet addiction, coping, loneliness and online social support: a comparison between male and female college students," *Journal of Educational Psychology*, vol. 41, no. 4, pp. 773–797, 2010.
- [30] J. M. E. Pennings and B. Wansink, "Channel contract behavior: the role of risk attitudes, risk perceptions, and channel members' market structures," *Journal of Business*, vol. 77, no. 4, pp. 697–724, 2004.
- [31] D. L. Ortega, H. H. Wang, L. Wu, and N. J. Olynk, "Modeling heterogeneity in consumer preferences for select food safety attributes in China," *Food Policy*, vol. 36, no. 2, pp. 318–324, 2011.
- [32] C. B. Pocol, V. Marinescu, D.-C. Dabija, and A. Amuza, "Clustering Generation Z university students based on daily fruit and vegetable consumption: empirical research in an emerging market," *British Food Journal*, vol. 123, no. 8, pp. 2705–2727, 2021.
- [33] J. Blom-Hoffman, C. Kelleher, T. J. Power, and S. S. Leff, "Promoting healthy food consumption among young

- children: evaluation of a multi-component nutrition education program,” *Journal of School Psychology*, vol. 42, no. 1, pp. 45–60, 2004.
- [34] J. Kim and J.-E. R. Lee, “The Facebook paths to happiness: effects of the number of Facebook friends and self-presentation on subjective well-being,” *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 6, pp. 359–364, 2011.
- [35] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, “Bayesian personalized ranking based on multiple-layer neighborhoods,” *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [36] M. Kornelis, J. de Jonge, L. Frewer, and H. Dagevos, “Consumer selection of food-safety information sources,” *Risk Analysis*, vol. 27, no. 2, pp. 327–335, 2007.

Research Article

The Influence of Author Degree Centrality and L-Index on Scientific Performance of Physical Education and Training Papers in China Based on the Perspective of Social Network Analysis

Bin Zhang ¹, Jian Wu ², Qian Huang ¹, Yujiao Tan ¹, Lu Zhang ¹, Qian Zheng ¹,
Yu Zhang ¹, Miao He ¹, and Wei Wang ¹

¹*Xi'an Physical Education University, Xi'an 710065, China*

²*Shanghai University of Sport, Shanghai 200438, China*

Correspondence should be addressed to Jian Wu; wujian@sus.edu.cn

Received 4 May 2021; Revised 10 August 2021; Accepted 28 August 2021; Published 30 September 2021

Academic Editor: Fei Xiong

Copyright © 2021 Bin 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.

With the increasing frequency of scientific cooperation, research on the impact of scientific cooperation on scholars' scientific performance has become a hot topic in academic research. This paper mainly uses quantitative research methods such as bibliometrics and social network analysis to analyze the correlation between degree centrality, L-index, and scientific performance of the Chinese author of Physical Education and Training Science and draws the following conclusions: (1) degree centrality is positively correlated with the number of papers, average citations per paper, and H-index and (2) L-index is positively correlated with the number of papers, average citations per paper, and H-index. Unlike previous studies on scientific collaboration networks and scholars' performance that focused on traditional network characteristics indicators, this study introduces a new network characteristics indicator, L-index, through which the degree of collaboration between authors and other important authors in the research field can be better assessed, which also provides a new direction for research related to research collaboration networks and scholars' performance; at the same time, it also provides a theoretical basis for subsequent research on the utility of scientific collaboration, a reference for the assessment of the research performance of scholars in other disciplines, and a theoretical reference for more scientific and comprehensive prediction of scholars' research performance in the future.

1. Introduction

With the development of society and the progress of science, scientific research becomes more and more complex. Price put forward the concept of "big science" for the first time, pointing out that modern scientific research is characterized by high investment intensity, multidisciplinary crossing, expensive experimental equipment, and complex research objectives [1]. Scientific cooperation is one of the effective ways to solve complex scientific research problems. Scientific cooperation refers to the knowledge production mode in which many people work together to complete the same research topic in the process of scientific research, and it is extremely important in the historical process of scientific

development. Especially at a time when science is increasingly comprehensive and complex, scientific collaboration not only allows researchers to share their knowledge in different fields, their experience and resources, and complex and expensive research equipment but also stimulates innovative thinking and increases the efficiency of research. Therefore, scientific cooperation has become more and more frequent, and scientific research has gradually shifted from individual research to team research. Scientific research cooperation has gradually become an unstoppable development trend.

As scientific cooperation has become a common phenomenon in scientific research, research on the effectiveness of scientific cooperation has gradually become an important

topic in the field of scientific cooperation. The effect of scientific cooperation on the scientific performance of scholars is an important aspect of the research on the effectiveness of scientific cooperation. As we all know, scientific performance is an important basis for the evaluation of scholars' academic levels. The higher the scientific performance of scholars, the richer their academic achievements and the greater the influence of their academic achievements. The evaluation of scholars' scientific performance not only can establish the correct value orientation of scientific research, promoting the sustainable and stable development of scientific research, but also can provide an objective basis for teacher recruitment, project application, government funding, and so on. By studying how scientific cooperation affects scholars' scientific performance, we can find out the factors affecting scholars' scientific performance and their internal correlation, so as to optimize the way of scientific cooperation and improve scholars' scientific performance.

As a discipline that encompasses both the social and natural science dimensions, the study of Physical Education and Training cannot be furthered and advanced without the reference significance brought by the author's scientific performance. This paper selects 10,386 papers in core journals of Physical Education and Training in the CSSCI database, and Python was used to visualize the correlation between degree centrality, L-index, and scientific performance. Stata was used to conduct linear regression on degree centrality, L-index, and scientific performance to discuss their correlation. Unlike previous studies on research collaboration networks and scholars' performance that focused on traditional network characteristics indicators, this study introduces a new network characteristics indicator, L-index, through which the degree of collaboration between authors and other important authors in the research field can be better assessed, which also provides a new direction for research related to research collaboration networks and scholars' performance; at the same time, it also provides a theoretical basis for subsequent research on the utility of scientific collaboration, a reference for the assessment of the research performance of scholars in other disciplines, and a theoretical reference for more scientific and comprehensive prediction of scholars' research performance in the future.

In addition to the introduction of the first part, this paper also includes the following sections. The second section introduces the theoretical basis of this research. The third section introduces the data source, the choice of variables, and the research method. The fourth section is the empirical analysis. Finally, the fifth section summarizes the conclusions and contributions of this study and points out the shortcomings in the study.

2. Theoretical Basis

2.1. The Scientific Cooperation. At present, academic circles have different definitions of scientific cooperation. Heffner believes that scientific research cooperation is a strong form of interaction between researchers in the process of scientific activities. They form an ideal model of scientific research

cooperation through ideological and intellectual exchanges [2]. Ziman and Schmitt believe that scientific cooperation is generated after scientific development has reached a certain stage and gradually stabilized, during which scientific cooperation plays an extremely important role in improving the output of scientific knowledge [3]. Katz and Martin believe that scientific cooperation is the work of researchers together for the common purpose of creating new knowledge [4]. Zhao and Wen believe that scientific cooperation is a kind of scientific activity in which two or more researchers or organizations cooperate together to maximize scientific research output in order to complete a common research task, and its essence is resource sharing between collaborators [5]. Based on the concept of scientific cooperation put forward by different scholars, the author believes that scientific cooperation means that researchers focus on a specific scientific problem through resource sharing and cooperation and finally jointly complete scientific tasks with clear objectives and jointly publish scientific results. Among them, the collaborative publication of research results is the ultimate expression of scientific collaboration.

2.2. Scientific Performance. Scientific performance refers to that in the process of scientific activities, the research subjects, driven by subjective motivation, objective needs, and possible original basis for cooperation, carry out various forms of cooperation and produce various direct or indirect benefits, such as the direct benefits of papers, patents monographs, and so on. Indirect benefits include the improvement of scholars' knowledge, skills, reputation, and influence. There are qualitative evaluation and quantitative evaluation to evaluate the research performance of scholars. The qualitative evaluation mainly refers to peer review, which is also the traditional way to evaluate the scientific performance of scholars. But peer review has its drawbacks. First, it is less efficient because it requires experts to score each scientific achievement and give a written explanation. In addition, peer review is highly subjective, and the review results are easily affected by subjective factors such as experts' personal interests, majors, and emotions, which make it difficult to guarantee fairness and justice. Quantitative evaluation has been proposed and widely applied with the continuous growth of scientific achievements and the continuous development of scientometrics and bibliometrics. Quantitative evaluation evaluates the research performance of scholars through clear quantitative characteristics, which improves the efficiency and objectivity of the evaluation of the scientific performance of scholars. At present, the research on quantitative evaluation of scholars' scientific performance has been in-depth, and its evaluation indicators are also diverse, mainly including: number of papers, average citations per paper, impact factors, number of collaborators, H-index, and so on. [6]. Integrating existing research, as a paper is the result of knowledge formed in the course of scientific activities, it is the most direct form of the output of scientific research. Its quantity and quality can, to a certain extent, relatively objectively reflect the contribution of scholars to the existing body of knowledge, and therefore,

the main reference for evaluating the performance of scholars is their published papers [7–9].

In this paper, the number of papers, average citations per paper, and H-index are selected as the evaluation indicators of scholars' scientific performance. The number of papers is the total number of papers published by each author. Although the number of papers does not fully represent the academic level and influence of a scholar, the production of high-quality papers is usually based on a certain number of publications. For example, Nobel laureates publish 13.1 papers per year on average when they are about 20 years old, while ordinary scholars publish 3.5 papers per year on average [10]. Average citations per paper refer to the ratio of the total number of citations per author to the number of papers. Average citations per paper are based on its quality. The more the average citations per paper receive, the higher the degree of "recognition" and the higher the quality of the paper is. Papers published by some authoritative scholars are cited as much as 200 times per year, while those published by ordinary scholars are less than 10 times [11]. H-index is a research performance evaluation index proposed by Hirsch that combines the number of papers published by scholars with average citations per paper, and H-index refers to the number of N papers published by a scholar, with H papers being cited at least H times each [12]. It not only quantifies research output but also combines the ability of researchers to publish with the impact of citation. Therefore, H-index is the most widely used indicator among the quantitative evaluation indicators to evaluate the scientific performance of scholars. Subsequent scholars have improved H-index on the basis of H-index and proposed new indicators for scientific performance evaluation, such as G-index (2006) [13], R-index and AR-index (2007) [14], M-index (2008) [15], EM-index (2017) [16], and so on. However, the concept of H-index is simple, and it is easy to calculate and has the characteristics of robustness and comprehensiveness, which makes it widely used in the evaluation of scientific performance.

2.3. Scientific Collaboration Network. The "network" is a collection of nodes and all their relationships with each other. The "scientific cooperation network" is a collection of researchers as nodes and the connections of researchers through scientific cooperation. In the scientific cooperation network, degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, PageRank, and other network characteristic indicators are commonly used to describe the position, function, or influence of nodes in the network. Studies have been conducted to examine the relationship between scientific collaboration networks and scholar performance in terms of these network characteristics indicators. Abbasi et al. used G-index as the dependent variable to measure the academic influence of scholars. The network characteristics such as degree centrality, closeness centrality, betweenness centrality, eigencentrality, and average relationship strength are taken as independent variables. They studied their relation and found that the degree centrality, eigencentrality, and average relation strength are

positively correlated with their academic influence (G-index). Closeness centrality and betweenness centrality are negatively correlated with their academic influence (G-index) [17]. From the perspective of social capital, Li et al. studied the relationship between degree centrality, closeness centrality, betweenness centrality of scientific cooperation network, and the average number of citations of scholars' papers. They found that degree centrality and closeness centrality had little correlation with the average number of citations, while betweenness centrality was positively correlated with the average number of citations [18]. Gonzalez-Brambila et al. found that having more direct contact with other scholars, being in the structural hole position of the scientific cooperation network, cooperating with scholars in other disciplines, and being in the center of the scientific cooperation network will increase the number of papers published by scholars [19]. Abbasi et al. proposed two new indicators, PDI and PTDI, on the basis of traditional social network indicators; studied the relationship between PDI and PTDI and scholars' research influence based on citation relationship; and found that PDI and PTDI were positively correlated with research influence based on citation relationship [20]. Damien et al. found that scholars who are in the bridge position of the scientific cooperation network will improve their scientific performance [21].

While these network characteristics indicators can describe to some extent the intrinsic relevance of collaboration in a research network, they do not provide a good description of the extent to which nodes in the network are connected to other important nodes. Among these commonly used network characteristic indicators, degree centrality is the most basic one. Degree centrality is a commonly used concept in social network analysis (SNA), also known as degree, and it is also the most intuitive centrality indicator in the study of scientific cooperation network. The higher the degree of a node is, the more nodes it is associated with, and the more important the node is in the network. In a cooperative network with g nodes, the degree centrality of node i is the total number of direct contacts between node i and other $g - 1$ nodes. If an author has a degree of 10, he is associated with 10 authors in his scientific collaboration network. While degree centrality is generally used to describe the extent to which a node is connected to other nodes in a social network, it only measures the number of nodes connected to a node, not the importance of the nodes connected to a node. Therefore, Korn et al. consider a new network characteristic metric based on degree centrality to measure the degree to which a node in the network is connected to other important nodes, namely L-index (lobby index), which refers that there are at least L nodes whose degrees are at least L out of N nodes connected to a node in the network [22]. L-index provides a new direction for the research on scientific cooperation network and scholar performance as well.

Based on this perspective, this paper explores the correlation between the degree and L-index of authorship of Physical Education and Training papers and scientific performance (number of papers, average citations per paper, and H-index), lays a theoretical foundation for subsequent

research related to the utility of research collaboration, provides a reference for the assessment of research performance of scholars in other disciplines, and provides a theoretical reference for more scientific and comprehensive prediction of scholars' research performance in the future.

2.4. Social Network Analysis. In the western society of the 1920s, social network analysis was gradually applied to the fields of sociology, anthropology, psychology, and so on. Simmel first used the concept of "network" in 1922. Brown first put forward the concept of "social network" in 1940. Barnes first transformed the social network into a systematic study by analyzing the social structure of a fishing village in Norway [23]. After decades of research and exploration, social network analysis has become a unique research method, which provided a new perspective for the study of social structure. Freeman summarized the four characteristics of social network research: the analysis of the specific structure formed between the subjects of the social network, applying statistical principles and computer technology as technical support, based on data in the network, and graphic language to express social networks [24]. Wellman, an analyst of the social network, pointed out that social network was a huge network composed of social relations between groups, and social network analysis explored the deep structure under the network, specifically the connection pattern hidden under the surface of the complex social network [25]. After 2000, social network analysis has been gradually recognized and widely used in China. Bao et al. first analyzed egocentric network using the method of social network analysis [26]. Hu and Deng analyzed interpersonal relationships using the structural hole theory [27]. Liu et al. analyzed the role of small-world theory in group analysis [28]. Social network analysis is also widely used in research related to scientific collaboration. Xu and Zhu applied social network analysis to citation analysis and analyzed the influence and clique of scholars through density, centrality, and other indicators [29]. Qiu and Wang analyzed the author's cooperative relationship using social networks and found out the potential cooperative groups [30]. Xing et al. visualized the relationship between keywords by constructing the overall keyword network and combining relevant indicators of the social network, so as to more intuitively understand the research status of relevant fields [31]. Generally speaking, the social network reflects the relationship between various elements. The core of social network analysis is to analyze the "relationship" of various elements and present the relationship between individuals from "micro" to "macro" one by one. The analysis method is based on mathematical modeling and network graph, and the main research content includes the importance of each participant in the network and its role in the whole network. In addition, social network analysis and bibliometrics can also be well complementary and integrated, which are widely used in the related research of scientific cooperation network. Many scholars analyze research collaborations through bibliometric methods, but in-depth research

analysis and the presentation of knowledge maps involve social network analysis methods, such as analyzing and visualizing the relationships between nodes in terms of network characteristics such as centrality, cohesive sub-groups, and density, in order to identify core authors, research methods, and disciplinary hotspots in subject areas.

The scientific collaboration network consisting of the authors of our Physical Education and Training papers studied in this paper is a typical social network. Among them, each node refers to the researchers participating in the scientific cooperation, and the edge between nodes refers to the scientific cooperation between the researchers, that is, through the scientific cooperation to jointly publish the paper.

3. Methods

3.1. The Data Source. In 1931, Bradford, a British bibliographer, first revealed the law of literature concentration and dispersion, that is, in a certain discipline, about one-third of the literature was published in 3.2% of the periodicals. In 1967, UNESCO researched the distribution of the literature and found that 75% of the literature appeared in only 10% of the journals. In 1971, Garfield, an American intelligence scientist and scientometrician and the founder of SCI, counted the distribution of references in journals and found that 24% of the citations appeared in 1.25% of the journals. All these indicate the existence of "core effect," that is, the existence of "core journals." The input documents of core journals are authoritative and representative scientific research achievements in various disciplines after strict selection.

The data used in this paper are from core journals of sports, and the data are searched through the CSSCI database. Input, respectively, in the "journal" column the name of 16 kinds of sports core periodicals: "Journal of Beijing Sport University," "Journal of Shanghai University of Sport," "Journal of Capital University of Physical Education and Sports," "Journal of Shandong Sport University," "Journal of Chengdu Sport University," "Journal of Tianjin University of Sport," "Journal of Xi'an Institute of Physical Education," "Journal of Wuhan Institute of Physical Education," "Journal of Shenyang Sport University," "Journal of Guangzhou Sport University," "Journal of Nanjing Sport Institute (Social Science)," "China Sport Science and Technology," "Sports Culture Guide," "Journal of Physical Education," "Sports & Science," and "China Sport Science." The subject type was set to "Physical Education," and the secondary subject type was set to "Physical Education and Training," with no restriction on publication time. Considering that the same Chinese characters will appear in the author's name, it is ambiguous to use only the name for calculation in the collected data set. In order to eliminate such ambiguity, this paper matches the name and work affiliation of the paper author at the same time, that is, two scholars with the same name and different affiliations are identified as two different scholars. Finally, the extracts saved all the literature names, authors, work affiliations of authors,

references, publication time, citation times, and other information and finally got 10,386 data.

3.2. Variable Selection and Measurement

3.2.1. The Dependent Variable. The dependent variable of this study is the scientific performance of all authors in the sample, which mainly includes three aspects: the number of papers, average citations per paper, and H-index.

(1) *The Number of Papers.* The number of papers refers to the total number of papers published by each author in his academic career. A certain number of papers is the basis for creating high-level scientific achievements.

(2) *Average Citations per Paper.* The average citations per paper are the ratio of the total number of citations per author to the number of papers. The higher an author's average citations per paper is, the higher the average quality of that author's research output and the more useful his or her results are to other scholars. The formula for calculating average citations per paper is as follows:

$$\text{Average citations per paper}_i = \frac{\text{the total number of citations}_i}{\text{the number of papers}_i}. \quad (1)$$

(3) *H-Index.* H-index refers to the number of N papers published by a scholar, with H papers being cited at least H times each [12]. H-index can accurately reflect a person's academic achievements. H-index not only quantifies the scientific output of scholars but also takes into account the ability of researchers to publish papers and the influence of citation. The higher an author's H-index is, the more influential and academically valuable articles he or she has published.

3.2.2. The Independent Variables. The independent variables for this study are the degree centrality and L-index of all authors in the sample.

(1) *Degree Centrality.* Degree centrality refers to the total number of direct connections between node i and other $g - 1$ nodes in a network with g nodes. The higher the degree centrality of a node is, the more nodes this node has contact with, and the more important this node is in the network. The calculation formula of degree centrality is as follows:

$$\text{Degree centrality}_i = \sum_{j=1}^g x_{ij} \quad (i \neq j). \quad (2)$$

(2) *L-Index.* L-index means the degree to which L of the N nodes in the network that are connected to a node is at least L [22]. L-index reflects the extent to which a node in the network is connected to other important nodes. The higher L-index of an author, the more authors with high centrality

he or she is associated with and the higher the quality of his or her collaborators.

3.3. The Research Methods

3.3.1. Establish the Scientific Cooperation Network. The scientific collaboration network in this study was constructed on the basis of our Physical Education and Training papers, which is a typical social network. In the scientific collaboration network, each node represents an author, and the edges between nodes indicate that two authors have collaborated on previous papers. The scientific collaboration network in this study is constructed by Python.

3.3.2. Measurement of the Independent and Dependent Variables. After the research collaboration network was constructed, the degree centrality and L-index of all authors in the sample were calculated using Python. Then using bibliometric methods, the research performance indicators such as the number of papers, average citations per paper, and H-index of all authors in the sample were calculated. The distribution of all independent and dependent variables and the correlation between the independent and dependent variables were then visualized and analyzed through Python.

3.3.3. Correlation Analysis of Independent Variables and Dependent Variables. The correlation between the independent and dependent variables was analyzed using Stata to explore in detail the effects of degree centrality and L-index on scientific performance indicators such as the number of papers, average citations per paper, and H-index.

4. Analysis and Results

4.1. Descriptive Statistics and the Distribution of Independent and Dependent Variables

4.1.1. Descriptive Statistics. By processing and analyzing all the 10,386 literature data of 16 core sports journals in China with the subject type of "Sports" and the secondary subject type of "Physical Education and Training," we obtained a total of 12,875 authors, each of whom corresponds to a node in the research collaboration network. Table 1 shows the mean, median, standard deviation, and minimum and maximum values of each variable in the research collaboration network.

As we can be seen from Table 1, the mean values of the number of papers, H-index, degree centrality, and L-index are 1.541, 1.395, 2.975, and 2.468, respectively, and their median values are 1, 1, 2, and 2, which are very close to their minimum values 1, 0, 1, and 1, respectively, and even the median of the number of papers is consistent with the minimum values. This shows that in the research field of Physical Education and Training in China, most of the authors published fewer papers, and the levels of H-index, degree centrality, and L-index are low. The average citation per paper is 23.848, and the median is 14, which is also far from the maximum citation amount of 1,268, indicating that

TABLE 1: The descriptive statistics of independent and dependent variables.

Variable	N	Mean	Median	Std. dev.	Min	Max
The number of papers	12,875	1.541	1	1.778	1	66
Average citation per paper	12,875	23.848	14	37.333	0	1,268
H-index	12,875	1.395	1	1.212	0	28
Degree	12,875	2.975	2	2.93	1	83
L-index	12,875	2.468	2	1.489	1	14
Career length	12,875	1.927	1	2.546	1	21

in the research field of Physical Education and Training in China, the citation number of papers published by most authors is relatively low.

4.1.2. Distribution of the Independent and Dependent Variables

(1) Distribution of the Dependent Variables.

(1) Distribution of the number of papers: Figure 1 shows the distribution of the number of papers published by the authors of Physical Education and Training in China. As can be seen from Figure 1, the overall distribution of the number of papers published by authors of Physical Education and Training in China is that the number decreases rapidly within a certain range, then tends to be stable, with the minimum value of 1 and the maximum value of 66. The largest number of authors who published one paper is 9,768, accounting for about 77% of the total number of authors. The number of authors who published less than 5 papers is 12,447, accounting for more than 96% of the total number of authors. The authors who published more than 10 papers are fewer than 100. And considering that the data were collected without limiting the publication date of the paper. The earliest recorded paper was published in 2000. In the 20 years since 2000, the total number of published papers in this field is only 10,386, and the average annual number of published papers is only about 500. At the author level, the vast majority of authors have published a small number of papers, and only a small number of authors have a high number of publications, all of which indicate that the overall level of scientific research in the subject area of Physical Education and Training in China is still relatively limited.

(2) The distribution of average citations per paper: Figure 2 shows the distribution of average citations per paper of the authors of Physical Education and Training in China. From Figure 2, we can find that the distribution of Physical Education and Training authors in China has more authors in the lower citation distribution range, and the distribution of authors gradually decreases as the number of citations increases. Although the number of authors with high citations is relatively limited, there are also authors with high citations per article, which can also indicate that there are authors in the field of Physical Education and Training in China who have

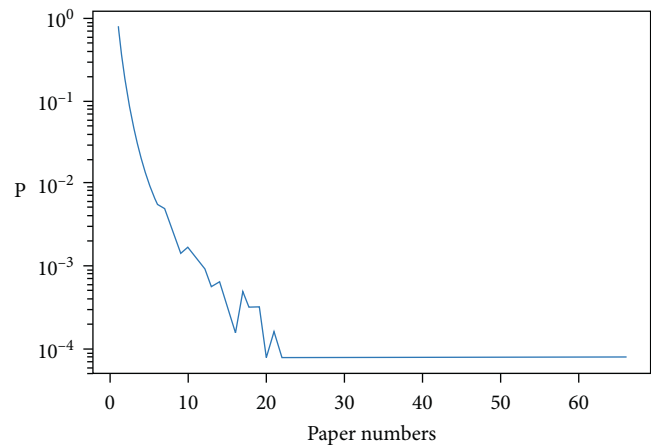


FIGURE 1: Distribution of the number of papers.

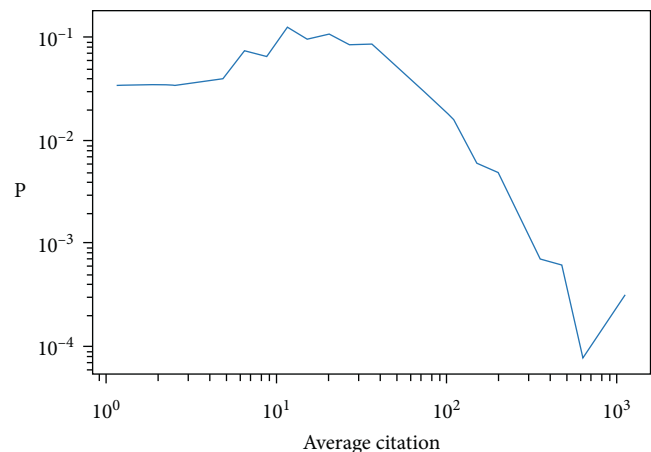


FIGURE 2: Distribution of average citations per paper.

published articles with high citations. That is, there are high impact and widely recognized scientific research results in the field of Physical Education and Training in China. However, from an overall perspective, the average number of citations per article of most authors in the field of Physical Education and Training research in China is still at a low level, and the overall level of scientific research in the discipline is still relatively limited.

(3) The distribution of H-index: Figure 3 shows the H-index distribution of the authors of Physical Education and Training in China. From Figure 3, we can find that H-index of Chinese Physical Education

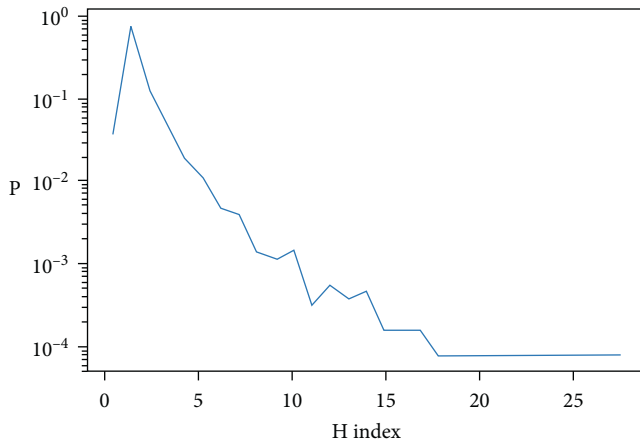


FIGURE 3: Distribution of H-index.

and Training paper authors is also on a downward trend, with a minimum value of 0. The trend of overall distribution is that the number of authors reaches the peak when H-index is 1. Then, with the increase of H-index, the number of authors decreases. When H-index is 10, the number of authors will reach a relatively low level. The number of authors with H-index 1 was the largest, with a total of 9,494, accounting for more than 73% of the total number of authors. The number of authors with a high H-index is very small, and the number of authors with an H-index higher than 10 is less than 1.5% of the total. It indicates that most authors have fewer published papers or the number of citations is not high, and only a few authors have a high number of published papers with a high number of citations. H-index is the main evaluation indicator to evaluate the academic influence, so it can explain that most of the authors in the field of Physical Education and Training in China have moderate academic influence, and only a few authors have relatively higher academic influence. The scientific research influence in the field of Physical Education and Training in China needs to be improved as a whole.

(2) *The Distribution of the Independent Variable.*

(1) The distribution of degree centrality: Figure 4 shows the distribution of degree centrality of the authors of Chinese Physical Education and Training. From Figure 4, we can find that the degree centrality distribution of the authors of Physical Education and Training papers in China is still on a decreasing trend, and from our data, we find that the minimum value of degree centrality is 1 and the maximum value is 83. With the increase of degree centrality, the number of authors decreases rapidly in the range of degree centrality from 1 to 10. Since our distribution is plotted in double logarithmic coordinates, the distribution of authors slowly decreases and plateaus as the degree centrality increases when the degree centrality is greater than 10. The authors with degree

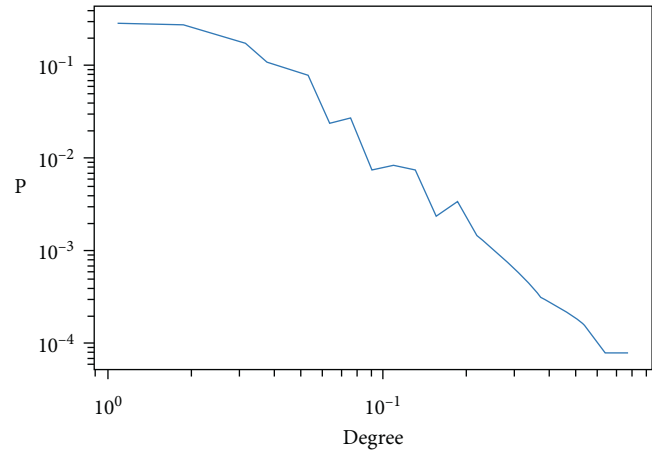


FIGURE 4: Distribution of degree centrality.

centrality 1, 2, and 3 have the largest number, with a total of 9,398 authors, accounting for more than 72% of all authors. The number of authors with degree centrality in the range of 1–10 is 12,610, accounting for more than 97% of the total number of authors, indicating that the degree centrality of most of the authors of Physical Education and Training in China is less than 10, that is, the number of collaborators of most of the authors is less than 10. It can also be seen that most of the authors in the field of Physical Education and Training in China have relatively fixed collaborators, and they generally have fixed research teams, especially small teams with less than 5 members. The number of authors whose degree centrality are more than 10 is few. Even the number of authors whose degree centrality is more than 20 is less than 70. It shows that a few authors may not have a fixed team, but they will take the initiative to seek cooperation with more of the authors, or the authors in the discipline of Physical Education and Training may have the higher reputation, and there are a number of authors or research groups that will actively seek to collaborate with them.

(2) The distribution of L-index: Figure 5 shows the distribution of L-index of the authors of Physical Education and Training in China. From Figure 5, we can find that the minimum value of L-index of authors of Physical Education and Training in China is 1, and the maximum value is 14. The overall distribution trend is that with the increase of L-index, the number of authors decreases. When L-index is 2, 1, and 3, the number of authors is the highest, with a total of 10,143, accounting for more than 78% of all authors. The number of authors whose L-index was less than 6 is 12,557, accounting for more than 90% of the total number of authors. It indicates that L-index of most of the authors of Physical Education and Training in China is less than 6, that is, most of the authors do not have at least 6 cooperators whose degree centrality is greater than 6. They have fewer collaborators or their collaborators had lower degree

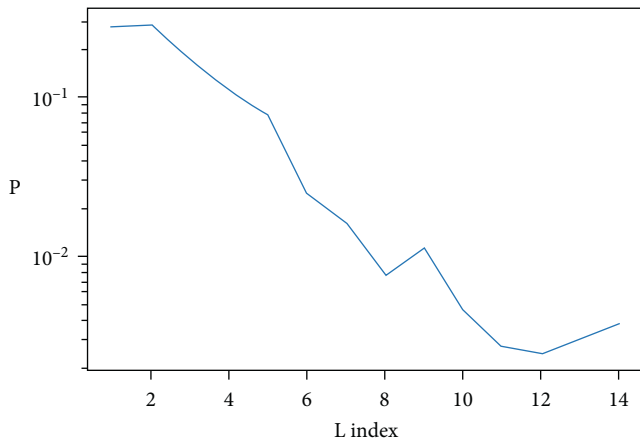


FIGURE 5: Distribution of L-index.

of centrality, and only a small percentage of the authors collaborate with authors who have the higher degree of centrality in the collaborative network. This shows that the authors in the field of physical education and training in China are still confined to the original small research team and can also reflect the slow development and limited research level in the field of physical education and training research in China.

4.2. Regression Analysis. The original data were initially tested for endogeneity. The explanatory variables were not correlated with the nuisance terms, and the sample size was large enough to be consistent with the “large-sample theory.” This paper uses large-sample OLS regression to analyze the original data, and to avoid bias in statistical inference due to the heteroskedasticity, robust standard errors are used instead of ordinary standard errors in the large-sample OLS regression. After the regression analysis, the reasonableness of the use of large-sample OLS regression will be further tested through robustness tests.

4.2.1. Regression Analysis of the Number of Papers and Degree Centrality. Figure 6 shows the scatter diagram and regression curve of the correlation between the number of papers and degree centrality by Python. Each point represents an author, and the red curve represents the correlation regression curve between the number of papers and degree centrality. As can be seen from Figure 6, with the increase of degree centrality, the number of papers also increases significantly. Most of the authors are distributed in the lower-left corner of the figure, that is, most of the authors have a low degree of centrality and published a small number of papers. Then Stata is used to make unary linear regression for the number of papers and degree centrality. The results are shown in Table 2.

As can be seen from Table 2, p value is less than 0.01; the regression coefficient is $0.347 > 0$, which indicates that the number of papers is positively correlated with degree centrality; and this relationship is significant at the confidence level of 0.01. It also indicates that with the increase of

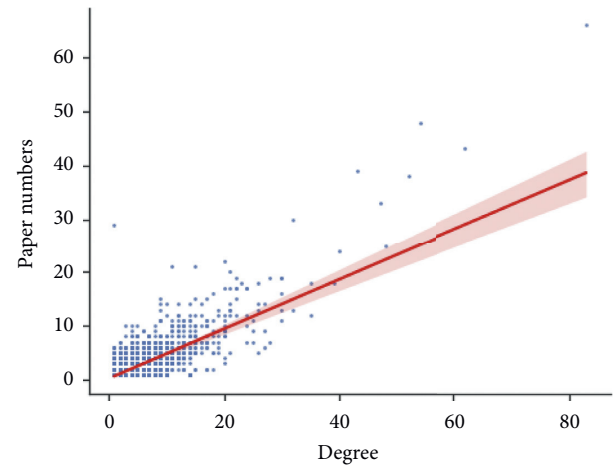


FIGURE 6: Scatter diagram and regression curve of correlation between the number of papers and degree centrality.

authors’ degree centrality, the number of published papers will be higher. In other words, if an author widely seeks to cooperate with more authors, the number of published papers will be increased.

4.2.2. Regression Analysis of Average Citations per Paper and Degree Centrality. The scatter diagram and regression curve of correlation between average citations per paper and degree centrality are shown in Figure 7. Each point represents an author, and the red curve represents the correlation regression curve between average citations per paper and degree centrality. Then Stata was used to make unary linear regression for average citations per article and degree centrality. The results are shown in Table 3:

As can be seen from Table 3, p -value is less than 0.01; the regression coefficient is $0.484 > 0$, which indicates that average citations per paper are positively correlated with degree centrality; and this relationship is significant at the confidence level of 0.01. It also indicates that with the increase of authors’ degree centrality, average citations per paper will also increase, that is, if the authors seek to cooperate with more authors extensively, the average citation of his or her papers will increase.

And Figure 7 shows that the authors with higher average citations per paper have lower degree centrality, which may be due to the number of samples with low degree centrality is relatively huge, and there are some authors without high degree centrality who publish the paper with higher influence. The number of samples with high degree centrality is relatively small, and their average citations per paper have a certain gap between the authors who publish papers with high influence and whose degree centrality is not high but still higher than the average level, which will not affect the positive correlation between average citations per paper and degree centrality.

4.2.3. Regression Analysis of H-Index and Degree Centrality. The scatter diagram and regression curve of the correlation between H-index and degree centrality are made by Python,

TABLE 2: Unary linear regression results of the number of papers and degree centrality.

The number of papers	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
Degree	0.347	0.038	9.17	≤0.001	0.273	0.421	***
Main journal	-0.005	0.002	-2.23	0.026	-0.009	-0.001	**
Start year	-0.015	0.002	-7.59	≤0.001	-0.019	-0.011	***
Career length	0.244	0.017	14.08	≤0.001	0.21	0.278	***
Constant	30.71	3.971	7.73	≤0.001	22.927	38.493	***
Mean dependent var		1.541		SD dependent var		1.778	
<i>R</i> -squared		0.675		Number of obs		12,875	
<i>F</i> -test		441.958		Prob > <i>F</i>		≤0.001	
Akaike crit. (AIC)		36,873.095		Bayesian crit. (BIC)		36,910.410	

****P* < 0.01, ***P* < 0.05, and **P* < 0.1.

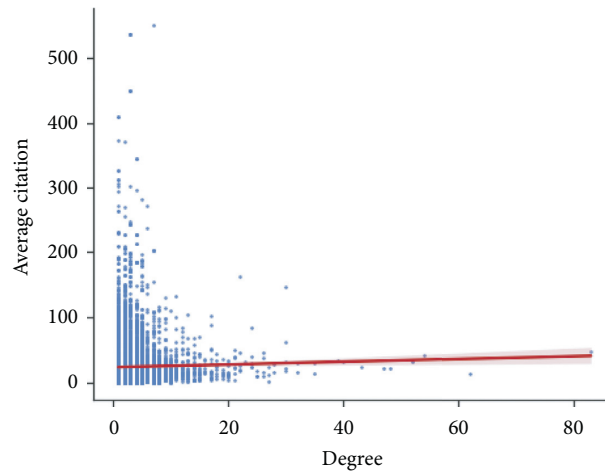


FIGURE 7: Scatter diagram and regression curve of correlation between average citations per paper and degree centrality.

TABLE 3: Unary linear regression results of average citations per paper and degree centrality.

Average citations per paper	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
Degree	0.484	0.106	4.55	≤0.001	0.276	0.693	***
Main journal	-0.283	0.093	-3.05	0.002	-0.465	-0.101	***
Start year	-1.621	0.046	-35.19	≤0.001	-1.711	-1.531	***
Career length	-0.754	0.111	-6.78	≤0.001	-0.972	-0.536	***
Constant	3,282.059	92.749	35.39	≤0.001	3,100.257	3,463.86	***
Mean dependent var		23.848		SD dependent var		37.333	
<i>R</i> -squared		0.056		Number of obs		12,875	
<i>F</i> -test		358.384		Prob > <i>F</i>		≤0.001	
Akaike crit. (AIC)		129,014.918		Bayesian crit. (BIC)		129,052.233	

****P* < 0.01, ***P* < 0.05, and **P* < 0.1.

as shown in Figure 8. Each point represents an author, and the red curve represents the correlation regression curve between H-index and degree centrality. As can be seen from Figure 8, with the increase of degree centrality, H-index also increases significantly. Most of the authors are distributed in the lower-left corner of the figure, that is, most of the authors have low degree centrality and low H-index. Then Stata was used to make unary linear regression of H-index and degree centrality. The results are shown in Table 4:

As can be seen from Table 4, *p*-value is less than 0.01; the regression coefficient is $0.188 > 0$, which indicates that H-index is positively correlated with degree centrality; and

this relationship is significant at the confidence level of 0.01. It also indicates that with the increase of degree centrality, the quantity and quality of published articles will also be higher. In other word, it also indicates that if an author widely seeks to collaborate with more authors, the quantity and quality of his published articles will be significantly increased, that is, his or her academic influence will be significantly increased.

4.2.4. Regression Analysis of the Number of Papers and L-Index. The scatter diagram and regression curve of the correlation between the number of papers and L-index are

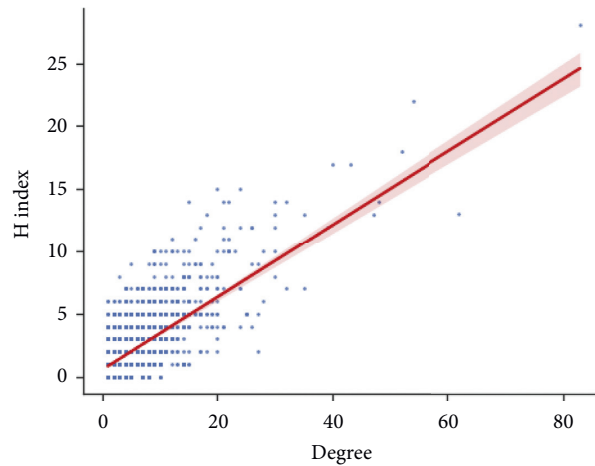


FIGURE 8: Scatter diagram and regression curve of correlation between H-index and degree centrality.

TABLE 4: Unary linear regression results of H-index and degree centrality.

H-index	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
Degree	0.188	0.011	16.47	≤0.001	0.166	0.211	***
Main journal	-0.01	0.002	-6.16	≤0.001	-0.013	-0.007	***
Start year	-0.021	0.001	-18.78	≤0.001	-0.023	-0.019	***
Career length	0.213	0.008	25.34	≤0.001	0.196	0.229	***
Constant	42.232	2.215	19.07	≤0.001	37.89	46.574	***
Mean dependent var		1.395		SD dependent var		1.212	
<i>R</i> -squared		0.659		Number of obs		12,875	
<i>F</i> -test		612.121		Prob > <i>F</i>		≤0.001	
Akaike crit. (AIC)		27,629.154		Bayesian crit. (BIC)		27,666.470	

*** $P < 0.01$, ** $P < 0.05$, and * $P < 0.1$.

made by Python, as shown in Figure 9. Each point represents an author, and the red curve represents the regression curve of correlation between the number of papers and L-index. As can be seen from Figure 9, with the increase of L-index, the number of papers also increases significantly. Most of the authors are distributed in the range where L-index is less than 10. When L-index is greater than 10, the distribution of the number of papers fluctuates due to the decrease of the sample size, but it does not affect the overall trend. Then Stata was used to make unary linear regression of the number of papers and L-index. The results are shown in Table 5.

As can be seen from Table 5, *p* value is less than 0.01; regression coefficient is 0.153 > 0, which indicates that there is a positive correlation between the number of papers and L-index; and this relationship is significant at the confidence level of 0.01. It also indicates that with the increase of the author's L-index, the number of published articles will be higher. In other words, if more authors seek to cooperate with more important and influential authors in the research collaboration network, that is, the more authors seek to cooperate with a higher L-index, the number of published papers will increase.

4.2.5. Regression Analysis of Average Citations per Paper and L-Index. The scatter diagram and regression curve of correlation between average citations per paper and L-index are

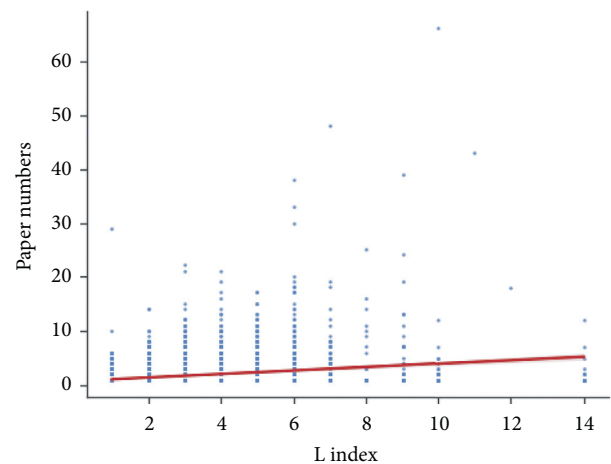


FIGURE 9: Scatter diagram and regression curve of correlation between the number of papers and L-index.

shown in Figure 10. Each point represents an author, and the red curve represents the correlation regression curve between average citations per paper and L-index. As can be seen from Figure 10, with the increase of L-index, the maximum of the average citations per paper decreases. According to the correlation regression curve between average citations per paper and L-index, average citations per paper and L-index do not show an obvious correlation. Then

TABLE 5: Unary linear regression results of the number of papers and L-index.

The number of papers	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
L-index	0.153	0.019	7.96	≤0.001	0.115	0.191	***
Main journal	−0.01	0.003	−3.62	≤0.001	−0.015	−0.004	***
Start year	−0.005	0.001	−3.98	≤0.001	−0.008	−0.003	***
Career length	0.442	0.02	22.26	≤0.001	0.403	0.481	***
Constant	11.229	2.71	4.14	≤0.001	5.918	16.54	***
Mean dependent var		1.541		SD dependent var		1.778	
<i>R</i> -squared		0.461		Number of obs		12,875	
<i>F</i> -test		152.352		Prob > <i>F</i>		≤0.001	
Akaike crit. (AIC)		43,397.556		Bayesian crit. (BIC)		43,434.871	

*** $P < 0.01$, ** $P < 0.05$, and * $P < 0.1$.

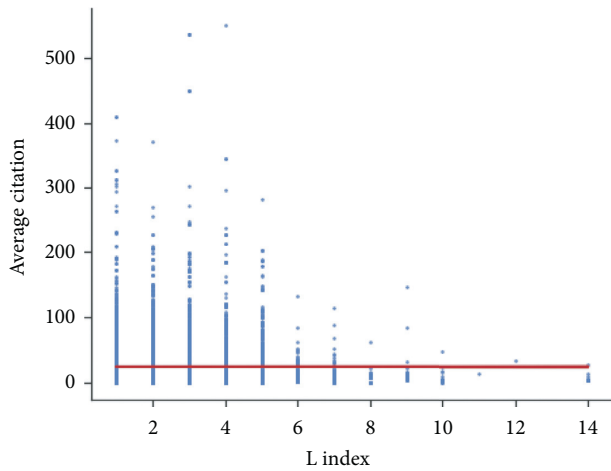


FIGURE 10: Scatter diagram and regression curve of correlation between average citations per paper and L-index.

Stata was used to make unary linear regression between average citations per paper and L-index. The results are shown in Table 6.

As can be seen from Table 6, p -value is less than 0.01; the regression coefficient is $0.926 > 0$, which indicates that there is a positive correlation between average citations per paper and L-index; and this relationship is significant at the confidence level of 0.01. It also indicates that with the increase of the author's L-index, the average citations per paper will be higher. In other words, if more authors seek to cooperate with more important and influential authors in the collaboration network, that is, the more authors seek to cooperate with a higher L-index, average citations per paper will increase.

4.2.6. Regression Analysis of H-Index and L-Index. The scatter diagram and regression curve of correlation between H-index and L-index are shown in Figure 11. Each point represents an author, and the red curve represents the regression curve of correlation between H-index and L-index. As can be seen from Figure 11, with the increase of L-index, H-index also increases. Then Stata was used to make unary linear regression of H-index and L-index. The results are shown in Table 7:

As can be seen from Table 7, the p -value is less than 0.01; the regression coefficient is $0.095 > 0$, which indicates that H-index is positively correlated with L-index; and this relationship is significant at the confidence level of 0.01. It also means that as the author's L-index goes up, his or her H-index also goes up. In other words, if more authors seek to cooperate with authors with higher L-index, that is, more authors seek to cooperate with more important and more influential authors in the scientific research collaboration network, the quantity and quality of their published papers will be improved.

4.3. Robustness Test. In reality, the research performance of the author will be affected by other variables except some of the author's own characteristics in the cooperation network, so his research performance is not a result of random allocation. In order to reduce the influence of other factors such as selection bias, this study uses the method of propensity score matching (PSM) to estimate the influence of degree centrality and L-index on the research performance of paper authors. When evaluating the research performance of the authors, we find a control group that is as similar as possible to the treatment group through the propensity score value and conduct a paired analysis; then the sample selection bias can be effectively reduced; and the observable factors such as control variables can be effectively removed. The author's research performance is affected, and the average treatment effect (ATT) after eliminating the selection bias can be obtained. Therefore, after selecting the most published journals, the starting year of the author's career, and the length of the author's career as matching variables, the PSM method can better avoid the general regression analysis of the degree centrality and the L-index on the research performance of the author. The estimated error in the past can effectively solve the endogenous problem. In addition, this study uses the nearest neighbor matching method to match the propensity value. This method can achieve the "one" to "all" (i.e., individuals in each treatment group and individuals in all control groups) matching and can also avoid degree centrality and L-index and the two-way causal effect of the author's research performance. The results of propensity score matching are shown in Table 8.

According to the empirical results in Table 8, taking the influence of degree centrality on the number of papers as an

TABLE 6: Unary linear regression results of average citations per paper and L-index.

Average citations per paper	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
L-index	0.926	0.216	4.29	≤0.001	0.503	1.35	***
Main journal	-0.288	0.093	-3.11	0.002	-0.47	-0.107	***
Start year	-1.635	0.047	-34.70	≤0.001	-1.728	-1.543	***
Career length	-0.584	0.095	-6.16	≤0.001	-0.769	-0.398	***
Constant	3,309.679	94.786	34.92	≤0.001	3,123.884	3,495.475	***
Mean dependent var		23.848		SD dependent var	37.333		
<i>R</i> -squared		0.056		Number of obs	12,875		
<i>F</i> -test		358.430		Prob > <i>F</i>	≤0.001		
Akaike crit. (AIC)		129,011.476		Bayesian crit. (BIC)	129,048.791		

*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

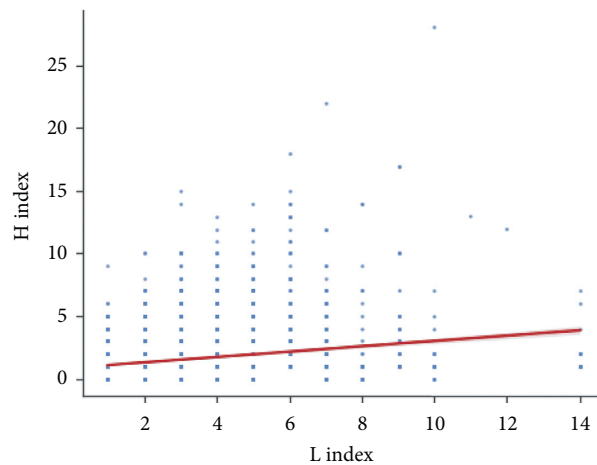


FIGURE 11: Scatter diagram and regression curve of correlation between H-index and L-index.

TABLE 7: Unary linear regression results of H-index and L-index.

H-index	Coeff.	Robust std. error	<i>t</i> -value	<i>p</i> -value	95% confidence interval		Sig
L-index	0.095	0.009	10.45	≤0.001	0.077	0.113	***
Main journal	-0.012	0.002	-6.71	≤0.001	-0.016	-0.009	***
Start year	-0.016	0.001	-14.94	≤0.001	-0.018	-0.014	***
Career length	0.319	0.01	31.11	≤0.001	0.298	0.339	***
Constant	32.61	2.137	15.26	≤0.001	28.421	36.799	***
Mean dependent var		1.395		SD dependent var	1.212		
<i>R</i> -squared		0.527		Number of obs	12,875		
<i>F</i> -test		326.115		Prob > <i>F</i>	≤0.001		
Akaike crit. (AIC)		31,861.910		Bayesian crit. (BIC)	31,899.225		

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

example, after matching, the average number of papers in the treatment group is 2.360; the average number of papers in the control group is 1.707; and the average treatment effect of ATT is 0.652 show great significance at 1% level. The test based on the PSM method shows that after controlling for the influence of other factors, the number of papers for authors with a higher level of degree centrality is 0.652 on average higher than that of authors with a lower level of degree centrality, which is about 38%, which means degree centrality. There is a significant improvement effect on the number of papers. In the same way, the L-index has a significant effect on the number of

papers, the degree centrality of the papers is cited, and the degree centrality and the L-index have a significant effect on the H-index.

In order to verify the sample characteristics of the main variables before and after the matching and the balance of the matching, the balance test was carried out according to different independent variables. The test results are shown in Tables 9 and 10. The standard deviation of each matching variable after the matching is small. According to Rosebaum and Rubin's point of view, the absolute value of the standard deviation value of the variable after the matching is significantly less than 20, which can be considered the

TABLE 8: Propensity score matching results.

Dependent variable	Scientific performance	Treatment group	Control group	ATT	Standard error	T-value
Degree centrality	The number of papers	2.360	1.707	0.652	0.048	13.35***
L-index		2.185	1.742	0.442	0.063	6.96***
Degree centrality	Average citations per article	25.061	22.344	2.717	1.130	2.40***
L-index		—	—	—	—	—
Degree centrality	H-index	2.044	1.577	0.467	0.034	13.62***
L-index		1.850	1.553	0.297	0.040	7.39***

TABLE 9: Degree centrality balance test.

Variable	Unmatched/matched	Mean		% reduction		t-test		V(T)/V(C)
		Treated	Control	% bias	bias	t	p > t	
Main journal	Unmatched	4.7676	5.0682	-8.0		-4.01	0.000	0.92*
	Matched	4.7887	4.7817	0.2	97.7	0.08	0.937	0.99
Start year	Unmatched	2009	2008.9	3.0		1.54	0.124	1.12*
	Matched	2009.1	2009.1	-0.3	90.5	-0.12	0.907	1.02
Career length	Unmatched	3.3256	1.4093	64.9		40.24	0.000	6.59*
	Matched	3.1639	3.1619	0.1	99.9	0.02	0.981	1.00

TABLE 10: L-index balance test.

Variable	Unmatched/matched	Mean		% reduction		t-test		V(T)/V(C)
		Treated	Control	% bias	bias	t	p > t	
Main journal	Unmatched	5.0425	4.9721	1.9		0.86	0.388	0.92*
	Matched	5.0469	4.9897	1.5	18.8	0.57	0.565	0.99
Start year	Unmatched	2009.5	2008.8	12.6		5.99	0.000	1.18*
	Matched	2009.5	2009.5	0.2	98.7	0.06	0.954	1.00
Career length	Unmatched	2.7577	1.703	35.4		19.50	0.000	3.04*
	Matched	2.7376	2.7461	-0.3	99.2	-0.09	0.931	0.99

matching method is suitable and the effect is better. Therefore, the propensity score matching result this time is more reliable.

The results of the robustness tests indicate that there is no endogeneity problem and also validate the rationality of using large-sample OLS regressions.

5. Research Conclusions and Implications

In this study, the correlation between the degree centrality and L-index of authors and scientific performance (number of papers, average citations per paper, and H-index) was analyzed in 16 core journals of Physical Education and Training, a second category discipline under the first category Sports, and the specific findings are as follows: (1) degree centrality is positively correlated with the number of papers, average citation per paper, and H-index and (2) L-index is positively correlated with the number of papers, average citations per paper, and H-index. From the results of this study, it can be seen that if the authors seek to cooperate with more authors extensively, especially with more influential authors, their research performance will be improved.

This study also has some theoretical contributions. This article analyzes the correlation between degree centrality, L-index of the authors in Physical Education and Training, and scientific performance. It lays a theoretical foundation for subsequent researches on the effectiveness of scientific

cooperation, provides a theoretical reference for predicting the scientific performance of scholars in a more scientific and comprehensive way in the future, and also provides a reference for the evaluation of the scientific performance of scholars in other disciplines.

In addition, this study also has some shortcomings. First of all, there is no in-depth study on the reasons why there is no correlation between average citations per paper and L-index. Second, the data of this study are 16 core journals of Physical Education and Training science in China, which define the subject area, and the research results may not be extrapolated to other fields. These need to be further explored in the follow-up research.

Data Availability

The data are available at the Chinese Social Sciences Citation Index (CSSCI) database (<http://cssrac.nju.edu.cn/>).

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Bin Zhang wrote the original draft, reviewed and edited the manuscript, and contributed literature collection. Jian Wu determined the framework of the paper, was responsible for

funding acquisition, and reviewed and edited the manuscript. Qian Huang wrote the original draft. Yujiao Tan and Lu Zhang reviewed and edited the manuscript. Qian Zheng, Yu Zhang, and Miao He contributed to the literature collection. Wei Wang was responsible for funding acquisition.

Acknowledgments

This work was supported in part by the National Social Science Foundation of China under Grant no. 20ATY007.

References

- [1] D. J. S. Price, "Little science, big science," *Von Der Studierstube Zur*, vol. 7, no. 3-6, pp. 443–458, 1963.
- [2] A. G. Heffner, "Funded research, multiple authorship, and subauthorship collaboration in four disciplines," *Scientometrics*, vol. 3, no. 1, pp. 5–12, 1981.
- [3] J. Ziman and R. W. Schmitt, "Prometheus bound: science in a dynamic steady state," *American Journal of Physics*, vol. 63, no. 5, pp. 476–477, 1995.
- [4] J. S. Katz and B. R. Martin, "What is research collaboration?" *Research Policy*, vol. 26, no. 1, pp. 1–18, 1997.
- [5] R. Zhao and F. Wen, "Scientific research collaboration and knowledge communication," *Library and Information Service*, vol. 55, no. 20, pp. 6–27, 2011.
- [6] Y. Okubo, *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples* OECD Publishing, Paris, France, 1997.
- [7] Wuyishan and L. Liang, "Some problems that should be paid attention to in quantitative evaluation of scientific research performance by using bibliometrics indicators," *Chinese Journal of Scientific and Technical Periodicals*, vol. 2, pp. 110–111, 2001.
- [8] Y. Zhang, "The influence factors should be used correctly to evaluate sci-tech periodicals and papers," *Acta Editologica*, vol. 4, pp. 214–215, 1998.
- [9] Wuyishan, "Some situations about SCI as a tool for scientific research performance evaluation," *Chinese Journal of Scientific and Technical Periodicals*, vol. 1, pp. 39–41, 2002.
- [10] J. Pang, *An. Scientific Metrological Research Methodology*, Scientific and Technical Documentation Press, Beijing, China, 1999.
- [11] G. Yang, *Metric Analysis on Scientific Papers and Evaluation of Scientific Performance for Military Medical University*, PLA Air Force Military Medical University, Xi'an, China, 2003.
- [12] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, 2005.
- [13] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.
- [14] J. Bihui, L. Liming, R. Rousseau, and L. Egghe, "The R- and AR-indices: complementing the h-index[J]," *Chinese Science Bulletin*, vol. 52, no. 6, pp. 855–863, 2007.
- [15] L. Bornmann and H. D. Daniel, *Are There Better Indices for Evaluation Purposes than the H Index? A Comparison of Nine Different Variants of the H Index Using Data from Biomedicine*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008.
- [16] B. Anand and S. Tripathi, "EM-index: a new measure to evaluate the scientific impact of scientists," *Scientometrics*, vol. 112, no. 1, pp. 659–677, 2017.
- [17] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures," *Journal of Informetrics*, vol. 5, no. 4, pp. 594–607, 2011.
- [18] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: a social capital perspective," *Research Policy*, vol. 42, no. 9, 2013.
- [19] C. N. Gonzalez-Brambila, F. M. Veloso, and D. Krackhardt, "The impact of network embeddedness on research output," *Research Policy*, vol. 42, no. 9, 2013.
- [20] A. Abbasi, R. T. Wigand, and L. Hossain, "Measuring social capital through network analysis and its influence on individual performance," *Library & Information Science Research*, vol. 36, no. 1, pp. 66–73, 2014.
- [21] C. Damien, D. Arnaud, L. Catherine, and M. Perroux, "The impact of a researcher's structural position on scientific performance: an empirical analysis," *PLoS One*, vol. 11, no. 8, Article ID e0161281, 2016.
- [22] A. Korn, A. Schubert, and A. Telcs, "Lobby index in networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 388, no. 11, pp. 2221–2226, 2009.
- [23] J. A. Barnes, *Class and Committees in a Norwegian Island Parish*, 1954.
- [24] L. C. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.
- [25] A. Herz, A. Müller, and B. Wellman, "Structural analysis: from method and metaphor to theory and substance," in *Social Structures: A Network Approach*, B. Wellman and S. D. Berkowitz, Eds., Cambridge University Press, Cambridge, UK, 2019.
- [26] C. Bao, X. Xie, and N. Shen, "Analysis of interpersonal networks," *Journal of The China Society for Scientific and Technical Information*, vol. 22, no. 3, pp. 365–374, 2003.
- [27] R. Hu and X. Deng, "Research on personal interpersonal network analysis system based on structure hole theory," *Journal of The China Society for Scientific and Technical Information*, vol. 24, no. 4, pp. 485–489, 2005.
- [28] Z. Liu, L. Yin, and D. Xu, "The application of complex network analysis method in collaborative research," *Science and Technology Management Research*, vol. 25, no. 12, pp. 267–269, 2005.
- [29] Y. Xu and Q. Zhu, "Demonstration study of social network analysis method in citation analysis," *Information Studies: Theory & Application*, vol. 2, pp. 184–188, 2008.
- [30] J. Qiu and F. Wang, "Analysis on author cooperation relationship of competitive intelligence research in China based on SNA," *Library Tribune*, vol. 30, no. 6, pp. 34–134, 2010.
- [31] J. Xing, W. Dong, D. Jia, and L. Yao, "Discussion on the study of digital library based on keywords network analysis," *Library Work and Study*, vol. 8, pp. 35–38, 2011.

Research Article

The Influence of Individual Characteristics on Cultural Consumption from the Perspective of Complex Social Network

Hui Liu,^{1,2} Shuang Lu,^{1,2} Ximeng Wang ,³ and Shaobo Long ^{4,5}

¹Center for Brand Leadership, Beijing 102488, China

²Business School, University of Chinese Academy of Social Sciences, Beijing 102488, China

³Cyber Finance Department, Postal Savings Bank of China, Beijing 100808, China

⁴Center for Public Economy & Public Policy, Chongqing 400044, China

⁵School of Public Affairs, Chongqing University, Chongqing 400044, China

Correspondence should be addressed to Shaobo Long; longshbcqu@126.com

Received 12 August 2021; Accepted 18 September 2021; Published 28 September 2021

Academic Editor: Xuzhen Zhu

Copyright © 2021 Hui 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 the era of the digital economy, social network as an important social capital has an important influence on individual consumption decision-making. This article uses the latest data from the China Household Finance Survey (CHFS) in 2017 to analyze the impact of personal characteristics on cultural consumption behavior under the influence of social networks from a theoretical and empirical perspective. Studies have shown that (1) social networks have a significant impact on cultural consumption; compared to gift money and social interaction, communication costs have a greater impact on cultural consumption; (2) communication costs have a greater impact on education consumption, entertainment consumption, and tourism consumption; (3) under the influence of social networks, individual characteristics have a significant impact on cultural consumption; (4) the higher the level of education, the easier it is for cultural consumption. There are intergenerational differences in cultural consumption expenditures of different age groups. It is easier for people to consume entertainment than the elderly.

1. Introduction

The Chinese economy is turning to a stage of high-quality development that focuses on the domestic big cycle and the international and domestic dual cycles promote each other. The 2020 “Government Work Report” clearly proposes and “firmly implements the strategy of expanding domestic demand,” highlighting the importance and urgency of domestic demand. In November 2020, the “Proposal of the Central Committee of the Communist Party of China on Formulating the Fourteenth Five-Year Plan for National Economic and Social Development and the Long-term Goals for 2035” (hereinafter referred to as the “Proposal”) for the first time put forward the concept of “High Quality of Life;” diversified cultural consumption is precisely the performance of the people’s high-quality life. The improvement of cultural consumption can not only improve the overall quality of life of residents but also force supply-side reforms and promote the transformation and upgrading of the

industrial structure, thereby accelerating the construction of China’s domestic economic cycle.

According to development economists such as Chenery, a country’s per capita GDP reaches US\$3,000; the cultural consumption expenditure accounts for about 23% of the total consumption expenditure. China’s per capita GDP in 2013 has exceeded 7,000 US dollars, but from 2013 to 2019, the proportion of per capita cultural consumption in total per capita consumption expenditure has been less than 12% (see Figure 1). It shows that the cultural consumption of Chinese residents is insufficient, and there is great potential for cultural consumption to improve. In addition, as shown in Figure 1, the percentage of per capita cultural consumption expenditure in per capita disposable income, the growth rate of residents’ daily consumption, and the growth rate of per capita cultural consumption have been in a state of fluctuation. It shows that other factors will also have an impact on residents’ cultural consumption expenditures, except income factors.

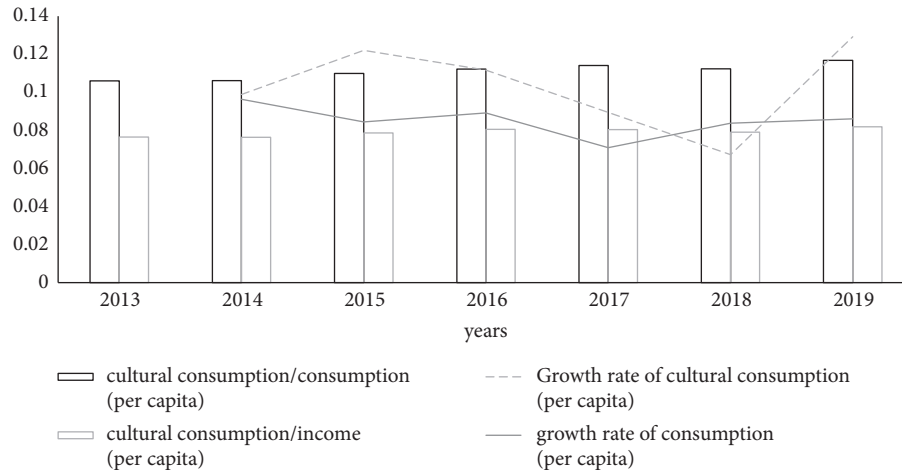


FIGURE 1: Growth of cultural consumption expenditure.

Social network analysis has gradually become an important method of economics research. A social network is a “relatively stable system of associations formed by the interaction between social individuals” [1]. It can bring benefits to the owner through the connection between individuals and has obvious social capital attributes. Social capital can effectively increase residents’ income [2], thereby increasing household consumption expenditure. Social network not only reflects the relationship between people but also is the best credit endorsement [3]. Social networks refer to “a relatively stable association system formed between social individuals because of interaction,” which can be used as an indicator to measure the level of household “relationship” [4]. The members of social networks are usually neighbors or have frequent contacts, which helps to alleviate various problems caused by information asymmetry [5], then establish a trust relationship, and deliver effective information. In addition, the social network is a relatively broad concept. The theoretical and empirical definitions are different. Therefore, the measurement of the social network is also different. Some use the number and type of relationships between businessmen as measurement indicators [6], some measure the popularity of social networks and the “neighbor trust index” [7], some discuss clan networks [8], and some scholars extend the relationship between relatives and friends to the number of gifts and gift change.

Cultural consumption is an important part of improving the quality of residents’ consumption and expanding domestic demand. There is a positive correlation between the expansion of tourism consumption and the improvement of residents’ cultural level, income growth, and the development of cultural industries [9]. The increase in consumer income has a limited effect on the expansion of cultural consumption demand [10]. The cultural consumption preferences of American residents are mainly affected by factors such as age and education [11]. The cultural consumption of college students is affected by personal characteristics, household economic status and education level, cultural consumption environment, and college students’

consumption concepts. Household culture and parents’ consumption concepts have a significant impact on the stratification of college students’ cultural consumption. The influencing factors of cultural consumption include individuals and collectives. In addition to paying attention to the symbolic value and hedonic value of cultural products, and their ability to create sensory and spiritual enjoyment, individual consumers also pay attention to the value of cultural products for sharing needs [12]. Therefore, this article believes that the study of cultural consumption should be a study under the condition of controlling personal characteristics. In addition, China is a typical society of human relations and will pay more attention to interpersonal relationships. Social networks are a typical feature of Chinese urban and rural society. Starting with social networks, it is a specific perspective for the analysis of the microfoundation of cultural consumption development.

In summary, considering the importance of social networks for cultural consumption. This research analyzes the impact of social networks on culture consumption, and under the influence of social networks, the impact of individual characteristics on cultural consumption explores the role of social networks in this impact. From theoretical and empirical analysis, we expect to explain the question of whether social networks can promote culture consumption and give targeted policy recommendations and satisfy the culture consumption needs of different households.

2. Theoretical Analysis and Hypothesis

2.1. Mechanism of Social Network on Culture Consumption. Duesenberry’s relative income hypothesis believes that consumers’ behaviors influence each other. If someone’s neighbor buys a new product that can improve the image of the household, he will be envious of the neighbor’s purchase behavior and will follow their neighbors to buy the same goods. The total utility obtained by consumers not only depends on their own consumption expenditures but also depends on the expenditures of other consumers, which means people have a tendency to compare [13]. The social

network will have an impact on consumers' personal consumption.

For individual consumers, their consumption choices are transformed into maximizing utility under certain budget constraints. The maximum utility is expressed as [14]

$$\max \sum_{t=0}^T U_t(p_t, C_{it}, z_{it}). \quad (1)$$

Consumers' intertemporal budget constraints are expressed as

$$A_{it+1} = (1+r)(A_{it} + Y_{it} - C_{it}). \quad (2)$$

$C_{it} = \sum_{k=1}^K P_t^k q_{it}^k$, P_t^k ($k = 1 \dots K$) denotes the corresponding price of cultural products or services q_{it}^k , C_{it} denotes individual cultural consumption, A_{it} denotes total assets owned by the individual, Y_{it} denotes individual income, and r denotes interest rate. This article draws on the practice of Blundell et al. [15] and sets the general form of utility as

$$U_t(p_t, C_{it}, z_{it}) = F_t(V_t(p_t, C_{it}, z_{it}^1), Z_{it}^2) + G(z_{it}^3). \quad (3)$$

$V_t(\cdot)$ denotes, in the same period, the choice between cultural consumption and savings. $U_t(\cdot)$ denotes, in different periods, an individual's allocation of cultural consumption. $F_t(\cdot)$ is a strictly increasing function, and $Z_{it} = (Z_{1it}, Z_{2it}, Z_{3it})$ is a condition vector reflecting the characteristics of cultural products or services. C denotes the individual culture consumption, and \bar{C} denotes Companion's cultural consumption. From formula (3), we could find the demand function of the same period or intertemporal distribution is independent of the function $F_t(\cdot)$, so according to Roy's identity,

$$q_{it}^k = -\frac{\partial V_t(\cdot)/\partial P_t^k}{\partial V_t(\cdot)/\partial C_{it}}. \quad (4)$$

The following Euler equation can be obtained as

$$\frac{\partial U_{t+1}(\cdot)}{\partial C_{it+1}} = (1+r)^{-1} \frac{\partial U_t(\cdot)}{\partial C_{it}}, \quad (5)$$

when $U_t(p_t, C_{it}, \{C_{nt}\}_{n=1, n \neq i}^N) = F_t(V_t(p_t, C_{it}, \{C_{nt}\}_{n=1, n \neq i}^N))$, no separation in the same period. Assume that

$$V_t(p_t, C_t, \bar{C}_t) = \frac{(C_{it}/a(p_t, \{C_{nt}\}_{n=1, n \neq i}^N))^{1-\gamma} - 1}{1-\gamma} \frac{1}{b(p_t, \{C_{nt}\}_{n=1, n \neq i}^N)} \prod_{n=1, n \neq i}^N C_{nt}^\theta, \quad (6)$$

$$U_t(\cdot) = F_t(V_t(\cdot)) = (1+\delta)^{-1} V_t(\cdot).$$

Let $a_t(\cdot) = a(p_t, \{C_{nt}\}_{n=1, n \neq i}^N)$ and $b_t(\cdot) = b(p_t, \{C_{nt}\}_{n=1, n \neq i}^N)$. By using Roy's identity, the budget constraint of cultural product or service j is

$$w_{it}^j = \frac{\partial \ln b(\cdot)}{\partial \ln P_t^j} \frac{1 - (C_{it}/a_t(\cdot))^{-(1-\gamma)}}{1-\gamma} + \frac{\partial \ln a_t(\cdot)}{\partial \ln P_t^j}. \quad (7)$$

When there is the price elasticity of cultural products or services consumed by peers changes, the individual's cultural consumption will be affected during the same period. Use a simple linear conversion mode:

$$\begin{aligned} \ln a(\cdot) &= a_0 + \sum_k (a_{0k} + a_{1k} \overline{\ln C_t}) \ln P_t^k + \frac{1}{2} \sum_k \sum_j \eta_{kj} \ln P_t^k P_t^j, \\ \ln b(\cdot) &= \sum_k (\beta_{0k} + \beta_{1k} \overline{\ln C_t}) \ln P_t^k. \end{aligned} \quad (8)$$

Then, the individual's expenditure on cultural products or services j depends on the cultural consumption of his peers. The cultural consumption of peers depends on the size of the coefficients a_{ij} and b_{ij} . If the above functional form is adopted, formula (7) can be transformed into

$$\begin{aligned} w_{it}^j &= a_{0j} + a_{1j} \overline{\ln C_t} + \sum_k \eta_{jk} \ln P_t^k \\ &+ (\beta_{0j} + \beta_{1j} \overline{\ln C_t}) \frac{1 - (C_{it}/a_t(\cdot))^{-(1-\gamma)}}{1-\gamma}. \end{aligned} \quad (9)$$

For intertemporal asset allocation, individual cultural consumption is also affected by peers ($\theta \neq 0$). That is, the social network influences cultural consumption:

$$\Delta \ln \frac{C_{it+1}}{a_{t+1}(\cdot)} \approx \gamma^{-1} \left((\gamma - \delta) - \Delta \ln b_{t+1}(\cdot) + \theta \Delta \frac{\overline{\ln C_{t+1}}}{a(P_{t+1})} \right). \quad (10)$$

In summary, the demand function of individual cultural consumption and Euler's equation are related to the cultural consumption of peers. Therefore, regardless of the same period or intertemporal period, cultural consumption has social network effects.

2.2. Theoretical Analysis and Hypothesis of Social Network on Culture Consumption. Urban agglomeration accelerates people's interaction, learning, and accumulation of human capital, provides a communication platform for people with professional knowledge and high skills, and then promotes

the formation of social network capital [16]. Social networks refer to “a relatively stable association system formed between social individuals because of interaction,” the more members, the more complex links, and the more communication with each other, the easier it is to form the social network capital, the higher the motivation for cultural consumption. And the word-of-mouth effect of cultural consumption could produce a positive spiral relationship. Social networks have three characteristics, namely, the scale of the household and friends network, the closeness of the household and friends network, and the supportability of the household and friends network. The gift money from families and friends could better reflect the characteristics of social networks. The popularization of mobile phones and the Internet have promoted the cultural consumption of urban residents. Therefore, we propose Hypothesis 1.

Hypothesis 1: social network has a significant positive impact on cultural consumption.

China is an urban-rural dualistic consumption structure. The household registration system has differentiated the social relationship networks, concepts, and behavioral norms of different household registration groups. As social capital, the social network has expanded the income gap between urban and rural households, which has different effects on the cultural consumption of households with different household registrations. Urban households have strong cultural consumption willingness and ability. Under the conditions of social interaction, household income has a positive impact on consumers' participation in the consumption of cultural products or services, and urban households are more active in participating in cultural consumption. Rural households have few resources, unstable employment, and heavy constraints on keeping promises [17], and social networks may have limited influence on their cultural consumption. Therefore, considering the urban-rural heterogeneity of the household, we propose Hypothesis 2.

Hypothesis 2: urban household registration has a higher probability of choosing cultural consumption than rural household registration.

From the perspective of social infection, consumers' susceptibility first drops and then rises with aging; that is, consumers aged under 18 are highly susceptible, and consumers aged from 19 to 23 are second, consumers aged from 24 to 31 are the lowest among the ages, and consumers aged over 31 are the highest. Consumers over the age of 31 are more influential, and the probability of influencing their peers is about 51% [18]. Age is an important factor that affects personal consumption behavior. Age has a significant impact on cultural consumption. Age is an important factor affecting personal consumption behavior, but it usually has different effects on different types of cultural consumption. The age effect of total cultural consumption often presents a “hump characteristic,” which is caused by the life characteristics of an individual and the life cycle of income [19]. In the early stages of life, individual consumption expenditure tends to increase with age, peak at the age of 45 to 55, and then decrease with age. Among the three categories of cultural consumption, education consumption also tends to

present “hump characteristics;” residents have to pay for themselves and their children's education expenditures; as they grow older, their education pressure will increase until the children reach adulthood. At the age of 50 years old, household education consumption reaches its highest. After 50 years old, the household's education expenditure begins to decline. At this time, with the increase of age, the household's education expenditure decreases. Entertainment consumption is often more popular with young people; young people buy more film and television works than other groups. Therefore, we propose Hypothesis 3.

Hypothesis 3a: cultural consumption increases before the age of 50 and then decreases after the age of 50.

Hypothesis 3b: education consumption increases before the age of 50 and then decreases after the age of 50.

Hypothesis 3c: entertainment consumption decreases with age.

Education has a significant impact on cultural consumption. Consumers with a high level of education have stronger cognitive abilities than consumers with a low level of education and will be more likely to obtain higher income growth and job security in the future. It is easier for them to accept new consumption concepts and lifestyles [20]. There is a social interaction effect in the influence of education level on individual behavioral decision-making [21]. The social network has a clustering effect [22]. The higher the social network clustering coefficient, the faster the speed of social infection and the wider the range. The size of the personal network of consumers with a bachelor's degree is twice that of a high school degree. The higher the level of education, the greater the possibility of mutual imitation between friends [23]. Consumers with different education levels have different social network attributes and scales, and the complexity of social networks has different effects on cultural consumption. Social interaction has a greater impact on highly educated consumers, making it easier to carry out development-centric consumption and enjoyment-centric consumption. Consumers with a bachelor's degree and a junior college degree are next. Consumers below high school and technical secondary school are restricted by their social resources and their cultural consumption level is also limited. Therefore, we propose Hypothesis 4.

Hypothesis 4a: consumers with a high school and technical secondary school degree or below have a lower probability of choosing cultural consumption.

Hypothesis 4b: consumers with bachelor's degrees and junior college degrees have a higher probability of choosing cultural consumption.

Hypothesis 4c: consumers with master's degrees and Ph.D. degrees have a higher probability of choosing cultural consumption.

A theoretical model is shown in Figure 2.

3. Materials and Methods

3.1. Data Processing. The data used in this article comes from the China Household Finance Survey and Research Center of Southwestern University of Finance and Economics, which conducted a nationwide survey of Chinese household

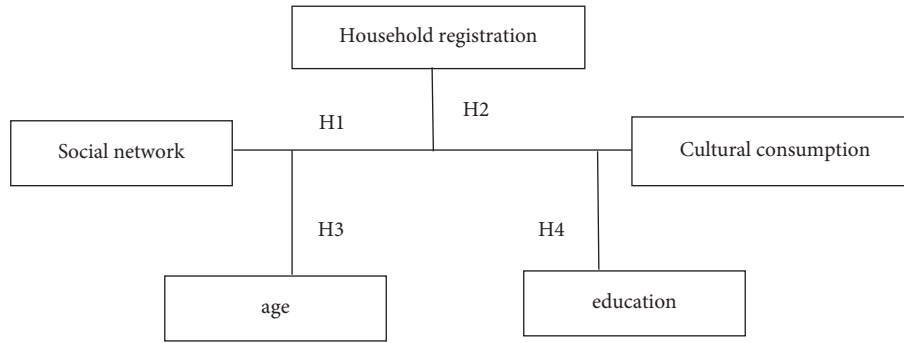


FIGURE 2: A theoretical model of influencing factors of individual characteristics on cultural consumption.

finance in 2017, referred to as CHFS2017. For more information about CHFS data, please refer to the relevant report issued by the China Household Finance Survey and Research Center. The sample covers 29 provinces (cities, autonomous regions) except Tibet, Hongkong, Macao, Xinjiang, and Inner Mongolia. The total sample is 40011; the main variables are selected according to the research objectives and the item settings of CHFS, including control variables that characterize social networks, cultural consumption, and cultural consumption, and cover characteristic variables at the household level.

Since the sample of household heads aged under 20 is very small, and people over 80 rarely carry cultural consumption, considering the balance of samples of various age groups, this article eliminates the data on household heads aged under 20 and over 80 years old. In addition, this paper also eliminates extreme data with zero cultural consumption and a negative logarithm of household total income. Finally, the samples of “Unable to judge,” “Missing,” “Inapplicable,” “Refusing to answer,” and others in the variables of social networks and culture consumption were eliminated, and 22454 valid samples were left.

3.2. Variable Selection and Measurement

3.2.1. Culture Consumption. This article selects cultural consumption as the main explanatory variable, takes the logarithm of the household’s annual cultural consumption to obtain cultural consumption variables, and then classifies cultural consumption to entertainment expenditure, education expenditure, and tourism expenditure. Because the proportions of these three parts are quite different, and education consumption tends to squeeze entertainment consumption and tourism consumption, this article firstly starts from the overall perspective and studies the influence of social networks and then studies the different effects of social networks on different types of cultural consumption. Regressing these three variables as dependent variables could analyze the structure of household cultural consumption. It can be seen from Table 1.

3.2.2. Social Networks. This article selects three indicators of gift money income and expenditure, communication costs, and social interaction as the proxy variables of the

social network. “Gift money” is the sum of gift money income and expenditure, which reflects an interactive process. Gift money expenditure is the household’s investment in and maintenance of the social network, and gift money income is the maintenance of the social network by other families; the sum of the two could reflect the scale, closeness, and supportability of the household’s social network. The gift money in this article comes from the CHFS questionnaire “Last year, your family gave out cash or noncash amounts to nonfamily members due to Spring Festival, Mid-Autumn Festival, etc.” “The amount of cash or noncash received during holidays and celebrations.”

In addition, in the era of the digital economy, communication between social network members is becoming more and more convenient, and communication expenditure could reflect the frequency and closeness of communication between social network members. Therefore, “communication cost” is another proxy variable of the household’s social network, and its data comes from “the household’s annual communication cost, network cost, and cable TV cost.” “Social interaction” reflects the positive situation of householders participating in company and community activities. Multiple variable proxy social network indicators could further verify the significance and robustness of the impact of social networks on cultural consumption.

3.2.3. Control Variables. Income and capital are important factors influencing consumption, so this article uses total household income and total household capital as control variables to avoid the endogenous problems caused by the lack of important variables in the model.

3.2.4. Independent Variables. Referring to studies such as Attanasio et al. [24] and Qin et al. [25], this article selects the household head’s gender, age, marital status, physical health, education, and household registration as control variables. The definition of variables and descriptive statistics are shown in Table 1. In addition, as cultural consumption is often greatly affected by leisure and retirees often have similar cultural consumption characteristics, this article sets the population over 60 years old as an age group.

TABLE 1: Definition of variables and descriptive statistics.

Variables	Definition	Mean	Standard deviation	Min	Max
Explained variables					
Cultural consumption	Household total cultural consumption last year	8.37181	1.768697	0	13.87378
Entertainment consumption	Household entertainment expenditure last year	7.012664	1.352275	2.484907	13.30468
Education consumption	Household education expenditure last year	8.61945	1.355617	2.079442	13.30468
Tourism consumption	Household tourist expenditure last year	8.488156	1.317347	1.609438	13.81551
Social networks					
Communication cost	The household's annual communication cost, network cost, and cable TV cost	7.644748	0.851387	3.583519	11.69525
Gift money	The sum of gift money income and expenditure	8.185994	1.229421	0	13.50285
Social interaction	If ever provided voluntary services: never = 0, ever = 1	0.257392	0.437207	0	1
Household characteristics					
Household size	Number of household members	3.627193	1.56269	1	15
Household income	Sum of nonfinancial assets and financial assets	11.11273	1.280413	0.04879	15.95532
Household assets	Total household income, including salary, business, transfer, and investment income	13.14457	1.668884	3.367296	17.97283
Householder's characteristics					
Gender	Female = 0, male = 1	0.797809	0.401643	0	1
Age	Survey year minus birth year	50.1679	13.31098	21	80
Education	Level of education	3.896682	1.7731	1	9
Marriage	Unmarried = 0, married = 1	0.891008	0.311636	0	1
Health	Very good = 1, good = 2, normal = 3, bad = 4, very bad = 5	2.459801	0.95662	1	5
Household registration	Urban = 0, village = 1	0.244055	0.429535	0	1
Age groups	Less than 31 years old = 0, 31~40 years old = 1, 41~50 years old = 2, 51~60 years old = 3, more than 60 years old = 4	2.396811	1.235731	0	4
Education groups	Below high school (technical secondary school) = 0; undergraduate and junior college = 1; master and doctor = 2	0.251984	0.472996	0	2

3.3. *Common Method.* This article uses the OLS model for regression analysis. The models are set as follows.

Model I is used to analyze the effect of social networks on cultural consumption, namely, Hypothesis 1. The specific regression equation is as follows:

$$C_i = \alpha + \gamma \cdot \text{social networks}_i + \beta \cdot \text{controls}_i + \varepsilon_i, \quad (11)$$

where i represents households. C_i denotes household cultural consumption behavior, including total cultural consumption, entertainment consumption, education consumption, and tourism consumption. social networks_i is independent variables, which represents the social network status of the household, including "Gift money," communication cost, and social interaction. controls_i is control variables, including total household income and total household assets. α represents a constant, γ represents the coefficient of social networks_i , β represents the coefficient of controls_i , and ε_i is the residual.

Model II: it is used to test the effect of the householder's individual characteristics on the household cultural consumption under the influence of social networks, that is, to verify Hypothesis 2~Hypothesis 4. The regression equation is as follows:

$$C_i = \alpha + \delta \cdot \text{individual}_i + \gamma \cdot \text{social networks}_i + \beta \cdot \text{controls}_i + \varepsilon_i, \quad (12)$$

where individual_i is independent variables, denoting the householder's individual characteristics, which contains

gender, age, education level, marital status, health status, and household registration. social networks_i and controls_i are control variables. δ represents the coefficient of individual_i .

4. Results and Discussion

Based on the hypothesis, this article firstly makes an empirical test on the influence of social networks on household cultural consumption, that is, to verify Hypothesis 1. After H1 is successfully proved, we then test the influence of the householder's personal characteristics on the household cultural consumption under the influence of the social network, namely, Hypothesis 2~Hypothesis 4.

4.1. Effects of Social Networks on Cultural Consumption.

The influence coefficients of communication expenses, gift money, and social interaction on cultural consumption are all positively significant at the level of 1%, which shows that cultural consumption increases as the degree of social network connection increases (Table 2). Therefore, Hypothesis 1 was proven.

In addition, we separate the social networks into three proxy variables: communication cost, gift money, and social interaction. As can be seen from Table 2, we find that the communication costs have the greatest impact on cultural consumption; for every 1% increase in communication cost, cultural consumption increases by 0.491%; cultural

TABLE 2: Impact of social networks on cultural consumption behavior.

	Cultural consumption		Entertainment consumption		Education consumption		Tourism consumption	
	Model1	Model2	Model1	Model2	Model1	Model2	Model1	Model2
Household income	0.178*** (0.0101)	0.0838*** (0.0119)	0.189*** (0.0119)	0.0920*** (0.0134)	0.118*** (0.00955)	0.0717*** (0.0117)	0.243*** (0.0138)	0.199*** (0.0164)
Household assets	0.234*** (0.00772)	0.144*** (0.00913)	0.168*** (0.00885)	0.0895*** (0.00996)	0.160*** (0.00786)	0.109*** (0.0100)	0.275*** (0.0104)	0.230*** (0.0119)
Communication cost		0.491*** (0.0162)		0.590*** (0.0177)		0.209*** (0.0176)		0.194*** (0.0209)
Gift money		0.0994*** (0.0106)		0.0746*** (0.0115)		0.0869*** (0.0113)		0.144*** (0.0131)
Social interaction		0.135*** (0.0273)		0.167*** (0.0266)		0.0922*** (0.0294)		0.0636** (0.0299)
Constant	3.320*** (0.109)	1.009*** (0.145)	2.585*** (0.133)	-0.559*** (0.168)	5.234*** (0.103)	4.085*** (0.146)	1.836*** (0.159)	0.178 (0.202)
<i>N</i>	22446	16409	11236	8796	14111	10400	7785	6199
<i>R</i> ²	0.093	0.153	0.093	0.214	0.074	0.095	0.191	0.229

Note. Standard errors are given in parentheses. * $P < 0.1$, ** $P < 0.05$, and *** $P < 0.01$.

consumption is secondly affected by social interaction. The social interaction variable is a dummy variable, and its coefficient in the cultural consumption model is 0.135. So, given the same household income, household assets, communication cost, and gift money, the household cultural consumption of people who have provided voluntary service is 13.5% higher than that of people who have never provided it. The impact of gift money is the least; for every 1% increase in gift money, household cultural consumption increases by 0.0994%.

We separate cultural consumption into three categories: entertainment consumption, education consumption, and tourism consumption. The result of Table 2 shows that communication cost has the greatest impact on entertainment consumption, followed by education consumption, and has the least impact on tourism consumption. Gift money has the largest impact on tourism consumption, and the influence coefficient on entertainment consumption is the smallest. Compared with entertainment consumption and education consumption, tourism consumption is less affected by social interaction factors. Among the three proxy variables of social networks, communication cost has the greatest impact on cultural consumption and its classified consumption [26, 27].

4.2. Effects of Householder's Individual Characteristics on Cultural Consumption under the Influence of Social Networks.

The results of Table 3 show that household registration has a significant impact on cultural consumption. Urban households will carry more cultural consumption than rural households; the cultural consumption of rural households is 19.1% lower than that of urban households. Education consumption of rural households is 15.9% lower than that of urban households, and travel consumption of rural households is 50.5% lower than that of urban households. At a significant level of 10%, rural households' entertainment consumption is 7.65% lower than that of urban households. It proves that Hypothesis 2 is valid.

As can be seen from Table 3, compared with the age group under 30, the cultural consumption of the 31–40 age group and the 41–50 age group is positively significant at the 1% level, and the regression coefficient of the 41–50 age group is larger. It shows that in people under 50 years old, the older the age group, the more cultural consumption. Compared with the age group under 30, the cultural consumption of the two oldest age groups is negatively significant. And the coefficient is smaller for seniors over 60, which shows that in people over 50, the older the age group, the less cultural consumption. This verifies Hypothesis 3a and shows that the age effect of cultural consumption does show a “hump characteristic.” In addition, Figure 3 shows the influence coefficients of the five age groups on cultural consumption. It can be seen that the age effect of education consumption shows a trend of first increasing and then decreasing; that is, early household education consumption increased with the age of the householder, and the later household education consumption decreased with the householder's age, which verifies Hypothesis 3b. Cultural consumption and education consumption keep the same trend with age, and both have hump characteristics. This is mainly because educational consumption accounts for the largest proportion of cultural consumption. Among all age groups, the entertainment consumption of the older age group is lower than that of the younger age group, which validates Hypothesis 3c, indicating that entertainment consumption decreased with age. Furthermore, the regression results also show that the age group “over 60” will spend more on tourism compared to the “51–60 years old” group. For the elderly, the cognitive effect of social interaction could improve their quality of life.

Compared with the high school and technical secondary school or below, the cultural consumption of the “undergraduate and junior college” education group ($\delta = 0.434$) and the “master and doctor” education group ($\delta = 0.736$) are positively significant at the 1% level. The result of Table 3 shows that the cultural consumption of high-level education

TABLE 3: Impact of householder's individual characteristics on cultural consumption under the influence of social networks.

	Cultural consumption		Entertainment consumption		Education consumption		Tourism consumption	
	Model3	Model4	Model3	Model4	Model3	Model4	Model3	Model4
Communication costs	0.491*** (0.0162)	0.347*** (0.0165)	0.590*** (0.0177)	0.481*** (0.0183)	0.209*** (0.0176)	0.200*** (0.0177)	0.194*** (0.0209)	0.274*** (0.0219)
Gift money	0.0994*** (0.0106)	0.110*** (0.0103)	0.0746*** (0.0115)	0.0729*** (0.0110)	0.0869*** (0.0113)	0.0843*** (0.0111)	0.144*** (0.0131)	0.136*** (0.0127)
Social interaction	0.135*** (0.0273)	0.101*** (0.0267)	0.167*** (0.0266)	0.0933*** (0.0258)	0.0922*** (0.0294)	0.0394 (0.0294)	0.0636** (0.0299)	-0.000509 (0.0292)
Household income	0.0838*** (0.0119)	0.0542*** (0.0118)	0.0920*** (0.0134)	0.0767*** (0.0132)	0.0717*** (0.0117)	0.0629*** (0.0118)	0.199*** (0.0164)	0.181*** (0.0164)
Household assets	0.144*** (0.00913)	0.133*** (0.00916)	0.0895*** (0.00996)	0.0967*** (0.00987)	0.109*** (0.0100)	0.0751*** (0.0103)	0.230*** (0.0119)	0.198*** (0.0119)
Gender		-0.131*** (0.0306)		-0.000493 (0.0303)		-0.138*** (0.0341)		-0.0987*** (0.0337)
Marriage		-0.0151 (0.0425)		-0.269*** (0.0428)		0.178*** (0.0521)		0.0591 (0.0495)
Health		-0.00239 (0.0131)		0.00254 (0.0143)		0.00606 (0.0138)		-0.0361** (0.0171)
Household registration		-0.191*** (0.0318)		-0.0765* (0.0408)		-0.159*** (0.0313)		-0.505*** (0.0526)
Household size		0.113*** (0.00887)		0.00745 (0.0100)		-0.0173* (0.00997)		-0.106*** (0.0131)
31~40		0.346*** (0.0527)		-0.180*** (0.0484)		0.134** (0.0616)		-0.00263 (0.0590)
41~50		0.487*** (0.0513)		-0.402*** (0.0482)		0.478*** (0.0606)		0.0509 (0.0585)
51~60		-0.0939* (0.0536)		-0.630*** (0.0509)		0.172*** (0.0655)		0.175*** (0.0602)
Over 60		-0.319*** (0.0531)		-0.909*** (0.0496)		0.0195 (0.0669)		0.255*** (0.0595)
Undergraduate and junior college		0.432*** (0.0315)		0.177*** (0.0296)		0.206*** (0.0349)		0.323*** (0.0331)
Master and doctor		0.732*** (0.0914)		0.354*** (0.0768)		0.429*** (0.0970)		0.476*** (0.0831)
Constant	1.009*** (0.145)	2.048*** (0.168)	-0.559*** (0.168)	1.045*** (0.183)	4.085*** (0.146)	4.491*** (0.180)	0.178 (0.202)	0.501** (0.224)
N	16409	16386	8796	8781	10400	10385	6199	6192
R ²	0.153	0.215	0.214	0.285	0.095	0.127	0.229	0.282

Note. Standard errors are given in parentheses. * $P < 0.1$, ** $P < 0.05$, and *** $P < 0.01$.

groups is greater than that of low-level education groups, and the cultural consumption of the “master and doctor” group is higher than that of the “undergraduate and college” group. In the same way, the expenditure on entertainment consumption, education consumption, and tourism consumption of the two high-level education groups is greater than that of the “high school and technical secondary school” group, and the three kinds of consumption of the “master and doctor” group are all higher than the “undergraduate and college” group. This verifies Hypothesis 4. As the education level of the householder increases, household cultural consumption will increase.

4.3. Extended Discussion. Other personal characteristics also affect household cultural consumption (Table 3). When the householder is a female, at a significant level of 1%, cultural

consumption is 13.1% lower, education consumption is 13.8% lower, and tourism consumption is 9.87% lower than that of male households. The gender of the household has no significant influence on entertainment consumption. As shown in Table 3, the household's marital status has no significant impact on the total cultural consumption and tourism consumption, but at a significant level of 1%, the unmarried householder produces more entertainment consumption and less education consumption than the married head. The health of households only has a significant impact on tourism consumption at the level of 5%. At a significant level of 1%, household size has a positive impact on household cultural consumption. For every 1% increase in household size, total household cultural consumption increases by 0.113%. But it has a negative impact on tourism consumption. For every 1% increase in household size, tourism consumption decreases by 0.106%.

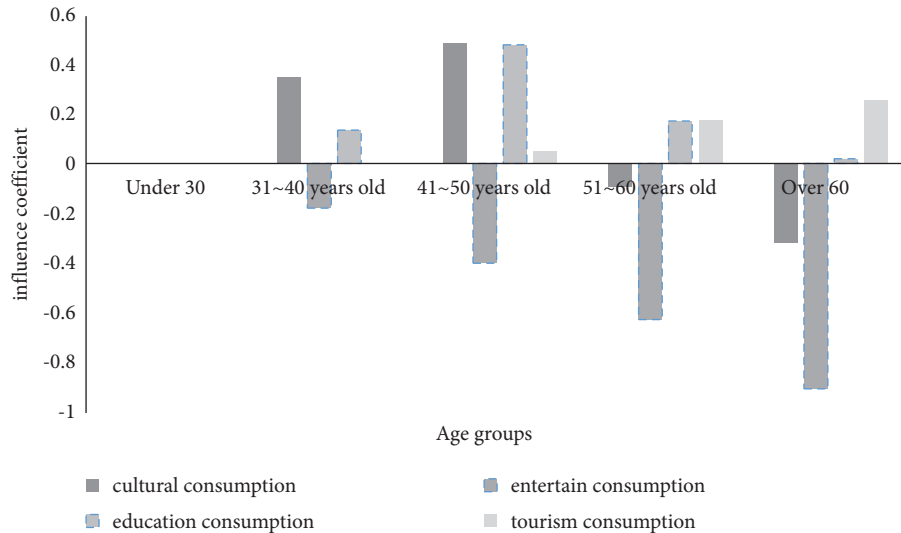


FIGURE 3: Impact of age groups on cultural consumption.

4.4. Robustness Discussion. This article uses Stata to test the correlation coefficient between the variables, as shown in Table 4. The correlation test results show that all correlation coefficients between each variable are less than 0.5, indicating that the correlation between the variables is low. At the same time, the variance inflation factor is used to test the multicollinearity problem, as shown in Table 5. It is found that the VIF value of the four models of cultural consumption, entertainment consumption, education consumption, and tourism consumption does not exceed 10, indicating that there is no multicollinearity.

4.5. Endogenous Discussion. The endogenous problem mainly comes from two aspects: omitted variables and two-way causality. In order to avoid the endogeneity caused by the omitted variables, this article uses three indicators of “Gift money,” “Communication costs,” and “Social interaction” to represent social network variables. To a certain extent, the robustness of the model is improved, but there may be a two-way causal relationship in this model; that is, while the communication cost impacts cultural consumption, it may also be affected by cultural consumption. In the era of the Internet, many kinds of cultural consumption are carried out online, so the greater the amount of cultural consumption and the richer the variety, the more Internet and TV fees need to be paid. In other words, cultural consumption generates communication costs. If this relationship exists, it will inevitably cause endogeneity, thereby reducing the robustness of the test results. Therefore, it is necessary to further discuss and analyze endogeneity.

In order to identify and eliminate endogeneity, it is necessary to find suitable instrumental variables for communication costs and then use the two-stage least squares method (2SLS) for further testing.

After trial and error, this article introduces the variable of “mobile phone type” as an instrumental variable for communication costs. The types of mobile phones are “no mobile phone,” “nonsmart phone,” and “smartphone.” “Smartphone” is set to variable 1 and other types are set to variable 0. Families with smartphones tend to incur more network expenses (communication costs), thereby having a stronger social network, which in turn affects cultural consumption. But at the same time, the residual’s phone type does not directly affect the cultural consumption of the household. Therefore, in theory, choosing this variable as an instrumental variable of communication costs meets the conditions of an effective instrumental variable.

Table 6 shows the results of the first stage regression of the instrumental variable. It shows that, in both cultural consumption and the three classification models, the regression coefficient of the instrumental variable (mobile phone type) is positive, and the regression coefficient is significant at the 1% statistical level, which indicates that the instrumental variable has good explanatory power for the endogenous explanatory variable and meets the conditions of relevance. The F-value in the four models is 391, 209, 229, and 163, which are all significantly higher than the empirical cut point of 10. Therefore, we can directly reject the null hypothesis that the variable “mobile phone type” is a weak instrumental variable of communication costs.

From the statistical results of the second stage in Table 7, after introducing instrumental variables to overcome the endogeneity of communication costs, the coefficients of communication costs on household cultural consumption, entertainment consumption, education consumption, and tourism consumption are still positive, and all have passed the significance test at the 1% statistical level.

Above all, communication costs indeed significantly affect household cultural consumption (and its classification). At the same time, compared with the previous

TABLE 4: Correlation coefficient between variables.

Variables	Communication cost	Gift money	Social interaction	Household income	Household assets	Gender	Marriage	Health	Hukou	Household size	Age groups	Education groups
Communication	1											
cost												
Gift money	0.2623	1										
Social interaction	0.1099	0.1138	1									
Household income	0.3448	0.3223	0.1454	1								
Household assets	0.3679	0.3217	0.1702	0.497	1							
Gender	0.0129	0.0041	-0.064	-0.0053	-0.0279	1						
Marriage	0.1042	0.0705	-0.053	0.1001	0.095	0.301	1					
Health	-0.1702	-0.1233	-0.0722	-0.1704	-0.2171	-0.028	-0.0268	1				
Hukou	-0.1751	-0.1388	-0.1394	-0.2655	-0.3093	0.1595	0.0735	0.1444	1			
Household size	0.161	-0.0697	-0.1198	0.0092	-0.0904	0.179	0.2747	0.0717	0.3341	1		
Age groups	-0.2187	-0.0489	-0.0511	-0.0505	-0.0326	-0.0336	0.0065	0.2464	0.0575	-0.0453	1	
Education groups	0.1968	0.1833	0.1987	0.3027	0.3021	-0.0515	-0.0459	-0.1692	-0.2729	-0.2102	-0.2763	1

TABLE 5: Variance inflation factor test.

	VIF			
	Cultural consumption	Entertainment consumption	Education consumption	Tourism consumption
Household income	1.58	1.49	1.72	1.51
Household assets	1.5	1.48	1.52	1.45
Household size	1.35	1.37	1.36	1.4
Communication cost	1.37	1.39	1.37	1.42
Gift money	1.2	1.22	1.2	1.14
Social interaction	1.07	1.05	1.09	1.04
Gender	1.13	1.11	1.14	1.08
Marriage	1.19	1.17	1.17	1.12
Health	1.14	1.12	1.16	1.1
Household registration	1.32	1.28	1.28	1.34
Age groups	1.34	1.35	1.31	1.35
Education groups	1.2	1.23	1.23	1.24
Average VIF	1.28	1.27	1.3	1.27

TABLE 6: First-stage regression of instrumental variable.

	Communication cost			
	Cultural consumption	Entertainment consumption	Education consumption	Tourism consumption
Phone type	0.343*** (0.0174)	0.357*** (0.0263)	0.311*** (0.0225)	0.350*** (0.0329)
Gift money	0.082*** (0.0049)	0.075*** (0.0065)	0.101*** (0.0063)	0.062*** (0.0073)
Social interaction	0.044*** (0.0120)	0.041*** (0.0148)	0.039*** (0.0153)	0.031*** (0.0165)
Household income	0.086*** (0.0063)	0.083*** (0.0088)	0.074*** (0.0072)	0.092*** (0.0108)
Household assets	0.100*** (0.0046)	0.087*** (0.0060)	0.124*** (0.0059)	0.088*** (0.0076)
Gender	-0.018 (0.0140)	-0.025 (0.0175)	-0.009 (0.0181)	-0.010 (0.0192)
Marriage	0.019 (0.0202)	0.017 (0.0250)	0.020 (0.0303)	-0.023 (0.0297)
Health	-0.027*** (0.0064)	-0.015* (0.0084)	-0.025** (0.0079)	-0.00696 (0.0103)
Hukou	-0.161*** (0.0160)	-0.11*** (0.0260)	-0.163*** (0.0185)	-0.109** (0.0331)
Household size	0.130*** (0.0044)	0.149*** (0.0060)	0.104*** (0.0058)	0.172*** (0.0075)
31~40	-0.177*** (0.0223)	-0.161*** (0.0257)	-0.168*** (0.0313)	-0.157*** (0.0327)
41~50	-0.179*** (0.0218)	-0.147*** (0.0252)	-0.173*** (0.0308)	-0.129*** (0.0319)
51~60	-0.185*** (0.0233)	-0.133*** (0.0276)	-0.167*** (0.0346)	-0.184*** (0.0337)
Over 60	-0.382*** (0.0238)	-0.392*** (0.0281)	-0.264*** (0.0363)	-0.383*** (0.0344)
Undergraduate and junior college	0.057*** (0.0138)	0.054** (0.0167)	0.018188 (0.0172)	0.037*** (0.0193)
Master and doctor	0.102*** (0.0347)	0.117** (0.0366)	0.107835 (0.0448)	0.101*** (0.0406)
Constant	4.257*** (0.0758)	4.469*** (0.1012)	4.035*** (0.0954)	4.438*** (0.1286)
<i>N</i>	16386	8781	10385	6192
<i>R</i> ²	0.2954	0.3046	0.2897	0.3207
<i>F</i> -value	391.34	209.61	229.76	163.95

Note. Standard errors are given in parentheses. * $P < 0.1$, ** $P < 0.05$, and *** $P < 0.01$.

TABLE 7: Second-stage regression of instrumental variables.

	Cultural consumption	Entertainment consumption	Education consumption	Tourism consumption
Communication cost	1.203*** (0.104)	1.065*** (0.125)	0.502*** (0.125)	0.880*** (0.179)
Gift money	0.0366*** (0.0141)	0.0272* (0.0155)	0.0529*** (0.0173)	0.0979*** (0.0177)
Social interaction	0.0479 (0.0295)	0.0615** (0.0278)	0.0227 (0.0303)	-0.0250 (0.0311)
Household income	-0.0258 (0.0159)	0.0256 (0.0182)	0.0392** (0.0163)	0.121*** (0.0249)
Household assets	0.0378** (0.0151)	0.0408** (0.0160)	0.0340* (0.0200)	0.141*** (0.0209)
Gender	-0.110*** (0.0331)	0.0184 (0.0323)	-0.134*** (0.0341)	-0.0906** (0.0366)
Marriage	-0.0406 (0.0459)	-0.282*** (0.0470)	0.170*** (0.0575)	0.0714 (0.0545)
Health	0.0265* (0.0146)	0.0141 (0.0156)	0.0160 (0.0148)	-0.0296 (0.0192)
Hukou	-0.0130 (0.0404)	0.0153 (0.0501)	-0.0993** (0.0405)	-0.412*** (0.0674)
Household size	0.00979 (0.0157)	-0.0774*** (0.0211)	-0.0462*** (0.0160)	-0.209*** (0.0341)
31~40	0.499*** (0.0597)	-0.0817 (0.0513)	0.186*** (0.0634)	0.0928 (0.0675)
41~50	0.645*** (0.0584)	-0.313*** (0.0516)	0.532*** (0.0631)	0.129** (0.0650)
51~60	0.101 (0.0624)	-0.534*** (0.0559)	0.239*** (0.0713)	0.294*** (0.0727)
Over 60	0.108 (0.0768)	-0.614*** (0.0809)	0.132 (0.0840)	0.535*** (0.103)
Undergraduate and junior college	0.376*** (0.0346)	0.141*** (0.0314)	0.200*** (0.0344)	0.296*** (0.0350)
Master and doctor	0.657*** (0.0990)	0.289*** (0.0748)	0.403*** (0.0965)	0.418*** (0.0808)
Constant	-1.685*** (0.482)	-1.658*** (0.605)	3.238*** (0.554)	-2.297*** (0.865)
<i>N</i>	16386	8781	10385	6192
<i>R</i> ²	0.087	0.202	0.102	0.194

Note. Standard errors are given in parentheses. * $P < 0.1$, ** $P < 0.05$, and *** $P < 0.01$.

method, new results are only different in terms of the size of the coefficients, and the sign of the coefficients has not changed.

5. Conclusions

This article studies the influence of individual characteristics on cultural consumption from the perspective of social networks. The article first theoretically analyzes the social network effect of cultural consumption under the life-cycle framework, then proposes hypotheses, and conducts empirical tests. The empirical results show that social network has a significant impact on cultural consumption. In addition, compared to gift money and social interaction, communication has a greater impact on cultural consumption. At the same time, we also find that different social

network variables have different effects on different cultural consumption, communication has the greatest impact on education consumption, entertainment consumption, and tourism consumption, and communication has a greater effect on entertainment consumption than education consumption and tourism consumption. Under the influence of social networks, individual characteristics have a significant impact on cultural consumption. Compared with rural households, urban households make more cultural consumption. The higher the educational level, the easier it is; the higher the degree, the easier it is to carry out cultural consumption activities; individuals with postgraduate and Ph.D. degrees have the highest cultural consumption expenditures. Different age groups have different perceptions of cultural consumption expenditures called intergenerational differences; the younger people are more likely to

engage in entertainment consumption than the elderly. The influence of other individual characteristics on cultural consumption is still significant. Gender has a significant impact on cultural consumption. Men's cultural consumption expenditure is higher; unmarried households produce more entertainment consumption than married households, and married households have more education consumption than unmarried households.

Although we have conducted an in-depth research on the relationship between social networks and culture consumption and the social network effect of the influence of individual characteristics on cultural consumption, this study has also some shortcomings. Firstly, this study selects cross-sectional sample data for analysis from an individual perspective, and it is difficult to obtain the dynamic process of variables. Secondly, due to the diversity and complexity of consumer behavior and social network and the availability of data, it is difficult to analyze the impact of the complexity of social networks on cultural consumption. In addition, this research mainly focuses on the moderating role of the social networks; future research could also conduct the other functions of social networks on culture consumption.

Data Availability

The data used in this article come from the China Household Finance Survey and Research Center of Southwestern University of Finance and Economics, which conducted a nationwide survey of Chinese household finance in 2017, referred to as CHFS2017 (<https://chfs.swufe.edu.cn/datacenter/apply.html>).

Conflicts of Interest

All authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors acknowledge the financial support from the National Social Science Foundation of China (no. 19BH156) and Innovation Project of the University of Chinese Academy of Social Sciences, Fundamental Research Funds for the Central Universities (Grant number 2019CDSKXYGG0042).

References

- [1] J. C. Mitchell, *The Concept and Use of Social Networks Social Networks in Urban Situations*, Bobbs-Merrill, Indianapolis, Indiana, 1969.
- [2] K. Munshi and M. Rosenzweig, "Traditional institutions meet the modern world: caste, gender and schooling choice in a globalizing economy," *The American Economic Review*, vol. 96, no. 4, pp. 1225–1252, 2006.
- [3] J. E. Stiglitz, "Peer monitoring and credit markets," *The World Bank Economic Review*, vol. 4, no. 3, pp. 351–366, 1990.
- [4] B. Wellman, "Physical place and cyberplace: the rise of personalized networking," *International Journal of Urban & Regional Research*, vol. 25, no. 2, pp. 227–252, 2015.
- [5] D. Karlan and, "Social connections and group banking," *Economic Journal*, vol. 117, pp. 52–84, 2007.
- [6] M. Fafchamps and B. Minten, "Returns to social network capital among traders," *Oxford Economic Papers*, vol. 54, no. No.2, pp. 173–206, 2002.
- [7] J. Isham and S. Kahkonen, *How Do Participation and Social Capital Affect Community-Based Water Projects? Evidence from Central Java*, Cambridge University Press, Cambridge, UK, 2002.
- [8] Y. Peng, "Kinship networks and entrepreneurs in China's transitional economy," *American Journal of Sociology*, vol. 109, no. 5, pp. 1045–1074, 2004.
- [9] G. Richards, "Production and consumption of European cultural tourism," *Annals of Tourism Research*, vol. 23, no. 2, pp. 261–283, 1996.
- [10] P. Brito and C. Barros, "Learning-by-Consuming and the dynamics of the demand and prices of cultural goods," *Journal of Cultural Economics*, vol. 29, no. 2, pp. 83–106, 2005.
- [11] T. Katz -Gerro, "Cultural consumption and social stratification: leisure activities, musical tastes, and social location," *Sociological Perspectives*, vol. 5, p. 55, 1999.
- [12] D. Bourgeon-Renault, C. Urbain, A. Gombault, M. Le Gall-Ely, and C. Petr, "An experiential approach to the consumption value of arts and culture: the case of museums and monuments," *International Journal of Arts Management*, vol. 9, no. 1, pp. 35–47, 2006.
- [13] J. S. Duesenberry, *Income-consumption Relations and Their Implications*, WW Norton & Company, New York, NY, USA, 1948.
- [14] S. Che and J. Gu, "The social network effect of cultural consumption," *Consumer Economics*, vol. 32, no. 6, pp. 51–58, 2016.
- [15] R. Blundell, M. Browning, and C. Meghir, "Consumer demand and the life-cycle allocation of household expenditures," *The Review of Economic Studies*, vol. 61, no. 1, pp. 57–80, 1994.
- [16] E. L. Glaeser and L. Edward, "Learning in cities," *Journal of Urban Economics*, vol. 46, no. 2, pp. 254–277, 1999.
- [17] A. Del-Río and G. Young, "The determinants of unsecured borrowing: evidence from the British household panel survey," *Bank of England Quarterly Bulletin*, vol. 4, no. 2, p. 187, 2005.
- [18] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.
- [19] X. Y. Chen, *Research on the Impact of Population Age Structure Changes on Residents' Consumption (In Chinese)*, China Social Sciences Press, vol. 4, no. 1, p. 165, Beijing, China, 2017.
- [20] E. Baek and G.-S. Hong, "Effects of family life-cycle stages on consumer debts," *Journal of Family and Economic Issues*, vol. 25, no. 3, pp. 359–385, 2004.
- [21] T. Li, L. Han, L. Zhang, and S. Rozelle, "Encouraging classroom peer interactions: Evidence from Chinese migrant schools," *Journal of Public Economics*, vol. 111, no. 3, pp. 29–45, 2014.
- [22] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.
- [23] N. A. Christakis and J. H. Fowler, *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, Little, Brown Spark, New York, NY, USA, 2009.
- [24] O. Attanasio, A. Barr, J. C. Cardenas, G. Genicot, and C. Meghir, "Risk pooling, risk preferences and social

- network,” *American Economic Journal: Applied Economics*, vol. 4, no. 2, pp. 134–167, 2012.
- [25] H. L. Qin, C. W. Li, and J. L. Wan, “Social capital, farmer heterogeneity and credit behavior-measurement analysis and empirical test based on CFPS data,” *Finance and Economy*, vol. 1, pp. 33–40, 2019.
- [26] F. Xiong, X. M. Wang, and S. R. Pan, “Social recommendation with evolutionary opinion dynamics,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [27] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, “Exploiting implicit influence from information propagation for social recommendation,” *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.

Research Article

Firms' Investment Behaviours in Temperature-Controlled Supply Chain Networks

Mengshuai Zhu,¹ Hao Chen,² Ximeng Wang,³ Yonghan Wang,¹ Chen Shen ¹
and Cong Zhu ¹

¹Agricultural Information Institute, Chinese Academy of Agricultural Sciences/Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China

²Key Laboratory of Agricultural Remote Sensing (AGRIRS), Ministry of Agriculture and Rural Affairs, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

³Cyber Finance Department, Postal Savings Bank of China, Beijing 100808, China

Correspondence should be addressed to Chen Shen; shenchen@caas.cn and Cong Zhu; zhucong@caas.cn

Received 8 June 2021; Accepted 19 July 2021; Published 20 September 2021

Academic Editor: Siew Ann Cheong

Copyright © 2021 Mengshuai Zhu 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.

Delivering high-quality food into markets is a vital expectation of modern customers. The significant increase in consumers' awareness of food freshness, nutrition, and safety makes the temperature-controlled supply chain (TCSC) the focus of food logistics safety. However, a large number of Chinese companies are still reluctant to invest in the food supply chain, resulting in a high rate of supply chain logistics loss. This research aims to establish an economic model to explain why these companies do not invest and under what conditions they will do. The results show that high economic investment is the main reason that hinders companies' willingness to build TCSC. Large companies with bigger production are more willing to invest in TCSC than small companies. Besides, larger companies running with high-quality products could get more profit while small companies operating with normal products are less competitive.

1. Introduction

Increased demand for cold chain products makes food logistics a vital issue for food security. China has a vast consumer population for agricultural production [1–3]. Because of the enormous demand for cold chain food, the total cold chain logistics demand reached 180 million tons in 2018. As a supply chain network connecting agricultural production and food consumption, a high-quality cold chain network can effectively transport food and raw materials between upstream companies and downstream consumers [4, 5]. Consequently, temperature-controlled supply chain (TCSC) networks that provide a series of equipment to keep food in ideal condition are adopted for food logistic and supply from production to consumption [6, 7]. TCSC has demonstrated exemplary performance in controlling food quality [8, 9], reducing food loss [10], improving health and

environment [9, 11], and promoting sustainability [2]. Furthermore, many innovative technologies [12], such as smart containers, are integrated into TCSC to make it play an efficient role in the food logistics system [13–15].

China's existing TCSC cannot meet its rapidly growing demand. Although the number of agricultural companies using TCSC has increased, there are still problems such as inadequate infrastructure [16, 17], uneven distribution of cold chain infrastructure [18, 19], and lack of integrated standards [20], which causes the enormous waste of fresh agricultural products [21–23]. Considering the efficient use of resources and sustainable development, the Chinese central government issued many policies and gave subsidies to promote the TCSC [24–27]. Yet, a large quantity of companies is still reluctant to invest in the construction of TCSC due to a lack of funds, professional management, and awareness of sustainable development [28, 29]. Therefore, it

is necessary to explore the reasons through economic and environmental analysis for improving comprehensive TCSC.

Previous literature mainly focuses on the construction and improvement of TCSC yet ignores the reasons why companies do not build it. Many researchers have studied the TCSC system and its optimism [8]. As a network guarantee to ensure the ideal state of food, TCSC involves coherent and complex procedures such as handling, packing, storage, and transportation [30, 31]. Furthermore, the researchers worked on optimizing the transportation route [32, 33], decision-making [34], minimizing the cost [35], reaching a supply chain network equilibrium [36], and innovative technologies [2]. Many researchers studied the behaviour in the network of upstream and downstream [37, 38]. However, the improvement of TCSC has not significantly prompted some Chinese companies to be willing to build it. Few studies in the past explained the reasons behind and explored the scenario that prompted TCSC construction.

This study aims to establish an economic model to explain why these companies do not invest and under what conditions they will do. First, we build an economic model to explain why the companies are not willing to invest in a comprehensive TCSC. Subsequently, a basic economic model with a supply chain network including producers, retailers, and consumers is adopted to generalise producers' and retailers' conditions to make the investment. Furthermore, we explore how government policies affect companies' decisions. Finally, we simulate how the companies can benefit from investment in TCSC.

2. Model Design

In this part, we simplify the TCSC into 3 layers (producers, retailers, and consumers) and 2 markets (wholesale markets where producers sell products to retailers and retail markets where retailers sell products to consumers) (Figure 1). There are millions of producers and retailers in agricultural product markets. In both markets, whether firms choose to

use the TCSC network could only affect the benefit of themselves and not the market price, so we assume both wholesale market and retail market are perfectly competitive.

2.1. Producers' Behaviour and Their Conditions of Investment.

Let q_i be the nonnegative production output of producer i , $c_i = c_i(q_i)$ be the production cost function of producer i , v_i be the loss rate of producers from producing to selling, and ρ_1 be the price of producers selling products to retailers. As the wholesale market is perfectly competitive, ρ_1 is constant, which means the behaviour of producers will not affect the market price.

The annual profit for producer i is

$$\rho_1 q_i (1 - v_i) - c_i(q_i). \quad (1)$$

Now, assume that producers are considering to invest in TCSC. TCSC has a similar structure but a different network. One reason is that producers with temperature-controlled equipment are quite different from others, so they cannot use the original chain. We use apostrophe to denote the variables with investment.

With TCSC, let q'_i be the production of producer i , $c'_{1i} = c'_{1i}(q'_i)$ be the production cost function of producer i , v'_i be the loss rate of producers from producing to selling, and ρ'_1 be the price of producers selling products to retailers. Also, in order to use TCSC, producers should invest a fixed cost, denoted as $f'_i(q'_i)$, and an additional variable cost, denoted as $c'_{2i} = c'_{2i}(q'_i)$. In addition, we have, $q'_i = q_i$, $c'_{1i} = c_i$, $v'_i < v_i$, and $\rho'_1 > \rho_1$.

Firms evaluate the profits with a time span of n years, and letting r be the discount rate, we could conclude producers' profit of no investment as

$$\pi_i = \sum_{t=0}^n \frac{\rho_1 q_i (1 - v_i) - c_i(q_i)}{(1+r)^t} = [\rho_1 q_i (1 - v_i) - c_i(q_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}}. \quad (2)$$

Also, producers' profit of investment is

$$\pi'_i = \sum_{t=0}^n \frac{\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)}{(1+r)^t} - f'_i(q'_i) = [\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_i(q'_i). \quad (3)$$

Producers will choose TCSC only when $\pi'_i \geq \pi_i$. As a result, we obtain the conditions of producers investing in TCSC:

$$[\rho'_1 q'_i (1 - v'_i) - \rho_1 q_i (1 - v_i)] - c'_{2i}(q'_i) \geq f'_i(q'_i) \frac{r(1+r)^{n-1}}{(1+r)^n - 1}. \quad (4)$$

2.2. Retailers' Behaviour and Their Conditions of Investment.

Compared with producers, retailers have a similar but a little more complex condition, since they should buy products from producers and be affected by producers' behaviours.

Let q_j be the amount of the products purchased by retailer j from producers, $c_j = c_j(q_j)$ be the cost function of retailer j , v_j be the loss rate of retailers from purchasing to selling, and ρ_2 be the price of retailers selling products to

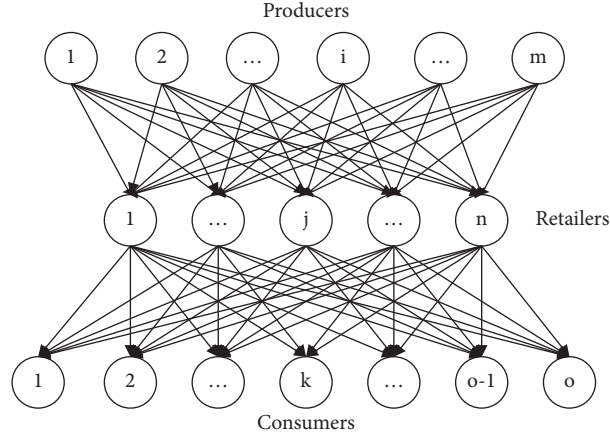


FIGURE 1: A simplified supply chain network including producers, retailers, and consumers.

consumers. Similarly, as the retail market is perfectly competitive, ρ_2 is also constant.

Besides operation cost, retailers should buy products from producers, so the annual profit for retailer j is

$$\rho_2 q_j (1 - v_j) - c_j(q_j) - \rho_1 q_j. \quad (5)$$

Considering investment on TCSC, let q'_j be the amount of the products purchased by retailer j from producers, $c_{1j}' =$

$c_{1j}'(q'_j)$ be the cost function of retailer j , v'_j be the loss rate of retailers from purchasing to selling, and ρ'_2 be the price of retailers selling products to consumers. If firms want to invest in new equipment and change into TCSC network, they should invest a fixed cost, denoted as $f'_j(q'_j)$, and an additional variable cost, denoted as $c_{2j}' = c_{2j}'(q'_j)$. We have $q'_j = q_j$, $c_{1j}' = c_j$, $v'_j < v_j$, and $\rho'_2 > \rho_2$.

The profit of retailers in n years without investment is

$$\pi_j = \sum_{r=0}^n \frac{\rho_2 q_j (1 - v_j) - \rho_1 q_j - c_{1j}(q_j)}{(1+r)^r} = [\rho_2 q_j (1 - v_j) - \rho_1 q_j - c_{1j}(q_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}}. \quad (6)$$

Also, the profit of retailers in n years with investment is

$$\pi'_j = \sum_{t=1}^n \frac{\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c_{1j}'(q'_j) - c_{2j}'(q'_j)}{(1+r)^t} - f'_j(q'_j) = [\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c_{1j}'(q'_j) - c_{2j}'(q'_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_j(q'_j). \quad (7)$$

Retailers will invest in TCSC only when $\pi'_j \geq \pi_j$, so we obtain the conditions of retailers investing in TCSC:

$$[\rho'_2 q'_j (1 - v'_j) - \rho_2 q_j (1 - v_j)] - [\rho'_1 q'_j - \rho_1 q_j] - c_{2j}'(q'_j) \geq f'_j(q'_j) \frac{r(1+r)^{n-1}}{(1+r)^n - 1}. \quad (8)$$

3. Government Policies and Firms' Behaviours

To encourage firms to use TCSC, both the central and local governments in China have issued many polices. For example, the central government announced that they will subsidise some firms with no more than 20 million yuan each, not exceeding 50% of total investment in 2020.

In this part, we will analyse how these policies will affect firms' behaviours.

3.1. One-Time Reward. Some local governments give rewards to firms for building cold supply chain networks, such as building cold storage and buying a freezer truck, after

finishing building. Let s denote the proportion of the reward of the fixed cost. Now, the profit of investment with one-time reward for producers is

$$\pi_i^{(s)} = [\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_i(q'_i) + s \frac{f'_i(q'_i)}{(1+r)^n}. \quad (9)$$

The profit of investment with one-time reward for retailers is

$$\pi_j^{(s)} = [\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c'_{1j}(q'_j) - c'_{2j}(q'_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_j(q'_j) + s \frac{f'_j(q'_j)}{(1+r)^n}. \quad (10)$$

Similarly, let $\pi_i^{(s)} > \pi_i$ and $\pi_j^{(s)} > \pi_j$, and we obtain the conditions of producers investing in TCSC with one-time reward:

$$[\rho'_1 q'_i (1 - v'_i) - \rho_1 q_i (1 - v_i)] - c'_{2i}(q'_i) \geq f'_i(q'_i) \frac{r(1+r)^{n-1} - sr(1+r)}{(1+r)^n - 1}. \quad (11)$$

Also, we obtain the conditions of retailers investing in TCSC with one-time reward:

$$[\rho'_2 q'_j (1 - v'_j) - \rho_2 q_j (1 - v_j)] - [\rho'_1 q'_j - \rho_1 q_j] - c'_{2j}(q'_j) \geq f'_j(q'_j) \frac{r(1+r)^{n-1} - sr(1+r)}{(1+r)^n - 1}. \quad (12)$$

3.2. Annual Subsidies. Some local governments give subsidy annually with a period lasting 3 or 5 years. Let a denote the proportion of annual subsidy and fixed investment and t_a

denote the time span for subsidy. Now, the profit of investment with annual subsidies for producers is

$$\pi_i^{(a)} = [\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_i(q'_i) + a f'_i(q'_i) \frac{(1+r)^{t_a} - 1}{r(1+r)^{t_a-1}}. \quad (13)$$

The profit of investment with annual subsidies for retailers is

$$\pi_j^{(a)} = [\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c'_{1j}(q'_j) - c'_{2j}(q'_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_j(q'_j) + a f'_j(q'_j) \frac{(1+r)^{t_a} - 1}{r(1+r)^{t_a-1}}. \quad (14)$$

Let $\pi_i^{(a)} > \pi_i$ and $\pi_j^{(a)} > \pi_j$, and we obtain the condition of producers investing in TCSC with governments' annual subsidies:

$$[\rho'_1 q'_i (1 - v'_i) - \rho_1 q_i (1 - v_i)] - c'_{2i}(q'_i) \geq f'_i(q'_i) \frac{r(1+r)^{n-1}}{(1+r)^n - 1} - a f'_i(q'_i) \frac{[(1+r)^{t_a} - 1] r(1+r)^{n-1}}{r(1+r)^{t_a-1} [(1+r)^n - 1]}. \quad (15)$$

Also, we obtain the condition of retailers investing in TCSC with governments' annual subsidies:

$$[\rho_2'q_j'(1-v_j') - \rho_2q_j(1-v_j)] - [\rho_1'q_j' - \rho_1q_j] - c_{2j}'(q_j') \geq f_j'(q_j) \frac{r(1+r)^{n-1}}{(1+r)^n - 1} - af_i'(q_i) \frac{[(1+r)^t - 1]r(1+r)^{n-1}}{r(1+r)^{t-1}[(1+r)^n - 1]} \quad (16)$$

4. Simulation and Analysis

To evaluate the model, we use simulation data to see how the results will change with different parameters. Table 1 gives the values of different parameters. Now, we evaluate the decisions of firms with different scales and different return period.

4.1. Decision-Making with Different Return Periods. At first, fixing firms' scale as constant, $q_i = q_i' = 50,000$ kg, and $q_j = q_j' = 300,000$ kg, the time is independent, and we have

$$\begin{aligned} \pi_i &= [\rho_1q_i(1-v_i) - c_i(q_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} = 55,000 * \frac{(1.04^n - 1)}{0.04 * 1.04^{n-1}}, \\ \pi_i' &= [\rho_1'q_i'(1-v_i') - c_{1i}'(q_i') - c_{2i}'(q_i')] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f_i'(q_i) = 110,000 * \frac{(1.04^n - 1)}{0.04 * 1.04^{n-1}} - 200,000, \\ \pi_j &= [\rho_2q_j(1-v_j) - \rho_1q_j - c_{1j}(q_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} = 300,000 * \frac{(1.04^n - 1)}{0.04 * 1.04^{n-1}}, \\ \pi_j' &= [\rho_2q_j'(1-v_j') - \rho_1q_j' - c_{1j}'(q_j') - c_{2j}'(q_j')] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f_j'(q_j') = 405,000 * \frac{(1.04^n - 1)}{0.04 * 1.04^{n-1}} - 300,000. \end{aligned} \quad (17)$$

Figures 2 and 3 demonstrate the profits changing with time for increasing of investment and no investment for producers and retailers. Let $\pi_i = \pi_i'$, and we can obtain the intersection $n = \log_{(26/25)}(143/123) \approx 3.84$ for producers and similarly $n = \log_{(26/25)}(91/81) \approx 2.97$ for retailers. It means that producers with a production of 50,000 kg/year will invest in TCSC if they hope to recover investment and get more profit within 4 or more years and retailers with a

production of 300,000 kg/year will invest in TCSC if they hope to recover investment and get more profit within 3 or more years.

4.2. Decision-Making with Different Scales for Fixed Return Period. Now, fixing time as constant $n = 6$, the firms' production is independent, and we have

$$\begin{aligned} \pi_i &= [\rho_1q_i(1-v_i) - c_i(q_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} \approx 6.00q_i, \\ \pi_i' &= [\rho_1'q_i'(1-v_i') - c_{1i}'(q_i') - c_{2i}'(q_i')] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f_i'(q_i) \approx 12.00q_i' - 200000, \\ \pi_j &= [\rho_2q_j(1-v_j) - \rho_1q_j - c_{1j}(q_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} \approx 5.45q_j, \\ \pi_j' &= [\rho_2'q_j'(1-v_j') - \rho_1'q_j' - c_{1j}'(q_j') - c_{2j}'(q_j')] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f_j'(q_j') \approx 7.36q_j' - 300000. \end{aligned} \quad (18)$$

TABLE 1: Data used to simulate the benefits change with variables of different values.

Variables	Value	Variables	Value
q_i	50,000 kg	q'_i	50,000 kg
v_i	0.15	v'_i	0.10
ρ_1	6 yuan/kg	ρ'_1	8 yuan/kg
c_i	4 yuan/kg	c'_{1i}	4 yuan/kg
f'_i	200,000 yuan	c'_{2i}	1 yuan/kg
q_j	300,000 kg	q'_j	300,000 kg
v_j	0.1	v'_j	0.05
ρ_2	10 yuan/kg	ρ'_2	13 yuan/kg
c_j	2 yuan/kg	c'_{1j}	2 yuan/kg
f'_j	300,000 yuan	c'_{2j}	1 yuan/kg
r	0.04		

Figures 4 and 5 demonstrate the profits changing with different firms' scales for producers and retailers of investment and no investment. Let $\pi_i = \pi'_i$, then we can obtain the intersection $q_i = q'_i \approx 33350$; let $\pi_j = \pi'_j$, then we can obtain $q_j = q'_j \approx 157221$. It means that if producers want to get more profits within 6 years, they should have a production bigger than 33350 kg/year and if retailers want to get

more profits within 6 years, they should have a production bigger than 157221 kg/year.

4.3. *One-Time Reward from the Government.* We use different values of s to calculate the changing situations of profits. Let $s = 0.1, 0.3, 0.5$. The profits of investment with one-time reward for producers and retailers are

$$\pi_i^{(s)} = [\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_i(q'_i) + s \frac{f'_i(q'_i)}{(1+r)^n},$$

$$\pi_j^{(s)} = [\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c'_{1j}(q'_j) - c'_{2j}(q'_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_j(q'_j) + s \frac{f'_j(q'_j)}{(1+r)^n}.$$

Figures 6 and 7 demonstrate the profits of producers and retailers with different rewards. For both producers and retailers, with increase in rewards, the profits increase and the intersections decrease. We can conclude that if governments give more rewards, more firms will choose to invest in TCSC.

4.4. *Annual Subsidies from the Government.* Let $a = 0.1$ and $t_a = 3, 4, 5$. Now, the profits of investment with annual subsidies for producers and retailers are

$$\pi_i^{(a)} = [\rho'_1 q'_i (1 - v'_i) - c'_{1i}(q'_i) - c'_{2i}(q'_i)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_i(q'_i) + a f'_i(q'_i) \frac{(1+r)^{t_a} - 1}{r(1+r)^{t_a-1}},$$

$$\pi_j^{(a)} = [\rho'_2 q'_j (1 - v'_j) - \rho'_1 q'_j - c'_{1j}(q'_j) - c'_{2j}(q'_j)] \frac{(1+r)^n - 1}{r(1+r)^{n-1}} - f'_j(q'_j) + a f'_j(q'_j) \frac{(1+r)^{t_a} - 1}{r(1+r)^{t_a-1}}.$$

Figures 8 and 9 show the profits of producers and retailers with the change of different subsidy period. Also, with more subsidies, both producers' and retailers' profits

increase and the intersections decrease. We can conclude that if governments give annual subsidies for a longer period, more firms will choose to invest in TCSC.

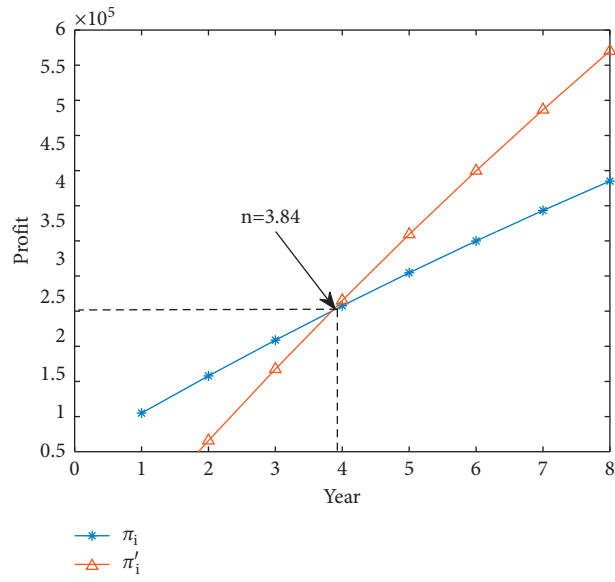


FIGURE 2: Profits of producers with $q_i = q'_i = 50,000$.

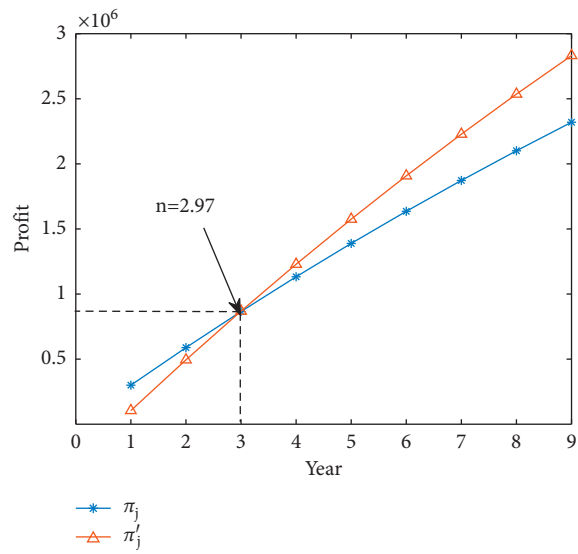


FIGURE 3: Profits of retailers with $q_j = q'_j = 300,000$.

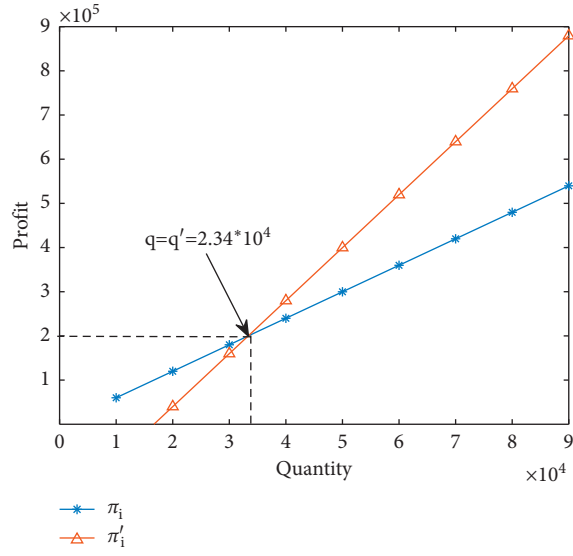


FIGURE 4: Profits of producers with different scales when $n = 6$.

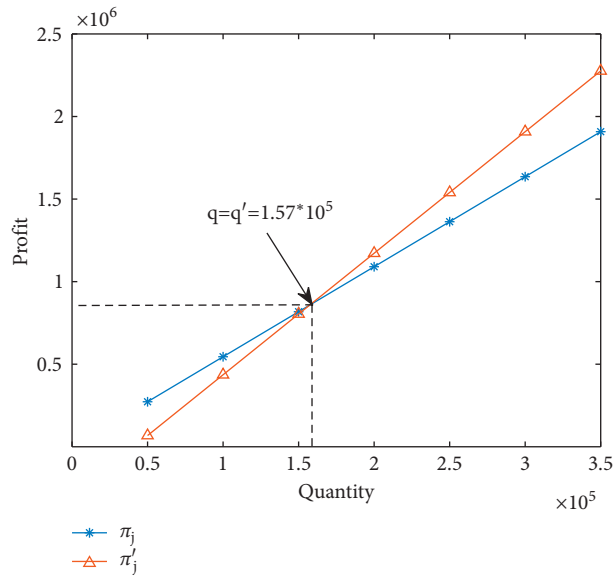


FIGURE 5: Profits of retailers with different scales when $n = 6$.

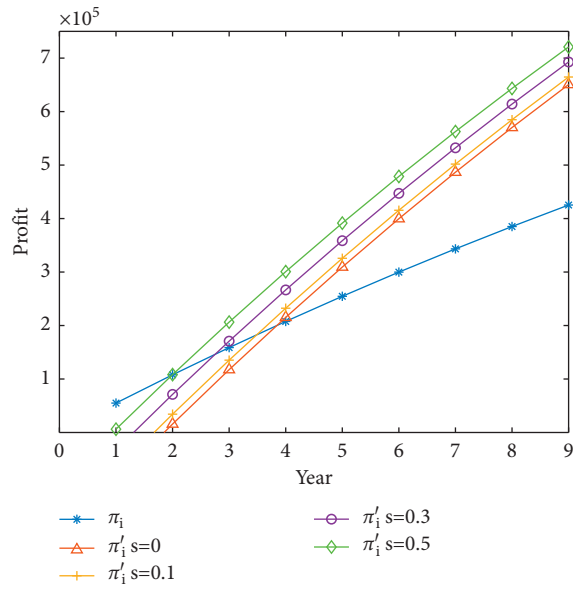


FIGURE 6: Profits of producers with $q_i = q'_i = 50,000$ and $s = 0, 0.1, 0.3, 0.5$.

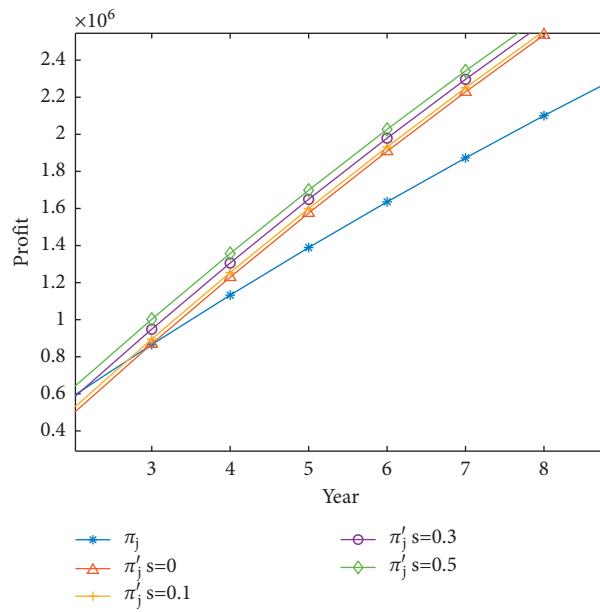


FIGURE 7: Profits of retailers with $q_j = q'_j = 300,000$ and $s = 0, 0.1, 0.3, 0.5$.

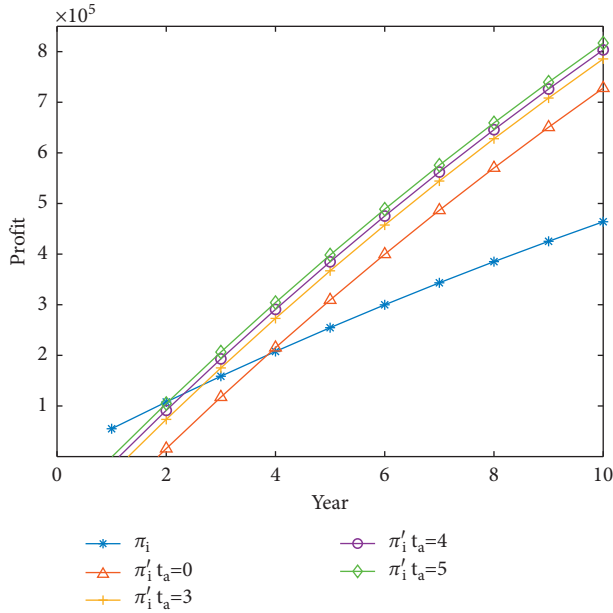


FIGURE 8: Profits of producers with $q_i = q'_i = 50,000$ and $t_a = 0, 3, 4, 5$.

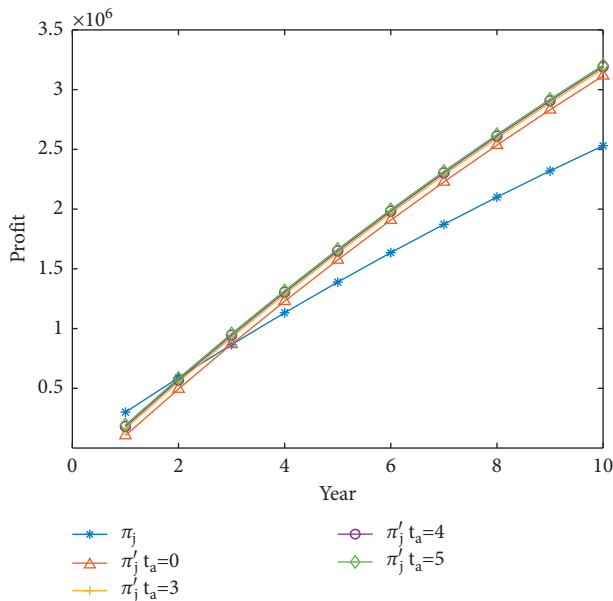


FIGURE 9: Profits of retailers with $q_j = q'_j = 300,000$ and $t_a = 0, 3, 4, 5$.

5. Conclusions

The TCSC is a necessary part for logistics to deliver food from producers to consumers. Besides improving the technologies of TCSC, studying firms' behaviours is also of great meaning. This paper has built an economic model to give explanations about why and when firms will choose to invest in TCSC. In general, firms with greater production tend to invest more in new equipment to get additional benefits. This will result in market

segmentation, where bigger firms running with high-quality products get more profit and smaller firms running with normal products are less competitive relatively.

There are still some parts that need to be improved. At first, we assume that both wholesale market and retail market are perfect competitive. With more and more firms choose to run with TCSC, the market price will be affected, and the assumption that both market prices are constant is too strict in a long run. Second, when considering the time span, we think prices will not change in different years, but food prices are always fluctuating. In some years, for some varieties, prices may even fluctuate heavily. The evaluation will become more complicated if we take this into account. Third, while doing the simulation, only some numbers are used to assess the model. It would be better if we do the simulation with a specific variety such as banana or cabbage.

Firms' behaviours are not affected only by economic benefits. In fact, every player in the supply chain networks will have influence. When considering the behaviours of competitors and collaborators, producers and retailers will be faced with a more complex situation. How to maximize their profits needs further discussion. Also, the Chinese government has issued policies supporting food cold chain logistics. Different policies give firms more choices, and how to maximize their profits will need more research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (71703159).

References

- [1] G. Cui and Q. Liu, "Regional market segments of China: opportunities and barriers in a big emerging market," *Journal of Consumer Marketing*, vol. 17, no. 1, pp. 55–72, 2000.
- [2] M. C. Dodd and J. J. Bouwer, *The Supply Value Chain of Fresh Produce from Field to Home*, Elsevier, Amsterdam, Netherlands, pp. 449–483, 2014.
- [3] B. Ghose, "Food security and food self-sufficiency in China: from past to 2050," *Food and Energy Security*, vol. 3, no. 2, pp. 86–95, 2014.
- [4] S. J. James and C. James, "The food cold-chain and climate change," *Food Research International*, vol. 43, no. 7, pp. 1944–1956, 2010.
- [5] S. Guohua, "Research on the fresh agricultural product supply chain coordination with supply disruptions," *Discrete Dynamics in Nature and Society*, vol. 2013, Article ID 416790, 9 pages, 2013.

- [6] D. Smith and L. Sparks, "Temperature controlled supply chains," *Food supply chain management*, vol. 1, pp. 179–198, 2004.
- [7] B. Behdani, Y. Fan, and J. M. Bloemhof, *Cool Chain and Temperature-Controlled Transport: An Overview of Concepts, Challenges, and Technologies*, Elsevier, Amsterdam, Netherlands, pp. 167–183, 2019.
- [8] A. Rong, R. Akkerman, and M. Grunow, "An optimization approach for managing fresh food quality throughout the supply chain," *International Journal of Production Economics*, vol. 131, no. 1, pp. 421–429, 2011.
- [9] M. M. Aung and Y. S. Chang, "Temperature management for the quality assurance of a perishable food supply chain," *Food Control*, vol. 40, pp. 198–207, 2014.
- [10] R. Jedermann, M. Nicometo, I. Uysal, and W. Lang, *Reducing Food Losses by Intelligent Food Logistics*, The Royal Society Publishing, London, UK, 2014.
- [11] D. Coulomb, "Refrigeration and cold chain serving the global food industry and creating a better future: two key IIR challenges for improved health and environment," *Trends in Food Science & Technology*, vol. 19, no. 8, pp. 413–417, 2008.
- [12] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, pp. 3804–3816, 2020.
- [13] R. Jedermann, L. Ruiz-Garcia, and W. Lang, "Spatial temperature profiling by semi-passive RFID loggers for perishable food transportation," *Computers and Electronics in Agriculture*, vol. 65, no. 2, pp. 145–154, 2009.
- [14] M. Lütjen, P. Dittmer, and M. Veigt, "Quality driven distribution of intelligent containers in cold chain logistics networks," *Production Engineering*, vol. 7, no. 2-3, pp. 291–297, 2013.
- [15] P. Vrat, R. Gupta, A. Bhatnagar, D. K. Pathak, and V. Fulzele, "Literature review analytics (LRA) on sustainable cold-chain for perishable food products: research trends and future directions," *Opsearch*, vol. 55, no. 3-4, pp. 601–627, 2018.
- [16] S. Negi and N. Anand, "Issues and challenges in the supply chain of fruits & vegetables sector in India: a review," *International Journal of Managing Value and Supply Chains*, vol. 6, no. 2, pp. 47–62, 2015.
- [17] K. Y. Wang and T. L. Yip, "Cold-chain systems in China and value-chain analysis," in *Finance and Risk Management for International Logistics and the Supply Chain*, pp. 217–241, Elsevier, Amsterdam, Netherlands, 2018.
- [18] K.-M. Tsai and K. Pawar, "Special issue on next-generation cold supply chain management: research, applications and challenges," *International Journal of Logistics Management*, vol. 29, no. 3, pp. 786–791, 2018.
- [19] Y. Dong, M. Xu, and S. A. Miller, "Overview of cold chain development in China and methods of studying its environmental impacts," *Environmental Research Communications*, vol. 2, Article ID 122002, 2021.
- [20] W. Qing-gang, "The current situation and the countermeasures of China's cold chain logistics development," *China Business and Market*, vol. 2, 2011.
- [21] F. Vanek and Y. Sun, "Transportation versus perishability in life cycle energy consumption: a case study of the temperature-controlled food product supply chain," *Transportation Research Part D: Transport and Environment*, vol. 13, no. 6, pp. 383–391, 2008.
- [22] K. Paritosh, S. K. Kushwaha, M. Yadav, N. Pareek, A. Chawade, and V. Vivekanand, "Food waste to energy: an overview of sustainable approaches for food waste management and nutrient recycling," *BioMed Research International*, vol. 2017, Article ID 2370927, 2017.
- [23] D. Dai, X. Wu, and F. Si, "Complexity analysis and control in time-delay vaccine supply chain considering cold chain transportation," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4392708, 15 pages, 2020.
- [24] D. B. Grant, C. Y. Wong, and A. Trautrim, *Sustainable Logistics and Supply Chain Management: Principles and Practices for Sustainable Operations and Management*, Kogan Page Publishers, London, UK, 2017.
- [25] J. Fernie and L. Sparks, *Logistics and Retail Management: Emerging Issues and New Challenges in the Retail Supply Chain*, Kogan page publishers, London, UK, 2018.
- [26] M. A. Cohen and H. L. Lee, "Designing the right global supply chain network," *Manufacturing & Service Operations Management*, vol. 22, no. 1, pp. 15–24, 2020.
- [27] N.-R. Xu and Z.-Q. Cai, "Research on the mechanism of cold chain logistics subsidy," *Journal of Chemistry*, vol. 2020, Article ID 4565094, 11 pages, 2020.
- [28] Y. He, H. Huang, D. Li, C. Shi, and S. J. Wu, "Quality and operations management in food supply chains: a literature review," *Journal of Food Quality*, vol. 2018, Article ID 7279491, 14 pages, 2018.
- [29] S. Liu and C. Zhang, "Optimization of urban cold chain transport routes under time-varying network conditions," *Journal of Advanced Transportation*, vol. 2021, Article ID 8817991, 16 pages, 2021.
- [30] H. R. El-Ramady, É. Domokos-Szabolcsy, N. A. Abdalla, H. S. Taha, and M. Fári, *Postharvest Management of Fruits and Vegetables Storage*, Springer International Publishing, New York, NY, USA, pp. 65–152, 2015.
- [31] Y. Wang, J. Yi, X. Zhu, J. Luo, and B. Ji, "Developing an ontology-based cold chain logistics monitoring and decision system," *Journal of Sensors*, vol. 2015, Article ID 231706, 8 pages, 2015.
- [32] Z. Zhao, X. Li, and X. Zhou, "Distribution route optimization for electric vehicles in urban cold chain logistics for fresh products under time-varying traffic conditions," *Mathematical Problems in Engineering*, vol. 2020, Article ID 9864935, 7 pages, 2020.
- [33] H. Xiong, "Research on cold chain logistics distribution route based on ant colony optimization algorithm," *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 752830, 6 pages, 2021.
- [34] A. Chaudhuri, I. Dukovska-Popovska, N. Subramanian, H. K. Chan, and R. Bai, "Decision-making in cold chain logistics using data analytics: a literature review," *International Journal of Logistics Management*, vol. 29, no. 3, pp. 839–861, 2018.
- [35] D. Nakandala, H. Lau, and J. Zhang, "Cost-optimization modelling for fresh food quality and transportation," *Industrial Management & Data Systems*, vol. 116, no. 3, pp. 564–583, 2016.
- [36] A. Nagurney, J. Dong, and D. Zhang, "A supply chain network equilibrium model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 38, no. 5, pp. 281–303, 2002.
- [37] F. Xiong and Y. Liu, "Opinion formation on social media: an empirical approach," *Chaos*, vol. 24, Article ID 013130, 2014.
- [38] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.

Research Article

EMM-CLODS: An Effective Microcluster and Minimal Pruning Clustering-Based Technique for Detecting Outliers in Data Streams

Mohamed Jaward Bah ¹, Hongzhi Wang ², Li-Hui Zhao ³, Ji Zhang ⁴, and Jie Xiao⁵

¹Zhejiang Lab, Hangzhou, China

²Harbin Institute of Technology, Harbin, China

³North University of China, Taiyuan, China

⁴University of Southern Queensland, Toowoomba, Australia

⁵Hangzhou Yugu Technology Co., Ltd., Hangzhou, China

Correspondence should be addressed to Ji Zhang; zhangji77@gmail.com

Received 7 July 2021; Revised 10 August 2021; Accepted 23 August 2021; Published 13 September 2021

Academic Editor: Fei Xiong

Copyright © 2021 Mohamed Jaward Bah 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.

Detecting outliers in data streams is a challenging problem since, in a data stream scenario, scanning the data multiple times is unfeasible, and the incoming streaming data keep evolving. Over the years, a common approach to outlier detection is using clustering-based methods, but these methods have inherent challenges and drawbacks. These include to effectively cluster sparse data points which has to do with the quality of clustering methods, dealing with continuous fast-incoming data streams, high memory and time consumption, and lack of high outlier detection accuracy. This paper aims at proposing an effective clustering-based approach to detect outliers in evolving data streams. We propose a new method called Effective Microcluster and Minimal pruning Clustering-based method for Outlier detection in Data Streams (EMM-CLODS). It is a clustering-based outlier detection approach that detects outliers in evolving data streams by first applying microclustering technique to cluster dense data points and effectively handle objects within a sliding window according to the relevance of their status to their respective neighbors or position. The analysis from our experimental studies on both synthetic and real-world datasets shows that the technique performs well with minimal memory and time consumption when compared to the other baseline algorithms, making it a very promising technique in dealing with outlier detection problems in data streams.

1. Introduction

In the current era, the need to detect abnormal behavior to reveal salient facts, observations, and realizing accurate predictions of data is extremely significant. Detecting outliers is one such important data mining task that aims at detecting objects that deviate from the expected pattern of the normal data. The process of detecting outliers is challenging due to the advancement in the digital age. For instance, with the revolution of data from traditional batch data, we have witnessed the advent of a large volume of data that is generated continuously at high speed and dynamically. These kinds of data are known as data streams and are

generated by many applications [1–3]. In contrast to traditional datasets, because of the nature of the data, it is not feasible to save in memory the whole data stream or run the data through multiple scans. This is because the data are massive and unbounded, have a varying rate, and continue to evolve.

A significant number of approaches have been proposed to detect outliers in data streams [8–11]. Among the different categories of proposed outlier detection methods, clustering-based approaches have shown to be popular in static data but yet one of the most challenging to adopt for outlier detection tasks in data streams. Although they have shown to be efficient for some outlier detection tasks, they lead to low

computational cost and high scalability in high-dimensional data [5, 12]. However, most of the prevailing data stream clustering approaches suffer from different drawbacks. They can be improved when we consider the spectrum of effectiveness and efficiency, for instance, to deal with the continuous fast-incoming data streams, higher computational demand in terms of its memory and time, the cluster quality, and the outlier detection rate. The process of clustering and detecting outliers in data streams is complicating since the clustering techniques often involve several parameters and operate in low- and high-dimensional spaces, constrained with excessive distance-based computation of object neighbors, noise, and so on. For this reason, clustering-based approach has varying performance for different application domains and data types. It is therefore imperative to design an effective method that will holistically address the issues and produce stable performance in detecting the outliers.

In spite of clustering's occasional challenges and caveats, it is still another good alternative and promising solution for detecting outliers. The advantage of clustering is that it allows for the use of limited amounts of time and memory, which is necessary when processing data streams. This is because clustering is the act of grouping elements using sets that provide the capability of grouping items that are similar to each other that curbs the need of redundant processing and over calculations. Clustering methods offer online and offline process support, which is usually used for data stream applications and is also flexible in adapting to the evolving nature of the data.

In this paper, we propose a new microclustering and minimal pruning clustering-based unsupervised outlier detection scheme to detect outliers in data streams while simultaneously addressing the mentioned challenges. The proposed approach involves different stages to adapt to the dynamic changes of data distribution that aims at eliminating the limitations of previously proposed methods. The newly propose method is called Effective Microcluster and Minimal pruning CLustering-based method for Outlier detection in Data Streams (EMM-CLODS), which is a clustering-based outlier detection approach. We call it CLODS for short and use this abbreviation instead of EMM-CLODS throughout the paper. It detects outliers from evolving data streams by first applying the microclustering technique to cluster dense data points. It then effectively handles objects within a sliding window according to the relevance of their status to their respective neighbors or position through minimal pruning technique.

In our data stream scenario, where the size of the dataset is potentially boundless, we process the data over a fixed period to reduce the complexity of the outlier detection task. When new incoming data points arrive, the microcluster technique is applied, which identifies objects that are more analogous to each other and that meet the fundamental prerequisite of the clustering methods. The methods scan the data once and adapt to the time changes as the streaming data evolve. It constantly and periodically updates incoming data, and the results are obtained. Finally, the CLODS reports key insights from these results to determine whether they are outliers or inliers. The advantages of the technique

are that it can effectively save time and memory, thanks to the microclustering technique and minimal pruning. It removes the need to compute every data point in and out of the cluster and store every data point in memory. In summary, the major contributions of this work are as follows:

- (i) We propose the CLODS, a new technique based on microclustering and minimal pruning of data points outside the clusters, to solve the problem of detecting outliers in continuous evolving data streams.
- (ii) We propose the concept of priority handling of evolving objects outside the clusters to minimize the memory and time consumption during the updating phase according to the relevance of their status to their respective neighbors or position.
- (iii) Our propose method can effectively optimize and solve the problems and challenges of time and memory constraints while maintaining its accuracy for detecting outliers in data streams.
- (iv) We demonstrate through an extensive experiment on some benchmark datasets the effectiveness of our method against some other methods used for the outlier detection process in data streams.

The rest of the paper is organized as follows: in Sections 2 and 3, we present the related work and problem formulation, respectively. In Section 4, we present in details the method we propose. In Section 5, we present the experimental studies including the results and discussion. Finally, in Section 6, we present the conclusion of the paper.

2. Related Work

Detecting outliers is a well-known domain in the data mining community, and it has been applied in a wide range of application areas [13, 14] and other domains such as community detection [15, 16]. It has been studied extensively [17–19]. In a recent survey [11], we classified outlier detection methods into diverse categories and have proposed effective methods among these categories to detect outliers in data streams [8, 11]. In progress to this study series, the clustering-based category has open research gaps and challenges. Proposing solutions and improving these methods will greatly contribute to the general body of outlier detection methods.

The clustering approach is an unsupervised data mining method that groups similar dense data points. Several methods using clustering techniques and its variant approaches have been proposed for outlier detection tasks. However, some earlier proposed clustering methods suffer from drawbacks such as the buffering of all data points in memory for future handling or, in some cases, not considering data points that often leads to poor clustering. There are a significant number of these methods concentrated on both static data and streaming data types [20, 21]. These methods mostly adopt the two-phase scheme: the online and offline phase. The majority of the earlier proposed method for stream data clustering deals with static

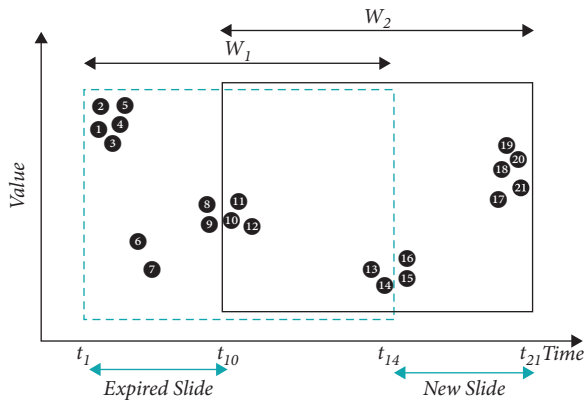


FIGURE 1: Streaming data.

clustering that is in a continuous form. One shortcoming of this kind of approach is that recent and outdated data are handled the same way. Several moving window models are proposed to solve this issue. For evolving data streams, Toshniwal and Yokita et al. [20] proposed a framework using simple k -means and the attribute weight to detect outliers, while Cao et al. [22] proposed a technique related to density-based clustering for evolving data streams. In their method, the incoming data are selected depending on the distance between their centers to either the outlier or potential core microcluster. In this case, with an increasing number of outliers, the clustering accuracy becomes a problem. Therefore, Liu et al. [23] proposed a new technique to address this drawback. Although they tried to address the issue, it comes at a high computational cost. To salvage the computational cost and improve the clustering and outlier detection accuracy, Kumar and Sharma [24] applied a technique that extracts the boundary points in the overlapped microclusters. Many other clustering techniques have been proposed for outlier detection processes, such as density-based microclustering [22, 25], grid-based clustering [6, 26], and partitioning algorithm for data streams [12, 21]. However, since this is a short paper, Table 1 briefly outlines some of these techniques in comparison to our method in terms of the summarization technique, evolving data model and outlier detection method.

Remarkably, from Table 1, no two methods share the same approach. Our work is the first to use microclustering in the sliding window model using outlier microcluster to handle continuously evolving objects with changing features. For a more comprehensive related work to clustering techniques for outlier detection, we recommend Wang et al. [11] survey paper.

3. Preliminaries and Problem Formulation

3.1. Notations and Definitions. The key symbols used in this paper include but not limited to the following in Table 2.

3.2. Definition of Key Terms

3.2.1. Outliers. For a dataset D of n points, $D = [d_1, d_2, \dots, d_n]$. Whenever the data point d_i or an entire set of data points d_1, d_2, \dots, d_n deviates drastically from these other sets, these points are considered outliers.

3.2.2. Neighbor. In the case of two data points d_i and d_n , a data point d_i is considered a neighbor of d_n if the distance between the two does not exceed the distance threshold value $R > 0$. In other words, if d_i is not further than R from d_n , then it is a neighbor of d_n . A data point d cannot be a neighbor of itself.

3.2.3. Sliding Window. In sliding window, the time-based window and the count-based window are two types of window models commonly used for data streams. The former takes into consideration the data points within the time interval of two identify data points, for instance, at point x and y , with t_x and t_y . The latter thus considers the count of the data points within a specified window size.

3.2.4. Microclusters. A microcluster is formed when a data point has a radius of $R/2$ from the center, and in a microcluster, the distance between two data points, let us assume d_1 and d_2 , should not exceed R .

The function of the microcluster in our technique is as follows: we applied the microclusters to minimize the range queries and minimize the distance-based computations. The microclusters eliminate the need for excessive range queries by storing the neighbor's data points in the microclusters. This, therefore, improves the underlying evaluation metrics: the memory and time consumption. The microclusters adopted in the proposed methods give the advantage of eliminating the need for range queries and in curbing the distance computations. In addition to only storing crucial inliers in memory, the microclusters also improve the memory constraints, since a single microcluster has the ability to obtain the neighborhood information of each object in the same cluster.

In Figure 1, we can see that $W_1 = t_1 - t_{14}$ and $W_2 = t_{10} - t_{21}$, where W_2 is the current window and W_1 is the expired window. The fast-incoming data points (dp) from 1 to 23 are the data streams. By definition, the data stream is an unlimited number of data points within a specific timestamp or unbounded sequence. That is, the data stream $i = S_i | 0 \leq t$, with $t = \text{time}$ and dp, $S_{i=1,2,n} = S_1, S_2, S_3, \dots, S_n$. Each dp within its window could have a neighbor or not, but it cannot be a neighbor on its own. The neighbor of any particular data point S_i must not exceed the required distance threshold R , from each other. For instance, in Figure 1, dp 1, 2, 4, 5 are neighbors of 3, while 17, 18, 20, 21 are neighbors of 19. The neighbors play a crucial role in the overall outlier detection process; therefore, we pay special attention to them.

In W_2 , or when the window slides, determining whether a data point is an outlier or inlier can create additional constraints due to the evolving nature of the data points. Some neighbors will expire, such as dp 8, 9 among 8 – 12, and become obsolete when the window slides. In the different window stages, the question of how to perform clustering, how to use minimal pruning to get the most significant data points, how to deal with incoming and expired dp, and what kind of clustering technique to apply comes up, and also, what requirements should the clustering technique meet to ensure that (1) the clusters capture more

TABLE 1: Some key clustering algorithms.

Method	Summarization technique	Evolving data model	Outlier detection
CluStream [4]	Microcluster	Tilted-time window	—
D-Stream [5]	Grid	Fading window	Sporadic grid
DenStream [6]	Microcluster	Fading window	Outlier microcluster
DENGRIS-Stream [7]	Grid	Sliding window	Sparse grid
Ours-CLODS	Microcluster	Sliding window	Outlier microcluster

TABLE 2: List of symbols with their interpretations.

Symbols	Interpretations
d_i	i -th data point, $i = 1, \dots, n$
R	Distance threshold
K	Number of neighbors
W	Window size
S	Window slide size
∞	Data streams
t_i	The specific time
d_{ci}	Data points in the current window
d_{ei}	Expired data points
O_d	Detected outlier/s

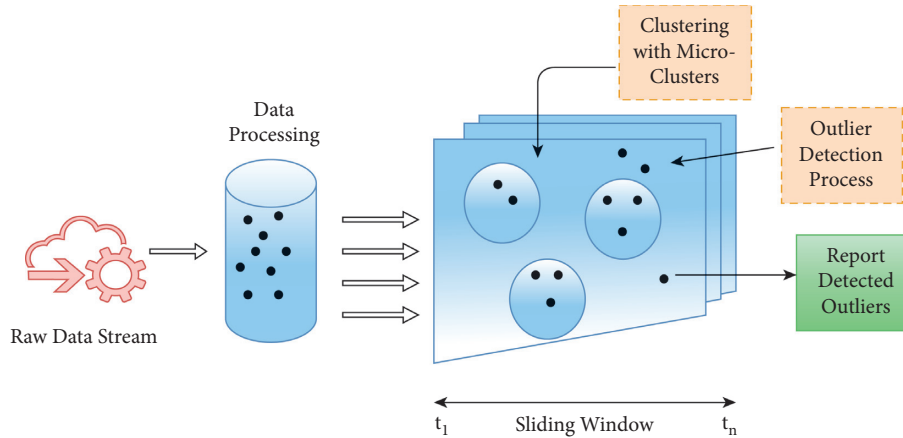


FIGURE 2: The framework of CLODS

dp and (2) the inliers or outliers are detected correctly and computed with the lowest computational cost possible.

3.3. Problem Formulation. Problem statement: the major goal of this paper is to present an improved solution to address the problem of effectively clustering and detecting outliers in fast-evolving data streams.

For new data streams arriving continuously, $S = \{S_t\}_{t=1,2,\dots}$, with dimensionality d at time t , and with evolving feature changes as the data speed increases, we need to design a robust approach that will deal with the evolving data streams by clustering incoming data streams effectively and simultaneously detect all outliers in the shortest conceivable time, with low memory usage, while maintaining high detection accuracy. Also, we handle data points outside the clusters while dealing with the fading of old clusters, new and expired data points, and detecting the outliers. The key challenge is that the actively evolving data point position continues to change due

to either the window slides or the arrival and expiration of some data points. This ultimately makes it complicating in addressing the overall problem. It will be a challenging task to process and remove data points one at a time as they arrive over the stream. It will incur a lot of time.

In addition, managing memory space presents another challenge since it is not possible to predict how many data points arrive and expire a priori. It becomes challenging to cluster essential data points and dynamically allocate space for the growing number of unknown data points that arrive and expire.

This brings us to the essential problem statement and question we address in this paper, how do we capture the data points that deviate from the others in streaming data which evolve as time progresses with these additional constraints:

- (i) The data point features might change over time.
- (ii) Prior unseen data point features might arrive over time.

4. The Proposed Methodology

4.1. Fundamentals of the Proposed Method. As data originate from their source in the form of fast continuous evolving data streams, they become challenging to cluster data points and effectively detect the outliers, as explained in the problem statement. There is a need for special attention on the clustering method and in handling both the inliers and outliers in this scenario. To do this, we propose a new framework, which involves different stages in order to detect the outliers efficiently while maintaining high accuracy. The newly proposed method called Effective Microcluster and Minimal pruning CLustering-based method for Outlier detection in Data Streams (EMM-CLOUDS) is a kind of clustering-based outlier detection approach that detects outliers in evolving data streams using microcluster and minimal pruning. This is done by first applying a microclustering technique to cluster dense data points and effectively handle the data points according to the relevance of their status to their respective neighbors or position in the window. We adopt the sliding window model, and within this model, the microclustering technique helps to cluster dense data points quickly and eliminate the need for a range query search. For the data points outside the clusters, an approximate probing is implemented by excluding a set of inliers whose significance in the computation is trivial in order to reduce the computation demand.

The CLOUDS makes use of both clustering and approximate probing of data points within the adopted sliding window model and minimal pruning of data points outside the clusters. It simultaneously discovers the outliers and deals with potential outliers outside the clusters, even when they continuously evolve as the data point changes state. In contrast to other conventional clustering-based approaches, it does not limit itself to detecting outliers in static data [2, 11, 27], and for those that support data streams, the clustering procedure is different [12, 20, 28, 29], or they are not clustering-based approaches [4, 8, 30]. Those with similar clustering techniques to ours use a different scheme to deal with data points within the window or adopt different window models [12, 27, 29]. Furthermore, the handling procedure of data points outside the microclusters is different. Unlike some of these methods [12, 20, 27, 28] that deal with every data point outside the microclusters equally, we focus especially on the relevance of data points with respect to its neighbors and position to determine its overall role in the outlier detection process. This is to ensure we identify potential outliers rather than data points that might be falsely labeled as outliers. This consequently saves time and memory constraints without a performance decline.

4.2. The Proposed Framework. Figure 2 shows an illustrative representation of the proposed framework. At the onset, objects in the form of data streams arrive continuously and in an unprecedented manner. We first filter the data through data processing to determine its characteristics. Then, we process the preprocessed data in the sliding window model. During a specified period in the sliding window, we apply

probing and clustering process together with pruning the data points outside the clusters and detect the outliers. During this phase, additional processing such as handling of crucial inliers and potential outliers, and handling of both active and expired data points as the window slides is done. In the final stage, the detected outliers are then reported.

Algorithm 1 gives the overall framework of CLOUDS, with line 3–5 depicting the processes. In Algorithms 2–4, details of algorithmic process are given to understand the whole CLOUDS algorithm. In Algorithm 5, we extend details of the different steps in Algorithm 1. In the first part, we perform preprocessing. The preprocessed data stream is then computed in the next stage. In processing data points within the window, in line 4 we determine whether they belong to a cluster. If not in a cluster, the relevance of their status with respect to the other members is checked in line 9. The data points outside the clusters and that are not relevant to their respective members can be applied to the function in the last stage and reported as an outlier as can be seen in line 11.

In Algorithm 2, the processing of new data points in the new sliding window is shown. We first discover the cluster and if there is a data point d_p within the cluster, we add the new data point or else initiate a new cluster accordingly (line 2–6), while in Algorithm 3, it shows the processing of the expired data. Similarly, as in 3, we first discover the cluster and if a data point is found in the cluster, we ensure that we check the d_p 's relevance status to the other data points before we add it into the cluster (line 4–5). If not, we try to remove it (line 7).

Lastly in Algorithm 4, we process and report the detected outliers. We first initialize the count (line 1), and if d_p is not in any cluster and less number of neighbors to form a cluster, it is returned as an outlier. If it has already expired, it is then removed from data points outside the microclusters.

4.3. The Data Stream Stage. In a data stream model, the input data are not accessible through random disk or memory, such as in the case of static data or batch data in standard databases, but rather arrive in the form of one or more continuous data streams. A data stream is an unlimited number of sequence data points $\infty_i = S_i | 0 \leq t$, within a specific timestamp or unbounded sequence with data points, $S_i = S_1, S_2, S_3, \dots, S_n$. They are infinite series of data points, S_{t-2}, S_{t-1}, S_t , observed at a particular time t . The streaming data have the following characteristics:

- (i) The data points of streaming data arrive incrementally in real-time. The streaming data are active since all inbound objects/items trigger actions on the data rather than being invited to participate.
- (ii) The system has no control over the order or sequence in which the items of the streaming data arrive.
- (iii) The streaming data have the possibility of unbounded numbers of data points.

The problem of detecting or mining outliers in such data with the abovementioned characteristics brings a number of significant implications. Firstly, to ensure that the results are continuously up-to-date, it is essential to analyze the incoming data within the shortest time and minimal memory

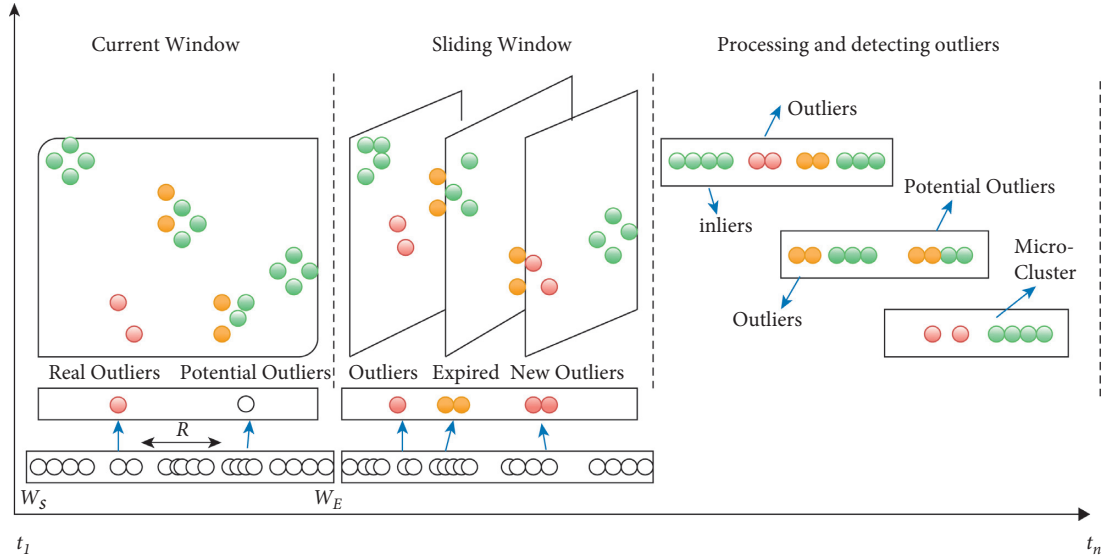


FIGURE 3: The different phases of processing the outliers in the sliding window.

Input: Preprocess Data Stream ∞ , Data point d_p , Parameters: {distance-threshold R , nearest-neighbor count K , sliding size S , Window Size W .}

Output: Outliers in sliding window

- (1) **Procedure:**
- (2) **While** the window slide or in $W_c \triangleright$ between period W_{start} to W_{end} when S arrives
- (3) Deal with data within W_c
- (4) Deal with new d_p, S and W
- (5) Deal with expired d_{ei}, S and W .
- (6) Report outliers, O_d
- (7) **end**

ALGORITHM 1: The CLODS algorithm.

usage. In the framework in Figure 3, the continuous infinite series of data points observed at a particular time t_1 is fed to the next stage.

4.4. Data Preprocessing Stage. As the incoming unbounded sequence of data arrives, it is impossible to store the entire data stream. Besides, to apply the clustering technique without taking note of the characteristics of the data makes the overall process more tedious. Therefore, we initially did some preprocessing based on the nature of the data to avoid assumptions about having clean and well-structured data and to tailor the data for our propose model. For instance, real-world datasets are highly susceptible to missing and inconsistent data. Such datasets may give rise to data quality issues, which in turn affects the overall result. During the data preprocessing and wrangling phase, we deal with the missing data and inconsistent data. Although outliers sometimes can influence the quality of the data, in this work we entirely avoid dealing with outliers since our primary goal is to detect outliers. For the missing data, we ensure that we ignore, fill manually, and compute values. For inconsistent data, we normalize the necessary datasets.

4.5. Sliding Window-Based Outlier Detection Stage. In this phase, we manage the evolving data streams; that is, we implement the CLODS and detect data points that deviate from their expected normal behavior when the window slides and expires, also when the data points will expire. We notice that it is not feasible to perform clustering on data streams during the all probable time. We handle the data points at different time windows. The process of exploring the evolving data stream during the different time windows provides the users with additional insights into the evolving nature and performance of the clusters. In terms of processing evolving data streams, different algorithms have adopted different window models. Some existing window models include the damped window model also known as the fading window model, the landmark window model, the tilted-time window model, and the sliding window model. In this paper, we use the sliding window model, in which the data are processed before the end of the streaming data window. This is as opposed to the landmark window model, which is adopted for cases where we want to mine the whole data stream history. It is suitable for static data settings. In the sliding window, the streaming data are considered from the current time to a certain range in its history. The key idea

```

(1) for  $d_p$  in new slide,  $S$  do
(2)  $c = \text{discoverCluster}$ 
(3) if  $d_p$  in  $C$  then
(4)  $c.\text{add}(d_p)$ 
(5) else
(6)  $\text{InitiateNewCluster}(d_p)$ 
(7) else if
(8) end for

```

ALGORITHM 2: *Process new data in the new slide window.

```

(1) for  $d_p$  in expired slide,  $S$  do
(2)  $c = \text{discoverCluster}$ 
(3) if  $d_p$  in  $C$  then
(4)  $\text{CheckRelevance}(d_p)$ 
(5)  $c.\text{add}(d_p)$ 
(6) else
(7)  $\text{remove}(d_p)$ 
(8) end If
(9) end for

```

ALGORITHM 3: *Process expired data point when slide expires.

```

(1)  $\text{Initiate outliers} = [ ]$ 
(2) Perform all functions
(3) for  $d_p$  in  $W$ ,  $S$  do
(4) if  $d_p$  cannot form a new cluster
(5)  $\text{add.Outlier}(d_p)$ 
(6) else
(7)  $\text{Processfunctions}$ 
(8) end if
(9) end for
(10) Return Outliers

```

ALGORITHM 4: *Process outlier W .

Input: Data Stream ∞ , Data point d_p , Parameters: {distance-threshold R , nearest-neighbor count K , sliding size S , window size W .}

Output: Outliers

```

(1) Procedure:  $\triangleright$  Preprocessing
(2) Perform Preprocessing  $\triangleright$  Process DataIn  $W_c$ 
(3) for for each  $d_p$  of preprocessed data in  $W$  do
(4)  $\text{DiscoverInClusters}$ 
(5) If  $d_p \geq k + 1$  neighbor then
(6)  $\text{InCluster}$ 
(7) elseif
(8)  $\text{NotIncluster}$ 
(9)  $\text{CheckRelevance to } d_i$ 
(10) else
(11)  $\text{ProcessNewData in } S$ 
(12) end if
(13) end for

```

ALGORITHM 5: Overall procedure of the CLODS.

in the sliding window is to do exhaustive analysis of the most up-to-date data items and summarized the outdated items.

As can be seen in Figure 3, in the second phase, we apply the clustering of the data stream in the sliding window model where data points expire as the window slides. Moreover, with an increasing time $t = t + \Delta t$, each data points' weight declines as it reaches the expiration point. In setting the window size for a distribution that fluctuates dynamically, we increased and set the window size large enough to minimize the effect caused by the dynamic change of the data. Consequently, this results in increased time usage, which undermines the performance of real-time computation. Eventually, it creates a challenge to find a balance between these two underlying issues.

In Figure 3, as the time increases $t = \{t_1, t_2, t_3, \dots, t_n\}$ within the time frame, some data points fade out and some data points change state depending on the window slide. Some evolving data points expire, some clusters dissolve, and new ones are created, and some data points might be classified wrongly as an outlier. Therefore, in designing CLODS, we consider the following prerequisite:

- (i) Firstly, we consider the status of the data points, i.e., whether they are in a cluster or not and whether data points outside the cluster can be viewed as an inlier or outlier.
- (ii) Secondly, we consider the distance between the clusters and data points outside the clusters, whether they are far or close to the clusters, and whether they can be viewed as an outlier or inlier.
- (iii) Thirdly, we consider whether the data points share a relationship with few other data points that form a cluster, and also, how to handle both the data points within and out of the clusters to detect the outliers accurately.
- (iv) Finally, we consider the characteristics of the summary information, and at what instance we should store or discard the summary information, and what to do with expired data points.

4.6. CLODS Clustering Phase. For a data stream with a set of continuous multidimensional data points S_1, S_n , arriving at different period t_1, \dots, t_n , we considered a set of active data points during the period t_1, \dots, t_n , which are the most recent n data points at the time in the sliding window. During the active period, we employ the microcluster concept, which is a fast-efficient method for clustering objects within the sliding window. We applied the idea of triangular inequality in metric space [30, 31], to guarantee the data points' distance between each other in the microclusters is less than the distance threshold R . Thus, confirming that every data point is labeled as an inlier within the microcluster. Among the labeled inliers, we store in memory only crucial inliers to avoid memory congestion, and it is impossible to store every object in memory. We stored each newly arrived object in a fix size buffer. If the buffer is full, we consider each data point in it as an inlier or outlier, depending on the weight of the objects in relation to its distance to the other objects. The

objects that are labeled as outliers are deleted in memory, while all newly incoming labeled inliers are maintained in the updated list. The different actions taken depend on the status of the data points in the different phases.

Figure 3 shows the different stages in the window model, which is divided into three partitions with the x -axis displaying the arrival time of the data points, while the ordinate depicts the number of data points with radius R . In the first partition, during the current window model space (W_{start} to W_{end}), we have a set of evolving data streams $\{s_1, s_3\}$ with fixed radius R , and a neighbor count threshold k from time interval t_1, \dots, t_n . In this partition, for $k = 2$, the microcluster technique is applied to cluster $K + 1$ data points for the objects in the window. These microclusters are data points within the radius of $R/2$ from the center and are not greater than the distance R between the two data points. The window contains four microclusters, c_1 to c_4 with radius $R/2$. The data points that are not within the microclusters are probable outliers depending on their status in relation to the other neighboring data points. To determine whether the probable data points will be labeled as an outlier or not, we consider both its ensuing and prior neighbors and, furthermore, its relative strength to its neighbors. Also, to consider which objects are stored in memory, we used a similar concept as in previous work [8] by storing the data points outside the microcluster in temporary memory while applying the minimal pruning to minimize the computational cost and demand. From Figure 3, the red marked data points show the outliers while the other data points, where $k \geq 2$, are marked in green.

In the next phase, some data points change state due to the sliding of the window, the appearance of new data points, and the expiration of some data points. These new changes create new challenges for detecting the outliers smoothly as compared to the previous phase. In this case, we have three sliding windows. In the first window, we have a single microcluster, outliers, and a full cluster that has some data points that their status will be potentially affected during the next slide. In the next window, at the onset, although two objects have expired, it does not dissolve the microcluster since it has $k + 1$ points. However, in the final window, the microcluster dissolves, which prompts the remaining data points to become outliers. When new data points arrive, they are added to their probable neighboring microclusters, provided it is not greater than the distance threshold R . Otherwise, it is added to the neighboring outlier cluster with more space. If none of the conditions exist, then a new marked outlier cluster is initialized. In the final stage, the figure vividly shows the status of the different data points. The green data points indicate the inliers, yellow expired data points, the orange points are those that have the propensity to change state, and the red are the detected outliers.

In terms of the memory usage, owing to the fast response and limited memory requirements in these kinds of environments, it is not practical to store the majority of the data, and it is impossible to store all the data in memory. Therefore, to salvage the situation, we minimize the memory consumption and stored relevant data points that aid the

overall clustering and outlier detection process. Furthermore, we minimized the number of rearranged micro-clusters as the update in memory is done. As the continuous incoming data arrive, we first determined whether it is in memory or not. If not, it is added to the temporary memory, and then an initialization process is done. The key inliers are temporarily stored in memory, and as the data evolve due to changes in window slides, an update is done with new data points replacing the older ones. We calculated the number of inliers, and all expired data points are deleted from memory to free the memory space. Finally, summary statistical information is obtained, and the outliers are then reported.

4.7. Outlier Detection Stage. The outlier detection process involves various phases. At the onset, we observe the potential outliers through the clusters. By definition, an outlier in an evolving data stream is a data point within the computational time frame that deviates from the clusters and lies beyond the distance threshold R with fewer than k neighbors in the dataset. In every window, data points that do not meet the deviation and threshold criteria are labeled as outliers, while the others are labeled as inliers. All potential outliers are initialized to one and stored in temporary memory. As new potential outliers accumulate, the longstanding vivid outliers stored in the outlier list are deleted from memory to free up space after processing. The detected outliers are reported, and the outlier list is updated.

5. Experiments and Results

In this section, we describe the experimental settings including the datasets, parameter settings, evaluation metrics, and the baseline methods and discuss the performance of i GAAL in comparison to the other models.

5.1. Experimental Setup

5.1.1. Environment. We did our experiment using Java to design the source code and ran it on Eclipse Java EE IDE on a PC running Windows 10 Operating System with 3.20 GHz X4 CPU, 8 GB of RAM, and Disk Space of 230 GB. One of the baseline algorithms is from previous work [8], and the other was prepared by Tran et al. [32]. The source code of some baseline methods and all related datasets can be found on the online repository [32].

5.1.2. Datasets. We use similar benchmark datasets that have been adopted in some previous studies [8, 32]. As shown in Table 3, we use three real-world datasets and one synthetic dataset that are openly accessible. The first dataset is the Forest Covertype (FC) [7, 32] which is openly available and can be found from the UCI Machine Learning Repository and has 581,012 records with a high-dimensional range of 1–55 attributes. The dataset comprises tree observations from four zones of the Roosevelt National Forest in Colorado. It has no remote sensing, as the entire observations are cartographic variables from $30\text{ m} \times 30\text{ m}$

sections of the forest. The FC dataset includes information on shadow coverage, tree type, distance to nearby landmarks, soil type, and local topography. The data are in raw form (not scaled) and contain binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

The second datasets adopted for our experiment are the tropical atmospheric ocean project (TAO) datasets [32, 33], which is a low-dimensional dataset with three attributes and 575, 648 records. The dataset is real-time data extracted from National Oceanic and Atmospheric Administration website [33]. TAO was established to get useful insights and forecast climate variations related to El Nino and the Southern Oscillation (ENSO). The phenomenon, ENSO, signifies the strongest year-to-year climate instability on the planet. Its events undoubtedly interrupt normal patterns of weather variability, thereby disturbing farming, transport, Pacific marine ecosystems, energy produce, and the livelihood of millions of people around the world.

The Stock dataset has only one attribute, and it is available from UPenn Wharton Research Data Services [34] with 1,048,575 records. The dataset shows Stock trading traces of about 1 million transactions throughout the trading hours per day. Since the Wharton Research Data Services is not easily accessible, the available data can be found on the online repository [32] together with the other datasets used in this experiment.

For the Synthetic dataset, we use the Gauss dataset [32]. The dataset is generated to produce streams with measured data distribution types and number of outliers. It is generated by mixing three Gaussian distributions and a random noise distribution, and it contains 1 million records with a single attribute. In each segment of the stream, the Gaussian distributed points and noise are randomly distributed.

5.1.3. Default Parameter Settings. Before performing our experiment, we take into consideration the slide size S , the window size W , the distance threshold R , and the neighboring count threshold K . The window size W is the key parameter which determines the volume of the data streams and number of accommodated clusters, while the slide S affects the speed and the remaining parameters help to determine whether the evolving data points are inliers or outliers or whether they belong to a cluster or not. The default value of W , S , R , and K is shown in Table 3 for the different datasets.

5.1.4. Evaluation Method. We evaluated our method using three evaluation metrics: the running time, memory usage, and the clustering quality. The running time is the time taken to complete the detection of outliers for each window slide. The memory usage is the record of the peak memory used during the outlier detection process, including the storage data for each window. Lastly, the clustering quality defines how accurately our approach clusters the datasets.

TABLE 3: Datasets with default values.

Dataset	Size (M)	Dim	W	S	R	K	Outlier rate (%)
FC	0.6	55	10,000	500	525	50	1
TAO	0.6	3	10,000	500	1.90	50	0.98
Stock	1.1	1	100,000	5,000	0.45	50	1
Gauss	1.0	1	100,000	5,000	0.028	50	0.96

5.1.5. Baseline Algorithms. We chose three state-of-the-art algorithms, MCODE [4, 35] Thresh_LEAP, MCMP for comparison with the CLODS. MCODE and Thresh_LEAP were the best performing among the existing methods [36] until the hybrid approach called MCMP [8] was proposed, which uses the strength of both techniques to boost the performance in solving outlier detection problems. In MCMP, the key difference when compared to the other baseline methods is in dealing with data points within the current window. MCMP implement uses the concept of strong and trivial inliers of dealing with the objects outside the microclusters. Thresh_LEAP in the majority cases is inferior to both MCODE and MCMP because of their lack of memory-efficient microclusters. It uses an index per slide for its neighbor search. Its minimal probing principle mitigates the expensive range queries and prioritizes the discovery of a minimal number of data points according to their arrival time. It has to continually re-evaluate and manage the data points in the updated list, which consequently increases its computational demand, while MCODE prunes out and minimizes outlier candidates. It uses an index structure called a microcluster that helps to prune out unqualified outlier candidates resourcefully. However, in MCODE, the absence of clearly distinguishing between the points outside the microclusters limits its potential to perform even better. Therefore, MCMP improves this shortcoming by using the strength of Thresh_LEAP minimal probing and the memory-efficient microcluster and introduces the concept of trivial and strong inliers. This consequently improves the overall performance both in terms of reducing time and memory consumption. However, the improved performance comes at a cost, and we noticed that the absence of the extensive distance-based computation of data points outside the microclusters thus would lower the time and memory usage when we focus mainly on the clustering and deal with those points according to the relevance of their respective neighbors. For in-depth understanding of the baseline methods, we request our audience to read the individual references.

5.2. Results and Discussion

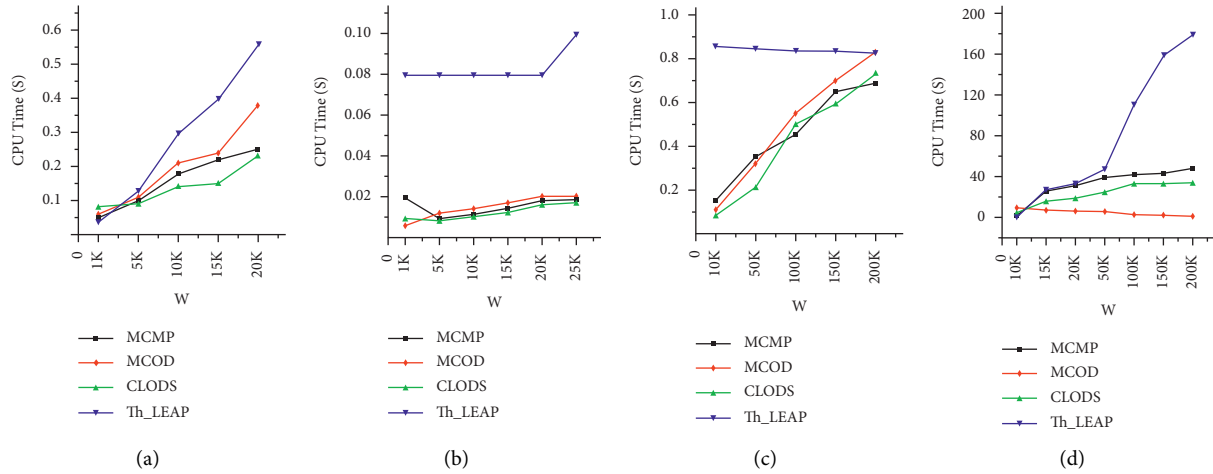
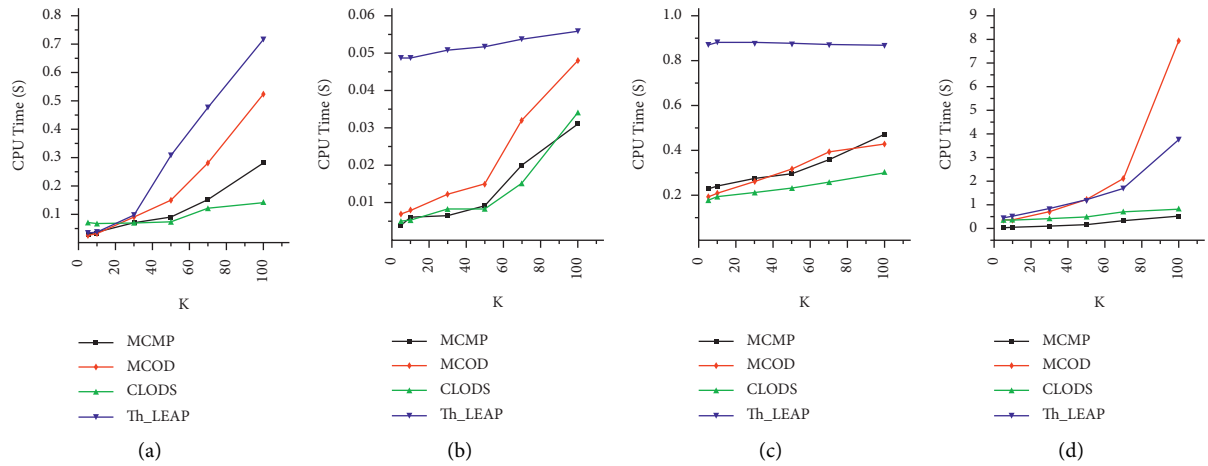
5.2.1. CPU Time. In order to observe the CPU time usage, we take into consideration the following: we vary the window size W , the distance threshold R , and the nearest neighbor count K .

Figure 4 shows the outcome of varying the window sizes W , from $10k-20k$ for FC and TAO and then $10k-200k$ for Stock and Gauss. The results are shown for fixed $K = 50$ and an approximate 1% outlier rate across the datasets. In

Figure 4, for all datasets, in most cases as W increases which means more data points to cluster and compute, the CPU time also increases (Figures 4(a) and 4(c)) except for Thresh_LEAP in Figures 4(b) and 4(c), and MCODE in Figure 4(d). The CLODS, similar to MCMP and MCODE in FC and TAO, shows a steady rise in all the datasets. However, in Gauss, when W is above 50K, we observe a sharp spike for Thresh_LEAP because fewer data points are captured since it does not have microclusters. Both CLODS and MCMP show the lowest CPU time usage when compared to the others since the use of index structures is absent. The CLODS ensures that significant inliers are stored in microclusters, which reduces the computational demand of performing range queries for every data point. Generally, we observe that when W is large enough, there is a tiny effect on the streaming data whose distribution changes dynamically. Nevertheless, if W becomes too large, then it will influence the responding time, and the time will greatly increase, which will, in turn, downgrade its performance.

Figure 5 illustrates the result of changing the neighbor count threshold k , from 1 to 100 across all the datasets. The results are shown for window size, $W = 10K$ for FC and TAO and $W = 100K$ for the remaining two datasets with other default parameters been maintained. In Figure 5, all the methods showed some changes across the different datasets since they depend on the neighbor count threshold k , which affects the outlier rate. From the figures, except for Thresh_LEAP in TAO and Stock (Figures 5(b) and 5(c)), which demands more probing to find k , the other methods showed very good time consumption with CLODS showing superior performance in the majority of the dataset. This is because, in the first three datasets, there are not many data points that fall within the clusters that will require additional computation. For Figures 5(a) and 5(d), an increase in k shows an increase in the time since more probing needs to be done. In Figure 5(d), we see that MCMP slightly outperforms CLODS because few clusters demand additional computation. Overall, our approach performs well for the datasets that have points whose neighbors are close to each other, which makes it easy for clustering and thus makes it easy to differentiate between vivid or false outliers and crucial or insignificant inliers. Consequently, it shows better performance than the others since it can do the least computation possible outside the clusters. The likelihood of getting enough neighbors to ensure the fast clustering process is relatively low for datasets with sparse data points. Therefore, there are fewer clusters in the synthetic dataset, which also results in increased processing time when compared to the real-world datasets.

Figure 6 displays the result and performance of varying the slide size S , from 1% of W to 100% of W . The

FIGURE 4: CPU time-varying W . (a) FC. (b) TAO. (c) Stock. (d) Gauss.FIGURE 5: CPU time-varying k . (a) FC. (b) TAO. (c) Stock. (d) Gauss.

slide size depicts the changes in the speed of the data stream. Across all the datasets, the value of k and R is maintained as in Table 3. In Figure 6, we can see that across the datasets the CLODS shows the lowest CPU time usage, while Thresh_LEAP incurs in the majority of the cases the highest CPU usage above that of MCOD and MCMP. In TAO and Stock datasets, we omit the trend of Thresh_LEAP since the CPU time incurs far greater than the others, and for the other two cases, it shows an abnormal trend when compared to the others. The CLODS and the other algorithm show an increase with increase in S/W . It confirms that an increase in S results in arrival and expiration of more data points, thereby consuming additional time. However, the CLODS showed improved performance compared to that of MCMP since it uses less time than MCMP and MCOD that tries to update its neighbors after the detection of strong and trivial inliers and in identifying the outliers. In addition, we can observe that the processing of new arriving data points in CLODS scales well to that of the expired data points when the

window size increases. In MCOD, for example, the time taken to process half of the data points outweighs the time for saving in discarding the expired data points. Overall, the slowest CPU time growth is shown across the datasets.

Figure 7 shows the effect of varying the distance threshold R through all the datasets, from 0–1000. The results are shown for slide size, $S=500$ for the first two datasets and $S=5K$ for Stock and Gauss. The other parameters are maintained as shown in Table 3. In each dataset, when the value of R is varied, it influences the outlier rate. For Figures 7(c) and 7(d), Thresh_LEAP incurs more time due to its trigger list, which makes it difficult to find neighbors. Overall, the CLODS showed better performance than the others and especially against MCMP since it has less distance computation when compared to MCMP that has to deal with strong and trivial inliers. The CLODS takes into consideration the relevance of K against each other rather than focusing on the influence of R . In Table 4, we notice that the outlier rate of R increases when default value of $R \leq 10\%$.

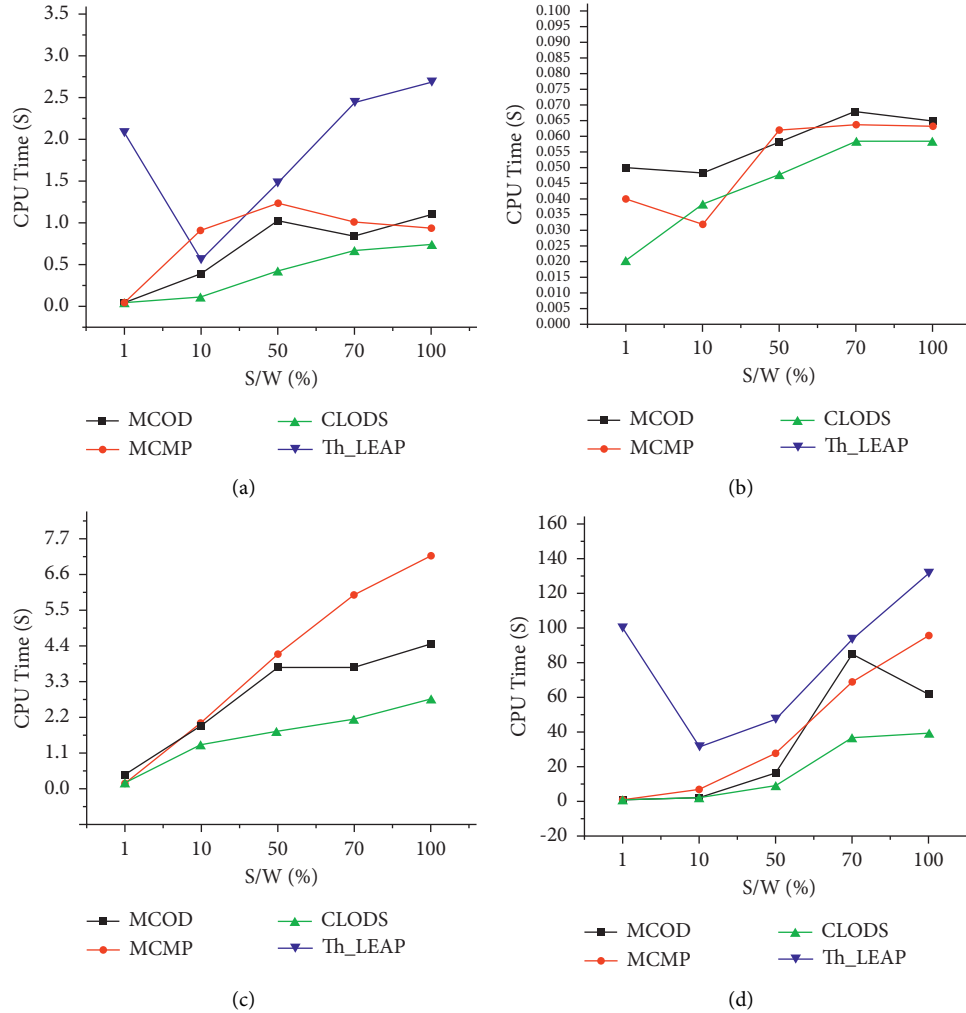
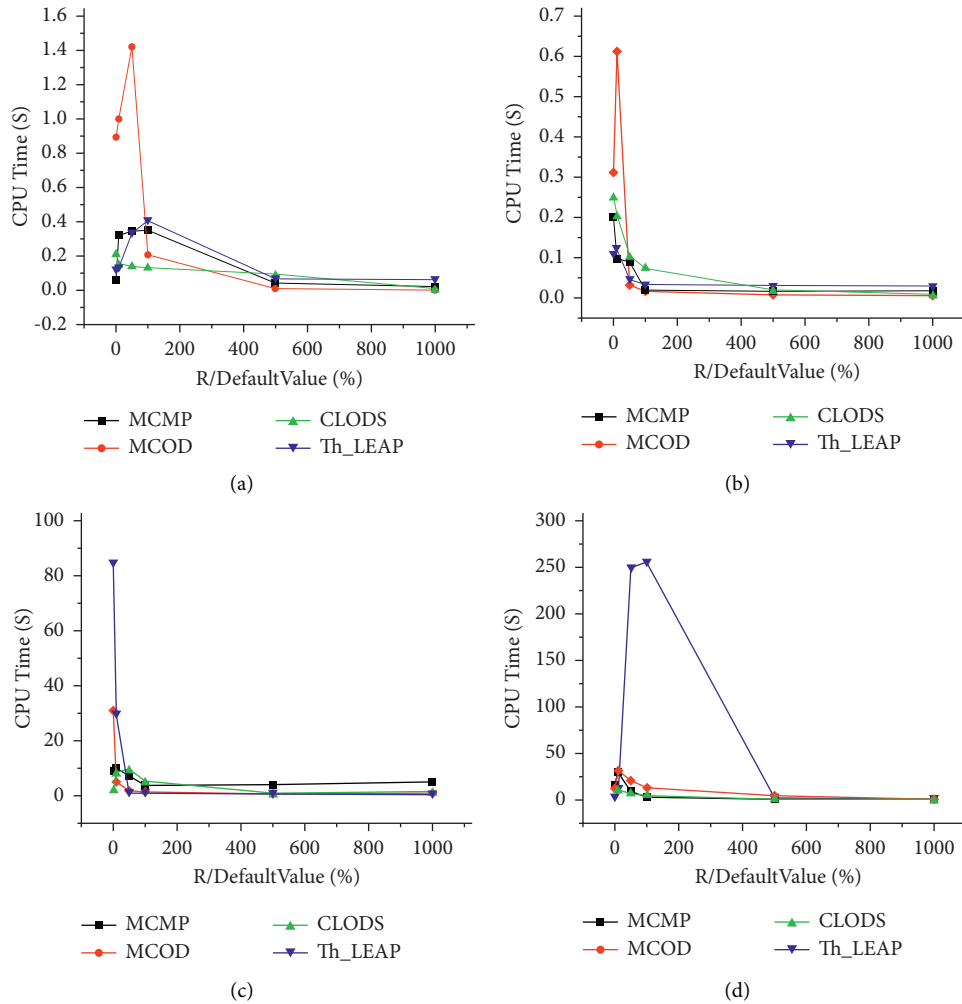


FIGURE 6: CPU time-varying S. (a) FC. (b) TAO. (c) Stock. (d) Gauss.

5.2.2. Memory Usage. In Figure 8, as the window sizes increase, it shows that more data points need to be processed, which result in an increase in memory usage for the majority of the datasets. More inliers will be in the microclusters, crucial inliers will be stored in temporary memory, and the objects' neighbors information will also be stored. From the figures, all the methods that make use of microclusters showed better performance across the datasets than Thresh_LEAP, which does not have the memory-efficient microcluster. Across all the datasets, it consumes more memory since its trigger list has to be redone every time the slides expire. In Figure 8(d), the Gauss datasets have few neighbors, and it shows an increase in memory usage for the various methods when compared to the other datasets, since finding the neighbors consumes the temporary memory. Our approach shows almost the same performance as MCOD since there is not much computation outside the microclusters like that in MCMP, which incurs slightly more memory. The CLODS, in the majority of cases, showed the least memory consumption due to freeing up space by deleting in memory detected outliers and queuing in temporary memory only significant inliers that are outside the microclusters.

When we vary the neighbor count threshold by increasing the value of k as shown in Figure 9, we expect more memory usage since k impacts the storing of the neighbors. For a few scenarios, it is almost stable, showing a small difference. For instance, in Figure 9(b), Thresh_LEAP difference does not exceed 1MB for $50 \text{ dp} \leq K \leq 20 \text{ dp}$, likewise for the other datasets in the same figure. The CLODS among the algorithms showed superior performance in most cases due to it is not entirely depending on K , as in the case of MCOD and MCMP. As K increases, more data points are not in microclusters, thereby occupying the temporary memory. For MCMP, the process of differentiating between the inliers utilizes some memory, while the CLODS only keeps a significant inlier in memory temporarily. One notable difference is in Figures 9(a) and 9(d) for Thresh_LEAP, which shows higher memory usage as compared to the others because of the neighbor count list that needs to be processed.

In Figure 10, when we vary the distance threshold R , there is no constant observable trend across the datasets. Overall, the CLODS together with the other algorithms does not make use of range queries; therefore, an increase in R does not result proportionally to an increase in memory usage. Initially, more memory is used for MCOD and

FIGURE 7: CPU time-varying R . (a) FC. (b) TAO. (c) Stock. (d) Gauss.TABLE 4: The outlier rate-varying R

R/default_R (%)	FC (%)	TAO (%)	Stock (%)	Gauss (%)
1	100.0	99.3	44.97	98.9
10	99.8	49.5	6.03	32.3
50	9.90	3.10	2.10	3.00
70	7.80	1.10	2.01	1.60
200	0.93	0.72	0.97	0.85
500	0.00	0.01	0.15	0.20
700	0.00	0.10	0.11	0.20
1000	0.00	0.10	0.07	0.20

MCMP since not many data points can be found in microclusters, and the additional computation to find neighbors occupies the memory. The CLODS showed, in most cases, better performance to some degree since it does not differentiate every outliers or inlier as in the case of MCMP, so it uses less memory at the start. In most cases, the decline in memory usage is because an increase in the value of R translates to more neighbors, which result in more objects within the microclusters and fewer data points outside the microclusters. This thus curbs the memory utilization.

Figure 11 shows the result of the memory usage when S increases. Across the datasets, the CLODS showed a decline in peak memory usage as S increases, likewise the other algorithms. The Thresh_LEAP case shows unique performance, since it differs from the others in how it processes its data points. Thresh_LEAP at the onset has higher peak memory consumption and continues to reduce further. The other memory-efficient microcluster algorithm including CLODS showed less memory consumption since it does not make use of trigger list as in Thresh_LEAP. Thanks to their microclusters, the CLODS is slightly superior to that of

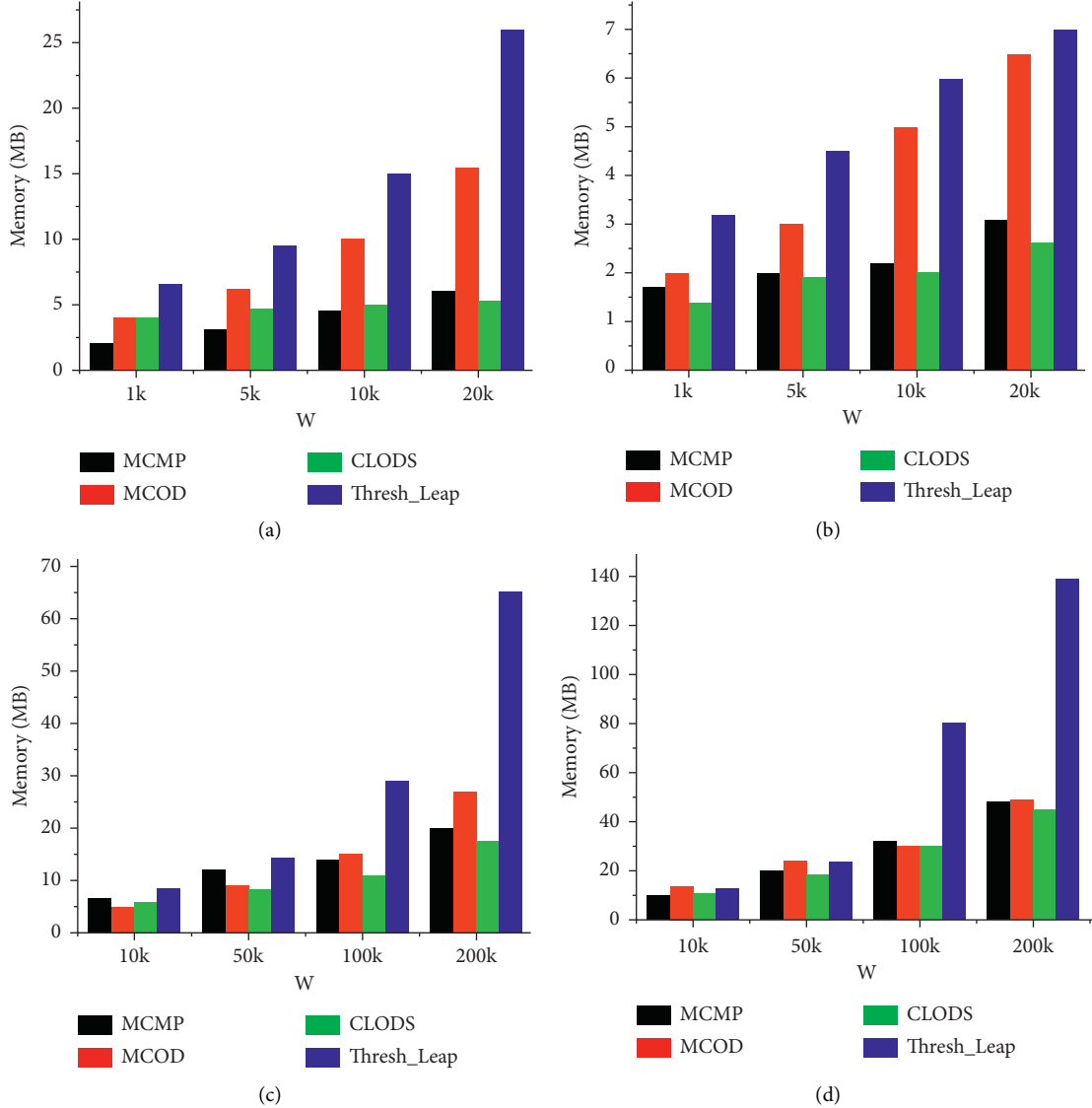


FIGURE 8: Memory-varying W . (a) FC. (b) TAO. (c) Stock. (d) Gauss.

MCMP in Figures 11(c) and 11(d), since storing the data points occupies the majority of the total memory. The absence of additional computation and to queue in memory trivial inliers gives it an advantage.

5.2.3. Space and Time Complexity. The complexity of the algorithm defines the running time and storage space needed by the algorithm in terms of its input size. The space complexity signifies the amount of memory space required by CLODS in its life cycle. To calculate the worst-case space required by CLODS, we take into consideration the space required to store the data and variables that are independent of the size of the problem. In Tables 5 and 6, we show the time and space complexity of the algorithms.

The time complexity in processing the data points within the current window in the worst-case scenario is the time cost of the function to discover whether the data point is in

cluster or not, which is $O(1-c)W$ and in checking the relevance of the d_p with respect to their neighbors in the sliding window. Since we are considering the worst-case scenario, we take into consideration the cost of computing this, which incurs higher cost than the processing of the new data points within the slide. The overall cost in this case is the cost of the data point in the window by the window slide size, that is $O(1-c)W * 1/S$. When the data points expire, in the worst-case, the process of removing expired data points within the slide does not cost as much as when we need to check the relevance of these objects and adding the data point if in cluster. In this case, the overall cost is $O(W/S \log k)$. Therefore, the overall time complexity is $O((1-c)W + (1-c)W/S + W/S \log k)$ which can be approximated to $O(W/S((1-c) + \log k))$. The time complexity of CLODS is better than that of MCMP because in CLODS the cost of checking the relevance of the neighbors to their respective neighbors is less than that of MCMP cost,

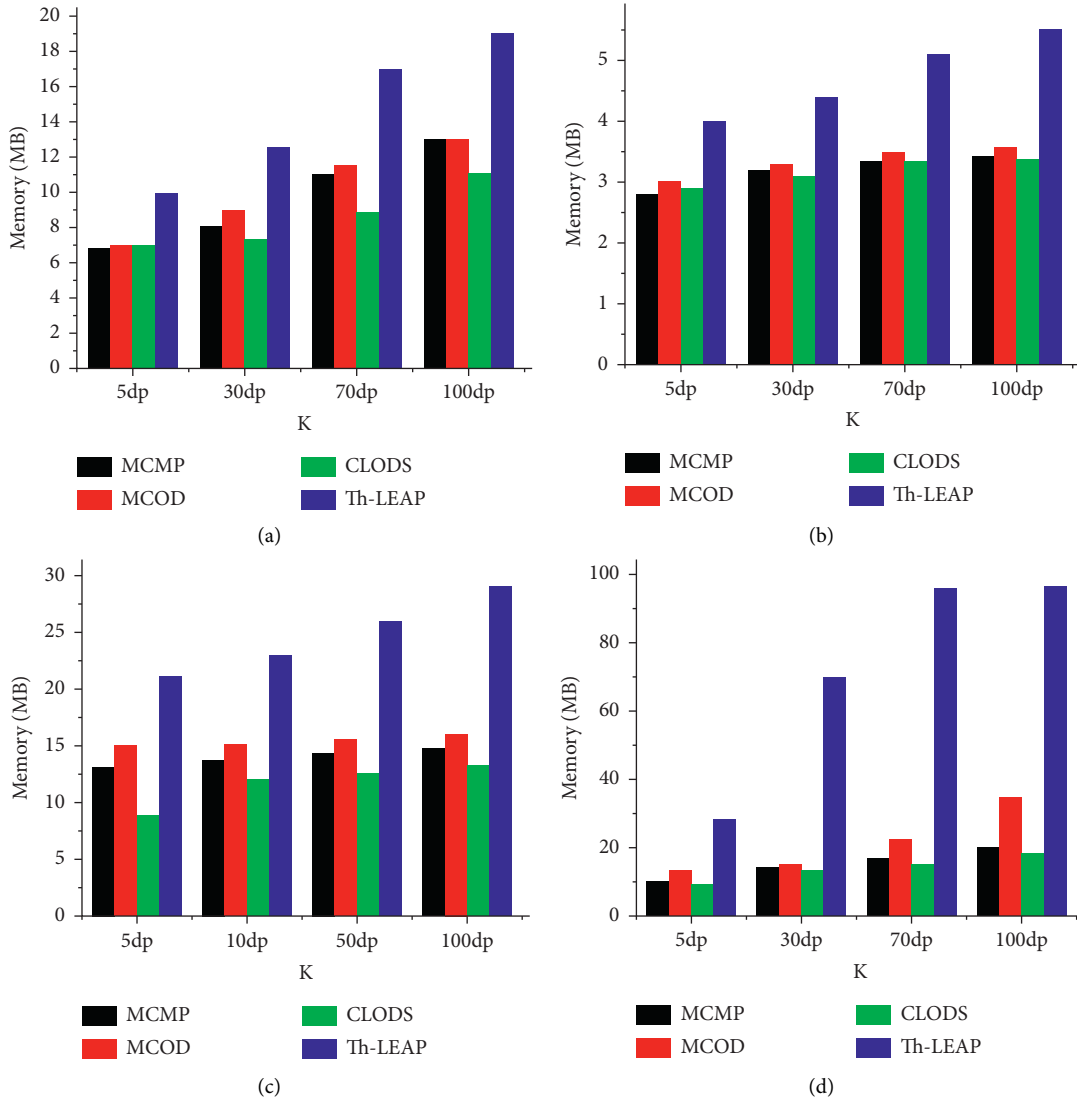


FIGURE 9: Memory-varying K . (a) FC. (b) TAO. (c) Stock. (d) Gauss.

which incurs additional cost due to the cost of differentiating between the strong and trivial inliers. We can see that the overall time complexity of MCMP is $O(W + W \log k + \log C_w + W \log C_w)$ which is approximately $O(W(\log C_w + \log k))$. This compared to the time complexity of the other two algorithms is almost the same as that of MCOD but superior to that of Thresh_LEAP. The reduction in the time complexity of MCMP confirms that microcluster using the concept of minimal probing by differentiating the strong and trivial inlier reduces the extra time required for computing data points outside the clusters as it minimizes the time complexity of recalculating and evaluating the all the inliers, as in the case of MCOD. Since differentiating between the inliers also incurs some amount of cost, however, this cost is less compared to the other way around.

In terms of the space complexity, a simple answer to the detection of continuous evolving outliers over streaming data in the window model will involve storing neighbors of

each data object in the current window. It is apparent such computation in the worst-case will result in a quadratic space requirement $O(n^2)$; therefore, for larger window size w , it will be practically unfeasible. For each data point d_i , instead of keeping all the preceding d_p and succeeding neighbors d_s , we store a number of d_s neighbors and at most k data point will suffice to detect the outliers for specific R and K . The space complexity for managing data points within the current window is $O(kW)$. We first calculate the size of the preceding neighbors d_p that corresponds to the unexpired data points. When the size is less than $k - d_s$, then d_i is labeled as an outlier. When the window slides and expired, the space required to keep the neighbor counts is similar to that of MCMP, that is, $O(W/S)$ since each data point within the window is not stored in for each W/S slide. However, in CLODS with in-depth analysis, we could say that it will slight outperform MCMP since the space complexity needed in PD to store extra trivial inliers is less than that of saving relevant inliers in queue of the memory. The

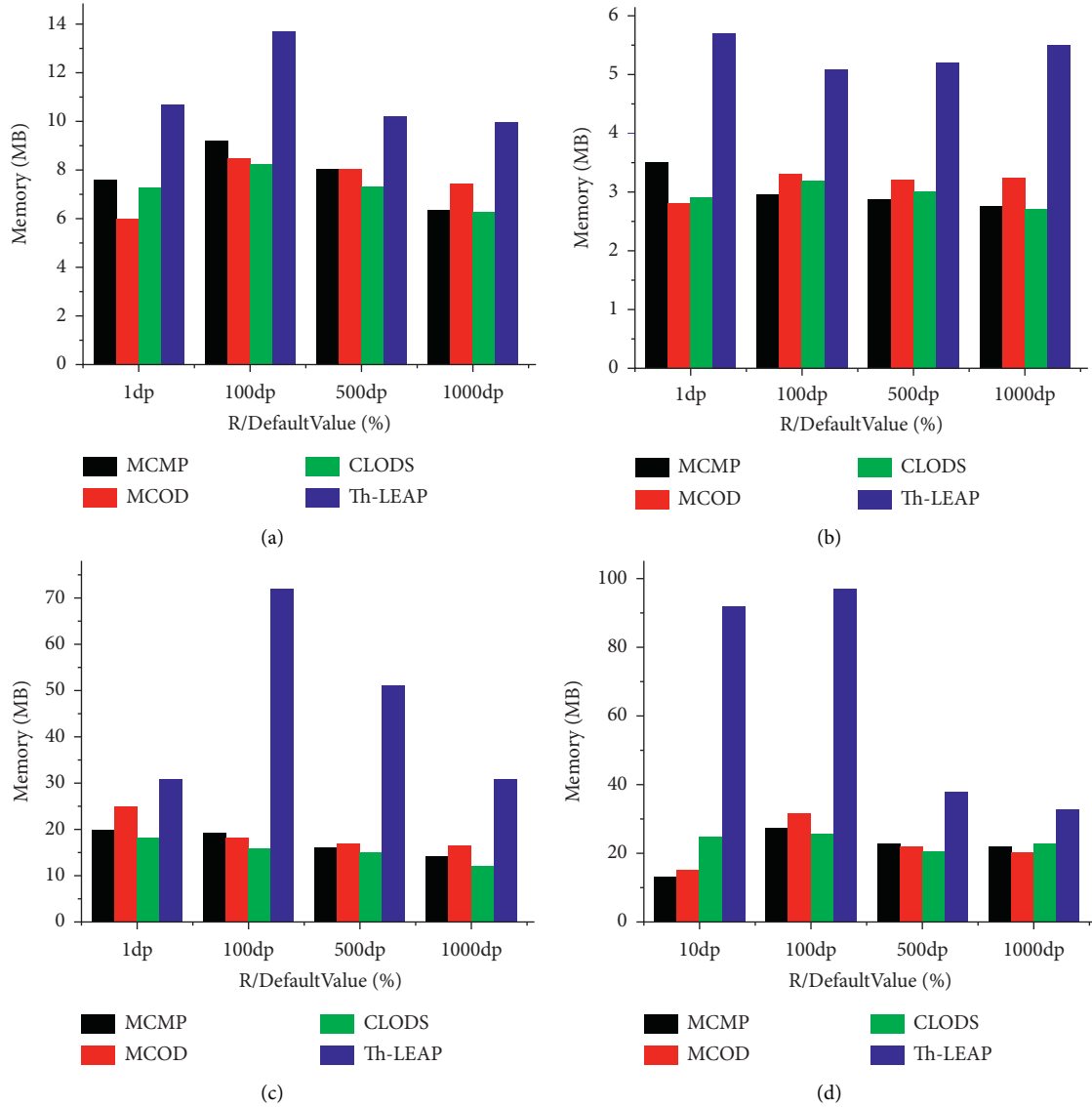


FIGURE 10: Memory-varying R. (a) FC. (b) TAO. (c) Stock. (d) Gauss.

overall worst space complexity of CLODS is $O(kW + W/S)$ which is almost the same as that of MCMP except that C_w in MCMP implies that during the expired window slides, the trivial inliers are stored in C , with $0 \leq c \leq 1$. Then, the number of data points within the window will be $(1 - c) * \text{Window}, W$. That is, the list of data point in PD will be $(1 - c) * W = C_w$. From Table 6, we can see that MCMP space complexity is also better than that of Thresh_LEAP and MCOD with $O(W^2/S)$ and $O(cW + (1c)kW)$, respectively. It is evident that the space needed for differentiating the inliers is negligible and better off compared to the space needed for data points outside microclusters to save the extra trivial inliers.

5.2.4. The Quality of Data Points in the Clusters. For clustering-based methods, an important metric to consider is the clustering quality, which affects the outlier detection rate in

the data streams. Figure 12 shows the effectiveness and clustering quality of CLODS against previous methods that also adopted microclustering technique. For the FC dataset in Figure 12(a), the percentage of clusters is relatively low since the distance between each object is sparse. In another case, for the Gauss dataset, the percentage is almost zero, with little or no data points participating in the micro-clusters. This is because, in this particular window, the dataset has few neighbors. MCMP shows inferior clustering quality when compared to both MCOD and CLODS because of its extra distance-based computation that involves computing and storing the strong and trivial inliers. In some instances, it influences the neighbor count threshold k 's relationship of the points outside the microclusters. The CLODS overall showed better clustering quality in almost all cases due to the absence of the extra computation that is involved in MCMP, and it ensures that clusters are generally formed on the basis of their relevance to their respective

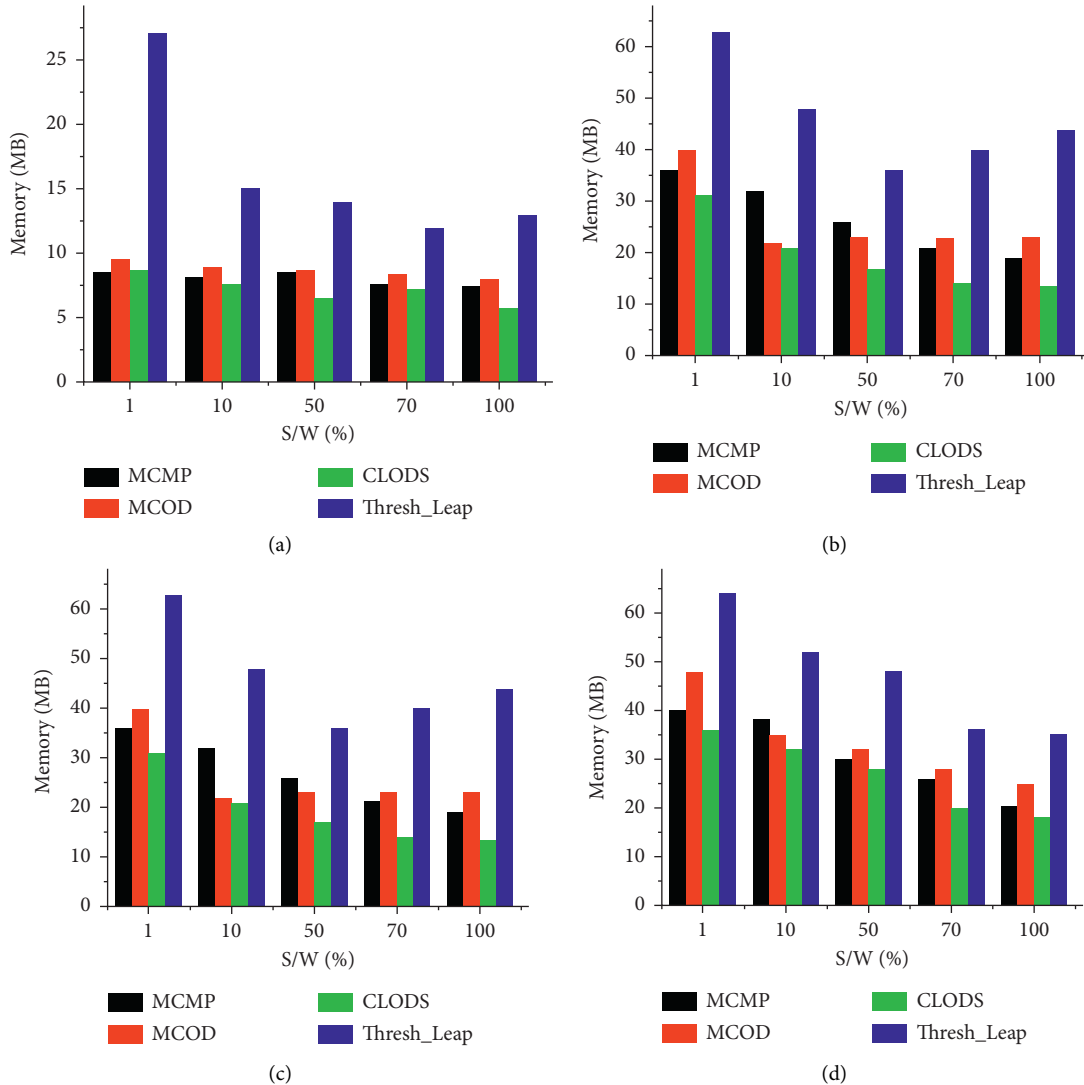


FIGURE 11: Memory-varying S. (a) FC. (b) TAO. (c) Stock. (d) Gauss.

TABLE 5: Time complexity analysis results.

Algorithms	Time complexity
Thresh_LEAP	$O(W^2 \log S/S)$
MCOD	$O((1-c)W \log((1-c)W) + kW \log K)$
MCMP	$O(W(\log Cw + \log k))$
CLODS	$O(W/S((1-c) + \log k))$

TABLE 6: Space complexity analysis results.

Algorithms	Space complexity
Thresh_LEAP	$O(W^2/S)$
MCOD	$O(cW + (1-c)kW)$
MCMP	$O(kC_w + W/S)$
CLODS	$O(kW + W/S)$

neighbors' position. This results in some cases of large percentage of data points discovered in the clusters, as can be seen in Figure 12 across the datasets.

5.2.5. Advantages of CLODS. CLODS through experiments has shown to outperform the existing methods in most cases and succeeded in curbing the computational cost in terms of the time taken and memory usage. It is a general solution used as a clustering-based outlier detection method for clustering evolving data streams based on microclusters and handling of objects within a sliding window according to the relevance of their status to their respective neighbors or position, excluding extended extra distance-based computation. The CLODS dynamically clusters data streams and offers support to meet flexible mining requirements. Furthermore, it has shown

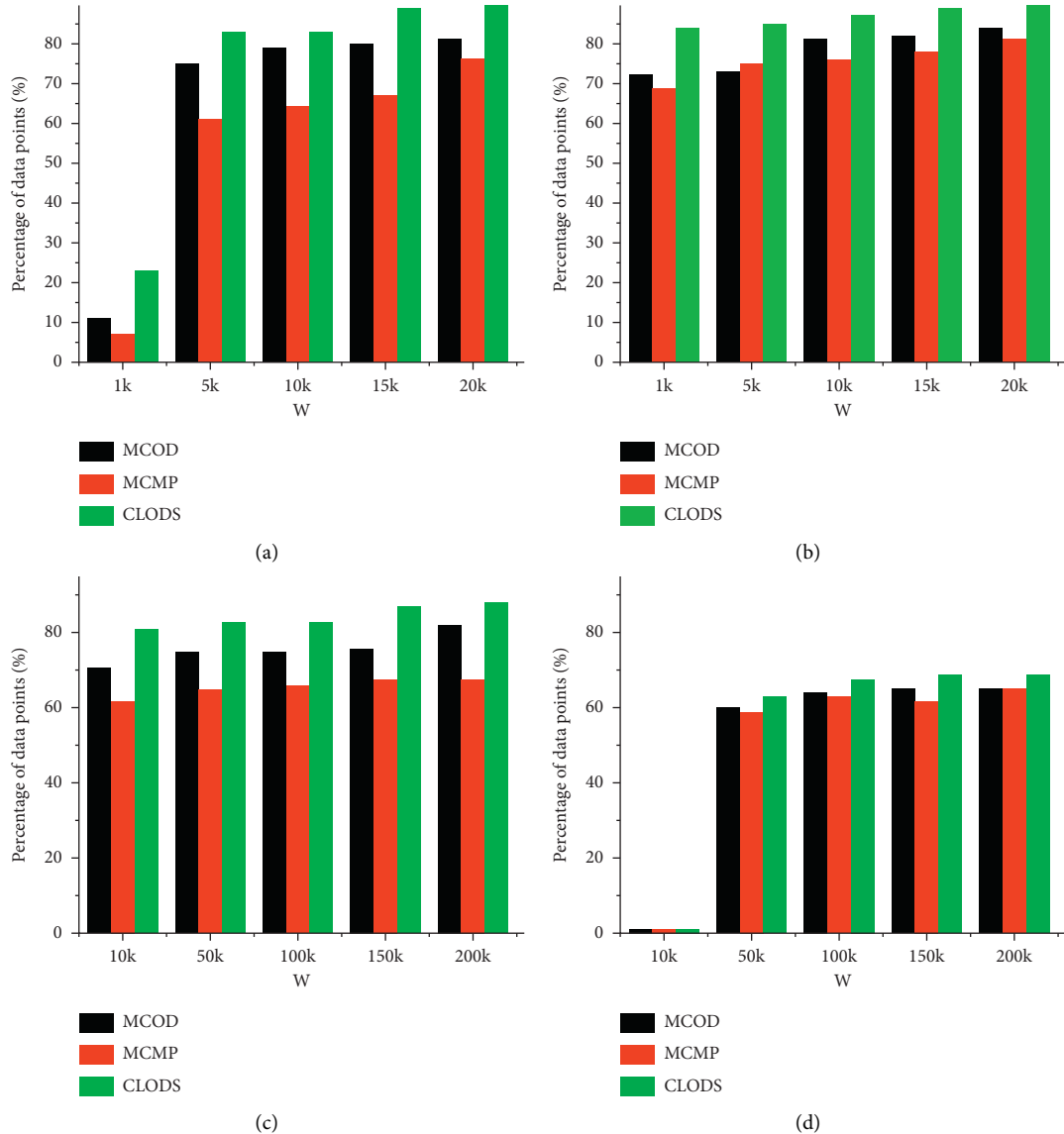


FIGURE 12: Comparison of the average percentage of data points in microclusters for MCOD, MCMP, and CLODS when we vary W . (a) FC. (b) TAO. (c) Stock. (d) FC.

robustness in the variation of the different performance parameters and its clustering quality with regard to the number of data points in its clusters. Finally, it has shown to be an effective method for detecting outliers.

6. Conclusion

Detecting outliers, which is the process of mining abnormal events from data, is a significant and challenging task. In this paper, we have proposed a clustering-based method called EMM-CLODS to address the problem of detecting outliers in continuous evolving data streams. The proposed method adopts the microcluster technique to group similar data points that are in proximity in the streaming data. It minimized the computational demand and showed an increase in the computational speed while it still maintained its effectiveness to detect outliers in the sliding window through minimal

computation of data points outside the microclusters. In terms of its memory usage, not all objects outside the microclusters were stored in memory, and likewise, expired outlier data points were deleted from memory to minimize the memory usage. From the experiments performed on both real and synthetic datasets, our method showed effectiveness in detecting outliers for continuous evolving data streams. In the majority of the cases, it shows superior performance in terms of both CPU and memory utilization when compared to the other baseline algorithms. It has shown to be a good technique for detection outliers in data streams as it is robust to the various parameter variations (W , R , and K).

Data Availability

The data and source code used to support the findings of this study have not been made available. However, all the

datasets except for the source code used have been clearly explained in the experimental section with links of where to directly access these data. Previously reported (FC, TAO, Stock, and Gauss) data were used to support this study and are available at <http://infolab.usc.edu/Luan/Outlier/>. These prior studies (and datasets) are cited at relevant places within the text.

Disclosure

Mohamed Jaward Bah, Hongzhi Wang, Li-Hui Zhao, and Ji Zhang are co-first authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

Acknowledgments

The authors would like to thank the support from the Postdoctoral Fund of Hangzhou City (no. 119001-UB2101S), PI Research Project of Zhejiang Lab (no. 111007-PI2001), Natural Science Foundation of China (no. 62172372 and no. U1866602), and Zhejiang Provincial Natural Science Foundation (no. LZ21F030001).

References

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, pp. 1–58, 2009.
- [2] X. Su and C. L. Tsai, "Outlier detection," *WIREs Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 261–268, 2011.
- [3] Ji Zhang, "Advancements of outlier detection: a survey," *ICST Transactions on Scalable Information Systems*, vol. 13, pp. 1–26, 2013.
- [4] L. Cao, Di Yang, Q. Wang, Y. Yu, J. Wang, and E. A. Rundensteiner, "Scalable distance-based outlier detection over high-volume data streams," in *Proceedings of the 2014 IEEE 30th International Conference on Data Engineering*, pp. 76–87, IEEE, Chicago, IL, USA, April 2014.
- [5] S. Guha, M. Adam, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: theory and practice," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 515–528, 2003.
- [6] Y. Chen and Li Tu, "Density-based clustering for real-time stream data," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, San Jose, CA, USA, August 2007.
- [7] S. Hettich and S. D. Bay, *The UCI KDD Archive Irvine*, Department of Information and Computer Science, University of California, Irvine, CA, USA, 1999, <http://kdd.ics.uci.edu>.
- [8] M. J. Bah, H. Wang, H. Mohamed, F. Zeshan, and H. Aljuaid, "An effective minimal probing approach with micro-cluster for distance-based outlier detection in data streams," *IEEE Access*, vol. 7, pp. 154922–154934, 2019.
- [9] L. Cao, J. Wang, and E. A. Rundensteiner, "Sharing-aware outlier analytics over high-volume data streams," in *Proceedings of the 2016 International Conference on Management of Data*, pp. 527–540, San Francisco, CA, USA, July 2016.
- [10] J. Tamboli and M. Shukla, "A survey of outlier detection algorithms for data streams," in *Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 3535–3540, IEEE, New Delhi, India, March 2016.
- [11] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: a survey," *Ieee Access*, vol. 7, pp. 107964–108000, 2019.
- [12] C. C. Aggarwal, P. S. Yu, J. Han, and J. Wang, "A framework for clustering evolving data streams," in *Proceedings 2003 VLDB Conference*, pp. 81–92, Berlin, Germany, September 2003.
- [13] P. Caroline Cynthia and S. Thomas George, "An outlier detection approach on credit card fraud detection using machine learning: a comparative analysis on supervised and unsupervised learning," in *Intelligence in Big Data Technologies—Beyond the Hype*, J. Dinesh Peter, S. L. Fernandes, and A. H. Alavi, Eds., Springer Singapore, Singapore, pp. 125–135, 2021.
- [14] M. E. Villa-Pérez, M. Á. Álvarez-Carmona, O. Loyola-González, M. A. Medina-Pérez, J. C. Velazco-Rossell, and K.-K. R. Choo, "Semi-supervised anomaly detection algorithms: a comparative summary and future research directions," *Knowledge-Based Systems*, vol. 218, Article ID 106878, 2021.
- [15] F. Liu, S. Xue, J. Wu et al., "Deep learning for community detection: progress, challenges and opportunities," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan, July 2020.
- [16] X. Su, S. Xue, F. Liu et al., "A comprehensive survey on community detection with deep learning," 2021.
- [17] A. Boukerche, L. Zheng, and A. Omar, "Outlier detection: methods, models, and classification," *ACM Computing Surveys*, vol. 53, no. 3, 2020.
- [18] X. Ma, J. Wu, S. Xue, J. Yang, Z. S. Quan, and H. Xiong, "A comprehensive survey on graph anomaly detection with deep learning," 2021, <http://arxiv.org/abs/2106.07178>.
- [19] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: a review," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–38, 2021.
- [20] D. Toshniwal and Yokita, "A framework for outlier detection in evolving data streams by weighting attributes in clustering," *Procedia Technology*, vol. 6, no. 2012, pp. 214–222, 2012.
- [21] A. Zhou, F. Cao, W. Qian, and C. Jin, "Tracking clusters in evolving data streams over sliding windows," *Knowledge and Information Systems*, vol. 15, no. 2, pp. 181–214, 2008.
- [22] F. Cao, M. Estert, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the 2006 SIAM International Conference on Data Mining*, pp. 328–339, SIAM, Bethesda, MD, USA, April 2006.
- [23] L.-x. Liu, Y.-f. Guo, J. Kang, and H. Huang, "A three-step clustering algorithm over an evolving data stream," in *Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, pp. 160–164, IEEE, Shanghai, China, November 2009.
- [24] M. Kumar and A. Sharma, "Mining of data stream using "DDenStream" clustering algorithm," in *Proceedings of the 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE)*, pp. 315–320, IEEE, Jaipur, India, December 2013.
- [25] A. Amini and T. Y. Wah, "A comparative study of density-based clustering algorithms on data streams: micro-clustering approaches," in *Intelligent Control and Innovative Computing*, pp. 275–287, Springer, Berlin, Germany, 2012.
- [26] A. Amini, T. Y. Wah, and Y. W. Teh, "DENGRIS-Stream: A density-grid based clustering algorithm for evolving data

- streams over sliding window,” in *Proceedings of the International Conference on Data Mining and Computer Engineering*, pp. 206–210, Visakhapatnam, India, January 2012.
- [27] L. Duan, L. Xu, Y. Liu, and J. Lee, “Cluster-based outlier detection,” *Annals of Operations Research*, vol. 168, pp. 151–168, 2009.
- [28] M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, “Efficient clustering-based outlier detection algorithm for dynamic data stream,” in *Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 298–304, IEEE, October 2008, Jinan, China.
- [29] A. Forestiero, C. Pizzuti, and G. Spezzano, “A single pass algorithm for clustering evolving data streams based on swarm intelligence,” *Data Mining and Knowledge Discovery*, vol. 26, no. 1, pp. 1–26, 2013.
- [30] M. S. Sadik and L. Gruenwald, “DBOD-DS: distance based outlier detection for data streams,” in *International Conference on Database and Expert Systems Applications*, pp. 122–136, Springer, Berlin, Germany, 2010.
- [31] M. B. Al-Zoubi, “An effective clustering-based approach for outlier detection,” *European Journal of Scientific Research*, vol. 28, no. 2, pp. 310–316, 2009.
- [32] L. Tran, L. Fan, and C. Shahabi, *Distance-Based Outlier Detection in Data Streams Repository*, Information Laboratory University of Southern California, Los Angeles, LA, USA.
- [33] Pacific Marine Environmental Laboratory. 2019. Wharton University of Pennsylvania. <https://infolab.usc.edu/Luan/Outlier/Datasets/tao.txt>.
- [34] Wharton Research Data Services, *Distance-Based Outlier Detection in Data Streams Repository*, Wharton Research Data Services, Philadelphia, PA, USA, 2020, <https://wrds-web.wharton.upenn.edu/wrds/>.
- [35] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsihclas, and Y. Manolopoulos, “Continuous monitoring of distance-based outliers over data streams,” in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, pp. 135–146, IEEE, Hannover, Germany, April 2011.
- [36] M. Shukla, Y. P. Kosta, and P. Chauhan, “Analysis and evaluation of outlier detection algorithms in data streams,” in *Proceedings of the 2015 International Conference on Computer, Communication and Control (IC4)*, pp. 1–8, IEEE, Indore, India, September 2015.

Research Article

Layer Information Similarity Concerned Network Embedding

Ruili Lu ¹, Pengfei Jiao ², Yinghui Wang ³, Huaming Wu ⁴, and Xue Chen ²

¹Tianjin International Engineering Institute, Tianjin University, Tianjin 300072, China

²Law School of Tianjin University, Tianjin 300072, China

³College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

⁴Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

Correspondence should be addressed to Xue Chen; xuechen@tju.edu.cn

Received 10 June 2021; Accepted 17 August 2021; Published 26 August 2021

Academic Editor: Fei Xiong

Copyright © 2021 Ruili Lu 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.

Great achievements have been made in network embedding based on single-layer networks. However, there are a variety of scenarios and systems that can be presented as multiplex networks, which can reveal more interesting patterns hidden in the data compared to single-layer networks. In the field of network embedding, in order to project the multiplex network into the latent space, it is necessary to consider richer structural information among network layers. However, current methods for multiplex network embedding mostly focus on the similarity of nodes in each layer of the network, while ignoring the similarity between different layers. In this paper, for multiplex network embedding, we propose a Layer Information Similarity Concerned Network Embedding (LISCNE) model considering the similarities between layers. Firstly, we introduce the common vector for each node shared by all layers and layer vectors for each layer where common vectors obtain the overall structure of the multiplex network and layer vectors learn semantics for each layer. We get the node embeddings in each layer by concatenating the common vectors and layer vectors with the consideration that the node embedding is related not only to the surrounding neighbors but also to the overall semantics. Furthermore, we define an index to formalize the similarity between different layers and the cross-network association. Constrained by layer similarity, the layer vectors with greater similarity are closer to each other and the aligned node embedding in these layers is also closer. To evaluate our proposed model, we conduct node classification and link prediction tasks to verify the effectiveness of our model, and the results show that LISCNE can achieve better or comparable performance compared to existing baseline methods.

1. Introduction

In the past few decades, network embedding has obtained remarkable achievements. The basic idea is converting a node into a low-dimensional space in which the network structure and properties can be preserved effectively. In the early period, traditional models such as MDS [1], Isomap [2], LLE [3], and LE [4] are mainly based on dimensionality reduction technologies. These models are not suitable for large networks due to their computational complexity. As Word2Vec [5] plays a vital role in the field of natural language processing, random walk-based methods that regard nodes in the network as words are proposed, such as DeepWalk [6] and Node2Vec [7]. In recent years, with the continuous development of deep learning, SDNE [8], DNGR

[9], and GCNs [10] have developed neural networks into network embedding models.

The methods mentioned above are all designed for single-layer networks. Figure 1(a) shows an example of a single-layer network, through which we can see that there is only one relation in the network. However, there are still many complex scenarios in the real world that cannot be described by single-layer networks. For example, the same set of individuals in social networks may participate in Twitter, Facebook, or Weibo for different purposes. Interactions in different social networks can be represented by a single-layer network. Each layer of the network has a specific relationship and specific semantics. However, these single-layer networks do not operate in isolation and there are always connections between them. Instead, these complex

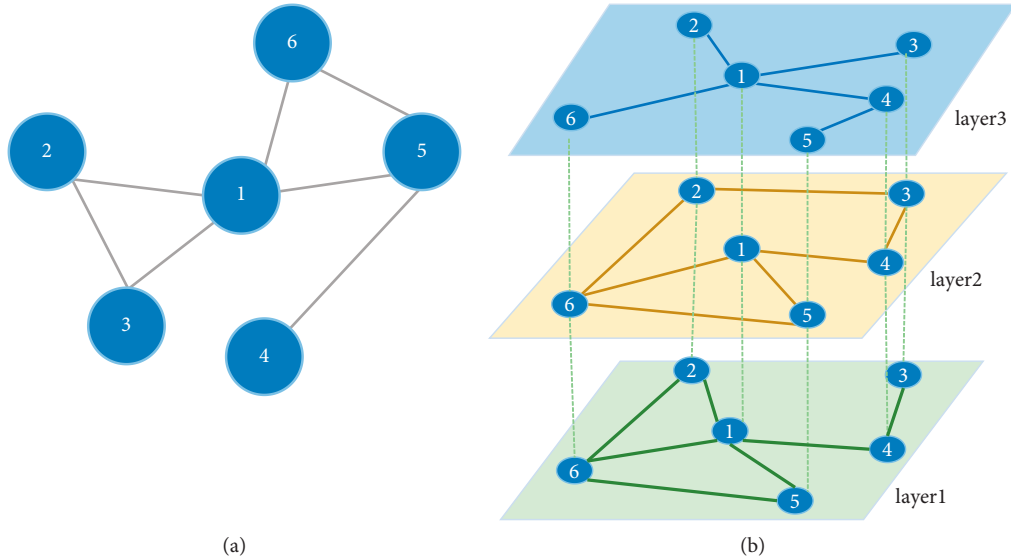


FIGURE 1: The toy examples of single-layer network and multiplex network. (a) Single-layer network. (b) Multiplex network.

scenes can be represented as a multiplex network, which is also a multilayer network in which layers share the same set of nodes. Each layer in a multiplex network represents a particular relationship of nodes, and the structure of each layer is typically associated. Figure 1(b) illustrates an example of an undirected multiplex network, and it has a unique structure in different layers while there also exist correlations between layers. Unlike the single-layer network, there are three relationships among a set of nodes, each of which describes a unique interaction in the given network structure. Multiplex relationships cannot be captured using single-layer methods. Therefore, it is necessary to conduct in-depth research on multiplex network embedding.

Compared with the single-layer network, one of the challenges for multiplex network embedding is how to aggregate the diverse types of structure in different layer networks without destroying their unique properties. To solve this problem, MCGE [11], MANE [12], and MVNE [13] use the tensor factorization to concurrently capture the main local structure and correlations between different layers. MNE [14] and MGCN [15] define one common vector shared by all layers to capture the shared information in all layer networks and low-dimensional node vectors in each layer to capture the unique properties. In addition to introducing the common vector, CrossMNA [16] also introduces a layer vector to extract the semantic meaning. One2Multi [17] uses one encoder to encode the most informative network from which we can extract the shared information and multiple decoders to reconstruct all layers learning the specific structure in each layer. DMNE [18] and MrMine [19] take advantage of the links between subgraphs or communities to learn the cross-network relationships.

While each layer in the multiplex network is constructed from different semantics and makes the structure of each

layer different, the varying relatedness between different semantics leads to diverse structural similarities between different layers. For example, we can observe from Figure 1(b) that layer2 and layer3 have more of the same edges between nodes compared to layer1, that is, the structure of layer2 and layers3 is more similar than that of layer1. Also, the similarity of any two layers is always different, which leads to the divergence in different layers of network analysis. It has been proved that considering inter-layer similarity can significantly improve the performance of link prediction [20] and community detection [21] in multiplex networks. Hence, it is an essential feature that should not be ignored in multiplex network embedding. However, the existing methods can obtain embedded representations of a multiplex network, and most of them fail to consider the similarities between different layers which is an important characteristic in the multiplex network.

To incorporate layer similarities when learning node vectors or layer vectors, we propose a novel model, Layer Information Similarity Concerned Network Embedding (LISCNE), and our model takes advantage of the common and local features in multiplex networks and exploits layer similarity at the same time.

Specifically, we firstly obtain node embeddings by concatenating common vector for each node shared by all layers and layer vector for each layer. Common vectors capture characteristics shared by cross layer by merging all the networks into a new single-layer network and training the common vector for each node in the new network. In addition, layer vectors learn the overall semantics for each layer. Then, to model the layer similarity, we define an index to formalize the similarity between different layers. With the constraint of layer similarities, we force the vectors with greater similarity to be closer.

The major contributions are summarized as follows:

- (i) After investigating the existing multiplex network embedding methods, we find that the methods consider the node connectivity among layers but ignore the inter-layer similarities.
- (ii) We propose a novel Layer Information Similarity Concerned Network Embedding (LISCNE) model, which effectively exploits the overall and local structure in multiplex networks and combines the concept of layer vectors with layer similarity at the same time.
- (iii) We conduct experiments to evaluate the proposed method using several real-world datasets on link prediction and node classification tasks. Compared with existing benchmark methods, LISCNE can achieve better or comparable performance.

2. Related Work

In this section, we review related work from two main aspects, namely, single-layer network embedding and multiplex network embedding.

2.1. Single-Layer Network Embedding. By assuming that the more similar the structure of nodes is, the closer their representation vectors are, the network embedding can learn latent low-dimensional representations for the nodes or links in a network. Earlier studies [2, 3, 22–24] were mainly based on matrix factorization. Isomap [2] obtained the shortest path d_{ij} between node i and node j by constructing a neighborhood graph with connectivity algorithms and then obtained the vector presentation by minimizing the function of $(d_{ij} - \|u_i - u_j\|)^2$. GraRep [24] defined a node transition probability and preserved k -order proximity. Inspired by Word2Vec [5], new types of methods [6, 7, 25, 26] using skip-gram model [27] have gradually emerged. The goal of the skip-gram model is to maximize the co-occurrence probability based on the context in a sentence:

$$\max \prod_{v_i \in V} \prod_{v_j \in \text{context}(v_i)} \Pr(\phi(v_i) | \phi(v_j)). \quad (1)$$

DeepWalk [6] regarded each vertex in the network as a word. It applied the Depth-First Sampling (DFS) strategy to obtain walk sequences when conducting random walks and performed the skip-gram algorithm for training the sequences. Node2Vec [7] employed a biased random walk strategy when getting the walk sequence. It defined two parameters p and q to adjust between BFS and DFS during random walks. Topo2Vec [26] used a greedy goal-based searching strategy to generate the node context and obtain the local and global topologically proximal nodes in a network. While these random walk-based methods cannot model the nonlinear structural information, some methods based on deep neural networks [8–10, 28–30] have been proposed. Both SDNE [8] and DNGR [9] used deep autoencoders, where SDNE used the encoder to preserve the first- and second-

order proximity of nodes, while DNGR captured higher-order proximity by using PPMI matrix which is indirectly transformed by the probabilistic co-occurrence matrix created by random surfing. GCNs [10] iteratively aggregated previous node embeddings and their neighbor embeddings to learn the new node embeddings. VGAE [28] was an inference model parameterized by a two-layer GCN. Pedronette and Latecki [31] proposed rank-based self-training to improve the accuracy of GCNs on semisupervised classification tasks. Recently, some novel algorithms [32–34] in the field of Contrastive Self-Supervised Learning have yielded good results. The core is to measure the similarities of sample pairs in a representation space, and the similarity between positive samples is much greater than the negative samples. These models are performed on the single-layer network. More discussion and methods for network embedding can be found in [35–38].

2.2. Multiplex Network Embedding. To better represent the multiplex networks used to describe the real-world data, there also exist various works for multiplex network embedding.

MCGE [11] applied tensor factorization and defined a multiview kernel tensor to obtain common latent factors that capture the global structure information. Random walks have been applied in network embedding [14, 19, 39–42]. MNE [14] learned two vectors for a node at the same time, i.e., a common vector sharing by all layers and a lower-dimensional vector for an individual layer. Then, it introduced a transformation matrix to align these two vectors. PMNE’s [39] network aggregation and result aggregation are essentially single-layer approaches. Considering the interactions between layers, the co-analysis method can traverse between layers with a probability r when taking a random walk. GATNE [43] proposed a unified framework to address the problem of embedding learning for attributed multiplex heterogeneous networks, and GATNE-T was a generalization of MNE [14] when training edge embeddings directly. MrMine [19] simultaneously learned the multinet network representation at three resolutions of network, subgraph, and nodes, and it further constructed cross-resolution including network-subgraph, subgraph-node, node-node context. HMNE [44] defined a heuristic 3D interactive walk and sampled sequences of node cross layers. It preserved cross-layer neighborhood of nodes and learned information of multitype relations into a unified embedding space.

MVE [45] learned the robust representation by promoting the collaboration of different layers and different weights which were assigned to layers during voting. CrossMNA [16] defined a network vector extracting the semantic meaning of the network and an inter-vector reflecting the common features of the anchor nodes in different networks. Then, these two vectors were added to form an intra-vector, which preserved the specific structural feature for a node in its selected network. MGCN [46] extended GCN to multiplex networks, which defined a general vector and dimension-specific vector to capture the common and individual layer information. TCMGC [47] developed a multilayer GCN to

capture the structure and multiview information. DMNE [18] used an encoder for all individual networks and regularized the cross-network embeddings through two types of loss functions to penalize the embedding inconsistency. DMGI [48] was an unsupervised model based on DGI [49]. In an individual layer, it performed the DGI algorithm to get the relation-type specific embedding and then took advantage of the multiplexity of the network by introducing consensus regularization and multiheaded attention mechanisms. MEGAN [50] was a multiplex GAN that designed a multilayer generator to model multilayer connectivity to generate fake samples and a node pair discriminator to enforce the generator to more accurately t the distribution of multilayer network connectivity. One2Multi [17] used the network with the most information as the input of encoder to learn the shared information of all the networks and then used a multidecoder to reconstruct the multiplex network from the shared information.

All the single-layer models mentioned above are effective for single network embedding; however, they do not consider the correlation in the multiplex network. In addition, the GCN-based multiplex network embedding models only consider the local information in the network, while other models ignore the similarities between layers. Our model combines the similarities between the layers and can simultaneously capture the local and global information in the network and the multiplex relationships between layers.

3. Notations and Problem Formulation

We begin with a formal definition of multiplex network, followed by the problem formulation. For the sake of clarity, the main notations are summarized in Table 1.

Definition 1 (multiplex network). A multiplex network consists of a set of N nodes $V = (v_1, v_2, \dots, v_N)$ and L layer. All layers share the same set of nodes V and the nodes form diverse structures in each layer. The structure layer l can be represented as ε_l . We denote this multiplex network as $G = \{G^1, G^2, \dots, G^L\} = \{V, \varepsilon\}$, where $\varepsilon = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L\}$.

Given such a multiplex network with L layers, the goal of our work is to learn low-dimensional embeddings $Z_i^l \in R^d$ for each node v_i on each individual network G^l , where d is the dimension of the embedding. The learned representations can be used as features in a variety of applications such as node classification and visualization, relationship mining, and link prediction. In our experiments, we perform both link prediction and node classification tasks to verify the effectiveness of the learned embedding.

4. Layer Information Similarity Concerned Network Embedding

As the nodes in each layer of the multiplex network are same, they shared the common information and the same node may show some similar features among layers. However, the structure among nodes in each layer is formed by different

TABLE 1: Main symbols and their definitions.

Symbol	Definition
L	The number of layers in the multiplex network
G^l	The network for layer l in the multiplex network
N	The number of nodes
V	The node set of the multiplex network
ε_l	The edge set of l -th network
r^l	The layer vector for l -th network
u_i	The common vector for node v_i
U	The common vector matrix for all nodes
Z_i^l	The embedding vector for node v_i in network G^l
d_1	The dimension of common vector
d_2	The dimension of layer vector
d	The dimension of final node vector
S	The similarity matrix between networks
$S^{\alpha\beta}$	The similarity between networks G^α and G^β

semantics and thus leads to quite diverse local structures of this node in each layer, and the varying relatedness between different semantics also leads to diverse structural similarities between different layers. In this paper, we propose LISCNE which models the common and local features in multiplex networks and exploits layer similarity at the same time.

Figure 2 illustrates the framework of LISCNE for a three-layer multiplex network. The architecture contains two components. The first part is modeling the common vector for all nodes that are shared by the counterpart nodes among different layers. The second part is learning the node embedding in each layer by integrating the common layer and layer vector introduced to capture distinct semantic information of different networks. The last part is describing the process of training layer vectors with layer similarities. The embedding for node v_i in layer l is defined as

$$Z_i^l = f(u_i + r^l), \quad (2)$$

where f is the map function integrating common vector and layer vector to get the final node presentation. In our model, we use concatenation as the map function. LISCNE specifies the relationship of different networks by the layer similarities, i.e., S^{12} indicates the index of structural similarity of network G_1 and network G_2 . By adding layer similarities to the layer vector, it can associate within-network and cross-network structure information.

Next, we will describe our model LISCNE in detail and introduce it in three parts: common feature modeling, learning node embedding in each layer, and integrating the similarity between layers.

4.1. Common Feature Modeling. In this part, we learn the common feature shared by the counterpart nodes among different layer networks in the multiplex network. Firstly, we use a network aggregation method to aggregate all layers into a new single-layer network, where multiple edges are not allowed. Specifically, we set the new network as $G^{\text{new}} = \{V, \varepsilon_{\text{new}}\}$, and for the edge in $\varepsilon_l \in \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L\}$, we add the edge in ε_{new} . The process is shown in Figure 3. Then, over the obtained new network, we learn the common vector

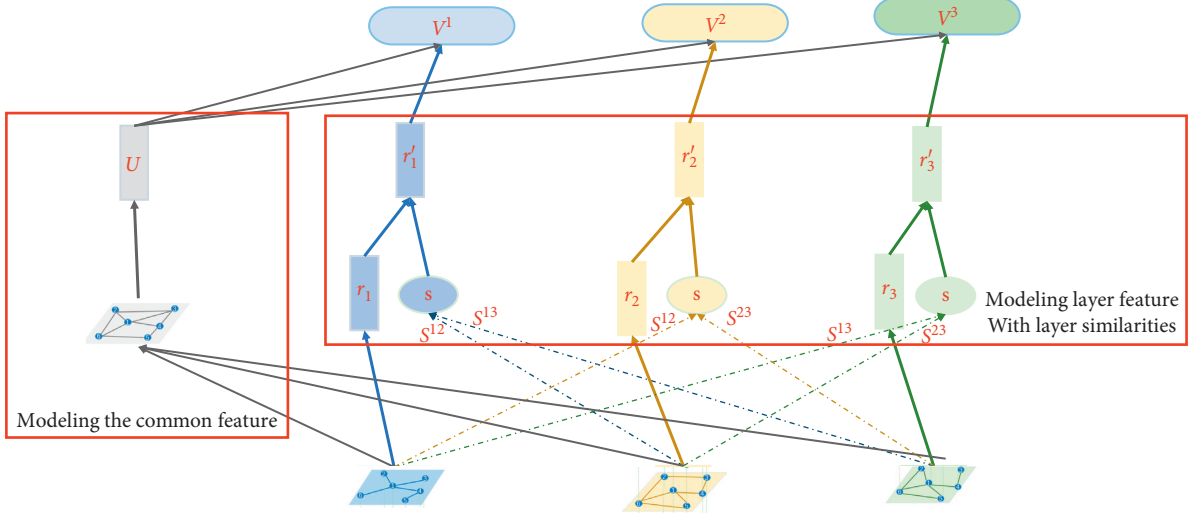


FIGURE 2: The simple framework of LISCNE.

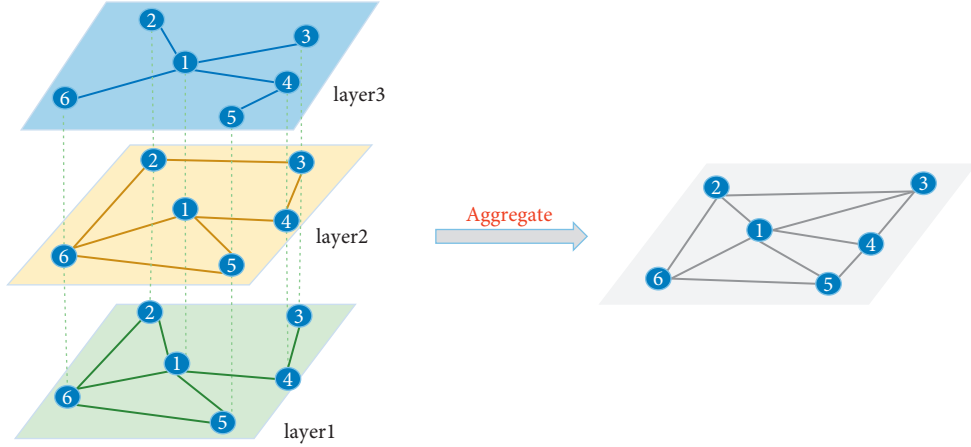


FIGURE 3: The process of aggregating the multiplex network into a single-layer network.

matrix U for all nodes. We take node v_i as an example; to get the common vector u_i , our goal is to maximize the probability of its neighbors' context in the given walk sequence:

$$\max P(v_{i-w}, \dots, v_{i+w} | v_i; u_i), \quad (3)$$

where w is half of the window size and the neighbors of v_i are v_{i-w}, \dots, v_{i+w} .

Based on the assumption of conditional independence and using the logarithmic probability, it can be further factorized as

$$L_1 = \sum_{v_i \in V} \sum_{i-c \leq j \leq i+c, j \neq i} \log P(v_j | v_i), \quad (4)$$

where $P(v_j | v_i)$ can be defined with a softmax function as

$$P(v_j | v_i) = \frac{\exp(u_j^T u_i)}{\sum_{k \in V} \exp(u_k^T u_i)}, \quad (5)$$

where u_i and u_j are the common vectors for the input node v_i and context v_j , respectively.

4.2. Learning Node Embedding in Each Layer. As discussed before, each layer in a multiplex network has distinct information, and to capture the specific structure for an individual network, we introduce the layer vector that maps single layers into a latent space, i.e., the layer vector r_l for the individual graph G^l . To obtain the overall structure of the multiplex network and layer vectors and learn semantics for each layer simultaneously, we get the node embedding for each layer by concatenating them. For a random node v_i , the embedding in layer G^l can be defined as $V_i^l = u_i \| r^l$.

To preserve local neighborhoods of nodes in each layer, our goal is to maximize the probability of specific neighbors' context in each individual layer:

$$L_2 = \sum_l \sum_{v_i \in V} \sum_{v_j \in C^l(v_i)} \log P(v_j | v_i; Z_i^l), \quad (6)$$

where $C^l(v_i)$ is the context of node v_i in layer G^l and $P(v_j | v_i; V_i^l)$ can be defined as

$$P(v_j|v_i; V_i^l) = \frac{\exp(Z_j^{lT} Z_i^l)}{\sum_{k \in V} \exp(Z_k^{lT} Z_i^l)}, \quad (7)$$

where V_i^l and V_j^l are the node embeddings for the input node v_i and context v_j , respectively.

4.3. Integrating the Similarity between Layer-Networks. The layer vector learned above can capture the distinct structure information within the layer, while in the multiplex network, there is another essential characteristic, which is the similarity between layers varying from layer to layer. Najari et al. [20] testified that incorporating the inter-layer similarities can improve the link prediction performance. Therefore, inspired by their study, we thought of using similarities to enhance embedding capabilities. We added constraints for layer vectors with the similarities between this layer and other layers. Through integrating into the layer similarity, we made the layer vector capture the cross-layer and within-layer information simultaneously.

Firstly, in our model, we used the Global Overlap Rate (GOR) algorithm to measure the similarity among layers in multiple networks. In detail, given two layers α and β in a multiplex network, an overlap edge means that the same node pair simultaneously exist in both networks. The global overlap between layers α and β is denoted by $S^{\alpha\beta}$, which represents the total number of overlapping edges observed in layers α and β . It can be formulated as

$$S^{\alpha\beta} = \frac{|\varepsilon_\alpha \cap \varepsilon_\beta|}{|\varepsilon_\alpha \cup \varepsilon_\beta|}, \quad (8)$$

where ε_α is the total number of edges in layer α . The range of $S^{\alpha\beta}$ is in $[0,1]$, and the higher the value, the more the similarity between layers. Particularly, $S^{\alpha\beta} = 0$ represents that there are no overlapping edges between layers, indicating that the layers are not related; otherwise, $S^{\alpha\beta} = 1$ means that the layers are completely correlated. The similarity $S^{\alpha\beta}$ between layer α and β is the same as similarity $S^{\beta\alpha}$ of layer β and α , and this can also be seen from equation (8).

After illustrating the definition, the next problem we should deal with is how to incorporate it into the model. To address this issue, we assume that if the structures of the two layers are more similar, their representation in the vector space should be closer. We force the following equation to obtain the minimum value:

$$L_3 = \sum_{\alpha\beta} \|r^\alpha - r^\beta\| S^{\alpha\beta}. \quad (9)$$

From equation (9), we can employ stochastic gradient descent to minimize L_3 function as follows:

$$r^l = r^l - \sum_{\beta} \frac{r^l - r^\beta}{\|r^l - r^\beta\|_2} S^{l\beta}. \quad (10)$$

4.4. Time Complexity Analysis. Our loss function includes two components. The first part is maximizing the probability-

specific neighbors' context in each individual layer to learn the node embedding in each layer, where the main processes of time consumption include getting random walk sequences and skip-gram training, just as the ordinary random walk algorithm. Assuming that the number of nodes is N , the number of edges in each layer is M , the walking length is T , and the number of walking sequences per node is t , the complexity of sampling all sequences is $O(M) + O(N^*T^*t)$. Besides, the complexity of optimization of N^*t sequences with the skip-gram model is $O(N \log N)$. Therefore, the time complexity of learning the node embedding in each layer is $O(M) + O(N) + O(N \log N)$. The second part is integrating the similarity between layer-networks. In this part, we exploit the structural similarity between pairs of two layers, and the time complexity is $O(L^*(L-1))$. In real-world network data, the number of layers of L is often very small. The time complexity of this part is relatively insignificant compared to that of the first part of learning node embedding in each layer. So, the overall time complexity of our model is $L^*(O(M) + tOn(N)q + hO_l(N \log N))$.

5. Experiments

In this section, we conduct experiments to validate the proposed LISCNE. To compare our model with some state-of-the-art single-layer embedding methods and multiplex network embedding methods, we perform link prediction and node classification tasks on several datasets with different types of networks.

5.1. Datasets. We employ five real-world multinet network datasets from three different fields: social, co-authorship, and genetic. The basic statistical information of the datasets is presented in Table 2.

All these datasets are downloaded from the CoMuNe lab's website (<https://comunelab.fbk.eu/data.php>). The detailed descriptions are as follows:

- (i) CKM [51]: by asking the physicians in Illinois, Bloomington, Quincy, and Galesburg three questions, this dataset is classified into three types of relationships. Its ground truth is related to node labels; therefore, we also use this dataset to perform the node classification task.
- (ii) PIERRE [52]: this dataset maps layers to different working tasks within the Pierre Auger Collaboration. Based on the keywords and contents of all submissions between 2010 and 2012, the multiplex network is divided into 16 layers.
- (iii) ARABIDOPSIS [53, 54]: based on BioGRID, this multiplex network considers genetic interactions of different types of organisms. The multiplex network used in the paper makes use of the following layers: direct interaction, physical association, additive genetic interaction defined by inequality, suppressive genetic interaction defined by inequality, synthetic genetic interaction defined by inequality, association, and colocalization.

TABLE 2: Statistics of datasets.

Dataset	Network type	Layers	Nodes	Edges	Directed/undirected
CKM	Social	3	246	1,551	Directed
PIERRE	Co-authorship	16	514	7,153	Undirected
ARABIDOPSIS	Genetic	7	6,980	18,654	Directed
MUS	Genetic	7	7,747	19,842	Directed
Arxiv	Co-authorship	13	14,489	59,026	Undirected

- (iv) MUS [53, 54]: the dataset is also based on BioGRID. The layers in this dataset are physical association, association, direct interaction, colocalization, additive genetic interaction defined by inequality, synthetic genetic interaction defined by inequality, and suppressive genetic interaction defined by inequality.
- (v) Arxiv [52]: choosing papers with “networks” in the title or abstract up to May 2014 in arxiv, the dataset is divided into 13 layers corresponding to different categories with 14,489 nodes.

5.2. *Baseline Models.* To show the performance of our model, the following six baseline models are implemented for comparison, which can be classified into single-layer network embedding and multiplex network embedding.

- (i) DeepWalk [6]: this is a classic single-layer network embedding method, which applies a random walk to get walk sequences and then conducts the skip-gram algorithm on the sequences to train the model.
- (ii) Node2Vec [7]: this is also a typical single-layer network embedding model, which utilizes two parameters to take control of the traverse probability in taking the random walk strategies.
- (iii) PMNE [39]: this is a multiplex network embedding model that consists of three methods, where network aggregation and result aggregation simply merge all networks or the embedding results of all networks into one, while co-analysis takes the interaction among layers. PMNE_n, PMNE_r, and PMNE_c are used to denote the network aggregation, result aggregation, and co-analysis, respectively.
- (iv) MNE [14]: this is a multiplex network embedding model that defines two different dimensional vectors for a node to capture the common information in the whole network and the specific features in a single layer, respectively.
- (v) CrossMNA [16]: this is a multiplex network embedding model and also a model for network alignment. It learns simultaneously inter-vector sharing by the anchor nodes in different networks and a network vector for each single layer.

5.3. *Experimental Setting.* For our model, we set both the common vector dimension and layer vector dimension to 100, and thus after concatenation, the final node embedding vector dimension is 200. For the sake of fairness, we set all

the dimensions of final vectors compared with our models as 200. Additionally, for DeepWalk, we set the walk to 20 and the walk length to 80 for each node taking a random walk. For Node2Vec, we empirically set $p = 2$ and $q = 0.5$. For PMME, we follow the default setting in the original paper, which sets α , p , and q to 0.5. For MNE, we set the additional vector dimension to 10 and the common vector dimension to 200. For CrossMNA, according to the original paper, we set the dimension of the inter-layer vector to 200 and the dimension of the network vector to 100.

5.4. *Evaluation Metrics.* We perform link prediction and node classification tasks to validate the efficiency of our model. For the link prediction task, we execute experiments in each layer and take the average as the final results. Then, we randomly divide datasets into testing sets and training sets. When predicting each positive edge, we also randomly sample unconnected node pairs as a negative edge. We adopt the ROC-AUC evaluation metric to test model performance, that is, the higher the value of AUC is, the better the model performs. For the node classification task, we train all data to get node embeddings of individual layers through our model and baseline models, get the average node embedding of all layers, and then inject the embeddings into a classifier to evaluate the effect. In our experiment, we select a logistic regression classifier and choose the $F1$ (weighted) and precision (weighted) as evaluation metrics.

5.5. *Performance on Link Prediction.* For single-layer methods, we train the node embedding for each layer and use it to predict links in the corresponding layer. For the three methods of PMNE, which take different strategies to aggregate the representations of all layers into one, we take the final node embedding to predict links in all layers. For all models, we average the AUC values of all relation types as final results. In experiments, we take five-fold cross-validation for all datasets.

The results are shown in Table 3, from which we can draw the following observations:

- (i) The proposed LISCNE model can stably outperform or achieve comparable performance with all the baseline methods. The results show that merging the layer similarity into models can exactly improve the performance.
- (ii) The multiplex network models almost perform better than single-layer models. Meanwhile, these single-layer models in different datasets vary a lot, e.g., in PIERRE dataset, DeepWalk and Node2Vec

TABLE 3: Results of link prediction on different datasets.

Model	CKM	PIERRE	ARABIDOPSIS	MUS	Arxiv
Node2Vec	0.707	0.572	0.525	0.651	0.753
DeepWalk	0.7	0.589	0.563	0.626	0.759
PMNE_n	0.781	0.8	0.828	0.867	0.872
PMNE_r	0.789	0.65	0.586	0.62	0.782
PMNE_c	0.763	0.516	0.529	0.563	0.599
MNE	0.785	0.791	0.765	0.779	0.834
CrossMNA	0.828	0.733	0.845	0.879	0.922
LISCNE	0.883	0.792	0.825	0.888	0.924

perform poorly. In other words, considering the intersection across layers is essential.

- (iii) LISCNE, CrossMNA, and MNE are more effective than PMNE’s three methods. PMNE models learn an overall vector for each node by aggregating all layers, while LISCNE, CrossMNA, and MNE all simultaneously define a vector to capture the common information and another vector to capture the distinct information about each specific layer.

5.6. Performance over Common Vector Embedding Dimension. Figure 4 shows the performance of our model as the embedding dimension of the common vector increases. It can be clearly seen from the figure that the larger the dimension, the better the prediction effect. When the dimension reaches 10, the curve tends to stabilize. Here, for the sake of both accuracy and computational complexity, we set the common vector dimension d_1 to 100.

5.7. Performance on Node Classification. In the node classification task, we choose the CKM dataset with reliable node labels to conduct the experiment and take the companies as the classification label. In addition, the ones injected into the classifier are average node vectors for node embeddings in individual layers. For single-layer network methods, we train all nodes in each layer and get the average of node vectors in each layer. For MNE, CrossMNA, and our model, we also get the average of node vectors of intra-vector in individual layers. Then, all the node representations and corresponding node labels in each layer are divided into training and testing datasets to train the classifier. In our experiment, we use a logistic classifier and evaluate the classification performance with the metrics accuracy, precision, and $F1$, respectively, which can be defined as follows:

$$\begin{aligned} \text{accuracy} &= \frac{TP + TN}{TP + FP + TN + FN}, \\ \text{precision} &= \frac{TP}{TP + FP}, \\ F1 &= \frac{2TP}{2TP + FN + FP}. \end{aligned} \quad (11)$$

As shown in Figure 5, the results prove the effectiveness of our model, where our model LISCNE can provide the best performance in terms of $F1$ and precision and achieve

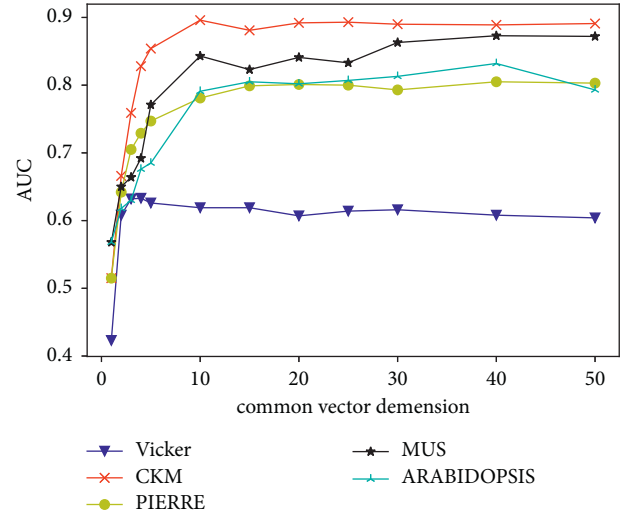
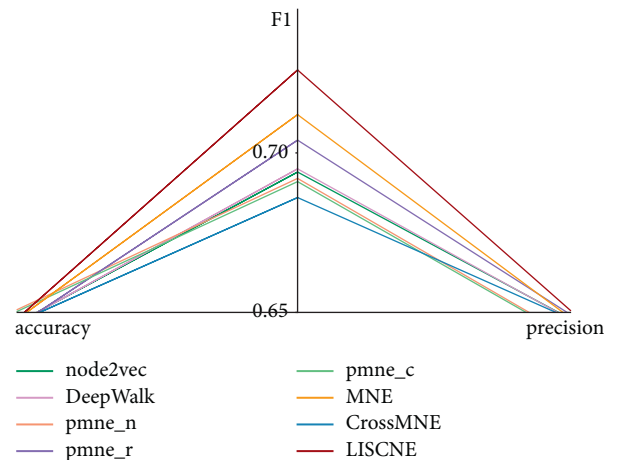
FIGURE 4: Performance over the dimension d_1 of common vector embedding.

FIGURE 5: The performance of node classification on CKM.

comparable accuracy with PMNE_n and PMNE_c. However, the effectiveness of multiplex network embedding models like CrossMNA on the link prediction task is not obvious. This may be because every model injected into the classifier is the average of node embedding in all layers. The effect of average is somewhat like aggregation and gets the shared information in all layers.

6. Conclusion and Future Work

In this paper, we propose an effective method called LISCNE for multiplex network embedding. LISCNE defines a common vector for all counterpart nodes in the multiplex network and also introduces a layer vector for each layer. Moreover, when learning layer vectors, it first merges the layer similarities to simultaneously capture intra-layer information and cross-network information. We have performed link prediction and node classification tasks to test LISCNE and conducted extensive experiments

to verify the effectiveness of our proposed model. This model is applicable to aligned networks and certain networks, in which one node in some network is only connected to one node in another network.

Unfortunately, this kind of network cannot cover lots of scenarios in the real world, e.g., the association between a collaboration graph of researchers and a citation graph of papers, where an author can cite papers on multiple topics. In the future, we will extend our model to more manifold networks, for example, one node in some network is connected to several nodes in another network through different weights.

Data Availability

The datasets used to support the results of this study can be available from <https://comunelab.fbk.eu/data.php>.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (61902278).

References

- [1] J. B. Kruskal, *Multidimensional Scaling*, SAGE, no. 11, Thousand Oaks, CA, USA, 1978.
- [2] J. B. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [4] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, vol. 14, pp. 585–591, 2002.
- [5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," <http://arxiv.org/abs/1301.3781>.
- [6] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, New York, NY, USA, August 2014.
- [7] A. Grover and J. Leskovec, "Node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, San Francisco, CA, USA, August 2016.
- [8] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234, San Francisco, CA, USA, August 2016.
- [9] S. Cao, W. Lu, and Q. Xu, "Deep neural networks for learning graph representations," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, pp. 1145–1152, AAAI Press, Phoenix, AZ, USA, February 2016.
- [10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," <http://arxiv.org/abs/1609.02907>.
- [11] G. Ma, L. He, C. T. Lu, and W. Shao, "Multi-view clustering with graph embedding for connectome analysis," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 127–136, Singapore, November 2017.
- [12] J. Li, C. Chen, H. Tong, and H. Liu, "Multi-layered network embedding," in *Proceedings of the 2018 SIAM International Conference on Data Mining*, SIAM, pp. 684–692, San Diego, CA, USA, May 2018.
- [13] Y. Sun, N. Bui, T. Y. Hsieh, and V. Honavar, "Multi-view network embedding via graph factorization clustering and co-regularized multi-view agreement," in *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1006–1013, IEEE, Singapore, November 2018.
- [14] H. Zhang, L. Qiu, L. Yi, and Y. Song, "Scalable multiplex network embedding," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, vol. 18, pp. 3082–3088, Stockholm, Sweden, July 2018.
- [15] M. Ghorbani, M. S. Baghshah, and H. R. Rabiee, "MGCN: semi-supervised classification in multi-layer graphs with graph convolutional networks," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 208–211, Vancouver, Canada, August 2019.
- [16] X. Chu, X. Fan, D. Yao, Z. Zhu, J. Huang, and J. Bi, "Cross-network embedding for multi-network alignment," in *Proceedings of the World Wide Web Conference*, pp. 273–284, San Francisco, CA, USA, May 2019.
- [17] S. Fan, X. Wang, C. Shi, E. Lu, K. Lin, and B. Wang, "One2multi graph autoencoder for multi-view graph clustering," in *Proceedings of the Web Conference 2020*, pp. 3070–3076, Taipei, Taiwan, April 2020.
- [18] J. Ni, S. Chang, X. Liu, and W. Cheng, "Co-regularized deep multi-network embedding," in *Proceedings of the 2018 World Wide Web Conference*, pp. 469–478, Lyon, France, April 2018.
- [19] B. Du and H. Tong, "Mrmine: multi-resolution multi-network embedding," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 479–488, Beijing, China, November 2019.
- [20] S. Najari, M. Salehi, V. Ranjbar, and M. Jalili, "Link prediction in multiplex networks based on interlayer similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 536, Article ID 120978, 2019.
- [21] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J. P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.
- [22] B. Shaw and T. Jebara, "Structure preserving embedding," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pp. 937–944, Association for Computing Machinery, New York, NY, USA, June 2009.
- [23] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, vol. 14, no. 6, pp. 585–591, 2001.
- [24] S. Cao, W. Lu, and Q. Xu, "Grarep: learning graph representations with global structural information," in *Proceedings of the 24th ACM International Conference on Information and*

- Knowledge Management, CIKM 2015*, pp. 891–900, ACM, Melbourne, Australia, October 2015.
- [25] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “Line: large-scale information network embedding,” in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077, Florence, Italy, May 2015.
 - [26] K. Mallick, S. Bandyopadhyay, S. Chakraborty, R. Choudhuri, and S. Bose, “Topo2vec: a novel node embedding generation based on network topology for link prediction,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1306–1317, 2019.
 - [27] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3111–3119, Lake Tahoe, Nevada, December 2013.
 - [28] T. N. Kipf and M. Welling, “Variational graph auto-encoders,” *Statistics*, vol. 1050, p. 21, 2016.
 - [29] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–1035, Long Beach, CA, USA, December 2017.
 - [30] H. Wang, J. Wang, J. Wang, and M. Zhao, “Graphgan: graph representation learning with generative adversarial nets,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2508–2515, New Orleans, LA, USA, February 2018.
 - [31] D. C. G. Pedronette and L. J. Latecki, “Rank-based self-training for graph convolutional networks,” *Information Processing & Management*, vol. 58, no. 2, Article ID 102443, 2021.
 - [32] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, “Deep graph infomax,” <http://arxiv.org/abs/1809.10341>.
 - [33] H. Hafidi, M. Ghogho, P. Ciblat, and A. Swami, “Graphcl: contrastive self-supervised learning of graph representations,” <http://arxiv.org/abs/2007.08025>.
 - [34] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang, “Deep graph contrastive representation learning,” <http://arxiv.org/abs/2006.04131>.
 - [35] P. Cui, X. Wang, J. Pei, and W. Zhu, “A survey on network embedding,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 5, pp. 833–852, 2018.
 - [36] H. Cai, V. W. Zheng, and K. C. C. Chang, “A comprehensive survey of graph embedding: problems, techniques, and applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1616–1637, 2018.
 - [37] P. Goyal and E. Ferrara, “Graph embedding techniques, applications, and performance: a survey,” *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.
 - [38] D. Jin, Z. Yu, P. Jiao, S. Pan, P. S. Yu, and W. Zhang, “A survey of community detection approaches: From statistical modeling to deep learning,” <http://arxiv.org/abs/2101.01669> CoRR abs/2101.01669.
 - [39] W. Liu, P. Y. Chen, S. Yeung, T. Suzumura, and L. Chen, “Principled multilayer network embedding,” in *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 134–141, IEEE, New Orleans, LA, USA, November 2017.
 - [40] M. Zitnik and J. Leskovec, “Predicting multicellular function through multi-layer tissue networks,” *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, 2017.
 - [41] A. Bagavathi and S. Krishnan, “Multi-net: a scalable multiplex network embedding framework,” in *Proceedings of the International Conference on Complex Networks and their Applications*, pp. 119–131, Springer, Basel, Switzerland, December 2018.
 - [42] C. Park, C. Yang, Q. Zhu, D. Kim, H. Yu, and J. Han, “Unsupervised differentiable multi-aspect network embedding,” <http://arxiv.org/abs/2006.04239>.
 - [43] Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou, and J. Tang, “Representation learning for attributed multiplex heterogeneous network,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’19*, pp. 1358–1368, Association for Computing Machinery, New York, NY, USA, July 2019.
 - [44] M. Gong, W. Liu, Y. Xie, Z. Tang, and M. Xu, “Heuristic 3D interactive walk for multilayer network embedding,” *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020.
 - [45] Q. Meng, T. Jian, J. Shang, R. Xiang, and J. Han, “An attention-based collaboration framework for multi-view network representation learning,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM ’17)*, pp. 1767–1776, Singapore, November 2017.
 - [46] Y. Ma, S. Wang, C. C. Aggarwal, D. Yin, and J. Tang, “Multi-dimensional graph convolutional networks,” in *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 657–665, SIAM, Champaign, IL, USA, May 2019.
 - [47] R. Wang, L. Li, X. Tao, X. Dong, P. Wang, and P. Liu, “Triobased collaborative multi-view graph clustering with multiple constraints,” *Information Processing & Management*, vol. 58, no. 3, Article ID 102466, 2021.
 - [48] C. Park, J. Han, and H. Yu, “Deep multiplex graph infomax: attentive multiplex network embedding using global information,” *Knowledge-Based Systems*, vol. 197, Article ID 105861, 2020.
 - [49] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, “Deep graph infomax,” ICLR, New Orleans, LA, USA, 2019.
 - [50] Y. Sun, S. Wang, T. Y. Hsieh, X. Tang, and V. Honavar, “Megan: a generative adversarial network for multi-view network embedding,” <http://arxiv.org/abs/1909.01084>.
 - [51] J. Coleman, E. Katz, and H. Menzel, “The diffusion of an innovation among physicians,” *Sociometry*, vol. 20, no. 4, pp. 253–270, 1957.
 - [52] D. Manlio, L. Andrea, A. Alex, and R. Martin, “Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems,” *Physical Review X*, vol. 5, no. 1, p. 11027, 2015.
 - [53] C. Stark, “Biogrid: a general repository for interaction datasets,” *Nucleic Acids Research*, vol. 34, no. 90001, pp. 535–539, 2006.
 - [54] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, “Structural reducibility of multilayer networks,” *Nature Communications*, vol. 6, no. 1, p. 6864, 2015.

Research Article

The Comprehensive Contributions of Endpoint Degree and Coreness in Link Prediction

Yang Tian ¹, Yanan Wang ², Hui Tian ¹ and Qimei Cui¹

¹State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

²School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Hui Tian; tianhui@bupt.edu.cn

Received 12 June 2021; Revised 14 July 2021; Accepted 27 July 2021; Published 11 August 2021

Academic Editor: Fei Xiong

Copyright © 2021 Yang Tian 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 past studies, researchers find that endpoint degree, H-index, and coreness can quantify the influence of endpoints in link prediction, especially the synthetical endpoint degree and H-index improve prediction performances compared with the traditional link prediction models. However, neither endpoint degree nor H-index can describe the aggregation degree of neighbors, which results in inaccurate expression of the endpoint influence intensity. Through abundant investigations, we find that researchers ignore the importance of coreness for the influence of endpoints. Meanwhile, we also find that the synthetical endpoint degree and coreness can not only describe the maximal connected subgraph of endpoints accurately but also express the endpoint influence intensity. In this paper, we propose the DCHI model by synthesizing endpoint degree and coreness and the HCHI model by synthesizing H-index and coreness on SRW-based models, respectively. Extensive simulations on twelve real benchmark datasets show that, in most cases, DCHI shows better prediction performances in link prediction than HCHI and other traditional models.

1. Introduction

The research of link prediction aims to find the lost, false, or possible links through the observed network structure and information [1–5]. Therefore, link prediction algorithms have been applied to many fields. For example, link prediction algorithms can remove noise of the networks [6]. Furthermore, link prediction algorithms also can be applied to friends' recommendation on online social networks [7–11] and products' recommendation on e-commerce websites [12–16]. Moreover, link prediction algorithms provide references for biological experiments, which can reduce the cost of experiments [17–19]. In addition, link prediction algorithms can reveal network evolution mechanism and organization pattern [20–22].

To reveal the structure of complex networks, researchers propose a large number of link prediction models. Specifically, the models based on local information have been gained more

attentions. For example, Kossinets [23] finds that two strangers become friends if they have more common friends in social networks. Newman [24] finds that two scientists have more likely to establish cooperation relationship in the future if they have more common cooperators. Based on this phenomenon, researchers propose the common neighbors' model (CN). Based on CN, some researchers propose improved models, such as Salton [25] and LHN-I [26]. Furthermore, according to the different similarity contributions of common neighbors, Adamic and Adar propose the AA model [27]. Zhou et al. [28] propose the resource-allocation model (RA). Moreover, Cannistraci et al. [29] propose that CN, AA, RA, and other algorithms can be weighted by local community information, which can further improve the performances of these models. However, the models based on common neighbors only consider the influence of endpoints on one-step paths. Though further research studies, Lü et al. [30] propose local path model (LP) through considering the influence of endpoints on three-

step paths. In addition, some models consider global information, such as Katz [31] and hierarchical structure model [32]. Besides, some models' consider quasi-local information, which can compensate the defect of low accuracy in local information and high-computational complexity in global information. For example, local random walk model (LRW) [33] considers a random walker within a quasi-local range, and superposed random walk model (SRW) [33] considers the effects of LRW with different path lengths. Based on SRW, HSRW [34] and CSRW [34] models consider the roles of H-index [35] and coreness [36] with different path lengths, respectively. Simple hybrid influence model (SHI) [37] synthetically considers the role of endpoint degree and H-index as hybrid influence with different path lengths.

At present, many link prediction models only consider the degree [33] of endpoints, such as Sørensen [38], LHN [26], LRW [33], and SRW [33]. These models illustrate that the source endpoint can effectively spread its influence to the target endpoint if the source endpoint has more neighbors to connect the target endpoint. Through abundant study, Lü et al. [39] find that H-index shows a better performance to quantify the influence of endpoint than degree and coreness. Zhu et al. [37] find that an endpoint possessing large synthetic degree and H-index can acquire a more extensive maximal connected subgraph, which can help the endpoint to attract other nodes. Through further investigations, we find that the endpoint influence can be expressed by the aggregation degree of neighbors. The large aggregation degree of neighbors illustrates that the endpoint has the extensive maximal connected subgraph, leading to attract more nodes. The aggregation degree of neighbors can only be quantified by coreness of endpoints. Thus, we synthesize the endpoint degree and coreness (or H-index and coreness) to quantify the endpoint influence and build the new link prediction models. Although the SHI model based on the synthetic degree and H-index has been explored, the synthetic endpoint degree and coreness (or the synthetic H-index and coreness) has not been fully verified.

Figure 1 shows a clear illustration. In Figure 1, endpoint b possesses degree = 6, H-index = 3, and coreness = 3, respectively. The influence intensity of endpoint b is size 6 in consideration of only degree. However, only the degree cannot express the depth and scope of the influence of endpoints accurately. Due to the role of coreness in influence of endpoints, synthesizing degree, and coreness or synthesizing H-index and coreness can better quantify the maximal connected subgraph of endpoints and the aggregation degree of neighbors. For endpoint b , the product of degree (H-index) and coreness is 18 (9). Obviously, degree and H-index indicate the different sizes of maximal connected subgraph belonging to endpoint b with the same coreness, leading to different influences of endpoints. Therefore, the prediction performance on the influence of endpoints based on the different quantification index needs to be further explored.

In real world, we find many phenomena to confirm our idea. For example, in Weibo, an ordinary individual possesses the limited influence because he/she only has many individual followers from colleagues, classmates, relatives, or friends, indicating that he/she only has large degree.

However, public figures possess extensive and strong influence because they have large number of fan club, indicating that they have large coreness to strengthen their influence. In scientists' collaboration network, if a scientist only cooperates with many scholars, meaning he/she has large degree but small coreness, the scientist cannot be known by more researchers and can hardly further attract them to cooperate. In e-commerce network, the applicability of products depends on purchase groups with similar identities, such as male/female group, student group, and teacher group, which shows the importance of aggregation degree. In paper-citation network, the value of a paper depends on the citation of researchers in the same field, not the citation of researchers in the different fields.

In summary, in this paper, we define the hybrid influence of synthetic degree and coreness (synthetic H-index and coreness) to redefine the SRW and propose two improved models DCHI and HCHI to further explore the accuracy of link prediction. Experimental results on twelve real networks show that DCHI exhibits better performances of link prediction.

The rest of this paper is organized as follows. In Section 2, we build two models based on the synthetic degree and coreness and the synthetic H-index and coreness, respectively. In Section 3, the thirteen benchmark experimental datasets are introduced. In Section 4, a link prediction metric and eight mainstream baselines are described, respectively. In Section 5, the experimental results are discussed. In Section 6, the conclusion is described.

2. Models Based on Hybrid Influence of Endpoints

Firstly, we study link prediction models in an undirected simple network $G(V, E)$, where E is the set of links ($|E|$ denotes the number of all edges.) and V refers to the set of nodes. Multiple links and self-connections are eliminated. For every pair of nodes, $x, y \in V$, a score, s_{xy} , is given to calculate the probability of their future connection. In this paper, we set the similarity value as a score directly, and a larger score illustrates that the potential link has more possibility to be found.

Secondly, we show two models based on the degree (SRW [33]) and the synthetic degree and H-index (SHI [37]) separately as follows.

2.1. SRW Model. Liu et al. [33] build the similarity model using random walk, which finds all intermediate nodes sequentially between two endpoints according to a Markov chain with one-step transmission probability $p_{xy} = a_{xy}/k_x$, where k_x represents the degree of node x and $a_{xy} = 1$ if node x successfully connects y and $a_{xy} = 0$ if not. The sequence of node with t -step between x and y is expressed as $\{x = x_0 = y_t, x_1 = y_{t-1}, \dots, x_{t-1} = y_1, x_t = y_0 = y\}$. Thus, the t -step transmission probability from x to y is denoted by $\pi_{xy}(t) = \prod_{i=0}^{t-1} p_{x_i x_{i+1}}$ and $\pi_{yx}(t) = \prod_{i=0}^{t-1} p_{y_i y_{i+1}}$. Importantly, Liu et al. consider the degree k_x and k_y , to quantify the influence of endpoints and define the SRW as

$$s_{xy}^{SRW}(t) = \sum_{l=2}^t \left[\frac{k_x}{2|E|} \cdot \pi_{xy}(l) + \frac{k_y}{2|E|} \cdot \pi_{yx}(l) \right], \quad (1)$$

where k_x and k_y denote the degree of endpoint x and y , respectively, and $|E|$ indicates the number of links in the network. $(k_x/2|E|)$ and $(k_y/2|E|)$ describe the influence of endpoint x and y , respectively.

2.2. SHI Model. Zhu et al. [37] find that the H-index can represent the maximal connected subgraph of endpoints and describe the influence intensity. Thus, Zhu et al. simply synthesize degree and H-index as the hybrid influence of endpoints and replace the degree in SRW to define a simple hybrid influence model (SHI) as

$$s_{xy}^{SHI}(t) = \sum_{l=2}^t \left[\frac{\sqrt{k_x \times h_x}}{2|E|} \pi_{xy}(l) + \frac{\sqrt{k_y \times h_y}}{2|E|} \pi_{yx}(l) \right], \quad (2)$$

where $(\sqrt{k_x \times h_x}/2|E|)$ and $(\sqrt{k_y \times h_y}/2|E|)$ denote the hybrid influence of node x and y based on synthetical degree and H-index, respectively.

Although the endpoint degree and H-index can quantify the endpoint influence, they only represent the number of neighbors and the maximal connected subgraph of endpoints separately, ignoring the influence intensity of endpoints. The influence intensity of endpoints can be expressed by the coreness of endpoints because the coreness can quantify the aggregation degree of neighbors which represents the endpoint influence intensity. Thus, we consider the role of coreness for endpoint influence. Finally, we build two models based on synthetical degree and coreness (DCHI) and synthetical H-index and coreness (HCHI) separately as follows.

2.3. DCHI Model. Through the explanation in Section 1 and the illustration in Figure 1, we synthesize degree and coreness to quantify the influence of endpoints and replace the degree in SRW to build a new model DCHI as

$$s_{xy}^{DCHI} = \sum_{l=2}^t \left[\frac{\sqrt{k_x \times c_x}}{2|E|} \pi_{xy}(l) + \frac{\sqrt{k_y \times c_y}}{2|E|} \pi_{yx}(l) \right], \quad (3)$$

where $(\sqrt{k_x \times c_x}/2|E|)$ and $(\sqrt{k_y \times c_y}/2|E|)$ denote the hybrid influence of node x and y based on synthetical degree and coreness, respectively.

2.4. HCHI Model. Furthermore, we synthesize H-index and coreness to quantify the influence and replace the degree in SRW to build a new model HCHI as

$$s_{xy}^{HCHI} = \sum_{l=2}^t \left[\frac{\sqrt{h_x \times c_x}}{2|E|} \pi_{xy}(l) + \frac{\sqrt{h_y \times c_y}}{2|E|} \pi_{yx}(l) \right], \quad (4)$$

where $(\sqrt{h_x \times c_x}/2|E|)$ and $(\sqrt{h_y \times c_y}/2|E|)$ denote the hybrid influence of node x and y based on synthetical H-index and coreness, respectively.

3. Experimental Data

In this section, we introduce 12 real network datasets to prepare the following experiments. (1) US Air97 (USAir) [40] represents the US airline network. (2) Yeast PPI (Yeast) [41] represents the yeast network of relationship between proteins. (3) Food Web (Food) [42] represents the relations of carbon exchanges in the cypress wetlands of Florida ecosystem. (4) Power Grid (Power) [43] represents the western US's electrical power transmission network. (5) NetScience (NS) [44] represents partnerships between scientists in publishing papers concerning the subject of networks. (6) Jazz [45] represents the networks of Jazz musicians. (7) e-mail network (e-mail) [46] represents e-mail communication network of University Rovira i Virgili (URV) in Spain. (8) Slavko [47] represents the friendship network of Slavko Zitnik on Facebook. (9) UC Irvine dealing with social network (UCsocial) [48] represents an online social network composed of students in the University of California, Irvine. (10) Infectious (Infec) [49] represents the offline contact network of visitors in the course of the exhibition named "Infectious: Stay Away" in the Science Gallery in Dublin, 2009. (11) EuroSiS web (EuroSiS) [50] represents interactions network between Science in Society actors from twelve European countries. (12) C. elegans (CE) [43] represents the network of neurons in the C. elegans worm. Table 1 lists the mentioned networks fundamental topological features.

To achieve preprocess, arcs are changed as nondirectional links, and loops and multiedges are eliminated to ensure the network unweighed and undirected. Subsequently, the largest linked simplified network subgraph is extracted to make sure the connectivity.

In the beginning, the set of network links is divided into the training set E^T containing 90% links in a random manner, and the testing set E^P containing 10% links, while the connectivity of E^T is ensured [1]. Besides, 30 divisions are identically and separately conducted on the network. Next, experimental processes are performed over the 30 separated training and testing sets, the averaged accuracy is achieved in a statistical manner, and metrics is recalled more than 30 times realization.

4. Experimental Methods

4.1. Metric. AUC [36], a metric of accuracy, can be interpreted as the probability that a potential link (a link in E^P) ranks a higher score than a nonexistent link (a link in $U \setminus E$, where U denotes the universal link set). In the specific implementation, among n independent comparisons if the potential link ranks higher in n' times and the same as the nonexistent link in n'' times, and the total score accumulates n' and $0.5n''$. After that, AUC expresses the averaged score over n -time comparisons as

$$AUC = \frac{n' + 0.5n''}{n}. \quad (5)$$

AUC evaluates the performance of a model globally. If all scores originate from an independent and identical

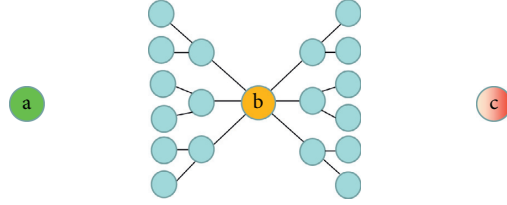


FIGURE 1: Illustration of the influences of endpoint b based on degree, H-index, and coreness. Endpoint b has degree = 6, H-index = 3, and coreness = 3. Node a and c represent two target endpoints, respectively. The blue nodes represent the medial nodes between endpoint b and two target endpoints.

distribution, the value should equal to 0.5. Therefore, the extent to which the accuracy exceeds 0.5 suggests how much better a model performs than pure chance.

4.2. *Baselines.* Comparatively, we introduce eight fundamental models as follows:

- (1) Common neighbors (CN) [24] describe the similarity between endpoints by calculating the number of common neighbors, defined as

$$s_{xy}^{\text{CN}} = |\Gamma(x) \cap \Gamma(y)|, \quad (6)$$

where $\Gamma(X)$, $X \in \{x, y\}$, represents the set of neighbors of endpoint X and $|\Gamma(x) \cap \Gamma(y)|$ refers to the number of common neighbors of endpoints x and y .

- (2) Adamic/Adar (AA) [27], based on CN, suppress the contributions of common neighbors with big degree by applying the inverse logarithm, which is defined as

$$s_{xy}^{\text{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k_z)}, \quad (7)$$

where k_z represents the degree of node z .

- (3) Resource-Allocation (RA) [28], analogous to AA, suppresses the large degree of common neighbors by applying the reciprocal of the degrees of common neighbors, which is defined as

$$s_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \quad (8)$$

- (4) Local Path Index (LP) [30] considers the similarity on two-step and three-step paths between endpoints simultaneously, with the two-step paths preferred, which are defined as

$$S^{\text{LP}} = A^2 + \varepsilon A^3, \quad (9)$$

where A represents the adjacency matrix and ε is a punishment parameter.

- (5) Superposed Random Walk (SRW) [33] is introduced in Section 2.

- (6) CSRW [34] exploits the coreness to quantify the influence of endpoint and replace the degree influence in SRW, which is defined as

$$s_{xy}^{\text{CSRW}}(t) = \sum_{l=2}^t \left[\frac{c_x}{2|E|} \cdot \pi_{xy}(l) + \frac{c_y}{2|E|} \cdot \pi_{yx}(l) \right], \quad (10)$$

where c_x and c_y represent the coreness of node x and y , respectively.

- (7) HSRW [34] exploits the H-index to quantify the influence of endpoint and replace the degree influence in SRW, defined as

$$s_{xy}^{\text{HSRW}}(t) = \sum_{l=2}^t \left[\frac{h_x}{2|E|} \cdot \pi_{xy}(l) + \frac{h_y}{2|E|} \cdot \pi_{yx}(l) \right], \quad (11)$$

where h_x and h_y represent the H-index of node x and y , respectively.

- (8) Simple hybrid influence (SHI) [33] is introduced in Section 2.

5. Results and Discussion

To explore the prediction performances of the proposed models, extensive simulations are conducted on 12 real datasets. Through comparisons with several main baselines in terms of accuracy metric, we obtain the experimental results on the models and discuss the findings in the following.

SHI, HCHI, and DCHI models mainly consider two aspects: random walk on paths and hybrid influences of endpoints. Through simulations, the experimental results show that the number of steps in random walk between endpoints can affect the accuracy of link prediction. For illustrating the changes of prediction accuracy on the number of steps t , we plot the relation curves in Figure 2.

In Figure 2, SHI (synthetical degree and H-index), HCHI (synthetical degree and coreness), and DCHI (synthetical H-index and coreness) models show their prediction performances on the random steps t , and they exhibit different optimal accuracies at certain number of steps t , respectively. Specifically, SHI shows optimal AUC values at $t = 15$ in food, power, NS, e-mail, UCsocial, and Eurosis, $t = 5$ in USAir and CE, $t = 3$ in yeast, $t = 2$ in Jazz, $t = 6$ in Slavko, and $t = 9$ in

TABLE 1: The basic topological features of the twelve benchmark networks. There are the properties: $|V|$ denoting the number of nodes, $|E|$ denoting the number of links, $\langle k \rangle$ denoting the average degree, $\langle d \rangle$ representing the average distance, C representing the clustering coefficient, r indicating the assortativity coefficient, and H indicating the degree heterogeneity and defined as $H = \langle k^2 \rangle / \langle k \rangle^2$.

Nets	$ V $	$ E $	$\langle k \rangle$	$\langle d \rangle$	C	r	H
Usair	332	2128	12.81	2.74	0.749	-0.208	3.36
Yeast	2370	10904	9.2	5.16	0.378	0.469	3.35
Food	128	2075	32.42	1.78	0.334	-0.112	1.24
Power	4941	6594	2.669	15.87	0.107	0.003	1.45
NS	1461	2742	3.75	5.82	0.878	0.461	1.85
Jazz	198	2742	27.7	2.24	0.633	0.02	1.4
E-mail	1133	5451	9.62	3.61	0.254	0.078	1.94
Slavko	334	2218	13.28	3.05	0.488	0.247	1.62
Ucsocial	1893	13825	14.62	3.06	0.138	-0.188	3.81
Infec	410	2765	13.49	3.63	0.467	0.226	1.39
Eurosis	1272	6454	10.15	3.86	0.382	-0.012	2.46
CE	453	2025	8.94	2.66	0.655	-0.225	4.49

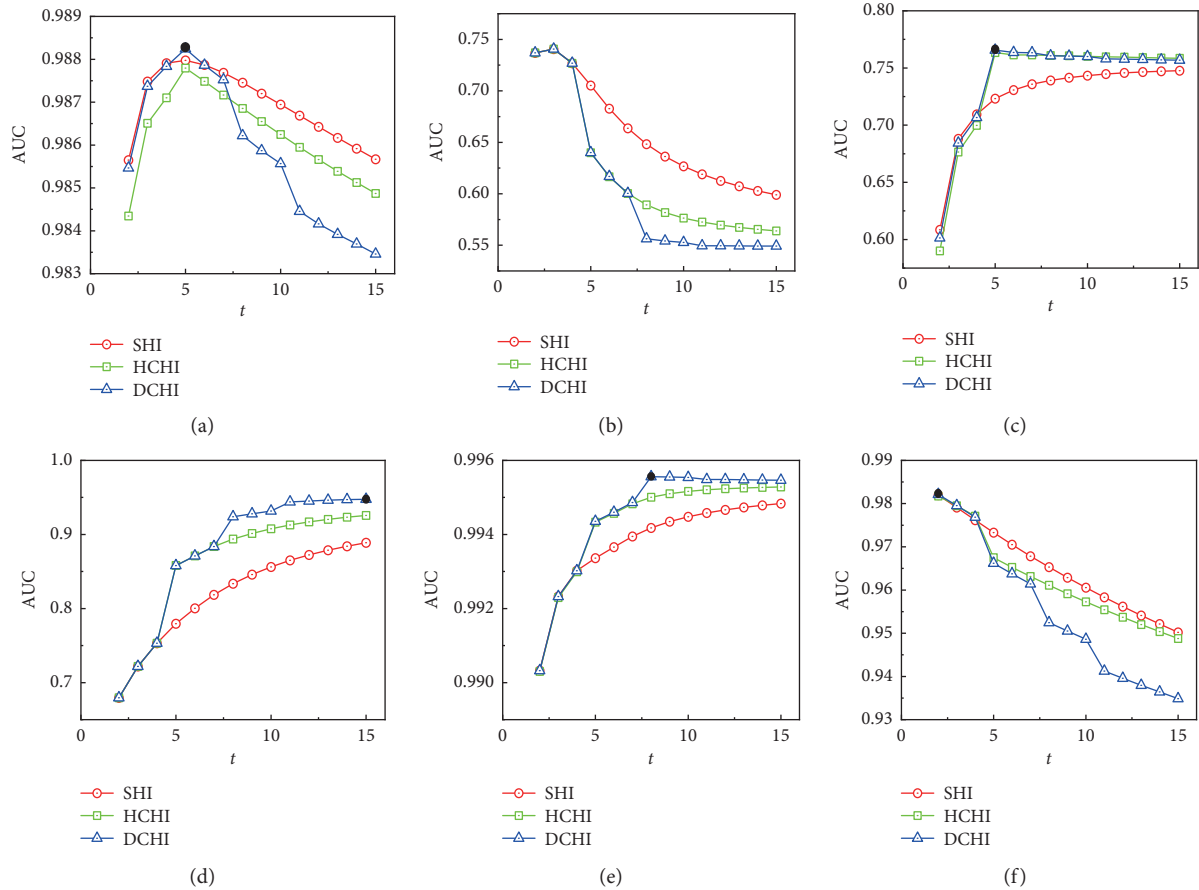


FIGURE 2: Continued.

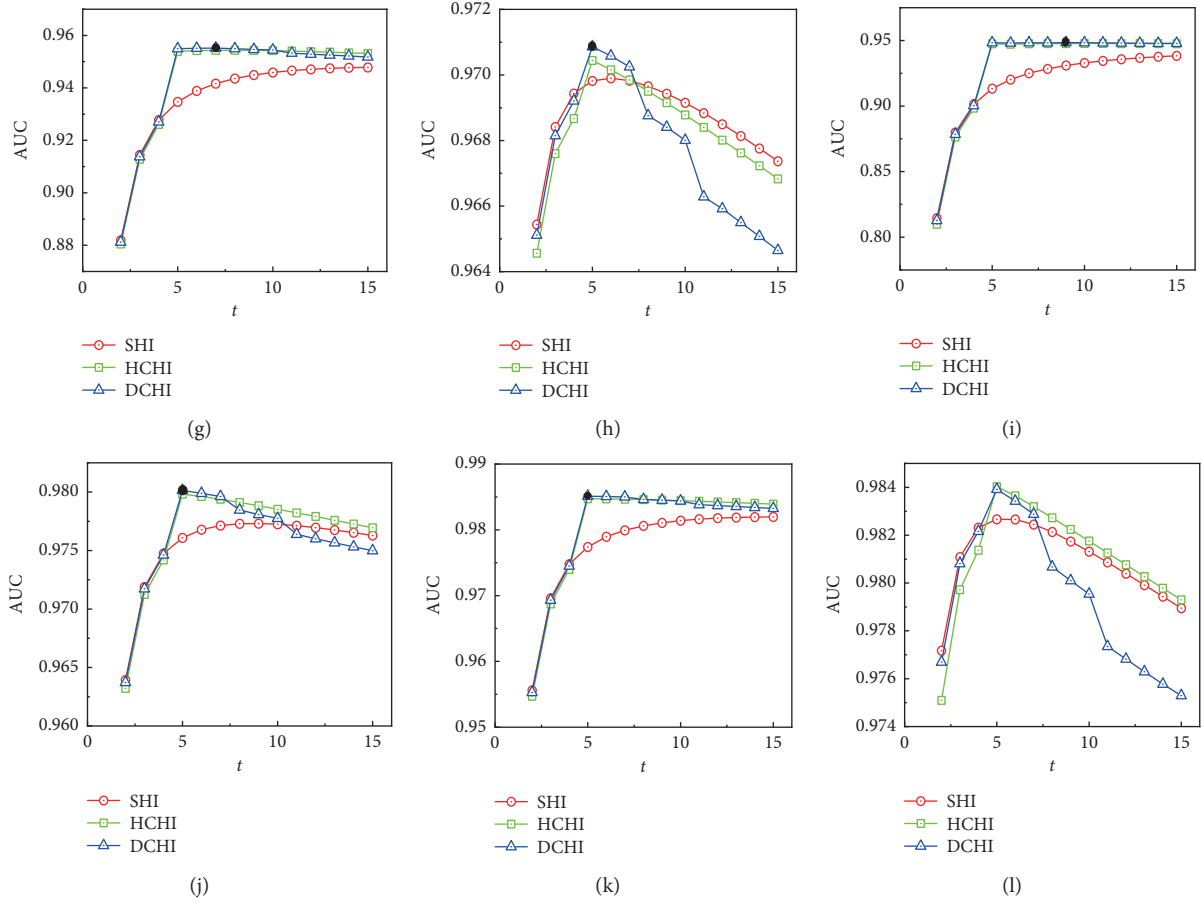


FIGURE 2: Pattern illustration of accuracy metric AUC on the number of random-walk steps (t). The relation curves between AUC and t in terms of SHI (red circles), HCHI (green squares), and DCHI (blue triangle) are provides on twelve real datasets. In most cases, DCHI obtains the most optimal AUC values with the least number of steps (t) which is 5 in (a) USAir; (b) yeast; (c) food; (d) power; (e) NS; (f) jazz; (g) e-mail; (h) Slavko; (i) UCsocial; (j) Infec; (k) Eurosis; (l) CE.

Infec. Obviously, the optimal number of steps on SHI mainly appears on the long path $t = 15$, illustrating that long paths can further facilitate the hybrid influence spreading based on the degree and the H-index. However, HCHI and DCHI all show optimal AUC values at $t = 5$ in USAir, Food, Slavko, Infec, Eurosis, and CE, illustrating that quasi-local paths can further facilitate the hybrid influence spreading based on H-index and coreness or degree and coreness. Importantly, we find that the influence concerning coreness can easily leak in the random-walk process on longer paths, which leads to weaken the intensity of influence spreading between endpoints. However, in power, the prediction performances of HCHI and DCHI reach the optimal value at $t = 15$ because power network includes large numbers of long paths with average distance $\langle d \rangle = 15.87$ much longer than other datasets (referring to Table 1). In addition, DCHI, compared with SHI and HCHI, has larger size of maximal connected subgraph and more paths to spread the hybrid influence of endpoints. Therefore, DCHI shows the best prediction performances in ten datasets (black mark on each dataset) except yeast and CE.

In addition, we compare HCHI and DCHI with eight link prediction models CN, AA, RA, LP, SRW, CSRW, HSRW, and SHI. To exhibit the experimental results, we show the averaged AUC values over 30 simulations in Table 2 for all models. The underlined bold fonts represent the best AUC values in each dataset and the numbers in parenthesis indicate the optimal random walk steps t , at which HCHI and DCHI obtain the optimal AUC values in eight datasets altogether.

As can be seen from Table 2, optimal values on seven datasets exist in DCHI with Power, NS, Jazz, Email, Slavko, Infec, and Eurosis. In contrast, local models CN, AA, and RA show worst prediction performances because they only consider the local paths and ignore the influence of endpoints. Then, optimal values on three datasets exist in LP with yeast, food, and UCsocial, illustrating that the quasi-local paths can limitedly promote the prediction performances. And then, SRW, CSRW, and HSRW also show worst performances because they only consider separately the contributions of degree, coreness, and H-index, meaning that degree, coreness, and H-index all cannot quantify the

TABLE 2: AUC on the twelve benchmark networks with $L = 100$, where L denotes the number of the candidate links to measure the prediction accuracy in each data set. Every data point is an average over 30 independent realizing processes, and every point represents a random 90%–10% division of training set and testing set. The values in the parentheses indicate the corresponding optimal number of steps. All the present results represent the optimal cases by (if any) adjusting the coefficients.

AUC	CN	AA	RA	LP	SRW	CSRW	HSRW	SHI	HCHI	DCHI
USair	0.977771	0.984243	0.986611	0.966917	0.988464 (4)	0.987171 (6)	0.987346 (5)	0.987975 (5)	0.987793 (5)	0.988243 (5)
Yeast	0.736946	0.737078	0.737075	0.742813	0.740707 (3)	0.740543 (3)	0.740576 (3)	0.740644 (3)	0.740553 (3)	0.740622 (3)
Food	0.616391	0.617681	0.619521	0.827879	0.749576 (15)	0.745886 (15)	0.745153 (15)	0.747621 (15)	0.763479 (5)	0.765610 (5)
Power	0.679613	0.679723	0.67968	0.763982	0.888964 (15)	0.888625 (15)	0.888724 (15)	0.888844 (15)	0.925632 (15)	0.947420 (15)
NS	0.990227	0.990345	0.990355	0.993998	0.994864 (15)	0.994783 (15)	0.994802 (15)	0.994836 (15)	0.995283 (15)	0.995562 (8)
Jazz	0.972242	0.976438	0.981334	0.935342	0.981301 (2)	0.981 (2)	0.982014 (2)	0.981992 (2)	0.981740 (2)	0.982131 (2)
E-mail	0.881955	0.883162	0.882471	0.942283	0.947927 (15)	0.946909 (15)	0.947445 (15)	0.947803 (15)	0.954343 (8)	0.955154 (7)
Slavko	0.964026	0.965934	0.965678	0.959088	0.970067 (6)	0.968883 (7)	0.969502 (6)	0.969898 (6)	0.970442 (5)	0.970861 (5)
UCsocial	0.813094	0.817405	0.81764	0.948516	0.939542 (15)	0.936628 (15)	0.937387 (15)	0.938368 (15)	0.947678 (11)	0.948375 (9)
Infec	0.962318	0.964223	0.964209	0.970787	0.977274 (8)	0.976935 (10)	0.977107 (9)	0.977311 (9)	0.979832 (5)	0.980139 (5)
Eurosis	0.955269	0.95659	0.956081	0.978458	0.981863 (14)	0.981535 (15)	0.98186 (15)	0.981694 (15)	0.984722 (5)	0.985111 (5)
CE	0.951545	0.977089	0.97905	0.932316	0.982963 (5)	0.982176 (7)	0.982329 (6)	0.982663 (5)	0.984026 (5)	0.983915 (5)

influence of endpoints comprehensively. Finally, we focus on the performances of SHI, HCHI, and DCHI. In twelve datasets, there are seven optimal performances in DCHI. DCHI, compared with SHI and HCHI, shows the effective influence of endpoints (e.g., extensive maximal connected subgraph of endpoints and aggregation degree of neighbors) and finds sufficient paths between two unconnected endpoints. Therefore, because the synthetical degree and coreness as hybrid influence can be a good quantification index, DCHI can better enhance prediction accuracy than SHI and HCHI in many cases of link prediction.

Besides, the low computation complexity is a necessary condition in link prediction. The time complexity of the product of two $N \times N$ matrices is $O(N^3)$. According to the definitions of the baseline models, CN, AA, and RA possess the time complexity of $O(N^3)$ and LP, SRW, CSRW, HSRW, and SHI have $M \times O(N^3)$ with coefficient M . Although HCHI and DCHI have the same time complexity $M \times O(N^3)$, two proposed models, especially DCHI, show greater performance improvement. Therefore, the proposed models show a better performance with no increase in complexity.

6. Conclusions

At present, researchers pay more attention to the contributions of the influence of endpoints for link prediction based on local, quasi-local, or global similarity. To quantify the influence of endpoints, researchers consider the degree, H-index, or coreness separately, which all cannot evaluate the influence of endpoints comprehensively. Specifically, the endpoint degree only represents the number of neighbors of endpoints, but cannot describe the maximal connected subgraph. The H-index can express the maximal connected subgraph of endpoints to quantify the influence scope. However, the endpoint degree and H-index cannot quantify the influence intensity of endpoints and result in incomplete influence expression. We find that the coreness can represent the aggregation degree of endpoints, which can quantify the influence intensity of endpoints accurately.

Through abundant investigations, we find that the synthetical degree and coreness and the synthetical H-index

and coreness can quantify the influence of endpoints accurately and comprehensively. Therefore, we synthesize degree (H-index) and coreness as the hybrid influence of endpoints and replace the degree in SRW to build two models DCHI and HCHI.

We explore the prediction performances of DCHI and HCHI by the comparisons among CN, AA, RA, LP, SRW, CSRW, HSRW, and SHI on twelve real datasets. As a result, we show that DCHI obviously outperform other models on the metric AUC and do not increase computational complexity. The outstanding improvement in accuracy illustrates the synthetical degree and coreness as hybrid influence of endpoints can describe the endpoint influence intensity accurately and can attract more nodes to produce links.

Although our models have been verified on the datasets, the models only make a simple synthesis between endpoint degree, H-index, and coreness. We find degrees differ in different networks, and so do H-indices and coreness. The network heterogeneity characterized by heterogeneous degrees, H-indices, and coreness directly results in heterogeneous influences. And, we find that endpoints in network with smaller heterogeneous influence can attract each other more likely. For such characteristic, we will further carry out research on the heterogeneous hybrid influence model based on DCHI and HCHI. In the future research studies, the impact of heterogeneous complex networks will become a crucial problem.

In addition, our study may provide new findings relating to link prediction based on similarity in future. Our research results can be applied to friends' recommendation, products' recommendation, scientists' cooperation, biological experiments, and so on.

Data Availability

The data used to support the findings of the study are available at <http://vlado.fmf.uni-lj.si/pub/networks/data/> and <http://snap.stanford.edu/data/index.html>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61471060) and Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

- [1] L. Lü and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and Its Applications*, vol. 390, 2010.
- [2] B. Barzel and A. L. Barabási, "Network link prediction by global silencing of indirect correlations," *Nature Biotechnology*, vol. 31, no. 8, pp. 720–725, 2013.
- [3] L. Lü, L. Pan, T. Zhou, Y. C. Zhang, and H. Stanley, "Toward link predictability of complex networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, 2325 pages, 2015.
- [4] Y. Wang, Y. Wang, X. Lin, and W. Wang, "The influence of network structural preference on link prediction," *Discrete Dynamics in Nature and Society*, vol. 2020, Article ID 6148273, 9 pages, 2020.
- [5] K. Chi, G. Yin, Y. Dong, and H. Dong, "Link prediction in dynamic networks based on the attraction force between nodes," *Knowledge-Based Systems*, vol. 181, Article ID 104792, 2019.
- [6] M. Newman, "Network structure from rich but noisy data," *Nature Physics*, vol. 14, 6 pages, 2018.
- [7] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," *ACM Transactions on the Web*, vol. 6, no. 2, pp. 1–33, 2012.
- [8] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: the state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1–38, 2014.
- [9] L. Pan, T. Zhou, L. Lü, and C. K. Hu, "Predicting missing links and identifying spurious links via likelihood analysis," *Scientific Reports*, vol. 6, no. 1, Article ID 22955, 2016.
- [10] D. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization," *Future Generation Computer Systems*, vol. 78, pp. 430–439, 2017.
- [11] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 1–13, 2018.
- [12] L. Lü, M. Medo, C. H. Yeung, Y. C. Zhang, and Z. T. Zhang, "Recommender systems," *Physics Reports-Review Section of Physics Letters*, vol. 519, pp. 1–49, 2012.
- [13] A. Kumar, S. S. Singh, K. Singh, and B. Biswas, "Link prediction techniques, applications, and performance: a survey," *Physica A: Statistical Mechanics and Its Applications*, vol. 553, Article ID 124289, 2020.
- [14] W. Wang, Q. H. Liu, J. Liang, Y. Hu, and T. Zhou, "Co-evolution spreading in complex networks," *Physics Reports*, vol. 820, pp. 1–15, 2019.
- [15] X. Chen, K. Gong, R. Wang, S. Cai, and W. Wang, "Effects of heterogeneous self-protection awareness on resource-epidemic coevolution dynamics," *Applied Mathematics and Computation*, vol. 385, Article ID 125428, 2020.
- [16] X. Chen, Q. Liu, R. Wang, Q. Li, and W. Wang, "Self-awareness-based resource allocation strategy for containment of epidemic spreading," *Complexity*, vol. 2020, Article ID 3256415, 12 pages, 2020.
- [17] M. Angeles Serrano and F. Sagués, "Network-based scoring system for genome-scale metabolic reconstructions," *BMC Systems Biology*, vol. 5, 76 pages, 2011.
- [18] M. Á. Serrano, M. Boguñá, and F. Sagués, "Uncovering the hidden geometry behind metabolic networks," *Molecular bioSystems*, vol. 8, no. 3, pp. 843–850, 2012.
- [19] X. Chen, R. Wang, D. Yang, J. Xian, and Q. Li, "Effects of the awareness-driven individual resource allocation on the epidemic dynamics," *Complexity*, vol. 2020, Article ID 8861493, 12 pages, 2020.
- [20] W. Wang, Q. Zhang, and T. Zhou, "Evaluating network models: a likelihood analysis," *Europhysics Letters (Epl)*, vol. 98, pp. 28004–28009, 2011.
- [21] F. Tan, Y. Xia, and B. Zhu, "Link prediction in complex networks: a mutual information perspective," *PLoS One*, vol. 9, no. 9, Article ID e107056, 2014.
- [22] B. Zhu and Y. Xia, "Link prediction in weighted networks: a weighted mutual information model," *PLoS One*, vol. 11, no. 2, Article ID e0148265, 2016.
- [23] G. Kossinets, "Effects of missing data in social networks," *Social Networks*, vol. 28, no. 3, pp. 247–268, 2006.
- [24] M. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, Article ID 025102, 2001.
- [25] M. Davenport, "Introduction to modern information retrieval," *Journal of the Medical Library Association: JMLA*, vol. 100, no. 1, 75 pages, 3rd edition, 2012.
- [26] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, Article ID 026120, 2006.
- [27] L. A. Adamic and A. Eytan, "Friends and neighbors on the web," *Social Networks*, vol. 25, pp. 211–230, 2003.
- [28] T. Zhou, L. Lü, and Y. C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.
- [29] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Scientific Reports*, vol. 3, no. 1, 1613 pages, 2013.
- [30] L. Lü, C. H. Lü, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, Article ID 046122, 2009.
- [31] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, pp. 39–43, 1953.
- [32] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.
- [33] W. Liu and L. Lü, "Link prediction based on local random walk," *Europhysic Letter*, vol. 89, Article ID 58007, 2010.
- [34] X. Zhu, Y. Yang, L. Li, and S. Cai, "Roles of degree, h-index and coreness in link prediction of complex networks," *International Journal of Modern Physics B*, vol. 32, Article ID 1850197, 2018.
- [35] J. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 102, pp. 16569–16572, 2005.
- [36] M. Kitsak, L. K. Gallos, S. Havlin, and F. Liljeros, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, pp. 888–893, 2010.

- [37] X. Zhu, W. Li, H. Tian, and S. Cai, “Hybrid influence of degree and h-index in the link prediction of complex networks,” *EPL*, vol. 122, Article ID 68003, 2018.
- [38] T. Sørensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons,” *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*, vol. 5, 1948.
- [39] L. Lü, T. Zhou, Q. M. Zhang, and H. Stanley, “The h-index of a network node and its relation to degree and coreness,” *Nature Communications*, vol. 7, Article ID 10168, 2016.
- [40] V. Batagelj and A. Mrvar, *Pajek-Program For Large Network Analysis*, University of Ljubljana, Ljubljana, Slovenia, 1999.
- [41] D. Bu, Y. Zhao, L. Cai, and H. Xue, “Topological structure analysis of the protein-protein interaction network in budding yeast,” *Nucleic Acids Research*, vol. 31, pp. 2443–2450, 2003.
- [42] C. Melian and J. Bascompte, “Food web cohesion,” *Ecology*, vol. 85, pp. 352–358, 2004.
- [43] G. Yan, T. Zhou, B. Hu, Z. Q. Fu, and B. Wang, “Efficient routing on complex networks,” *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, Article ID 046108, 2006.
- [44] P. Holme and M. Newman, “Nonequilibrium phase transition in the coevolution of networks and opinions,” *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, Article ID 056108, 2006.
- [45] P. Gleiser and L. Danon, “Community structure in jazz,” *Advances in Complex Systems*, vol. 6, 2003.
- [46] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, “Self-similar community structure in a network of human interactions,” *Physical Review E*, vol. 68, 2003.
- [47] T. Opsahl and P. Panzarasa, “Clustering in weighted networks,” *Social Networks*, vol. 31, 2009.
- [48] N. Blagus, L. Subelj, and M. Bajec, “Self-similar scaling of density in complex real-world networks,” *Physica A: Statistical Mechanics and Its Applications*, vol. 391, 2012.
- [49] L. Isella, J. Stehle, A. Barrat, C. Cattuto, and J. F. Pinton, “What’s in a crowd? analysis of face-to-face behavioral networks,” *Journal of Theoretical Biology*, vol. 271, 2011.
- [50] B. Ermiş, E. Acar, and A. Cemgil, “Link prediction in heterogeneous data via generalized coupled tensor factorization,” *Data Mining and Knowledge Discovery*, vol. 29, pp. 203–236, 2013.

Research Article

Link Prediction Based on the Derivation of Mapping Entropy

Hefei Hu ¹, Yanan Wang,¹ Zheng Li,² Yang Tian ³ and Yuemei Ren⁴

¹School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

²Network Department of China Mobile Wenzhou Branch, Wenzhou 325000, China

³State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

⁴School of Electronic Information Engineering, Henan Polytechnic Institute, Nanyang 473000, China

Correspondence should be addressed to Hefei Hu; hufefei@bupt.edu.cn

Received 10 June 2021; Accepted 27 July 2021; Published 2 August 2021

Academic Editor: Fei Xiong

Copyright © 2021 Hefei Hu 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 algorithms based on topological similarity play an important role in link prediction. However, most of traditional algorithms based on the influences of nodes only consider the degrees of the endpoints which ignore the differences in contribution of neighbors. Through generous explorations, we propose the DME (derivation of mapping entropy) model concerning the mapping relationship between the node and its neighbors to access the influence of the node appropriately. Abundant experiments on nine real networks suggest that the model can improve precision in link prediction and perform better than traditional algorithms obviously with no increase in time complexity.

1. Introduction

A large number of complicated systems in nature can be described by complex networks [1]. The nodes in the network represent the individuals in the real system, and the edges connecting two nodes represent the relationships of the individuals in the real system. The existing networks in the world can be divided into social networks, biological networks [2, 3], and so on. Link prediction evaluates the possibility of the link between two nodes in the network by the known network structure or node attributes. Through link prediction, we can find the links existing but unknown in the network which misses some data. Besides, we can also predict the possible links in the coming evolution of the network [4]. Link prediction plays an important role in practical application. For instance, through link prediction, the unknown interaction between proteins is predicted, which avoids the high experimental cost [5]. Furthermore, it also plays a role in user recommendation [6].

In the early period, most researchers engaged in link prediction focused on the similarity of the attributes of nodes such as age, occupation, interest, and so on [2, 7] to

judge the possibility of links. This method can achieve high-precision prediction. However, it is difficult to extract the attributes of nodes in complex networks, and the reliability of information is hard to assure [8]. So, researchers turned their attention to the study of network structure [9, 10] which has relatively low computational complexity.

Algorithms based on the similarity of the network structure can be divided into three categories: local similarity algorithm, global similarity algorithm, and quasilocal similarity algorithm according to the path length [11]. The core idea of local similarity is common neighbors. On this basis, considering the influence of endpoints from different angles, a variety of local similarity indices is derived. For instance, common neighbor (CN) index [12] considers that if two nodes have more common neighbors, then they are more likely to have connected edges. Adamic-Adar (AA) index [13] holds the idea that the common neighbor with small degree has a greater contribution. Accordingly, each node is given a weight. Because these local similarity indices only consider the local structure of the network which lead to low precision, the third-order and higher-order path similarity indices were proposed such as Katz index [14]. Katz index

considers all paths of two unconnected nodes in the case of short path priority. There is no doubt that this greatly increases the computational complexity. Compared with the above two kinds of algorithms, the quasilocal similarity algorithm with moderate complexity and precision is more and more widely applied. Superposed random walk (SRW) index [15] is one of the indices based on the Markov model.

In traditional algorithms, only the degree of endpoint is considered when we evaluate the influence of it. This considers the influences of neighbor nodes to the same extent which loses the impacts of indirect neighbors [16, 17]. In fact, due to the different degrees of neighbors, their influence on the endpoint should be different. The larger the degree, the greater the influence. However, taking global nodes into account will increase the complexity of the algorithm, and the result is not necessarily good. Because the influence of endpoint is limited, it only has great influence on nearby neighbors. Therefore, this paper proposes to use the derivation of mapping entropy (DME) of node to represent the influence, which represents the mapping relation between a node and its neighbors. It considers not only the weight of the node but also the weight of its neighbors. Figure 1 shows a clear illustration.

On the basis of above discussion, we improve the SRW model, taking the influence of indirect neighbors into account. Through extensive experiments on nine complex networks, the results show that DME can achieve higher precision than traditional algorithms in most cases.

The rest of paper is organized as follows. In Section 2, we propose a new model based on the DME index. In Section 3, we introduce 9 complex networks and experimental approaches. In Section 4, five classic models are introduced as reference. In Section 5, results and analysis are presented. In Section 6, we arrive at a conclusion of our study.

2. Model Based on the Derivation of Mapping Entropy

2.1. Network Model. $G(V, E)$ is defined as a network, where V is the node set and E is the edge set. The total number of nodes is N and the total number of edges is E . The universal set U can have $(N \times (N - 1))/2$ links. The method of link prediction is to give a score s_{xy} to each pair of unconnected nodes which indicates the likelihood of connecting the two nodes. Then, all unconnected nodes are arrayed in descending order of score. The node pair in the top represents that the two nodes are the most likely to generate a connection. In order to test the precision of the algorithm, the known edge set E is divided into training set E^T and testing set E^P . Only the testing set can be used to calculate scores. Obviously, $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. We define an edge belonging to U but not to E as an inexistent edge. In this paper, we use precision [18] to measure the accuracy of link prediction algorithm, which describes the proportion of real links in the top- L links with highest scores. If there are m real links in top- L links, the precision of the algorithm can be expressed as

$$\text{precision}(L) = \frac{l}{L}. \quad (1)$$

In order to simplify the model, we use undirected and unweighted networks.

2.2. Superposed Random Walk (SRW) Model. The SRW model inspired from the LRW model considers random walk between endpoint x and y , making the nodes nearby more likely to connect to the target node [19]. It is defined as

$$s_{xy}^{\text{SRW}}(t) = \sum_{l=2}^t \left[\frac{k_x}{2|E|} \times \pi_{xy}(l) + \frac{k_y}{2|E|} \times \pi_{yx}(l) \right], \quad (2)$$

where the initial density vector $\vec{\pi}_{xy}(0) = \vec{e}_x$ and it evolves as $\vec{\pi}_{xy}(t+1) = P^T \times \vec{\pi}_{xy}(t)$. P represents the probability transition matrix with $p_{xy} = (a_{xy}/k_x)$, and $a_{xy} = 1$ when the link exists; if not, $a_{xy} = 0$. Besides, t denotes the time steps.

2.3. Derivation of Mapping Entropy (DME) Model. Inspired by Shannon entropy, the information entropy [20] of the network can be expressed as

$$E = - \sum_{i=1}^N \text{DC}_i \log \text{DC}_i, \quad (3)$$

where DC_i is the degree centrality of node i . A node and its neighbors construct a subnetwork. The local entropy (LE) [21] of the subnetwork originated from endpoint v_i is shown in the following formula:

$$\text{LE}_i = - \sum_{j=1}^M \text{DC}_j \log \text{DC}_j, \quad (4)$$

where DC_j is the degree centrality of node v_j , which belongs to the neighbor set M of node v_i . Taking the mapping relation between a node and its neighbors into account, we can obtain the mapping entropy (ME):

$$\text{ME}_i = -\text{DC}_i \sum_{j=1}^M \log \text{DC}_j, \quad (5)$$

where DC_i is the degree centrality of node v_i and DC_j is the degree centrality of one of the neighbors of node v_i .

Inspired by ME index, we introduce the derivation of mapping entropy: DME, which is defined by interleaving the degrees of node v_i and v_j .

$$\text{DME}_i = k_i \sum_{j=1}^M \lg k_j. \quad (6)$$

The definition considers both the degrees of the node and the degrees of its neighbors which takes the influence of indirect neighbors into account. This may be useful for distinguishing the importance of neighbors. Based on the SRW model, we consider using the DME index to replace the influence of the endpoint, which can perform better than the ME model introduced later through experiments based on the superposed random walk. The model is defined as

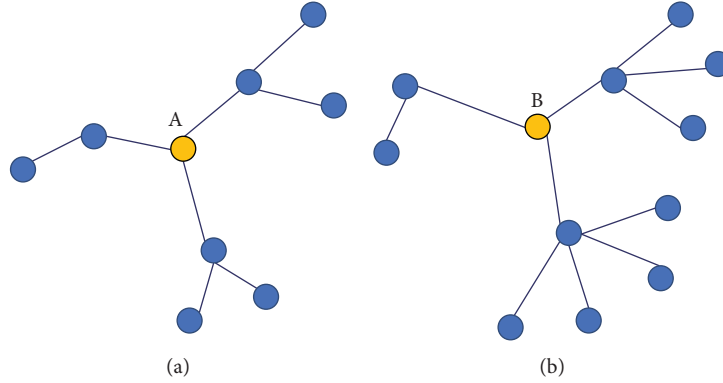


FIGURE 1: Sketch maps of influence based on the derivation of mapping entropy. As is shown, the degree of node A in subgraph (a) is equal to the degree of node B in subgraph (b). In traditional models, they are considered to have the same influence. Nevertheless, in the DME model, the differences in the contribution of their neighbors are taken into account. We think that node B has more influence. By calculation, we quantify the influence of A and B as 3.77, 4, 81.

$$s_{xy}^{\text{DME}}(t) = \sum_{l=2}^t \left[\frac{k_x \sum_{j=1}^M \lg k_j}{2|E|} \times \pi_{xy}(l) + \frac{k_y \sum_{i=1}^N \lg k_i}{2|E|} \times \pi_{yx}(l) \right]. \quad (7)$$

As mentioned above, for better comparison, we also apply the ME index into the SRW model and the ME model as shown below.

$$s_{xy}^{\text{ME}}(t) = \sum_{l=2}^t \left[\frac{-\text{DC}_x \sum_{j=1}^M \log \text{DC}_j}{2|E|} \times \pi_{xy}(l) + \frac{-\text{DC}_y \sum_{i=1}^N \log \text{DC}_i}{2|E|} \times \pi_{yx}(l) \right]. \quad (8)$$

3. Experimental Data

In order to confirm the validity of the DME model, we conduct abundant experiments on nine real networks. They are listed as follows: (1) US Air (USAir), describing the network of the US air transportation system [22]; (2) Yeast PPI (Yeast), expressing the protein-protein relationship network of yeast [23]; (3) Food Web (Food), describing the association of carbon exchanges in the cypress wetlands of Florida ecosystem [24]; (4) Power Grid (Power), expressing the electrical power transportation network in the west of the US [25]; (5) UC Irvine, representing social network (Ucsocial), describing an online social network composed of students of the University of California, Irvine [26]; (6) Jazz, indicating the collaborative relationships among jazz musicians [27]; (7) EuroSiS Web (EuroSiS), showing the interactions between Science in Society actors from 12 European countries [28]; (8) Router, referring to the transmission of data packets between the routers in Internet [4]; and (9) King James, coming from the datasets Lü et al. collected. The fundamental topological features of the mentioned networks are listed in Table 1.

Our model is applied to the undirected and unweighted connected networks. Accordingly, we make arcs turn into undirected links. Besides, we delete the loops and multiple connections. Subsequently, the maximal edge-connected graph is extracted from each raw dataset to guarantee the connectivity of the whole.

Before the experiment, the edge set E of the nine networks is divided into two parts E^T and E^P randomly. The training set E^T contains 90% of the whole edge set. The testing set E^P contains 10%. The connectivity of the E^T is guaranteed by the means of adding edges randomly to the minimum spanning tree until the training set contains 90% links. Next, 30 groups of separate experimental data for each network are divided in the same size. Then, they are applied for the averaged precision by statistical methods to avoid the randomness of results.

4. Reference Standard

In order to highlight the superiority of our algorithm, five classic methods are listed as follows.

- (1) In common neighbor (CN) [12], the similarity is judged by the number of neighbors shared by node x and node y , which is defined as

$$s_{xy}^{\text{CN}} = |\Gamma(x) \cap \Gamma(y)|, \quad (9)$$

where $\Gamma(x)$ represents the set of neighbors of node x . Besides, $|\Gamma(x) \cap \Gamma(y)|$ refers to the amount of common neighbors of nodes x and y .

- (2) Preferential attachment (PA) [29] considers that the probability of a new link linked to the node x is proportional to k_x , so the probability between node x

TABLE 1: The basic statistical features of the 9 benchmark networks.

Nets	$ V $	$ E $	$\langle k \rangle$	$\langle d \rangle$	C	r	H
USAir	332	2126	12.81	2.74	0.749	-0.208	3.36
Yeast	2364	10898	9.2	5.16	0.378	0.470	3.35
Food	128	2075	32.42	1.78	0.335	-0.112	1.24
Power	4941	6594	2.669	15.87	0.107	0.003	1.45
Ucsocial	1893	13835	14.62	3.06	0.138	-0.188	3.81
Jazz	198	2742	27.7	6.45	0.618	0.02	1.4
EuroSiS	1272	6454	10.15	3.86	0.382	-0.012	2.46
Router	5021	6257	2.49	6.45	0.033	-0.138	5.50
King James	1707	9059	10.61	3.38	0.710	-0.052	3.92

$|V|$ denotes the number of nodes in the network, $|E|$ represents the total number of links, $\langle k \rangle$ is the average degree of all nodes, $\langle d \rangle$ indicates the average distance among nodes, C represents the clustering coefficient, r is the associativity coefficient, and H denotes the degree heterogeneity defined as $H = (\langle k^2 \rangle / \langle k \rangle^2)$.

and y is proportional to $k_x \times k_y$. The index is defined as

$$s_{xy}^{\text{PA}} = k_x \times k_y. \quad (10)$$

This index does not require the information of the neighborhood of each endpoint. Therefore, it has low computational complexity.

- (3) In Adamic-Adar (AA) [13], the idea is that the contribution of the node with small degree is greater. So, each node is given a weight value equaling to $1/(\log k_z)$ where k_z is the degree of a node from common neighbor set. The similarity is defined as

$$s_{xy}^{\text{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k_z)}, \quad (11)$$

where k_z represents the degree of common neighbor z .

- (4) Resource allocation (RA) [30], derived from AA, considers the resource allocation of network. Each node is given a weight value which is equal to $1/(k_z)$, and the similarity is defined as

$$s_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \quad (12)$$

- (5) Superposed random walk (SRW) has been discussed in Section 2 in detail.

5. Results and Analysis

In order to prove the effectiveness of the DME model, abundant experiments have been carried out in nine real networks. The results are shown as follows.

In Figure 2, we plot the variation of the average precision with random walk steps obtained by SRW, ME, and DME in nine networks in the case of $L=100$. We can see that DME performs better obviously in 8 of the 9 networks than SRW. Furthermore, compared to both SRW and ME models, the DME model achieves the maximum precision in 6 of the 9 networks. Because the ME index reflects the

robustness of local network, it is more suitable for applying in network attacks to represent the importance of nodes. Therefore, we arrive at a conclusion that DME can achieve the highest accuracy in most cases when the random walk step t is optimal. Besides, it can reach the maximum precision in the minimum number of steps so that it can reduce the computation with the same precision.

Table 2 contains the detailed description of Figure 2. Furthermore, it also compares our model with other five classical models. The maximum precision is emphasized in bold and the corresponding step is in the parentheses. As is shown, the DME model reaches the highest precision in 6 of the 9 networks under the condition of $L=100$ compared with other five traditional models.

For ensuring the integrity of the experiment, we also conduct experiments in the case of $L=50$. The results are shown in Table 3. We italicize the values when the DME model is more exact than SRW. There are still 6 networks. Nevertheless, the advantage is not obvious when compared with other five models comprehensively. This means the DME model performs better in the top 100 links than 50 links. Actually, L is often defined as a large number to avoid random error.

The reason why the DME model can have an excellent performance is that it takes the mapping relationship between a node and its neighbors into comprehensive consideration. In this way, the differences in contribution of neighbors (i.e., the influences of indirect neighbors) are included, so that the model can assess the importance of endpoint better.

Though the DME model can achieve preferable performance in most datasets experimented by us, it also has no superiority in a few networks such as Jazz. By analyzing the topological characteristics of these networks, we find that they usually have the same features. The model we propose may be not suitable for the networks with good associativity coefficient and high clustering coefficient. We infer the reason is that the differences in contribution of neighbor nodes in such networks cannot be well reflected.

Furthermore, time complexity is also a significant factor to evaluate an index. For instance, CN index has $O(N^3)$

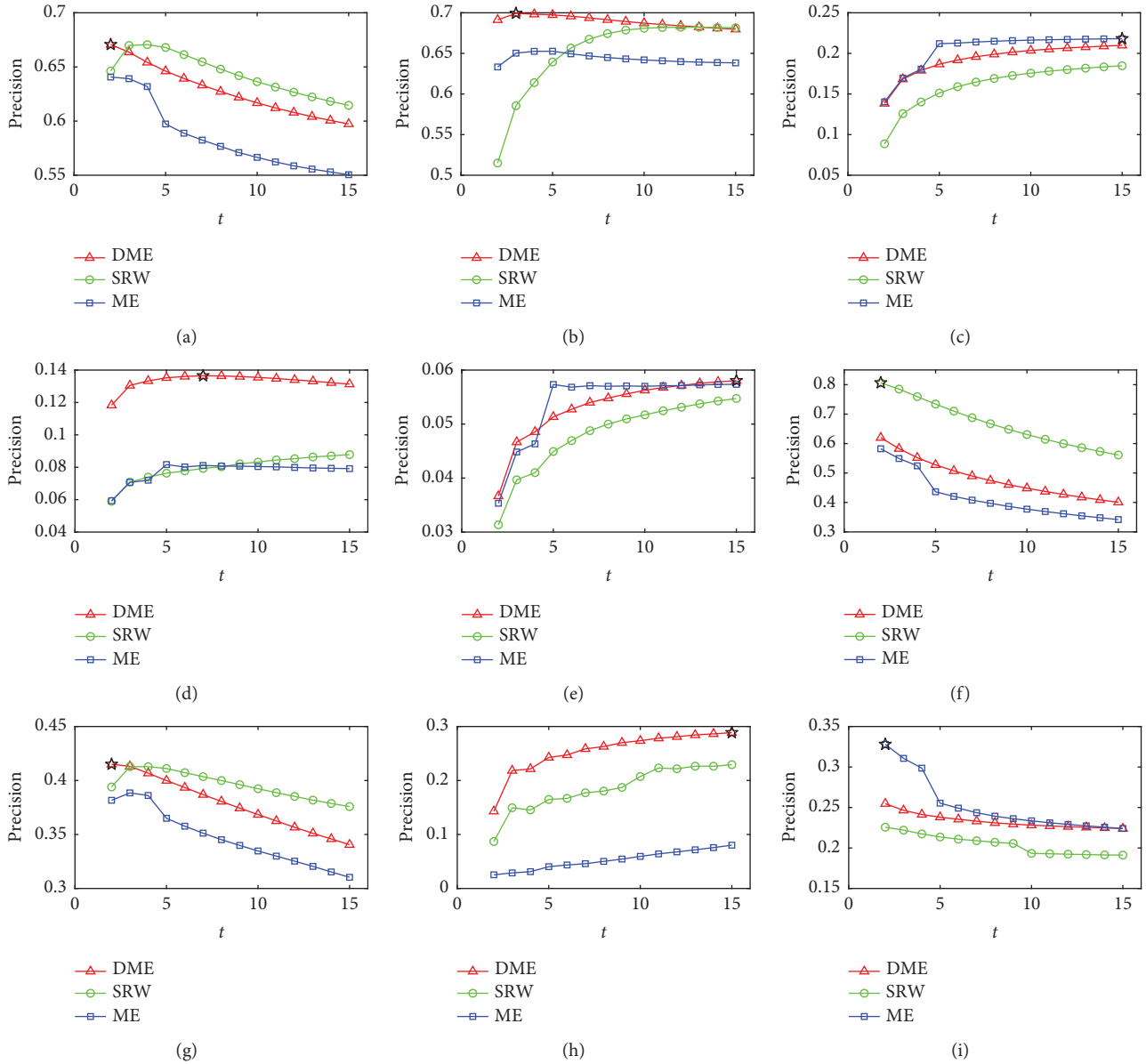


FIGURE 2: Precision of SRW (green circles), ME (blue rectangles), and DME (red triangles) versus number of random walk steps t with $L = 100$ in 9 real datasets. The highest precision in each network is marked by a black five-angled star. As is shown, DME performs better obviously in 8 of the 9 networks than SRW. Furthermore, compared to SRW and ME models, the DME model achieves the maximum precision in 6 of the 9 networks. (a) USAir. (b) Yeast. (c) Food. (d) Power. (e) Ucsocial. (f) Jazz. (g) EuroSiS. (h) Router. (i) King James.

TABLE 2: Precision on the nine benchmark networks in the case of $L = 100$.

Precision	CN	PA	AA	RA	SRW	DME
USAir	0.635	0.479667	0.654333	0.64	0.670583 (4)	0.670667 (2)
Yeast	0.728	0.555	0.722667	0.515	0.682242 (12)	0.699333 (3)
Food	0.086	0.211333	0.089667	0.088667	0.184810 (15)	0.209952 (15)
Power	0.124333	0.017333	0.081333	0.057333	0.087810 (15)	0.136500 (7)
Ucsocial	0.039333	0.057333	0.037333	0.031333	0.054738 (15)	0.058048 (15)
Jazz	0.816	0.168667	0.835667	0.806333	0.806333 (2)	0.62 (2)
EuroSiS	0.456333	0.049333	0.46667	0.394	0.412833 (3)	0.415 (2)
Router	0.132667	0.016	0.102333	0.075	0.229278 (15)	0.288690 (15)
King James	0.404333	0.190667	0.554333	0.82867	0.225667 (2)	0.254667 (2)

Each data point is an average over 30 independently divided datasets. The number in bold means the highest precision. The value enclosed in the parentheses represents the step corresponding to optimal precision.

TABLE 3: Precision on the nine benchmark networks with $L = 50$.

Precision	CN	PA	AA	RA	SRW	DME
USAir	0.821333	0.646667	0.82533	0.753333	0.8016 (6)	0.806889 (4)
Yeast	0.779333	0.594667	0.763333	0.470667	0.813429 (15)	0.817333 (15)
Food	0.118667	0.274	0.120667	0.116667	0.268 (15)	0.275762 (15)
Power	0.212	0.008667	0.059333	0.045333	0.119333 (15)	0.187778 (7)
Ucsocial	0.050667	0.077667	0.043333	0.041333	0.070667 (15)	0.064667 (15)
Jazz	0.895333	0.222667	0.906667	0.876667	0.0.876667 (2)	0.716667 (2)
EuroSiS	0.593333	0.074	0.560667	0.437333	0.466667 (5)	0.507333 (2)
Router	0.17	0.014667	0.121333	0.065333	0.134 (15)	0.196810 (15)
King James	0.552667	0.269333	0.69	0.898667	0.898667 (2)	0.502 (2)

Each data point is an average over 30 independent datasets, each of which is randomly divided into training set and test set with 90% and 10% probability. The value enclosed in the parentheses represents the step corresponding to optimal precision. Besides, the italicized values illustrate that the DME model performs better than the SRW model.

computational complexity while the complexity of RA is $2 \times O(N^3)$. The complexity of SRW which considers local path is $M \times O(N^3)$, and M is far less than N^3 . The model we introduce has the same time complexity of $M \times O(N^3)$ as SRW but can realize higher precision.

6. Conclusions

Existing link prediction algorithms on the basis of structural similarity mostly focus on the paths or the influences of nodes with only their degrees considered. Because the differences in contribution of neighbors are not considered, the precision of algorithm is limited. Through analysis, we propose the derivation of mapping entropy (DME) model, which interleaves the degrees of node and its neighbors. We investigate our model in comparison of CN, PA, AA, RA, SRW, and ME models on nine real datasets. The results indicate that the DME model prominently performs better than other six models and can achieve maximum precision in the minimum number of steps which reduces the computation with the same precision. Furthermore, the DME model does not increase time complexity.

The DME model proposed in our study reveals the effectiveness of distinguishing the differences in neighbors' contribution. This finding can provide a reference for future research. However, we only take the influence of indirect neighbors into account and ignore other factors such as coreness and H-index which can describe the maximal connected subgraph. Besides, we do not know the performance of the DME model in weighted and directed networks.

The results of our research are meaningful, and they are of great significance to the practical application of academic research. We can apply it in recommendation system, social cooperation network, information and communication technology, potential interactions in biological networks, and so on. Significantly, this work can inspire further work to add other factors such as H-index on the basis of our model and optimize the DME model in weighted and directed networks.

Data Availability

The datasets used in this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (nos. 61821007 and 62090015) and Science and Technology Key Project of Henan Province (Research on Key Technologies of Link Prediction in Complex Networks) (no. 202102311007).

References

- [1] W. Wang, M. Tang, H. F. Zhang, and Y. C. Lai, "Dynamics of social contagions with memory of nonredundant information," *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 92, Article ID 012820, 2015.
- [2] L. Lu and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2010.
- [3] W. Wang, Y. Ma, T. Wu, Y. Dai, X. Chen, and L. A. Braunstein, "Containing misinformation spreading in temporal social networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, Article ID 123131, 2019.
- [4] R. Pech, D. Hao, L. Pan, H. Cheng, and T. Zhou, "Link prediction via matrix completion," *EPL*, vol. 117, Article ID 38002, 2016.
- [5] X. Zhu, W. Li, H. Tian, and S. Cai, "Hybrid influence of degree and h-index in the link prediction of complex networks," *EPL (Europhysics Letters)*, vol. 122, 2018.
- [6] Y. Li, P. Luo, Z. P. Fan, K. Chen, and J. Liu, "A utility-based link prediction method in social networks," *European Journal of Operational Research*, vol. 260, 2017.
- [7] Q. Sun, R. Hu, Y. Zhao, Y. Yao, and Y. Fan, "An improved link prediction algorithm based on degrees and similarities of nodes," in *Proceedings of IEEE/ACIS International Conference on Computer Information Science*, Wuhan, China, May 2017.
- [8] A. De, N. Ganguly, and S. Chakrabarti, "Discriminative link prediction using local links, node features and community structure," 2013, arXiv.
- [9] J.-X. Yang and X.-D. Zhang, "Revealing how network structure affects accuracy of link prediction," *The European Physical Journal B*, vol. 90, no. 8, p. 157, 2017.
- [10] X. Chen, K. Gong, R. Wang, S. Cai, and W. Wang, "Effects of heterogeneous self-protection awareness on resource-

- epidemic coevolution dynamics,” *Applied Mathematics and Computation*, vol. 385, Article ID 125428, 2020.
- [11] H. Liu, Z. Hu, H. Haddadi, and H. Tian, “Hidden link prediction based on node centrality and weak ties,” *EPL (Europhysics Letters)*, vol. 101, no. 1, Article ID 18004, 2013.
- [12] F. Lorrain and H. C. White, “Structural equivalence of individuals in social networks,” *Social Networks*, vol. 1, pp. 67–98, 1977.
- [13] Y. Yao, R. Zhang, F. Yang et al., “Link prediction via layer relevance of multiplex networks,” *International Journal of Modern Physics C*, vol. 28, no. 8, Article ID 1750101, 2017.
- [14] Q. Zhang, M. Li, and Y. Deng, “Measure the structure similarity of nodes in complex networks based on relative entropy,” *Physica A Statistical Mechanics Its Applications*, vol. 491, 2017.
- [15] W. Liu and L. Lü, “Link prediction based on local random walk,” *EPL (Europhysics Letters)*, vol. 89, Article ID 58007, 2010.
- [16] X. Chen, Q. Liu, R. Wang, Q. Li, and W. Wang, “Self-awareness-based resource allocation strategy for containment of epidemic spreading,” *Complexity*, vol. 2020, Article ID 3256415, 12 pages, 2020.
- [17] X. Chen, R. Wang, D. Yang, J. Xian, and Q. Li, “Effects of the awareness-driven individual resource allocation on the epidemic dynamics,” *Complexity*, vol. 2020, Article ID 8861493, 12 pages, 2020.
- [18] L. Lin-Yuan, “Link prediction on complex networks,” *Journal of University of Electronic Science and Technology of China*, vol. 39, pp. 651–661, 2010.
- [19] Y. Liu, T. Li, and X. Xu, “Link prediction by multiple motifs in directed networks,” *IEEE Access*, vol. 8, pp. 174–183, 2020.
- [20] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang, “Deep belief network-based approaches for link prediction in signed social networks,” *Entropy*, vol. 17, no. 4, pp. 2140–2169, 2015.
- [21] T. Nie, Z. Guo, K. Zhao, and M. Zhe, “Using mapping entropy to identify node centrality in complex networks,” *Physica, A. Statistical Mechanics and Its Applications*, vol. 453, 2016.
- [22] Z. Shan, “Link prediction based on local information considering preferential attachment,” *Physica A: Statistical Mechanics and Its Applications*, vol. 443, pp. 537–542, 2016.
- [23] R. Pech, D. Hao, Y. L. Lee, Y. Yuan, and T. Zhou, “Link prediction via linear optimization,” *Physica A: Statistical Mechanics and Its Applications*, vol. 528, 2018.
- [24] Z. Wu, Y. Lin, J. Wang, and S. Gregory, “Link prediction with node clustering coefficient,” *Physica A: Statistical Mechanics and its Applications*, vol. 452, no. C, pp. 1–8, 2015.
- [25] L. Jiao, F. Liu, J. Wu, and L. Ding, “Prediction of missing links based on community relevance and ruler inference,” *Knowledge Based Systems*, vol. 98, pp. 200–215, 2016.
- [26] W. Chao, V. Satuluri, and S. Parthasarathy, “Local probabilistic models for link prediction,” in *Proceedings of IEEE International Conference on Data Mining*, Omaha, NE, USA, October 2007.
- [27] L. Lü, C. H. Jin, and T. Zhou, “Similarity index based on local paths for link prediction of complex networks,” *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, Article ID 046122, 2009.
- [28] B. Ermiş, E. Acar, and A. T. Cemgil, “Link prediction in heterogeneous data via generalized coupled tensor factorization,” *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 203–236, 2015.
- [29] V. Martinez, F. Berzal, and J. C. Cubero, “A survey of link prediction in complex networks,” *Acm Computing Surveys*, vol. 49, pp. 69.1–69.33, 2017.
- [30] L. Ji and G. Deng, “Link prediction in a user–object network based on time-weighted resource allocation,” *Physica A Statistical Mechanics Its Applications*, vol. 388, pp. 3643–3650, 2009.

Research Article

Dual-Channel Reasoning Model for Complex Question Answering

Xing Cao,^{1,2} Yun Liu ,^{1,2} Bo Hu,^{1,2} and Yu Zhang^{1,2}

¹School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

²Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing 100044, China

Correspondence should be addressed to Yun Liu; liuyun@bjtu.edu.cn

Received 10 May 2021; Revised 23 June 2021; Accepted 8 July 2021; Published 26 July 2021

Academic Editor: Xuzhen Zhu

Copyright © 2021 Xing Cao 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.

Multihop question answering has attracted extensive studies in recent years because of the emergence of human annotated datasets and associated leaderboards. Recent studies have revealed that question answering systems learn to exploit annotation artifacts and other biases in current datasets. Therefore, a model with strong interpretability should not only predict the final answer, but more importantly find the supporting facts' sentences necessary to answer complex questions, also known as evidence sentences. Most existing methods predict the final answer and evidence sentences in sequence or simultaneously, which inhibits the ability of models to predict the path of reasoning. In this paper, we propose a dual-channel reasoning architecture, where two reasoning channels predict the final answer and supporting facts' sentences, respectively, while sharing the contextual embedding layer. The two reasoning channels can simply use the same reasoning structure without additional network designs. Through experimental analysis based on public question answering datasets, we demonstrate the effectiveness of our proposed method

1. Introduction

One of the long-standing goals of natural language processing (NLP) is to enable machines to have the ability to understand natural language and make inferences in textual data. Many applications, such as dialogue systems [1, 2], recommendation systems [3–5], question answering [6, 7], and sentiment analysis [8], aim to explore the machine ability to understand textual data. Question answering, abbreviated as QA, has emerged as an important natural language processing task because it provides a quantifiable way to evaluate an NLP system's capability on language understanding and reasoning and its commercial value for real-world applications.

Most previous works have focused on questions answering only from a single paragraph, known as single-hop QA [9]. Although recent advances in QA and machine reading comprehension (MRC) had surpassed human performance on some single-hop datasets [10, 11], those datasets have gaps from real-world scenarios. In the real

world, there are a lot of complex questions that need to be answered through multiple steps of reasoning by aggregating information distributed in multiple paragraphs, named multihop QA [12].

Jiang and Bansal [13] pointed out that because examples include reasoning shortcuts, some models may directly locate the answer by word-matching the question with sentences in the context. For the complex question “what was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?” The context contains the sentence “Peter Bolesław Schmeichel..... and was voted the IFFHS World's Best Goalkeeper in 1992 and 1993.” At this time, the model may find the correct answer “World's Best Goalkeeper” through simple word matching, but does not infer that Peter Bolesław Schmeichel is the father of Kasper Schmeichel. Therefore, to enhance the interpretability of models and avoid answering complex questions through reasoning shortcuts, our study considers that, in addition to predicting the correct answer, it is also important to extract evidence sentences. However, most of the existing works only focused on improving the

accuracy of the model to answer complex questions, but pay less attention to the ability of the model on predicting the inference path.

An example from HotpotQA is illustrated in Figure 1. Ten paragraphs are given to answer complex questions (“*what government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?*”); the model first needs to identify passage 2 (P2) and passage 6 (P6) above as relevant paragraphs to correctly answer the question.

The first sentence of P6 and P2 are evidence sentences, which lead to the next-hop paragraph and the predicted answer, respectively. However, it is more difficult for the model to predict correct and complete evidence sentences than to answer a complex question because the question often does not contain information about the intermediate answer, such as “Shirley Temple” (green font) in Figure 1.

Most existing methods [14–16] predicted the final answer and the supporting facts in sequence or simultaneously, and the architecture of these methods is designed primarily to predict the right answer. In this paper, we propose a novel dual-channel reasoning architecture for complex question answering. Concretely, complex questions and documents pass through the word embedding layer and contextual embedding layer in succession. Thereafter, the output of the contextual embedding layer is the input of two reasoning channels: one for predicting the answer span or answer type, and the other for predicting evidence sentences.

Our contributions can be summarized as follows:

- (1) We propose the dual-channel reasoning architecture, which is a novel architecture for the complex question answering task. The results of the experiment show that the dual-channel reasoning architecture is suitable for many kinds of existing neural network models, such as graph-based models.
- (2) We perform comprehensive experiments on multihop QA datasets, and our proposed method outperforms previous approaches on complex questions, especially on the task of extracting evidence sentences. We conducted a detailed visual analysis of the baseline model and two channels in the dual-channel architecture and further explored the differences in the distribution of attention heat maps of several models.

2. Related Work

2.1. Multihop Question Answering over Knowledge Base. Knowledge-based question answering (KBQA) computes answers to natural language questions based on a knowledge base. Besides the traditional methods of defining templates and rules, KBQA methods can be mainly divided into two branches: semantic parsing (SP) based and information retrieval (IR) based. Semantic parsing methods focus on translating complex natural language questions into the executable query graph over the knowledge base. Lan and Jiang [17] proposed a modified staged query graph generation method by allowing longer relation paths. Sun et al. [18] proposed a novel skeleton grammar that uses the BERT-based

parsing algorithm to improve the downstream fine-semantic parsing. To avoid generating noisy candidate queries, Chen et al. [19] proposed abstract query graphs (AQG) to describe query structures. The IR-based model first extracts the subject entities mentioned in the question and links them to the knowledge base [20]. Then, the subgraph centered on the subject entity is extracted, and all nodes in the subgraph are selected as candidate answers. Chen et al. [21] used a novel bidirectional attentional memory network to simulate the bidirectional interactive flow between a question and a knowledge base. Xu et al. [22] enhanced KV-MemNNs models by a new query updating strategy to perform interpretable reasoning for complex questions.

2.2. Multihop Question Answering over Text. Currently, there are two mainstream branches for complex question answering over textual data. The first direction is to apply the previous neural networks that are successful in single-hop QA tasks to multihop QA tasks. The Bidirectional Attention Flow (Bi-DAF) network proposed by Seo et al. [23] has achieved state-of-the-art results in single-hop QA datasets. Yang et al. [12] proposed the multihop dataset HotpotQA and used the model with the Bi-DAF module as the core which was used as the baseline model of this dataset. Zhong et al. [24] proposed a model combination of coarse-grained reading and fine-grained reading. The query-focused extractor model proposed by Nishida et al. [16] regards evidence extraction as a query-focused summarization task and reformulates the query in each hop. Because the semantics of the questions in the multihop QA task is more complex, it is difficult for the Bi-DAF module to fully understand the semantics. Min et al. [25] addressed HotpotQA by decomposing its multihop questions into single-hop sub-questions to achieve better performance and interpretability. Jiang and Bansal [26] proposed a self-assembling modular model to make multihop reasoning and support fact selection more interpretable. However, their model needs to be trained by using a large amount of manually labeled data, which is undoubtedly expensive. Because answers to complex questions require aggregating information from multiple paragraphs and BERT cannot encode all documents at once, Bhargav et al. [27] proposed translucent answer prediction architecture to effectively capture the local context and the global interactions between the sentences.

The other direction is based on graph neural networks (GNNs) [28]. Graph is an effective way to represent complex relationships between entities and to obtain relationship information. Ding et al. [29] used the implicit extraction module and explicit reasoning module to build the reasoning process into a cognitive graph. Inspired by human’s step-by-step reasoning behavior, Qiu et al. [15] proposed a dynamically fused graph network that can predict the subgraphs dynamically at each reasoning step. The multi-level graph network can represent the information in the context in more detail. The hierarchical graph network (HGN) proposed by Fang et al. [14] captures clues from different granularity levels and weaves heterogeneous nodes into a single unified graph.

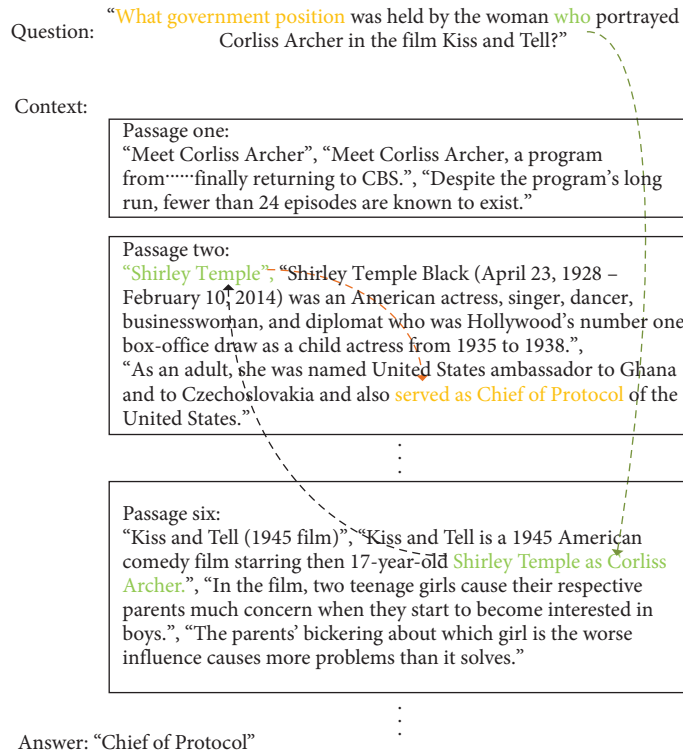


FIGURE 1: Example of multihop datasets HotpotQA.

3. Task Formulation

Suppose we are given a set of training data $\{C_i, Q_i, A_i, \text{Sup}_i\}$, where each context C_i is composed of many documents $\{P_1, P_2, \dots, P_n\}$ related to the question and is regarded as one connected text $C_i = \{x_1, x_2, \dots, x_T\}$, and $Q_i = \{q_1, q_2, \dots, q_J\}$ is regarded as a complex query; context C_i and query Q_i have T words and J words, respectively.

The goal of the task is to design models to predict A_i and Sup_i . A_i includes answer type A_T and answer string A_S ; the answer type A_T is selected from the answer candidates, such as "yes/no/span." The answer string A_S is a short span in context, which is determined by predicting the positions of the start indexes and the end indexes when there are not enough answer candidates to answer Q , expressed by $\langle \text{start}_i, \text{end}_i \rangle$. Sup_i is regarded as evidence sentences, and supporting facts include more than one sentence in C_i , expressed by $\langle \text{paragraph title, sentence indexes} \rangle$.

4. Solution Approach

4.1. Process Overview. We describe the dual-channel reasoning architecture in this section. Our proposed model consists of four components that are the input module, contextual module, reasoning module, and prediction module. To test the applicability of the proposed architecture, the input module, contextual module, and reasoning module, respectively, adopt different current mainstream neural networks. The overall dual-channel reasoning architecture is illustrated in Figure 2.

4.2. Input Module. An input question $Q_i = \{q_1, q_2, \dots, q_J\}$ and context $C_i = \{x_1, x_2, \dots, x_T\}$ are represented as sequences of word embeddings and character embeddings, respectively. The concatenation of the character and word embedding vectors is passed to the highway network, and the outputs of the highway network are two matrices $X_1 \in \mathbb{R}^{T \times d_1}$ for the context and $Q_1 \in \mathbb{R}^{J \times d_1}$ for the query, where d_1 is the dimension after fusion of the word embedding and character embedding. In addition, the input module can also use a pretrained model, BERT. The query Q_i and the context C_i are concatenated, and they pass the resulting sequence to a pretrained BERT model to obtain representations $X_2 \in \mathbb{R}^{T \times d_2}$ for the context and $Q_2 \in \mathbb{R}^{J \times d_2}$ for the query, where d_2 is the size of BERT hidden states.

4.3. Contextual Module. To model the temporal interactions between words in context and question, bidirectional long short-term memory (Bi-LSTM) networks are applied above the input module. The output representation of Bi-LSTM are $U \in \mathbb{R}^{J \times 2d_1}$ and $H \in \mathbb{R}^{T \times 2d_1}$, where $2d_1$ denotes the output dimension. For the graph neural network method, identifying supporting entities and the text span of potential answers from the output of BERT are used as nodes in the graph. Undirected edges are defined according to the positional properties of every node pair.

4.4. Reasoning Module. The reasoning module includes the context-query interaction layer and modeling layer. The typical implementation of the context-query interaction

layer is Bi-DAF. Bi-DAF is responsible for connecting and integrating the information of context and query words. Finally, the contextual module output and the vectors computed by the context-query interaction layer are combined to yield G :

$$\begin{aligned}
 G_{t,:} &= \tilde{\beta}(H_{t,:}, \tilde{U}_{t,:}, \tilde{H}_{t,:}), \\
 \tilde{\beta}(h, \tilde{u}, \tilde{h}) &= [h; \tilde{u}; h \circ \tilde{u}; \tilde{h} \circ \tilde{u}], \\
 \tilde{U}_{t,:} &= \sum_j a_{t,j} U_{j,:}, \\
 a_{t,:} &= \text{softmax}(S_{t,:}), \quad a_{t,:} \in \mathbb{R}^T, \\
 S_{tj} &= [h + u + \alpha(H, U)], \quad S_{tj} \in \mathbb{R}^{T \times J}, \\
 \alpha(H, U) &= U^T H, \alpha(H, U) \in \mathbb{R}^{T \times J}, \\
 h &= \text{linear}(H), \quad h \in \mathbb{R}^{T \times 1}, \\
 u &= \text{permute}(\text{linear}(U)), \quad u \in \mathbb{R}^{1 \times J}, \\
 \tilde{h} &= \sum_t b_t H_{t,:}, \quad \tilde{h} \in \mathbb{R}^{2^d}, \\
 b &= \text{softmax}(\max_{\text{col}}(S)), \quad b \in \mathbb{R}^J.
 \end{aligned} \tag{1}$$

where \tilde{h} is tiled T times across the column, thus giving $\tilde{H} \in \mathbb{R}^{T \times 2^d}$, $[\cdot; \cdot]$ is vector concatenation across row, S is the similarity matrix, and \tilde{U} and \tilde{H} represent the output of context-to-query attention and query-to-context attention, respectively. The output G of the context-query interaction layer is taken as the input to the modeling layer, which encodes the query-aware representations of context words. We use one layer of the bidirectional GRU to capture the interaction among the context words conditioned on the query. Since multiple documents contain thousands of words, the long-distance dependency problem is obvious, so a self-attention module is added to alleviate this problem.

For the graph neural network method, graph attention networks, graph recurrent networks, and graph convolutional networks, their variants can propagate messages across different entity nodes in graphs and update the vector representation of the original entity.

4.5. Prediction Module. The prediction module consists of four homogeneous Bi-GRU and linear layers. Corresponding to the channels used to predict the answer are three sets of Bi-GRU and linear layers, and they have three output dimensions, including (1) the start indexes of the answer, (2) the end indexes of the answer, and (3) the answer type. The prediction module corresponding to the evidence sentences extraction channel only outputs the supporting sentences predicted by the model.

5. Experiments

5.1. Datasets. HotpotQA is a recently introduced multihop QA dataset with 113k Wikipedia-based question-answer pairs. HotpotQA has two benchmark settings, namely, distractor setting and full wiki setting. In the distractor setting, for each example, there are two golden paragraphs

related to complex questions and eight unrelated ones. The two gold paragraphs and the eight distractors are shuffled before they are fed to the model. Full wiki setting requires the model to answer the question given the first paragraph of all Wikipedia articles, in which no specified golden paragraphs are given. Here, we focus on the HotpotQA dataset under the distractor setting to challenge the model to find the true supporting facts in the presence of noise. For the full wiki setting where all Wikipedia articles are given as input, we consider the bottleneck to be about information retrieval, thus we do not include the full wiki setting in our experiments. In HotpotQA, only the training and validation data are publicly available, while the test data are hidden. For further analysis, we report only the performance on the validation set, as we do not want to probe the unseen test set by frequent submissions. According to the observations from our experiments and previous works, the validation score is well correlated with the test score.

5.2. Model Comparison. We compared the results with those of the three categories' model. The first category is the model which follows the feature interaction framework, such as models with Bi-DAF as the core component, specifically, NMN, etc. The second category is the reasoning model based on a graph neural network, such as KGNN and DFGN. The third category is the pretrained model, such as BERT.

5.2.1. Baseline. The baseline model was proposed in the original HotpotQA paper. The network architecture is composed of context and question embedding layer, contextual embedding layer, modeling layer, and prediction layer from the bottom to the top.

5.2.2. NMN. NMN is a self-assembling modular model for multihop QA. Four atomic neural modules were designed, namely, Find, Relocate, Compare, and NoOp, where four neural modules were dynamically assembled to make multihop reasoning and support fact selection more interpretable.

5.2.3. KGNN. KGNN is a knowledge-enhanced graph neural network (KGNN), which performs reasoning over multiple paragraphs.

5.2.4. DFGN. DFGN is a dynamically fused graph network that can predict the subgraphs dynamically and update query at each reasoning step.

5.2.5. BERT. BERT has been shown to be successful on many NLP tasks, and recent papers have also examined complex QA using the BERT model.

5.2.6. Coarse-Grained Decomposition Strategy. To solve the problem that the original Bi-DAF module cannot obtain a query-aware context representation correctly for complex

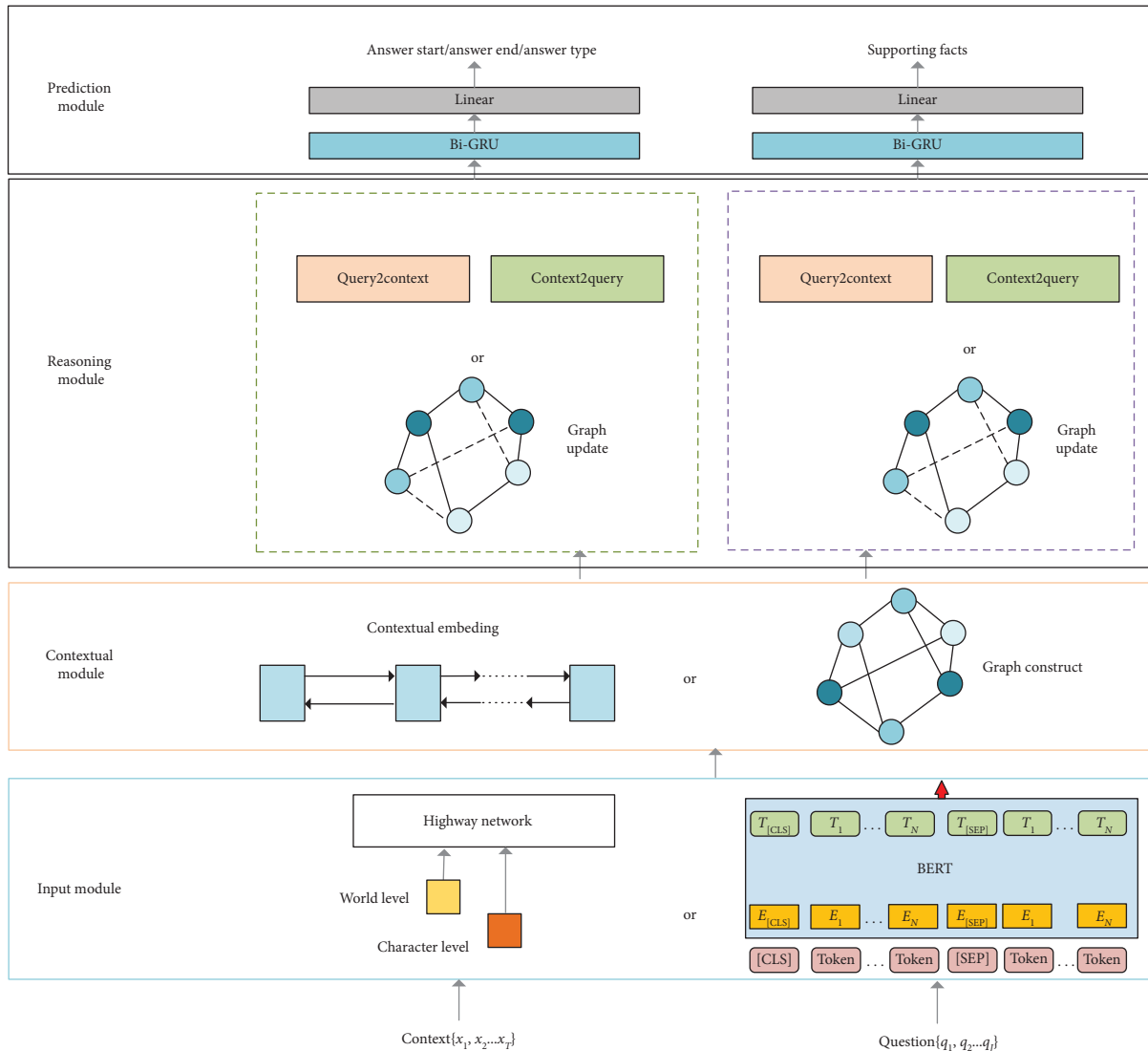


FIGURE 2: Overview of the dual-channel reasoning architecture.

questions, Cao and Liu [6] proposed the coarse-grained decomposition strategy, named CGDe strategy. The CGDe is responsible for decomposing complex questions and generating a new question which contains the semantics of intermediate answers that appear in the text to a certain extent.

5.2.7. Fine-Grained Interaction Strategy. Cao and Liu [6] proposed the fine-grained interaction strategy to solve deficiencies of vanilla Query2Context, named FGIn strategy. Instead of max pooling operation, softmax is used for each column of the attention matrix, and then, the document vector is dotted with each column weight. The method obtains J vector matrices of size $(T, 2d)$, where J is the number of words in the question. Finally, the J matrices are added to obtain the output matrix of the same size as the original Query2Context module. Comprehensive experiments showed that FGIn strategy predicts the number of evidence sentences more accurately than the baseline.

5.3. Implementation Details. To prove that our model components and model architecture have absolute performance advantages over the baseline model, we reimplemented the architecture described in the works by Yang et al. [12] and Qiu et al. [15].

5.3.1. Baseline Model for HotpotQA Dataset. We use the standard 300-dimensional pretrained GloVe as word embeddings. The dimensions of hidden states in Bi-GRU are set as $d=80$. Using the Adam optimizer, with a minibatch size of 32 and an initial learning rate of 0.01, an early stopping strategy is adopted, with patience = 1.

5.3.2. Dynamically Fused Graph Network. We also used a pretrained BERT model as the encoder, d is 768. All the hidden state dimensions are set to 300 using the Adam optimizer and an initial learning rate of 0.0001.

5.4. Main Results. The performance of multihop QA on HotpotQA is evaluated by using the exact match (EM) and F1 as two evaluation metrics for answer prediction and evidence sentences extraction. Exact match (EM) means that the answer or evidence sentences predicted by the model are exactly the same as the golden label. Joint EM is 1 only if the answer string and supporting facts are both strictly correct. The calculation formula of Joint F1 is

$$\begin{aligned} P^{(\text{joint})} &= P^{(\text{ans})} P^{(\text{sup})}, \\ R^{(\text{joint})} &= R^{(\text{ans})} R^{(\text{sup})}, \\ \text{Joint}F_1 &= \frac{2P^{(\text{joint})}R^{(\text{joint})}}{P^{(\text{joint})} + R^{(\text{joint})}}. \end{aligned} \quad (2)$$

To verify the general applicability of dual-channel reasoning in various neural network models, we apply dual-channel reasoning architecture to the feature interaction framework model and graph-based model, respectively. Correspondingly, we selected the baseline model proposed by Yang et al. [12] and the DFGN model proposed by Qiu et al. [15], which are the Baseline-Dual model and DFGN_Dual model in Table 1, respectively.

We integrate the CGDe strategy and the FGIn strategy proposed by Cao and Liu [6] into the dual-channel architecture. The CGDe strategy is conducive to finding the answer, so the channel for predicting the answer in the dual-channel architecture uses the CGDe strategy, and the other reasoning channel uses the baseline reasoning module. Similarly, FGIn is conducive to extracting evidence sentences, and the FGIn strategy is used for supporting facts prediction channels, which means that the dual-channel architecture is FGIn-Baseline.

We compare our approach with several previously published models and present our results in Table 1, where * represents the result of our reimplementing of the model. As shown in Table 1, all the results of our proposed model are superior to those of the baseline model, especially in supporting fact prediction tasks, both EM_{sup} and F1_{sup} have greatly improved. It is worth noting that although our model does not use any pretrained language model such as BERT for encoding, it outperformed the methods that used BERT such as DFGN, DFGN/BERT, and BERT Plus in the supporting facts prediction task.

5.5. Ablation Studies. In this paper, a dual-channel reasoning architecture is designed for complex question answering. To study the contributions of the dual-channel structure and these two strategies to the performance of our model, we perform an ablation experiment on the HotpotQA datasets.

As shown in Table 2, the three models in the dual-channel reasoning architecture are superior to all single-channel models on all metrics of supporting the fact prediction task (see the bottom of Table 2). Table 2 shows that when the baseline only performs answer prediction or supporting fact prediction tasks, both EM and F1 metrics are

higher than models that simultaneously perform answer prediction and supporting fact prediction. It shows that when the single-channel reasoning structure is adopted, the two tasks not only do not promote each other but also reduce the model’s ability to extract evidence. Using the dual-channel reasoning structure, the two tasks promote each other, and the score of supporting facts’ extraction tasks is higher than those of complex methods that use graph neural networks and pretrained language models. In the CGDe-Baseline architecture, there is a significant improvement in the indicators on the answer prediction task, while the performance on the supporting facts prediction task drops slightly. As Cao and Liu [6] concluded, the CGDe model’s ability to predict supporting facts is limited because the new question generated contains the intermediate answer required for the first subquestion, so the support sentence that answers the first question may not be predicted as a supporting fact. In the CGDe-Baseline architecture, the performance of the supporting facts prediction task is also affected, which further proves that there is a soft interaction between two reasoning channels in the dual-channel reasoning architecture.

5.6. Analysis and Visualization. In this section, we conduct a series of visual analyses with different settings using our approach.

For a more intuitive analysis, on the HotpotQA validation set, we evaluate the baseline model proposed by Yang et al. [12], the dual-channel reasoning model, and the model that only performs answer prediction task or supporting facts prediction task. At the same time, heat maps of the attention matrices of these models are generated. As the heat map of the model that only performs the answer prediction task (Figure 3) shows, the phrase “*who portrayed Corliss Archer in the film Kiss and Tell?*” used to describe constraints in the complex question has low correlation with all words in the document (the part within the red frame in the figure). This means that the model only answers part of the question, and complex questions are mistakenly regarded as simple questions. Similar to Figure 3, the phrase “*who portrayed Corliss Archer in the film Kiss and Tell?*” in Figures 4 and 5 is also low in correlation with the words in the document, but the correlation in Figure 4 is better than that in Figure 3, and the correlation in Figure 5 is better than that in Figure 4.

The reason for the high correlation of the corresponding positions in Figure 4 is that the baseline model also extracts evidence sentences while predicting the answer, using a single-channel reasoning structure. The correlation of corresponding words shown in Figure 5 is further superior to that shown in Figure 4, indicating that the supporting facts prediction task has a greater impact on the answer prediction task in dual-channel reasoning architecture. It is worth noting that although the EM_{ans} and F1_{ans} values of the only-ans model are slightly higher than those of the baseline model, it may be that the only-ans model mistakenly regards complex questions as simple questions and happens to find the correct answer by using the reasoning shortcut pointed out by Jiang and Bansal [13].

TABLE 1: The performance of our model and competing approaches on the HotpotQA dataset.

Model	Answer		Sup fact		Joint	
	EM	F1	EM	F1	EM	F1
Baseline	45.60	59.02	20.32	64.49	10.83	40.16
NMN	49.58	62.71	—	—	—	—
KGNN	50.81	65.75	38.74	76.69	22.40	52.82
BERT Plus	55.84	69.76	42.88	80.74	27.13	58.23
DFGN	56.31	69.69	51.50	81.62	33.62	59.82
DFGN*	55.19	68.68	49.72	80.67	31.53	58.26
DFGN/BERT	55.17	68.49	49.85	81.06	31.87	58.23
<i>Our model</i>						
Baseline-Dual	49.56	64.15	47.61	83.45	26.44	55.41
CGDe-Baseline	51.23	65.42	46.71	83.03	27.69	55.77
FGIn-Baseline	50.42	64.82	48.93	84.10	27.95	56.47
DFGN_Dual	55.42	68.90	50.70	81.70	31.76	58.83

TABLE 2: Ablation results on the HotpotQA dataset.

Model	Answer		Sup fact		Joint	
	EM	F1	EM	F1	EM	F1
<i>Baseline</i>						
—Yang et al.	45.60	59.02	20.32	64.49	10.83	40.16
—Yang et al.*	49.06	63.62	32.01	75.40	18.12	50.36
—Only ans	49.88	64.42	—	—	—	—
—Only sup	—	—	43.15	79.61	—	—
<i>Single channel</i>						
CGDe/FGIn	51.04	65.57	39.45	79.80	23.49	54.79
Only CGDe	50.81	65.20	38.78	79.35	22.62	53.86
Only FGIn	49.72	64.39	41.03	80.72	23.03	53.97
<i>Dual channel</i>						
—Baseline-Dual	49.56	64.15	47.61	83.45	26.44	55.41
—CGDe-Baseline	51.23	65.42	46.71	83.03	26.79	55.77
—FGIn-Baseline	50.42	64.82	48.93	84.10	27.95	56.47

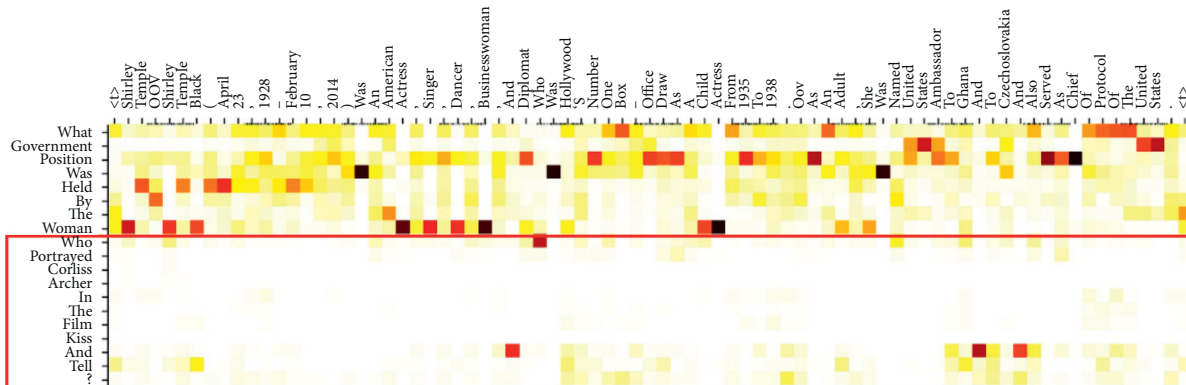


FIGURE 3: Attention heat map of the only-ans model.

The attention heat map shown in Figures 6 and 7 is no longer very sparse in the corresponding part of the phrase “who portrayed Corliss Archer in the film Kiss and Tell?,” indicating that the model has further captured the semantics of the phrase. This is very important for the model to extract evidence sentences because “who portrayed Corliss Archer in the film Kiss and Tell?” is a constraint on the complex question.

The main difference between our dual-channel reasoning model and the single-channel reasoning model is the supporting facts prediction task. Figure 8 reveals that the reason for the high EM and F1 is that the dual-channel reasoning model (baseline-baseline) rarely extracts too many supporting facts. That is, it predicts the number of evidence sentences more accurately than the baseline model. In addition, Figure 8 shows that the dual-channel reasoning model has a similar

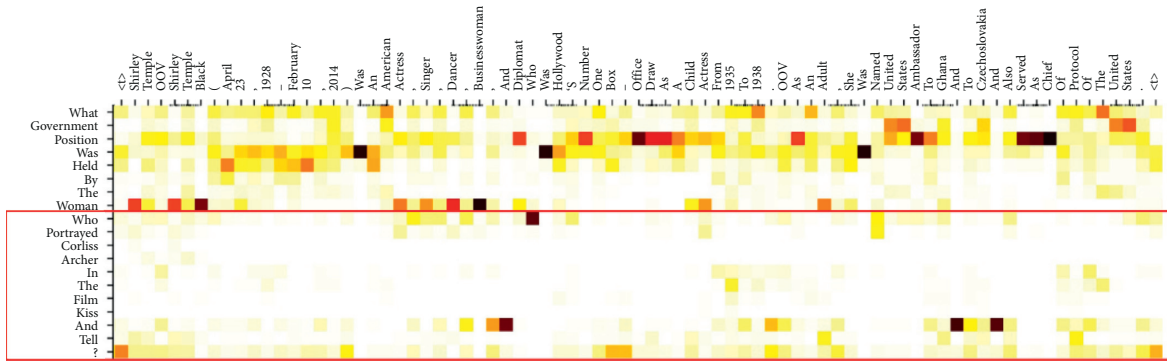


FIGURE 4: Attention heat map of the baseline model.

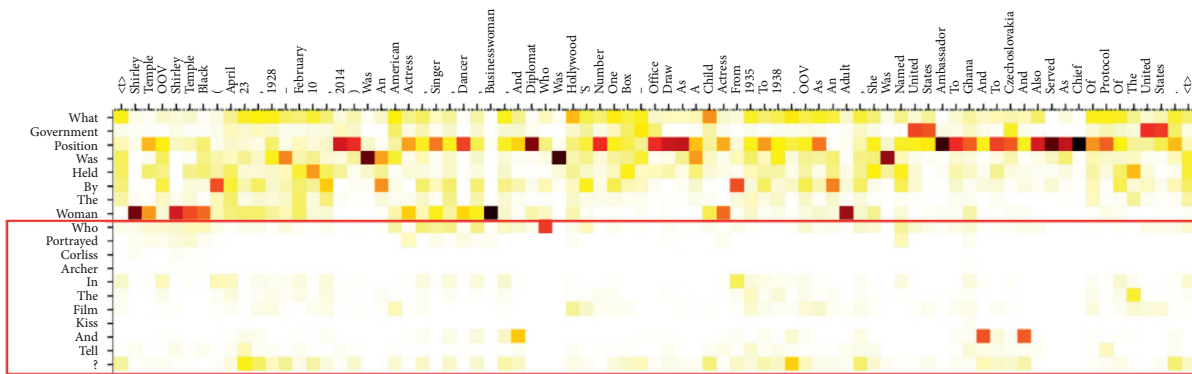


FIGURE 5: Attention heat map of the dual/ans model.

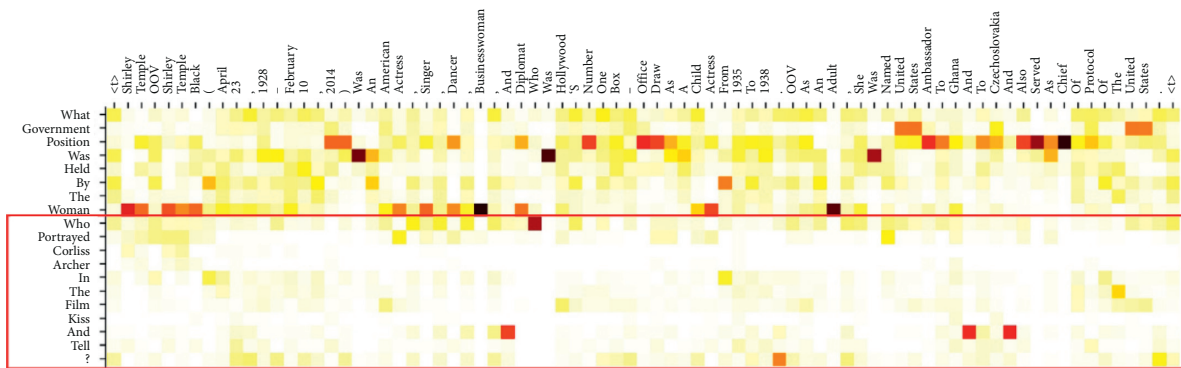


FIGURE 6: Attention heat map of the dual/sup model.

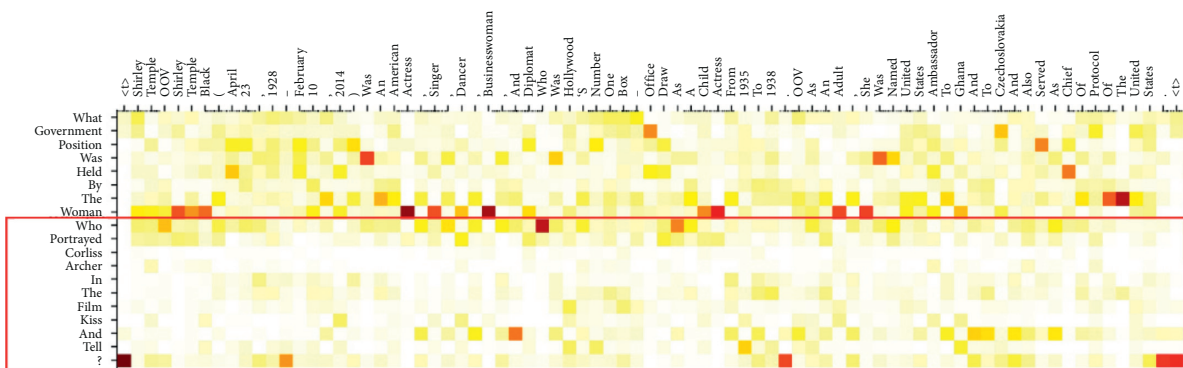


FIGURE 7: Attention heat map of the only-sup model.

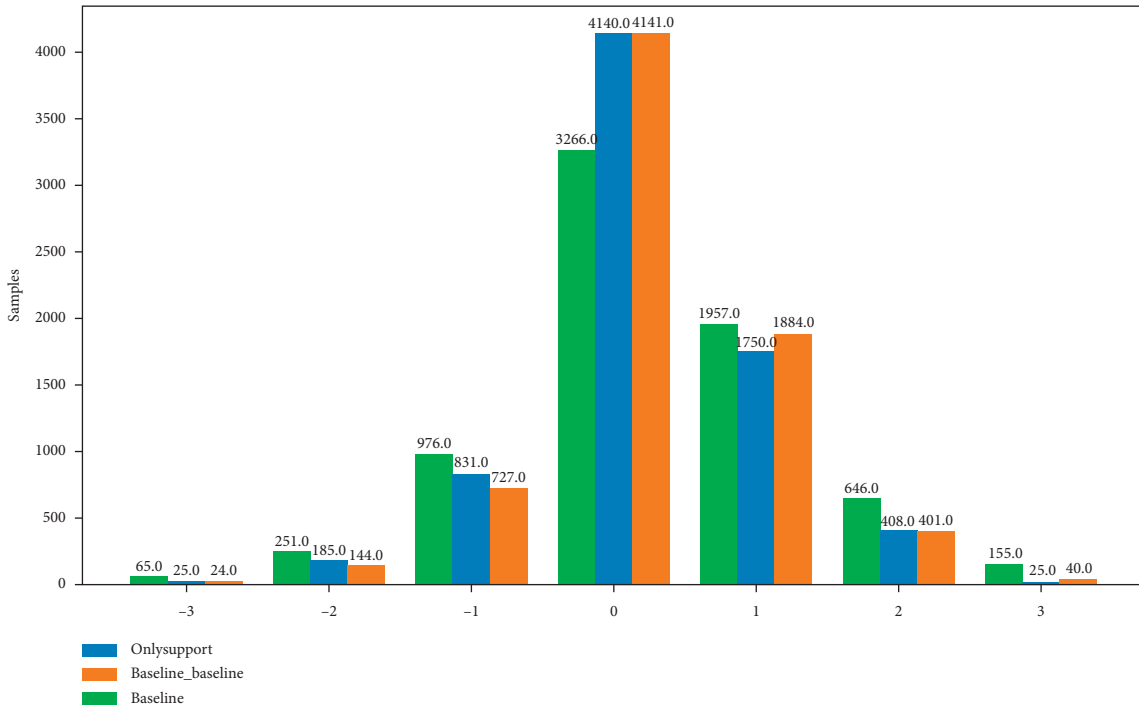


FIGURE 8: Number of predictions minus number of gold sentences.

distribution to the model that only performs supporting fact prediction tasks, and in the left half of the graph, the number of the latter is generally higher than the former, while in the right half of the graph, the opposite is true. This situation shows that the dual-channel reasoning model tends to predict more evidence sentences than the only-sup model.

To further explore the advantages of the dual-channel reasoning model, we calculated Kendall’s tau correlation between the number of predicted evidence sentences of the three models and the gold evidence sentences. As shown in Figure 9, the dual-channel reasoning model improves EM_{sup} , $F1_{sup}$, precision, recall, and Kendall’s tau.

We also introduced the two strategies of CGDe and FGI into the dual-channel reasoning architecture. Similarly, Figures 10 and 11, respectively, show the number of predictions minus the number of gold sentences of several models and the scores of all evaluation metrics.

In supporting facts prediction task, the model with the FGI strategy performs best; the model with the CGDe strategy performs slightly lower than the other two models. The reason for this result is that CGDe decomposes complex questions so that it is easy to ignore evidence sentences, while FGI can better represent each word in multiple documents. In contrast, the answer prediction ability of the model containing the CGDe strategy is significantly stronger than the other two models. Finally, the dual-channel inference model is not affected by the difference in the proportion of the two training losses when the two tasks are jointly optimized.

Tang et al. [30] used a neural decomposition model [25] to generate subquestions for multihop questions to explain the reasoning process of the question answering system to answer complex questions. In order to be able to further

evaluate the ability of our proposed dual-channel reasoning architecture to perform true multihop reasoning, we evaluated the dual-channel reasoning model, the single-channel reasoning model, and the only-answer model on the subquestion datasets proposed by Tang et al.

As shown in Table 3, the complex question is decomposed into two subquestions. Tang et al. [30] divided complex questions in the HotpotQA verification set into two subquestions and extracted the answers of subquestion 1 from the original text. Then, they saved the answer to subquestion 1, subquestion 1, and all contexts to the JSON file Dev_sub1, and they saved the answer to subquestion 2 (also the final answer to the original complex question), subquestion 2, and all contexts to the JSON file Dev_sub2. The model is trained using the HotpotQA dataset and tested on three validation sets (Dev_ori, Dev_sub1, and Dev_sub2) to evaluate the ability of different models to answer subquestions.

As shown in Table 4, under the first three columns, *correct* represents the model answered correctly and *wrong* represents the model answered incorrectly. For example, the sixth line indicates that the model correctly answers the first subquestion but incorrectly answers the complex question and the second subquestion. For all experiments, we measure EM scores for question, question_{sub1}, and question_{sub2} on 1,000 human-verified examples. When the answer predicted by the model is the same as the correct answer (both the start index and the end index are predicted correctly), the score is 1. The last three columns in Table 4 indicate the number of examples in the corresponding situation. For example, the number of examples in which the dual-channel model answers all complex questions and subquestions correctly is 282.

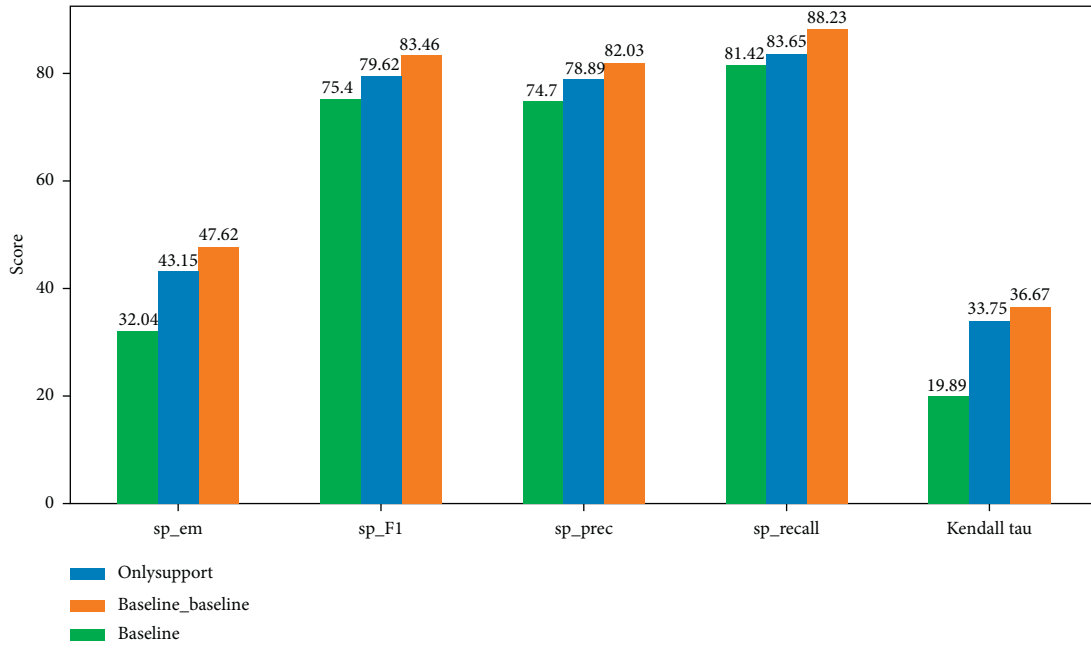


FIGURE 9: Evaluation metrics.

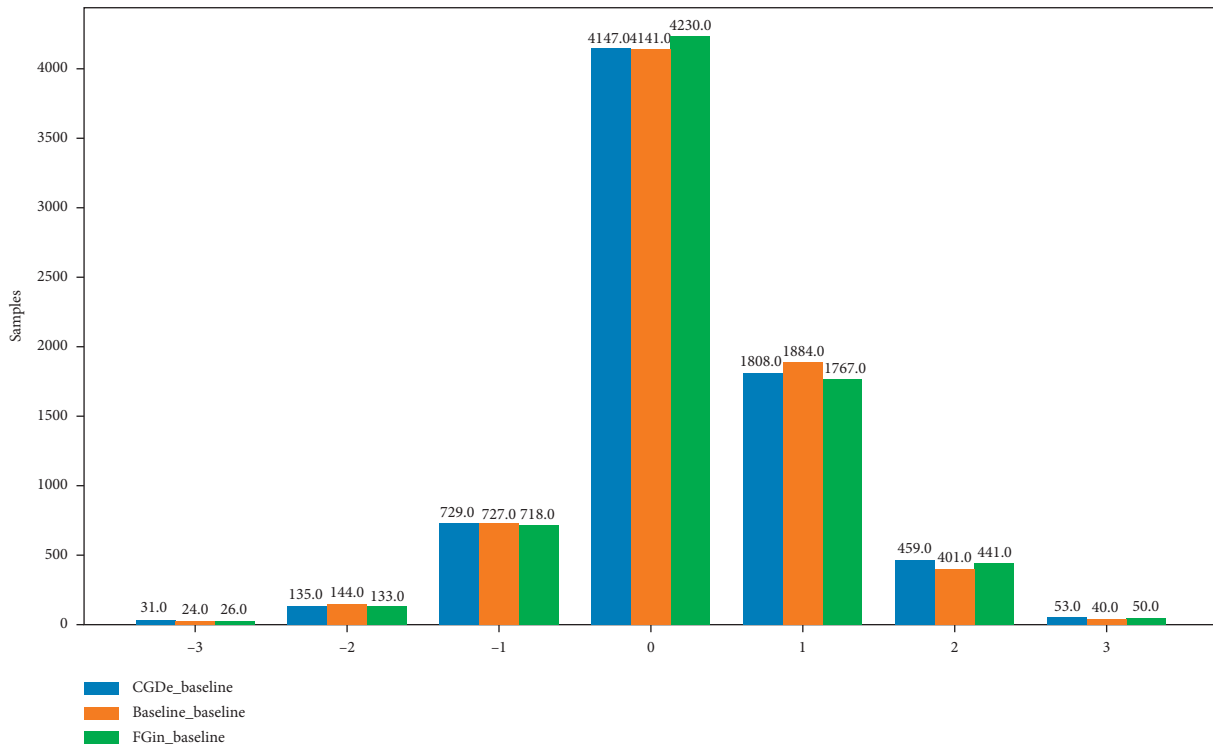


FIGURE 10: Number of predictions minus number of gold sentences (with two strategies).

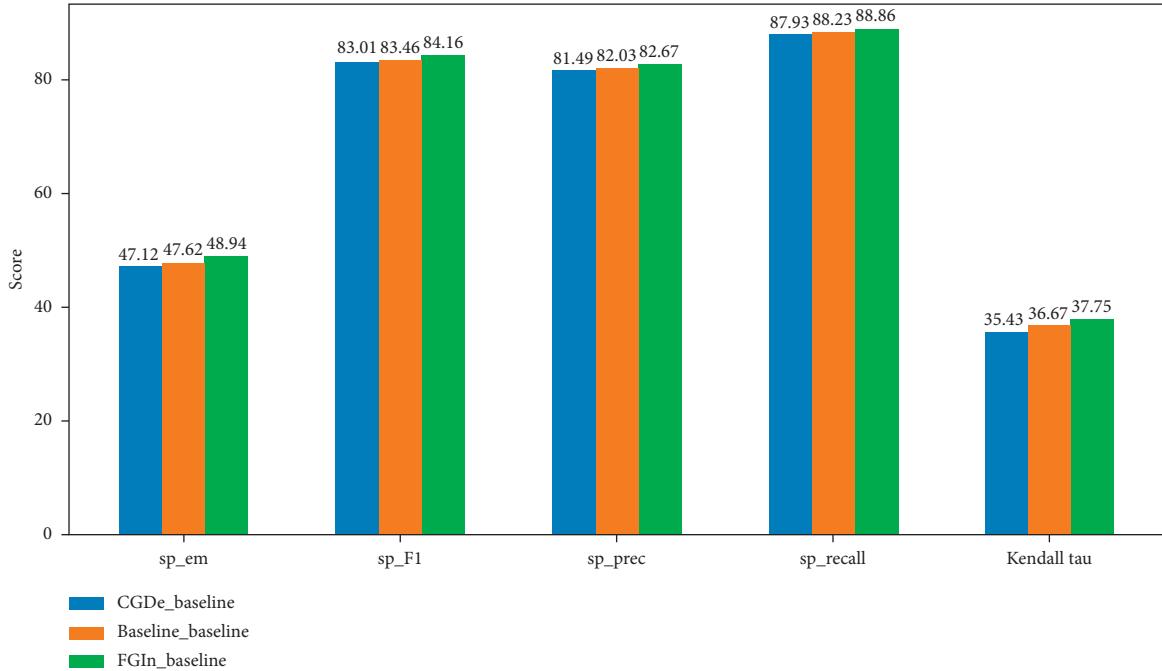


FIGURE 11: Evaluation metrics (with two strategies).

TABLE 3: An example in the subquestion dataset.

Dev_ori:
Complex question: what government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Dev_sub1:
Subquestion 1: which woman portrayed Corliss Archer in the film Kiss and Tell?
Dev_sub2:
Subquestion 2: what government position was held by Shirley Temple?

TABLE 4: Categorical EM statistics (%) of subquestion evaluation for the three models.

Question	question _{sub1}	question _{sub2}	Baseline model	Dual-channel model	Only-ans model
Correct	Correct	Correct	26.7	28.2	26.3
Correct	Correct	Wrong	8.6	6.0	8.2
Correct	Wrong	Correct	14.6	14.7	17.4
Correct	Wrong	Wrong	4.8	5.4	4.3
Wrong	Correct	Correct	2.9	3.2	4.0
Wrong	Correct	Wrong	24.7	21.4	21.4
Wrong	Wrong	Correct	1.9	3.0	2.2
Wrong	Wrong	Wrong	15.8	18.1	16.2

As shown in Table 4, the dual-channel model has the largest number of examples that can correctly answer complex questions and subquestions. The model has the least number of examples when only one subquestion can be answered correctly, and the complex question is still answered correctly because this situation is not consistent with common sense.

6. Conclusion and Future Work

In this paper, we propose a dual-channel reasoning architecture for complex question answering. The dual-channel reasoning architecture is applied to the feature interaction framework and graph-based models to verify its general applicability. In the experiments, we show that our models

significantly and consistently outperform the baseline model, especially in supporting fact prediction tasks. After more detailed experimental analysis, it is proved that the dual-channel reasoning structure has stronger step-by-step reasoning ability than the single-channel reasoning structure. In the future, we believe that the following issue will be worth studying. For the dual-channel reasoning architecture, the interaction strategy between the two channels, such as the soft parameter sharing of the homogeneous neural network components of the two channels, is worthy of further study.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the National Key Research and Development Program of China (Grant no. 2018YFC0832304) and Fundamental Research Funds for the Central Universities (Grant no. 2020YJS012).

References

- [1] L. Qin, X. Xu, W. Che, Y. Zhang, and T. Liu, "Dynamic fusion network for multi-domain end-to-end task-oriented dialog," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6344–6354, Seattle, WA, USA, April 2020.
- [2] Y. Dai, H. Li, C. Tang, Y. Li, J. Sun, and X. Zhu, "Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 609–618, Seattle, WA, USA, July 2020.
- [3] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [4] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [5] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [6] X. Cao and Y. Liu, "Coarse-grained decomposition and fine-grained interaction for multi-hop question answering," *Journal of Intelligent Information Systems*, 2021, <https://arxiv.org/abs/2101.05988>.
- [7] Y. Feldman and R. El-Yaniv, "Multi-hop paragraph retrieval for open-domain question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2296–2309, Florence, Italy, July 2019.
- [8] Y. Fu and Y. Liu, "CGSPN: cascading gated self-attention and phrase-attention network for sentence modeling," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 147–168, 2021.
- [9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, TX, USA, November 2016.
- [10] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018.
- [11] S. Reddy, D. Chen, and C. D. Manning, "CoQA: a conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.
- [12] Z. Yang, P. Qi, S. Zhang et al., "HotpotQA: a dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October 2018.
- [13] Y. Jiang and M. Bansal, "Avoiding reasoning shortcuts: adversarial evaluation, training, and model development for multi-hop QA," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2726–2736, Florence, Italy, July 2019.
- [14] Y. Fang, S. Sun, Z. Gan, R. Pillai, S. Wang, and J. Liu, "Hierarchical graph network for multi-hop question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8823–8838, November 2020.
- [15] L. Qiu, Y. Xiao, Y. Qu et al., "Dynamically fused graph network for multi-hop reasoning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6140–6150, Florence, Italy, July 2019.
- [16] K. Nishida, K. Nishida, M. Nagata et al., "Answering while summarizing: multi-task learning for multi-hop QA with evidence extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2335–2345, Florence, Italy, July 2019.
- [17] Y. Lan and J. Jiang, "Query graph generation for answering multi-hop complex questions from knowledge bases," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Seattle, WA, USA, July 2020.
- [18] Y. Sun, L. Zhang, G. Cheng, and Y. Qu, "SPARQA: skeleton-based semantic parsing for complex questions over knowledge bases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8952–8959, New York, NY, USA, February 2020.
- [19] Y. Chen, H. Li, Y. Hua, and G. Qi, "Formal query building with query structure prediction for complex question answering over knowledge base," in *International Joint Conference on Artificial Intelligence (IJCAI)*, Yokohama, Japan, July 2020.
- [20] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, "A survey on complex question answering over knowledge base: recent advances and challenges," 2020, <https://arxiv.org/abs/2007.13069>.
- [21] Y. Chen, L. Wu, and M. J. Zaki, "Bidirectional attentive memory networks for question answering over knowledge bases," in *Proceedings of the 2019 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2913–2923, Minneapolis, MN, USA, June 2019.
- [22] K. Xu, Y. Lai, Y. Feng, and Z. Wang, “Enhancing key-value memory neural networks for knowledge based question answering,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2937–2947, Minneapolis, MN, USA, June 2019.
- [23] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bidirectional attention flow for machine comprehension,” 2016, <https://arxiv.org/abs/1611.01603>.
- [24] V. Zhong, C. Xiong, N. S. Keskar, and R. Socher, “Coarse-grain fine-grain coattention network for multi-evidence question answering,” 2019, <https://arxiv.org/abs/1901.00603>.
- [25] S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi, “Multi-hop reading comprehension through question decomposition and rescoring,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6097–6109, Florence, Italy, June 2019.
- [26] Y. Jiang and M. Bansal, “Self-assembling modular networks for interpretable multi-hop reasoning,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4464–4474, Hong Kong, China, November 2019.
- [27] G. P. S. Bhargav, M. Glass, D. Garg et al., “Translucent answer predictions in multi-hop reading comprehension,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 7700–7707, New York, NY, USA, February 2020.
- [28] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” 2018, <https://arxiv.org/abs/1810.00826>.
- [29] M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang, “Cognitive graph for multi-hop reading comprehension at scale,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2694–2703, Florence, Italy, July 2019.
- [30] Y. Tang, H. T. Ng, and A. K. H. Tung, “Do multi-hop question answering systems know how to answer the single-hop sub-questions?,” 2020, <https://arxiv.org/abs/2002.09919>.

Research Article

Privacy-Preserving Efficient Data Retrieval in IoMT Based on Low-Cost Fog Computing

Na Wang ¹, Yuanyuan Cai ², Junsong Fu,³ and Jie Xu³

¹School of Cyber Science and Technology, Beihang University, Beijing, China

²National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, China

³School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China

Correspondence should be addressed to Yuanyuan Cai; caiyuanyuan@btbu.edu.cn

Received 9 May 2021; Accepted 12 June 2021; Published 22 June 2021

Academic Editor: Fei Xiong

Copyright © 2021 Na 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 rapid development of Internet of Medical Things (IoMT) is remarkable. However, IoMT faces many problems including privacy disclosure, long delay of service orders, low retrieval efficiency of medical data, and high energy cost of fog computing. For these, this paper proposes a data privacy protection and efficient retrieval scheme for IoMT based on low-cost fog computing. First, a fog computing system is located between a cloud server and medical workers, for processing data retrieval requests of medical workers and orders for controlling medical devices. Simultaneously, it preprocesses physiological data of patients uploaded by IoMT, collates them into various data sets, and transmits them to medical institutions in this way. It makes the entire execution process of low latency and efficient. Second, multidimensional physiological data are of great value, and we use ciphertext retrieval to protect privacy of patient data in this paper. In addition, this paper uses range tree to build an index for storing physiological data vectors, and meanwhile a range retrieval method is also proposed to improve data search efficiency. Finally, bat algorithm (BA) is designed to allocate cost on a fog server group for significant energy cost reduction. Extensive experiments are conducted to demonstrate the efficiency of the proposed scheme.

1. Introduction

Until now, more than 150 million people worldwide have been suffering from COVID-19, resulting in more than 3 million deaths. Mortality rates of COVID-19 are different across the world. In areas of poverty and lack of medical resources, more death occurrences because of the exhaustion of medical resources pose a significant threat to health and safety of healthcare workers. To protect medical workers, it is particularly important for them to use IoMT to manage, diagnose, and provide treatment advices to patients remotely. The medical devices or health detection sensors are connected with a computer which is then connected to a cloud sever via a network. Physiological data of patients such as blood sugar, blood pressure, heart rate, and neutrophils are uploaded to the computer and cloud server for storage. Retrieving relevant data on a cloud sever enables healthcare

workers to remotely interact with patients for pathological analysis and effective diagnosis.

At present, the industrial standards of Internet of Medical Things have been improved, and technological innovations have been emerging fast. These have made personalized medicine increasingly popular. Medical workers are able to analyze patients' physiological data remotely in their own environment. The development of this remote approach depends on advancement of sensing, monitoring, and data processing technologies. This paper uses the Internet of Medical Things to deal with the challenges. Under the premise of protecting privacy of patients' data, the patients' data are to be retrieved and analyzed, through remote monitoring of patients, diagnosis, analysis, and provision of telemedicine suggestions. Apparently, the Internet of Medical Things involves a large number of detection devices or sensors. They are widely distributed and create a large amount of data.

In order to manage these IoMT devices and organize data more efficiently, some medical institutions and medical workers outsource large-scale data to cloud servers [1–4]. However, a “distance” between medical workers and cloud servers likely leads to a high delay for medical workers to diagnose and analyze diseases. At the same time, with development of intelligent medical treatment, medical workers need immediate and efficient retrieval of patients’ data. This is reasonable considering that certain diseases such as heart disease and stroke have a rapid onset, requiring immediate and rapid diagnosis by medical workers. It is difficult for existing solutions to meet all the above actual needs.

In order to reduce network latency, He et al. [5] use fog computing to efficiently utilize cloud resources in the network, which brings sustainability to data processing in IoMT. Fog computing was first proposed by Professor Stolfo of Columbia University. Cisco redefined fog computing and proposed an application method, which made fog computing famous. The fog computing system is located between the cloud service and the medical worker, with features of low latency, high computing efficiency, and decentralization. The architecture of fog computing system is close to the edge of network and presents the characteristics of distribution. Because fog computing is “closer” to healthcare workers than to cloud services, fog computing preprocesses requests sent by medical workers and then uploads them to cloud servers for retrieval. Therefore, efficiency and latency are both considered. It is better than retrieval based only on cloud services. Therefore, fog computing based IoMT emerges [6–9].

Because cloud server is “curious and honest” [10], traditional retrieval is generally based on plaintext retrieval. However, directly uploading patients’ physiological data or index to a cloud server leads to privacy disclosure of patients. If the encrypted data are uploaded to a cloud server and the retrieval is not processed based on ciphertext, then medical workers need to decrypt the ciphertext before retrieving it. Development of ciphertext retrieval based on multiple keywords improves efficiency and accuracy of retrieval [11, 12]. It also ensures the privacy and security of medical workers, promoting retrieval services based on the IoMT to a certain extent.

Different from existing ciphertext retrieval schemes in the cloud computing, this paper constructs a ciphertext retrieval scheme based on the fog computing system. Our scheme adopts a vector space model. Each physiological data is regarded as a point in a high-dimensional space, and a corresponding data vector is generated by a medical institution. At the same time, data vectors are preprocessed to construct a range tree index, so as to improve the efficiency of retrieval. Finally, the data set and the range tree index are encrypted and sent to the cloud server for storage. After the medical workers are authorized by the medical institution, the query vector is uploaded to the fog computing system. The fog computing system is “safe and reliable” and it shares the security key with medical workers. It is also responsible for encrypting the query vector sent by medical workers. The fog computing system then sends a query trapdoor and a retrieval range vector to the cloud server. By range retrieval

on cloud server, ciphertext data are returned to the fog computing system. Finally, the fog computing server sends decrypted data to medical workers for disease diagnosis and analysis.

This paper adopts a two-layer fog computing architecture. The first layer is composed of high-end intelligent devices, such as routers, switches, and gateways, which are used to collect data from IoMT devices. The second layer is a fog computing server group made up of multiple high-performance servers that process data and execute orders sent by the healthcare workers and the IoMT. However, with high efficiency and low delay, management of resources becomes challenging in the face of frequent command requests from medical workers and IoMT. In addition, the fog computing system has a risk of high energy cost while executing orders. The energy cost mainly comes from the execution environment, refrigeration equipment, and power regulation. For fog computing system, high energy cost is a key issue. Deploying the system costs a lot of energy. The main source of energy is fossil fuel, which potentially causes a serious greenhouse effect. Therefore, optimizing orders configuration of server improves efficiency of fog computing system.

The main contributions of this paper are summarized as follows:

- (1) This paper proposes an IoMT retrieval service based on fog computing, which enables medical workers to efficiently obtain IoMT data. The fog computing system makes the data transmission of IoMT devices and the retrieval request of medical workers efficient with low latency.
- (2) A range tree is adopted to construct the index of data, which significantly improves retrieval efficiency. In order to prevent the privacy leakage of medical workers during retrieval, a ciphertext retrieval scheme based on multibody feature data was proposed. At the same time, the scheme also improves retrieval accuracy.
- (3) Within the fog computing system, a scheduling algorithm is designed to reduce energy cost. This algorithm can not only ensure the high efficiency of retrieval in our scheme, but also significantly reduce energy cost of the system.
- (4) In-depth analysis of efficiency and accuracy of the retrieval data of the scheme is provided in this paper. Moreover, we conduct simulation on the actual data set. Simulation results show that the proposed scheme achieves high efficiency and accuracy, while significantly reducing energy cost.

The rest of this paper is organized as follows: in Section 2, this paper introduces relevant research and illustrates innovation of this paper’s scheme. In Section 3, the architecture and system model of the IoMT are described, and functions of their parts are introduced. Finally, the threat model and symbol description are introduced. In Section 4, the algorithm based on ciphertext retrieval is introduced in detail. The construction method of range tree index and

functions of the application layer of IoMT by using range retrieval are also introduced. Section 5 presents bat algorithm (BA) which allocates resources for orders on the fog computing system. In Section 6, security, retrieval efficiency, and energy cost of the scheme are simulated and analyzed, and the rationality and effectiveness of the scheme are proved. Finally, the article is summarized in Section 7.

2. Related Work

Previous IoMT research focused on applications of physical testing equipment. For example, Hijazi et al. [13] employed IoMT in detection of heart sound, through signal processing and auxiliary diagnosis, but they did not mention how to retrieve data of patients, and data privacy protection. Redlarski et al. [14] proposed a machine learning algorithm to filter and analyze patient data uploaded by the IoMT to achieve disease warning. However, this scheme, without fog computing system, provides privacy protection with low efficiency and long delay. Rizk et al. [15] elaborated on privacy protection, but they did not propose how to retrieve data. Mishra et al. [16] proposed a fog computing service scheduling algorithm and discussed how to reduce energy cost of fog computing system, but it was not combined with a retrieval scheme of the IoMT.

Previous studies on patients' data analysis and retrieval are commonly based on plaintext, which violates patients' privacy rights. Therefore, the current research direction turns to ciphertext retrieval. At present, research studies on ciphertext retrieval are mainly based on cloud services, and there are few ciphertext retrieval studies based on fog computing system and even fewer ciphertext retrieval studies on IoMT. Cao et al. [17] first proposed a privacy-preserving multikeyword ranked search over encrypted data in cloud computing (MRSE). In this scheme, the secure KNN algorithm is used for retrieval, and a reversible matrix and random split indicator are used to encrypt data vectors and retrieval request vectors, so as to realize ciphertext retrieval. However, the index is not processed effectively in this scheme, which results in low retrieval efficiency. Fu et al. [12] realized personalized search by encrypting and outsourcing data. In the scheme, the index is partitioned into blocks, and then the index tree is constructed by blocks. In addition, the interest model of medical workers is added to improve the retrieval efficiency. However, their retrieval accuracy is low because of the truncated index tree. Xia et al. [10] proposed a secure dynamic multikeyword sorting search scheme based on cloud data. The scheme is designed based on a clustering algorithm, which clusters the data first and then builds a tree to improve retrieval efficiency. However, due to different values of the clustering algorithm, the clustering and retrieval results are different, resulting in inaccurate retrieval.

In applications of IoMT, previous studies [5, 15, 16, 18] likely involved no ciphertext retrieval and no fog computing. In this paper, we study and propose relevant schemes to protect the privacy of patients, improve the retrieval efficiency of medical workers, and reduce the delay of data transmission. At the same time, this paper adopts the

scheduling algorithm of fog computing system, which reduces the energy cost of the whole system. In terms of retrieving encrypted data, Cao et al. [17] proposed the concept of multikeyword ciphertext retrieval, but there was no efficient index processing, resulting in low retrieval efficiency. The scheme proposed by Fu et al. [12] divided the index into blocks, but the blocks are simple and the structure is complex, which results in low retrieval efficiency. Xia et al.'s scheme [10] uses clustering algorithm, but this scheme leads to low search accuracy. In this paper, we improve the efficiency and accuracy of retrieval.

3. Problem Description

3.1. Architecture of the Internet of Medical Things. Figure 1 shows the architecture of the Internet of Medical Things used in this paper. The architecture in the figure is divided into three layers: perceptual layer, network layer, and application layer. The following is an introduction to functions of each layer:

- (1) Perceptual layer is composed of medical sensors, identification QR codes, transmission paths, and gateway. Patients' blood pressure, blood sugar, neutrophils, and other data are collected by sensors, and they are identified by two-dimensional code. At last, the data are uploaded to the network layer for further processing from WIFI and other transmission channels.
- (2) Network layer is composed of fog computing system and public cloud server. The fog computing system is responsible for collecting the patient's physiological data. First, the cloud server normalizes them and then generates a data vector F_i in which each value represents a physical feature of the patient. Finally, the cloud computing system sends the patient's data vector set \mathcal{F} to medical institutions for encryption before uploading them to the cloud server, which aims to protect the patient's privacy. The cloud server is responsible for collecting and storing the encrypted physiological data of patients and the encrypted index. Medical workers retrieve the physiological data of patients through the range and use them for medical analysis in the cloud.
- (3) Application layer is responsible for realizing the specific functions of the IoMT. This paper mainly proposes three specific functions.

First, the application layer manages patient information, by collecting physiological data of patients and establishing a patient information database to centrally manage patients.

Second, medical workers retrieve suspected patients' information from the cloud by inputting range of physiological data of a disease's characteristics. For example, the body temperature of COVID-19 patients is severally above 37.4°C , with creatinine value of more than $100\ \mu\text{mol/L}$ and interleukin-6 of $150\ \text{pg/mL}$. Previous retrieval schemes, such as KNN algorithms, retrieve the most relevant first k value. However, normal physiological indexes of a human body are

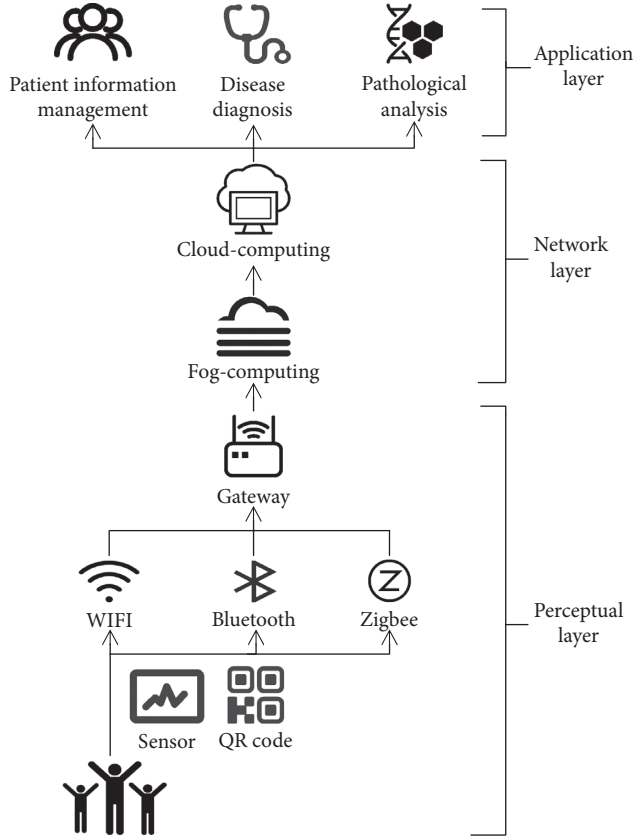


FIGURE 1: Architecture of the Internet of Medical Things (IoMT).

more likely a range rather than a specific value. Therefore, this paper uses range search, which requires the medical workers to input an appropriate range to preliminarily screen out suspected patients, and then conduct further detailed screening. This enables remote diagnosis and treatment, with protection for health workers from infection.

Third, the application layer also implements pathological analysis. For example, in the case of an unknown disease, healthcare workers want to know the physical characteristics of the patient with the disease. The medical workers identify the patient with the disease by “coloring” the patient and searching for the range of certain physiological indicators. If there is a cluster of color-coded patients within a certain range, this indicates that the pathological characteristics of the disease are related to the physiological index.

3.2. System Model. Figure 2 is the system model of the scheme designed in this paper. The following is a functional introduction of each part of the model:

Medical institutions

The medical institution encrypts the data set \mathcal{F} transmitted by the fog computing system as \mathcal{E} , then constructs a range tree index according to \mathcal{F} , and encrypts it as \mathcal{I} . The medical institution then delivers the encrypted data set \mathcal{E} to the cloud server and the

encrypted index \mathcal{I} to the fog computing system. In addition, medical institutions transfer the shared keys to trusted medical workers.

Medical workers

When the medical worker obtains the shared key of the medical institution, the medical worker transfers the retrieval range, key k , and query vector Q to the fog computing system according to their own retrieval demands. The fog computing system returns required data to healthcare workers after it completes the retrieval on the cloud server. In addition, medical workers issue orders to fog computing systems to remotely control IoMT devices.

IoMT device or sensor

An IoMT device or sensor collects physiological data from a patient and uploads it to a fog computing system. The fog computing system then preprocesses the data and transmits it to a medical institution. In addition, medical workers remotely control IoMT devices through fog computing system, such as adding or deleting devices and adjusting patients’ intelligent medical devices.

Fog computing system

The fog computing system is “safe and reliable” and it is responsible for receiving data uploaded by IoMT devices or sensors. The fog computing system then collates the data and sends it to the medical institution to update the data set. In addition, the fog computing system receives the retrieval range and query vector Q sent by the medical worker. Because the fog computing system is safe and reliable, it shares the key with the medical workers. The system encrypts the query vector and generates trapdoor according to the key shared by the medical workers. At the same time, the fog computing system generates the retrieval range vector R according to the retrieval range and then uploads it and query trapdoor T_Q to the cloud server for retrieval. It is also responsible for receiving the results returned by the cloud server. Finally, the fog computing system uses the key to decrypt the data set and send it to the medical workers, minimizing the workload of the medical workers.

Public cloud server

The public cloud server is responsible for storing the encrypted data set \mathcal{E} and encrypted range tree index \mathcal{I} uploaded by the medical institution. In addition, the cloud server receives query trap T_Q and retrieval range vector R according to the retrieval range. Then, the cloud server uploads the retrieval range and query trapdoor T_Q to the cloud server for retrieval, and it is also responsible for receiving the results returned by the cloud server.

3.3. Threat Model. In this paper, cloud servers are “curious and honest” and follow the orders of healthcare workers. At the same time, they “curiously” analyze the data retrieved by

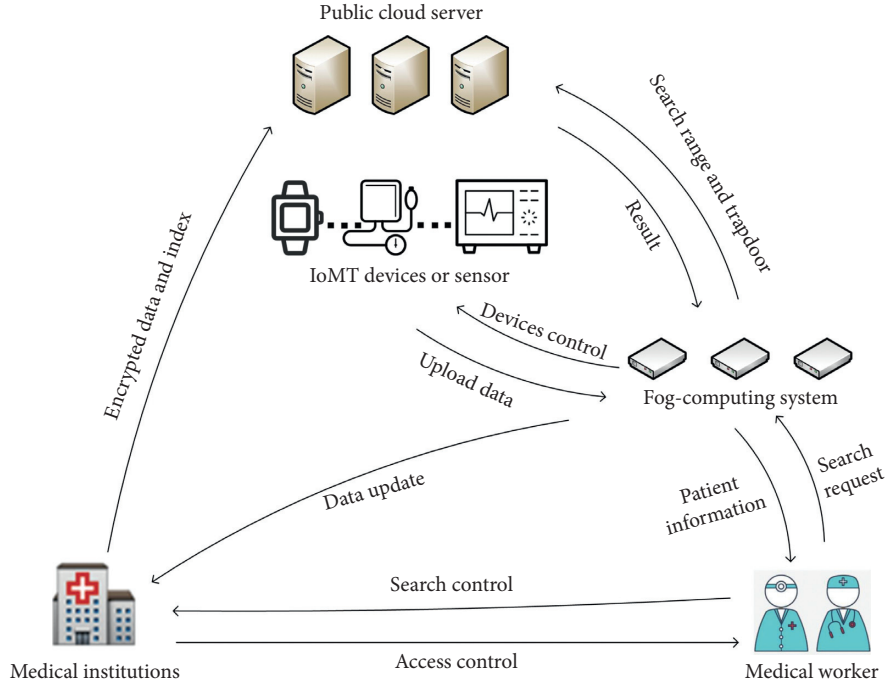


FIGURE 2: System model.

medical workers, which eventually leads to disclosure of the privacy of medical workers and data. Based on the available information of cloud server, this paper establishes two threat models:

Known ciphertext model

The cloud server obtains the encrypted data set \mathcal{E} and encrypted index \mathcal{F} sent from the medical institution. They know nothing else and only attack ciphertext to gain privacy.

Known background knowledge model

Under the condition of known background knowledge model, the cloud server analyzes the retrieval process of medical workers. Then, it tries to find the connection between ciphertext and index by statistical information of medical workers' search records and get the connection between keyword frequency and physical characteristics data.

3.4. Symbol Description. For convenience, some notations are first defined as follows:

- (i) \mathcal{F} : the plaintext data set $\mathcal{F} = \{F_1, F_2, \dots, F_m\}$ contains m patient physiological data.
- (ii) \mathcal{E} : data set \mathcal{F} encryption form $\mathcal{E} = \{C_1, C_2, \dots, C_m\}$, a total of m encrypted data.
- (iii) F_i : plaintext data vector $F_i = \{f_{i1}, f_{i2}, \dots, f_{in}\}$, where each dimension is the normalized value of the patient's body index data.
- (iv) \mathcal{S} : the range tree index is built by the data vector $F_i (0 < i \leq m)$, and then the range tree index is encrypted to get \mathcal{S} .

- (v) Q : if a dimension of query vector Q is 1, it represents the medical worker to retrieve the physiological index data. If the value is 0, it indicates that the physiological indicator data was not retrieved.
- (vi) T_Q : the query vector Q is encrypted to generate trapdoor T_Q .
- (vii) R : the retrieval range of the patient's physiological index data was generated by the fog computing system according to the retrieval range sent by the medical workers to generate the retrieval range vector $R = \{e_{w_1}, e_{w_2}, \dots, e_{w_n}\}$.
- (viii) VM : the fog computing system VM is composed of m high-performance servers, $VM = \{V_1, V_2, \dots, V_m\}$.
- (ix) S : the list of orders sent to the fog computing system by an IoMT device or medical worker consists of n orders. $S = \{s_1, s_2, \dots, s_n\}$.

4. Secure Storage and Retrieval of Medical Data

4.1. Framework of Ciphertext Retrieval. The IoMT devices upload data and send it to the fog computing system, which is responsible for collating these data and generating data sets. Then, the fog computing system sends the data sets to the medical institution to update them. The medical institution uses random numbers; a pair of reversible matrices M_1^T, M_2^T ; master key sk ; and key k to encrypt the data vector set \mathcal{F} and the range tree index constructed from the data vector. Then, it sends the data set and index to the cloud server. Medical workers submit retrieval request to fog service system. The fog server generates trapdoor by encrypting query vector and uploads it to the cloud along

with the range search vector R . At last, the results are returned to the medical worker. The scheme framework designed in this paper mainly includes the following algorithms:

- (i) Key generation $(1^{l(n)}) \rightarrow (\text{sk}, k)$: this step mainly generates a master key sk and key k to encrypt index and data, respectively.
- (ii) Constructing an encrypted index $(\mathcal{F}, \text{sk}) \rightarrow \mathcal{I}$: the medical institution first uses the data vector set to construct a range tree index and then encrypts it with a secure algorithm to get index \mathcal{I} .
- (iii) Data encryption $(\mathcal{F}, k) \rightarrow \mathcal{E}$: the medical institution encrypts the data set using a symmetric encryption algorithm to obtain the ciphertext set \mathcal{E} .
- (iv) Trapdoor generation $(Q, \text{sk}) \rightarrow T_Q$: the fog computing system generates a query trapdoor T_Q based on the query vector sent by the medical worker and the key shared by the medical worker.
- (v) Retrieval $(T_Q, \mathcal{I}, R) \rightarrow C_Q$: in this process, the cloud server receives the query trapdoor T_Q and the retrieval range vector R from the fog computing system, and then it retrieves the corresponding ciphertext data C_Q in the range. Finally, it sends C_Q to the fog computing system.
- (vi) Decryption $(C_Q, k) \rightarrow F_Q$: the cloud server returns the retrieved encrypted data to the fog computing system. The fog computing system decrypts the data according to the key shared with the medical workers and sends it to the medical workers.

The following is a detailed description of the main algorithms in the scheme architecture of this paper:

Key generation $(1^{l(n)})$

The medical institution generates an $(n + u + 1)$ dimensional split indicator vector H , where each element is a random 1 or 0. At the same time, the medical institution generates two $(n + u + 1)$ -dimensional reversible matrices M_1^T and M_2^T , where each element is a random integer. In this paper, the master key $\text{sk} = \{H, M_1^T, M_2^T\}$. In addition, the medical institution selects an n -bit pseudosequence to generate the data encryption key k .

Building an encrypted index (\mathcal{F}, sk)

Medical institutions construct range tree index according to \mathcal{F} and then extend the n -dimensional vector A of each node in the range index tree to the dimension vector \bar{A} of $(n + u + 1)$, in which the dimensions from $n + 1$ -th to $n + u$ -th are set as random integers, and the dimension $n + u + 1$ -th is set as 1. Then, the medical institution uses the split indicator vector H to split \bar{A} . If $H[i] = 0$, then $\bar{A}'[i] = \bar{A}''[i] = \bar{A}[i]$; If $H[i] = 1$, then $\bar{A}'[i]$ is g' random number, $\bar{A}''[i] = \bar{A}[i] - g'$. Finally, the medical institution obtains the encrypted index $\mathcal{I} = \{M_1^T \bar{A}', M_2^T \bar{A}''\}$ and sends it to the cloud server.

Data encryption (\mathcal{F}, k)

Symmetric encryption algorithm (for example, AES encryption) is adopted in medical institutions [17] to encrypt the plaintext data set \mathcal{F} , and the encrypted ciphertext set \mathcal{E} is outsourced to the cloud server.

Generating (Q, sk) by trap gate

The healthcare worker generates the query vector Q and sends it to the fog computing system. Similarly, the fog computing system first extends the n -dimensional query vector Q to the $(n + u + 1)$ dimensional vector \bar{Q} . It randomly selects b values between the $n + 1$ -th dimension and the $n + u$ -th dimension and set them to 1. It sets the remaining values to 0 and the value of the $n + u + 1$ -th dimension to a random number $t \in [0, 1]$. The fog computing system generates $\bar{Q} = (r \cdot Q, t)$ by multiplying the previous $(n + u)$ dimension vector by a random number r and then splits \bar{Q} according to the split vector H . When $H[i] = 1$, $\bar{Q}'[i] = \bar{Q}''[i] = \bar{Q}[i]$. When $H[i] = 0$, $\bar{Q}'[i]$ is a random number g , $\bar{Q}''[i] = \bar{Q}[i] - g$. Finally, the fog computing system generates an encrypted retrieval trap $T_Q = \{M_1^{-1} \bar{Q}', M_2^{-1} \bar{Q}''\}$ and sends it to the cloud server.

Retrieving (T_Q, \mathcal{I}, R)

The fog computing system sends trap T_Q and range retrieval vector R to the cloud server, where $R[i]$ ($i = 1, 2, \dots, n$) represents the retrieval range of the physiological data of the patient by the medical worker. The cloud server retrieves the list of data required by medical workers in the range tree based on the range tree index \mathcal{I} , the query trap T_Q , and the dynamically updated retrieval range R . The retrieval process of range trees is described in detail in Section 5.2 of this paper. The physiological data are calculated as follows:

$$\text{Physiological value} = T_Q \cdot I$$

$$\begin{aligned} &= \{M_1^{-1} \bar{Q}', M_2^{-1} \bar{Q}''\} \cdot \{M_1^T \bar{A}', M_2^T \bar{A}''\} \\ &= \bar{Q}' \cdot \bar{A}' + \bar{Q}'' \cdot \bar{A}'' \\ &= Q \cdot A. \end{aligned}$$

(1)

4.2. Structure of Range Tree Index. Range tree is an improvement of kd-tree. Although range tree needs more storage space than kd-tree, it has a significant improvement in retrieval efficiency. Because the cloud server has a large amount of storage space, it is not necessary to consider the storage space taken by the range tree. The scheme designed in this paper mainly considers efficiency and accuracy of medical workers' data query, so the range tree is adopted to construct the index. The construction process is as follows:

- (1) For the data set \mathcal{F} (data vector set $\mathcal{F} = \{F_1, F_2, \dots, F_m\}$), this paper constructs a balanced binary search tree TR from bottom to top according to the first vital sign data of all data vectors, and the data vector is stored in the leaf node.

- (2) In a subtree of a nonleaf node L_1 in the balanced binary search tree TR , the data vector set corresponding to all leaf nodes under this subtree forms a subset \mathcal{F}' (namely, $\mathcal{F}' \subset \mathcal{F}$) of \mathcal{F} , which is called the regular subset corresponding to L_1 and denoted as $\mathcal{F}(L_1)$.
- (3) The data vectors in the regular subset of nonleaf node L_1 are organized according to their second vital sign data to establish the second dimensional range tree and form a joint structure with the first dimensional range tree. Nonleaf node L_1 has a pointer to the root of the new tree $TR_1(L_1)$ and the regular subset of the leaf nodes under this node is the subset \mathcal{F}'' of the previous one-dimensional data vector set, namely, $\mathcal{F}'' \subset \mathcal{F}'$.
- (4) After recursive (2) and (3) steps, an n -dimensional range tree is constructed, as shown in Figure 3.

4.3. Data Retrieval on the High-Dimensional Range Tree. The retrieval process of the scheme designed in this paper is elaborated as follows:

- (1) After the cloud server receives the query trap T_Q , it conducts the range retrieval according to the first dimension of the retrieval range vector R from the range root node in the first dimension of index \mathcal{S} . If the corresponding value of the query trap gate T_Q is 1, then the range retrieval is carried out in the first dimensional range tree to find the required data vector (leaf node).
- (2) Based on the set of leaf nodes found in the first dimension, this paper takes them as a regular subset and finds their lowest common ancestor nodes easily through middle order traversal. Then, from this lowest common ancestor node to the balanced binary search tree in the next dimension, the process excludes leaf nodes that are not included in the first dimension range search. Finally, apply the steps in (1) to find the required data set. If the query gate $T_Q[i] = 0$ (that is, the medical worker does not retrieve the physical characteristics data of the dimension), the i -th dimension is retrieved directly from the root node to the tree of the $i + 1$ dimension. In addition, in order to prevent too many retrieval times and a narrow retrieval range width d value and for other reasons, the datum is not searchable. Therefore, this paper presets a minimum returned data quantity k . If the number of leaf nodes retrieved in the balanced binary search tree TR of a dimension is less than k , then the backtracking algorithm is called to find the leaf node closest to the lower bound of the retrieval range until k data is retrieved. Then, the retrieval is stopped and k data is returned. This guarantees that at least k data are returned per retrieval.
- (3) The cloud server recurses the above two steps, and the required data vector set is found after range retrieval. Then, according to the id of these data, the

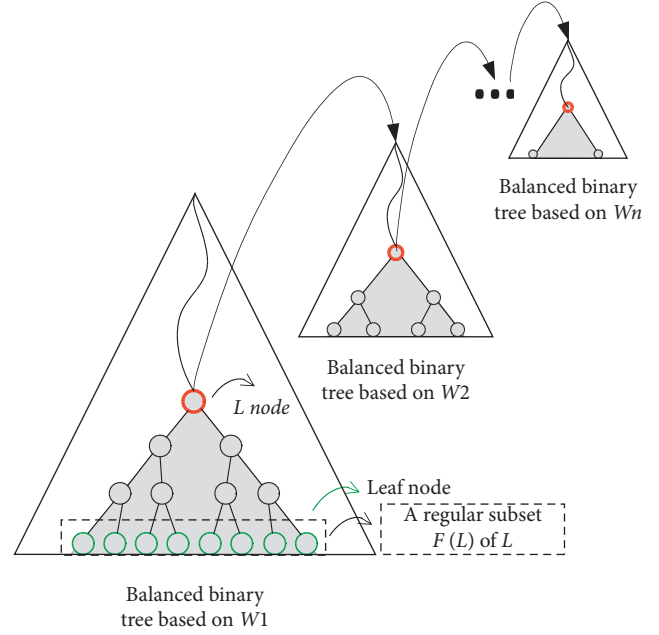


FIGURE 3: High-dimensional range tree construction.

corresponding ciphertext is returned to the fog computing system. Finally, the fog computing system decrypts it and sends it to medical workers.

The advantages of using range tree index and range retrieval are described below:

- (1) Different from the retrieval scheme of Euler distance, this paper adopts the range retrieval scheme, which is more suitable for range tree index with higher efficiency. Meanwhile, in the medical field, most of the physiological data is a certain range rather than a specific value, and hence using the range retrieval is more suitable for the IoMT.
- (2) After retrieving the balanced binary search tree of each dimension, a part of leaf nodes (i.e., the data vectors corresponding to leaf nodes) is excluded. It only needs to be retrieved in the balanced binary search tree composed of this regular subset of dimension. In this way, we significantly improve retrieval efficiency.
- (3) As for range retrieval, currently most schemes adopt kd-tree. Simulation results show that the range tree retrieval efficiency is higher than kd-tree.

4.4. The Application Layer Function of IoMT Realized by Using Range Search. In this paper, the encrypted data is outsourced to the cloud server, which not only preserves the data of patients, but also protects the privacy of patients. In addition, the range tree index is constructed according to the patient's data vector set. In this way, the function of centralized management of patient information in the application layer of IoMT is realized, and the efficiency of retrieving patient information is improved.

For the disease diagnosis function of the IoMT application layer, medical workers are required to input the query vector and multidimensional physical signs of a disease data range in the fog computing system. The fog computing system normalizes the data range and uploads it to the cloud server for retrieval according to the retrieval vector generated on the trapdoor. Finally, information about suspected patients is returned to help medical workers diagnose the disease.

For example, the diagnosis of uremia has three important indicators: the glomerular filtration rate is less than 15 ml/min, the serum creatinine is greater than or equal to 707 umol/L, and the serum potassium is less than 3.5 mmol/L. By uploading the data range of these indicators, medical workers receive information of the suspected patient within this range. The proposed scheme also helps medical workers to carry out pathological analysis. For a disease with unknown pathology, medical institutions first “color” the patients with the disease. Medical workers select appropriate body index data for range retrieval. If a large number of color-coded patients appear in the search results within a certain range, it is determined that the disease has this physiological characteristic. This helps medical workers achieve function of pathological analysis.

4.5. Update of the Range Tree. IoMT devices generate data and upload them to fog computing systems, where the data are processed into a data set and transmitted to the medical institution for updating. Medical institutions encrypt the data and upload them to cloud servers.

4.5.1. Node Insertion. Medical institutions calculate the data vector based on the updated data and insert each dimension of the data vector into the balanced binary tree according to the bottom-up rule.

4.5.2. Node Deletion. Medical institutions delete nodes in each dimension of the range tree according to the deletion rules of balanced binary tree. The node deletion of the range tree is completed after recursion to the dimension.

5. Efficient Processing of Medical Data

5.1. The Order Assignment Problem of Fog Computing System. As shown in Figure 4, this paper adopts the two-layer fog computing architecture. The first layer is composed of intelligent devices, such as routers, switches, and gateways, which are used to collect data sent by IoMT devices.

The second layer is the fog computing server group VM composed of multiple high-performance servers, which is used to process the collected data and execute the orders sent by medical workers and IoMT devices, such as the retrieval request of medical workers and the data upload request of IoMT devices. Each server V_i has a unique ID_{v_i} , main memory, bandwidth, and storage. Each order S_i is calculated

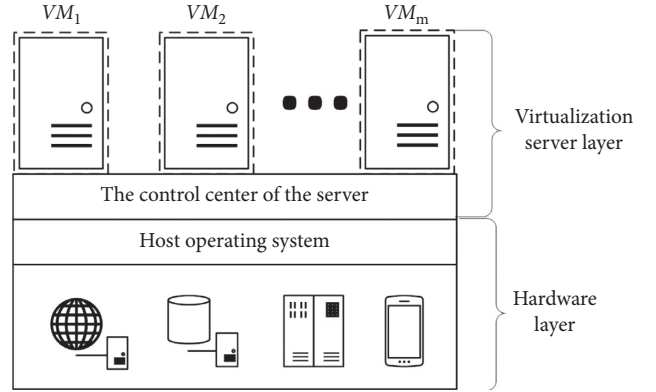


FIGURE 4: The architecture of fog computing system.

by the order ID_{s_i} , and the workload is calculated in millions of orders per second (MIPS). For each order S_i , only one fog server V_i is assigned, and service migration is not allowed until the order is completed. Therefore, for the order request list S to be assigned to the sever group VM , how to allocate the VM on the premise of meeting *sla* [16] without reducing QoS (quality of service) becomes a new challenge.

In this paper, the following assumptions are made for the fog computing system:

- (1) Stipulate the expected running time of an order S_i on fog server V_j as ETC_{ij} . In addition, all order requests are independent and heterogeneous.
- (2) The resource capabilities of all VM are heterogeneous.
- (3) An order is only allowed to execute on one V_i .

Order list S is composed of n heterogeneous orders, and the order length L_i of each order S_i is represented by millions of orders (MI). These orders run in a server group VM with m servers, so you can build an $n \times m$ ETC matrix (see Figure 5).

In addition, each fog server V_j has a processing speed P_j , so in the ETC matrix, the order S_i in server V_j has $ETC_{ij} = (L_i/P_jX)$.

Suppose the energy consumed by the server V_j in the fog computing server group (Joule) is expressed as follows: First, the paper gives the energy consumed by V_j per unit length (J/MI):

$$E_{v_j} = \begin{cases} \beta_j, & \text{if } V_j \text{ is active,} \\ \alpha_j, & \text{if } V_j \text{ is not active.} \end{cases} \quad (2)$$

Let X_{ij} be the decision variable for whether to assign orders to a particular VM . If the order S_i is assigned to a server in the fog compute server group V_j , then the X_{ij} value is 1; otherwise, it is 0. Then, the total execution time of all orders of the j -th server V_j is

$$ET_j = \sum_{i=1}^n X_{ij} \times ETC_{ij}. \quad (3)$$

Makespan M is the maximum time value for all VM :

ETC matrix					
	VM_1	VM_2	VM_3	...	VM_m
S_1	ETC_{11}	ETC_{12}	ETC_{13}	...	ETC_{1m}
S_2	ETC_{21}	ETC_{22}	ETC_{23}	...	ETC_{2m}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
S_n	ETC_{n1}	ETC_{n2}	ETC_{n3}	...	ETC_{nm}

FIGURE 5: ETC matrix.

$$M = \max(ET_j), \quad 1 \leq j \leq m. \quad (4)$$

The total energy cost of V_j is

$$E(V_j) = [ET_j \times \beta_j + (M - ET_j) \times \alpha_j] \times MIPS_j. \quad (5)$$

Thus, the energy consumed by the entire fog computing system can be obtained:

$$\varepsilon = \sum_{j=1}^m E(V_j). \quad (6)$$

The goal of this paper is to minimize the total cost ξ , which is a two-objective problem. On the one hand, it is hoped that the energy cost of fog service system is reduced as far as possible; on the other hand, it is hoped that the makespan M of fog service processing orders is shortened and hence the efficiency of fog computing system processing orders is improvable. In this paper, the total energy cost is expressed as

$$\text{Minimize } \xi = M \times \sigma + \varepsilon \times (1 - \sigma). \quad (7)$$

In this paper, a penalty value σ is set to represent the importance that medical workers attach to efficiency and energy cost. If healthcare workers are focusing on reducing energy use, they try to set a low σ value. If efficiency is important, set a high σ value.

From the above analysis, it is observed that assigning the order queue to the fog server group is an NP-hard problem [16] to minimize energy cost. In order to solve this problem, a service allocation algorithm is designed based on Algorithm 1.

5.2. Efficient Order Assignment Based on Bat Algorithm.

This section discusses the order allocation problem of fog computing system. The retrieval request of medical workers and the upload request of IoMT device data both belong to the order. In the process of handling these orders, if the orders are not allocated in advance, the fog computing system consumes a lot of energy. In order to reduce energy consumption, bat algorithm (BA) [16] is adopted to assign orders to the fog computing system.

Suppose that, for a server group VM consisting of m servers, the order sequence S consisting of n orders enters into the server group VM . An order assignment vector is specified in the following as shown in Figure 6.

Algorithm 1 is a bionic computing technology that imitates bats to capture food and avoid obstacles by emitting ultrasonic pulses and acquiring echoes at night. It is stipulated that bats (i.e., commands) fly at speed v_i at position x_i , and each command has a fixed frequency Γ_{\min} , signal strength G , and signal pulse rate z . By iteratively calculating and updating the position and speed of the order, an optimal allocation scheme is finally converged.

In the first step of the algorithm, an order set is initialized, and the order assignment vector, velocity vector, signal strength G_0 of the initial order, and pulse rate z_0 of the initial order are specified. Then, the total energy cost is calculated according to the initial input service allocation vector, and the initial position is set as the optimal position. In steps 4–7 of the algorithm, the orders calculate the new position and speed in terms of frequency, speed, and previous position. In the 8–10 steps of the algorithm, if the generated random number is greater than the signal pulse rate of the order, the current optimal distribution vector is slightly disturbed to generate a new optimal distribution vector.

In steps 11–15 of the algorithm, if the generated random number is less than the signal strength of the instruction and the frequency is greater than the frequency of the previous iteration, then the change of the optimal distribution vector is accepted, the signal strength of the instruction is attenuated, and the signal pulse rate of the instruction is increased. Consequently, the algorithm is iterated to find the optimal solution.

6. Analysis and Simulation of Efficiency and Energy Cost

In this section, the retrieval efficiency and energy cost proposed by the scheme are analyzed theoretically and verified by simulation. In terms of retrieval efficiency, this paper adopts the common corpus on the network as the data set and uses C++ for simulation. In the energy cost simulation, this paper adopts MATLAB 2019a for simulation. The simulated hardware environment is Intel Core i5-8300H CPU, 8 GB memory, and Microsoft Window 10 operating system.

6.1. Safety Analysis. In this paper, the symmetric encryption algorithm AES is adopted to encrypt the data set \mathcal{F} and generate ciphertext data set \mathcal{C} , which is uploaded to the public cloud server, effectively ensuring the security of the patient's physiological data itself. Then, reversible matrices M_1 and M_2 are generated randomly, and the index of range tree and query vector Q are encrypted to generate secure index \mathcal{I} and query trap T_Q . Then, the fog computing server uploads them to the public cloud server. Since the space of the key matrix is infinite, each randomly generated key matrix has only one reversible matrix. The probability that the public cloud server correctly forges the key matrix to crack the security index \mathcal{I} and query trap T_Q is almost 0, effectively ensuring the security of the information contained in the range tree index and query vector. Under the

Input: ETC matrix, VM processing speed, maximum number of iterations.

Output: the order assignment results for the VM (fog computing server group), the total cost ξ .

- (1) Initialize random order set: a random order assignment vector x , velocity vector v and frequency of each order Γ , order signal strength G , order signal pulse rate z . In addition, I has a random variable $\eta \in (0, 1)$;
- (2) Use formula (7) to calculate total energy cost ξ ;
- (3) x_{best} is the optimal location for each order;
- (4) Update the position and speed of the order according to steps (5), (6) and (7);
- (5) $\Gamma = \Gamma_{\min} + (\Gamma_{\max} - \Gamma_{\min})\eta$;
- (6) $v_i^t = v_i^{t-1} + (x_i^{t-1} - x_{\text{best}})\Gamma$;
- (7) $x_i^t = x_i^{t-1} + v_i^t$;
- (8) **if** ($\text{rand}_1(0, 1) > z_i$) **then**
- (9) $x_i = x_{\text{best}} + 1$;
- (10) **end if**
- (11) **if** $\text{rand}_2(0, 1) < G_i$ and $\Gamma_i^t > \Gamma_i^{t-1}$ **then**
- (12) $x_i = x_i^t$;
- (13) $G_i^{t+1} = \text{rand}_3(0, 1) \cdot G_i^t$;
- (14) $z_i^{t+1} = z_i^0 [1 - e^{-\text{rand}_4(0,1)^t}]$;
- (15) **end if**
- (16) Repeat steps 4–15 for each order;
- (17) Repeat steps 2–16 until a satisfactory convergence result or a maximum number of iterations is achieved.

ALGORITHM 1: BA.

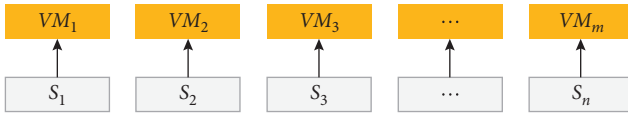


FIGURE 6: Order assignment vector.

known ciphertext model, the public cloud server only obtains ciphertext data set \mathcal{C} , security index \mathcal{S} , and query trapdoor T_Q . It is not allowed to obtain any useful data information unless it is ensured that the master key sk and key k are not artificially disclosed. In such cases, the scheme is safe.

In order to further prevent public cloud servers from mining and leaking data privacy information based on known background knowledge (that is, based on the internal connection between the security index and the query trapdoor), the mater key sk in our scheme is a random split indicator vector H , which is used to randomly split the expanded index vector \bar{A} and query vector \bar{Q} . At the same time, random numbers g' and g are introduced in this random splitting process. Through such a series of operations, it is ensured that multiple regional tree indexes and query vectors are unrelated. Even if the medical workers repeat the same query operation for many times, the query trapdoor received by the public cloud server is different, which displays the unlinkability of the query trapdoor and effectively resists the statistical analysis attack. Consequently, our scheme is also safe for the known background knowledge model.

6.2. Retrieval Efficiency Evaluation. In the simulation of retrieval efficiency, the scheme in this paper is compared with the kd-tree scheme [19], the binary balanced tree scheme in literature [10], and the MRSE scheme in literature [17]. For comparison, the simulation of our scheme is to set

the number of physical feature data records to be queried to 5. We analyze the time complexity of index construction and retrieval of range tree.

Theorem 1. *Given a data vector set consisting of n data vectors, the storage space occupied by the corresponding p -dimensional range tree is $O(n \log_n^{p-1} n)$.*

Proof. In each dimension of the range tree, each element in the leaf node set (i.e., data F_i) is stored only once at each depth. For a set of data vectors consisting of n data vectors, the height of constructing an equilibrium binary search tree is $\log n$. Since the storage space of each dimension in the range tree of each data vector is $O(\log n)$, the storage space required to construct a one-dimensional range tree of n data vectors is $O(n \log n)$. For a p -dimensional range tree, each dimension of storage space needs $O(n \log n)$, so the size of storage space required for a p -dimensional range tree is p times that of a one-dimensional range tree, that is, $O(n \log_n^{p-1} n)$. \square

Theorem 2. *The p -dimensional range tree is constituted by n data vectors, and the query time complexity is $O(\log^p n + k)$ when it is retrieved in the range tree.*

Proof. To retrieve a p -dimensional range tree, the first dimension of the range tree should be retrieved (that is, a balanced binary search tree retrieval process), and the time complexity required is $O(\log n)$. Next, search the remaining $p - 1$ dimensional range tree to obtain the following time complexity relationship:

$$O_p(n) = O(\log n) + O(\log n) \times O_{p-1}(n), \quad (8)$$

where $O_p(n)$ represents the lookup time complexity of p -dimensional range tree and $O_{p-1}(n)$ represents the lookup

time of $p - 1$ -dimensional range tree. Then, according to (9), it is deduced as follows:

$$O_2(n) = O(\log 2n). \quad (9)$$

From the recursive formula (9) the time complexity of dimensional range tree retrieval is $O(\log^p n)$. In addition, this paper needs to record the obtained k data; the total time complexity is $O(\log^p n + k)$.

The retrieval efficiency of the scheme is mainly determined by the index structure of the data set and the score of calculated data similarity. For the range tree index in this paper, the number of queries for medical workers is set to 5 in the simulation. This paper first analyzes the effect of retrieval range width d on retrieval efficiency of medical workers. If the d value is large, the result is returned with low accuracy. If the d value is small, fewer nodes are retrieved in the range. In order to return k results, the scheme needs to invoke the backtracking algorithm many times, which reduces the retrieval efficiency. The relation between the retrieval time and the retrieval range width d of our scheme is shown in Figure 7. As seen from Figure 7, if the d value is larger, the retrieval time is shorter; if the d value is smaller, the retrieval time is longer. In addition, it is shown in Figure 7 that the return of k data also has a great influence on retrieval time. When the retrieval range width d value is greater than 0.03, the retrieval time tends to converge. In order to improve retrieval efficiency, it is necessary to select a high d value, but this leads to reduction of retrieval accuracy. Therefore, the most appropriate retrieval range width d value for medical workers is between 0.03 and 0.05 in order to balance the retrieval efficiency and accuracy under the scheme proposed. A d value 0.03 is selected for this simulation.

The retrieval efficiency of the proposed scheme is compared with that of other schemes. In the scheme proposed by Cao et al. [17], tree-type index structure was not used, and the retrieval time increased linearly with the increase of the amount of patient data, resulting in lower retrieval efficiency compared with other schemes. Although both Xia et al. [10] and the scheme in this paper adopted balanced binary search tree for retrieval, the multidimensional link structure of range tree was embedded in this paper, with high retrieval efficiency, as shown in Figure 8. Since both this study and [10] adopt a tree structure to construct indexes, the retrieval time is prolonged with the increase of data quantity in a logarithmic manner. Figure 8 shows that the retrieval time of the scheme is less than that in [10] and the retrieval efficiency is higher. Chen et al. [19] designed an index structure based on kd-tree, whose time complexity is close to $O(\sqrt{n}^p + k)$, which is higher than the time complexity $O(\log^p n + k)$ of our scheme. Our scheme has less retrieval time and higher retrieval efficiency than the scheme in [19], as shown in Figure 8. Both the scheme in this paper and the schemes in [10, 19] show a linear increase in the retrieval time with the increase in the number of return orders, while the retrieval time in [17] is almost unaffected by the increase in the number of orders, as shown in Figure 9. However, because the range tree index structure is used in

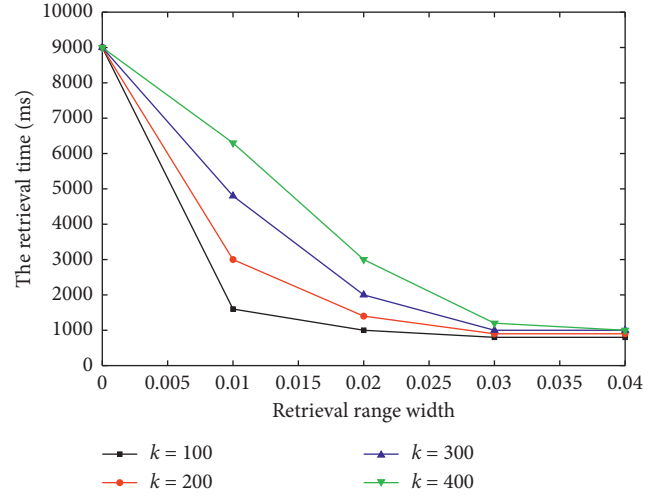


FIGURE 7: Relationship between retrieval time and retrieval range width.

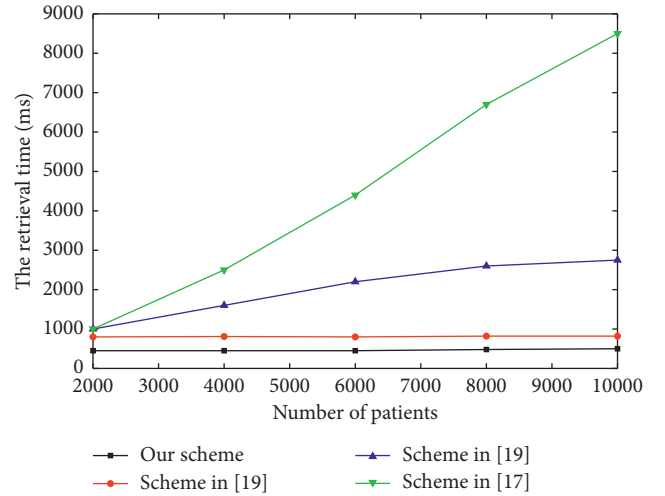


FIGURE 8: The retrieval time changes with the number of patients m ($k = 100, d = 0.03$).

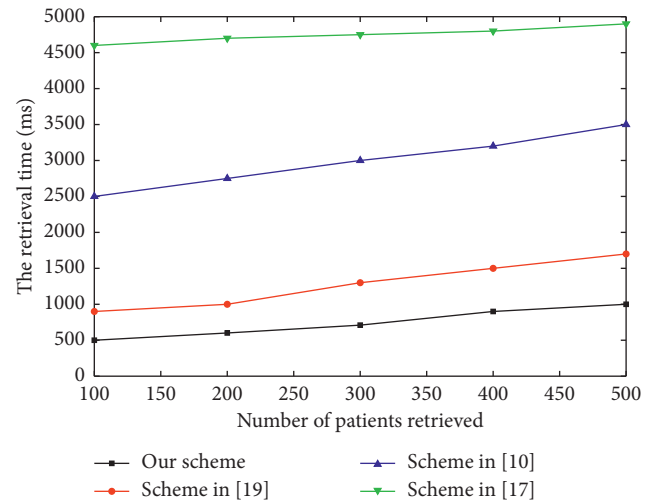


FIGURE 9: The retrieval time varies with the number of data records returned ($m = 6000, d = 0.03$).

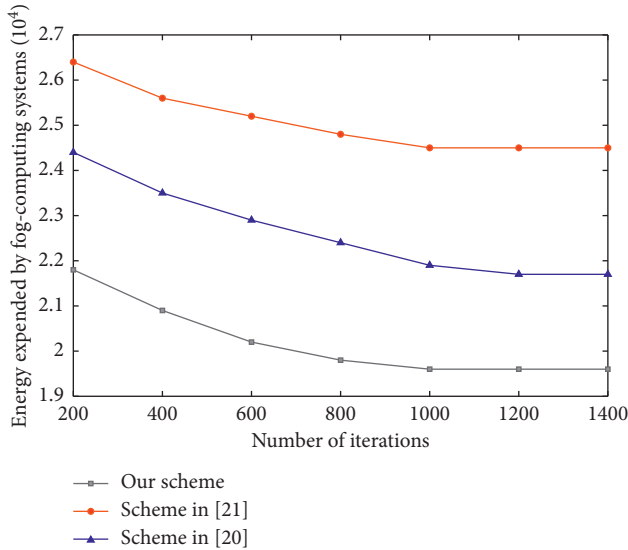


FIGURE 10: Iterative convergence of scheduling algorithm (number of orders = 600).

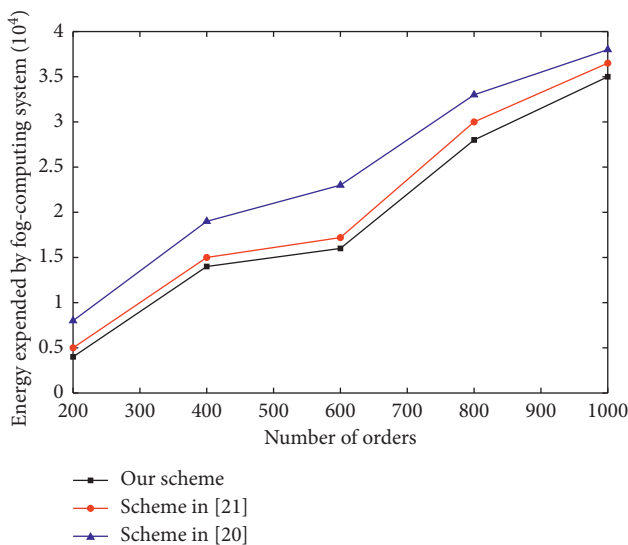


FIGURE 11: The relationship between the total cost of fog computing system and the number of orders.

this paper, the retrieval time is much less than that of other schemes, which further indicates that the retrieval efficiency of this paper is higher. To sum up, the retrieval efficiency of the scheme in this paper is superior to those from [10, 17, 19]. \square

6.3. Analysis of Energy Cost. In this paper, MATLAB 2019a is used to implement the algorithm for server allocation. We set $\beta = 10^{-8} \times (\text{MIPS})$ and $\alpha = 0.6 \times \beta (\text{J})$ in (2), the penalty value M is 0.5, and the number of server group VMs in the fog computing system is 10. The scheme in this paper compares the energy cost with PSO algorithm in [20] and artificial bee colony algorithm in [21].

Figure 10 shows that different schemes have different energy costs and convergence rates with the increase of algorithm iteration times. As seen from Figure 10, the convergence speed of their scheme [20] is slow, and the convergence energy cost is high. In [21], although the convergence energy cost is low, the convergence speed is the slowest, which reduces the efficiency of order processing. Compared with the schemes in [20, 21], our scheme has a higher convergence speed and lower convergence energy cost. The relationship between the energy cost of the fog computing system and the number of orders is shown in Figure 11. With the increase of the number of orders, the scheme in this paper produces lower energy cost than that of the schemes in [20, 21].

7. Conclusion

The traditional retrieval scheme of IoMT is faced with problems such as privacy leakage, low retrieval efficiency, and high system power cost. This paper proposes a data retrieval and analysis service scheme of IoMT under privacy protection based on low-cost fog computing. We set up a fog computing system between the IoMT and cloud services to not only improve the data retrieval efficiency, but also reduce the service delay. We adopt range tree to construct data index and form a multidimensional range tree structure. This improves the retrieval efficiency on the premise of ensuring the index security and the unlinkability of the portal. Moreover, this paper adopts an algorithm to allocate resources for the orders on the fog server group, which significantly reduces the energy cost of the system while ensuring system efficiency. The simulation results show that the proposed scheme not only improves the retrieval efficiency and accuracy, but also significantly reduces energy cost compared with the existing schemes.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request (email: 1711023984@qq.com).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (nos. 62001055 and 61802025); Beijing Natural Science Foundation (no. 4204107); Funds of “YinLing” (no. A02B01C03-201902D0); Open Project Program of National Engineering Laboratory for Agri-Product Quality Traceability; and Beijing Technology and Business University (BTBU), under grant AQT-2020-YB4.

References

- [1] N. Wang, J. Fu, B. K. Bhargava, and J. Zeng, "Efficient retrieval over documents encrypted by attributes in cloud computing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2653–2667, 2018.
- [2] H. Kim, J. Shin, Y. Song, and J. Chang, "Privacy-preserving association rule mining algorithm for encrypted data in cloud computing," in *Proceedings of the 2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pp. 487–489, Milan, Italy, July 2019.
- [3] X. Shi and S. Hu, "Fuzzy multi-keyword query on encrypted data in the cloud," in *Proceedings of the 2016 4th Intl Conference on Applied Computing and Information Technology/3rd Intl Conference on Computational Science/Intelligence and Applied Informatics/1st Intl Conference on Big Data, Cloud Computing, Data Science and Engineering (ACIT-CSII-BCD)*, pp. 419–425, Las Vegas, NV, USA, December 2016.
- [4] P. Pandiaraja and P. Vijayakumar, "Efficient multi-keyword search over encrypted data in untrusted cloud environment," in *Proceedings of the 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, pp. 251–256, Tindivanam, February 2017.
- [5] S. He, B. Cheng, H. Wang, Y. Huang, and J. Chen, "Proactive personalized services through fog-cloud computing in large-scale IoT-based healthcare application," *China Communications*, vol. 14, no. 11, pp. 1–16, 2017.
- [6] Q. Li, J. Zhao, Y. Gong, and Q. Zhang, "Energy-efficient computation offloading and resource allocation in fog computing for internet of everything," *China Communications*, vol. 16, no. 3, pp. 32–41, 2019.
- [7] H. K. Apat, B. S. Compt, K. Bhaire, and P. Maiti, "An optimal task scheduling towards minimized cost and response time in fog computing infrastructure," in *Proceedings of the 2019 International Conference on Information Technology (ICIT)*, pp. 160–165, Bhubaneswar, India, December 2019.
- [8] S. K. Datta, C. Bonnet, and J. Haerri, "Fog computing architecture to enable consumer centric internet of things services," in *Proceedings of the 2015 International Symposium on Consumer Electronics (ISCE)*, pp. 1–2, Madrid, Spain, June 2015.
- [9] K. H. Abdulkareem, M. A. Mohammed, S. S. Gunasekaran et al., "A review of fog computing and machine learning: concepts, applications, challenges, and open issues," *IEEE Access*, vol. 7, pp. 153123–153140, 2019.
- [10] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 340–352, 2016.
- [11] C. Chen, X. Zhu, P. Shen et al., "An efficient privacy-preserving ranked keyword search method," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 4, pp. 951–963, 2016.
- [12] Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 9, pp. 2546–2559, 2016.
- [13] S. Hijazi, A. Page, B. Kantarci, and T. Soyata, "Machine learning in cardiac health monitoring and decision support," *Computer*, vol. 49, no. 11, pp. 38–48, 2016.
- [14] G. Redlarski, D. Gradolewski, and A. Palkowski, "A system for heart sounds classification," *PLoS One*, vol. 9, no. 11, Article ID e112673, 2014.
- [15] D. Rizk, R. Rizk, and S. Hsu, "Applied layered-security model to IoMT," in *Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, p. 227, Shenzhen, China, July 2019.
- [16] S. K. Mishra, D. Puthal, J. J. P. C. Rodrigues, B. Sahoo, and E. Dutkiewicz, "Sustainable service allocation using a meta-heuristic technique in a fog server for industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4497–4506, 2018.
- [17] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, 2014.
- [18] M. A. Khan and F. Algarni, "A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS," *IEEE Access*, vol. 8, pp. 122259–122269, 2020.
- [19] L. Hu, S. Nooshabadi, and M. Ahmadi, "Massively parallel KD-tree construction and nearest neighbor search algorithms," in *Proceedings of the 2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2752–2755, Lisbon, Portugal, July 2015.
- [20] R. Pahlevi, M. A. Murti, and E. Susanto, "The implementation of PID using particle swarm optimization algorithm on networked control system," in *Proceedings of the 2014 International Conference on Industrial Automation, Information and Communications Technology*, pp. 35–38, Bali, Indonesia, August 2014.
- [21] Z. Zhang, W. Su, and K. Zhou, "Airborne radar sub array partitioning method based on artificial bee colony algorithm," in *Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 484–489, Chengdu, China, March 2019.

Research Article

The Sustainability of Knowledge-Sharing Behavior Based on the Theory of Planned Behavior in Q&A Social Network Community

Xin Feng ¹, Lijie Wang ², Yue Yan ³, Qi Zhang ², Liming Sun ⁴, Jiangfei Chen ⁵,
and Ye Wu ⁶

¹School of Economics and Management, Yanshan University, Qinhuangdao 066004, China

²Business School, University of International Business and Economics, Beijing 100029, China

³Huaxin College, Hebei GEO University, Shijiazhuang 050031, China

⁴School of Language and Culture, Hebei GEO University, Shijiazhuang 050031, China

⁵Miami University, Middletown 45042, OH, USA

⁶School of Journalism and Communication, Beijing Normal University, Beijing 100875, China

Correspondence should be addressed to Lijie Wang; wanglijie@uibe.edu.cn

Received 13 May 2021; Accepted 10 June 2021; Published 19 June 2021

Academic Editor: Xuzhen Zhu

Copyright © 2021 Xin Feng 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 the update and iteration of Internet technology, the socialized Q&A (question-and-answer) platform realizes the cross-border dissemination of knowledge with the main purpose of disseminating and sharing knowledge. Zhihu, as a knowledge-sharing platform that relies on the user-generated content model to maintain operation and the strong willingness of users to share knowledge, plays a key role in the development of the community. Currently, social Q&A platforms are facing problems such as low user participation rate and gradual decrease in the number of active users. It is very important and urgent to explore the factors that affect users' willingness to share knowledge. In response to this problem, this paper builds a theoretical model of the factors that are affecting users' willingness to share knowledge and uses questionnaire research methods to conduct research design and collect sample data and uses methods such as correlation analysis and structural equation modeling to verify the model and hypothesis. The research results show that the theoretical model of planned behavior has strong explanatory power and self-efficacy and material rewards have a positive effect on knowledge-sharing attitudes. Finally, according to some research results, this paper shows that, with the change of time, young people have different needs for knowledge sharing than before, while realizing self-worth through sharing experience, and we also hope to protect own interests and hope that there are more factors to encourage more users to share their knowledge and experience. Therefore, we propose that the platform can use incentive mechanisms to promote knowledge sharing while helping sharers realize their self-worth, improve the existing functions of the platform, or carry out activities to encourage users to participate, so as to achieve the purpose of knowledge sharing and maintain the operation of the Zhihu platform.

1. Introduction

1.1. Research Significance. More and more domestic and foreign researchers have begun to focus on the study of knowledge sharing; in whether traditional community or virtual community, information flow and knowledge sharing play an important role in the development of communities.

Through the analysis of Zhihu users' attitudes, motivations, and other factors affecting knowledge sharing, a knowledge community model that provides the sustainability of knowledge-sharing behaviors is constructed. Researching and summarizing the influencing factors of the willingness of community users to share knowledge can guide people to continue knowledge sharing and information exchange in

the Zhihu community and propose suggestions to promote community knowledge sharing to the socialized question-and-answer community represented by Zhihu.

1.2. Research Purpose. Whether in traditional communities or virtual communities, the flow of information and the development of knowledge communities are crucial. At present, Zhihu has become a typical representative of the knowledge community. This article uses Zhihu users as an example to study and analyze the factors that affect the knowledge-sharing willingness of users in the knowledge community. Based on the summary of domestic and foreign research, this article combines the theory of planned behavior (TPB) with Zhihu community users as the research object and discusses the influencing factors of Zhihu users' willingness to share. This research constructs a theoretical research model, defines research variables and proposes hypotheses, designs and develops measurement tools and questionnaires, conducts data analysis and verification of research hypotheses on the questionnaire survey data, and then discusses and researches the results.

1.3. Status of Domestic and Foreign Research

1.3.1. Foreign Research Trends. Research on knowledge-sharing platforms by foreign scholars started early. Harper et al. [1] believed that the knowledge-sharing platform does not limit the types of questions and the domains they belong to and allows users to ask and answer a wide range of questions without violating the law. This widens the scope of users' discussions, and the platform can contact more users with various types of questions. The platform becomes more flexible and active, and users have the possibility of reaching more knowledge. Driven by knowledge and curiosity, users will continue to be active. Choi and Yi [2] took the respondents of the Yahoo community as the research object and studied the factors that affect their knowledge-sharing behavior. It found that altruism, ideology, self-efficacy, and pleasure significantly positively affect the willingness to share knowledge. These factors belong to the intrinsic motivation of behavior, altruist for personality beliefs, and likes to help others. In terms of ideology, when people's opinions, values, etc., collide with other opinions, whether it is a good collision or a vicious collision, it is more likely to lead to the generation of knowledge-sharing behavior. Knowledge-sharing behavior is also related to the user's history of success or failure. The more success stories a person has, the higher the sense of accomplishment, the higher the degree of trust in knowledge, and the higher the probability of knowledge-sharing behavior, which is the positive impact of self-efficacy. Emotionally, a happy mood is more likely to cause people to share behaviors, not just knowledge-sharing behaviors, and a cold mood is not easy to produce knowledge-sharing behaviors and is more likely to produce talk and silent behaviors. Cronk [3] combines social capital theory with a positive emotional tone to form a new comprehensive model. The experimental results show that positive emotional tone, trust, and common vision will have

a positive impact on knowledge sharing. At the same time, the role of trust and social interaction on knowledge sharing is regulated by the positive emotional tone. The idea of trust and interaction comes from emotional judgments, and positive emotions promote the generation of trust and interactive ideas. In the theory of social capital, the one with the other in a friendly and noncompetitive cooperative relations will be more prone to knowledge sharing.

As the influence of socialized knowledge sharing gradually grows, researchers have made relevant research on the influence of knowledge-sharing behavior. Alexander and Nick [4] believed that self-interest, community culture, and trust among members are positively related to the willingness to share knowledge. When there are sufficient self-interest conditions, such as answering questions to earn reward points or accumulate personal fame, such conditions will lead some users to actively answer questions; when answering questions becomes a very common phenomenon in community platforms, users are accustomed to answering questions to each other and even indulging in the joy of answering questions, which can also be called community culture. This is undoubtedly the cultural environment conducive to the emergence of knowledge-sharing behavior. Meng et al. [5] found that self-efficacy, trust, and result expectations are the main factors affecting knowledge sharing in virtual communities; on the virtual social platform, people do not know each other. While seeking knowledge to get a certain answer, another question will follow, which is the correctness of the answer. This scenario is very common in virtual communities, and sometimes wrong answers will occur. Very bad influences, such as asking about the database addition operation, when someone inadvertently answered the database deletion operation in a joke, will cause very serious consequences and make the trust between people in the virtual community disappear. When there are certain official certifications or correct certification marks, knowing the expected results will make it easier for people to accept and leave a trust mark in the virtual community, which is conducive to long-term active knowledge-sharing behavior. According to the TPB, Hassandoust et al. research [6] on members' knowledge-sharing intention found that attitude and perceived behavior control have a significant impact on knowledge-sharing intention, while subjective norms have no significant impact on knowledge-sharing intention. Attitude drives behavior, and knowledge-sharing behavior actually occurs when members feel the need to share knowledge. When members are subjectively required to share knowledge, it is more likely to become an imposed burden or task, and only a small probability of good knowledge-sharing behavior will occur. According to the research on the influence of social network relationship and TPB on knowledge-sharing intention and behavior, Chen et al. [7] showed that subjective norms, students' attitude, self-efficacy, and social relations have a significant positive correlation with knowledge-sharing intention and indirectly influence knowledge-sharing behavior through knowledge-sharing intention. Tao [8] constructed a virtual community knowledge-sharing influencing factor model based on social cognition theory and further explored

self-efficacy, trust, perceived compatibility, and perceived relative advantages. The influence mechanism of these four factors on knowledge-sharing behavior has been found to have a significant impact on knowledge-sharing behavior. Trust positively affects self-efficacy, and there is a significant relationship between perceived relative advantage and perceived compatibility.

1.3.2. Domestic Research Trends. Chen Juan, based on the theory of self-determination, used regression analysis to study the influencing factors of the user experience. The results show that the improvement of visual attractiveness and demand satisfaction will enhance the user experience, and emotions play a moderating role in it. Vision brings the user an intuitive feeling, which is the first impression during the experience, and the comfortable visual effect improves the user experience. Inspired by the six-degree segmentation theory, Liu and Jia proposed a knowledge contribution behavior research model suitable for Zhihu users on the basis of the knowledge-sharing cross graph [9]. She believes that trust, incentive system, community culture, and results are expected to positively affect the knowledge-sharing behavior of Zhihu users. The trust is high; with the question of answering rewards, the official certification mark, and celebrities answering, these will be more popular in the community.

According to the TPB, Li Zhihong introduced social cognition theory and social exchange theory and constructed theoretical models to obtain trust, result expectation, self-efficacy, and altruism, which were significantly related to the willingness of members to share knowledge. Based on the theory of social capital, Zhang Nai and Zhou Nianxi found that social interaction in virtual communities can have a positive impact on knowledge-sharing behavior. Based on action control theory and TPB, Jiang Peizhen found that subjective norms, self-efficacy, and users' attitudes towards knowledge sharing are significantly positively correlated, and knowledge-sharing attitudes have a direct positive impact on knowledge-sharing behavior. Liu et al. [10] proceeded from the social and cultural dimensions and individual psychological dimensions and proposed the influence of personal outcome expectations, self-efficacy, relationships, and other influencing factors on the willingness to share knowledge in virtual communities. According to research, personal outcome expectations, relationships, etc. have a significant impact on the willingness to share knowledge in virtual communities, while self-efficacy has no significant impact on personal outcome expectations. Zhao Yuxiang integrated social exchange theory, technology acceptance model theory, social capital theory, etc., constructed an integrated theoretical model, and proposed three factors that affect users' use of blogs: perceived usefulness, perceived ease of use, and exchange costs.

2. The Definition of Theoretical Concepts

2.1. Theory of Planned Behavior

2.1.1. Theory of Planned Behavior's Concept. The theory of planned behavior (TPB) is derived from the Theory of Reasoned Action (TRA), which emphasizes that human

behavior is based on rational volitional control and mainly determined by attitudes and subjective norms. The TPB was put forward by Icek Ajzen. Ajzen believed that all factors that may affect behavior indirectly influence the performance of behavior through behavioral intention. Fu [11] conducted an empirical study on the influencing factors of individuals' and organizations' willingness to explore knowledge sharing and found that knowledge sharing is significantly influenced by organizational image and altruism to a large extent. In conclusion, a large number of studies have proved that the TPB is applicable to the study of knowledge sharing.

2.1.2. TPB's Research Status. Based on TPB, Ajzen [12] put forward the theory of decomposing planned behavior. They believe that there are multiple concepts in attitude, subjective norms, and perceptual behavior control, decomposing attitudes into emotional and instrumental, and dividing subjective norms into command elements and descriptive elements; perceptual behavior control includes two aspects of self-efficacy and controllability. According to the research on the TPB, Keats et al. [13] pointed out that, when predicting behavior and intention, the occurrence of individual behavior is affected by the expectation of others. Zahra and Mohammad and Chen [14, 15] studied knowledge-sharing behavior using TRA or TPB theoretical model, and the research showed that knowledge-sharing intention has a direct influence on knowledge-sharing behavior. Bello and Oyekunle [16] studied that knowledge-sharing behavior is influenced by attitude, intention, and motivation by using the TPB. Knowledge-sharing attitude is significantly related to intention, and the knowledge-sharing intention is also significantly related to behavior. Jin et al. [17] used TRA to study the process of knowledge sharing in the communication industry and found that perceived knowledge ownership and material incentives have a positive impact on individuals' willingness and behavior of knowledge sharing. Based on the TPB, Zhao [18] integrated the key variables of expectation confirmation theory, social cognition theory, and social capital theory, extended the original variables of planned behavior theory, and constructed a new theoretical model of planned behavior. Based on this theoretical model, this study will build a new theoretical model to study the sustainability of community users' knowledge sharing.

2.2. Knowledge Sharing

2.2.1. The Concept of Knowledge Sharing. Knowledge sharing is the most complicated link in all aspects of knowledge management, and it is also the key step for successful knowledge management. Knowledge sharing is also a kind of dissemination behavior; from the contribution and flow of knowledge to the adoption, digestion, and absorption of knowledge, people can acquire knowledge from others through this behavior. The concept of knowledge sharing is not entirely uniform. Pilerot [19] put forward that knowledge sharing can be explained and understood by words such as transfer and giving. Tong et al. [20] defines

knowledge sharing as the process of exchanging tacit knowledge and creating new knowledge among friends, families, organizations, or communities. Hao et al. [21], a domestic scholar, pointed out that easiness is the basis of knowledge-sharing behavior, and it is the marginal increasing utility of knowledge that promotes the emergence of the knowledge-sharing phenomenon.

2.2.2. Influencing Factors of Knowledge Sharing. The research results of Chan et al. [22] found that, in virtual communities, users' sense of community, helpfulness, and image finally promoted members' knowledge sharing. In the degree of knowledge sharing, time and interest in discussion topics played a moderating role. Koh and Kim [23] found that, by encouraging knowledge sharing among virtual community users, virtual community providers could enhance community activity and participation, enhance the business value of the community, and finally establish the loyalty of virtual community members to virtual community providers. Based on social exchange theory, Yang [24] verified that reciprocity and trust had positive effects on knowledge-sharing behavior by using regression analysis method. Based on rational behavior theory, Yang [25], combined with incentive theory and social exchange theory as a supplement, studied the mechanism of knowledge-sharing willingness of users who have used social Q&A websites and found that altruism and reciprocity have a positive effect on knowledge-sharing attitude; sharing attitude and self-efficacy have a significant effect on knowledge-sharing willingness, while subjective norms have not shown a significant effect on willingness in this research. Trust also has a significant positive effect on sharing attitude and willingness.

3. Construction of Theoretical Model and Research Hypothesis

3.1. Construction of Theoretical Model. In the research of behavior prediction, many scholars and experts believe that behavioral intention is closer to behavior than attitude, belief, and feeling. Therefore, understanding an individual's intention for a specific behavior is an important prerequisite for predicting whether an individual will perform the behavior. Fishbein and Ajzen [26] believe that attitude, subjective norms, and cognition of behavior control are important psychological factors that directly affect behavioral intention. On this basis, Zhang [27] put forward four influencing factors of knowledge sharing among individuals in mobile Internet, which are expected organizational reward, reciprocal benefit, knowledge self-efficacy, and altruistic beliefs, and established the following model (Figure 1).

Based on Figure 1, influencing factors of group recognition and material reward are added which are shown in Figure 2. Wang et al. [28] believe that group recognition is an important factor for users to build trust and increase influence in social networks. According to the social exchange theory of Blau and Li [29], all human behaviors are

dominated by a certain exchange activity that can bring rewards and remunerations; therefore, all social activities of human beings can be summed up as an exchange. At present, China's knowledge-based communities often use material rewards to control the behavior of this kind of spontaneous knowledge sharing.

This research is guided by the TPB, constructs a theoretical model of Zhihu users' sharing willingness, studies the influence of attitude and perceived behavior control variables on behavior willingness, and studies the characteristics of Zhihu community users. Introduce reciprocity, group recognition, self-efficacy, altruistic beliefs, and material rewards, and study the influence of these variables on the knowledge-sharing willingness.

3.2. Research Hypothesis Establishment.

- (1) The influence of reciprocity on knowledge-sharing attitude. Reciprocity as a benefit also inspires responsibility and trust among individuals. Studies as early as the 1990s have found that reciprocity is a significant motivator for knowledge sharing [30]. A large number of subsequent studies have also demonstrated the impact of reciprocal benefits on knowledge sharing. For example, Bock and Kim [31] elaborated the impact of expected reciprocal relationship on knowledge-sharing attitude. Through empirical research, Lin [32] found that reciprocal benefits significantly affected employees' attitude and intention of knowledge-sharing behavior. Therefore, this paper proposes the following hypotheses:

H1: reciprocity positively influences knowledge-sharing attitude.

- (2) The influence of group recognition on knowledge-sharing attitude. Group recognition is an important factor that affects trust and social influence in social networks. Group recognition is a measure of a person's degree of social recognition. When group recognition is low, it means that the person's social credibility is low. When group recognition is high, it means that this person is recognized by the majority of people in society and also recognizes the knowledge shared by him.

H2: group recognition has a positive effect on knowledge-sharing willingness.

- (3) The influence of self-efficacy on willingness to share knowledge. Self-efficacy will affect individuals' intention and cognitive ability to do something. Users' subjective judgment formed after knowledge sharing is their self-assessment of their ability to provide valuable knowledge to other users. The perception of self-efficacy plays an important role in influencing people's motivation and behavior. Runhaar and Sanders [33] showed that self-efficacy had a significant predictive effect on knowledge-sharing behavior, and people with high self-efficacy had higher

willingness and behavior for knowledge sharing. The research of Hsu et al. [34] found that self-efficacy, as intrinsic motivation, can directly or indirectly affect the willingness to share knowledge in virtual communities. Individuals with high self-efficacy have stronger willingness to share knowledge. Therefore, the following hypothesis is proposed:

H3: self-efficacy has a positive influence on willingness to share knowledge.

- (4) The influence of altruistic beliefs on knowledge-sharing willingness. In social question-and-answer communities, a large number of people answer questions for other users as contributors, choosing to ask questions and then expounding their own views. Gan et al. [35] found that altruistic beliefs can promote users' knowledge-sharing behavior in the question-and-answer community. The typical network altruistic behavior is that the respondent, as the contributor, answers the questions of others, which needs to meet the needs of the recipient when the specific value is met. Therefore, this paper proposes the following hypotheses:

H4: altruistic beliefs have a positive effect on knowledge-sharing willingness.

- (5) The influence of material rewards on knowledge-sharing willingness. To complete a high-quality article, users need to spend a lot of time editing text or images or even models. Knowledge sharing requires efforts, and if efforts are proportional to returns, users' willingness to share knowledge can be promoted. High-quality knowledge sharing requires early accumulation and investment of a lot of time, energy, and money. Therefore, the following assumptions are listed:

H5: material rewards have a positive effect on knowledge-sharing willingness.

3.3. Questionnaire Design

3.3.1. Questionnaire Design Instructions. In this research, Likert five-level scale was used to measure variables. The questionnaire was composed of three parts:

The first part is the introduction, which clarifies the purpose of this research.

The second part is the basic personal information, which mainly includes the time, age, gender, and education level of the respondents on Zhihu.

The third part is the measurement question of the variables in this research. The respondents are required to give a grade of 1–5 based on their personal experience, and the corresponding levels are *very disapproving*, *disapproving*, *indifferent*, *approving*, and *very approving*. Table 1 shows the questionnaire variable design, and the questionnaire item design is shown in Table 2.

3.3.2. Questionnaire Item Design. Questionnaire item design is provided in Table 2.

4. Data Collection and Analysis

4.1. Sample Profile and Descriptive Analysis. In this research, the data were collected by questionnaire, and the survey methods used in this questionnaire are all distributed in the form of network questionnaire. It is mainly distributed to the respondents in the form of links and two-dimensional codes. The regulations for the respondents are user of Zhihu. Specifically, since we defined the specific users of the questionnaire survey as college students, the age of the users can be confirmed as the young user group. At the same time, in order to ensure the randomness of the questionnaire, we randomly distributed questionnaires on major social platforms to call on more young user groups to participate in the survey. Among them, we linked the questionnaire to some Zhihu users in Zhihu community randomly and randomly selected and distributed questionnaires in multiple groups on Sina Weibo, QQ, and WeChat. A total of 280 questionnaires were distributed, and 271 valid questionnaires were collected. The specific questionnaire survey results are shown in Table 3.

Based on the descriptive statistics of the above survey results, it can be concluded that the survey objects have the following characteristics:

- (1) Service life of Zhihu: 18.5% of the respondents come to Zhihu for 1–6 months, 37.6% for 1-2 years, 26.2% for 3-4 years, and 17.7% for more than 5 years, which ensures that the respondents are all users who have used Zhihu.
- (2) Gender characteristics: the sample size of boys is 106, accounting for 39.1% of the total sample. The sample size of female students is 165, accounting for 60.9% of the total number of samples. Relatively speaking, females account for a large proportion.
- (3) Age characteristics: most of the respondents are 21–30 years old, accounting for 74.9%, mainly because most of the respondents are school students, so the age structure is more consistent.
- (4) Education level: the target of the questionnaire is mainly college students, so the sample size of undergraduates is 187, accounting for 69.0%, those with junior college or below is 11.1%, those with master's degree account for 12.9%, and those with doctor's degree account for 7.0% of the total sample size.

4.2. Reliability and Validity Analysis of the Questionnaire. Firstly, this paper uses SPSS 25.0 to analyze the reliability of the survey data. Secondly, Excel, SPSS, and AMOS software were used to conduct validity analysis of the questionnaire and scale to verify the validity of the scale contents. Finally, the goodness of fit of the model was verified by confirmatory factor analysis.

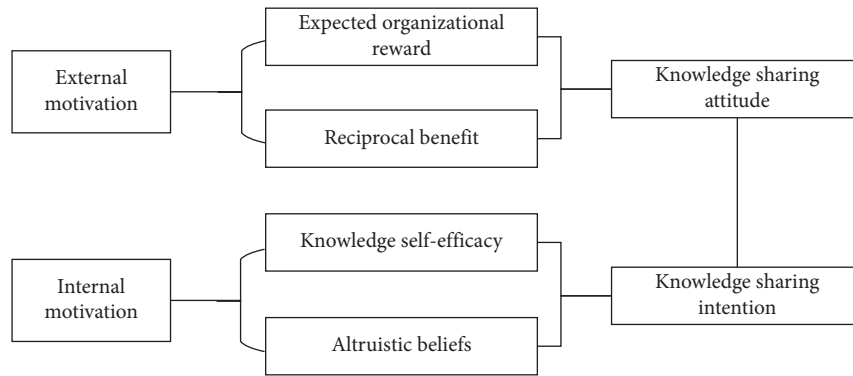


FIGURE 1: A theoretical model of influencing factors of individual knowledge sharing in the mobile Internet environment.

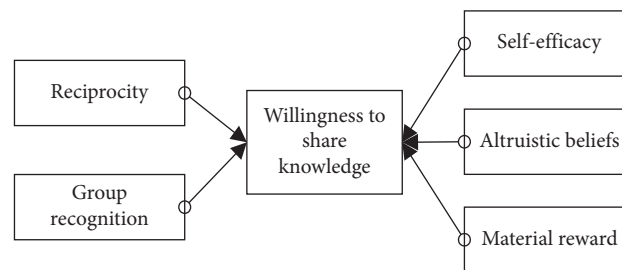


FIGURE 2: Theoretical model of Zhihu users' sharing willingness.

TABLE 1: Questionnaire variable design.

Variable	Variable definition	Reference
Reciprocity	Respondents exchange knowledge contribution behaviors in the community	Shen et al. [36]
Group recognition	The trust established by users in the social network and the satisfaction after sharing knowledge promote sharing behavior	Wang et al. [28]
Self-efficacy	Personal confidence in their ability to provide knowledge	Hao et al. [21]
Altruistic beliefs	Contributors choose to answer the questioner's questions	Gan et al. [35]
Material rewards	An incentive to influential users for their efforts	Atreyi Kankanhalli et al.
Knowledge-sharing willingness	An individual's judgment of the subjective probability of taking knowledge-sharing behavior, which reflects the individual's willingness for a specific behavior	Bock and Kim [31]

4.2.1. Reliability Analysis. Reliability analysis is mainly used to test the stability or consistency of the analysis results, that is, through multiple analysis of the questionnaire data, to observe whether the analysis results are consistent, in order to determine the authenticity or reliability of the empirical research results. Generally speaking, reliability tests are conducted in different time periods, among different respondents and raters. The higher the consistency of the final results is, the higher the reliability of the questionnaire is. Reliability can be divided into retest reliability, duplicate reliability, internal consistency reliability, and rater reliability. In this study, Cronbach's alpha reliability measurement method, which is widely used in previous studies and has a high degree of recognition, is adopted to conduct the reliability analysis of the questionnaire. Cronbach's alpha can fully reflect the implementation of the final scores of all grid question indicators in the scale. It can be seen from

Tables 4 and 5 that the overall reliability of these variables is 0.938 and the Cronbach- α coefficient of each variable is higher than the minimum standard 0.6, indicating that the scale used in this study has good reliability and internal consistency, and the design is reasonable.

4.2.2. Validity Test. Validity refers to the degree to which the measurement tool can accurately measure the required measurement characteristics, mainly including content validity and construction validity. Content validity means that the content of the scale is appropriate and representative, and the item distribution is reasonable, which can reflect the characteristics of the measurement indicators. For the content validity, this study mainly refers to the questionnaire items designed by some published articles to modify and invited teachers to evaluate the designed

TABLE 2: Questionnaire item design.

Variable	Item	Reference
<i>Reciprocity</i>	While answering other people's questions, I hope they can also answer my questions When I share knowledge, I hope to get knowledge from the platform	Shen et al. [36]
<i>Group recognition</i>	I think sharing knowledge on the platform can be recognized by others A user's trust in me is the recognition of me Sharing useful knowledge will increase others' recognition of me	Wang et al. [28]
<i>Self-efficacy</i>	I think I have the ability to share knowledge in the Zhihu community I think I have the ability to share knowledge in Zhihu community I feel confident that I could provide something that the rest of the community found valuable The knowledge I share will be helpful to other users in the community	Hao et al. [21]
<i>Altruistic beliefs</i>	I want to share my knowledge, experience, insight or skills with others I answer questions to help people in need	Gan et al. [35]
<i>Material rewards</i>	I hope to get substantial rewards after sharing knowledge After sharing knowledge, I think it is very important to get the corresponding commission After sharing knowledge, I think the user's subscription is very important I hope that sharing valuable knowledge can be rewarded I hope that the shared knowledge will be rewarded when it is useful to others A certain amount of pay and influence in the community can encourage me to share knowledge	Atreyi Kankanhalli et al.
<i>Knowledge-sharing willingness</i>	I would like to share knowledge and information more frequently on the Zhihu community I am willing to often share my experience on the Zhihu community I am willing to share my knowledge more frequently with other users I am willing to comment and reply other users' answers I would like to use the community's reply, comment and other features regularly	Bock and Kim [31]

TABLE 3: Descriptive statistics of knowledge-sharing willingness survey.

Project	Category	Frequency	Percentage
<i>Time</i>	1-6 months	50	18.5
	1-2 years	102	37.6
	3-4 years	71	26.2
	5 years and above	48	17.7
<i>Gender</i>	Male	106	39.1
	Female	165	60.9
<i>Age</i>	0-20 years old	39	14.4
	21-30 years old	203	74.9
	31-40 years old	25	9.2
	Over 40 years old	4	1.5
<i>Education level</i>	Junior college or below	30	11.1
	Bachelor degree	187	69.0
	Master's degree	35	12.9
	Doctor's degree	11	7.0
Total		271	100.0

questionnaire and give relevant suggestions. The construction validity refers to the degree to which the actual test results explain the measurement index. As for the construction validity, this study will mainly test the construction validity of the questionnaire scale through factor analysis. Before factor analysis, KMO measure and Bartlett sphere test should be used to verify the validity of the questionnaire. Cerny and Kaiser [37] believed that studies showed that when KMO value was between 0.6 and 1, it was suitable for factor analysis, and when KMO value was below 0.6, it was

not suitable for factor analysis. In this paper, the factor analysis method was used to test the validity of the measurement model and scale and obtain the Bartlett sphere test and the KMO calculation result table. It can be seen from Table 6 that the significance level of the Chi-square value of Bartlett spherical test is 0.000 and KMO is 0.931, indicating that the scale has good validity and is suitable for factor analysis.

As can be seen from Table 7, the factor loading is about 0.5, indicating that the factor convergence effect is good.

TABLE 4: Overall reliability analysis data.

Cronbach's alpha variable	Cronbach's alpha based on standardized items	Number of items
0.938	0.939	20

TABLE 5: Reliability analysis of each variable.

Variable	Number of items	Cronbach α 's coefficient
Material rewards	6	0.895
Knowledge-sharing willingness	4	0.817
Group recognition	3	0.797
Self-efficacy	3	0.820
Reciprocity	2	0.786
Altruistic beliefs	2	0.690

TABLE 6: KMO and Bartlett test.

KMO sampling appropriateness number	0.931
Bartlett sphericity test	Approximate Chi-square Degree of freedom Significance
	3067.643 190 0.000

TABLE 7: Factor matrix.

	Composition					
	1	2	3	4	5	6
A01	0.772					
A02	0.763					
A03	0.753					
A04	0.748					
A05	0.743					
A06	0.572					
A07		0.730				
A08		0.704				
A09		0.594				
A10		0.562				
A11			0.746			
A12			0.717			
A13			0.552			
A14				0.800		
A15				0.754		
A16				0.533		
A17					0.840	
A18					0.776	
A19						0.638
A20						0.625

Combined with the theoretical structure and factor analysis results, six factors were finally extracted.

4.3. Experimental Hypothesis Validation Analysis

4.3.1. Simulation Fitting Degree Analysis. This study mainly uses AMOS 26 to process structural equation model analysis. After modeling by AMOS, the measurement model of the structural equation model in this study included 6 latent variables, which were represented by ellipses. A total of 20 observed variables were represented by a square. There are

21 residuals, which were represented by circles. The final model is shown in Figure 3.

Table 8 shows the data of the overall fit degree of the model, which is analyzed from two aspects: absolute fit degree and value-added fit degree. It can be seen from the data in the table that the indexes of the two aspects of this model meet the evaluation criteria and the overall fitting degree is good. The structural equation model agrees well with the original data of the sample and meets the standard, indicating that the research model in Figure 3 can evaluate the research problem of influencing factors of Zhihu users' willingness to share knowledge.

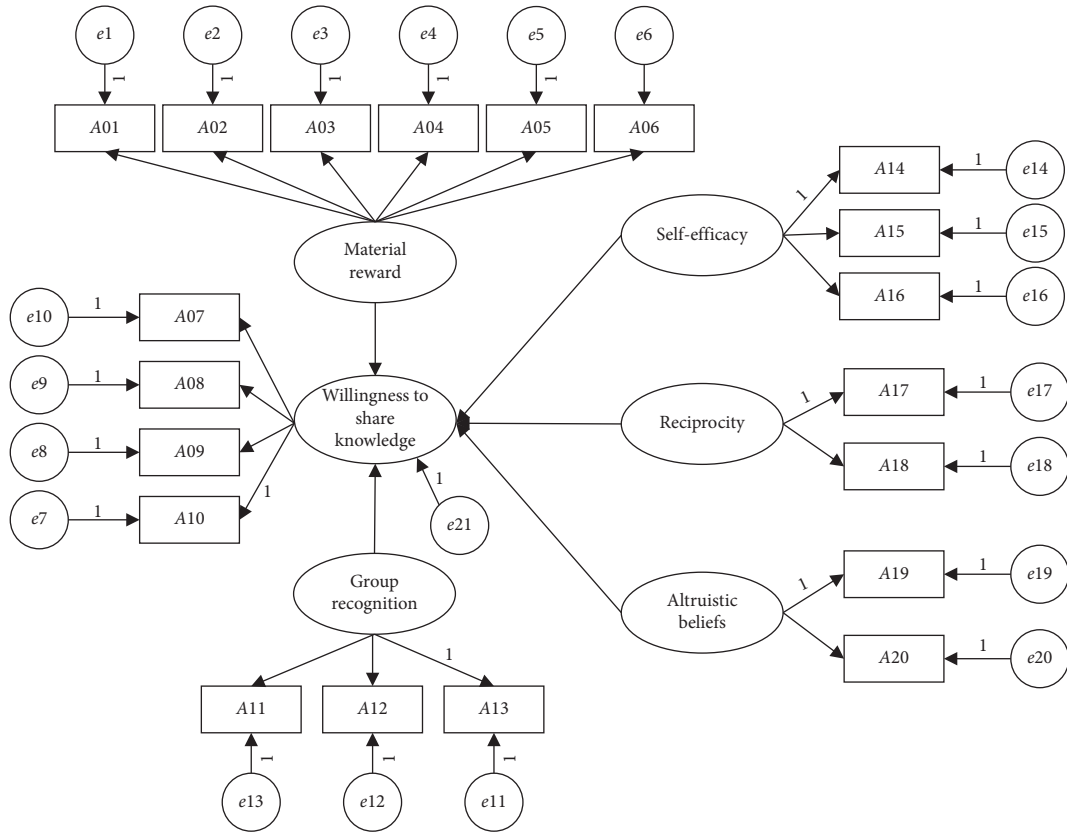


FIGURE 3: Structural equation model.

TABLE 8: Model fit degree.

	Statistical test volume	Name	Adaptation standard	Test result data	Model adaptation judgment
<i>Absolute adaptation index</i>	$\chi^2/d.f.$	Chi-square/degree of freedom value	<3	1.733	Yes
	RMR	Residual root mean square	<0.05	0.025	Yes
	RMSEA	Approximate residual root mean square	<0.08	0.052	Yes
	GFI	Benign adaptation index	>0.9	0.913	Yes
	AGFI	Adjusted benign adaptation index	>0.9	0.877	No
<i>Value-added adaptation index</i>	NFI	Rule adaptation index	>0.9	0.918	Yes
	RFI	Relative adaptation index	>0.9	0.896	No
	IFI	Value-added adaptation index	>0.9	0.964	Yes
	CFI	Comparative adaptation index	>0.9	0.963	Yes

Wen et al. research [38] shows that as long as the model is a good fit according to several criteria (including Chi-square criterion), the model can be considered acceptable from some perspectives on the premise that other indexes should also be referred to and cannot be too far from the boundary value. Therefore, it can be considered that the fitting degree of this model is acceptable. At the same time, the maximum likelihood estimation method is used to estimate the parameters of the model, and the path analysis of the structural equation model is used to calculate the standardized path coefficients among the potential variables. In addition, the standardized path regression coefficients

among model variables are studied in this paper to more intuitively explain the results of hypothesis verification of this research (Figure 4).

4.3.2. Hypothesis Test Result. It can be seen from Tables 8 and 9 that the hypotheses proposed in this paper are partially valid. The experimental results of the knowledge-sharing willingness research model constructed in this paper show that the path coefficients of self-efficacy and material rewards are positive and the significant p value is high, so they have a positive impact on the knowledge-sharing willingness, while

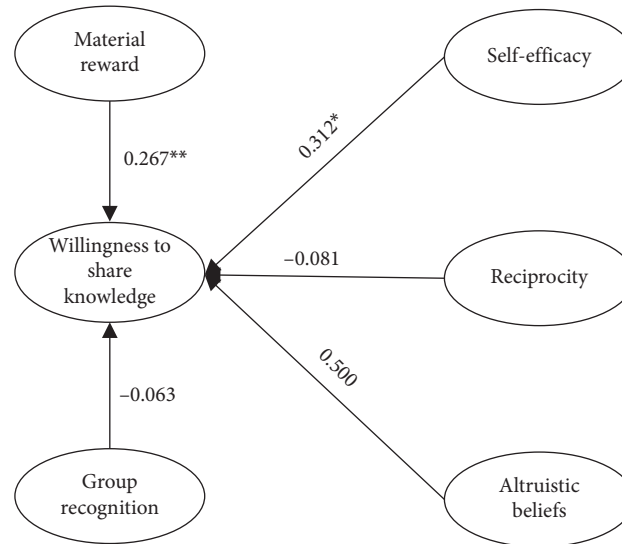


FIGURE 4: Structural equation model analysis results. Note: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

TABLE 9: Assumption verification result.

Hypothesis	Verification path	Path coefficient	p value	Validation results
H1	Reciprocity \rightarrow knowledge-sharing willingness	-0.081	0.374	Invalid
H2	Group recognition \rightarrow knowledge-sharing willingness	-0.063	0.810	Invalid
H3	Self-efficacy \rightarrow knowledge-sharing willingness	0.312	0.050	Valid
H4	Altruistic belief \rightarrow knowledge-sharing willingness	0.500	0.182	Invalid
H5	Material reward \rightarrow knowledge-sharing willingness	0.267	0.005	Valid

reciprocity and group recognition have no positive effect on knowledge-sharing willingness. According to the data, although altruistic beliefs have an impact on the willingness to share knowledge, the significance and p value are not high, which does not indicate that it has a positive effect on the willingness to share knowledge.

5. Conclusions

Based on the literature research and the basic situation of Zhihu's use, this study proposes the influencing factor model of Zhihu users' willingness to share knowledge based on the theory of planned behavior. The knowledge-sharing willingness of Zhihu users was studied through multiple factors, 271 pieces valid data from Zhihu users were collected by using questionnaires, and the research hypotheses were verified by using the structural equation model to verify the influence of reciprocity, social capital, group recognition, self-efficacy, and altruistic beliefs on the knowledge-sharing willingness. The research results show that self-efficacy and material rewards have a positive effect on the knowledge-sharing willingness, altruistic beliefs are not significant, and reciprocity and group recognition have negative effects.

Since the respondents are all representative, young people between the ages of 20 and 30 occupy most of the power of interpretation in this study. Because the respondents are mainly randomly selected college students, the accuracy of the proportion of young people in the age characteristics of the questionnaire data can be determined,

which covers the main user groups of Zhihu users. Compared with the continuous development of Zhihu users and the continuous penetration of new and old users, the sample data in this study is more representative of the knowledge-sharing willingness of some young users. Data analysis results show that most young people at present are more inclined to the positive impact of self-efficacy and material rewards, which shows that many young people in society now hope to get some opportunities to prove their social value and are good at sharing their own experiences. However, because, in this era of information sharing, a lot of knowledge has been transported and cannot reflect the value of the sharer, young people now also hope more to get some external rewards in return for their efforts. This is because high-quality knowledge sharing requires the sharers to accumulate in the early stage and even invest a lot of time and energy, and material rewards can compensate and motivate the sharers. In this regard, today's knowledge community platforms must not only give sharers the opportunity to show their self-worth, but also learn to use appropriate rewards and incentives for users to share knowledge in order to maintain the operation of the community platform.

The three influencing factors of altruistic beliefs, reciprocity, and group recognition do not have a positive effect on the knowledge-sharing willingness, but relevant literature has confirmed that these three influencing factors have a positive effect on the knowledge-sharing willingness. Research shows that the data in this questionnaire is not rigorous enough or that Zhihu's new and old users are

constantly changing. Users of different age groups have different perceptions of Zhihu. The research results only represent the behavioral willingness of the group we studied. Therefore, the impact of the three influencing factors of altruistic beliefs, reciprocity, and group recognition on the knowledge-sharing willingness, as well as the knowledge-sharing behavior of different groups, remains to be investigated.

Data Availability

The data presented in this study are available on request from the corresponding author.

Conflicts of Interest

All authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors acknowledge the financial support from the National Natural Science Foundation of China (no. 11905042), 333 Funded Project of “333 Talent Project” in Hebei Province (no. A202001015), and Key Project of Humanities and Social Sciences Research for the Colleges and Universities of Hebei Province (no. SD2021017).

References

- [1] F. M. Harper, D. Raban, and S. Rafaei, “Predictors of answer quality in online Q&A sites,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 865–874, ACM, Florence, Italy, April 2008.
- [2] N. Choi and K. Yi, “Raising the general public’s awareness and adoption of open source software through social Q&A interactions,” *Online Information Review*, vol. 39, no. 1, pp. 119–139, 2015.
- [3] R. Cronk, “Knowledge sharing in the online environment: emotional intelligence, social capital, and intellectual capital relationships,” in *Proceedings of the European Conference On Knowledge Management*, pp. 232–239, Barcelona, Spain, September 2017.
- [4] S. Alexander and B. Nick, “Negotiate, reciprocate, or cooperate? The impact of exchange modes on inter-employee knowledge sharing,” *Journal of Knowledge Management*, vol. 20, no. 4, pp. 687–712, 2016.
- [5] H. S. Meng, L. Teresa, and M. C. Chun, “Knowledge sharing in virtual communities: the relationship between trust, self-efficacy and outcome expectations,” *International Journal of Human-Computer Studies*, vol. 65, no. 2, pp. 153–169, 2007.
- [6] F. Hassandoust, M. F. Kazerouni, and V. Perumal, “Socio-behavioral factors in virtual knowledge sharing,” *International Journal of Knowledge-Based Organizations*, vol. 2, no. 2, pp. 40–53, 2012.
- [7] Y. L. Chen, N. S. Chen, and Kinshuk, “Examining the factors influencing participants’ knowledge sharing behavior in virtual learning communities,” *Journal of Educational Technology & Society*, vol. 12, no. 1, pp. 134–148, 2009.
- [8] Z. Tao, “Explaining virtual community user knowledge sharing based on social cognitive theory,” *Wireless Communications Networking & Mobile Computing-WiCOM*, vol. 31, pp. 9538–9541, 2008.
- [9] Y. F. Liu and F. F. Jia, “Research on user knowledge-sharing behavior based on SNS,” *Information Science*, vol. 35, no. 1, pp. 41–46, 2017.
- [10] R. Liu, P. Tian, and W. J. Wang, “An empirical study on influencing factors of knowledge-sharing in virtual communities in Chinese cultural context,” *Information Science*, vol. 30, no. 6, pp. 866–872, 2012.
- [11] Z. R. Fu, *Research On Knowledge Sharing Attitudes Of Information Personnel in Organizations-Multilevel Analysis Model Method*, Central University, Taiwan, Taoyuan, 2005.
- [12] I. Ajzen, “Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior,” *Journal of Applied Social Psychology*, vol. 32, no. 4, pp. 665–683, 2002.
- [13] M. R. Keats, S. N. Culos-Reed, K. S. Courneya, and M. McBride, “Understanding physical activity in adolescent cancer survivors: an application of the theory of planned behavior,” *Psycho-Oncology*, vol. 16, no. 5, pp. 448–457, 2007.
- [14] T. Zahra and M. Mohammad, “Knowledge sharing behaviour and its predictors,” *Industrial Management & Data Systems*, vol. 110, no. 4, pp. 611–631, 2010.
- [15] C.-C. Chen, “Factors affecting high school teachers’ knowledge-sharing behaviors,” *Social Behavior and Personality: An International Journal*, vol. 39, no. 7, pp. 993–1008, 2011.
- [16] O. W. Bello and R. A. Oyekunle, “Attitude, perceptions, motivation towards knowledge sharing: views from universities in kwara state,” *African Journal of Library*, vol. 24, no. 2, pp. 123–134, 2014.
- [17] H. Jin, Z. Yang, and F. Feng, “Research on material incentives, knowledge ownership and organizational knowledge sharing,” *Science Research*, vol. 29, no. 7, pp. 1036–1045, 2011.
- [18] W. J. Zhao, “Research on Sustainable Behavior of Knowledge Sharing in Virtual Community,” *Central China Normal University*, Hubei, China, 2012.
- [19] O. Pilerot, “Information sharing in the field of design research,” *Information Research: An International Electronic Journal*, vol. 20, no. 1, pp. 26–45, 2015.
- [20] C. Tong, W. I. W. Tak, and A. Wong, “The impact of knowledge sharing on the relationship between organizational culture and job satisfaction: the perception of information communication and technology (ICT) practitioners in Hong Kong,” *International Journal of Human Resource Studies*, vol. 5, no. 1, pp. 19–47, 2015.
- [21] Q. Hao, C. Jin, and K. Wei, “The influence mechanism of knowledge-sharing behavior of virtual team members: the perspective of personal and environment interaction,” *Technological Progress and Countermeasures*, vol. 36, no. 7, pp. 138–144, 2019.
- [22] C. M. L. Chan, M. Bhandar, H.-C. Chan, and L.-B. Oh, “Recognition and participation in a virtual community,” in *Proceedings of the 37th Annual Hawaii International Conference On System Sciences*, p. 10, Big Island, HI, USA, February 2004.
- [23] J. Koh and Y.-G. Kim, “Sense of virtual community: a conceptual framework and empirical validation,” *International Journal of Electronic Commerce*, vol. 8, no. 2, pp. 75–94, 2003.
- [24] Z. B. Yang, “Research on influencing factors and interaction of knowledge-sharing in socialized question-and-answer website,” *Management Case Studies and Reviews*, vol. 9, no. 3, pp. 212–223, 2016.
- [25] H. J. Yang, “An empirical study on the influencing factors of users’ willingness to contribute to social Q&A,” *Library Science Research*, vol. 14, pp. 29–38, 2014.

- [26] M. A. Fishbein and I. Ajzen, *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*, Reading, MA: Addison-Wesley, Boston, MA, USA, 1975.
- [27] Z. Zhang, *An Empirical Study on the Influencing Factors of Individual Knowledge-Sharing in the Mobile Internet Environment*, University of Science and Technology of China, Hefei, China, 2015.
- [28] X. L. Wang, Q. Y. Zhang, and Q. Liu, "Path analysis of the influence of knowledge-sharing and knowledge innovation on organizational performance," *Journal of Inner Mongolia University of Technology (Natural Science Edition)*, vol. 33, no. 4, pp. 315–319, 2014.
- [29] P. M. Blau and G. W. Li, *Exchange and Power in Social Life*, Commercial Press, Beijing, China, 2008.
- [30] T. Connolly and B. K. Thorn, *Discretionary Databases: Theory, Date and Implications*, Sage Publications, New York, NY, USA, 1990.
- [31] G. W. Bock and Y.-G. Kim, "Breaking the myths of rewards," *Information Resources Management Journal*, vol. 15, no. 2, pp. 14–21, 2002.
- [32] H.-F. Lin, "Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions," *Journal of Information Science*, vol. 33, no. 2, pp. 135–149, 2007.
- [33] P. Runhaar and K. Sanders, "Promoting teachers' knowledge sharing: the fostering roles of occupational self-efficacy and human resources management," *Educational Management Administration & Leadership*, vol. 27, no. 5, pp. 1–20, 2014.
- [34] M.-H. Hsu, T. L. Ju, C.-H. Yen, and C.-M. Chang, "Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations," *International Journal of Human-Computer Studies*, vol. 65, no. 2, pp. 153–169, 2007.
- [35] C. M. Gan, W. J. Wang, and P. Tian, "Research on the psychological incentives of knowledge exchange and sharing in academic blogs," *Journal of Library Science in China*, vol. 38, no. 3, pp. 91–99, 2012.
- [36] Y. F. Shen, B. Liao, and Y. Xu, "Research on the influence of reputation system on knowledge sharing activities in social Q&A communities," *Journal of Information*, vol. 37, no. 11, pp. 1154–1163, 2018.
- [37] B. A. Cerny and H. F. Kaiser, "A study of a measure of sampling adequacy for factor-analytic correlation matrices," *Multivariate Behavioral Research*, vol. 12, no. 1, pp. 43–47, 1977.
- [38] Z. L. Wen, J. T. Hou, and H. Marsh, "Structural equation model testing: fitting index and chi-square criterion," *Psychological News*, vol. 2, pp. 186–194, 2004.

Review Article

Topic Detection and Tracking Techniques on Twitter: A Systematic Review

Meysam Asgari-Chenaghlu ¹, Mohammad-Reza Feizi-Derakhshi ², Leili Farzinvasht ³,
Mohammad-Ali Balafar ³ and Cina Motamed⁴

¹Department of Computer Engineering, University of Tabriz, Tabriz, Iran

²Computerized Intelligence Systems Laboratory, Department of Computer Engineering, University of Tabriz, Tabriz, Iran

³Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

⁴Department of Computer Science, University of Orléans, Orléans, France

Correspondence should be addressed to Mohammad-Reza Feizi-Derakhshi; mfeizi@tabrizu.ac.ir

Received 14 February 2021; Revised 9 May 2021; Accepted 8 June 2021; Published 18 June 2021

Academic Editor: Fei Xiong

Copyright © 2021 Meysam Asgari-Chenaghlu 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.

Social networks are real-time platforms formed by users involving conversations and interactions. This phenomenon of the new information era results in a very huge amount of data in different forms and modalities such as text, images, videos, and voice. The data with such characteristics are also known as big data with 5-V properties and in some cases are also referred to as social big data. To find useful information from such valuable data, many researchers tried to address different aspects of it for different modalities. In the case of text, NLP researchers conducted many research studies and scientific works to extract valuable information such as topics. Many enlightening works on different platforms of social media, like Twitter, tried to address the problem of finding important topics from different aspects and utilized it to propose solutions for diverse use cases. The importance of Twitter in this scope lies in its content and the behavior of its users. For example, it is also known as first-hand news reporting social media which has been a news reporting and informing platform even for political influencers or catastrophic news reporting. In this review article, we cover more than 50 research articles in the scope of topic detection from Twitter. We also address deep learning-based methods.

1. Introduction

Topic detection and tracking, which is also called TDT, is techniques and methods used for detecting news or document related topics best fitting their relevant intellectual material and also tracking these events or detected topics through dedicated media. Topic detection is a summarization problem that must fulfill certain demands. Topic as a summarized tag-set of an input document is different from an event which in most cases is a real-world phenomenon with certain spatial and temporal properties [1, 2]. This tiny difference between a topic and an event becomes more clear when talking about social networks. Identification of ongoing events on media can be expressed as *detection* while tracking of these events and storyboarding is *tracking*. This

so called media can be a single document, group of multiple documents, or even a social media like Twitter. Topic detection and tracking has been widely applied to documents, offline corpus, and newswire, including a pilot study running from 1996 till 1997 and sponsored by DARPA [3].

Social media services like *Twitter*, *Facebook*, *Google+*, and *LinkedIn* play an important role in information exchange. In case of *Twitter*, the data exchange metrics predict that 7,454 tweets are sent per second which are about 644,025,600 tweets per day [4]. This metric for 2013 was reported by Twitter officials to be more than 500,000,000 per day [5]. Importance of this large amount of data that has large variety of topics which users tend to talk about comes to light when researchers revealed that users are most likely to talk about real-world events in social media networks

more than traditional *news* and *blogging* media. Detection of topics on these short messages can make a more describing insight of users opinions about named events and real-world occurrences.

A new research area of this TDT race has begun while new social media like *Twitter* has come to existence. *Twitter* by its nature is composed of users instantly sending short posts called *tweets*. These *tweets* can be daily life messages of a user such as “*i ate a pizza! yaaay!*”; important messages from a technical society like “*Ubuntu 16.10 release date is soon!*”; or even a political message like “*WikiLeaks operative: Clinton campaign emails came from inside leaks, not Russian hackers.*” These messages are often tagged with specific word to make it addressable and fetchable. Figure 1 shows an example of tagging in *Twitter*. However, mostly this tag does not show much relation between desired news and topics, only a user’s point of view in relation to his/her *tweet*. One message can be about voting while another is related to feeding ducks and both are tagged as *#DuckTales*. This issue can be addressed as *variety* from big data aspect and *ambiguity* from natural language processing aspect. Moreover, detection of a real-world event with large *volume* and *velocity* of data requires more research than finding an event on selected and filtered datasets [6]. Another problem with this media is noisiness of posted tweets. These tweets, unlike news articles and intellectual documents, are not well written and contain misspelling, grammatical errors, and even words or expressions like “*yaaaaaay*” that are not literary. Expressed problems of this media make TDT task much harder.

Data mining and artificial intelligence community has seen many research works done in this scope which show promising leverage compared to each other. Many of these works are based upon simple *bag of words* model while others keep searching on *probabilistic topic models* and still some of them look for sudden change in monitored properties. The common part of them all is the use of natural language processing techniques and methods instead of character level stochastic *n*-gram models.

These methodologies have come to aid in accomplishing the task of detecting and tracking events, and topics on social media streamlines are emerging to answer couple of questions such as the following:

- (i) What everybody talks about in a specific time?
- (ii) What is trending?
- (iii) What happens somewhere on Earth?
- (iv) Also, dynamic answered questions which have temporal and spatial properties with great increase of public interest.

In order to find most related articles to this scope, we used Google Scholar academic search engine. First we prepared our search keywords that are listed as follows:

- (i) Topic detection
- (ii) Twitter topic detection



FIGURE 1: Some *tweets* related to a hashtag: *#DuckTales*.

- (iii) Twitter event detection
- (iv) Twitter event extraction
- (v) Twitter topic extraction
- (vi) Twitter topic tracking
- (vii) Twitter event tracking.
- (viii) Twitter trending topic.
- (ix) Twitter trending event

We used citation per year metric to get an overall metric of importance of each article from an academic point of view. We used a threshold of two for this metric and eliminated articles that had less than two citations per year. In case of new articles, such as the ones that are published in the past two years, we did not remove them from the list even if they have less than two citations per year. In order to make sure that the unrelated articles are eliminated from the list, we read the title and abstract of each article and eliminated ones that are not related to our review title. Afterwards, we categorized the remaining articles based on their novelty and methodology. The remaining articles are the ones used to conduct this research.

This review article is organized as follows: Firstly, Section 2 describes *Twitter* as a service. Section 3 categorizes and explains existing methods and models. In Section 4, pre-processing as a general step which is common between methods is explained. Section 5 details the methods and approaches based on different categorizations. Section 6 provides a general discussion about data and evaluation issues. At the end, Section 7 concludes the paper.

2. Twitter

Twitter is described in the current section and its respective features are detailed. In Section 2.1, this microblogging service and its data types are explained. Section 2.2 discusses the details of TDT task obstacles in case of Twitter. Finally, in Section 2.3, the social big data tools are explained and detailed.

2.1. Twitter Microblogging Service. Twitter as one of the largest social blogging services is the world's fifteenth website, and in the United States of America it is the ninth and has been linked by over 6,087,240 websites (extracted from Alexa website). Its services include posting of short text messages on online Twitter platform which also enables users to track posted short messages of other users by *following* them. These short messages are called tweets that may contain a GIF image, a short textual message containing 140 characters or less including some *emojis* or only text, an image, or a poll. All of these parts are listed as parts of a *tweet*:

- (1) A short text message composed of 140 characters or less that can contain emojis
- (2) An image
- (3) A GIF describing short text message, feeling, or anything else
- (4) A poll question with predefined answers (only one of last three parts of a *tweet* can be used)

Twitter allows users to communicate in their respective social network with other users by these tweets. They can share their ideas, feelings, polling questions, pictures, and anything else that has no contradiction with its rules. A tweet posted on Twitter can be seen by other users by default unless users change their privacy settings to make it readable to only followers list or specific people.

A *Mention* or *Reply* tweet can be made by using "@" symbol before a user name. These replies or mentions create a more social web service by helping users to interact with and reply to each other. Retweet is also another feature of Twitter that allows users to resend or forward another user's tweet to their respective followers. Hashtag is also another feature of Twitter that helps users categorize their tweets with use of a "#" sign and a word related to the posted tweet; this simple keyword style helps in tweets retrieval and categorization and is also used by Twitter to detect trending events.

Twitter also provides an application programming interface (Twitter API) that enables developers and researchers to access its streaming tweets. This streaming can be filtered out by location, specific keyword, author, etc.

2.2. Challenges of Twitter for Event Detection and Tracking Task. Twitter as a great information source that is described in the earlier section has enormous information

retrieval issues that make event detection and tracking task in its growing social network much harder. Twitter streams usually contain large amount of rumor tweets that have been generated by users or spammers. These fable, fiction, and in most cases mendacity tweets greatly affect the performance of event detector and tracker systems. Another issue arises when most of tweets are related to daily life of users, that is, about their personal information and daily activities. In some cases such as elections, these daily activities can be used to retrieve good information, but in the case of general event detection, they are not so much helpful. For a good event detector and tracker system, it is necessary to separate this irregular and polluted information from useful information.

Twitter messages are as short as 140 characters as the maximum size, which raises another problem. These short messages must be grouped or preprocessed to make a longer stream of tweets. Event detection and tracking in general long documents and newswire is much easier in terms of sparsity and irrelevance of documents than in the case of short blogging services such as Twitter. Most of Twitter posts contain grammatical errors and misspellings that make it harder compared to regular newswire. Twitter, as a source of user generated data, mostly contains many unseen words that are only seen in short messages. As an example of such words and abbreviations, we can name the word "OMG" which is equivalent to "Oh My God"; such words are used and generated frequently by users. Users also add misspelling and lengthen to such words, which results in a very unpleasant issue.

All of the mentioned problems are also added to big data 3-V model in which a large *variety* of *velocity* data along with big *volume* are generated and need to be processed just-in-time to be monitored and tracked. This 3-V model is much more generalized than the 5-V model that is defined as follows:

- (i) *Volume* denotes large amount in terms of tally about data that is streamed or generated. Processing, grouping, clustering, and making useful information out of large scale data are crucial in information retrieval applications and also in case of Twitter-like social networks.
- (ii) *Velocity* indicates speed of data generation or transfer. Streaming and online data sources such as Twitter possess this property in which real-time information extraction applications are needed to fit this kind of speed.
- (iii) *Variety* is called difference of data gathered from a data source in which various data types are generated and collected to be processed. In case of Twitter, this data is different because users' generated data types are about distinct topics and events.
- (iv) *Value* describes the process of information extraction from big data sources. It is also known as big data analysis that in case of Twitter is noted as *big social data analysis*.

- (v) *Veracity* refers to correctness and accuracy of information extracted from a big data source. It is also known as data quality [7]. This quality is poor for some tweets (user generated daily life tweets) while it is rich in case of Twitter newswire accounts (such as a news channel related Twitter account that only posts rich tweets about real-world events).

2.3. *Social Big Data Tools*. Many tools for different applications of social big data analysis, storage, database systems, cluster computing, web crawler, data integration, parallel data flow, and complex event processing are presented by different companies. These tools are trivial for today's big data analysis and of course for Twitter data analysis. Some methodologies in this review use some of these tools while others do not:

- (1) *Lucene* is a free and open-source information retrieval java library that has been ported to other programming languages such as PHP, C#, C++, Python, and Ruby. Indexing, searching, and recommendation are other capabilities of this tool. It has its own mini-query syntax which is easy to grasp, and its nature helps researchers and information retrieval industry to use it as a free and open-source Apache foundation tool [8, 9].
- (2) *Apache Storm* is another free and open-source real-time computing system. It can reliably process unbounded streams of data for real-time applications. It is simple and can be used with any programming language [10].
- (3) *NoSQL databases* such as MongoDB are designed to store and retrieve any data with big data properties in large scales. Social data storage and retrieval require NoSQL databases to perform computing tasks [11].

Other tools and programming languages can be used in this particular job, but the main properties of social big data require making use of the described tools as their relativity.

3. Categorization of Methods

Existing methods for event detection and tracking task in Twitter can be categorized in different ways based on diverse points of view. One of these categorizations distinguishes between methods that *only detect* versus methods that *detect and track* events. Some of existing methods only detect while others track detected events and make storyline of detected topics based on timeline of tweets. The first one is also known as *topic detector* while the other one realizing importance of tracking is an *event detector and tracker*, respectively, abbreviated as *TD* and *EDT*.

Another categorization is raised when different methods use different Twitter data sources. Some use offline datasets for detection and/or tracking while others make use of online Twitter API. This distinction of data acquisition for training and testing part of algorithms raises a comparison error when comparing performance and results accuracy of existing methods.

Two other categories for event detection and tracking are known as retrospective event detection and new event detection. These two are abbreviated as *RED* and *NED*. The main focus of RED is to discover previously unidentified events from offline datasets and documents while NED is focused on finding new events in online data streams. For TDT tasks, these two concepts are broadly investigated, and many research articles have been published to fulfill this task. From Twitter point of view, event discovery algorithm can be either NED or RED. Iterative clustering algorithms such as *k-means* are a common practice in RED category. Firstly, a document, sentence, or short tweet is selected as an entity and other entities are compared to the first one; if it is close enough in terms of distance in vector space, then both are merged to form bigger cluster; if not, a new cluster is created and this object is assigned to that new one. This process continues until all objects (documents/sentences/tweets) are finished. In contrast to RED, NED does not have any initial query or cluster; thus, it must provide some decision rules between new or old events. TF-IDF metric is used in some practices to compare new streams and old ones. In some cases a time attribute is also added to close clusters when specific time is passed; for example, after three days, no further tweets are added to that specific cluster.

“New” and “retrospective” terms belong to *document-pivot* techniques in which algorithms are designed to investigate textual properties of related objects. These techniques aim to provide some metrics to compute similarity of objects based on their textual and linguistic properties.

Being in contradiction to document-pivot methods, *feature-pivot* methods aim to find rapidly growing property in detection stream. This so called *bursty activity* with rising frequency describes a new event fortuity. For example, maybe a huge rise in hashtag usage frequency in Twitter is due to a new event which is happening or has been occurred recently.

Some Twitter event detection and tracking methods use predefined information about users or administrators interests. These methods are known as specified event detectors. Some other techniques do not need any information about events to be tracked and detected and find the real-world occurrences, topics, and events by their properties in frequency raise pick or in terms of similarities. These two distinct methodologies are known as *specified event* and *unspecified event* detection and tracking systems.

As described in this section, many categorizations are drawn for event detector systems; these categorizations lack the main methodology part of algorithms. Section 5.1 describes a new categorization and explains existing methods under this categorization. Table 1 shows a list of methodologies that are studied through this manuscript.

4. Preprocessing

Preprocessing of data in data mining related applications is a common practice while it is also inevitable in the Twitter event detection task. This task includes parts such as data normalization, removal of noisy data, and amendment. NLP tasks require grammatically correct text with certain

TABLE 1: Twitter topic/event detection/tracking related studies.

Reference	Detection method	Detection type		Detection task		Data collection Dataset	Detection task
		Event	Topic	RED	NED		
[12]	Naïve Bayes classifier		✓		✓	Twitter API, handpicked users	Hot news detection
[13]	BScore based BOW clustering	✓			✓	Twitter API (offline)	Disaster and story detection
[14]	BOW distance similarity	✓			✓	Twitter API	FSD (first story detection)
[15]	BNgram and TF-IDF		✓	✓		Offline datasets	Topic detection
[16]	Cross checking via Wikipedia	✓			✓	Twitter API, Wikipedia	Hot news detection
[17]	Formal concept analysis		✓		✓	RepLab 2013 dataset	Topic detection
[18]	FPM (frequent pattern mining)	✓			✓	Twitter API	Event detection
[19]	FPM		✓	✓		Super Tuesday/FA Cup/US elections	Topic detection
[20]	FPM (hierarchical clustering)		✓		✓	Topic dataset from CLEAr system	Topic detection
[21]	FPM (TF-IDF & n -gram improved)	✓			✓	Twitter API	Event detection
[22]	GPU improved TF-IDF approximation		✓	✓		Offline dataset	Topic detection
[23]	BOW similarity	✓			✓	Offline dataset	Topic detection
[24]	Word embedding					SemEval dataset	Twitter sentiment classification
[25]	Spatiotemporal detection	✓			✓	Offline dataset	Targeted-domain event detection
[26]	Clustering of temporal & spatial features	✓			✓	Twitter API	Event detection
[27]	Geographical regularity estimation	✓			✓	Twitter API	Geosocial event detection
[28]	BOW clustering	✓			✓	Twitter API	Event detection & analysis
[29]	Probabilistic modeling	✓			✓	Twitter API	Early disaster detection
[30]	FPM	✓			✓	Offline dataset	Event detection
[31]	Heartbeat graph	✓			✓	Super Tuesday/FA Cup/US elections	Topic/event detection
[32]	Enhanced heartbeat graph	✓			✓	Super Tuesday/FA Cup/US elections	Topic/event detection
[33]	Sentence BERT/streaming graph mining		✓	✓	✓	Super Tuesday/FA Cup/US elections	Topic/event detection
[34]	Universal sentence encoder		✓	✓	✓	COVID-19 dataset	COVID-19 topics
[35]	TF-IDF, CCA, and BTM		✓	✓	✓	Twitter API	Trend ranking
[36]	LDA, USE, and SBERT	✓			✓	COVID-19 dataset	COVID-19 topics
[37]	Autoencoder and fuzzy c -means		✓	✓		Berita	Trend ranking

properties. Preprocessing is one of the main parts of social big data analysis subtasks. Short tweets communicated through Twitter service as described before need to be processed to be ready for further event detection computations. Removal of stop words and punctuation marks is a crucial step in preprocessing of natural language processing related data mining tasks [38]. Identification of URLs and emojis is also needed. Regular expressions can be used to detect URLs in short messages.

In some cases, stemming is also applied for unification of processed words while non-target-language words are also vanished in this process. Elimination of non-target-language words helps improve extracted topic to be in a target language. Tokenization is also another part of preprocessing that gives unique tokens to each word in a tweet. This part of preprocessing is more crucial in TF-IDF (Term Frequency-Inverse Document Frequency) related models.

Some methodologies like *EvenTweet* [26] use WordNet [39] check as part of their preprocessing. This WordNet dictionary lookup improves correctness of preprocessing

output; thus, no non-English and incorrect words will be used for event detection task. Slang word translation is also used to translate user generated words into their formal meaning. *NoSlang* website is also a common tool for this task [40].

Common information retrieval processes from Twitter or any other online web-based data sources require special preprocessing techniques. One of these techniques is removal of unwanted and trashy character sets such as HTML tags. Sometimes these trashy looking character sets seem to be useful (in case of encoding and critical information related to data). White space and punctuation marks that are also called white spaces need to be sorted out. An example of these occurrences is *Ph.D.* that has ambiguity of end of sentence; another example is *\$5.79*.

The main concepts of a clean and clear text are *Word Token* and *Word Type*. The first one refers to occurrences of words that are numbered while the latter one implies unique words that are entries of a table called *vocabulary* list. Tokenizing a text is a natural language processing task aimed

at tokenizing words and giving them unique numbers in sentence which later will be used by tasks such as stemming or part of speech tagging.

As discussed so far, preprocessing is an essential and inevitable part of any natural language processing algorithm, and in case of Twitter TDT task it is also demanded.

5. Event Detection and Tracking Task in Twitter

Event detection and tracking task in Twitter is a well investigated research issue. This section provides details of approaches that are applied to this problem.

5.1. Event Detection in Twitter: Methodological Categorization. Event detection and tracking in most of cases is composed of known data mining methods that have been used before in different areas. Such algorithms and methods are combined with NLP techniques to obtain better results over testing process of algorithms. In this subsection we try to categorize existing algorithms for this task with respect to their utilized data mining and NLP methods.

5.1.1. Bag of Words Methods. Inclusive methods of this category mainly use TF – IDF metric to extract final topic related to tweets, and any other features of a sentence like its part of speech tags are disregarded. Term Frequency-Inverse Document Frequency, abbreviated as TF – IDF, is a common metric among most of topic detection or extraction methods and is described as (1) and (2). Respectively, t and d in these equations refer to term and document, which in case of the latter can be assumed as a single document containing more than a tweet, maybe couple of tweets or just a single tweet which can also be referred to as a message. Furthermore, $\text{count}(t \text{ in } d)$ represents counting occurrences of term t in document/message d while $\text{count}(d \text{ has } t)$ denotes counting documents/messages that have at least one occurrence of t .

A similarity metric is used with utilization of TF – IDF to compare two separate tweets in [41]. This similarity metric described in (3) is used as a score function to group new messages; a message that does not belong to any group is considered to be a new group. New groups are populated in order of classification of new messages with respect to score function. To avoid unrelated messages to first one in a group, all messages are compared to first message and top k messages.

$$\text{tf}(t, d) = \frac{\text{count}(t \text{ in } d)}{\text{size}(d)}, \quad (1)$$

$$\text{idf}(t) = 1 + \log \frac{N}{\text{count}(d \text{ has } t)}, \quad (2)$$

$$\text{similarity}(d_1, d_2) = \sum_{t \in d_1} \text{tf}(t, d_2) \times \text{idf}(t) \times \text{boost}(t). \quad (3)$$

Another method described in [12] represents a new architecture for news related TDT task from Twitter. In this architecture, a cosine similarity measure is utilized along

with TF-IDF representation of tweets to accomplish this task. This similarity measure is computed between tweet t and cluster c . Equation (4) shows related mathematical expression. Feature vectors of \overrightarrow{FV}_t and \overrightarrow{FV}_c are obtained from TF – IDF model of messages. A Gaussian attenuator is then applied to this similarity measure to place impact of temporal dimension in clustering. This weight makes sure that no old clusters and messages get twisted. This architecture makes use of hand selected users which are most likely to post news and also a sampling and tracking system.

The *BNgram* model that is introduced in [15] along with sentiment classification and part of speech tagging forms a trending topic detection system. *BNgram* model in this research is similar to [41] with small differences that imply boost factor. If this factor is set to 1.5, then n -gram model holds named entity; otherwise, it is a small number, and the respective model does not hold a named entity. Based on n -gram TF-IDF, all tweets are scored and, based on these scores, are then clustered into respective clusters. This scoring and clustering process is conducted in time windows, and in each time step, tweets related to a time window are compared to others that have been posted earlier. The proposed method has been trained on some handpicked datasets collected from Twitter API which were related to sports (the Cricket World Cup 2015), medicine (Swine Flu 2015), and bills (Land Acquisition Bill). Compared to frequent pattern mining methods, this method seems to be a simpler algorithm in terms of software implementation with good results in terms of output topics on some cases that shamefully are not expressed as F-measure, precision, recall, or any related metrics. The only social big data tool that this method uses is Lucene for keyword indexing.

“Bieber no more!” is title of another article in these criteria which uses simple nearest neighbor among tweet hashtags to find dissimilarity of previously seen events and new ones [16]. This first story detection system utilizes Wikipedia as a source of information. Wikipedia is a multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on a model of openly editable content. Wikipedia page view helps to find out if an event occurred recently or it is just a false positive detected by this system. Simple use of nearest neighbor among hashtags of multiple tweets and utilization of Wikipedia are expanded to a multistream first story detection system. This system works in the same manner of single-stream first story detection with the only difference being in vector space modeling. This vector space modeling between tweets and Wikipedia pages checks the following: if any new event occurred, it is reflected as pick user page views in Wikipedia; if it was a false positive, no pick view on Wikipedia-related page happens.

Another first story detection system is proposed in [14]. This system makes use of an improved version of locality sensitive hashing (LSH) within a $(1 + \epsilon) \times r$ distance of query point for Twitter first story detection. Time and space bounding narrow nearest neighbor finding problem. This problem arises when huge amount of user tweets are posted per day, and the goal is to find out if they point to a new story/event or a previously seen one; storing all of these data

and finding nearest neighbor between them are almost impossible. Time bounding refers to using a time window instead of computing all data from all times while space bounding points to solving this problem among limited number of tweets. Similarity of a tweet compared to previous ones shows if it is new or not, and this task guides proposed system to open a new story or keep it the way it was.

Another way of extracting answers for 4-W question, *Who, What, When, and Where*, is proposed in [42] which uses a new data representation method called *named entity vector*. This data representation vector along with *term vector* is integrated as a *mixed vector* to obtain results.

$$\text{cosine_similarity}(t, c) = \frac{\overrightarrow{FV}_t \cdot \overrightarrow{FV}_c}{\|\overrightarrow{FV}_t\| \times \|\overrightarrow{FV}_c\|}. \quad (4)$$

Term Frequency-Inverse Document Frequency (TF-IDF), Combined Component Approach (CCA), and Bitern Topic Model (BTM) are the main approaches addressed in [35]. Ranking trends is aimed to be solved by authors by using these models and features.

5.1.2. Probabilistic Models and Classifiers. Probabilistic topics models and classifiers that are described in this section are used to model and classify Twitter datasets or streamlines. One of these approaches that is presented in [23] uses a Naïve Bayesian classifier called NB-Text to satisfy this requirement. This probabilistic method is trained over 2,600,000 Twitter messages annotated by humans posted on 2010. This dataset is labeled for training and testing phases. Firstly, a classifier called RW-Tweet is trained to distinguish between real-world and non-real-world events. Weka toolkit [43] along with extracted cluster level features is used to train classification model. This Naïve Bayesian classifier treats all messages in a cluster as a single document and uses TF-IDF metric as features. Cluster level event features such as temporal, social, Twitter central, and topical features are utilized for this classifier.

TwitterStand is the name of another system proposed in [12] that clusters events by a Naïve Bayesian classifier. This can deal with noise and fragmentation. Noise, according to the authors, is clusters that are not relevant to real-world events; thus, reliable news sources as seeds are used instead of regular users, which weakens this system. This assumption is true when news sources post news in real-time, but the nature of social media has proven that users are the real people who happen to be a part of event or disaster. On the other hand, fragmentation refers to duplicate clusters that mean the very same event. Periodic checking of duplicate clusters overcomes this problem on the system. Event geolocating of this system makes it stronger and more useful.

5.1.3. Formal Concept Analysis. Formal concept analysis has been used by [17] in an unsupervised fashion. RepLab 2013 dataset [44] is used to evaluate this system. Formal concept as it is known from literature is an approach for finding relations between data that is almost hidden in its nature.

This relation can be defined between objects and their attributes.

Extent: if we see A as a set of objects (itemset), then it is called an extent

Intent: if B is a set of all attributes of set A, then it is called intent

Formal concept analysis in this way is formalization of extension and intention to find the most related items that possess important attributes in share.

In [17], tweets are seen as objects and their terms are attributes, which makes this methodology very similar to the ones described in Section 5.1.4 as FPM methods. The proposed method tries to find *concept lattices* in unstructured data of tweets, which shows good reliability and sensitivity. A set of tweets in proposed setup of this work are assumed to be objects while terms (words) are attributes. A relation indicates that a term has been used in a tweet. Formal concepts extracted from concept lattice show topics. Some of these concepts are discarded to have better topic. Small concept lattice and terms are computable with this methodology while bigger size of corpus and tweets and vast number of terms lead to a huge lattice. In such a case, a term selection strategy is required to narrow down this problem. Most shared attribute selection strategies drop least shared attributes (terms). This balanced version of algorithm utilizes term frequency of each attribute. This term frequency (tf) shows a threshold of selecting which term should be used in concept lattice. In each iteration, terms with highest tf are selected, and objects (tweets) with less than two terms in their attributes are discarded. Last iteration of this fine-tuned strategy outputs the attributes with highest tf and objects that possess them. Last step of this framework is to actually make topics out of these lattices. However, the previous step has reduced the potential concept lattices to be candidates of final topic. Stability concept that has been previously proposed in [45] indicates how much concept intent depends on objects available in extent. This reduction with keeping stability helps to form topic.

5.1.4. Frequent Pattern Mining Methods. Frequent pattern mining methods have been applied to TDT task in Twitter. Frequent pattern mining (FPM) as indicated by its name is concept of finding frequent itemsets in a database or any related data storage. A simple example of these frequently repeated patterns is described as a set of coffee and donuts which are in most of cases bought together [46].

In [19], a FPM algorithm is introduced for Twitter offline dataset and compared to other relative studies. FP-growth algorithm with small modifications and utilization of similarity metric is applied to form a set of related tweets that form a topic. Cooccurrence patterns between terms that are larger than two constitute main contribution of this work. Three phases of topic extraction in this method are term selection, cooccurrence-vector formation, and post-processing. First stage indicates that likelihood of terms occurrence in a corpus is major concern. A probability such as $P(\text{term}|\text{corpus})$ is obtained in this phase, and between a

new corpus and this reference corpus, this likelihood is compared with ratio of $(P(\text{term}|\text{corpus}_{\text{new}}))/ (P(\text{term}|\text{corpus}_{\text{ref}}))$. This ratio is a metric to show how a term frequency is changing. Higher ratio means higher frequency of appearance, and thus this term can appear in the final topic. Next phase constructs S and D matrices that are later used for frequent pattern mining. Matrix D_s shows how many terms of S appear in several documents while D_t shows how many times a term appears in several documents. Cosine similarity between these two matrices indicates how a term is suitable for adding to final topic. A sigmoid function is used to limit this similarity and act like a threshold. Final phase of this algorithm is a cleaning stage to remove duplicated topics.

Moreover, a similar method that uses FPM to detect social events from Twitter is introduced in [21]. At first step, the K most relevant terms of current set of tweets such C_{curr} are selected by means of highest appearance likelihood. After this step, the soft version of FPM with utilization of sigmoid function as a threshold computes similarity. Social aspects such as event, spam, and past event are introduced to evaluate performance of system. This system performs on live Twitter streamline.

The idea of burst pattern mining that is introduced in [20] is used to construct burst topic user graph with other various features. These features are *tweet number*, *retweet ratio*, *reply ratio*, *user number*, *overlap user ratio*, *big user ratio*, *burst number*, *burst interval*, and *burst time interval*. Macro and micro burst patterns are defined as bellow as main contributions of this work.

Macro burst pattern is finding all clusters in BT in which BT is a burst topic set, and this task is accomplished with the use of a distance measure among features.

Micro burst pattern is finding all subgraphs in user graph G such that $\text{sup}_G(GS) > \text{threshold}$.

This algorithm starts with finding set S that contains all frequent edges, and with use of DFS (Deep First Search algorithm), the subgraph extension algorithm eliminates nodes that do not satisfy the support threshold (τ). The subgraph extension algorithm is executed recursively to extend frequent subgraphs.

Association rule mining (ARL) is another approach of frequent pattern mining in relational databases that has been used in [18] to detect events in Twitter. ARL has two parts: antecedent and consequent. An antecedent is an item that is found in data while a consequent is an item found in combination with the antecedent [47]. These can be named as if/then (antecedent/consequent) patterns with help of criteria support to identify the strongest and most important relations between items in data. In [18], two main equations are used to match rules with regard to their similarity; they are adopted from [48]. Emerging rules as a contribution of this work are proposed to identify breaking news. US Elections dataset has been used to evaluate the proposed methodology that shows good results in terms of F-measure, recall, and precision.

Tracking dynamics of words in terms of graph, or converting sentences into graph representation and trying to understand the spikes inside, is a very useful method. The

graph heartbeat model, introduced by [31], and its enhanced version [32] are all based on this fact. They used graph analytics to detect the emerging events from Twitter data stream by using graph based formulation and spike detection. This spike detection that is called heartbeat model is a mathematical formulation of matrix analysis during detection of events from Twitter social media.

5.1.5. Signal Transformation-Based Approaches. Signal transformation based approaches, such as *Fourier* and *wavelet* transforms, apply spectral analysis techniques to categorize features for different event properties. DFT (discrete Fourier transform) methodology that has been applied in [49] converts burst in time domain to spike in frequency domain. This spike only shows a bursty event, not its period. Thus, a mixture of Gaussian models for identifying time period of these feature bursts have been applied. Fourier transform is given in (5) which is invertible, and its inverse transform that leads to the $y_f(t)$ function is given in (6).

$$X_k = \sum_{t=1}^T y_f(t) e^{-(-2\pi i/T)(k-1)t}, \quad k = 1, 2, \dots, T, \quad (5)$$

$$y_f(t) = \frac{1}{T} \sum_{k=1}^T X_k e^{(2\pi i/T)(k-1)t}, \quad t = 1, 2, \dots, T. \quad (6)$$

With these prerequisites known, the dominant period spectrum can be explained further; this period is assumed to be a period in which the specified frequency reaches its maximum activeness or, in other words, it is bursty. These specifications tempted the authors of [49] to categorize all features into four main types, *HH*, *HL*, *LH*, and *LL* (the first letter shows Dominant Power Spectrum, and the second letter indicates dominant period in which H means high and L means low). Detecting periodic feature bursts is accomplished by aid of a Gaussian mixture.

Reference [30] presented a new online event detector in news streams with utilization of statistical significant tests of n -gram word frequency within a time frame. Three definitions given in the original manuscript are *textual data stream*, *alphabet*, and *time frame* that are, respectively, described as a sequence of text samples S_t that is sorted by t (time), English words (such as “president” and “coffee”), and a time range starting from t_0 and ending at T in form of $[t_0, t_0 + T]$. In this terminology, an event is described to be a change in the source of text stream which is a surprising rise in n -gram frequency. Computed p value for n -gram hypotheses gives a clear insight about the correctness of the null hypothesis that is stated to be “two individual textual datasets of two time frames are generated from one source.” Due to vast variety of n -grams, a suffix tree is also proposed to store the n -gram. Computed frequency is stored in this new data structure, and another algorithm runs over the tree to calculate and store p values along with it.

Clustering of discrete wavelet signals of words generated from Twitter is also another approach that is used in [50]. Unlike Fourier transform, wavelet transformations are

localized in both time and frequency domain and hence able to identify the time and the duration of a bursty event within the signal. Wavelet signal transformation transforms signal from time domain to time and scale domain. A wavelet family is defined in

$$\psi_{a,b}(t) = |a|^{-1/2} \psi\left(\frac{t-b}{a}\right), \quad (a, b \in \mathbb{R}), a \neq 0. \quad (7)$$

Wavelet energy, entropy, and H-measure are also other discrete wavelet transformation parts that give useful information about the signal. H-measure is normalized Shannon wavelet entropy that shows distribution of signal over different scales. The proposed *EDCoW* algorithm (Event Detection with Clustering of Wavelet-based signals) has three main components of signal construction, cross correlation computation, and modularity-based graph partitioning.

First step computes DF-IDF (DF is not TF and it means document frequency rather than term frequency) shown in the following equation:

$$df - idf(\text{term}) = \frac{N_{\text{tweets}}(\text{term})}{N(\text{term})} \times \log \frac{\sum_{i=1}^{T_c} N(i)}{\sum_{i=1}^{T_c} N_{\text{Tweets}}(i)}. \quad (8)$$

A raise of DF-IDF metric is also reflected as a raise in wavelet entropy of this metric. Cross correlation of two different signals is used to group words/terms that happen to have raise in their wavelet entropy together, meaning that these terms have been used together in a topic that previously seen in a raise or happened to be an event candidate. This clustering methodology is suitable for signal transformed detection. A modular sparse matrix is formed at the last phase of this work to detect events by clustering the weighted matrix. This matrix is called M and is in form of $G(V, E, W)$ in which V is vertices, E is edges, and W is weights of the graph G .

A similar method is [51] which uses LDA and hashtag occurrences. This method, unlike [50], uses hashtags to build wavelet signals. LDA is used to form the final topic model. Another difference between this work and [50] is summarization of extracted events that is done with the aid of LDA topic inference and seems to show promising results but cuts off the tweet data and reduces it to hashtags. This reduction harms the algorithm but improves its speed compared to the latter one.

5.1.6. Geoevent Detection Methods. Methods that are described earlier try to only answer the question “What is happening?” However, there is another question yet to be answered: “Where it happened?” Geolocation of an event expresses more insights of a detected event. In [25], a spatiotemporal event detection scheme is proposed; it detects events along with their occurrence time and also geolocation. Some definitions need to be known before further description of algorithm; these definitions are *spatiotemporal event* and *article*.

Spatiotemporal event is a real-world incident that happened at location l and time t which is denoted by event $_{l,t}$. Domain is known to be set of events that fit into a categorization such as music and civil.

Article set of targeted domains can be open or closed. A closed article such as A_p denotes an article related to topic p , and a_x can be a news report from that article.

This manuscript suggests two types for tweet categorization in order to classify tweets as related/unrelated to event. A *positive tweet* is a related tweet to event, and in contrast a *negative tweet* is simply an unrelated tweet to the event. With all this setup, we can dive into the concept of *label*. A tweet label is known to be a triple of $z = (x, Y^{(x)}, \hat{Y}^{(x)})$, where x indicates event, $Y^{(x)}$ indicates related tweets, and $\hat{Y}^{(x)}$ expresses unrelated tweets. Label generation is task of classifying labels of specific topic that are also related/unrelated to the event. After this step is completed, the next step of proposed work is spatiotemporal event detection. This last step inputs a label set on a specific topic that is given from previous step and the real-time Twitter stream and outputs the online event sets of targeted domain that are happening or happened in location l at time t .

First step of this work consists of feature extraction and relevancy ranking. The relevancy ranking step ranks tweets based on how they are relevant to event in terms of textual and spatial similarity. These ranked features are then used by a tweet classifier that is a SVM-based (Support Vector Machine) classifier. Event location estimation is the latest step of this scheme to estimate actual location of classified tweets.

TEDAS is another spatiotemporal event detection system originally proposed in [28]. This system has three main phases: detecting new events, ranking events according to their importance, and generating spatial and temporal patterns of detected and highest ranked events. Java and PHP along with MySQL are utilized to make this system that also makes use of Lucene, Twitter API, and Google Maps to output final user friendly output. Crime and disaster related tweets are subject to this system. A query based use of Twitter API has been applied to obtain tweets. A set of rules for query are needed, so some simple rules are used to obtain tweets, and later these rules are populated with the help of obtained tweets. Twitter and crime or disaster based features help the next phase of this system to classify the obtained tweets; this classifier has accuracy of 80% as authors indicate. The last phase of this scheme uses content, user, and usage related features to rank the detected events while previous phase is focused on guessing the location of user. The first assumption is that the location of user is in his GPS-tagged tweets if there are any; if not, his/her friends are more likely to be close to him. The last assumption says his/her location is mentioned in his tweets for at least once. One of the main problems of this location guessing is that in the case of second and third assumptions, the extracted information can be false.

The idea of social sensors that has been used in [29] is proposed to find the location of real-world disasters in Twitter. The definition of event according to the authors is

an arbitrary classification of a space/time region. As the earlier method, this scheme also makes use of SVM as classifier with three features of types A, B, and C that, respectively, are known as statistical, keyword, and word context features. Each tweet is known to be a sensory value, and users are the sensors of this scheme. They tweet about the event, meaning that they are sensors and sensed values are posted as tweets. This report is helpful to detect the real-world disasters such as earthquakes. Real problem of this assumption is that there is a possibility of error when a user posts unrelated tweets that seem to be relevant; an example of these according to authors can be this tweet: "My boss is shaking hands with someone!" Shaking as a primary keyword is used in this tweet but it does not mean that the Earth is shaking. Other features of previous part make error possibility lower, but still there is a chance. Two spatial and temporal models are proposed to clarify the assumptions. These models rely on tweet time stamp and GPS stamp. The evaluation and experimental results show that the system shows over 60 percent accuracy on two related queries. This valuable system is used as an earthquake warning system in Japan that in time can save lives of several people.

5.1.7. Deep Learning-Based Methods. Transfer learning in deep learning and specially NLP by using new methods and approaches such as Transformers enabled researchers to use pretrained models for various problems. Topic detection and tracking from Twitter is also one of these problems that researchers tried to solve by transfer learning based models such as BERT. TopicBERT is one these methods that utilizes BERT for semantic similarity combined with streaming graph mining [33]. The proposed architecture is composed of a deep named entity recognition model [52], a graph database to store the nodes, and a semantic similarity extraction tool (SBERT). The whole system works in a combined manner in which the different parts constantly try to update the underlying graph database, and an extraction system using probability of clusters and probability of words gets the topics at each moment. This system beats state-of-the-art methods on three different datasets and is one of first methods that used Transformers for topic detection and tracking from Twitter.

Combination of semantic vector representation of tweets with clustering algorithms is another methodology that is investigated in [34]. The authors show that utilization of a good semantic feature extractor in form of a dense vector can be quiet useful when dealing with problems such as topic detection. They have used COVID-19 dataset from Twitter and detected topics relatively. Another similar method for COVID-19 is proposed in [36]. The authors propose to use Sentence BERT and Universal Sentence Encoder (USE) for sentiment analysis in combination with LDA based topic detection.

Autoencoder based fuzzy c-means algorithm is presented by [37]. Autoencoder is used for representation of tweets while fuzzy c-means is the clustering part of method. The authors report their results on Berita dataset which is an Indonesian news dataset from Twitter.

Utilization of these methods, which are all based on deep learning, is a new field in NLP, specially transfer learning based ones that use Transformers to have a semantic understanding of text. This semantic understanding is a missing part of other methods. The semantic clustering used by various methods can categorize texts with different words into a single cluster if they have close meaning. Language models and pretrained transformer based architectures that can capture semantic similarity such as SBERT and USE are successful examples of these approaches. These approaches are well known for their ability to understand complex sentences. In case of USE, it can even match sentences from different languages to each other if they carry the same semantic meaning. Compared to non-deep learning based methods, these approaches provide a semantic way to TDT task in Twitter.

5.1.8. Performance Improvements. Recently modeling data as image and processing it on graphic cards constitute a useful view to fasten data processing and obtain real-time or at least near real-time results. As it has been described before, TF-IDF has been used widely used for TEDT task. Methodology of fastening data processing presented in [22] uses an approximation way to figure out the TF-IDF metric. Similar to FPM methods (Section 5.1.4), it uses a sort-based algorithm to find frequent items (tweets). The described algorithm is inspired by [53].

The first step of this algorithm is to find the most frequent itemsets. If we assume that set of B contains all of ordered pairs, the next step is to reduce these items by their id or just simply add the pairs that have the same id. The last step would be to divide them to total count of itemsets, and the result would be TF. The whole process of this algorithm can be run in parallel on a dedicated GPU which gives it more computing power than regular CPUs and is more suitable for real-time computation of TDT task, because other algorithms are weak on this aspect and most of them are applicable to offline datasets.

5.1.9. Deep Learning Short Sentence Sentiment Classification: A Post-TEDT Phase. The main difference of algorithms and machine learning methods described in this section is that they do not detect topics or track events on Twitter. Instead, they can be recommended after event or topic detection phase in which the overall sentiment of users is averaged on the detected topic. This output can give great analytical information. Algorithms, machine learning roadways, and neural networks categorized in this subsection are post-topic/event detection step with regard to deep learning.

Recently, with emerging growth of deep learning methods in NLP tasks, short sentence classification and sentiment analysis of these sentences have seen a major change of methods and applications. Deep learning, as suggested by its name, allows computational models to have a lot of abstraction layers for data representation [54]. Raise of unified architectures of multilayer neural networks for NLP tasks seems to be a promising methodology to solve many unsolved problems in this scope [55] while word

embeddings such GloVe [56] and Word2Vec [57] suggest new vector representation of words that also possess sentimental property of dedicated words and can be applied in terms of matrix calculus.

Sentiment analysis of short sentences has been focused on by many researcher from many aspects such as short sentences (CharSCNN) [58]. On the other hand, distinct characteristics of corpora obtained from Twitter led researchers to find new algorithms of sentiment analysis and sentence classification tasks in Twitter which are foundation of topic and event detection in Twitter using these new research outcomes.

Like other word embedding algorithms, CharSCNN in its first layer transforms the input words into encoded vectors representing distinct words. Any word such as W that has been encoded into a vector in previous layers is separated in terms of its characters, and each character is encoded into another vector such as r_m^{chr} . Matrix vector multiplication of set $\{r_1^{\text{chr}}, r_2^{\text{chr}}, \dots, r_n^{\text{chr}}\}$ gives r^{chr} for each character that would be character embedding in this layer. Sentence level representation and scoring are applied as described in character and word level. CharSCNN has been applied to two distinct short sentence datasets of Movie Reviews and Twitter posts with word embedding size of 30.

Sentiment-specific word embedding for Twitter sentiment classification that is proposed in [24] uses $C\&W$ method of [59]. Three different neural networks (SSWE_h, SSWE_r, and unified model of SSWE_u) are proposed in this manuscript for different strategies to overcome task of Twitter sentiment classification.

5.2. Specified versus Unspecified. Based on available information about an event that is to be detected, an event detection method can be categorized as specified or unspecified. Unspecified methods mainly rely on detecting temporal signs of Twitter such as bursts or trends. These methods have no prior information about an event, and thus they need to classify relative events based on bursty properties and cluster them. Specified event detection systems, unlike previous ones, need some information of an event that can be its occurrence time, type, description, and venue. These features can be exploited by adapting traditional information retrieval and extraction techniques (such as filtering, query generation and expansion, clustering, and information aggregation) to the unique characteristics of tweets. The next subsections categorize existing methods based on this terminology.

5.2.1. Unspecified Event Detection. User driven Twitter short posts sometimes contain very important information about real-world events that are published by users before news media websites and TV/radio channels. These short but important posts are unknown to event detector system and also not predefined by any supervisor. A raise in Twitter temporal and signal patterns can reveal this fact. For example, a sudden and unexpected raise in use of a keyword or hashtag may show a sudden attraction to that topic, and somehow that might reveal a real-world event. An ambiguity

occurs due to this setup while some frequent hashtags and keywords about daily life tweets are detected as unseen and new event. An efficient unspecified event detection algorithm must deal with this kind of ambiguity.

In [60], an event detection system called TwitterMonitor is proposed. TwitterMonitor identifies emerging topics in real-time in Twitter and provides meaningful analytical information that can be further used to extract a topic to detected event. A StreamListener listens to Twitter API data stream and detects bursty keywords; these keywords are then grouped and along with an index are passed into Trend Analysis module. All of described steps form the backend of system while a user interface sums up all of information and presents it to user. Other implementations such as AllTop, Radian6, Scout Lab, Sysomos, Thoora, and TwitScoop have a user interface to represent information gathered from different social media, newswire, and other data stream lines to the front end user.

TwitterStand is another electronic medium that, with use of Naïve Bayes classifier, separates news from irrelevant user generated tweets [12]. Cosine similarity metric along with TF-IDF weighing classifies the cleansed events. A breaking news detection system also fits this scope that has been previously described [41]. This method collects, groups, ranks, and tracks breaking news from Twitter by sampling tweets and indexing them using Apache Lucene.

First story detection (FSD) system proposed in [14] uses a thread based ranking algorithm to assign a novelty score to tweets and then clusters tweets based on cosine similarity between them. Each tweet is assigned to a thread if it is close to tweets in that thread; otherwise, a new thread is made for this new category. The bigger similarity threshold results in thin categories that are mostly the same while lower threshold results in fat threads.

5.2.2. Specified Event Detection. Specified event detection terminology deletes the question “What is happening?” It simply tends to find “where” or “when” it is happening. The first part of query is known to system, and the latter parts are yet to be answered.

Researchers of Yahoo! Labs in [61] tried to find controversial events that users tend to disbelieve or have opposing opinions about. Controversial event detection is process of detecting events and ranking them according to their controversy. The authors proposed three models for this task: direct model, two-step pipeline model, and two-step blended model. Direct model scores event based on a machine learning regression based algorithm, two-step pipeline model detects events from the snapshot and then scores them based on the controversy, and the soft model of the described one is the two-step blended model. Twitter based news buzz and news and web controversy features are the main feature classes used by this system. This system is user negative opinion mining rather than an event detection system while it still detects events based on entity query.

The very same authors of [61] described another system in [62] that also extracts descriptors from Twitter about the events. Gradient boosted decision trees in a supervised

machine learning fashion are employed to form two main models that authors described: EventBasic and EventAboutness.

Many other methods that are categorized as in this subsection are described earlier and are put together in a cumulative manner in Table 1.

5.3. Unsupervised versus Supervised. Machine learning algorithms are trained in both supervised and unsupervised fashions. This means that a training task can be accomplished using labeled data and the machine learning algorithm is assigned to learn the labels from tagged data, while in the unsupervised methodology, it is accomplished by learning by categorization of unknown data labels that are later to be scored. The unsupervised machine learning algorithms have harder job to do in terms of learning with unknown labels. This subsection describes the unsupervised and supervised algorithms for Twitter TEDT task; other algorithms that are described in previous sections are discarded.

5.3.1. Unsupervised Algorithms. Twitter event detection algorithms that use unsupervised machine learning concepts mostly rely on clustering algorithms. As was described earlier, NED is a term used to identify new event detection systems that, contrary to RED (retrospective event detection), detect and identify new events, while the latter one detects and identifies specified events. Unsupervised methods are highly recommended for tasks that require clustering of unknown categories that exactly fit the NED domain. Furthermore, there is no prior information about the number of classes to be categorized because of dynamic nature of user activities in social networks.

5.3.2. Supervised Algorithms. Supervision of a clustering algorithm that needs labeled data to classify the user generated real-life events has a close relation to RED category. As described earlier, the RED algorithms tend to classify the known events while supervision needs labeled data in its training phase. This terminology has many shortcomings in real-world applications such as event detection system. A system that is aimed to find and track real-world incidents cannot be trained in supervised fashion; this is because of unknown events that yet to come and absence of information about their quantity and entity.

6. Data and Evaluation Issues

Twitter by its nature possesses unstructured and unlabeled data stream that can be obtained from online or offline sources. Online Twitter data source is the Twitter API, and offline data is the offline Twitter data obtained from different snapshots. These snapshots possess better properties to evaluate differences between algorithms or systems that aim to find events or topics on Twitter. Evaluation of an online Twitter event extraction system is doable if the input data is the same input data snapshot that is recorded.

Another drawback of event detection and tracking algorithms that has indirect relation to the previous issue, is the event detection time. Suppose that two algorithms or systems such as A_1 and A_2 both have the same precision and recall on finding events and tracking them on Twitter data snapshot but have different detection times. Detection time is defined as the time it takes for a typical algorithm to detect and identify events and track them. If these times (that is related to time complexity) are the same, we can assume that both algorithms are the same, but in case of different times, the near real-time algorithm should be used and preferred. This metric is not reported in any of the works that have been studied in this manuscript, but it seems an essential step to define a real-time event detection and tracking system. In the case of offline systems, this metric is not important.

Both of the evaluation issues described earlier heavily affect the process of evaluation. The Defense Advanced Research Projects Agency (DARPA) published the results of a competition named “The DARPA Twitter BOT Challenge” [63]. The contestants of this competition were the big companies of information technology industry (SentiMetrix, IBM, USC, DESPIC, B. Fusion, G. Tech). A mathematical scoring system was used to score the bots created by contestants. Equation (9) defines this scoring system. This competition aimed to create bots that can identify fake users (bots) that are posting on Twitter and creating influence. However, the relevance of this research is important, and it is related to event detection and tracking system because the scoring system used in this competition is a usual artificial intelligence related measuring system which also points to speed.

$$\text{Final Score}(t) = \text{Hits}(t) - 0.25 \times \text{Misses}(t) + \text{Speed}. \quad (9)$$

A related scoring system to event detection systems according to (9) can be extracted. The very same manner of speed in evaluation of event detection system is also used in [64] to measure quality of systems.

Duplicity of detected events or topics is also another drawback. Mis-detection of events and identification of a nonevent phenomenon also constitute a huge problem. The reason this issue possesses bigger threads is that a real-time disaster informing system can be fooled and mis-detect a disaster or even not detect it.

With all of these in mind, an evaluation/scoring system for TEDT requires quantities of HITS, MISSES, recall, precision, and speed to be calculated on a specific data snapshot of Twitter. Otherwise, the systems cannot be compared to each other. A typical scoring system can be known as 10 with α, β as weights. Other scores of Score_2 and Score_3 are the precision and recall of algorithm on the dataset.

$$\text{Score}_1(t) = \alpha \times \text{Hits}(t) - \beta \times \text{Misses}(t) + \text{Speed}. \quad (10)$$

7. Conclusion

Twitter as one of the biggest social networks and micro-blogging services enables users to post and share their

thoughts, daily life posts, and news about real-world events. Many of these users' posts are related events are real-world incidents and some are rumor, meaningless, and plot information. Unfolding these real-world events and extracting them from Twitter need real-time systems with high accuracy and precision. Evaluation of systems faces many issues such as data and evaluation metric problems. In this article, we studied some TEDT systems that aim to find, detect, extract, and track real-world incidents from Twitter and also described the problems related to evaluating such systems. Many categorizations were proposed to classify these algorithms and methods that are also presented in this article; in addition, another categorization based on the methodology of the relying algorithms is proposed in this article. Finally, this article discussed a postdetection methodology proposed as deep learning short sentence classification that can be useful after detection of events.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] J. Allan, *Introduction to Topic Detection and Tracking*, Springer, Berlin, Germany, 2002.
- [2] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*, Vol. 12, Springer Science & Business Media, Berlin, Germany, 2012.
- [3] J. Allan, J. G. Carbonell, D. George, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, Lansdowne, VA, USA, February 1998.
- [4] Twitter Usage Statistics, 2017, InternetLiveStats.com.
- [5] Twitter Tweets Per Day Statistics, 2013, <https://blog.twitter.com/2013/new-tweets-per-second-record-and-how>.
- [6] M. James, M. Chui, B. Brown et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, New York, NY, USA, 2011.
- [7] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.
- [8] A. Bialecki, R. Muir, and I. Grant, "Apache lucene 4," in *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*, Portland, OR, USA, August 2012.
- [9] UIMA Apache, Apache Software Foundation, 2011, <https://java.apache.org>.
- [10] Apache, Apache Storm, 2013.
- [11] MongoDB, MongoDB, 2013.
- [12] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: news in tweets," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 42–51, ACM, Seattle, WA, USA, January 2009.
- [13] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 120–123, IEEE, Toronto, Canada, August 2010.
- [14] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Proceedings of the Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 181–189, Association for Computational Linguistics, Los Angeles, CA, USA, June 2010.
- [15] S. D. Tembhurnikar and N. N. Patil, "Topic detection using bngam method and sentiment analysis on twitter dataset," in *Proceedings of the 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, pp. 1–6, Noida, India, September 2015.
- [16] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis, "Bieber no more: first story detection using twitter and wikipedia," in *Proceedings of the SIGIR 2012 Workshop on Time-Aware Information Access*, Portland, OR, USA, August 2012.
- [17] J. Cigarrán, Á. Castellanos, and A. García-Serrano, "A step forward for topic detection in twitter: an FCA-based approach," *Expert Systems with Applications*, vol. 57, pp. 21–36, 2016.
- [18] M. Adedoyin-Olowe, M. M. Gaber, C. M. Dancausa, F. Stahl, and J. B. Gomes, "A rule dynamics approach to event detection in twitter with its application to sports and politics," *Expert Systems with Applications*, vol. 55, pp. 351–360, 2016.
- [19] G. Petkos, S. Papadopoulos, L. Aiello, S. Ryan, and Y. Kompatsiaris, "A soft frequent pattern mining approach for textual topic detection," in *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, p. 25, June 2014.
- [20] G. Dong, W. Yang, F. Zhu, and W. Wang, "Discovering burst patterns of burst topic in twitter," *Computers & Electrical Engineering*, vol. 58, pp. 551–559, 2017.
- [21] S. Gaglio, G. Lo Re, and M. Morana, "Real-time detection of twitter social events from the user's perspective," in *Proceedings of the 2015 IEEE International Conference on Communications (ICC)*, pp. 1207–1212, IEEE, London, UK, June 2015.
- [22] U. Erra, S. Senatore, F. Minnella, and G. Caggianese, "Approximate TF-IDF based on topic extraction from massive message stream using the GPU," *Information Sciences*, vol. 292, pp. 143–161, 2015.
- [23] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: real-world event identification on twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, vol. 11, pp. 438–441, Barcelona, Spain, July 2011.
- [24] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1555–1565, Baltimore, MD, USA, June 2014.
- [25] H. Ting, F. Chen, L. Zhao, C. T. Lu, and N. Ramakrishnan, "Automatic targeted-domain spatiotemporal event detection in twitter," *GeoInformatica*, vol. 20, no. 4, pp. 765–795, 2016.
- [26] H. Abdelhaq, C. Sengstock, and M. Gertz, "EvenTweet: online localized event detection from twitter," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1326–1329, 2013.
- [27] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 1–10, ACM, San Jose, CA, USA, November 2010.
- [28] R. Li, K. H. Lei, R. Khadiwala, and K. C. C. Chang, "Tedas: a twitter-based event detection and analysis system," in *Proceedings of the 2012 IEEE 28th International Conference on*

- Data Engineering (ICDE)*, pp. 1273–1276, IEEE, Arlington, VA, USA, April 2012.
- [29] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860, ACM, Raleigh, NC, USA, April 2010.
- [30] T. Snowsill, F. Nicart, M. Stefani, T. De Bie, and N. Cristianini, “Finding surprising patterns in textual data streams,” in *Proceedings of the 2010 2nd International Workshop on Cognitive Information Processing (CIP)*, pp. 405–410, IEEE, Elba, Italy, June 2010.
- [31] Z. Saeed, R. A. Abbasi, A. Sadaf, M. I. Razzak, and G. Xu, “Text Stream to temporal network—a dynamic heartbeat graph to detect emerging events on twitter,” in *Proceedings of the Advances in Knowledge Discovery and Data Mining in Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 534–545, Springer, Melbourne, Australia, June 2018.
- [32] Z. Saeed, R. A. Abbasi, I. Razzak, O. Maqbool, A. Sadaf, and G. Xu, “Enhanced heartbeat graph for emerging event detection on twitter using time series networks,” *Expert Systems with Applications*, vol. 136, pp. 115–132, 2019.
- [33] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvas, M. A. Balafar, and C. Motamed, “Topicbert: a transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection,” 2020, <https://arxiv.org/abs/2008.06877>.
- [34] M. Asgari-Chenaghlu, N. Nikzad-Khasmakhi, and S. Minaee, “Covid-transformer: detecting trending topics on twitter using universal sentence encoder,” 2020, <https://arxiv.org/abs/2009.03947>.
- [35] H. U. Khan, S. Nasir, K. Nasim, D. Shabbir, and A. Mahmood, “Twitter trends: a ranking algorithm analysis on real time data,” *Expert Systems with Applications*, vol. 164, Article ID 113990, 2021.
- [36] K. Garcia and L. Berton, “Topic detection and sentiment analysis in twitter content related to Covid-19 from Brazil and the USA,” *Applied Soft Computing*, vol. 101, Article ID 107057, 2021.
- [37] H. Murfi, N. Rosaline, and N. Hariadi, “Deep autoencoder-based fuzzy c-means for topic detection,” 2021, <https://arxiv.org/abs/2102.02636>.
- [38] J. Leskovec, A. Rajaraman, and J. David Ullman, *Mining of Massive Datasets*, Cambridge University Press, Cambridge, UK, 2014.
- [39] G. A. Miller, “WordNet,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [40] NoSlang.com, 2017.
- [41] S. Phuvipadawat and T. Murata, “Breaking news detection and tracking in twitter,” in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology—Volume 3, WI-IAT ’10*, pp. 120–123, IEEE Computer Society, Washington, DC, USA, August 2010.
- [42] G. Kumaran and J. Allan, “Text classification and named entities for new event detection,” in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 297–304, ACM, Sheffield, UK, July 2004.
- [43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- [44] E. Amigó, J. C. De Albornoz, I. Chugur et al., “Overview of replab 2013: evaluating online reputation monitoring systems,” in *Proceedings of the Lecture Notes in Computer Science in International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 333–352, Springer, Valencia, Spain, September 2013.
- [45] S. O. Kuznetsov, “On stability of a formal concept,” *Annals of Mathematics and Artificial Intelligence*, vol. 49, no. 1–4, pp. 101–115, 2007.
- [46] C. C. Aggarwal and J. Han, *Frequent Pattern Mining*, Springer, Berlin, Germany, 2014.
- [47] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, “Algorithms for association rule mining—a general survey and comparison,” *ACM Sigkdd Explorations Newsletter*, vol. 2, no. 1, pp. 58–64, 2000.
- [48] D. R. Liu, M. J. Shih, C. J. Liao, and C. H. Lai, “Mining the change of event trends for decision support in environmental scanning,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 972–984, 2009.
- [49] H. Qi, K. Chang, and E. P. Lim, “Analyzing feature trajectories for event detection,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 207–214, ACM, Amsterdam, The Netherlands, July 2007.
- [50] J. Weng and B. S. Lee, “Event detection in twitter,” in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, vol. 11, pp. 401–408, Barcelona, Spain, July 2011.
- [51] M. Cordeiro, “Twitter event detection: combining wavelet analysis and topic inference summarization,” in *Proceedings of the Doctoral Symposium on Informatics Engineering*, Porto, Portugal, January 2012.
- [52] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L. Farzinvas, M. A. Balafar, and C. Motamed, “A multimodal deep learning approach for named entity recognition from social media,” 2020, <https://arxiv.org/abs/2001.06888>.
- [53] U. Erra and B. Frola, “Frequent items mining acceleration exploiting fast parallel sorting on the GPU,” *Procedia Computer Science*, vol. 9, pp. 86–95, 2012, Proceedings of the International Conference on Computational Science, ICCS.
- [54] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [55] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning,” in *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167, ACM, Helsinki, Finland, June 2008.
- [56] J. Pennington, R. Socher, and C. D. Manning, “GloVe: global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014.
- [57] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, <https://arxiv.org/abs/1301.3781>.
- [58] C. N. Dos Santos and M. Gatti, “Deep convolutional neural networks for sentiment analysis of short texts,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78, Dublin, Ireland, August 2014.
- [59] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.
- [60] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,” in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 1155–1158, ACM, Indianapolis, IN, USA, June 2010.

- [61] A. M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM International Conference on Information and knowledge Management*, pp. 1873–1876, ACM, Toronto, Canada, October 2010.
- [62] A. M. Popescu, M. Pennacchiotti, and D. Paranjpe, "Extracting events and event descriptions from twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web*, pp. 105-106, ACM, Hyderabad, India, March 2011.
- [63] V. S. Subrahmanian, A. Azaria, S. Durst et al., "The darpa twitter bot challenge," *Computer*, vol. 49, no. 6, pp. 38–46, 2016.
- [64] A. Weiler, M. Grossniklaus, and M. H. Scholl, "Editorial: survey and experimental analysis of event detection techniques for twitter," *The Computer Journal*, vol. 60, no. 3, pp. 329–346, 2016.

Research Article

Credit Behaviors of Rural Households in the Perspective of Complex Social Networks

Qiang Zhao, Yue Shen , and Chaoqian Li

School of Economics and Finance, Xi'an Jiaotong University, Xi'an 710064, China

Correspondence should be addressed to Yue Shen; 154271416@qq.com

Received 29 March 2021; Revised 5 May 2021; Accepted 25 May 2021; Published 4 June 2021

Academic Editor: Fei Xiong

Copyright © 2021 Qiang 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.

With the increasing number of social networks emerging and evolving, the influence of social networks on human behavior is now again a subject of discussion in academe. Dynamics in social networks, such as opinion formation and information sharing, are restricting or proliferating members' behavior on social networks, while new social network dynamics are created by interpersonal contacts and interactions. Based on this and against the backdrop of unfavourable rural credit development, this article uses CHFS data to discuss the whole and heterogeneous impact of social networks on rural household credit behavior. The results show that (1) social networks can effectively promote rural household credit behavior; (2) social networks have a significant positive impact on both formal credit and informal credit, but the influence of the latter is stronger; (3) both emotional networks and instrumental networks have a positive impact on formal credit and informal credit, and their influences are stronger on informal credit; (4) the influence of emotional network is stronger than instrumental networks on either formal credit or informal credit.

1. Introduction

For the past sixteen years, the documents of the Central Committee of the Communist Party of China (CPC) have been focused on the issue of “agriculture, rural areas and farmers.” From promoting supply-side structural reform in agriculture to combating poverty and implementing the strategy of rural revitalization, rural financial reform is always involved. With the state attaching great importance to the priority development of agriculture and rural areas, China has initially formed a multisubject and multilevel rural financial service system, including policy-oriented, commercial financial institutions, formal finance with Rural Credit Cooperatives as the main body, and informal finance with private lending, pawnshops, and loan companies as the main body. China's rural financial system has been growing and improving. However, at the same time, rural households face financial constraints both in their daily lives and in agricultural production; their funds are in great demand [1]. According to the China Rural Financial Services Report, the balance of rural household loans stood at 10.34 trillion yuan at the end of 2019, up 12.1% year on year but down 1.8%

from the end of last year. Among these, borrowing from relatives and friends remains the main solution, with a high proportion of 68.3%, compared with 24.4% from formal financial institutions. From the attitude of rural households to loans, 73.8% had no willingness to use credit and insisted on doing things according to the money they had on hand. Another 12.7% of rural households only consider loans for agricultural operations, while 9.2% consider loans for both household and agricultural operations. From the above data, the attitude towards credit use for rural households in China is still refusal and rejection; even if it is necessary, rural households prefer to borrow money from relatives and friends.

Why is this happening? On the one hand, rural land is owned by collectives while rural households have lower economic level, fewer household assets, and lack of corresponding collateral in the formal credit market [2], coupled with weak awareness of rural household credit, which cause serious information asymmetry between rural households and formal financial institutions, leading to adverse selection, moral hazard, and so on [3]. These above make rural households subject to more serious formal credit

constraints. On the other hand, China has always attached great importance to “human relations,” which are a typical relationship-oriented group. In social networks, if one party is short of funds, the first choice is to borrow money from his relatives and friends and the other is often the willingness to borrow money free because of “human relations” [4]. In addition, in rural areas, social networks not only can reflect a person’s relationships but also, to some extent, can be used as a symbol of their credit and asset levels [5]. Social networks refer to “a relatively stable association system formed between social individuals because of interaction,” which can be used as an indicator to measure the level of rural family “relationship” [4]. Using social networks of rural households can not only reflect the strength of informal credit but also alleviate the problem of information asymmetry between rural households and financial institutions to reduce the moral hazard and adverse selection of rural households [5]. The reasons are the stronger the social networks, the greater the reputation loss when breaking the contract, in return reducing social networks’ strength and making the cost of breaking the contract higher [6]; besides, rural households in the same social networks can improve each other’s loan repayment ability and credit standing [7].

In recent years, scholars have made an extensive and deep analysis on the problem of rural household credit in China and made theoretical and empirical analysis on the basic characteristics, influencing factors, and economic effects. Most works of literature on social networks and rural household credit thought that social networks have a significant role in promoting rural household credit behavior. Yang et al. discussed the relationship between social networks and farmers’ private credit demand behavior and pointed out that social networks increased farmers’ private credit demand and further promoted farmers’ private credit behavior [1]. Hu and Chen focused on farmers’ lending behavior and found that the tighter social networks are, the more lending behavior farmers would take [2]. Shoji et al., based on the data of Sri Lanka, deeply discussed the relationship between social capital formation and credit access and concluded that the quicker the social network formed, the easier the credit access [8]. Zhou analyzed whether the number of brothers of householders would affect families’ saving rate in the financial market, and the research revealed householders with more brothers had a higher saving rate for they prepared to help each other; this meant more relatives would decrease the possibility of householders to use credit [9]. Many scholars have broadened their research perspectives on the strength and weakness of social network relationships [10], the structure of social capital [11], and the quality of social capital [12] and obtained different conclusions. Therefore, with the situation of previous researches, we found that though contents in previous research are extensive, neither the theoretical level nor the empirical level has been able to draw a consistent conclusion on the specific role of social networks on rural household credit behavior [12, 13], some research conclusions were contradictory [7, 8], and some studies that found the relationship and degree of social networks on rural household credit

behavior were heterogeneous [10–12]. Thus, further research is necessary.

Considering the importance of social networks for rural economic development and the increasing formal credit demand, easing credit constraints, and raising the viability of formal credit institutions, this article studies the influence of social networks on rural household credit behavior based on social network theory, including the influence on formal credit behavior and informal credit behavior. What is more, it discusses the influence from emotional and instrumental heterogeneous perspectives. From theoretical and empirical analysis, we expect to explain the question of whether social networks can promote rural household credit behavior and how the relationship varies under the heterogeneous situation, finally expand China’s rural financial system, and improve the availability of rural household financial services to give targeted policy recommendations and satisfy the credit needs of rural households.

2. Theoretical Analysis and Hypothesis

2.1. Mechanism of Social Network on Rural Household Credit Behavior. For debtors, there are usually two levels of decision-making in their lending behavior: the first level is whether to borrow and the second level is whether to keep the contract, as shown in Figure 1. Among them, the return of households that do not borrow is zero, while the return of keeping contract after borrowing is R_1 and the return of breaching contract after borrowing is R_2 .

Usually, the decision tree is analyzed from far to near, so we first use the game method to judge the second-level decision (whether to keep the contract or not). Social networks have punishment mechanisms and reputation incentive effects [9]. When rural households borrow money and sign a contract, social networks would use the reputation mechanism to regulate and restrain it. If debtors violate the terms of the contract, they would be regarded as dishonest and ungrateful, making the debtors fall under psychological pressure and moral condemnation, thus encouraging the debtors to take the contract seriously. From the perspective of game theory, there is often a prisoner’s dilemma between debtors and creditors, which leads to credit constraints. Meanwhile, social networks can use their own punishment mechanism and reputation incentives; thus, the prisoner’s dilemma game with finite duration is transformed into a repeated game with an infinite duration. In fact, the rural household credit behavior is often conducted many times; both debtors and creditors need to enter the market repeatedly and deal with different opponents, so the game behavior of rural household credit undoubtedly has the repeated game characteristic of the infinite boundary.

Firstly, we assume that the rural household credit behavior is an infinite repeated game; secondly, there is a penalty mechanism in social networks, which can prevent the participants from defaulting easily; finally, we assume that the returns of both debtors and creditors are observable. The return on rural household credit is shown in Table 1, where $b < a$. The bottom right of the table is a game equilibrium formed by the rational behavior of both debtors and

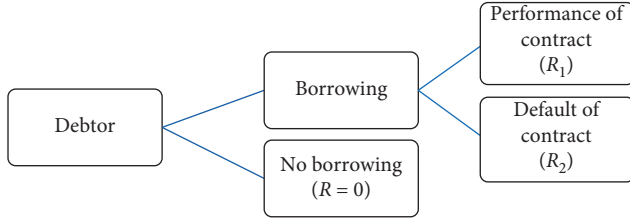


FIGURE 1: Decision tree of rural household credit behavior.

TABLE 1: Game strategy and benefits of participants.

		Creditor	
		Formal debt	Informal debt
Debtor	Breach of contract	$(-c, 0)$	$(a-c, 0)$
	Performance of contract	$(0, a)$	(b, b)

creditors, both sides benefit from b , but it is not stable. The upper left is in the formal credit situation; if debtors choose to default, due to the existence of social networks' punishment mechanism, debtors will have a loss of c unit, including reputation loss or financial loss, and creditors will

suffer a loss due to the debtors' default behavior, the return is zero or even negative; if debtors perform, creditors get a unit of return. In informal credit, if debtors violate, they will get a unit of revenue but also pay c unit of social network penalty cost and creditors' revenue will be zero.

Because we assume the game behavior of rural households is an infinite repeated game, the total income of both sides is the discounted sum of the income in each period; that is,

$$R_i = \frac{1 - \delta}{1 - \delta^{t+1}} \sum_{t=0}^{\infty} \delta^t r_i(\theta_t), \quad (1)$$

where R_i is the total income of debtors and creditors and δ is the discount factor, $0 \leq \delta \leq 1$, and $r_i(\theta_t)$ is the income of debtors and creditors in t period. At this point, we analyze debtors' behavior. If debtors refuse to default, the total return is

$$R_1 = (1 - \delta)(b + b\delta + b\delta^2 + \dots + b\delta^t + \dots). \quad (2)$$

The limit of equation (2) is b . On the other hand, if debtors always default in the t period, then total return is

$$R_2 = (1 - \delta)(b + b\delta + b\delta^2 + \dots + b\delta^{t-1} + a\delta^t - 2c\delta^t + a\delta^{t+1} - 2c\delta^{t+1} + \dots). \quad (3)$$

Simplifying equation (3) and calculating limit, we get

$$R_2 = b - \delta^t(2\delta - 2c - b). \quad (4)$$

Debtors choose to keep the contract only when $R_1 > R_2$, which means

$$\delta > c + \frac{b}{2}. \quad (5)$$

If the discount factor is understood as the patience of the players in the game [1], equation (5) shows that the degree of punishment c can increase the discount factor and the debtors' patience and make them confident in their potential future earnings, thus reducing the incentive to default. That is, the greater the social networks penalty, the greater the debtors' patience and the smaller the incentive to default. Generally, the richer the social networks, the faster and farther the information transfer and the greater the penalties. Therefore, social networks help to promote rural household compliance behavior. In the second level of decision-making, households with stronger social networks are more likely to choose to keep the contract.

Furthermore, we make the first-level decision (whether to borrow or not). Rural households will choose to perform under the influence of a sufficiently large social network, which means $R_1 > R_2$ and $R_1 = b > 0$. Under this circumstance, as opposed to not borrowing and no earning, rural households will choose to borrow. That is, the stronger the social networks are, the more likely the rural households are to borrow and the more likely they are to repay.

2.2. Theoretical Analysis and Hypothesis of Social Network on Rural Household Credit Behavior. According to the credit rationing theory, debtors must have enough assets to obtain a loan [14]. Rural household credit behavior is affected by two factors: on the one hand, rural household own capital is low or not, which makes it difficult to obtain loans, whether formal or informal. On the other hand, since the main source of income for rural households is work or agricultural management, work is vulnerable to the impact of farmers' own health, climate [15], etc. and agricultural production is vulnerable to climate and natural disasters and is inherently high-risk [16], so rural households do not have easy access to finance. As a traditional "hidden guarantee mechanism," social networks can be used as a "guarantee" for rural households with little or no free capital, which makes it easier to obtain financial credit [17]. In both formal and informal credit, members of a social network can measure each other's free capital and the marginal productivity of the debtors' efforts, as they interact closely with each other; thus, the cost of supervision is low and the supervision is powerful [18]. In addition, the sharing function of social networks can also increase household access to financial information, reduce the information asymmetry of their credit behavior, and further promote their credit behavior [9]. In terms of both credit rationing theory and information asymmetry theory, the more socially networked rural households are, the more active they are in crediting and the more likely they are to be financially supported [1]. Therefore, we propose Hypothesis 1.

Hypothesis 1. Social networks can effectively promote rural household credit behavior.

Most previous studies consistently show that social capital is beneficial to easing rural household credit constraints and improving the availability of rural household credit [1, 15, 16]. Social network, as an implicit guarantee mechanism, can fill the gap when rural families carry out formal credit. For example, credit cooperatives do not need collateral and require borrowers to provide group guarantee. Generally, five to seven people form a mutual guarantee group, and the members are responsible for the liabilities of other people in the group. If someone in the group defaults on the loan, the other people in the same group will not be able to obtain new loans. However, at the same time, there is still a gap between this way of guarantee and the real material guarantee, and rural families may still be constrained by formal credit rationing [18, 19]. For this reason, rural households often choose to obtain funds through relatives, friends, and private loans [12, 20, 21], while family and friends are the embodiment of social networks. In general, the richer a family's social networks, the more friends and relatives it has, and the more likely it is to borrow. Among informal credits, due to the small rural area, the close association among the members of social networks, and the low cost of supervision, the moral hazard and adverse selection are effectively avoided [22]; there is considerable literature to suggest that social networks can alleviate the credit rationing problems caused by information asymmetry in rural areas [23]. To maintain household reputation under the rapid spread of public opinion word-of-mouth, social network members restrict their compliance with the loan contract to a certain extent, which plays an important role in the risk control of informal finance [22]. Thus, we propose Hypothesis 2.

Hypothesis 2. Social networks have a stronger positive impact on informal credit than formal credit.

With the continuous development of China's social economy, population mobility and nonagricultural employment have increased, and the implementation of family planning and late marriage policies made rural household size become smaller. With these, traditional family values have been affected and China's rural households are no longer small farmers and have begun to pay more attention to individual values and interests [24]. This leads them to change from survival rationality to social rationality, which brings about the subsequent change of the coverage and intensity of the rural family social networks [25]. In the traditional rural social relations, the rural household social networks are mainly based on the emotional relationship, which is strongly dependent on the family relationship and blood relationship with stability and relative nonselectivity [26]. With the development of the society, the social circle of rural households expanded beyond the blood relationship and instrumental social networks, which based on the professional or classmate relationship began to be established. According to their own purposes and needs, members of society will establish more externalized social

relations and develop instrumental social networks based on mutual interests [27].

Emotional social networks are based on the family concept and the blood relationship and have strong stability and nonselectivity. What is more, violation of group rules and the likelihood of exclusion are lower in emotional social networks, which also means that institutional attributes of emotional social networks are not strong. Therefore, under emotional social networks, rural households are more likely to engage in informal credit behavior. On the one hand, because of China's strong family values and kinship ties, debtors are less likely to refuse loans from those in social networks. Moreover, the cost of borrowing is lower in an emotional society [28]; on the other hand, for formal financial credit, the role of the institution is stronger and the role of emotion is weaker [29]; most emotional social networks provide direct financial support rather than surety support. Previous studies have shown that even relatives and most families are reluctant to provide surety for others [1, 2, 10, 11]. Therefore, we propose Hypothesis 3.

Hypothesis 3. Emotional social networks have a stronger positive impact on informal credit behavior than on formal credit behavior.

As for instrumental social networks, the rural households are relatively more selective, and their maintenance is stronger [27]. Households pay more attention to get more information about employment, finance, and development from instrumental social networks; rural households can use this information more flexibly and scientifically to make better credit practices [28]. Compared to others, rural households with strong instrumental social networks can increase their access to credit by reducing information asymmetry [3], thus helping rural households improve their management and smooth risks and increase the income and the repayment ability of rural households. Based on this, we propose Hypothesis 4.

Hypothesis 4. Instrumental social networks have a stronger positive impact on formal credit behavior than on informal credit behavior.

3. Materials and Methods

3.1. Subject and Data Processing. Data for this article are from the China Household Finance Survey (CHFS) conducted by the Southwestern University of Finance and Economics, covering 29 provinces (cities, autonomous regions) except Tibet, Hongkong, Macao, Xinjiang, and Inner Mongolia. Using the proportional sampling method, 262 counties (cities, districts) were selected; then, four communities (villages and neighbourhood committees) were randomly selected from each county (cities, districts), and finally, 20–50 families were randomly selected from each community. CHFS adopts several measures to control errors such as nonsampling errors and investigates household credit, financial assets, and income at household, individual, and regional levels with high coverage and complete data on

variables. Therefore, as far as the subject of this study is concerned, it has a good representation.

In this article, CHFS 2017 survey data were used as a source sample; the data were screened and cleared to determine the sample, whose process is as follows: firstly, the sample of urban household registration was excluded for this article mainly studies whether social networks affect the rural household credit behavior; secondly, minors and the elderly over 65 are often unable to borrow or have a weak willingness to borrow for reasons of trust, repayment ability, credit requirement restriction, etc. Considering the focus of this study, a sample between the ages of 16 and 60 years was chosen as the research object. Finally, the samples of “Unable to judge,” “Missing,” “Inapplicable,” “Refusing to answer,” and others in the variables of social networks credit behavior were eliminated, and 3,037 valid samples were left.

3.2. Variable Selection and Measurement

3.2.1. Household Credit Behavior. In this article, household credit behavior is divided into formal credit and informal credit. Formal credit refers to household credit to financial institutions such as banks, including agricultural, industrial, and commercial production and business and real estate, automobiles, education, and other loans. Informal credit refers to the borrowing or lending from relatives, friends, and nongovernmental credit organizations based on the production and management of agriculture, industry and commerce, real estate, automobile, education, etc. CHFS questionnaire on informal credit questions includes the following: in addition to bank or credit union loans, is your family still have informal credits not paid off due to industrial and commercial production? How much money was borrowed? Apart from bank or credit union loans, do you currently have outstanding informal loans from your family due to the purchase of your car or children’s education? How much money was borrowed?.

3.2.2. Social Networks. Lin measured social networks with the heterogeneity of members in a social network, the most typical resources in the network, the roof of the network, and the net difference [30]. Gui and Huang obtained seven indicators of social networks based on data analysis, which are local social network, community trust, community belonging, voluntarism, community cohesion, reciprocity and general trust, and nonlocal social interaction [31]. According to previous literature, the indicators of social networks can be summarized as follows: community activity participation and community prestige; gift expenditure, gift income, gift exchange, and communication cost; the sum of cash and noncash income and expenditure on holidays and events; the number of relatives and friends visiting during the spring festival; the frequency to become the host through the “face-saving mechanism” principle; the number of relatives and friends who work in the government or in the city; the number of brothers and sisters and whether they are registered in this municipality; the four ways migrant workers looking for work (government organization,

nongovernmental organization, introduction of relatives and friends, and spontaneous search) [27, 32, 33]. These measurements of social networks are novel and have taken the possible endogenous problems into account, but some of them are too complicated to calculate; some of them are simply quantitative indicators that are not appropriate. Based on the available survey data and previous similar literature, this article measures social networks by gift-money exchange, which is the sum of gift-money expenditure and gift-money income. In addition, according to the definition of emotional network and instrumental network, gift-money exchange with parents, children, and other relatives is regarded as emotional networks, while gift-money exchange with classmates and friends is regarded as instrumental networks.

3.2.3. Control Variables. Referring to studies such as Attanasio et al. and Wan et al., this article selects the household head’s gender, age, physical health, education, marital status, political outlook, risk attitude, work status, family size, household income, and the provinces as control variables. To maintain the characteristics of the data and avoid bias, we take the logarithm of the original data. The definition of variables and descriptive statistics are shown in Table 2.

In addition, the statistics on the specific behavior of rural household credit (Figure 2) show that in the survey samples, residents’ funding needs are often met by informal credit, with a proportion of 82.4%, and only 17.6% of households get funding from banks and other formal financial institutions. This is consistent with reality. Based on the data of fixed observation points in rural areas, we find that only less than 20% of rural households borrow from formal financial institutions. The reasons may be that for most rural households, they lack enough collateral or credit records; therefore, formal financial institutions usually cannot estimate the possibility of their default. As a result, information asymmetry leads to high transaction costs, and formal financial institutions are reluctant to lend to these households. At the same time, in informal credit behavior, 65.1% of households borrow funds and the leftover is informal lending behavior. This may be because, on the one hand, rural households have low income and usually need to borrow money to meet consumption expenditure or production activities; on the other hand, it may be because our statistics focus on borrowing activities rather than lending activities, so the data show that in informal credit behavior, borrowing activities are the main ones.

In terms of the amount of money spent on maintaining household social networks (Figure 3(a)), most households spend between 1,000 yuan and 5,000 yuan with a proportion of 45.99%. In terms of the amount of money spent on maintaining emotional and instrumental networks (Figures 3(b) and 3(c)), most households spent between 1,000 and 5,000 yuan on maintaining emotional networks, while most spent between 5,000 and 10,000 yuan on instrumental networks. Overall, the latter is more expensive to maintain.

TABLE 2: Definition of variables and descriptive statistics.

Variables	Definition	Mean	Standard deviation	Min	Max
Social network	The sum of gift-money expenditure and income	7.906	1.339	0	12.612
Emotional network	The exchange of gift-money between parents, children, and other relatives	0.068	0.794	0	11.617
Instrumental network	The exchange of gift-money between peers, work partners and other friends	5.937	3.602	0	12.612
Formal credit	The actual amount borrowed from banks and other financial institutions	0.965	3.144	0	17.727
Informal credit	The actual amount borrowed from relatives, friends, and informal institutions	4.171	5.104	0	16.705
Household credit behavior	Whether to extend credit: yes = 1, no = 0; Credit amount: the sum of formal credit and informal credit	4.665	5.277	0	17.727
Gender	Male = 1, female = 0	0.471	0.499	0	1
Age	Survey year minus birth year	37.845	12.901	16	60
Health	Good = 1, bad = 0	0.843	0.363	0	1
Education	Level of education	3.110	1.436	1	9
Marriage	Married = 1, others = 0	0.754	0.430	0	1
Job	Employed = 1, others = 0	0.688	0.463	0	1
Politics	Party of CPC = 1, others = 0	0.020	0.142	0	1
Risk attitude	Risk aversion = 0, risk neutral = 1, risk preference = 2	0.416	0.676	0	2
Family size	Number of family members	1.794	0.406	1	3
Family income	Total household income	10.411	1.359	0	15.424

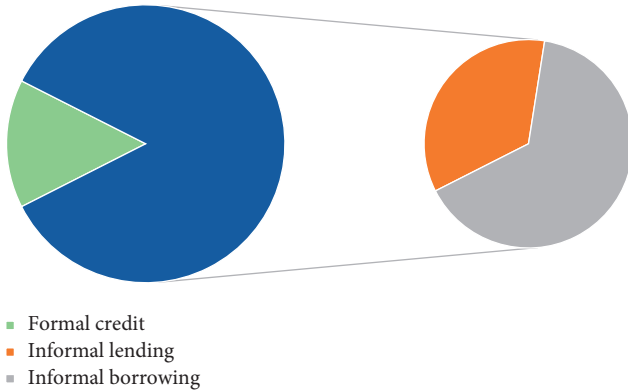


FIGURE 2: Distribution of different types of rural household credit behavior.

3.3. Common Method

3.3.1. Probit Model. The probit model is used to analyze the effect of social networks on whether rural households participate in credit behavior. The model is set as follows:

$$\text{Prob}(y_D = 1) = \alpha_1 \text{social_network} + \beta_1 X_i + \mu_i, \quad (6)$$

where X_i is control variables, μ_i is the residual, $\mu_i \sim N(0, \sigma^2)$, y_D is the dumb variable of the credit behavior, $y_D = 1$ indicates that households credit behavior occurs, and $y_D = 0$ means no households credit behavior occurs.

3.3.2. Panel Data Model

$$y_{ir} = \alpha_{1ir} \text{social_network}_{ir} + \beta_{1ir} X_{ir} + \gamma_r + \mu_{ir}. \quad (7)$$

Y_{ir} denotes household credit behavior, including total credit scale, formal credit scale, and informal credit scale. i represents households, r represents the province, and γ_{ir} is the fixed effect of provincial classification.

3.3.3. Two-Stage Least Squares. In addition, the variables of social networks may be endogenous, which may come from two aspects: on the one hand, households participate in formal credit or informal credit, which may lead to changes in social networks; for example, the increase in demand for family participation in formal and informal credit will spawn more exchange of gifts and social contacts, thus expanding its social network resources. On the other hand, social networks and household formal and informal credit practices and amounts may be influenced by factors, such as local cultural background and customs, which are not observable. Therefore, a key problem to be dealt with in this article is the endogeneity of social networks. After repeated tests, this article holds that communication cost can be used as an instrumental variable for it represents the level of communication with others and can subtly increase their own social network resources, while it has no close relationship with variables in residual. The model is as follows:

$$\begin{aligned} y_{ir} &= \alpha_{1ir} \text{social_networkIV}_{ir} + \beta_{1ir} X_{ir} + \gamma_r + \mu_{ir}, \\ \text{social_networkIV}_{ir} &= \alpha_{2ir} \text{social_network}_{ir} + \beta_{2ir} X_{ir} + \gamma_r + \mu_{ir}, \end{aligned} \quad (8)$$



FIGURE 3: The money spent on maintaining social networks. (a) Social networks. (b) Emotional network. (c) Instrumental network.

where *Social_networkIV* is the instrumental variable of *Social_network*.

4. Results and Discussion

4.1. *Main Results.* The probit model is used to test the relationship between social networks empirically and

whether rural households take credit or not (Table 3 (1)). The results show that when other conditions remain unchanged, for every 1% increase in rural household social network relationship maintenance expenditure, the probability of households choosing credit behavior increases by 0.489%; that is, social network relationships do

TABLE 3: Impact of social networks on rural household credit behavior.

	(1) Whether credit or not	(2) Credit scale	(3) Credit scale
Social networks	0.489*** (0.072)	0.596*** (0.075)	0.684*** (0.099)
Education	-0.046* (0.019)	0.085 (0.078)	-0.199* (0.101)
Job	0.093 (0.053)	0.465* (0.223)	0.442 (0.252)
Gender	0.086 (0.047)	0.200 (0.199)	0.460* (0.231)
Health	-0.355*** (0.067)	-0.280*** (0.028)	-0.661** (0.315)
Age	-0.015*** (0.002)	-0.045*** (0.010)	-0.075*** (0.012)
Family size	0.091 (0.057)	0.307 (0.249)	0.558* (0.277)
Marriage	0.020 (0.067)	0.554* (0.273)	0.141 (0.325)
Politics	0.103 (0.164)	1.261 (0.682)	0.676 (0.788)
Family income	-0.117*** (0.027)	0.063 (0.081)	-0.510** (0.161)
Risk attitude	0.034 (0.034)	0.233 (0.141)	0.121 (0.162)
Constant	-1.962*** (0.391)	0.564 (1.119)	-7.720** (2.391)
<i>N</i>	3037	3037	3037

Note: standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

help promote household credit behavior. Meanwhile, from the empirical results of social networks on the scale of rural household credit (Table 3 (2)), social networks also have a significant positive impact on the scale of household credit. For every 1% increase in social network maintenance expenditure, the household credit scale increases by 0.596%.

Due to the possible endogenous social network variables, 2SLS is used to estimate instrumental variable regression (Table 3 (3)). In the Durbin Wu Hausman endogeneity test, the D-Wu Hausman statistic is 3.782, and it is significant at the level of 10%, which indicates that there is endogeneity in the variable of social networks. The communication cost is selected as the instrumental variable, and the Wald F value is 60.616, which means the instrumental variable has passed the validity test. The results show that when communication fees are used as social network instrumental variables; for every 1% increase in communication fees, the scale of household credit increases by 0.684%. In summary, Hypothesis 1 has been verified, and social networks have a significant positive impact on rural household credit behavior.

In addition to the main variables, most of the control variables also show good statistical characteristics. Educational experience has a significant negative impact on household credit behavior. Educational experience increases one level, the likelihood of households choosing to use credit

increase by 4.6% and household credit scale by 19.9%; health status, age, and total household income also have significant negative impacts. Compared with residents with good health, households with average or poor health are more likely to credit and borrow higher amounts. This may be due to the relatively less income of residents with average or poor health [23]. Their families need to pay a higher amount for health; the older the person is, the less likely it is to borrow and the smaller the amount of borrowing is, which may be because the older they are, the less they can work for the remaining years, and their repayment ability is limited or they may not be able to meet the credit conditions; the higher the total household income is, the less they will choose to borrow and the lower the borrowing amount is. This is in line with reality: under normal circumstances, when family income can cover family expenditures, there is no need for families to make additional borrowing.

4.2. Results of Social Network on Formal Credit and Informal Credit. As the characteristics and requirements of formal credit and informal credit are different, the influence of social networks on them may also be different. Therefore, this article makes an empirical test on the relationship between social networks and formal credit or informal credit; the results are in Table 4. Columns (1) and (2) are estimated by panel fixed effect, and columns (3) and (4) are estimated by 2SLS.

The results of Table 4 (1) and (3) show that social networks have a significant positive impact on the formal credit behavior of rural households. The panel estimation results show that, under other unchanged conditions, for every 1% increase in social network maintenance expenditures, the amount of formal credit will increase by 0.173% (Table 4 (1)). The results of 2SLS (Table 4 (3)) are consistent with the panel fixed effect, which shows the coefficient is 0.980. The coefficient of the former increases nearly five times of the latter, which means that the endogeneity of social networks is worth considering. Educational experience, marital status, and family income have a significant positive impact on the amount of formal family credit: for each level of education experience, the amount of household formal credit increases by 33%. This may be because households with high academic qualifications are more likely to meet formal financial requirements [18]; compared to unmarried families, married families can obtain higher financial loans, and the amount is 52.9% higher than that of unmarried families. This may be because married families have more stable income and stronger repayment ability; for every 1% increase in family income, the credit amount increases by 0.132%. This is also because families with higher income have stronger repayment ability and easier access to loans [17].

The results of Table 4 (2) and (4) show that social networks also have a significant positive impact on informal credit. Panel estimation shows that for every 1% increase in social networks maintenance spending, the amount of informal credit increases by 0.531% (Table 4 (2)). The results of 2SLS (Table 4 (4)) are consistent with the panel fixed effect, which shows the coefficient is 0.414. The coefficient of the

TABLE 4: Impact of social networks on formal and informal credit behavior.

	(1) Formal credit	(2) Informal credit	(3) Formal credit	(4) Informal credit
Social networks	0.173*** (0.048)	0.531*** (0.072)	0.980** (0.298)	0.414*** (0.473)
Education	0.330*** (0.050)	-0.107 (0.075)	0.243*** (0.060)	-0.365*** (0.096)
Job	-0.223 (0.143)	0.669** (0.214)	-0.226 (0.151)	0.623** (0.239)
Gender	0.107 (0.127)	0.155 (0.191)	0.217 (0.138)	0.366 (0.219)
Health	-0.120 (0.180)	-0.237*** (0.027)	-0.272 (0.189)	-0.555* (0.299)
Age	-0.004 (0.007)	-0.043*** (0.010)	-0.013 (0.007)	-0.070*** (0.012)
Family size	0.081 (0.159)	0.444 (0.239)	0.149 (0.165)	0.670* (0.263)
Marriage	0.529** (0.175)	0.310 (0.262)	0.377 (0.193)	-0.051 (0.308)
Politics	-0.278 (0.438)	1.329* (0.655)	-0.608 (0.471)	0.786 (0.748)
Family income	0.132* (0.052)	-0.039 (0.078)	-0.085 (0.096)	-0.564*** (0.153)
Risk attitude	-0.025 (0.091)	0.192 (0.136)	-0.078 (0.096)	0.099 (0.153)
Constant	-2.816*** (0.717)	1.831 (1.076)	-6.169*** (1.428)	-5.556* (2.269)
<i>N</i>	3037	3037	3037	3037

Note: standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

former is close to the latter, which means the endogeneity of social networks in informal credit is not obvious. Work status, health status, and age also have significant impacts on the amount of informal credit: compared with nonworking residents, the amount of informal credit obtained by working residents is 66.9% higher. This is because working residents have income and a source of repayment [18]; the improvement of health status reduces the amount of informal credit. The amount of loans borrowed by residents with good health is 23.7% less than the amount borrowed by residents with average or poor health. This may be because residents in good conditions have more opportunities to find a job, and their medical expenses are less, so their loan amount needs are less [19]; the older the age, the smaller the amount of informal credit, and the amount of borrowing decreases by 4.3% for each year of increase. This may be because as residents get older, their ability to repay gradually decreases, so the number of loans borrowed decreases.

In addition, comparing the impact of social networks on formal credit and informal credit (Figure 4), it can be found that social networks have a stronger impact on informal credit no matter it is linear fitting (Figure 4(a)) or curve fitting (Figure 4(b)). Hypothesis 2 has been verified. This is consistent with the research of scholars such as He et al. [15] and Yang et al. [1]: on the one hand, since formal credit requires mortgage or guarantee, although social networks can be used as a kind of "hidden" guarantee and supervision mechanism; it is incompatible with practical material. There are still gaps in guarantees, and rural households may still be restricted by formal credit rationing and cannot get loans; on

the other hand, plentiful social networks mean more relatives, friends, and closer contacts between social network members; this can effectively avoid moral hazard and adverse selection [3]. Moreover, the rapid spread of public opinion in social networks can also restrain debtors from complying with loan contracts [33]. Therefore, the enrichment of social networks can effectively strengthen rural informal credit.

4.3. Results of Emotional Network and Instrumental Network on Household Credit Behavior. In social networks, differences in connection strength between individuals form different types of networks, strong connections form emotional networks, and weak connections form instrumental networks. This section empirically examines the relationship between different types of networks and household credit behavior. The results are shown in Table 5.

Columns (1) and (2) in the table are the influence of emotional networks and instrumental networks on the total household credit amount. The results show that emotional networks and instrumental networks have significant positive impacts on the amount of household credit with coefficients of 0.423 and 0.169, respectively, indicating that with other unchanged conditions, each time emotional networks maintenance expenditure increases by 1%, the credit amount increased by 0.423%, and for every 1% increase in instrumental networks maintenance expenditure, the credit amount increased by 0.169%.

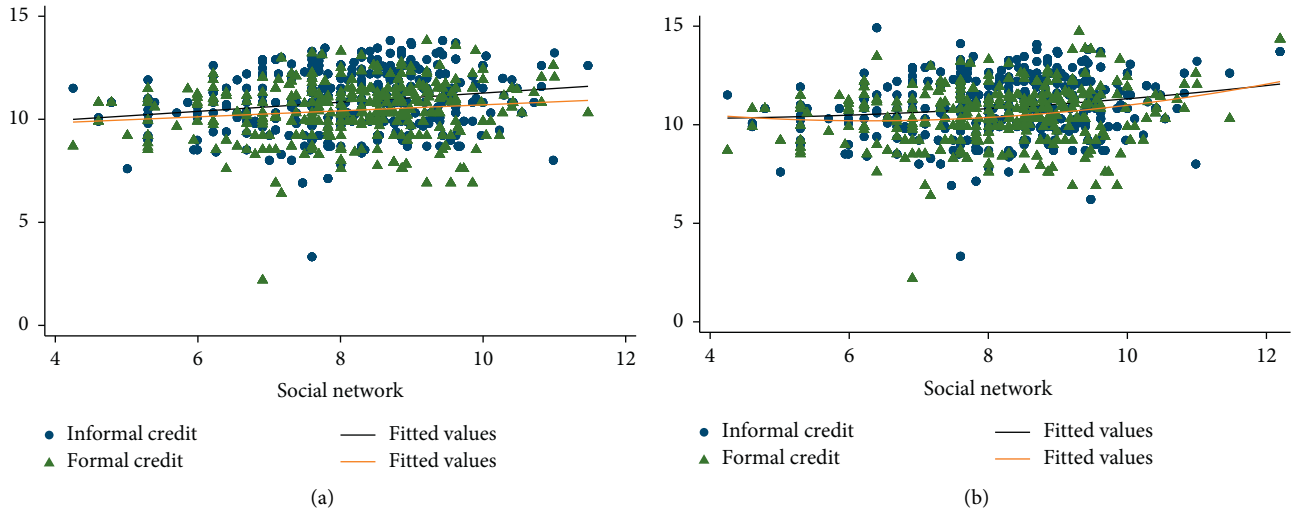


FIGURE 4: Slope graph of social networks on formal and informal credit behavior. (a) Linear fitting; (b) curve fitting.

TABLE 5: Impact of emotional and instrumental social networks on rural household credit behavior.

	(1) Household credit	(2) Household credit	(3) Formal credit	(4) Formal credit	(5) Informal credit	(6) Informal credit
Emotional networks	0.423*** (0.095)		0.229*** (0.059)		0.278** (0.091)	
Instrumental networks		0.169*** (0.024)		0.061*** (0.015)		0.142*** (0.023)
Education	0.111 (0.068)	0.088 (0.068)	0.292*** (0.042)	0.287*** (0.042)	-0.058 (0.065)	-0.082 (0.065)
Job	0.480* (0.195)	0.460* (0.195)	-0.115 (0.120)	-0.132 (0.120)	0.628*** (0.187)	0.615** (0.187)
Gender	-0.062 (0.174)	-0.050 (0.173)	-0.053 (0.107)	-0.057 (0.107)	-0.005 (0.167)	0.009 (0.166)
Health	-1.337*** (0.244)	-1.331*** (0.243)	-0.048 (0.150)	-0.047 (0.150)	-1.416*** (0.234)	-1.410*** (0.233)
Age	-0.038*** (0.009)	-0.042*** (0.009)	0.003 (0.005)	0.001 (0.005)	-0.040*** (0.009)	-0.043*** (0.009)
Family size	0.230 (0.216)	0.217 (0.215)	0.026 (0.132)	0.023 (0.132)	0.317 (0.207)	0.305 (0.206)
Marriage	0.552* (0.233)	0.534* (0.232)	0.294* (0.143)	0.311* (0.143)	0.422 (0.224)	0.392 (0.223)
Politics	1.446* (0.572)	1.662** (0.569)	0.089 (0.353)	0.188 (0.352)	1.199* (0.549)	1.358* (0.546)
Family income	0.170** (0.065)	0.119 (0.066)	0.185*** (0.040)	0.170*** (0.041)	0.065 (0.063)	0.021 (0.063)
Risk attitude	0.119 (0.123)	0.135 (0.123)	-0.033 (0.076)	-0.032 (0.076)	0.112 (0.118)	0.126 (0.118)
Constant	3.859*** (0.892)	3.664*** (0.889)	-2.032*** (0.549)	-2.139*** (0.549)	4.879*** (0.856)	4.737*** (0.853)
N	3984	3974	4001	3991	3984	3974

Note: standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Columns (3) and (4) are the influence of emotional networks and instrumental networks on the amount of household formal credit. The results show that emotional networks and instrumental networks also have significant positive impacts on the amount of household formal credit, with coefficients of 0.229 and 0.061, respectively. Columns

(5) and (6) are the influence of emotional networks and instrumental networks on the amount of household informal credit. The results show that emotional networks and instrumental networks have significant positive impacts on the amount of household informal credit, with coefficients of 0.278 and 0.142, respectively.

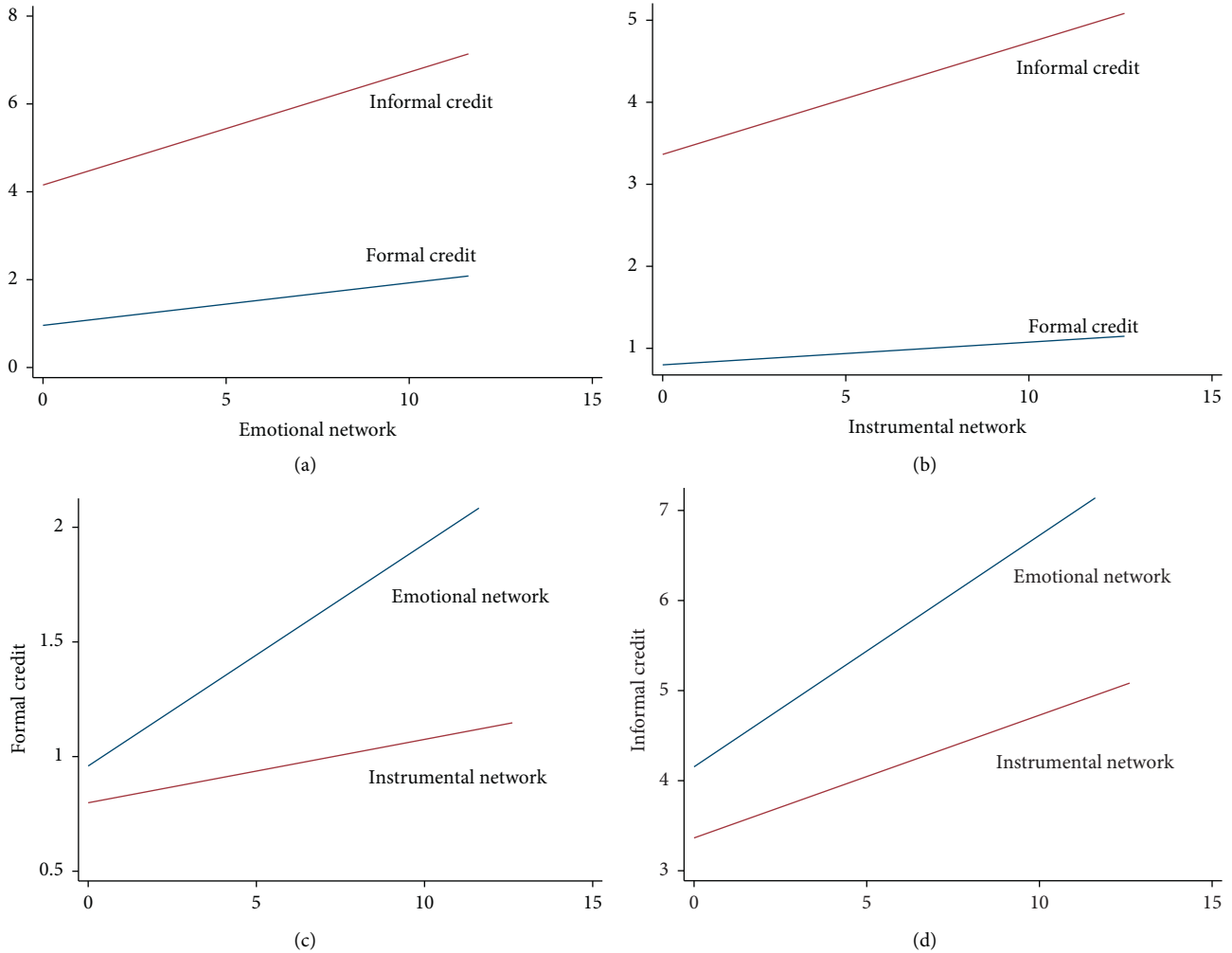


FIGURE 5: Slope graph of social networks effect on credit behavior. (a) Emotional networks and (b) instrumental networks effect on formal and informal credit. Emotional networks and instrumental networks effect on (c) formal and (d) informal credit.

Summarizing the impact of different types of networks on distinctive credit amounts (Figure 5), we find that whether it is an emotional network or an instrumental network, it has a significant positive impact on household credit behavior. Hypothesis 1 has been verified again. In addition, compared with formal credit behavior, emotional networks have a stronger positive impact on informal credit behavior (Figure 5(a)), Hypothesis 3 is verified; compared with formal credit, the positive influence of instrumental networks on informal credit behavior is stronger (Figure 5(b)), Hypothesis 4 is not verified. This may be because in social networks, households are more likely to borrow directly from individuals in the network and seldom make formal credit through individual relationships [20]. In addition, even if there are credit personnel from financial institutions in household instrumental networks, households must meet the formal credit conditions, which are not met by the invisible guarantee mechanism of social networks. Therefore, in general, the tool-based network has a stronger influence on informal credit. In addition, from Figures 5(c) and 5(d), we find that whether in formal credit or informal credit, the

positive effect of emotional networks is stronger than that of instrumental networks.

5. Conclusions

This research discusses the impact of social networks on rural household credit behavior and explores the heterogeneity of the impact from differential networks and distinct credit behavior. The article first analyzes the influence of social networks on household credit behavior decisions based on game theory and further proposes hypotheses and conducts empirical tests. The results show that social networks can actively promote rural household choice of credit behavior and increase their credit amount. In addition, compared with formal household credit behavior, social networks have a stronger positive role in promoting informal credit behavior. This conclusion has also been confirmed in the influence of emotional networks and instrumental networks on household credit behavior. That is, both emotional and instrumental networks have a stronger influence on informal credit behavior than formal credit. At the same time, we also find that the positive effect of

emotional networks is stronger than that of instrumental networks on either formal credit or informal credit. Therefore, we suggest that in the process of giving play to the role of rural financial credit, social networks should be actively used to transmit and supervise, especially to strengthen the establishment of emotional networks, which has an important practical effect on the development of formal financial credit and informal credit. Specifically, first, the implementation of finance credit should pay attention to the characteristics of the village acquaintance society in order to solve the problem of high transaction cost caused by information asymmetry. Secondly, when analyzing the effect of inclusive financial projects, the dynamic impact of external policies on the overall equilibrium should be considered for the factors that made mutual fund projects successful (such as the social network of villages) may change with the implementation of the projects, and then have a global impact on the policy effect. Third, by combining the formal system with the informal system, the government should create a more complete credit market environment and ease the credit constraints. Especially for farmers who lack collateral, it should fully consider the guarantee function of social networks with geographical and kinship relationship so as to solve the problem of loan difficulty in rural areas.

Although we have conducted an in-depth analysis of the relationship between social networks and rural household credit behavior, this study also has some shortcomings. First, this study mainly uses cross-sectional sample data for analysis from an individual perspective, and it is difficult to obtain the dynamic process of variables in the individual development process. Secondly, this research only studies the direct relationship between social networks and household credit behaviors. The mechanism of how social networks affect household credit behaviors has not yet been explored. Future research could analyze the influence mechanism in detail and conduct empirical tests on it to understand more specifically the extent and path of the influence; finally, this study only considers rural samples and does not consider the influence of social networks on urban household credit behavior. Due to the dual development of rural households and urban households, their credit needs are different. Future research could also conduct heterogeneous research on the relationship between urban and rural family social networks and credit behavior.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (71974157 and 72003113).

References

- [1] R. D. Yang, B. K. Chen, and S. E. Zhu, "Research on Farmers' private credit demand behavior from the perspective of social network," *Economic Research Journal*, vol. 11, pp. 116–129, 2011.
- [2] F. Hu and Y. Y. Chen, "Social networks and farmers' lending behavior: evidence from the CFPS," *Financial Research*, vol. 12, pp. 178–192, 2012.
- [3] J. E. Stiglitz and A. Weiss, "Credit rationing in markets with imperfect information," *American Economic Review*, vol. 71, no. 3, pp. 393–410, 1981.
- [4] B. Wellman, "Physical place and cyberspace: the rise of personalized networking," *International Journal of Urban & Regional Research*, vol. 25, no. 2, pp. 227–252, 2015.
- [5] J. E. Stiglitz, "Peer monitoring and credit markets," *The World Bank Economic Review*, vol. 4, no. 3, pp. 351–366, 1990.
- [6] N. H. Nie, "Sociability, interpersonal relations, and the Internet: reconciling conflicting findings," *American Behavioral Scientist*, vol. 45, no. 3, pp. 420–435, 2001.
- [7] D. S. Karlan, "Social connections and group banking," *Economic Journal*, vol. 117, no. 517, pp. 52–84, 2007.
- [8] M. Shoji, K. Aoyagi, R. Kasahara, Y. Sawada, and M. Ueyama, "Social capital formation and credit access: evidence from Sri Lanka," *World Development*, vol. 40, no. 12, pp. 2522–2536, 2012.
- [9] W. Zhou, "Brothers, household financial markets and savings rate in China," *Journal of Development Economics*, vol. 111, pp. 34–47, 2014.
- [10] C. Zhou and H. J. Yue, "A study on Chinese farmers' family credit behavior from the perspective of social network," *Journal of Xiangtan University (Philosophy and Social Science Edition)*, vol. 5, pp. 77–82, 2017.
- [11] H. L. Qin, C. W. Li, and J. L. Wan, "Social capital, farmer heterogeneity and credit behavior-measurement analysis and empirical test based on CFPS data," *Finance and Economy*, vol. 1, pp. 33–40, 2019.
- [12] X. Chen and S. Chen, "The quality of social capital and the availability of farmers' loans: an analysis based on professional reputation," *Jiangxi Social Sciences*, vol. 38, no. 5, pp. 218–226, 2018.
- [13] G. Tenzin, K. Otsuka, and K. Natsuda, "Can social capital reduce poverty? a study of rural households in eastern Bhutan," *Asian Economic Journal*, vol. 29, no. 3, pp. 243–264, 2015.
- [14] C. Okten, "Social networks and credit access in Indonesia," *World Development*, vol. 32, no. 7, pp. 1225–1246, 2004.
- [15] G. W. He, J. He, and P. Guo, "Further discussion on farmers' credit demand and credit availability," *Agricultural Economic Problems*, vol. 2, pp. 38–49, 2018.
- [16] Z. Y. Xu and H. Yang, "Analysis of farmers' credit behavior tendency and its influencing factors-based on the survey of 1664 farmers in 11 western provinces," *China Soft Science Magazine*, vol. 3, pp. 45–56, 2014.
- [17] B. Wydick, H. K. Hayes, and S. H. Kempf, "Social networks, neighborhood effects, and credit access: evidence from rural Guatemala," *World Development*, vol. 39, no. 6, pp. 974–982, 2011.
- [18] X. Gine, "Access to capital in rural Thailand: an estimated model of formal vs. informal credit," *Journal of Development Economics*, vol. 96, no. 1, pp. 16–29, 2011.
- [19] P. D. Khoi, C. Gan, G. V. Nartea, and D. A. Cohen, "Formal and informal rural credit in the Mekong River Delta of

- Vietnam: interaction and accessibility,” *Journal of Asian Economics*, vol. 26, pp. 1–13, 2013.
- [20] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, “Exploiting implicit influence from information propagation for social recommendation,” *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [21] T. M. Chau, E. C. Gan, and B. D. Hu, “Credit constraints and their impact on farm household welfare: evidence from Vietnam’s North Central Coast region,” *International Journal of Social Economics*, vol. 43, no. 8, pp. 782–803, 2016.
- [22] O. Atanasio, A. Barr, and J. C. Cardenas, “Risk pooling, risk preferences and social network,” *American Economic Journal: Applied Economics*, vol. 4, no. 2, pp. 134–167, 2012.
- [23] Y. Wang, M. J. Jinhong, and Z. Yuan, “Social network and credit market: evidence from China,” *Journal of Financial Research*, vol. 412, no. 10, pp. 116–132, 2014.
- [24] H. C. Xia, Z. X. Yan, and Z. Liang, “Impacts of social capital on farmers’ credit behavior: cases of Zhangye city, Gannan Tibetan autonomous Prefecture and Linxia Hui autonomous Prefecture in Gansu province,” *Arid Land Geography*, vol. 37, no. 4, pp. 831–837, 2014.
- [25] R. Bathish, D. Best, and M. Savic, “Is it me or should my friends take the credit? The role of social networks and social identity in recovery from addiction,” *Journal of Applied Social Psychology*, vol. 47, no. 1, pp. 126–141, 2017.
- [26] J. Lin, B. Y. Wu, and Z. D. Li, “The efficient social networks in household credit: friendship or kinship?” *Journal of Financial Research*, vol. 427, no. 1, pp. 130–144, 2016.
- [27] S. Wang, “Instrumental networking and social network building: how horizontal networking and upward networking create social capital,” *Acta Psychologica Sinica*, vol. 49, no. 1, pp. 116–127, 2017.
- [28] S. Roberts and R. Dunbar, “Communication in social networks: effects of kinship, network size, and emotional closeness,” *Personal Relationships*, vol. 18, no. 3, pp. 31–45, 2011.
- [29] F. Xiong, X. M. Wang, and S. R. Pan, “Social recommendation with evolutionary opinion dynamics,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [30] N. Lin, “Social networks and status attainment,” *Annual Review of Sociology*, vol. 25, no. 1, pp. 467–487, 1999.
- [31] Y. Gui and R. G. Huang, “Measurement of community social capital: a study based on empirical data,” *Sociological Research*, vol. 3, pp. 122–142, 2008.
- [32] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, “Bayesian personalized ranking based on multiple-layer neighborhoods,” *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [33] J. L. Wan, C. W. Li, and H. L. Qin, “Binary finance, social network and household lending—an empirical test based on repeated game model and CFPS data,” *Journal of Financial Development Research*, vol. 7, pp. 42–49, 2018.

Research Article

Efficient Data Transmission for Community Detection Algorithm Based on Node Similarity in Opportunistic Social Networks

Aizimaiti Xiaokaiti,^{1,2} Yurong Qian ^{1,2} and Jia Wu ³

¹Software College, Xinjiang University, Urumqi 830000, China

²Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830046, China

³School of Computer Science and Engineering, Central South University, Changsha 410083, China

Correspondence should be addressed to Yurong Qian; qyr@xju.edu.cn and Jia Wu; jiawu5110@163.com

Received 11 March 2021; Revised 13 April 2021; Accepted 17 April 2021; Published 29 May 2021

Academic Editor: Fei Xiong

Copyright © 2021 Aizimaiti Xiaokaiti 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 the rapid development of 5G era, the number of messages on the network has increased sharply. The traditional opportunistic networks algorithm has some shortcomings in processing data. Most traditional algorithms divide the nodes into communities and then perform data transmission according to the divided communities. However, these algorithms do not consider enough nodes' characteristics in the communities' division, and two positively related nodes may divide into different communities. Therefore, how to accurately divide the community is still a challenging issue. We propose an efficient data transmission strategy for community detection (EDCD) algorithm. When dividing communities, we use mobile edge computing to combine network topology attributes with social attributes. When forwarding the message, we select optimal relay node as transmission according to the coefficients of channels. In the simulation experiment, we analyze the efficiency of the algorithm in four different real datasets. The results show that the algorithm has good performance in terms of delivery ratio and routing overhead.

1. Introduction

With the booming of information technology and the popularization of wireless network equipment [1], people have a growing demand for the network. As a fresh type of self-organizing network [2], an opportunistic social network has attracted researchers' attention [3]. There is no complete end-to-end path between nodes in opportunistic social networks [4]; it uses the encounter opportunities brought by node movement to communicate hop by hop [5]. At present, opportunistic social network has widespread use in various fields, such as mobile phones [6], handheld electronic devices [7], vehicular networks with mobile intelligent devices on the road [8], wildlife tracking [9], and network transmission in remote areas [10].

The traditional social network method to deal with data transmission faces significant challenges [11], which will become an obstacle to the information exchange and sharing [12]. To enhance data transmission in a 5G wireless network

[13], we should design a more convenient model to achieve data forwarding flexibly [14]. The user terminal equipment needs to transmit a large amount of data and needs to calculate these intensive tasks [15]. To enhance wireless devices' computer ability, mobile edge computing (MEC) is proposed [16–18]. Because the mobile edge server locates at the edge of the wireless network and closer to the users, it can efficiently provide the surrounding users' services and integrate the concept of opportunistic social networks into mobile edge computing, to reduce the consumption of source nodes [19].

However, each node has many social attributes [20]. They represent the relationship among different users, and the connections between nodes in the same community are more than closer [21]. So, the network nodes can be divided into communities by their different attributes to improve the algorithm's performance [22]. The existing algorithms do not fully consider nodes' characteristics, so there is a large space for improvement in community detection accuracy

and efficiency [23]. That is why it is necessary to propose an efficient community detection algorithm.

Opportunistic social network uses the strategy of “storing-carrying-forwarding” to handle the energy consumption problem in the data transmission process [24]. Messages are forwarded through encounter opportunities produced by node movement. In this paper, the network topology attributes and social attributes are used to measure the similarity between nodes, and the hierarchical clustering method effectively divides the community [25]. In the process of data transmission, if the mobile device does not have a suitable transmission target, the message will occupy a lot of cache, and the data transmission in the community is likely to wait a long time and cause the delay in transmission [26]. After dividing the community, we need to further establish the weight distribution between nodes and community to reduce the time complexity and overhead cost and construct a set of candidate relay nodes based on the relationship between information forwarders and adjacent nodes. From the perspective of minimizing bit error rate, the channel coefficients of the two channels from the source node to the relay node and the relay node to the destination node are analyzed. This must select the optimal relay node from the set of candidate relay nodes as transmission. In summary, we propose an efficient data transmission strategy for community detection in opportunistic social network using mobile edge computing combined with network topology and social attributes. The transmission strategy is divided into two periods: the initialization period and the routing period.

The contributions of this research study are as follows:

- (1) Initialization period: using network topology attributes and social attributes to measure the similarity between nodes, a community detection algorithm is proposed through hierarchical clustering.
- (2) Routing period: based on the relationship between the message forwarder and the adjacent nodes, a set of candidate relay nodes is constructed. By analyzing the channel coefficients of the source node to the relay node and the relay node to the destination node, a method for selecting the optimal relay node is proposed.
- (3) Simulation results show that the algorithm EDCD proposed in this paper has good performance such as delivery ratio, routing overhead, and average end-to-end delay in different real datasets.

2. Related Works

Many researchers have conducted research on routing and forwarding algorithms in opportunistic social networks and proposed very effective approaches in different application scenarios in recent years. Many research methods have focused on algorithm research. Routing algorithms can be roughly divided into two sorts: existing social-ignorant algorithms and existing social-aware algorithms [27].

Existing social-ignorant algorithms mean that social message relating to nodes will not make adaptable messaging

decisions in the process of data transmission. Vahdat and Becker [28] proposed the epidemic routing algorithm. Epidemic algorithm is essentially a flooding algorithm, and each node forwards information to all its neighbors. However, there are a lot of message copies in the network, which will consume many network resources. Sisodiya et al. [29] proposed a flood routing algorithm, that is, spray and wait algorithm, which divides the information forwarding process into two steps. The first step is to copy the message and the transmission process is in the second step. It can easily lead to ultratransmission delay and data redundancy.

Sharma et al. [30] proposed a routing protocol named MLProph, which uses machine learning (ML) algorithms, namely, decision trees and neural networks, to determine the probability of successful message delivery, but this algorithm has great limitations. Tang et al. [31] proposed a scheme based on reinforcement learning (RL), which can apply to opportunistic routing transmissions that require high reliability and low latency. However, this opportunistic routing scheme can only be used for specific scenarios and is not for all networks. Wu et al. [32] proposed the algorithm that adjusts the cache by analyzing the importance of message propagation. This algorithm has a small routing overhead, but to avoid deleting the cached data, the data shares by adjacent nodes will cause data redundancy.

Social-aware algorithms refer to the social relationship between nodes to measure the transmission relevance between nodes. Yan et al. [33] established an effective data transmission strategy (ENPSR), which uses the priority of nodes and social relationships in opportunistic social networks. Obtain the data transmission priority by measuring the social attributes and historical information of the node. Then use the forecast plan to determine the appropriate message delivery decision. Wu and Chen [34] proposed an optimal routing scheme for cooperative nodes based on opportunistic network features. This scheme can use in social networks. By reliability, availability, and weighting factors are used as the weights of human activities to obtain the optimal cooperative node, but the algorithm has a high routing overhead. Drăgan et al. [35] proposed that nodes can be divided into several communities according to their intimacy and the time together. This community detection method does not fully consider all of the nodes in the community.

Zeng et al. [36] proposed a social-based clustering and routing scheme, in which each node selects the nodes with close social relationships to form a local cluster, but this can cause data redundancy issues. Liu et al. [37] proposed an algorithm using node similarity (FCNS) based on fuzzy routing and forwarding. This algorithm has good performance in data transfer ratio and routing overhead but high transmission delay. Niu et al. [38] proposed a predictive and extended routing protocol, which uses Markov chain as a node mobility model to realize the social characteristics of nodes. It does not consider node communication between different places, and nodes just upload and send message in the same place.

Because the abovementioned traditional methods do not fully consider node characteristics and other problems, this

paper proposes a model that combined with the network topology and social attributes to detect community and analyze the channel coefficients of source node to relay node and relay node to destination node to select optimal relay node as information transmission in opportunistic social networks. This model can effectively handle the challenge of improving data transmission and has good performance of low delay and low routing overhead.

3. Model Design

In opportunistic social networks, we can define the topological structure $G = (V, E, w)$, where V is the node of the network and E is the edge set in the network reflecting the relationship between the nodes. $E = \{(m, n) | m \in V, n \in V\}$, m and n are nodes, and w is the weight of the edges of node m and node n . On the basis of the division of the community, we make $C = \{C1, C2, \dots, Cn\}$, which require more edges between vertices in each community subgraph. We consider that there will be differences between nodes and the number of encounters between nodes to weight each edge. This paper proposes to measure the similarity between different nodes in terms of network topology attributes and social attributes. The greater the similarity is between nodes, the more likely they are to belong to the same community.

Firstly, we must reasonably define the similarity between nodes. For a real social network, and considering the network topology, we also need to consider the social attributes between nodes. We must collect the data of the node, and the process is shown in Figure 1. The nodes information collection method is that the base station collects all node information in the area within a period of time. When the node has a transmission task, request the probability table of the source node and the destination node from the base station that has collected the information and use edge computing to transmit decision information to reduce node's workload. Because many communities can usually only share messages based on one or two nodes, there must be enough cache to improve data transmission efficiency. The node requires obtaining the position, speed, and moving direction of itself and the destination node. However, the encounter of nodes in opportunistic social networks is random. Combining the characteristics of node movement to calculate the probability of node encounters, in this paper, PE_{mn} means the probability of nodes m and n meeting in a period of time t , and the node meeting interval time obeys the exponential distribution; then the probability of node m and node n meeting within the sensing range is

$$PE_{mn}(t) = 1 - e^{-\lambda_{mn}(t)}, \quad (1)$$

where m is the source node, n is the destination node, $\lambda_{mn} = (1/\Delta t_{mn})$ is the encounter strength of node m and node n , and Δt_{mn} is the average time between node m and node n :

$$\Delta t_{mn} = \frac{1}{n} \sum_{k=0}^n (t_{mn}^{k+1} - t_{mn}^k), \quad (2)$$

where t_{mn}^k is the time of the k th encounter, and we define $t_{mn}^0 = 0$. In short, combine the formula to get

$$PE_{mn}(t) = 1 - e^{-\lambda_{mn}(t)} = 1 - \exp\left[\frac{-n(t_{\text{init}} - t_0)}{\sum_{k=0}^n (t_{mn}^{k+1} - t_{mn}^k)}\right], \quad (3)$$

where $t_r = t_{\text{init}} - t_0$ is the remaining time to live of the message, t_{init} is the initial time to live of the message, and t_0 is the current time the message has been alive.

Secondly, construct the encounter probability matrix. The number of encounters between nodes to a certain extent only reflects the number of encounters of the node in a period of time.

$$MT = \begin{pmatrix} PE_{11} & \dots & PE_{1n} \\ \vdots & \ddots & \vdots \\ PE_{m1} & \dots & PE_{mn} \end{pmatrix}, \quad (4)$$

Use the number of encounters between nodes to weight each edge. $w = \{w_{mn} | e_{mn} \in E\}$ is the set of edge weights, where w_{mn} is the number of encounters between two nodes. $MT = \{PE_{mn}(t)\}$ represents the $n * n$ encounter matrix of node m and node n in a period of time t .

$$w_{mn} = PE_{mn}(t) \times \sum_{\text{on}}^n w_{m,\text{on}}, \quad (5)$$

where $w_{m,\text{on}}$ represents the number of encounters the node m has met with other nodes within a certain period of time.

In opportunistic social networks, network topology attributes reflect the status of the network. It requires more edges between the vertices in each community subgraph.

- (1) The strength of nodes describes how close the node is to the surrounding network, and the node strength is equal to the degree of the node, that is, the number of neighbor nodes. The defined formula is

$$ND_{(m,n)} = \frac{|NC_m - NC_n|}{|NC_m + NC_n|}, \quad (6)$$

where $ND_{(m,n)}$ is the node connection strength between node m and node n . NC_m is the set of neighbor nodes connected to a node m in current times, and NC_n is the set of neighbor nodes connected to a node n in current times. We have to consider that two nodes may share a set of similar neighbor nodes, so the higher the relationship between them, the higher the probability of data transmission.

- (2) The direct connection strength represents the influence of the direct connection between two nodes. When there is an edge between two nodes, the edge weight measures the strength of the connection between them. We define the sum of the weights of all edges adjacent to node m as $s(m) = \sum_{n \in \theta(m)} w_{mn}$, where $\theta(m)$ is the set of neighbor nodes of m . For any $\theta(m)$, there is a relationship between node m and node n . So the formula for direct connection strength is as follows:

$$DC_{(m,n)} = \frac{w_{mn}}{s(m) + s(n) - w_{mn}}, \quad (7)$$

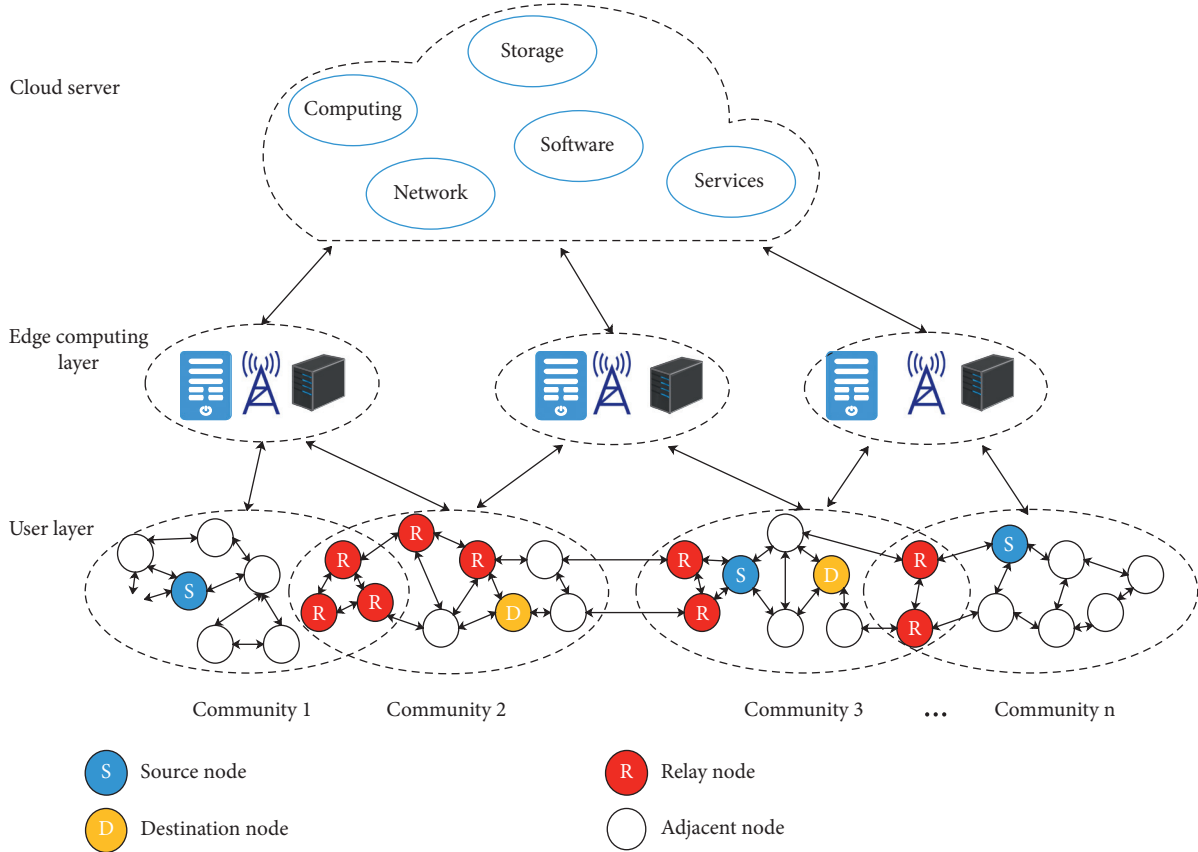


FIGURE 1: Schematic diagram of community node information.

where $DC_{(m,n)}$ is the strength of the direct connection between two nodes and is also the ratio of the weight of the two nodes to the weight of their adjacent edge.

- (3) The indirect connection strength indicates the influence of the indirect connection between two nodes; just as when node m and node n have a common adjacent node p , then node m and node n also have a certain chance to connect. The more adjacent nodes that two nodes have in common, the closer the two nodes are. So the formula for indirect connection strength is as follows:

$$IC_{(m,n)} = \sum_{p \in s(m) \cap s(n)} \frac{w_{mp} + w_{np}}{s(m) + s(n) - w_{mn}}, \quad (8)$$

where w_{mp} and w_{np} represent the connection strength between node m and node n through node p , and the indirect connection strength between nodes is the sum of the strengths of all common neighbor connections. That is to say, the more common adjacent nodes the two nodes have, the greater the indirect connection strength is.

In the network topology attributes, we classify the possible relationships between two nodes into the following four types, where we use $SMD_{tp(m,n)}$ to express topological similarity between node m and node n .

- (a) No direct and no indirect connection:

$$SMD_{tp(m,n)} = \alpha ND_{(m,n)}. \quad (9)$$

- (b) Indirect but no direct connection:

$$SMD_{tp(m,n)} = \eta IC_{(m,n)} + \alpha ND_{(m,n)}. \quad (10)$$

- (c) Direct but no indirect connection:

$$SMD_{tp(m,n)} = \varphi DC_{(m,n)} + \alpha ND_{(m,n)}. \quad (11)$$

- (d) Direct and indirect connection:

$$SMD_{tp(m,n)} = \varphi DC_{(m,n)} + \eta IC_{(m,n)} + \alpha ND_{(m,n)}, \quad (12)$$

where α is the coefficient of the strength of node, φ is the coefficient of the direct connection strength, and η is the coefficient of the indirect connection strength. The higher the topological similarity between nodes, the greater the chance of communication between nodes, which can improve data transmission efficiency.

The social attributes between nodes measure the social similarity between two nodes.

- (1) The geographic relevance of nodes: the node has mobile characteristics; the mobile node's trajectory information is used to analyze the geographic location correlation of the node. The trajectory

information refers to the geographic location information of the sensing area. The sensing area is the area where the node can transmit messages within a certain range. Specifically, in the time period T , if the nodes' geographical locations are close, it means that the probability of node information transmission is high; that is to say, the probability of meeting in the same area will also be increased. The geographical correlation between nodes can be expressed as

$$GD_{(m,n)} = \frac{\sum_{m=1}^k \sum_{n=1}^k A(R_m^r, R_n^r)}{T}, \quad (13)$$

where $GD_{(m,n)}$ is the geographic relevance of nodes, $A(R_m^r, R_n^r)$ represents the similarity function of node m and node n at position A , R_m^r represents rm th trajectory information of node m , and R_n^r represents rn th trajectory information of node n .

$$A(R_m^r, R_n^r) = \max\{E_m^r, E_n^r\} - \min\{Q_m^r, Q_n^r\}, \quad (14)$$

where $\max\{E_m^r, E_n^r\}$ takes the maximum value between E_m^r and E_n^r , E_m^r is the time when node m enters the sensing area for the rm th time, and E_n^r is the time when node n enters the sensing area for the rn th time. $\min\{Q_m^r, Q_n^r\}$ represents take the minimum value between Q_m^r and Q_n^r , Q_m^r is the time when node m quits the sensing area for the rm th time, and Q_n^r is the time when node n quits the sensing area for the rn th time.

- (2) The interesting relevance of nodes: users with common interests will visit the same business. Naturally, mobile users with the same interests will spend more time and energy communicating together. The information transmission between nodes will be carried out between mobile users with the same interest in the time period T . The interesting relevance between nodes can be expressed as

$$IR_{(m,n)} = \frac{\sum_{k=1}^{k_n} T_{m,n}^k}{\sum_{k=1}^{k_n-1} T_{m,otn}^{k-1}}, \quad (15)$$

where $IR_{(m,n)}$ represents the interesting relevance between node m and node n . $T_{m,n}^k$ represents the ratio of time occupied by node m and node n during the k th transmission of information in time period T . $T_{m,otn}^{k-1}$ represents the ratio of the time occupied by node m and other nodes except node n in the k -1th transmission information in time period T .

- (3) The separating time relevance of nodes: two nodes can make a connection and communicate. The average interval between two nodes can be defined as the time interval when two nodes meet each other. If there is no communication for a long time, the relationship between the two nodes is not close enough. Conversely, a shorter separation means that the two nodes are closely related. The separating time relevance of nodes can be expressed as

$$AS_{(m,n)} = \frac{T_{m,n}^k - T_{m,n}^1}{k \times T}, \quad (16)$$

where $AS_{(m,n)}$ represents the separate time relevance of node m and node n to convey information. $T_{m,n}^k$ is the time of the k th transmission of information in the time interval T . $T_{m,n}^1$ is the time of the first transmission of information in the time interval T .

Through the above calculation of social attribute values, we can quantify the relationship between node m and node n . $SR_{(m,n)}$ represents the similarity of social attributes as follows:

$$SR_{(m,n)} = \beta GD_{(m,n)} + \delta IR_{(m,n)} + \rho AS_{(m,n)}, \quad (17)$$

where β is the coefficient of the geographic relevance of nodes, δ is the coefficient of the interesting relevance of nodes, and ρ is the coefficient of the separating time relevance of nodes. The higher the node's social attribute value, the higher the closeness between the nodes and the higher the probability of encountering communication, which will improve the efficiency of information transfer between nodes.

Node similarity is affected by the network topology and social attributes. $NS_{(m,n)}$ represents the similarity between node m and node n . Correspondingly, in this paper, we define node similarity to be composed of network topology and social attributes, and the node similarity formula is

$$NS_{(m,n)} = \phi SMD_{tp(m,n)} + (1 - \phi)SR_{(m,n)}. \quad (18)$$

Through the above description, we can know the relationship between nodes more accurately. The higher the node similarity, the more frequent the communication between nodes. Source node can accurately find the relay node and then transmit information to the destination node by establishing a community [39]. The information transmission in this process is more efficient, and the time delay reduces.

The nodes within the same community are closely connected. Community detection is essentially the clustering of nodes with a tight structure in the network. This paper uses a hierarchical clustering algorithm to divide the community. Lead in modularity Q , which is used to measure the degree of community division. The fast unfolding algorithm considering data scale, running time, and other aspects of the community division results is ideal. The algorithm is stable and will continuously merge nodes to construct new graphs, which significantly reduces the calculation amount. The algorithm steps are as follows:

Step 1: initialize and calculate the node similarity; divide each node into the community where the adjacent node is located. As shown in Figure 2, the source node S is in community one. We try to move the node S to community two and community three. Calculate the corresponding modularity value, and move the node S to the corresponding community with the largest change value. We lead in modularity Q to measure the

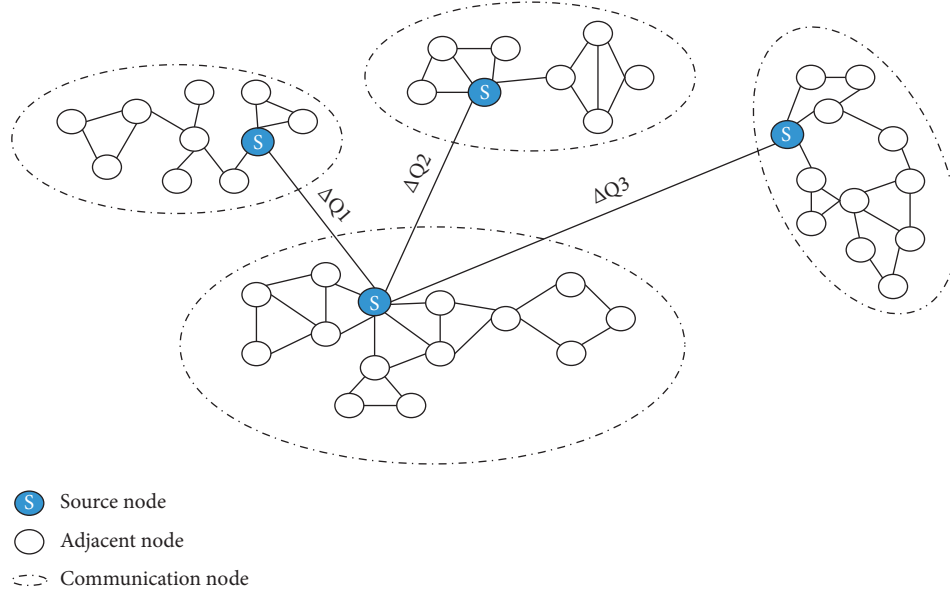


FIGURE 2: Schematic diagram of nodes moving between communities.

degree of community division. The specific calculation formula is as follows:

$$Q = \sum c \frac{\sum \text{in}}{2l} - \left(\frac{\sum \text{tot}}{2l} \right)^2, \quad (19)$$

where Q is the modularity, $\sum \text{in}$ represents the number of connections within the community, $\sum \text{tot}$ represents the sum of degrees of all nodes in the community, and l is the sum of weights in the network.

Step 2: select each node one by one, and calculate the modularity gain divided into the community where the adjacent point is located. ΔQ represents modularity gain, and the calculation formula is as follows:

$$\Delta Q = \left[\frac{\sum \text{in} + 2k_{m,\text{in}}}{2l} - \left(\frac{\sum \text{tot} + k_m}{2l} \right)^2 \right] - \left[\frac{\sum \text{in}}{2l} - \left(\frac{\sum \text{tot}}{2l} \right)^2 - \left(\frac{k_m}{2l} \right)^2 \right], \quad (20)$$

where $k_{m,\text{in}}$ is the sum of weights from node m to the community and k_m is the sum of the weights of node m . After calculating the modularity gain, we have to determine whether it is a positive number; if it is a positive number, it will be divided into the corresponding community; otherwise, no division will be made.

Step 3: repeat Step 2 until the node's community no longer changes.

Step 4: construct a new graph; each point in the new graph is each community divided in Step 3; continue to execute until the community structure does not change.

This paper roughly divides the above algorithm steps into two stages:

Stage 1: divide each node into the community where the adjacent node is located so that the modularity value becomes more immense.

Stage 2: the communities divided in the first stage are aggregated into one point, and the network is reconstructed until the structure of the network no longer changes.

This paper draws on the hierarchical clustering idea of the fast unfolding algorithm. We use network topology attributes and social attributes to express node similarity and comprehensively calculate node similarity to update network weights. In the first stage of node merging, we form an initial community to merge and improve the overall modularity and then calculate modularity gain; if ΔQ is positive then the two communities are merged; otherwise they will not be merged. The modularity gain is calculated repeatedly, and the final division result is output.

Nodes have the characteristics of random movement, and it is vital to establish a community. In opportunistic social network, many communities can usually deliver messages based on only one or two nodes. If these nodes do not have enough cache or overhead, data transmission in the community is likely to wait a long time. Therefore, after we divide the community, we need to establish further the weight distribution between the nodes and the community reconstruction so as to reduce the time complexity and overhead cost better. Below we will prove the changes in the community of the source node during the movement.

We define at time t , Q is the degree of modularity of the community, E_{we} is the total weight of weight, E_{wec} is total weight of the edges of community c , D_m is the degree of node m in community c , and ΔE_{we} is the increment of edge weight.

$$Q(t) = \frac{E_{wec}}{E_{we}} - \frac{D_m^2}{4E_{we}^2}. \quad (21)$$

Proposition 1. *In opportunistic social networks, the weight of the edge made by a node with other adjacent nodes in the network increases; the community relevance also will increase.*

Proof. With time t , the modularity in the community is $Q(t)$.

When the time increases to $t + 1$, the modularity change in the community can be expressed as

$$\begin{aligned} Q(t+1) &= \frac{E_{wec} + \Delta E_{we}}{E_{we} + \Delta E_{we}} - \frac{(D_m + 2\Delta E_{we})^2}{4(E_{we} + \Delta E_{we})^2} \\ Q(t+1) - Q(t) &= \frac{E_{wec} + \Delta E_{we}}{E_{we} + \Delta E_{we}} - \frac{(D_m + 2\Delta E_{we})^2}{4(E_{we} + \Delta E_{we})^2} - \left(\frac{E_{wec}}{E_{we}} - \frac{D_m^2}{4E_{we}^2} \right) \\ &= \frac{(4E_{we}^3 E_{wec} + 4E_{we}^2 \Delta E_{we} E_{wec} + 4E_{we}^3 \Delta E_{we} + 4E_{we}^2 \Delta E_{we}^2) - (E_{we}^2 D_m^2 + 4E_{we}^2 \Delta E_{we}^2 + 4E_{we}^2 D_m \Delta E_{we})}{4E_{we}^2 (E_{we} + \Delta E_{we})^2} \\ &\quad - \left(\frac{(4E_{we}^3 E_{wec} + 4E_{we} \Delta E_{we} E_{wec} + 8E_{we}^2 \Delta E_{we} E_{wec}) - (D_m^2 E_{we}^2 + D_m^2 \Delta E_{we}^2 + 2D_m^2 E_{we}^2 \Delta E_{we}^2)}{4E_{we}^2 (E_{we} + \Delta E_{we})^2} \right) \\ &\geq \frac{4E_{we}^3 \Delta E_{we} - 6E_{we}^2 D_m \Delta E_{we} + 2E_{we}^2 D_m \Delta E_{we} - 2E_{we}^2 D_m \Delta E_{we} + (D_m \Delta E_{we})^2}{4E_{we}^2 (E_{we} + \Delta E_{we})^2} \\ &= \Delta E_{we} \frac{4E_{we}^3 \Delta E_{we} - 6E_{we}^2 D_m + 2E_{we}^2 D_m - 2E_{we}^2 D_m \Delta E_{we} + D_m^2 \Delta E_{we}}{4E_{we}^2 (E_{we} + \Delta E_{we})^2} \\ &= \Delta E_{we} \frac{(2E_{we}^2 - 2E_{we} D_m - D_m \Delta E_{we}) \times (2E_{we} - D_m)}{4E_{we}^2 (E_{we} + \Delta E_{we})^2}. \end{aligned} \quad (22)$$

We can get $\Delta E_{we} > 0$, so we just need proof $(2E_{we}^2 - 2E_{we} D_m - D_m \Delta E_{we}) \times (2E_{we} - D_m) > 0$.
In other words,

$$\begin{cases} 2E_{we}^2 - 2E_{we} D_m - D_m \Delta E_{we} > 0, \\ 2E_{we} - D_m > 0, \end{cases}$$

$$\begin{cases} 2E_{we}^2 - 2E_{we} D_m - D_m \Delta E_{we} > 0, \\ 2E_{we} - D_m > 0, \\ \Delta E_{we} > 0, \end{cases}$$

$$\begin{cases} 0 < \Delta E_{we} < 2E_{we} \left(\frac{E_{we}}{D_m} - 1 \right), \\ 2E_{we} \left(\frac{E_{we}}{D_m} - 1 \right) > 0, \\ D_m < 2E_{we}, \end{cases}$$

$$\begin{cases} 0 < \Delta E_{we} < 2E_{we} \left(\frac{E_{we}}{D_m} - 1 \right), \\ D_m < E_{we}. \end{cases} \quad (23)$$

It is known that $2Q$ is the total of nodes in the network, and no community in the network appears more than $2Q$. In short, we are aware that increasing the weight can increase the community's relevance in opportunistic social networks. For this paper, the weight will affect the community's relevance in opportunistic social networks, and the proposition holds. \square

Proposition 2. *If the weight of an edge of two communities increases, node m is in community A, U_{commB}^- will be increased, and U_{commA}^+ will be decreased. The community corresponding to the node m will change, and the weight of an edge between the node and the community is ΔE_{we} ($\Delta E_{we} > 0$); if the weight of the edge can be changed, the result of the community will also change.*

Proof. Before the weight changes, for node m ,

$$\begin{cases} U_{\text{comm } A}^+ = e_A^m - \frac{D_m(D_A - D_m)}{2E_{\text{we}}}, \\ U_{\text{comm } B}^- = e_B^m - \frac{D_m D_B}{2E_{\text{we}}}. \end{cases} \quad (24)$$

After the weight changes, for node m ,

$$\begin{cases} \overline{U}_{\text{comm } A}^- = e_A^m - \frac{(D_m + \Delta E_{\text{we}})(D_A - D_m)}{2(E_{\text{we}} + \Delta E_{\text{we}})}, \\ \overline{U}_{\text{comm } B}^- = (e_B^m + \Delta E_{\text{we}}) - \frac{(D_m + \Delta E_{\text{we}})(D_B + \Delta E_{\text{we}})}{2(E_{\text{we}} + \Delta E_{\text{we}})}, \end{cases} \quad (25)$$

$$\begin{aligned} \overline{U}_{\text{comm } B}^- - U_{\text{comm } B}^- &= \frac{2E_{\text{we}}\Delta E_{\text{we}} + D_m D_B}{2E_{\text{we}}} - \frac{D_m D_B + \Delta E_{\text{we}} D_B + \Delta E_{\text{we}} D_m + \Delta E_{\text{we}}^2}{2(E_{\text{we}} + \Delta E_{\text{we}})} \\ &\geq \frac{2E_{\text{we}}\Delta E_{\text{we}} + D_m D_B}{2(E_{\text{we}} + \Delta E_{\text{we}})} - \frac{D_m D_B + \Delta E_{\text{we}} D_B + \Delta E_{\text{we}} D_m + \Delta E_{\text{we}}^2}{2(E_{\text{we}} + \Delta E_{\text{we}})} \\ &= \frac{2E_{\text{we}} - D_B - D_m - \Delta E_{\text{we}}}{2(E_{\text{we}} + \Delta E_{\text{we}})} \cdot \Delta E_{\text{we}}. \end{aligned}$$

Because $E_{\text{we}} > 0$, when $2E_{\text{we}} = \sum di$,

$$2E_{\text{we}} > 2E_{\text{we}} - D_B - D_m - \Delta E_{\text{we}} \cdot \overline{U}_{\text{comm } B}^- - U_{\text{comm } B}^- > 0. \quad (26)$$

All in all, if the weight of one side increases, then for node m $U_{\text{comm } B}^-$ increases. Then,

$$\overline{U}_{\text{comm } A}^+ - U_{\text{comm } A}^+ = \Delta E_{\text{we}} (D_A - D_m) \frac{D_m - E_{\text{we}}}{2E_{\text{we}}(E_{\text{we}} + \Delta E_{\text{we}})}. \quad (27)$$

Because $\Delta E_{\text{we}} > 0$, $E_{\text{we}} > 0$, $D_A - D_m > 0$, for all edges in the network $\sum D_i > D_m D_m - E_{\text{we}} < 0$, $\overline{U}_{\text{comm } A}^+ - U_{\text{comm } A}^+ < 0$.

If the weight of one side increases, then $U_{\text{comm } A}^+$ decreases.

If the weight of an edge of two communities increases, node m is in community A , then $U_{\text{comm } B}^-$ will increase and $U_{\text{comm } A}^+$ will decrease. \square

Proposition 3. *If node m and node n are connected, and one of the nodes has one and only one edge, when the weight between node m and node n drops, the community will not divide.*

Proof. Let us assume that the community is divided; then the following three conditions must be met:

$$\begin{cases} E_{\text{we},m} + E_{\text{we},n} < E_{\text{we}}, \\ \frac{u_m}{E_{\text{we}}} - \frac{D_i^2}{4E_{\text{we}}^2} + \frac{u_n}{E_{\text{we}}} - \frac{D_j^2}{4E_{\text{we}}^2} < \frac{D_i + D_j + w_{mn}}{E_{\text{we}}} - \frac{(D_i + D_j)^2}{4E_{\text{we}}^2}, \\ w_{mn} > \frac{D_i D_j}{2E_{\text{we}}}. \end{cases} \quad (28)$$

As the weight changes, the formula can also be expressed as

$$\begin{cases} E_{\text{we},m}^* + E_{\text{we},n}^* > E_{\text{we}}^*, \\ w_{mn} < \Delta E_{\text{we}} + \frac{D_i D_j + D_m \Delta E_{\text{we}} + \Delta E_{\text{we}}^2}{2(E_{\text{we}} + \Delta E_{\text{we}})}, \end{cases} \quad (29)$$

$$\frac{D_i D_j}{2E_{\text{we}}} < w_{mn} < \frac{D_i(D_j + \Delta E_{\text{we}})}{2(E_{\text{we}} + \Delta E_{\text{we}})} = \frac{D_i D_j + D_i \Delta E_{\text{we}}}{2(E_{\text{we}} + \Delta E_{\text{we}})}.$$

So, it can be seen from the above proof and we conclude that

$(D_i D_j / 2E_{\text{we}}) < w_{mn} < \Delta E_{\text{we}} + ((D_i D_j + D_m \Delta E_{\text{we}} + \Delta E_{\text{we}}^2) / (2(E_{\text{we}} + \Delta E_{\text{we}})))$ is false.

For a node in opportunistic social networks, if it has only one edge connected to another node, the community will not divide when the weight between the two nodes decreases.

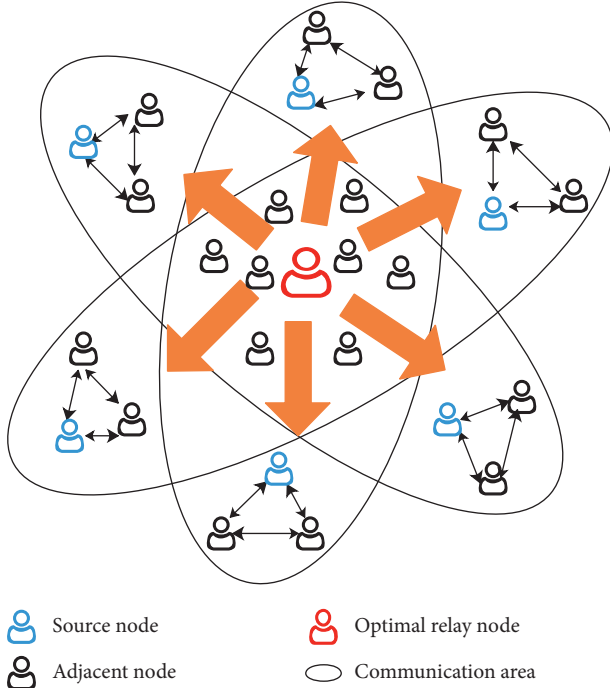


FIGURE 3: Information transfer model between communities.

After community detection, we construct a set of candidate relay nodes according to the relationship between the information forwarder and adjacent nodes. Select the optimal relay node from the set of candidate relay nodes to undertake the transmission task. Therefore, selecting one or more relays among multiple relay nodes to participate in transmission has become our concern. As shown in Figure 3, when the community is established and transmitted between each community, it is necessary to find a reliable relay node to transmit information. To achieve higher efficiency, construct a set of candidate relay nodes from the neighbor nodes of the source node; from the perspective of minimizing the bit error rate, this paper analyzes the channel coefficients of the two segments of the source node to the relay node and the relay node to the destination node and chooses the AF protocol as the relay node's forwarding method, which is suitable for the information transmission process of various channel qualities [40]. Calculate the sum of the channel coefficients of the channel corresponding to each relay node, and find the largest coefficient of the relay node, which is the optimal relay node and will improve the efficiency of information transmission.

Let us suppose there are a source node S , destination node D , and relay nodes R_1, R_2, \dots, R_n , when transferring information between communities. The communication model is as shown in Figure 4. In this case, the channels from the source node to the destination node and the source node to each relay node are all Rayleigh fading channels, which obey the Rayleigh distribution. We assume that the channel coefficient from the source node to the destination node is $C_{s,d}$, the channel coefficient from the source node to the n th relay node is $C_{s,rn}$, and the channel coefficient from the n th relay node to the destination node is $C_{rn,d}$.

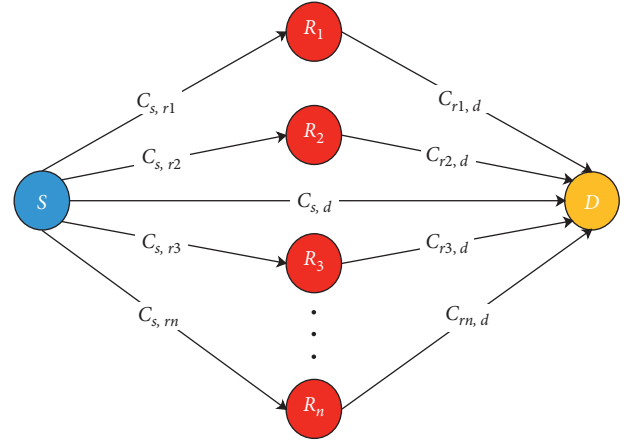


FIGURE 4: The communication model of the source node to the destination node.

The transmit power of the source node is P_1 , and the transmission power of the relay node is P_2 . When there is a direct transmission from the source node to the destination node, the power $P = P_1 + P_2$. When the source node sends information i to the destination node and the relay node is with power P_1 , noise from the source node to destination node is $V_{s,d}$ and noise from the source node to the relay node is $V_{s,r}$. So information received by the relay node and the destination node is as follows:

$$\begin{aligned} RM_{s,d} &= \sqrt{P_1} C_{s,d} \cdot i + V_{s,d}, \\ RM_{s,r} &= \sqrt{P_1} C_{s,r} \cdot i + V_{s,r}. \end{aligned} \quad (30)$$

In the AF protocol, when the relay node receives the signal from the source node and forwards it to the destination node, it will amplify the received signal, and the scaling factor is

$$\chi = \frac{1}{\sqrt{P_1 |C_{s,r}|^2 + N_0}}. \quad (31)$$

We can know that the signal from the relay node to the destination node is $\chi_{s,r}$, and then the information sent by the relay node to the destination node is

$$RM_{r,d} = \sqrt{P_2} C_{r,d} \cdot (\chi_{s,r}) + V_{r,d}. \quad (32)$$

This paper's focus on selecting the optimal relay node is how to find an optimal relay node that makes the channel coefficients of the source node to the relay node and the relay node to the destination node larger.

The channel coefficient matrix from the source node to the relay node is A , and the channel coefficient matrix from the relay node to the destination node is B . Then,

$$\begin{aligned} A_{1 \times n} &= [C_{s,r1}, C_{s,r2}, C_{s,r3}, \dots, C_{s,rn}], \\ B_{1 \times n} &= [C_{r1,d}, C_{r2,d}, C_{r3,d}, \dots, C_{rn,d}]. \end{aligned} \quad (33)$$

We define a threshold for the number of candidate relay nodes ψ and set $\psi \leq 100$; we have to consider the following situations:

- (1) If $n \geq \psi$, compare the channel coefficients of each relay node corresponding to matrices A and B , find the smaller of the two, and store the smaller value in the matrix S .

$$S_{1 \times n} = [C_1, C_2, C_3, \dots, C_n]. \quad (34)$$

Sort the matrix elements S from largest to smallest, select the first m relay nodes with a larger C_i value from them, and store them in the matrix T and $C_i = \min\{C_{s,ri}, C_{ri,d}\}$, where C_i is the smaller value of the channel coefficient of the two channels corresponding to the relay node r_i .

$$T_{1 \times n} = [r_1, r_2, r_3, \dots, r_i, \dots, r_n], \quad (35)$$

where r_i is one of the first m elements in the matrix after sorting. The value of m largely depends on the number of candidate relay nodes n , (m/n) the larger the value, the lower the bit error rate. Bit error rate refers to the index of the accuracy of data transmission within a specified time.

$$\text{SER} = \frac{\text{SER}_{te}}{\text{SER}_T} * 100\%, \quad (36)$$

where SER is bit error rate, SER_{te} is the bit errors in transmission, and SER_T is the total number of codes transmitted. We add the two channel coefficients of these m relay nodes, and the relay node with the largest sum is the optimal relay node as follows:

$$R = \{ri | \max(C_{s,ri} + C_{ri,d}), i = 1, 2, \dots, m\}. \quad (37)$$

- (2) Otherwise, when the number of candidate relay nodes is less than the threshold, we must pay attention to the accuracy of being selected as the optimal relay node; calculate the sum of the channel coefficients of the channel corresponding to each relay node and the relay node with the largest sum, which is the optimal relay node.

Based on the above definition, we propose an efficient data transmission algorithm EDCCD and the algorithm steps are as follows:

Step 1: calculate the encounter probability of node m and node n , construct the encounter probability matrix, and use the number of encounters between nodes to weight each edge.

Step 2: define node similarity, which is composed of network topology attributes and social attributes. Network topology attributes are composed of the strength of node, the direct connection strength, and the indirect connection strength. Social attributes are composed of the geographic relevance of nodes, the interesting relevance of nodes, and the separating time relevance of nodes.

Step 3: use a hierarchical clustering algorithm to divide the community and lead in the modularity Q . The modularity is used to measure the degree of community

division. And the fast unfolding algorithm is used to calculate the node similarity to update the network weight comprehensively.

Step 4: from the perspective of minimizing the bit error rate, after the community is divided into a multihop wireless network, construct a set of candidate relay nodes based on the relationship between the information forwarder and adjacent nodes and select the optimal relay node from the set of candidate relay nodes to undertake the transmission task. Analyze the channel coefficients of the channels from the source node to the relay node and the relay node to the destination node, and select the AF protocol as the relay node forwarding method for routing and forwarding.

To enhance the understanding and readability of the entire algorithm, the specific calculation flowchart of the EDCCD algorithm is shown in Figure 5. Algorithm 1 gives the initialization and community establishment phase of the proposed algorithm, and Algorithm 2 presents the routing and forwarding phase of the proposed algorithm. \square

4. Simulation and Analysis

To assess the performance of the EDCCD, we use a simulation tool called ONE (Opportunistic Network Environment) [41] and we compare with the following four typical routing algorithms.

Spray and wait [29]: this algorithm sprays the copies to the network and waits for these nodes to reach the destination node. The number of copies of the algorithm will affect performance, reduce the message delivery success rate, and increase the delivery delay.

SCR (Social-based Clustering and Routing Scheme) [36]: this algorithm is a useful measurement method of social relations between nodes in mobile opportunistic network, and is a novel social-based clustering and routing scheme.

SECM (status estimation and cache management) [42]: the algorithm uses state estimation and cache management methods to identify surrounding neighbors to evaluate the transmission probability between nodes, to ensure that they have high transmission, and to achieve the purpose of adjusting the cache.

EIMST (effective information transmission based on socialization nodes) [2]: the algorithm is based on social nodes to achieve effective information transmission. According to the defined stop time, when $t < h$, the node forwards the message with the most excellent probability, and when $t > h$, the node stops sending the message.

Download the real datasets from the network repository to experiments. According to the data information required for data transmission in opportunistic social networks, and choose pages-government [43], wiki-elec [44], advogato [45], and slashdot [46] four datasets for simulation experiments. The characteristic information of the four experimental datasets is shown in Table 1.

In the simulation experiment, we set the following metrics according to the characteristics of data transmission. The EDCCD algorithm and the other four algorithms run in

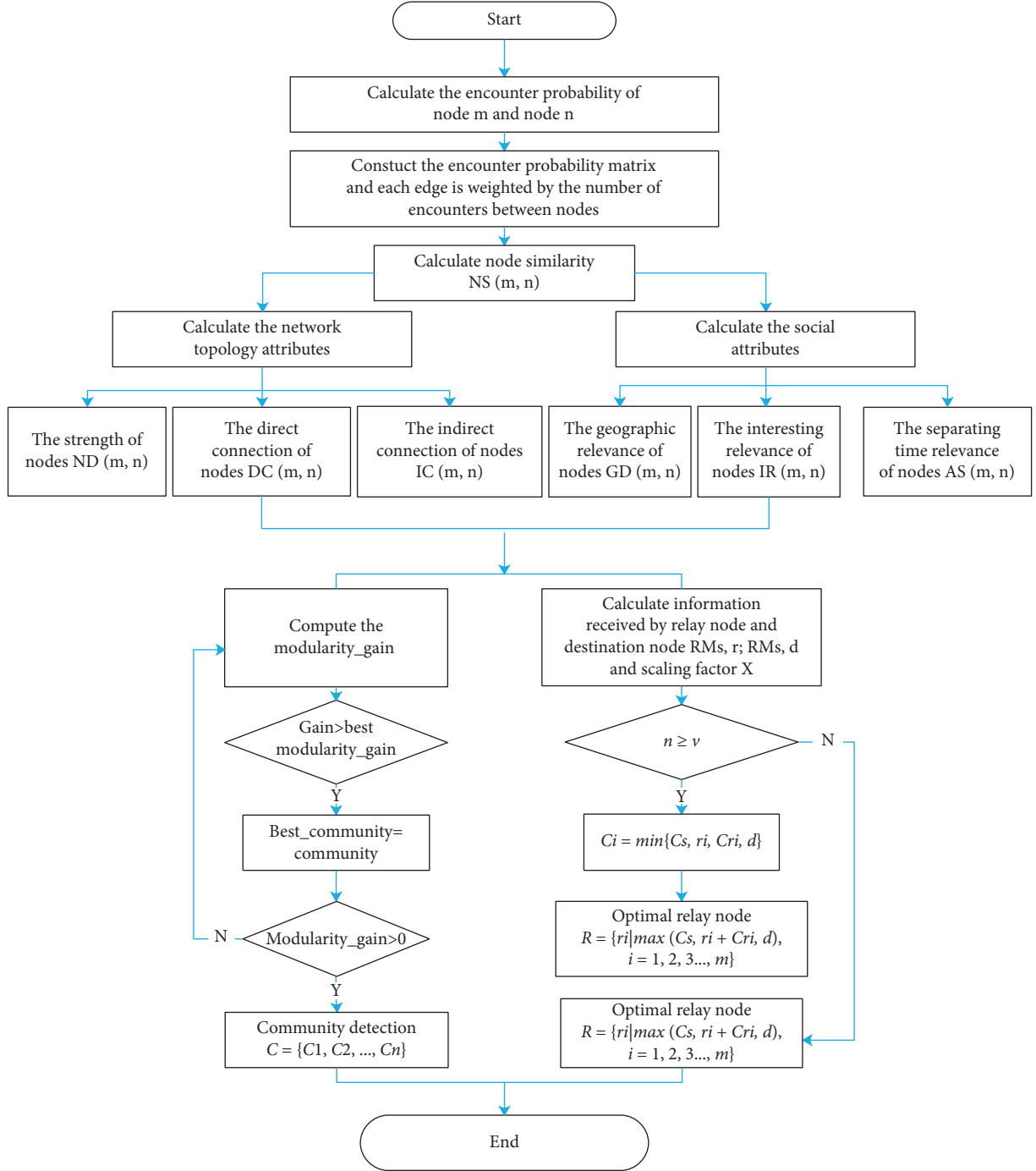


FIGURE 5: The process of EDCD algorithms.

the same simulation environment to compare their performance.

- (1) Delivery ratio: probability of choosing a suitable node as the next-hop node, represented as follows:

$$D_{\text{node}} = \frac{D_{\text{receive}}}{D_{\text{send}}}, \quad (38)$$

where D_{receive} is the number of messages received by the destination node and D_{send} is the total number of sent messages.

- (2) Routing overhead indicates the overhead between nodes when transmitting information, represented as follows:

$$Rd = \frac{R_{\text{sum}} - R_{\text{suc}}}{R_{\text{sum}}}, \quad (39)$$

where R_{sum} is the total time of the transmission between nodes and R_{suc} is the time to transmit a successful message between nodes.

- (3) Average end-to-end delay: express the delay in selecting the optimal next hop.


```

Input:  $G = (V, E, w)ND_{(m,n)}DC_{(m,n)}IC_{(m,n)}GD_{(m,n)}IR_{(m,n)},AS_{(m,n)}$ 
Output:  $C = \{C1, C2, \dots, Cn\}$ 
(1) Begin
(2) Initialize every node as a cluster;
(3) Calculate the encounter probability and times of node  $m$  and node  $n$  in a period of time  $t$ ;
(4)  $SMD_{tp(m,n)}$  Get Network topology ( $ND_{(m,n)}, DC_{(m,n)}, IC_{(m,n)}$ )
(5)  $SR_{(m,n)}$  Get Social relationship ( $GD_{(m,n)}, IR_{(m,n)}, AS_{(m,n)}$ )
(6)  $NS_{(m,n)} = SMD_{tp(m,n)} + SR_{(m,n)}$ //Compute the node similarity
First_phase:
(7) Initialize (self, nodes, edges):
(8) for ( $i=0; i \leq n; i++$ )
(9) self.communities = { $n1, n2, n3$ };
(10) partition = self.first_phase (network);
(11)  $q = q + self.s\_in[i]/2l - self.s\_tot[i]/2l$ ;
(12) End for
(13) Compute modularity_gain (self, node,  $c, k\_i\_in$ ):
(14) return  $2 * k\_m\_in - self.s\_tot[c] * self.k\_m[node]/self.m$ ;
(15) If (gain > best modularity_gain)
(16) best_community = community;
(17) best_partition[best_community].append (node);
(18) self.communities[node] = best_community;
(19) End If
Second_phase:
(20) for ( $i=0; i < partition.length; i++$ )
(21) Self.communities = (nodes, edges);
(22) In_order (nodes, edges);
(23) If (modularity_gain > 0)
(24) return  $C = \{C1, C2, \dots, Cn\}$ ;
(25) else
(26) return First_phase;
(27) End If
(28) End for
(29) END

```

ALGORITHM 1: Initialize community detection.

$$D_d = \frac{D_{\text{sum}}}{D_{\text{suc}}}, \quad (40)$$

where D_{sum} is the total delay of per node and D_{suc} is the total number of nodes successfully receiving messages.

The correlation between the time and delivery ratio in four different real datasets is shown in Figures 6–9. Figure 6 shows the delivery ratio of spray and wait, SCR, SECM, EIMST, and EDCD algorithms in pages-government dataset. We can infer that when the simulation time is less than one day, the advantages of the algorithm EDCD are not apparent in the four real datasets. However, as the simulation time increases, we can find that the transmission rate of the EDCD algorithm is always bigger than other algorithms. EDCD algorithm divides the community by node similarity, and the effective nodes in the community carry out data transmission, so the data delivery ratio is better than the other four algorithms. The relationship between the delivery ratio and the simulation time in wiki-elec dataset is shown in Figure 7. The SCR algorithms deliver information to nodes, and the community by using the flooding method leads to mass information missing.

The delivery ratio of SECM is 0.65–0.78. EIMST and EDCD algorithm's delivery ratio is higher than the other. EIMST algorithm controls the time interval of delivery information that improves the transmission and receiving of effective information, and its delivery ratio reached 0.66–0.81. Due to the adoption of the EDCD algorithm combining network topology and social attributes, the algorithm's transmission rate is the highest among all algorithms, reaching 0.67–0.84.

The correlation between the delivery ratio and simulation time in advogato dataset is shown in Figure 8. We see that the algorithm with the highest delivery rate is the EDCD algorithm, reaching 0.85–0.88. The spray and wait algorithm uses flooding to transmit information at community nodes, a large amount of information is lost, and the delivery rate is the lowest, only 0.67–0.70. Figure 9 shows the relationship between time and delivery ratio in slashdot dataset. The dataset with the largest number of nodes in the four datasets is slashdot dataset. When the simulation time is less than one and a half days, each dataset's delivery ratio is rising sharply, and the time is up to three days; only the EDCD and EIMST algorithms' delivery ratio is rising. This is because, in slashdot dataset, the two algorithms quantify the social attributes in 5G environment of nodes. On the whole, in the

Input: source node S , relay node R_1, R_2, \dots, R_n , destination node D ; power of source node P_1 ; power of relay node P_2 ;
Output: optimal relay node R ;

```

(1) Begin
(2) Power of Destination node  $P = P_1 + P_2$ ;
(3) Calculate Information received by relay node and destination node  $RM_{s,r}, RM_{s,d}$ ;
(4) Amplify the received signal and calculate Scaling factor  $\chi$ ;
(5) Threshold of number of candidate relay nodes  $\psi$ ;
(6) Function BER = AF_Simulation (max_SNR);
(7) for (snr = 0; snr ≤ max_SNR; snr++)
(8)   for ( $i = 0; i \leq n; i++$ )
(9)      $V = 1/(10^{(snr/10)})$ ; //  $V$  is the variance and the noise energy is normalized
(10)    sig = randsrc(1, N, [0 1]); // Generating binary input sequences
(11)    sig_mod = QpskMapping(sig); // The input binary sequence is QPSK modulated
(12)  End for
(13)  If ( $n > \psi$ )
(14)     $C_i = \min\{C_{s,r_i}, C_{r_i,d}\}$ ;
(15)    Optimal relay node  $R = \{r_i | \max(C_{s,r_i} + C_{r_i,d}), i = 1, 2, \dots, m\}$ 
(16)  else
(17)    Optimal relay node  $R = \{r_i | \max(C_{s,r_i} + C_{r_i,d}), i = 1, 2, \dots, m\}$ 
(18)  End If
(19) End for
(20) END

```

ALGORITHM 2: Routing and forwarding.

TABLE 1: Characteristics of the experimental datasets.

Dataset	Pages-government	Wiki-elec	Advogato	Slashdot
Duration (days)	3.5	11	4	4
Message transmission rate (kbps)	250	250	250	250
Number of experimental devices	7100	8000	5200	70000
Buffer size (M)	5	5	5	5
TTL (time to live)	0.5 days	1day	60 min	2days
Node initial energy (J)	200	200	200	200
The sending frequency of a data packet(s)	35	30	35	25

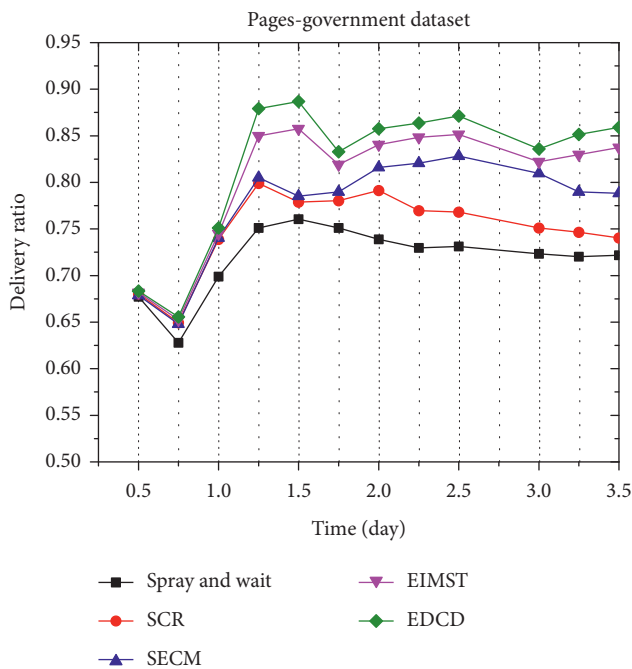


FIGURE 6: Comparison of algorithms delivery ratio in pages-gov dataset.

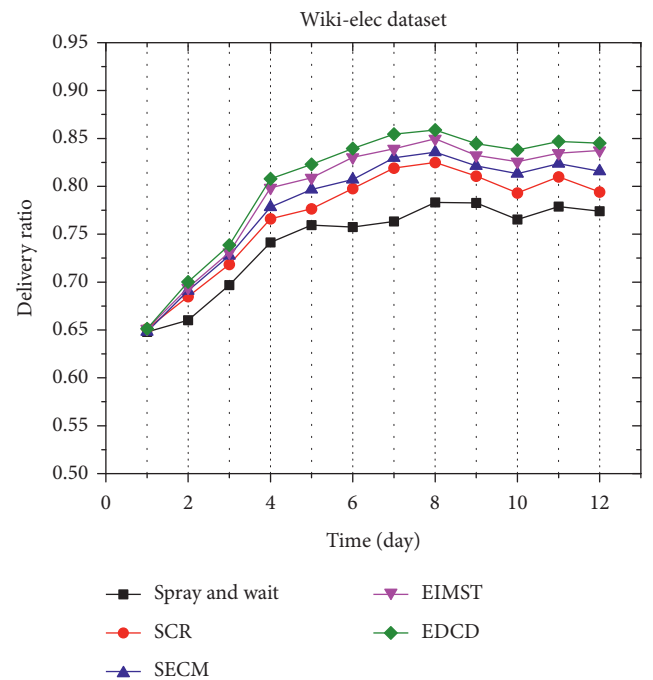


FIGURE 7: Comparison of algorithms delivery ratio in wiki-elec dataset.

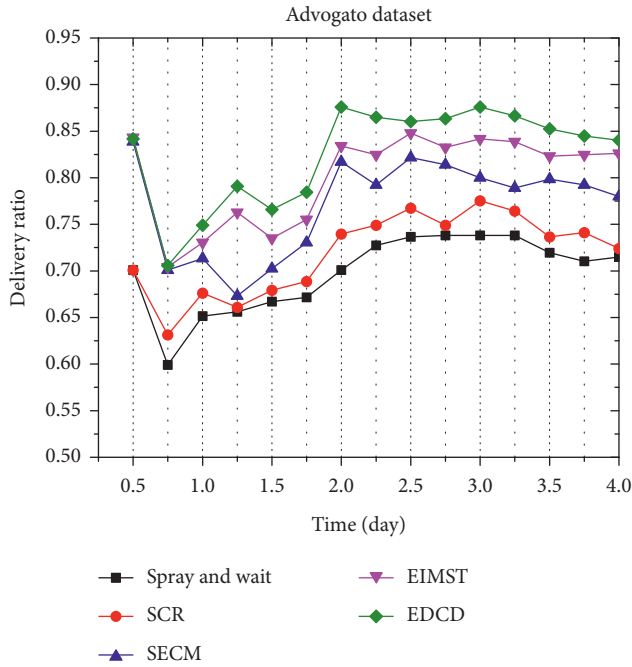


FIGURE 8: Comparison of algorithms delivery ratio in advogato dataset.

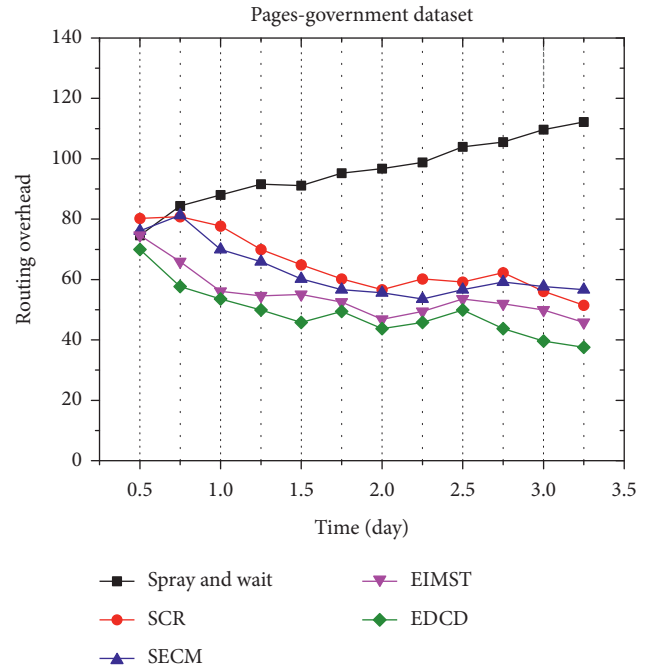


FIGURE 10: Comparison of algorithms routing overhead in pages-government dataset.

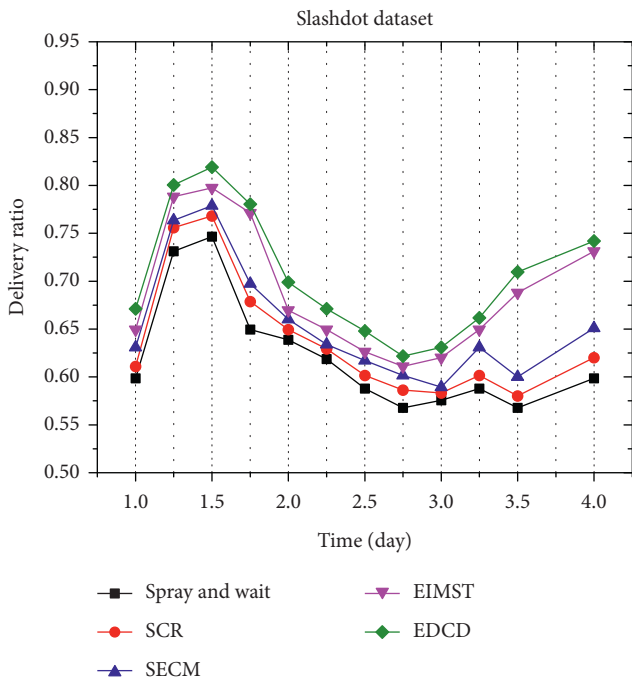


FIGURE 9: Comparison of algorithms delivery ratio in slashdot dataset.

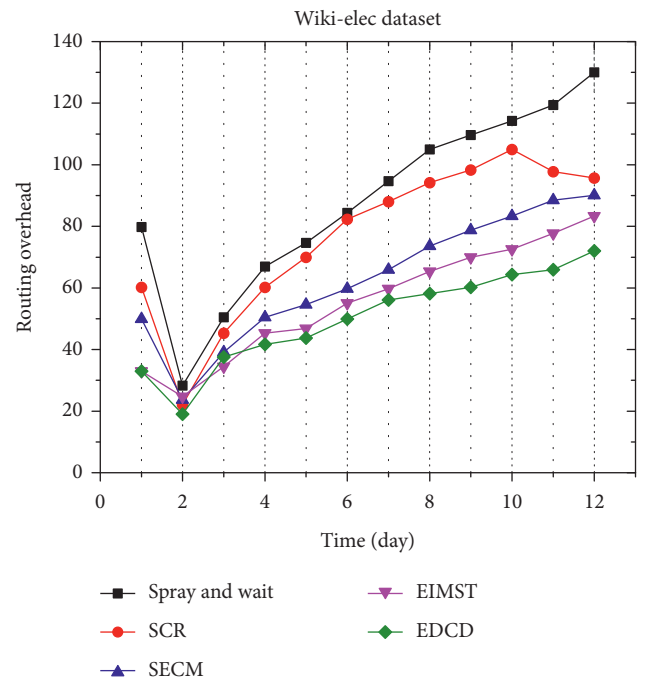


FIGURE 11: Comparison of algorithms routing overhead in wiki-elec dataset.

EDCD algorithm, the delivery ratio is 0.76 on average, which is higher than the other algorithms.

The correlation between the time and routing overhead in four different real datasets is shown in Figures 10–13. The comparison of the routing overhead between these five

different algorithms in pages-government dataset is shown in Figure 10. The average routing overhead of the EDCD algorithm is always kept to the lowest. The algorithm uses the node similarity to divide the community and uses the optimal relay node strategy to forward information. The routing overhead of the EDCD algorithm is maintained between 40 and 65.

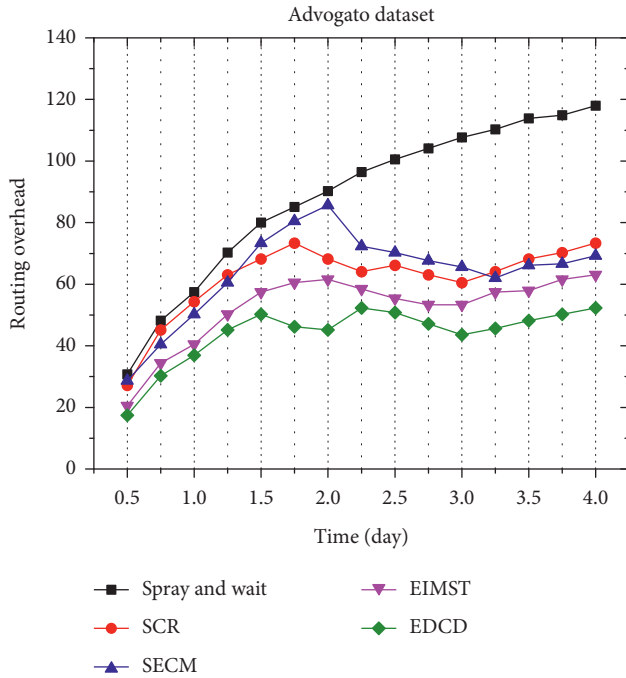


FIGURE 12: Comparison of algorithms routing overhead in advogato dataset.

Figure 11 shows the association between routing overhead and time in wiki-elec dataset. In the spray and wait algorithm, redundant message group copies require a lot of time and resources, which is the main reason for the vast routing overhead. In the SCR algorithm, each node only forwards a copy of the message to the node with the destination node as a cluster member, ignoring the current availability of the next-hop node, which will cause overhead. In the SECM algorithm, because the node injects many redundant data, the overhead will be large. In the EIMST algorithm, information and buffer space can be effectively managed, but it consumes some unavailable node resources. In terms of routing overhead, EDCD always performs best among these five algorithms. Figure 12 shows the relationship between time and routing overhead in advogato dataset. Compared with other algorithms, EDCD algorithms select the optimal relay node and set up the weight distribution between nodes and community to reduce the overhead cost. Regarding the spray and wait algorithms, a lot of redundant information use lot of computing resources. For SCR and SECM algorithms, the cooperation mechanism is conducive to the reasonable allocation of computing resources, so the cost of these two algorithms is in the middle level. EIMST does not fully consider the transmission preference of nodes, so its performance is worse than that of EDCD algorithm.

The relationship between routing overhead and time in slashdot dataset is shown in Figure 13. From the chart, we can see that the routing overhead increases sharply at first, nearly stably by the time it reaches the third day. The routing overhead of the spray and wait algorithm increases dramatically; a large number of data copies are generated in

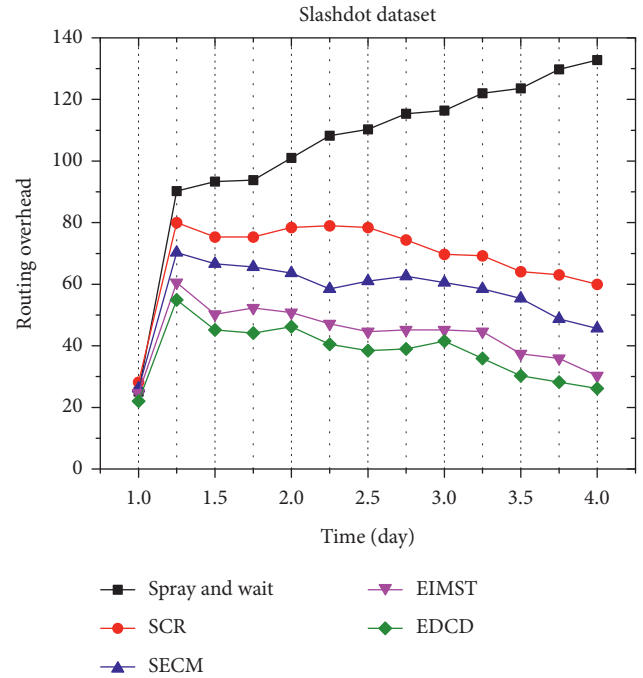


FIGURE 13: Comparison of algorithms routing overhead in slashdot dataset.

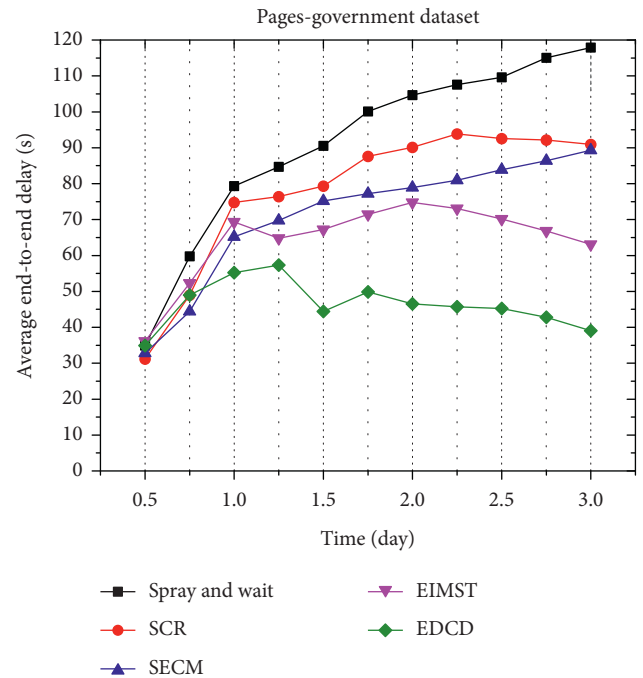


FIGURE 14: Comparison of algorithms average end-to-end delay in pages-government dataset.

slashdot dataset with a large number of nodes, and these need to be processed, so the routing overhead is higher than other algorithms.

The association between the time and average end-to-end delay in four different real datasets is shown in Figures 14–17. The relationship between the average end-to-

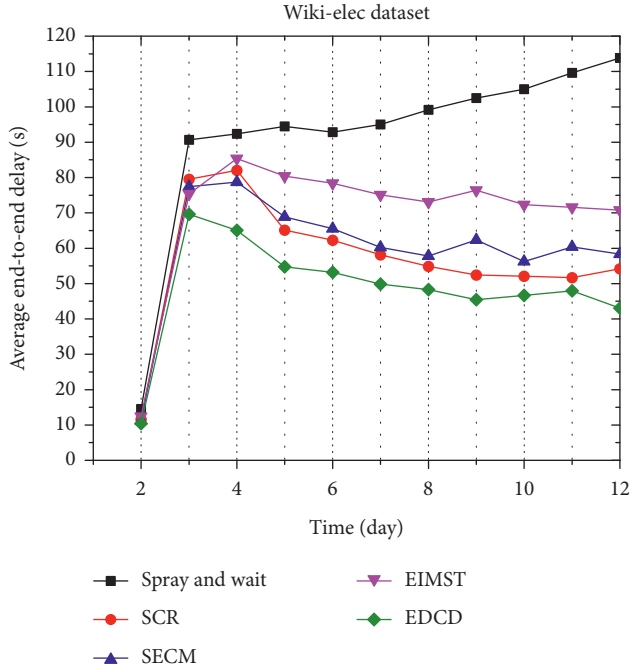


FIGURE 15: Comparison of algorithms average end-to-end delay in wiki-elec dataset.

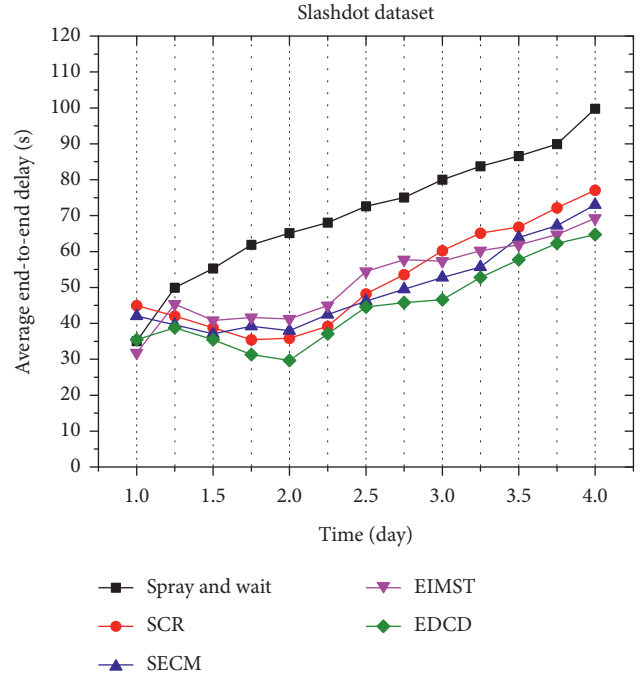


FIGURE 17: Comparison of algorithms average end-to-end delay in slashdot dataset.

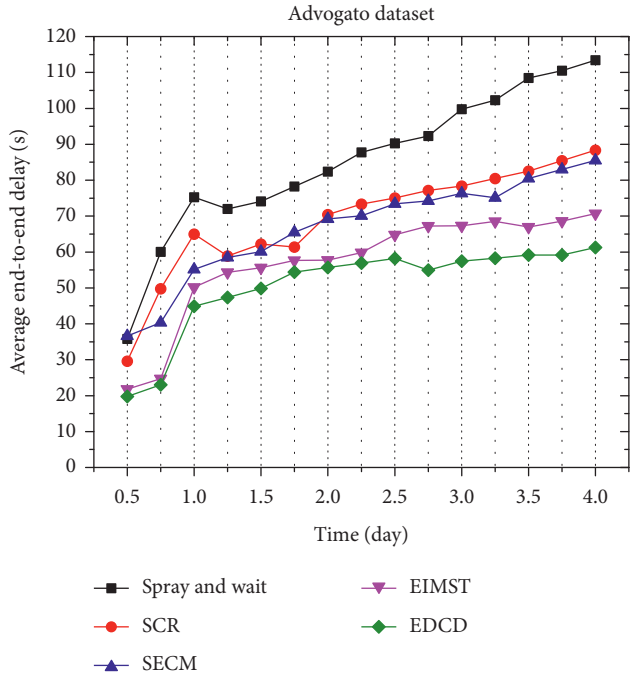


FIGURE 16: Comparison of algorithms average end-to-end delay in advogato dataset.

end delay and time of each algorithm in pages-government dataset is shown in Figure 14. Compared with the other four algorithms, the EDCD algorithm has the lowest average end-to-end delay.

Since the EDCD algorithm proposes a strategy for dividing communities by analyzing the comprehensive

characteristics of nodes, it can reduce inefficient nodes that are not helpful to the transmission process, reducing the average end-to-end delay. The spray and wait algorithm has more message copies, which will cause corresponding delays. The SCR algorithm effectively forwards the copy of the message to the destination node, so the transmission delay is lower than the spray and wait algorithm. SECM algorithm will also increase the cache of node before data transmission, so there will be a corresponding delay.

Figure 15 shows the association between routing overhead and time in wiki-elec dataset. We can see that the EIMST algorithm's delay is higher than that in other datasets but lower in the rest of the datasets. Because the EIMST algorithm applies node based on information management, there are more nodes in the wiki-elec dataset, and the delay increases as the simulation time increases. In short, the average end-to-end delay of the EDCD algorithm in wiki-elec dataset is lower than the other four algorithms.

Figure 16 shows the relationship between average end-to-end delay and time in advogato dataset. To be specific, spray and wait algorithm's maximum delay could reach 95 because this method remarkably increased routing and message forwarding delays. The SCR and SECM algorithms have lower delays than the spray and wait algorithm because both algorithms effectively controlled a lot of message copies. Besides, the SCR algorithm implemented community division and information management. In contrast, the SECM algorithm effectively utilized the cooperation mechanism between nodes to utilize the nodes' cache space reasonably to reduce the delay in the message forwarding process.

The average end-to-end delay of the EIMST algorithm was also significantly lower than the other algorithms.

Figure 17 shows the correlation between the average end-to-end delay and time in slashdot dataset. In a dataset with many nodes, we can see in the figure that the average end-to-end delay of the EIMST algorithm is significantly higher than other datasets. That is why the EIMST algorithm implements community detection. However, the effect is general when processing large amounts of data. The algorithm EDCD proposed in this paper has a lower latency in different real datasets than other algorithms.

5. Conclusions

In this study, an effective data transmission scheme in opportunistic social networks that uses mobile edge computing combined with network topology attributes and social attributes to measure node similarity to divide communities and select the optimal relay node. This algorithm is mainly based on the idea that the closeness between nodes in the community is higher than that exterior in the community and provides a method for selecting the optimal relay node according to the sum of channel coefficients in the process of transmitting information. The simulation experiment results show that the strategy has good performance in different real datasets such as delivery ratio, routing overhead, and average end-to-end delay. The EDCD algorithm can be used to the 5G data transmission scene and can cope with the challenges of stability and continuity required by data in the interactive process through efficient community division and information transmission. In future work, we will enhance the related performance of the algorithm and will further study the security of data transmission in opportunistic social networks.

Data Availability

The data used to support the findings of this study are currently under embargo, while the research findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Science Foundation of China under Grant 61966035 (Research on Super-Resolution Reconstruction of Remote Sensing Images Based on Deep Learning of Spatio-Temporal Spectrum Features), by the Intelligent Multi-Modal Information Processing Project (XJEDU2017T002), by the International Cooperation Project of the Autonomous Region's Science and Technology Department's "Data-driven China-Russia Cloud Computing Sharing Platform Construction" No. 2020E01023.

References

- [1] G. Yu and J. Wu, "Content caching based on mobility prediction and joint user Prefetch in Mobile edge networks,"

- Peer-to-Peer Networking and Applications*, vol. 13, no. 5, pp. 1839–1852, 2020.
- [2] J. Wu, Z. Chen, and M. Zhao, "Effective information transmission based on socialization nodes in opportunistic networks," *Computer Networks*, vol. 129, pp. 297–305, 2017.
- [3] J. Wu, Z. Chen, and M. Zhao, "Community recombination and duplication node traverse algorithm in opportunistic social networks," *Peer-to-Peer Networking and Applications*, vol. 13, no. 3, pp. 940–947, 2020.
- [4] Y. Cai, S. Pan, X. Wang, H. Chen, X. Cai, and M. Zuo, "Measuring distance-based semantic similarity using meronymy and hyponymy relations," *Neural Computing and Applications*, vol. 32, no. 8, pp. 3521–3534, 2018.
- [5] J. Wu, X. Tian, and Y. Tan, "Hospital evaluation mechanism based on mobile health for IoT system in social networks," *Computers in Biology and Medicine*, vol. 109, pp. 138–147, 2019.
- [6] J. Luo, J. Wu, and Y. Wu, "Advanced data delivery strategy based on multiperceived community with IoT in social complex networks," *Complexity*, vol. 2020, pp. 1–15, Article ID 3576542, 2020.
- [7] X. Zhu, Q. Yang, H. Tian, J. Ma, and W. Wang, "Contagion of information on two-layered weighted complex network," *IEEE Access*, vol. 7, pp. 155064–155074, 2019.
- [8] H. Zhang, Z. Chen, J. Wu, and K. Liu, "FRRF: a fuzzy reasoning routing-forwarding algorithm using mobile device similarity in mobile edge computing-based opportunistic mobile social networks," *IEEE Access*, vol. 7, pp. 35874–35889, 2019.
- [9] Y. I. N. Sheng, W. U. Jia, and Y. U. Genghua, "Low energy consumption routing algorithm based on message importance in opportunistic social networks," *Peer-to-Peer Networking and Applications*, vol. 14, no. 2, pp. 948–961, 2021.
- [10] W. U. Jia, Q. U. Jingge, and Y. U. Genghua, "Behavior prediction based on interest characteristic and user communication in opportunistic social networks," *Peer-to-Peer Networking and Applications*, vol. 14, no. 2, pp. 1006–1018, 2021.
- [11] E. P. N. Karunanayake, "Optimal relay node placement to improve design optimal relay node placement to improve expected life time in wireless sensor network design," 2020.
- [12] F. Xiong, Y. Liu, and H. F. Zhang, "Multi-source information diffusion in online social networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015, no. 7, 2015.
- [13] W. Y. B. Lim, "Federated learning in mobile edge networks: a comprehensive survey," *arXiv*, vol. 22, no. 3, pp. 2031–2063, 2019.
- [14] S. Pan, J. Wu, X. Zhu, G. Long, and C. Zhang, "Task sensitive feature exploration and learning for multitask graph classification," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 744–758, 2017.
- [15] S. Wang, X. Chang, X. Li, Q. Z. Sheng, and W. Chen, "Multi-task support vector machines for feature selection with shared knowledge discovery," *Signal Processing*, vol. 120, pp. 746–753, 2016.
- [16] Z. Gao, J. Meng, Q. Wang, and Y. Yang, "Data offloading for deadline-varying tasks in mobile edge computing," in *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, pp. 1479–1484, Guangzhou, China, October 2018.
- [17] W. Shi, L. Zhai, M. Ouyang, and J. Zhang, "A mobile edge computing server deployment scheme in wireless mesh

- network,” in *Proceedings of the 2019 IEEE/CIC International Conference on Communications Workshops in China (ICCC Workshops)*, pp. 25–29, Changchun, China, August 2019.
- [18] A. Adebayo, D. B. Rawat, L. Ni, and M. Song, “Group-Query-as-a-Service for secure low-latency opportunistic RF spectrum access in mobile edge computing enabled wireless networks,” in *Proceedings of the 2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–7, Hangzhou, China, July 2018.
- [19] J. Wu, Z. Chen, and M. Zhao, “An efficient data packet iteration and transmission algorithm in opportunistic social networks,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 8, pp. 3141–3153, 2020.
- [20] F. Xiong, Y. Liu, and J. Cheng, “Modeling and predicting opinion formation with trust propagation in online social networks,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 44, pp. 513–524, 2017.
- [21] Y. He, F. R. Yu, N. Zhao, and H. Yin, “Secure social networks in 5G systems with mobile edge computing, caching, and device-to-device communications,” *IEEE Wireless Communications*, vol. 25, no. 3, pp. 103–109, 2018.
- [22] Y. A. N. G. Weiyu, W. U. Jia, and J. Luo, “Effective data transmission and control base on social communication in social opportunistic complex networks,” *Complexity*, vol. 2020, Article ID 3721579, 13 pages, 2020.
- [23] J. Wu, Z. Chen, and M. Zhao, “Weight distribution and community reconstitution based on communities communications in social opportunistic networks,” *Peer-to-Peer Networking and Applications*, vol. 12, no. 1, pp. 158–166, 2019.
- [24] X. Li and J. Wu, “Node-oriented secure data transmission algorithm based on IoT system in social networks,” *IEEE Communications Letters*, vol. 24, no. 12, pp. 2898–2902, 2020.
- [25] X. Zhu, J. Ma, X. Su, H. Tian, W. Wang, and S. Cai, “Information spreading on weighted multiplex social network,” *Complexity*, vol. 2019, Article ID 5920187, 2019 pages.
- [26] Z. Zhang, Y. Liu, X. Chen et al., “Sequential optimization for efficient high-quality object proposal generation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1209–1223, 2018.
- [27] Y. Yan, Z. Chen, J. Wu, L. Wang, K. Liu, and Y. Wu, “Effective data transmission strategy based on node socialization in opportunistic social networks,” *IEEE Access*, vol. 7, pp. 22144–22160, 2019.
- [28] A. Vahdat and D. Becker, *Epidemic Routing for Partially Connected Ad Hoc Networks*, *Handbook of Systemic Auto-immune Diseases*, Elsevier, Amsterdam, Netherlands, 2000.
- [29] S. Sisodiya, P. Sharma, and S. K. Tiwari, “A new modified spray and wait routing algorithm for heterogeneous delay tolerant network,” in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 843–848, Palladam, India, February 2017.
- [30] D. K. Sharma, S. K. Dhurandher, I. Woungang, R. K. Srivastava, A. Mohananeey, and J. J. P. C. Rodrigues, “A machine learning-based protocol for efficient routing in opportunistic networks,” *IEEE Systems Journal*, vol. 12, no. 3, pp. 2207–2213, 2018.
- [31] K. Tang, C. Li, H. Xiong, J. Zou, and P. Frossard, “Reinforcement learning-based opportunistic routing for live video streaming over multi-hop wireless networks,” in *Proceedings of the 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, London, UK, May 2019.
- [32] J. Wu, Z. Chen, and M. Zhao, “Information cache management and data transmission algorithm in opportunistic social networks,” *Wireless Networks*, vol. 25, no. 6, pp. 2977–2988, 2019.
- [33] Y. Yan, Z. Chen, J. Wu, L. Wang, K. Liu, and P. Zheng, “An effective transmission strategy exploiting node preference and social relations in opportunistic social networks,” *IEEE Access*, vol. 7, pp. 58186–58199, 2019.
- [34] J. Wu and Z. Chen, “Human activity optimal cooperation objects selection routing scheme in opportunistic networks communication,” *Wireless Personal Communications*, vol. 95, no. 3, pp. 3357–3375, 2017.
- [35] R. Drăgan, R. I. Ciobanu, and C. Dobre, “Leader election in opportunistic networks,” in *Proceedings of the 2017 IEEE 16th International Symposium on Parallel and Distributed Computing (ISPDC)*, Innsbruck, Austria, July 2017.
- [36] F. Zeng, N. Zhao, and W. Li, “Effective social relationship measurement and cluster based routing in mobile opportunistic networks,” *Sensors*, vol. 17, no. 5, pp. 1109–1119, 2017.
- [37] K. Liu, Z. Chen, J. Wu, and L. Wang, “FCNS: a fuzzy routing-forwarding algorithm exploiting comprehensive node similarity in opportunistic social networks,” *Symmetry*, vol. 10, no. 8, pp. 338–8, 2018.
- [38] J. Niu, J. Guo, Q. Cai, N. Sadeh, and S. Guo, “Predict and spread: an efficient routing algorithm for opportunistic networking,” in *Proceedings of the 2011 IEEE Wireless Communications and Networking Conference*, pp. 498–503, Cancun, Mexico, March 2011.
- [39] Z. Zhang and V. Saligrama, “Zero-shot learning via joint latent similarity embedding,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6034–6042, Seattle, WA, USA, June 2016.
- [40] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, “Semi-supervised multiple feature analysis for action recognition,” *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 289–298, 2014.
- [41] A. Keränen, J. Ott, and T. Kärkkäinen, “The ONE simulator for DTN protocol evaluation,” in *Proceedings of the Second International ICST Conference on Simulation Tools and Techniques*, Rome, Italy, March 2009.
- [42] J. Wu, Z. Chen, and M. Zhao, “SECM: status estimation and cache management algorithm in opportunistic networks,” *The Journal of Supercomputing*, vol. 75, no. 5, pp. 2629–2647, 2019.
- [43] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, “Gemsec,” in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 65–72, Vancouver, Canada, August 2019.
- [44] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks in social media,” in *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, pp. 1361–1370, Atlanta, GA, USA, April 2010.
- [45] P. Massa, M. Salvetti, and D. Tomasoni, “Bowling alone and trust decline in social network sites,” in *Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pp. 658–663, Chengdu, China, December 2009.
- [46] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, “Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters,” *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

Research Article

The Hotspots of Sports Science and the Effects of Knowledge Network on Scientific Performance Based on Bibliometrics and Social Network Analysis

Linxiao Ma ¹, Yuzhu Wang ², Yue Wang ³, Ning Li ³, Sai-Fu Fung ⁴, Lu Zhang ³
and Qian Zheng ³

¹Sport College, Xi'an University of Architecture and Technology, Xi'an 710311, China

²Shandong Sport University, Jinan 250063, China

³Xi'an Physical Education University, Xi'an 710065, China

⁴City University of Hong Kong, Kowloon Tong, Hong Kong

Correspondence should be addressed to Yuzhu Wang; wangyuzhu@sdpei.edu.cn

Received 18 March 2021; Revised 17 April 2021; Accepted 8 May 2021; Published 24 May 2021

Academic Editor: Fei Xiong

Copyright © 2021 Linxiao Ma 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 study, we sorted out the research hotspots in sports science by bibliometric method and also used social network analysis to explore the relationship between knowledge networks and their scientific performance. We found 38 high-frequency keywords with obvious curricular nature or classical direction of sports science research and 4 high-frequency research groups. The topics of hotspots covered the secondary disciplines of sports science: physical education and training, national traditional sports, sports human science, and sports humanities and sociology. However, sports human science research is less; therefore, accelerating the research of sports human science is the focus of future research. Meanwhile, we use social network structure analysis (i.e., centrality, clustering coefficient, PageRank, and structural holes) to study the relationship between knowledge elements in knowledge networks and their scientific performance. In addition to betweenness centrality, the closeness centrality, clustering coefficient, and structural holes of knowledge elements are significantly and positively related to their influence. In the relationship between knowledge elements and productivity, betweenness centrality and closeness centrality show significant positive correlations, and clustering coefficient and structural hole show significant negative correlations. Therefore, knowledge networks can be used to predict the scientific performance of knowledge elements.

1. Introduction

In the context of the big data era, human beings produce large-scale behavioral data, and the development of computer technology enables the generated data to be stored [1]. By collecting, cleaning, and mining data to reveal the characteristics contained in the data, we can better understand human behavior and social interaction and provide new perspectives and methods for sociological research. With the rapid growth in the number of academic achievements of researchers, it is difficult for researchers to explore the research hotspots in their subject areas and the mechanisms underlying the impact

of knowledge themes, and knowledge graph is a bibliometric visualization method developed on the basis of social network theory. It combines knowledge and methods from disciplines such as graph theory and information visualization techniques. It is able to show the knowledge development process and structural relationships and helps to understand the research hotspots and status quo of subject areas, etc. [2, 3]. For the investigation of the intrinsic mechanisms of the influence of knowledge topics, social networks provide a good research perspective, which helps us to understand the strengths and weaknesses of the scientific research performance of knowledge elements by network features.

2. Data Collection and Measure

2.1. Data Collection. According to the Annual Report on the Impact Factor of Chinese Academic Journals (2020 Edition), developed and published by the China Research Center for Scientific Bibliometric Evaluation and Tsinghua University Library, ten journals entered Q1 area. They were “Sports Science,” “Journal of Beijing Sport University,” “Journal of Shanghai Sport Institute,” “Journal of Wuhan Sport Institute,” “Sports Journal,” “China Sports Science and Technology,” “Sports and Science,” “Sports Science Research,” “Journal of Chengdu Sport Institute,” and “Sports Culture Guide.” In 2018, Journal of Nanjing Sport Institute (Social Science) was renamed “Sports Science Research.” Its paper had been published only for three years, so the data of “Sports Science Research” were discarded. In this paper, nine journals were selected from 2000 to 2020 and the paper type was academic journals, and the search time was December 11, 2020. There were 45,472 journal papers that met the criteria, and 43465 papers remained after excluding those with empty keywords. We filtered papers with 0 citations when examining the impact of knowledge networks on scientific performance.

2.2. Measure

2.2.1. Measure High-Frequency Keywords. Paper keywords are the summary of the paper content, which can express the research content of the paper more accurately. Through the statistics of the frequency of the occurrence of paper keywords, it can reveal the research hotspots and evolution trends in the subject field. Chu reveals that the research hotspots in the field of knowledge management in the recent ten years focus on knowledge management, knowledge sharing, tacit knowledge, library, knowledge management, systematic knowledge, knowledge economy, enterprise, knowledge transfer, knowledge service, explicit knowledge, and knowledge map through the word frequency statistics of paper keywords [4]. Li and Jiang analyzed the hot research topics of sports science in the recent five years, including competitive sports, sports management, mass sports, school sports, physical education, sports culture, sports history, traditional national sports, sports economy, and sports industry [5]. Jiang et al. conducted keyword analysis based on CSSCI database of sports humanities and sociology and found that the research hotspots of sports humanities and social sciences in the past five years were sports teaching, sports culture, competitive sports, sports industry and Olympics, with a decreasing trend in sports teaching research and an increasing trend in sports culture research [6]. This provides a reference for sports humanities and social science researchers.

Donohue proposed a method to distinguish between high- and low-frequency keywords [7]. The dividing line between high- and low-frequency keyword frequencies is defined as follows:

$$TF = \frac{-1 + \sqrt{1 + 8TF_1}}{2}, \quad (1)$$

where TF refers to term frequency and TF_1 is the number of keywords with a term frequency of 1.

2.2.2. Measuring Knowledge Network Features. Social network analysis, also known as structural analysis, is a set of norms and methods to analyze the relational structure of social networks and their attributes [8–10], which helps us to measure the importance of nodes using their positional attributes. Common measures are centrality, clustering coefficient, and structural holes [11–15]. Abbasi et al. found a significant correlation between the attributes of authors in collaborative networks and g -index [11]. The papers are sorted in descending order by the number of citations, and when the cumulative number of citations is equal to the square of the order number, the order number is the g -index [15]. Yan and Ding found that the centrality of authors in co-authorship networks was significantly correlated with citation counts [13]. Guan et al. demonstrated that structural holes in knowledge networks are positively related to citation, and centrality has an inverted U-shaped relationship with citation, but they did not focus on the impact of knowledge networks on productivity [14]. Network features and performance studies mostly focus on co-authorship networks [11–13, 16–18], and fewer studies involve knowledge networks [14], and the measurement metrics are not comprehensive enough. In this paper, we use the centrality, clustering coefficients, and structural holes of knowledge network elements to explore their relationship with scientific research performance (productivity and impact).

(1) Centrality

(1) Degree centrality

Degree centrality is defined as the number of nodes that are directly connected to a node [11]. For node i in the network, its degree centrality is calculated as follows:

$$DC_i = \sum_{j=1}^n x_{ij} (i \neq j), \quad (2)$$

DC_i is the degree of node i , n is the number of nodes in the network, and j is all nodes in the network except node i . When i is adjacent to j , $x_{ij} = 1$, and when i is not adjacent to j , $x_{ij} = 0$.

(2) Betweenness centrality

The idea of betweenness centrality: if a node is located on multiple shortest paths of other nodes, then the node is at the core of the network and has a large betweenness centrality [11]. For node i in the network, its betweenness centrality is calculated as follows:

$$BC_i = \sum \frac{d_{mn}(i)}{d_{mn}}, \quad (3)$$

BC_i is the betweenness centrality of node i , $d_{mn}(i)$ is the number of shortest paths of node m and node n through node i , and d_{mn} is the number of shortest paths between node m and node n .

(3) Closeness centrality

Closeness centrality is the inverse of the cumulative shortest path distance from a node to all other nodes [11], which is calculated as follows:

$$CC_i = \sum_{j=1}^n \frac{1}{d(i, j)} \quad (i \neq j), \quad (4)$$

CC_i is the closeness centrality of node i , n is the number of nodes in the network, j is all the nodes in the network except node i , and $d(i, j)$ is the distance between node i and node j .

(2) *Clustering Coefficient*. Clustering coefficient is a coefficient used to describe the degree of clustering between nodes in a graph, i.e., the degree of interconnection between neighbors of a node. The clustering coefficient is divided into global clustering coefficient and local clustering coefficient [19]. In this paper, we study the clustering coefficient of nodes, i.e., local clustering coefficient. For an undirected graph $G=(V, E)$, V is the set of nodes and E is the set of edges. For node i , define the set of neighboring nodes as N_i , then the clustering coefficient of node i is

$$C_i = \frac{2|\{j, k \in N_i; e_{jk} \in E\}|}{k_i(k_i - 1)}. \quad (5)$$

Here, C_i is the clustering coefficient of node i , node j and node k are the neighboring nodes of node i , e_{jk} is the edge in the undirected graph G , and $k_i(k_i - 1)$ is the number of possible connections in the N_i .

(3) *Measuring PageRank*. PageRank algorithm, also known as network ranking algorithm, is used to measure the importance of web pages [20, 21]. Yu and Lu used PageRank algorithm to reveal the basic vocabulary within the discipline [22]. Gu and Xu proposed LTWPR (located and TF-weighted PageRank) algorithm based on PageRank algorithm, which can extract keywords of text more accurately [20]. The SQT-PageRank core patent discovery method proposed by Xuis is superior to the PageRank algorithm [23]. In this study, paper keywords are used to calculate ranking, that is, the text topic ranking algorithm TopicRank (TR) [24]. TR is defined as follows:

$$TR(k_i) = \frac{1-d}{N} + d * \sum_{k_j \in M(k_i)} \frac{TR(k_j) * Weight(k_j)}{Degree(k_j)}, \quad (6)$$

where $TR(k_i)$ is the TopicRank value of keyword i , d is the damping factor, which is generally 0.85 by default, N is the

total number of keywords, $M(k_i)$ is the set of keywords connected with keyword i , Degree (k_j) is the degree of keyword j , and Weight (k_j) is the weight of edge (k_i, k_j).

(4) *Structure Hole and Constraint*. Burt proposed the structure hole theory, which means that one or some individuals in a social network are directly connected to some individuals but not to others, i.e., there is no direct relationship or the relationship is intermittent, and the network as a whole appears as if there is a hole in the network structure [25].

In structural hole theory, the “limit degree” of an individual is the ability of that individual to use the structural hole in his or her own network. The larger the value is, the stronger the node’s constraint is. Thus, nodes have access to fewer sources of information, which is not conducive to generating structural advantages. For a node’s structural hole, it is obtained using 1-constraint. For node i and its neighborhood N_i , the constraint of i is calculated as follows [25]:

$$\text{Constraint}_i = \sum_{j \in N_i} \left(p_{ij} + \sum_q p_{iq} p_{qj} \right)^2, \quad q \neq i, j, \quad (7)$$

where j is the neighboring nodes of node i , q is the nodes other than node i and j , p_{ij} represents the proportion of direct links between nodes i and j to the total links of node i , and $p_{iq} p_{qj}$ refers to the proportion of links of node i indirectly connected to node j through node q to the total links of node i .

2.2.3. *Measuring Scientific Performance*. There are two common ways to measure scientific performance: productivity and impact [18, 26, 27]. We use the average citation count to measure the impact of knowledge network elements, and the number of articles of knowledge network elements is chosen as the productivity measure.

3. Data Statistics

Figure 1 shows the trend of the number of papers issued by the Q1 area journals of sports science. From 2000 to 2020, the number of papers issued by the Q1 area journals of sports science in China was 45472. From the figure, we can find that the annual numbers of paper showed a trend of rising and then falling. Before 2007, it had an upward trend. In 2007, a total of 2998 papers were published. Then, the number of annual papers decline after 2007. A study was conducted using the volume of articles published in 567 journals (approximately 1.05 million papers) across 25 disciplines included in the CSSCI during the period 2010–2019. It finds that the number of C-journal papers in all disciplines declined from 2010 to 2019. The decline in sports science, political science, economics, library, intelligence and literature, etc., is obvious. The number of papers published in sport science journals decreased by 43%, which was the largest decrease among all subjects. The decline in the number of articles published in C-journal may be due to a

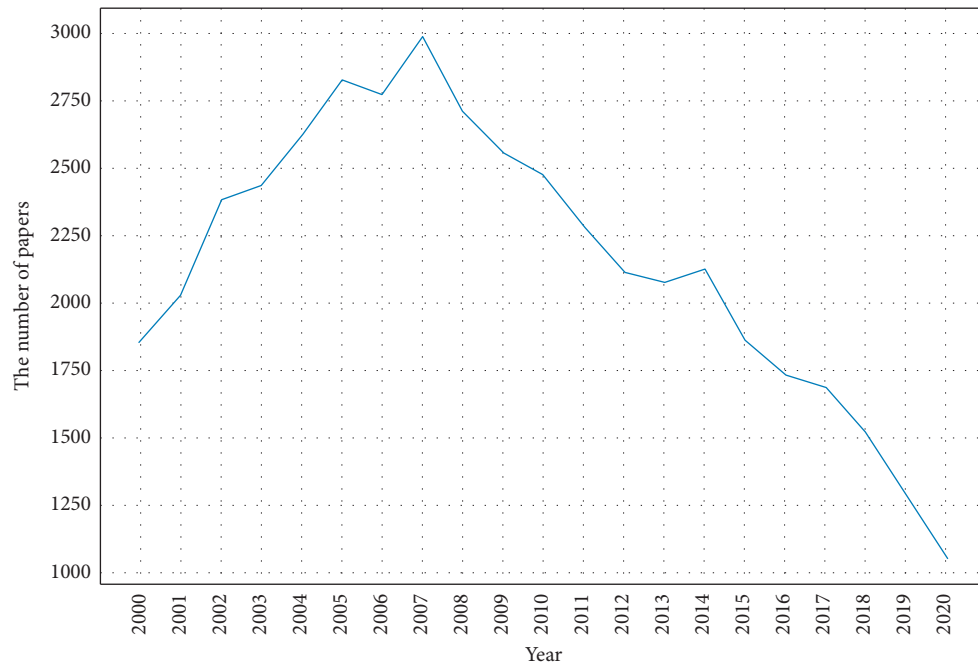


FIGURE 1: The trend of papers (2000–2020).

shift in the journal’s strategy from pursuing high production to high quality.

4. Analysis of Research Hotspots

With the development of social network research, social network visualization software has gradually increased, such as UCINET, CiteSpace, Pajek, and VOSviewer. The advantage of VOSviewer is the “cooccurrence clustering,” i.e., the simultaneous occurrence of two things indicates their relatedness to a certain extent. The more the number of simultaneous occurrence is, the greater the relatedness is [28]. Run the VOSviewer software for cooccurrence of the processed data, as shown in Figure 2. From 2000 to 2020, the keywords with frequency greater than 150 were screened, and the keyword pairs with cooccurrence frequency greater than 3 were constructed to cooccurrence network.

Table 1 is the calculation of the dividing line formula of high- and low-frequency keywords based on Donohue, and the table of high-frequency keywords whose keyword frequency exceeds the minimum threshold of 268. In the keyword frequency statistics, we removed the keywords with low specific reference (such as China, sports, research, analysis, influence, etc.) and merged the synonyms (such as 2008 Olympic Games, 29th Olympic Games, and Beijing Olympic Games; colleges and universities; and the Olympic Movement and the Olympics)

4.1. Microperspective: Hotspots Analysis. Table 1 shows that the research hot topics of sports science are competitive sports, school sports, sports management, physical education, mass sports, sports culture, national traditional sports, sports industry, sports history, sports economy, sports teaching, Olympic movement and Olympic Games,

national fitness, rat, college sports, animals experiment, etc. The high-frequency keywords of sports management, physical education, national traditional sports, sports history, and sports economy have obvious disciplinary nature, while competitive sports, school sports, mass sports, sports culture, sports industry, physical education, and animal experiments are all classical directions of sports science research in China. The research hot groups are college students, athletes, teenagers, and physical education teachers. The research hot programs include traditional national sports of martial arts, competitive sports of football, and table tennis.

Competitive sports is the keyword with the highest frequency in the recent 20 years. Since the development of competitive sports in China, competitive sports has always been the important research direction of sports in China and has become an absolute hotspot in the field of sports research in China. As shown in Figure 3, the competitive sports showed a trend of rising and then falling, with a rapid increase around 2008. By analyzing the papers about competitive sports in the past 20 years, the research on competitive sports has become an absolute upsurge in this Olympic cycle since the Beijing Olympic Games was held in China. As time goes by, the popularity of competitive sports has declined. The research about competitive sports focuses on the high-quality development of competitive sports in the new era [29], technology-led competitive sports [30], and the Winter Olympics [31, 32], and it is not limited to the study of the development status and path of competitive sports in the perspective of the Olympic Games.

Figure 4 shows the trend of keyword, mass sports. Mass sports is also one of the research hotspots in the Q1 area of sports science in China from 2000 to 2020. Before 2008, there was less research on mass sports, and after 2008, it

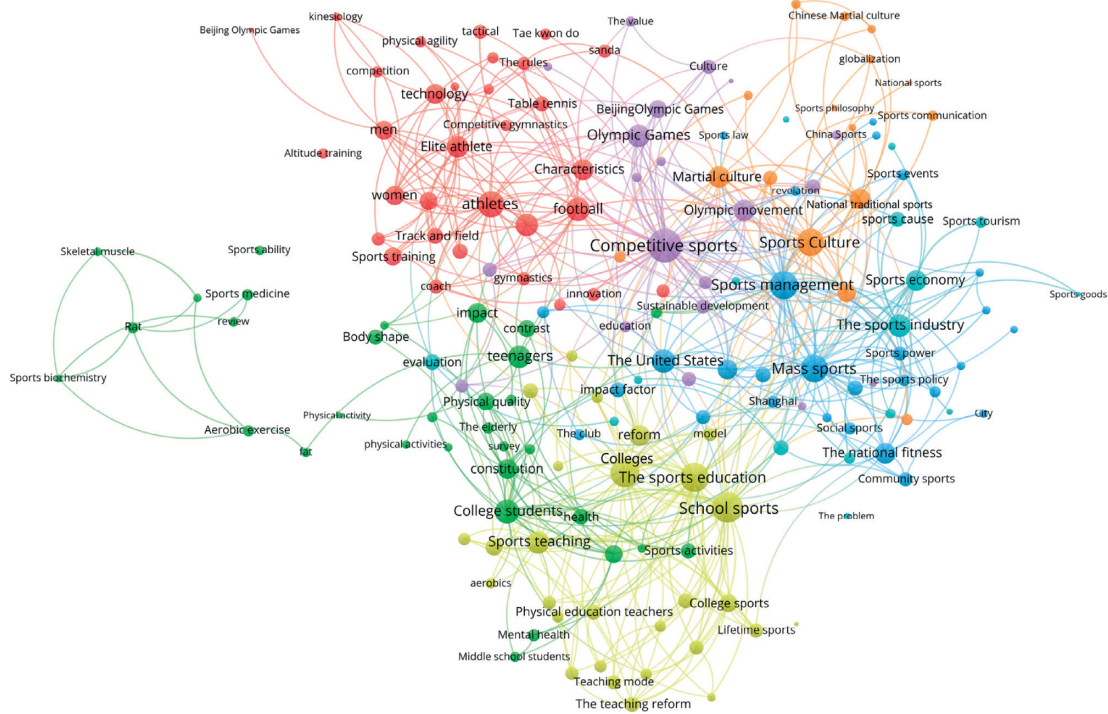


FIGURE 2: Keyword cooccurrence network.

TABLE 1: High-frequency keywords.

No.	Keyword	Frequency
1	Competitive sports	1853
2	School sports	1316
3	Sports management	1262
4	Physical education	1243
5	Mass sports	1240
6	Sports culture	1104
7	Traditional national sports	993
8	Colleges	962
9	Sports industry	924
10	College students	839
11	Athletes	808
12	Martial arts	797
13	Sports history	797
14	Sports economy	778
15	Sports teaching	741
16	Olympic movement	590
17	The Olympic games	589
18	Football	585
19	National fitness	511
20	Beijing Olympic games	496
21	The United States	471
22	Rat	459
23	Teenagers	446
24	Basketball	446
25	Excellent athletes	411
26	Sports training	387
27	Physical education curriculum	373
28	PE teachers	362
29	Women	343
30	Physical exercise	329

TABLE 1: Continued.

No.	Keyword	Frequency
31	Constitution	319
32	Sports sociology	299
33	Japan	297
34	Men	287
35	Culture	286
36	College sports	283
37	Animal experiments	280
38	Table Tennis	277

entered an upward phase, reaching its peak in 2014–2015. In recent years, many Chinese scholars have focused their research on mass sports, such as national fitness [33] and leisure sports [34]. The research in this field is in line with the characteristics of this era. However, it is different from competitive sports. The research focus of mass sports is still in its initial stage, with the majority of qualitative research, such as countermeasures research [33, 35] and development research [34].

Figure 5 is the trend of keyword, sports management. Sports management is one of the disciplines of sports science. It applies management theories and methods to study the coordination of sports organizations in order to achieve the predetermined goals of sports [36], thus came into being related research on sports management. To be specific, sports management is an activity process in which managers in sports organizations coordinate the activities of others and play the role of various resources to achieve predetermined goals through the implementation of planning,

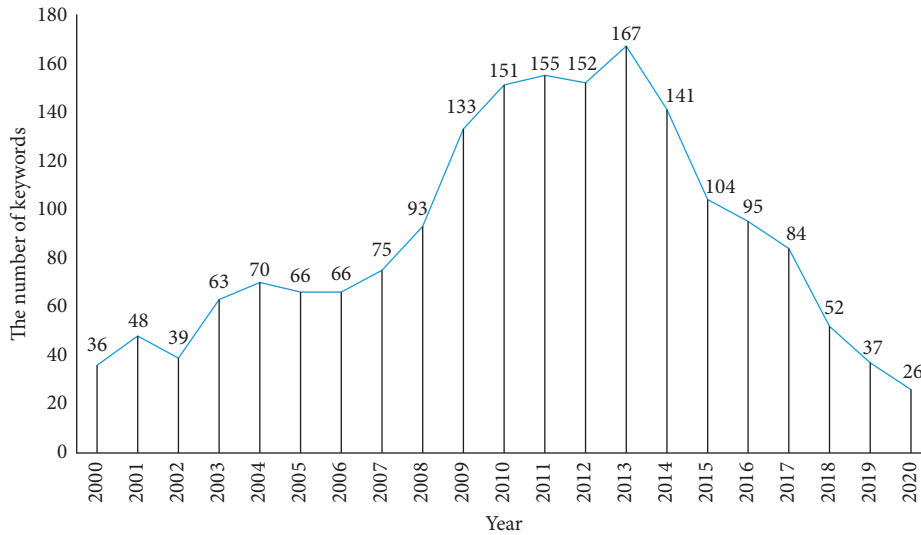


FIGURE 3: The trend of keyword, competitive sports.

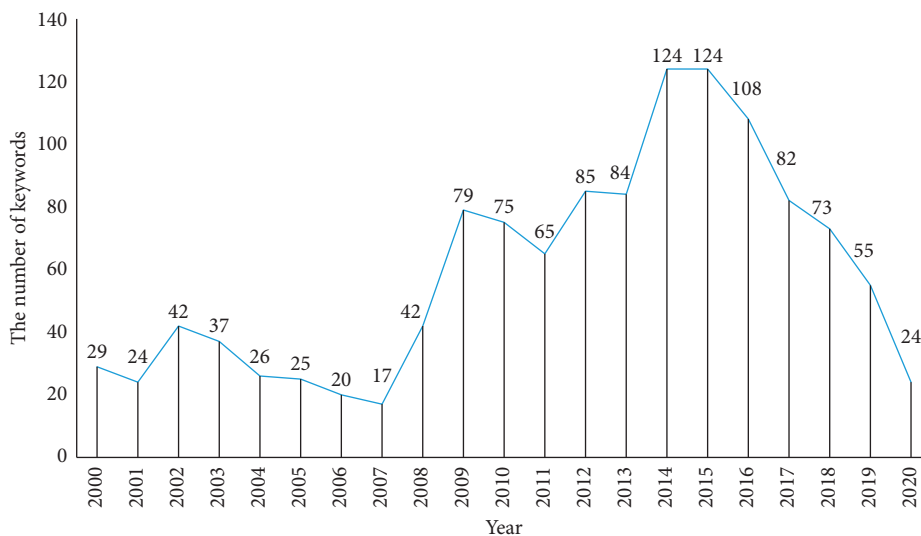


FIGURE 4: The trend of keyword, mass sports.

organization, leadership, and control functions for the object of sports management [36]. In sports management, mass sports management, competitive sports management, and school sports management are the main research contents [37]. The change trend of sports management is same as that of mass sports and competitive sports. After 2008, the number of its related studies increased rapidly and then declined after 2013–2014. Compared with school sports, there are more researches on mass sports and competitive sports in sports management. The research on competitive sports in sports management focuses on the research on management system [38, 39] and the research on development [40, 41], which deeply analyzes the institutional problems of competitive sports, optimizes the development path of competitive sports, and promotes the vigorous development of competitive sports. Research on mass sports in sports management focuses on sports policies [42, 43],

community sports [44], and the elderly sports and the disabled sports [45, 46]. From the perspective of management, it analyzes how China develops mass sports and how to give correct policy guidance to mass sports.

Figure 6 shows the trend of keyword, school sports and physical education. As a keyword of high frequency, school sports has always been one of the research key points in the field of sports science. As an important branch of sports science, school sports and its derivative keywords account for almost half of the high-frequency keywords. As shown in Figure 6, the emergence of school sports is bound to accompany the emergence of physical education, and they complement each other and develop together. In the past 20 years, it has been found that the research on school sports has a fluctuation. Many scholars' researches on school sports have mainly focused on teaching reform [47–49]. With the continuous development of sports teaching reform research,

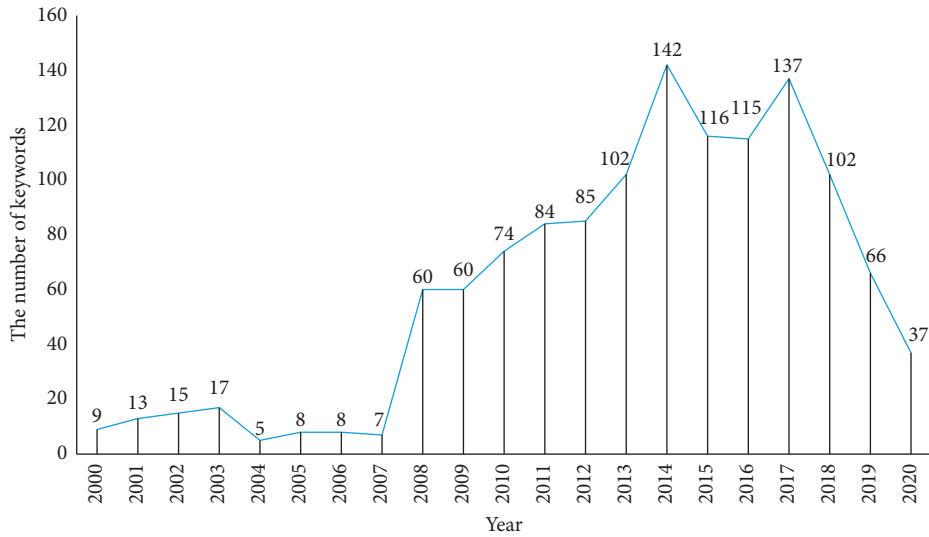


FIGURE 5: The trend of keyword, sports management.

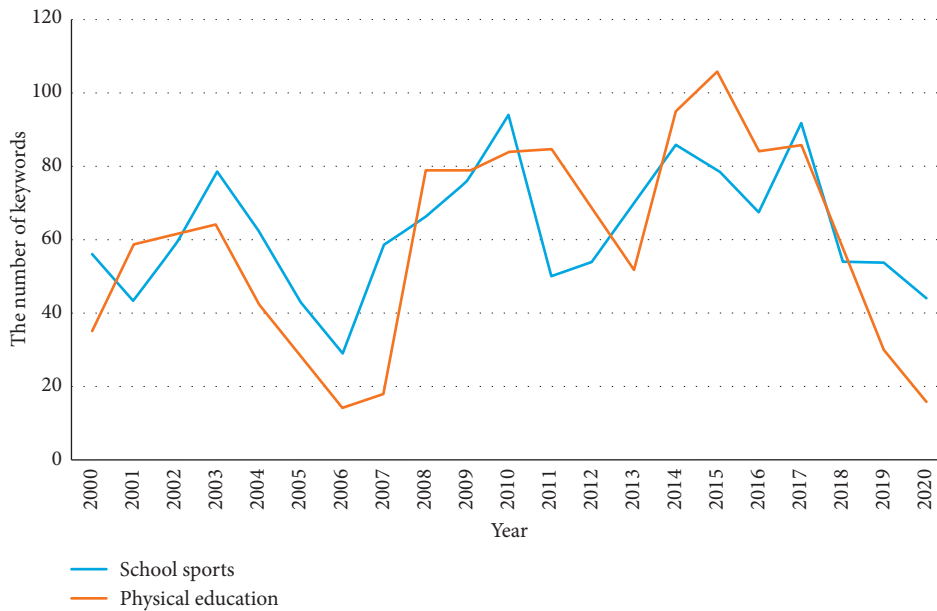


FIGURE 6: The trend of keyword, school sports and physical education.

almost all aspects of school sports are covered from teaching idea to teaching pattern.

4.2. *Macroperspective: Cluster Analysis.* According to the clustering results of VOSviewer in Figure 2, it can be seen that the research hotspots of sports journals in Q1 area of China from 2000 to 2020 mainly focus on six categories: red clustering belongs to the category of sports training; purple clustering belongs to the category of competitive sports; orange clustering belongs to the category of national traditional sports; sports medicine and sports biochemistry of green clustering belong to the category of sports human science; blue clustering belongs to the category of mass sports and sports management; cyan clustering belongs to

the category of sports industry and sports economy; yellow clustering belongs to the category of school sports and physical education.

- (1) Sports training of red clustering and physical education of yellow clustering belong to physical education and training. Physical education and training is an interdisciplinary and comprehensive discipline composed of traditional physical teaching theories and methods and sports training, which mainly research the system of basic theories and methods of physical education and sports training [50]. As a subdiscipline of physical education and training, sports training is an important part of competitive sports. Athletes are trained scientifically under the

guidance of coaches to improve their competitive level and sports results, so as to show their ability and spirit in the competition and win glory for the country. Physical education achieves the goal of education by promoting people's physical and mental development through sports; its research focuses on school sports and college physical education.

- (2) The orange clustering belongs to national traditional sports. Traditional sports is a kind of fitness and entertainment activity with strong ethnic cultural color. It mainly focuses on physical movement and the understanding of the human body of various nationalities and also a special education method used by people to seek to enhance physical skills training [51]. According to the clustering results of VOSviewer, the research topics of national traditional sports are mainly composed of sports culture, traditional sports, and the Olympic Games. The inheritance and internationalization of traditional national sports have always been the focus of traditional sports research in China.

Fan and Yu studied the inheritance of traditional martial arts and analyzed its current situation, problems, and countermeasures [52]. Li and Guo studied the inheritance and international dissemination of martial arts. They pointed out that martial arts culture should give full play to its advantages and be the pioneer in the process of international dissemination of Chinese martial arts and explore the frontier position of the international promotion of Chinese martial arts [53]. In addition to the research on the inheritance and internationalization of martial arts, the Chinese sports science has a lot of researches on the Olympic-entering of the traditional national sport, martial arts. Its essence of those researches is still the internationalization and inheritance of martial arts. Liu et al. deeply analyzed the difficulty of Olympic-entering of martial arts because of the low popularity of martial arts and the esoteric martial arts culture [54]. Hong et al. pointed out that the cultural differences between China and Western countries and the degree of international development made martial arts difficult to become an official event in the Olympic Games [55]. As one of the traditional sports programs, we should take the cultural connotation as the core, promote its traditional spirit, popularize the education of martial arts, and move towards the forefront of the internationalization of national traditional sports.

- (3) Green clustering belongs to sports human science. Sports medicine and sports biochemistry are subordinate to the secondary disciplines of sports human science. Sports human science is a subject which applies the theory and method of anatomy, physiology, nutrition, and biochemistry to study the influence of sports on human body shape, structure, and physiological function, as well as the law and

measure of health care in sports. As one of the secondary disciplines of sports science, human sports science has fewer hot topics compared with physical education and training, sports humanities and sociology, and national traditional sports from the perspective of knowledge map of keywords cooccurrence.

- (4) The purple, blue, and cyan clustering all belong to sports humanities and sociology. Sports humanities and sociology is a comprehensive discipline developed on the basis of the humanities and social sciences, which studies the essential issues of sports, such as the relationship between sports and human beings, sports and society, and the basic laws, through the theories and methods of the humanities and social sciences [56]. The main research directions of sports humanities and sociology include competitive sports, school sports, social sports (national fitness problem), sports industry, and market, and sports management is also one of the research directions of sports humanities and sociology [57].

Competitive sports are a process of sports activities that maximizes the potential of athletes in physical, psychological, and intellectual aspects; while developing the body comprehensively, the main purpose is to climb the peak of sports technique and create excellent sports performance [58]. Competitive sports is an important way to highlight the national sports strength. Athletes as the main body of competitive sports, its cultivation and training have always been a hot topic [59, 60].

School sports is a discipline that combines sports with education and cultivate teachers to follow the national policy of all-round development of moral, intellectual, physical, social, and aesthetic education. Based on the characteristics of the students' physical and mental development, by means of proper physical exercises and health knowledge, through the physical education curriculum, physical exercise, sports competition, and so on, school sports is a planned and organized education activity, which dedicates to strengthen students' physique and cultivate students' consciousness, interest, habit and ability of lifetime sports, and help them become a socialist builders and defenders with comprehensive development in moral, intelligence, sports, aesthetics, and labour education [61]. Related concepts of physical education run through the work of school sports; the research topics are also centered on the two large pieces of education and teaching. Education pays attention to process, teaching focuses on results, and research focuses on teaching mode, effect, and reform, promoting the development of school sports to achieve the aim of students' all-round development. For example, Zhou et al. pointed out that the close connection between teaching materials and society and life should be strengthened in their research on the reform of physical education content in universities, and that there should be not only competitive programs but also fitness, entertainment, and traditional

Chinese national sports, so as to realize a good transition from school sports to social sports [62].

Sports Management and Mass Sports. Sports management is a discipline that applies management theories and methods to study the coordination of sports organizations to achieve predetermined sports goals [37]. Mass sports refer to sports activities with a wide range of contents and various forms that are voluntarily participated by ordinary people for the purpose of physical improvement, physical fitness, entertainment, leisure, and social interaction and generally do not seek to achieve high level of athletic performance [63]. Research on mass sports in sports management focuses on analyzing the realistic conditions and restrictive factors of mass sports in China from the perspective of management, so as to optimize the development path of mass sports in China [42–46].

Sports Industry and Economy. Sports industry refers to the collection of the same kind of economic activities and the synthesis of the same kind of economic sectors that provide sports products for the society. Although China’s sports industry started late, it has developed rapidly. The field and scale of the industry have been expanded; the quality benefits have also been improved significantly. On the basis of the sound modernization system of sports industry, the high-quality development of sports industry has become the focus of research. Li and Liu analyzed the development concept, topics context, and logical thinking of the sports industry [64]. Guo and Ren analyzed the inner logic and basic connotation of the high-quality development in the new era and pointed out the path selection of high-quality development of sports industry: promoting the supply-side structural reform in sports industry, strengthening sports market entity, perfecting construction of sports market system, promoting the efficient integration of sports industry and other industries and the spatial distribution optimization, and improving the system of development and policy in sports industry [65].

To sum up, the research hotspots of Q1 area journals in the past 20 years involve physical education and training, sports humanities and sociology, sports human science, and national traditional sports.

5. Analysis of the Relationship between Knowledge Networks and Scientific Performance Correlations

5.1. Descriptive Statistics of Variables. Table 2 is a descriptive statistical table of variables. For the sample, the mean value of the number of articles is 28.6, standard deviation 85.7, minimum value 6, and maximum value 2407. The mean value of citation counts is 21.7. The mean values of sample degree centrality, betweenness centrality, and closeness centrality are 42.7, 2994.6, and 0.4, respectively. The mean value of PageRank for the sample is only 0.00024, with a standard deviation of 0.00045. The means of clustering coefficient and structural holes are 0.23 and 0.945, respectively.

TABLE 2: Descriptive statistics of variables.

Variable	Mean	Std. dev.	Min	Max
Paper counts	28.64119	85.66474	6	2405
Average citation	21.68361	11.71496	2.5	148.25
Degree centrality	42.72149	68.7952	3	1537
Betweenness centrality	2994.606	22082.97	0.723307	1012664
Closeness centrality	0.4125856	0.038428	0.286623	0.3840
PageRanks	0.0002411	0.000454	0.000049	0.013906
Clustering	0.2286381	0.0823648	0	0.7
Structural holes	0.9457353	0.0335163	0.6375047	0.9974905

*Obs = 4147.

5.2. Correlation Analysis of Independent Variables.

Table 3 shows the correlation test results of the independent variables and the correlation coefficients between closeness centrality and degree centrality; betweenness centrality and degree centrality are high; closeness centrality and betweenness centrality Pearson correlation coefficient is low, only 0.3321. Leydesdorff measured degree centrality and betweenness centrality in a network with sample size of 7379 and degree centrality and closeness centrality, and closeness centrality and betweenness centrality correlation coefficients were 0.509, 0.651, and 0.210, respectively. From the table, we can also find that the correlation coefficients between PageRank and three centralities are high, which is an expected result. The PageRank uses the degree of the neighboring nodes of the direct node to calculate and thus the two show high correlation. From the variance inflation factor (VIF) test, we can find that there is a serious problem of multicollinearity in the model, so we removed the variables degree centrality and PageRanks. After removing the variables and conducting the VIF test again, the VIF of all variables is less than 5.

5.3. Associations between Knowledge Network and Research Influence.

Table 4 shows the multiple regression analysis of knowledge network features and performance. From the table, we can learn that only the correlation of betweenness centrality and average citation counts is not significant, and the rest of the results are significant at different levels. Although the correlation of betweenness centrality and paper counts is significant, the correlation coefficient was only 0.0032. In the multiple regression of knowledge network features and impact, we can find that when a node is in the core position of the network, i.e., when the closeness centrality goes higher, its impact will increase. Similarly, when the nodes of a knowledge network are clustered into groups or in the structural hole position, they are conducive to the higher citations and increased impact. In the multiple regression of knowledge network features and productivity, closeness centrality also shows a significant positive correlation with productivity. Interestingly, clustering coefficients and structural holes present different results from those in the impact analysis. When knowledge network nodes cluster into groups with other nodes, they instead reduce their productivity, and when nodes are in structure holes, their

TABLE 3: Independent variable correlation test and VIF.

		1	2	3	4	5	VIF
1	Degree centrality	—					39.75
2	Betweenness centrality	0.8139	—				8.41
3	Closeness centrality	0.6509	0.3321	—			3.86
4	PageRanks	0.9723	0.8996	0.5446	—		54.20
5	Clustering	-0.5139	-0.2454	-0.4942	-0.4341	—	2.01
6	Structural holes	0.4862	0.1788	0.7743	0.3840	-0.6369	3.32

TABLE 4: Multiple regression analysis.

Knowledge network measure (Obs = 4147)	Average citation		Paper counts	
	β	Std. err.	β	Std. err.
Betweenness centrality	-0.0000146	8.69e-06	0.0031839**	0.0000254
Closeness centrality	65.16445**	7.646053	394.5624**	22.35862
Clustering	10.01557**	2.826052	-129.9039**	8.263953
Structural holes	29.53742*	9.575936	-151.2**	28.00199

*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

productivity decreases. This may be due to the fact that the nodes in the structure hole position are only associated with a small number of nodes and are not hotspots.

6. Discussion and Conclusion

The research in this paper is helpful to understand the research status quo and development situation of the field and provide decision-making and reference information for the selection of thesis topics, research projects, and discipline planning in sports science. Through keywords frequency analysis and social network analysis method, we can draw the following conclusions:

First, through word frequency method, we find 38 research hotspots of Q1 area journals of sports science in China, such as competitive sports, school sports, sports management, sports education, and mass sports. There are four research focus groups: college students, athletes, teenagers, and physical education teachers, and we found three hot research programs: martial arts in traditional national sports and football and table tennis in competitive sports. Therefore, it is necessary to understand the current situation of sports science and maintain the development of physical education and training, sports humanities, and sociology and national traditional sports. At the same time, we need to accelerate the development of sports human science, which is the focus of future research.

Secondly, the characteristics of knowledge network elements were significantly associated with scientific research performance, except for betweenness centrality, which was not significantly related to average citation.

Data Availability

The original data used in this paper are from the Chinese Social Sciences Citation Index Database.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Linxiao Ma was responsible for writing the original draft, review and editing, and literature collection. Yuzhu Wang was responsible for determining the framework of paper, funding acquisition, and review and editing. Yue Wang was responsible for writing the original draft. Ning Li was responsible for review and editing. Sai-fu Fung was responsible for funding acquisition. Lu Zhang and Qian Zheng did calculations. All authors discussed and interpreted results.

Acknowledgments

This work was supported in part by the National Key R&D Plan (2020YFC2007003).

References

- [1] D. Lazer, A. Pentland, L. Adamic et al., "Social science: computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.
- [2] Y. Chen, W. Tian, and J. Wu, "Research on the visual analysis method of topic domain research hotspots tracking and trend forecasting," *Information Theory and Practice*, vol. 40, no. 6, pp. 117–121, 2017.
- [3] F. Luo and B. Lu, "Dynamic analysis of China's circular economy research from the perspective of knowledge mapping," *Journal of Statistics and Information*, vol. 32, no. 3, pp. 109–113, 2017.
- [4] J. Chu and Q. Qian, "Research on knowledge management based on word frequency analysis in recent 10 years," *Information Science*, vol. 32, no. 10, pp. 156–160, 2014.
- [5] Y. Li and C. Jiang, "Hotspots and future trends of physical education research in China in recent five years," *Journal of Wuhan Institute of Physical Education*, vol. 53, no. 4, pp. 19–25, 2019.

- [6] L. Jiang, Z. Wang, and X. Wang, "Key words analysis of sports humanistic sociology papers based on CSSCI," *Journal of Southwest University for Nationalities (Humanities and Social Sciences Edition)*, vol. 35, no. 1, pp. 229–238, 2014.
- [7] J. C. Donohue, *Understanding Scientific Literatures: A Bibliometric Approach*, Massachusetts Institute of Technology Press, Cambridge, MA, 1972.
- [8] C. Lyndon, *Freeman. The History of Social Network Analysis*, China People's University Press, Beijing, China, 2008.
- [9] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 10, pp. 3804–3816, 2020.
- [10] Y. Hu, F. Xiong, S. Pan, X. Xiong, L. Wang, and H. Chen, "Bayesian personalized ranking based on multiple-layer neighborhoods," *Information Sciences*, vol. 542, pp. 156–176, 2021.
- [11] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures," *Journal of Informetrics*, vol. 5, no. 4, pp. 594–607, 2011.
- [12] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: a social capital perspective," *Research Policy*, vol. 42, no. 9, pp. 1515–1530, 2013.
- [13] E. Yan and D. Ying, *Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009.
- [14] J. Guan, Y. Yan, and J. J. Zhang, "The impact of collaboration and knowledge networks on citations," *Journal of Informetrics*, vol. 11, no. 2, pp. 407–422, 2017.
- [15] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.
- [16] A. Abbasi, R. T. Wigand, and L. Hossain, "Measuring social capital through network analysis and its influence on individual performance," *Library & Information Science Research*, vol. 36, no. 1, pp. 66–73, 2014.
- [17] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, vol. 50, no. 10, pp. 4186–4199, 2020.
- [18] S. Guoan, J. Luo, J. Sun, F. Rong, and C. Zhang, "Analysis of the academic impact of articles in journal of Xi'an Jiaotong university (social science edition)," *Journal of Statistics and Information*, vol. 24, no. 07, pp. 92–96, 2009.
- [19] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.
- [20] Y. Gu and M. Xu, "News keyword extraction algorithm based on PageRank," *Journal of University of Electronic Science and Technology of China*, vol. 46, no. 5, pp. 777–783, 2017.
- [21] X. Guo and X. Zhang, "Hot topics and future prospects of education research on "one belt and one road"-based on CSSCI literature analysis (2013–2019)," *Journal of Xi'an University of Finance and Economics*, vol. 33, no. 4, pp. 98–104, 2020.
- [22] F. Yu and W. Lu, "Key words co-occurrence network perspective of discipline basic vocabulary discovery," *Library and Information Service*, vol. 63, no. 9, pp. 95–100, 2019.
- [23] C. Xu, *Research on Core Patent Discovery Based on Improved PageRank Algorithm*, Shanxi University, Taiyuan, China, 2020.
- [24] H. Zhao, "Centrality and power embodiments: a study on the generation path of network media power based on social network analysis," *Journalism and Communication Research*, vol. 20, no. 03, pp. 50–63+127, 2013.
- [25] R. S. Burt, *Structural Holes*, Harvard University Press, Cambridge, MA, USA, 1992.
- [26] J. Guan and L. Pang, "Bidirectional relationship between network position and knowledge creation in Scientometrics," *Scientometrics*, vol. 115, no. 1, pp. 201–222, 2018.
- [27] C. Damien, D. Arnaud, L. Catherine et al., "The impact of a researcher's structural position on scientific performance: an empirical analysis," *PLoS One*, vol. 11, no. 8, Article ID e0161281, 2016.
- [28] L. Zuo and X. Xiao, "Comparison of knowledge graph visualization tools VosViewer and NWB tool," *Information Science*, vol. 33, no. 2, pp. 95–99, 2015.
- [29] G. Yang, "On the new development of Chinese competitive sports in the new era," *Sports Culture Guide*, vol. 3, pp. 11–16, 2019.
- [30] W. Tian and F. Lianshi, "Research on the relationship between sports science and technology innovation and competitive sports training level," *Journal of Beijing Sport University*, vol. 6, pp. 825–827, 2003.
- [31] X. Yao, "Experience and inspiration of ice and snow resources development in host cities of the Winter Olympic Games," *Sports Culture Guide*, vol. 6, pp. 18–23, 2019.
- [32] Y. Du and B. Sun, "Research on the sustainable development of winter Olympic games host area: a case study of whistler resort in Vancouver," *Sports Culture Guide*, vol. 2, pp. 23–28, 2018.
- [33] M. Chen, "Basic countermeasures for implementing national fitness program in developed and underdeveloped cities," *Journal of Physical Education*, vol. 4, pp. 24–26, 2000.
- [34] L. Luo, "On the development of leisure sports industry in China from the interactive relationship between industry and culture," *Journal of Beijing Sport University*, vol. 12, pp. 1645–1647, 2006.
- [35] J. Yu and X. Zuo, "A comparative study of leisure sports thought between China and west," *Sports Culture Guide*, vol. 6, pp. 68–69, 2008.
- [36] L. Burney, *Sports Management: Fundamentals and Applications*, East China Normal University Press, Shanghai, China, 4th edition, 2009.
- [37] Q. Zuo, Z. Huang, J. Su et al., *Sports Management*, Beijing Normal University Press, Beijing, China, 2010.
- [38] Lu Yuan, "Institutional cost of the current management system of competitive sports in China," *Journal of Physical Education*, vol. 17, no. 3, pp. 7–12, 2010.
- [39] Y. Zhu, M. Su, B. Dai, and H. Huang, "Research on competitive advantage of Chinese competitive sports," *Sports Culture Guide*, vol. 2, pp. 31–35, 2010.
- [40] X. Han, H. Shen, and B. Zheng, "Research on the development of Chinese competitive sports in 60 years after the founding of the People's Republic of China," *Sports Culture Guide*, vol. 8, pp. 55–57, 2009.
- [41] B. Huang, "Theory of factors in the development of competitive sports system in China," *Journal of Sport Culture Tribune*, vol. 5, pp. 5–8, 2015.
- [42] T. Wen, "Mass sports development strategy research in China," *Journal of Sport Culture Tribune*, vol. 9, pp. 4–7, 2010.
- [43] H. Dong and X. Fang, "Consideration on the benefit coordination between mass sports and competitive sports," *Sports Culture Guide*, vol. 8, pp. 16–24, 2009.
- [44] Z. Jin-guo, "Social stratification and community sports development," *Sports Culture Guide*, vol. 9, pp. 34–37, 2011.

- [45] Y. Liu, W. Tang, J. He, and H. Pan, "Research on sports rights of disabled persons in China," *Sports Culture Guide*, vol. 3, pp. 17–20, 2010.
- [46] Y. Liu, "De-structuring and reconstruction of public service system of sports for the elderly in China," *Sports Culture Guide*, vol. 2, pp. 5–8, 2014.
- [47] Z. He, "College physical education reform and college students physical education ability cultivation," *Journal of Wuhan Institute of Physical Education*, vol. 3, pp. 110–112, 2000.
- [48] B. Gao, M. Xu, R. Li, and B. Wang, "The reform of physical education teaching evaluation in colleges and universities," *Journal of Physical Education*, vol. 6, pp. 77–80, 2003.
- [49] X. Xie and J. Mao, "Thoughts on the path of sports reform in China," *Sports Culture Guide*, vol. 3, pp. 5–7, 2009.
- [50] M. Jing, "On the research object and the nature of the discipline of physical education and training," *Sports and Science*, vol. 33, no. 5, pp. 104–107, 2012.
- [51] L. Cui, "On national traditional fitness sports and national fitness sports," *Journal of Shandong Institute of Physical Education*, vol. 4, pp. 44–50, 1998.
- [52] T. Fan and D. Yu, "The current situation, problems and countermeasures of traditional martial arts inheritance - based on the perspective of intangible cultural heritage," *Journal of Nanjing Sports College (Social Science Edition)*, vol. 29, no. 1, pp. 27–31, 2015.
- [53] J. Li and Z. Guo, "Regional martial arts heritage and international dissemination of Chinese martial arts: history and reality of martial arts culture in the Lingnan Pearl River Delta," *Journal of Wuhan Institute of Physical Education*, vol. 44, no. 3, pp. 56–60, 2010.
- [54] X. Liu, W. Mao, and Y. Bai, "The difficulty of martial arts entering Olympic Games," *Journal of Shandong Institute of Physical Education*, vol. 2, pp. 31–32, 2005.
- [55] H. Hong and L. Zhang, "Reflections on some theoretical issues of martial arts entering in the Olympic Games," *Journal of Sports Culture*, vol. 9, pp. 49–51, 2007.
- [56] Lu Yuan, "Discipline integration and research frontier of sports humanities and sociology," *Journal of Physical Education*, vol. 1, pp. 4–7, 2005.
- [57] L. Y. Zhen, *Sports Sociology*, Page Higher Education Press, Beijing, China, 3rd edition, 2010.
- [58] Z. Yan, Z. He, and X. Li, "Research on the training path of reserve talents in football power," *Sports Culture Guide*, vol. 8, pp. 26–28, 2007.
- [59] R. Liu and L. Pang, "Research on the training of reserve talents in competitive sports in China," *China Sports Science and Technology*, vol. 53, no. 4, pp. 42–47, 2017.
- [60] Jinyu, S. Pan et al., "Current situation and countermeasure on cultivation of reserve talents of competitive sports in China," *And Sports and Science*, vol. 5, pp. 82–86, 2006.
- [61] Y. Huang, *Journal of Guangzhou Institute of Physical Education*, vol. 36, no. 3, pp. 121–124, 2016.
- [62] P. Zhou, R. Yu, and L. Dong, "Reform of physical education teaching content in colleges and universities," *Physical Culture and History*, vol. 6, p. 44, 2000.
- [63] W. Xing, "Analysis on the current situation and development trend of mass sports research in China," *Martial Arts Research*, vol. 4, no. 2, pp. 137–147, 2019.
- [64] R. Li and N. Liu, "Theoretical framework and logical path: research on high-quality development of sports industry in China," *Journal of Tianjin Institute of Physical Education*, vol. 35, no. 6, pp. 651–657, 2020.
- [65] H. Guo and B. Ren, "High-quality development of Chinese sports industry in the new era: logical generation and path selection," *Journal of Xi'an Institute of Physical Education*, vol. 37, no. 3, pp. 291–297, 2020.

Research Article

Power Control Algorithm Based on a Cooperative Game in User-Centric Unmanned Aerial Vehicle Group

Yuexia Zhang ^{1,2,3} and Pengfei Zhang ¹

¹School of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing 100101, China

²Key Laboratory of Modern Measurement & Control Technology, Ministry of Education, Beijing Information Science & Technology University, Beijing 100101, China

³Beijing Key Laboratory of High Dynamic Navigation Technology, University of Beijing Information Science & Technology, Beijing, 100101, China

Correspondence should be addressed to Yuexia Zhang; zhangyuexia@bistu.edu.cn

Received 8 April 2021; Accepted 11 May 2021; Published 21 May 2021

Academic Editor: Fei Xiong

Copyright © 2021 Yuexia Zhang and Pengfei Zhang. 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 quality of service (QoS) of a user in user-centric unmanned aerial vehicle group (UUAVG) is degraded by complex cochannel interference; hence, a cooperative game power control (CGPC) algorithm in UUAVG is proposed. The algorithm helps to establish a downlink power control model of the UUAVG, construct a product of the signal to interference noise ratio function of each user as a utility function of the cooperative game, and deduce the optimal power control scheme using the Lagrange function. This scheme reduces the interference of the service unmanned aerial vehicle (UAV) to edge users and improves the communication quality of all the users as well as the throughput of the entire system. Simulation results show that the average throughput of the CGPC algorithm improved by 10.32% compared with the traditional Stackelberg Game based Nonuniform Pricing Power Control (SGNPPC) algorithm. This shows that the CGPC algorithm can effectively reduce the transmission power of the cooperative UAV and enhance the capacity of the entire system, ensuring communication quality.

1. Introduction

With the rapidly increasing number of wireless devices and growth of user traffic demand, improving the throughput of a communication system has gained considerable attention [1]. Wireless networks are effective solutions to improve system throughput because a large number of unmanned aerial vehicles (UAVs) groups can be deployed. The user-centric UAV group (UUAVG) can not only improve system capacity but also achieve seamless coverage and enhance the quality of service (QoS) of the system [2]. The UUAVG network organises a dynamic UAVG for each user; the unmanned aerial vehicles group (UAVG) is composed of all potential UAVs near a user, so that the user can perceive that the network follows them [3]. UUAVG is defined as the network architecture of serving user by the “decellular”

method [4]. When the user moves, the UAVG will dynamically add and delete UAV members according to the user location to meet user requirements in the UAVG network centre [5]. Compared with traditional cellular network, UUAVG network structure can effectively improve network coverage, reduce the interrupt probability by 30%, and improve the spectrum efficiency by 5%–15% [6, 7]. Although the UUAVG improves the system throughput by deploying a large number of UAVs, the cochannel interference problem is severe and reduces the QoS for the users. Therefore, it is essential to realise methods to reduce this cochannel interference.

At present, many scholars have studied the complex interference problem of UUAVG, and power control is regarded as the most effective interference suppression method. In [8], an orthogonal frequency division

multiplexing- (OFDM-) based UAVG communication link resource allocation algorithm was introduced. This algorithm considers subcarrier and power allocations. It maximises the system capacity, while accounting for user fairness. However, the algorithm is complicated and has a low convergence speed. In [9], the local optimal solution is obtained by joint optimisation of uplink cell association and power allocation, which effectively improves the network performance. In [10], aiming at the problem of UAV access and base station bandwidth allocation, a hierarchical game power control algorithm is proposed. This algorithm can effectively solve the problem of UAV access and base station bandwidth allocation. Nash equilibrium is proved by theory, but the problem is modeled as a noncooperative game, sacrificing the overall performance of the system. In [11], a power control algorithm based on mean field game and deep reinforcement learning is proposed. The algorithm first transforms the power control problem into discrete mean constant game and then uses neural network to solve the optimal transmit power. Although this algorithm can effectively improve the energy efficiency of UAV, its complexity is much higher than those of other algorithms. In [12], a noncooperative game distributed power control algorithm was introduced, which could effectively improve the system throughput and reduce cochannel interference; however, the proposed algorithm cannot achieve optimal system capacity. In [13], a power control algorithm based on noncooperative game is proposed. The algorithm adopts a nonuniform pricing method and sets different prices for different base stations. The scheme can effectively improve the system performance in UAVG network structure. Although the above studies reduce interference to a certain extent, they are not user-centric and cannot achieve optimal system throughput.

To solve the above problems, a cooperative game power control algorithm in UUAVG is proposed by this study. First, the system model of the UUAVG downlink power control is established. Considering the interference to user QoS and service UAV to edge users, a new cooperative game utility function is proposed, and the optimal power control scheme is solved using the Lagrange function. The simulation results show that the cooperative game power control (CGPC) algorithm can reduce the user transmitted power and interuser interference and improve system throughput.

2. System Model

The model of UUAVG downlink power control system is shown in Figure 1. Assuming that there is a service user (SU) in the model, the dynamic UAVG is formed with the SU as the centre, and the coverage range of the UAVG available to the SU is represented by the dotted circle. It is assumed that there is only one service unmanned aerial vehicle (SUAV) in the UAVG. Suppose that there are n cooperative unmanned aerial vehicles (CUAV), denoted as CUAV $_i$ ($i = 1, 2, \dots, i, \dots, j, \dots, N$); in the coverage of the UAVG, assume that there are n edge users (EU) at the same time, denoted as EU $_i$

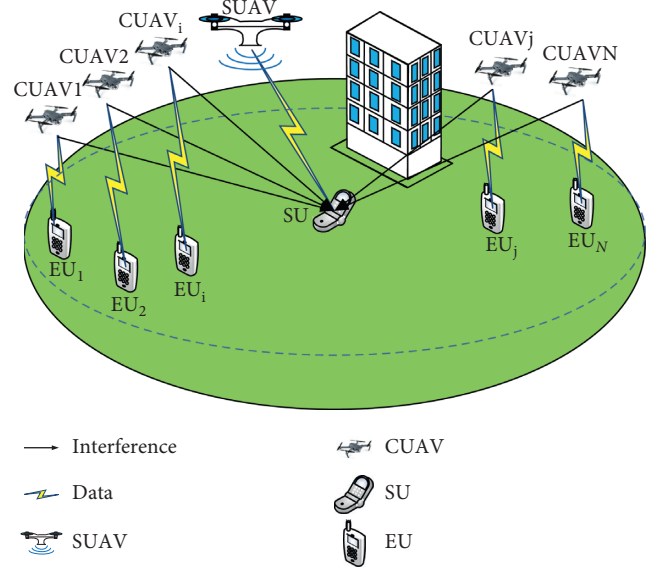


FIGURE 1: Downlink power control system model of UUAVG.

($i = 1, 2, \dots, i, \dots, j, \dots, N$), and that they are far away from the SUAV, with the service base station of EU $_i$ represented as CUAV $_i$. Let the vertical height of the SUAV and CUAV from the ground be H_0 . Owing to the scarcity of spectrum availability and to improve frequency utilisation, when deploying the UAVG, the downlink communication frequencies of the SUAV and SU are the same as those of the CUAV $_i$ and EU $_i$. Therefore, the interference between the SUAV and CUAV and those between the CUAVs are inevitable.

For each user, they are in the center of the network, and UAVs around the user are dynamically composed of UAVG. Each user has a unique UAVG to provide services, but each UAV may belong to a different UAVG at the same time. Therefore, each UAV may be both a user's SUAV and another user's CUAV at the same time. For users, if a UAVG is the user's own UAVG, it is called a service user of this UAVG; otherwise, it is called an edge user of this UAVG.

Let p_t and p_i denote the transmitted power levels of SUAV and CUAV $_i$, and let γ_i denote the signal to interference noise ratio (SINR) of the corresponding EU $_i$.

$$\gamma_i(p_i) = \frac{p_i h_{ii}}{\sum_{j=1, j \neq i}^N p_j h_{ji} + p_t g_{si} + \sigma^2}. \quad (1)$$

The downlink channel gain between CUAV $_i$ and EU $_i$ is denoted as h_{ii} , the channel gain between CUAV and EU $_i$ is denoted as h_{ji} , and the channel gain between SUAV and EU $_i$ is denoted as g_{si} . The power of the additive white Gaussian noise on EU $_i$ is σ^2 . The distance between CUAV $_i$ and EU $_i$ is denoted as d_{ii} and that between SUAV and EU $_i$ is denoted as r_{si} .

The horizontal distance between CUAV $_i$ and EU $_i$ is denoted as D_{ii} and that between SUAV and EU $_i$ is denoted as R_{si} . Using the triangle Pythagorean theorem, we can get d_{ii} and r_{si} .

$$\begin{aligned} d_{ii} &= \sqrt{H_0^2 + D_{ii}^2}, \\ r_{si} &= \sqrt{H_0^2 + R_{si}^2}. \end{aligned} \quad (2)$$

3. CGPC Algorithm

3.1. CGPC Game Elements. In the noncooperative game, each participant will choose selfishness to maximise their benefits, thereby causing interference to the other participants and reducing the benefits to other users. Meanwhile, a cooperative game is meant to maximise the interests of the collective and is a competition model in which the participants seek to maximise their interests. Therefore, a cooperative game will bring more benefits to the system.

The three basic elements of the cooperative game are the participants, strategy set, and utility function. Participants in the game are represented as I : the set of CUAVs participating in the game, where $I = \{1, 2, \dots, N\}$. The strategy set is represented as Ω_i : decision of each participant can be expressed as $\{p_1, p_2, \dots, p_i, \dots, p_n\}$, $0 \leq p_i \leq p_{\text{CUAV}}^{\max}$ (here, p_{CUAV}^{\max} represents the maximum allowable transmitted power of the CUAV), and each decision is independent of the others. The utility function is given as U_i , which represents the preference of EUi for a certain strategy.

3.2. Utility Function. In the traditional cooperative game algorithm, the utility function only considers the user SINR. In this study, we consider not only the SINR of users but also the influence of SUAV on EUi. A new utility function is divided into two types according to whether the SINR of users is greater than the threshold SINR:

$$U_i(p_i, \gamma_i) = \begin{cases} \prod_{i=1}^m g_{si} \frac{\gamma_i - \gamma_{\text{EU}}^{\min}}{\gamma_i}, & \gamma_i \geq \gamma_{\text{EU}}^{\min}, \\ 0, & \gamma_i < \gamma_{\text{EU}}^{\min}. \end{cases} \quad (3)$$

In the first case, when user $\gamma_i \geq \gamma_{\text{EU}}^{\min}$, $U_i(p_i, \gamma_i) = \prod_{i=1}^m g_{si} (\gamma_i - \gamma_{\text{EU}}^{\min} / \gamma_i)$, where m is defined as the number of $\gamma_i \geq \gamma_{\text{EU}}^{\min}$ in all UAVs, and $\gamma_{\text{EU}}^{\min}$ is defined as the minimum SINR required for EU communication. $\gamma_i - \gamma_{\text{EU}}^{\min}$ guaranteed communication QoS of EU. Only when $\gamma_i \geq \gamma_{\text{EU}}^{\min}$ can the utility function be guaranteed to be positive. $(\gamma_i - \gamma_{\text{EU}}^{\min} / \gamma_i)$ can make the power change of user EUi smoother and reduce the times of games. It can be observed from (3) that the utility function EUi is related to g_{si} and $\gamma_{\text{EU}}^{\min}$. The utility function considers the SINR of EUi and the channel gain of the SUAV and EUi. The physical meaning of utility function is to ensure the maximisation of $(\gamma_i - \gamma_{\text{EU}}^{\min} / \gamma_i)$ of all EUi.

In the second case, when user $\gamma_i < \gamma_{\text{EU}}^{\min}$, γ_i cannot be guaranteed the normal communication requirements of EUi, so the utility function of EUi is defined as 0.

The objective function of power control based on the cooperative game in CGPC can be expressed as follows:

$$\begin{cases} \max U(p_i, \gamma_i) = g_{si} \prod_{i=1}^m \frac{\gamma_i - \gamma_{\text{EU}}^{\min}}{\gamma_i}, \\ \text{s.t.} \begin{cases} 0 \leq p_i \leq p_{\text{CUAV}}^{\max}, \\ \sum_{j=1, j \neq i}^m h_{ji} p_i \leq T, \\ \gamma_i \geq \gamma_{\text{EU}}^{\min}. \end{cases} \end{cases} \quad (4)$$

Here, T is the threshold value for which EUi can withstand interference from the UAVG. In theory, we set $\gamma_{\text{EU}}^{\min}$ and determine the interference threshold T according to the channel gain and transmit power from EUi to CUAV. $(p_i h_{ii} / T) \geq \gamma_{\text{EU}}^{\min}$, if and only if $\gamma_i = \gamma_{\text{EU}}^{\min}$; we can calculate the threshold $T = (p_i h_{ii} / \gamma_{\text{EU}}^{\min})$. The physical meaning of power control in CGPC indicates that the utility function of the EU reaches the maximum when the interference from CUAV to SU does not exceed the maximum interference threshold that the SU can bear.

3.3. Nash Solution. Equation (4) can be further modified by considering the logarithm of the objective function of (4) and using the properties of logarithmic functions, $\ln \prod_{i=1}^N \gamma_i = \sum_{i=1}^N \ln \gamma_i$, as follows:

$$\begin{cases} \max U(p_i, \gamma_i) = g_{si} \prod_{i=1}^m \ln \left(\frac{\gamma_i - \gamma_{\text{EU}}^{\min}}{\gamma_i} \right), \\ \text{s.t.} \begin{cases} 0 \leq p_i \leq p_{\text{CUAV}}^{\max}, \\ \sum_{j=1, j \neq i}^m h_{ji} p_i \leq T, \\ \gamma_i \geq \gamma_{\text{EU}}^{\min}. \end{cases} \end{cases} \quad (5)$$

Therefore, the problem of $U_i(p_i, \gamma_i)$ maximisation is now transformed into the problem of $u_i(p_i, \gamma_i)$ maximisation, and $u_i = \ln U_i$.

Theorem 1. Model equivalence theorem. Suppose that the decision set for the utility function set has a one-to-one mapping relationship that satisfies the concave function characteristics; that is, $u_i = \ln U_i$ satisfies the characteristics of the concave function. Thus, it can be concluded that the models in (4) and (5) are equivalent, and the same solution is obtained.

Equation (5) is an optimisation problem with multiple constraints, and the Lagrange factor $\lambda, \mu_i, \eta_i, \varepsilon_i$ is introduced to construct a Lagrange function. Among them, Lagrange factor is the limiting factor set when solving Lagrange function. Through Lagrange factor, the objective function

and constraint conditions are connected together to construct a new Lagrange function.

Therefore, Lagrange function can be expressed as follows:

$$L(p_i, \gamma_i) = g_{si} \sum_{i=1}^m \ln \left(\frac{\gamma_i - \gamma_{EU}^{\min}}{\gamma_i} \right) - \lambda \left(\sum_{j=1, j \neq i}^m h_{ji} p_i - T \right) - \mu_i (\gamma_i - \gamma_{EU}^{\min}) - \eta_i (p_{CUAV}^{\max} - p_i) - \varepsilon_i (0 - p_i). \quad (6)$$

According to (6), the partial derivative of the Lagrange function p_i is calculated, and the optimal solution is

$$p_i^{(k+1)} = \frac{\gamma_{EU}^{\min}}{\gamma_i} p_i^{(k)} + \frac{g_{si}}{\mu_i (\gamma_i / p_i^{(k)}) - \lambda (\sum_{j=1, j \neq i}^m h_{ji} p_i - T) - (\eta_i + \varepsilon_i)}. \quad (8)$$

According to (8), the first term $(\gamma_{EU}^{\min} / \gamma_i) p_i^{(k)}$ on the right is consistent with the traditional SINR balancing algorithm [12], and it is expressed that $p_i^{(k+1)}$ will gradually stabilise when γ_i approaches γ_{EU}^{\min} . The second term $(g_{si} / \mu_i (\gamma_i / p_i^{(k)}) - \lambda (\sum_{j=1, j \neq i}^m h_{ji} p_i - T) - (\eta_i + \varepsilon_i))$ is a fine-tuning term comprising the Lagrangian factors, channel gain, and current SINR. This proves the correctness of the algorithm. It not only ensures that it is similar to the traditional SINR balancing algorithm but also makes fine-tuning on the basis of the traditional SINR balancing algorithm. The algorithm in this paper is based on UAV transmit power polling, not on time.

In this paper, the throughput of the system is characterised by the sum of the average reachable rates of all users, where the average reachable rate of each user can be expressed as

$$R_i = B \log_2 (1 + \gamma_i), \quad (9)$$

where R_i is the user reachable rate and the bandwidth is $B = 1$ Hz. Therefore, the system throughput of each iteration can be expressed as follows:

$$S_t^{(k)} = \sum_{i=1}^N R_i^{(k)}, \quad (10)$$

where $S_t^{(k)}$ is the throughput of the k -th iteration. $R_i^{(k)}$ is the user reachable rate of user i at the k -th iteration.

4. Performance Evaluation

In this section, we present verification of the feasibility of the CGPC algorithm through simulations and calibrations for the theoretical analysis results and simulation results.

4.1. Simulation Configuration and Parameters. According to [11–13], the height of the UAVG is selected as $H_0 = 80$ m, and the horizontal projection coverage area of the UAVG is selected in a circular area of radius 300 m, where the horizontal distance from the SU to SUAV is set as $[0, 100]$ m,

obtained by setting $(\partial L(p_i, \gamma_i) / \partial p_i) = 0$; so we can get p_i as follows:

$$p_i = \frac{\gamma_{EU}^{\min}}{\gamma_i} p_i + \frac{g_{si}}{\mu_i (\partial \gamma_i / \partial p_i) - \lambda (\sum_{j=1, j \neq i}^m h_{ji} p_i - T) - (\eta_i + \varepsilon_i)}. \quad (7)$$

According to the SINR definition, $(\partial \gamma_i / \partial p_i) = (\gamma_i / p_i)$ holds. The fixed-point iterative method is used for iteration; after k iterations, the transmitted power of CUAV i is as follows:

and the horizontal distance from EU to SUAV is set as $(100, 300]$. Suppose that there is one SUAV and four CUAVs, namely, CUAV1, CUAV2, CUAV3, and CUAV4, in this circular area. In the circular region, there is one SU and four EU, namely, EU1, EU2, EU3, and EU4. The distance relationship between the EUs and SUAV is expressed as $R_{s1} \leq R_{s2} \leq R_{s3} \leq R_{s4}$.

Suppose that the transmitted power of the SUAV is 30 dBm, and the initial transmitted power of the CUAV is 20 dBm. Assuming noise threshold $T = 1 \times 10^{-10}$ W, the Lagrange factors are obtained as $\lambda = 1 \times 10^{16}$ and $\mu_i = \eta_i = \varepsilon_i = 10$ [11–13].

4.2. Simulation Results and Analysis. Figure 2 illustrates the curves for the CUAV transmitted power changing with the number of iterations. The abscissa in the figure is the number of iterations, and the ordinate is the transmitted power from the CUAV. It can be observed from the figure that the initial transmitting power of CUAVs is 0.1 W. With the increase of iteration times, the transmitting power of CUAV decreases gradually and reaches a stable state after many games. The attenuation rate of CUAV1 is the fastest and that of CUAV4 is the slowest. This is due to the different distances between EU and SU. The distance between EU1 and SU is relatively close, and the downlink communication power of CUAV1 and EU1 has a greater impact on SU. The power attenuation of CUAV1 is relatively large. The distance between EU4 and SU is relatively long, and the influence of CUAV 4 and EU4 downlink communication power on SU is relatively small. Therefore, the power attenuation of CUAV 4 is relatively small. This is consistent with the result of formula (3).

Figure 3 presents a comparison of the CUAV transmitted power for different algorithms. The abscissa is the distance between the EU and SUAV, and the ordinate is the transmitted power from the CUAV. The curves in the figure represent the results of the K-G algorithm [12] and CGPC algorithm. It can be observed that the transmitted power of the CUAV increases with the distance between the EU and SUAV. When the distance between EU and CUAV is close,

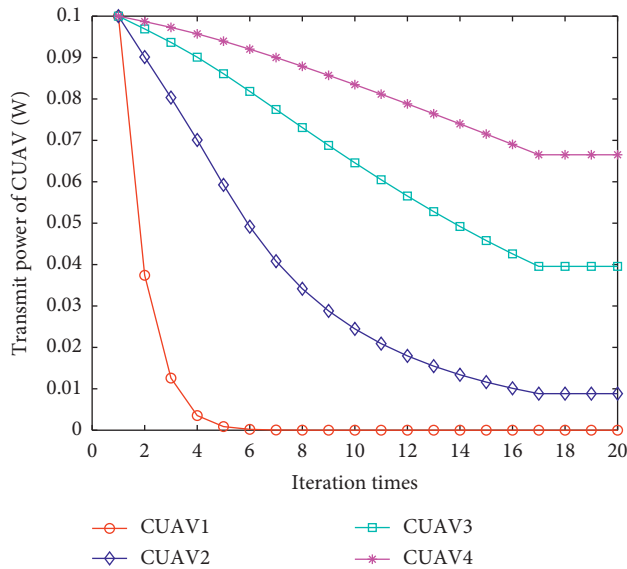


FIGURE 2: Transmitted power from CUAUV for number of iterations.

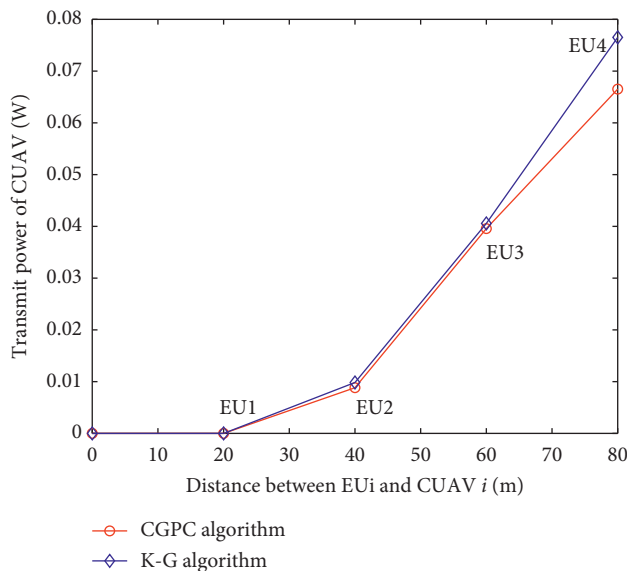


FIGURE 3: Comparison of cooperative UAV transmitted power under different algorithms.

the effect of the proposed algorithm is equivalent to that of the traditional K-G algorithm. However, after 60 meters, it obviously reflects the advantages of this algorithm, which can effectively reduce the transmission power of CUAUV. This is because the design of utility function considers not only the user's QoS but also the influence of different distance EU on CUAUV. The farther the distance between EU and CUAUV, the more obvious the advantages of this algorithm. Comparing these two algorithms under the same conditions, the transmitted power of the CUAUV based on the CGPC algorithm is observed to be lower than that of the K-G algorithm, which reduces the cofrequency interference.

Figure 4 demonstrates the curves for the system average throughput according to the number of CUAUVs for different

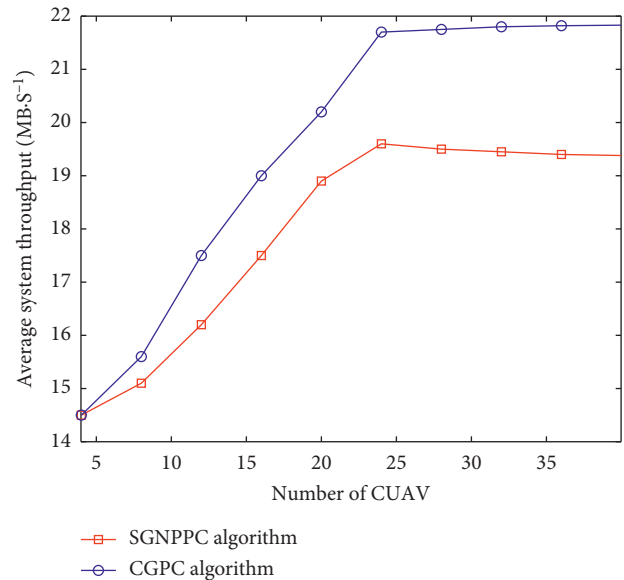


FIGURE 4: Transmitted power for the system average throughput CUAUV for different algorithms.

algorithms; the abscissa is the number of CUAUVs, and the ordinate is the average throughput of the system. The curves represent results for the Stackelberg Game based Nonunified Pricing Power Control (SGNPPC) algorithm [13] and CGPC algorithm for the average system throughput. It can be observed that the average throughput of the two systems increases as the number of CUAUVs increases. When the number of CUAUVs reaches 24, the average throughput of the system is close to saturation. Thereafter, the system throughput of the SGNPPC algorithm gradually decreases with increase in the number of CUAUVs; this is because more interference is inserted with the increase in the number of CUAUVs; conversely, the throughput of the CGPC algorithm increases with the increase in the number of CUAUVs, but the rate of increase is reduced. This deceleration can be attributed to the adoption of cooperative games to reduce the interference between the CUAUVs. Compared with the SGNPPC algorithm, the average throughput of the CGPC algorithm increased by 10.32%.

5. Conclusion

In this study, a CGPC algorithm is proposed in the scenario of the UUAUVG. The algorithm considers not only the SINR of the user but also the influence of the different distances between the EU and SUAV on the EU. The complex Nash theorem utility function is transformed into an easily solved optimisation problem using the principle of model equivalence; further, the Lagrange function is constructed, and the optimal transmitted power is obtained iteratively. The convergence of the algorithm is proven by simulation, and the average throughput of the system is improved. In the future work, we will consider the joint optimisation of distributed power control and user rate control, as well as the trajectory design and layout design of UAV.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the subproject of National Key Research and Development Plan in 2020 (no. 2020YFC1511704), the National Natural Science Foundation of China (Grant no. 61971048), Beijing Science and Technology Project (Grant no. Z191100001419012), and scientific research level improvement project to promote the colleges connotation development of Beijing Information Science & Technology University in 2020 (no. 2020KYNH212).

References

- [1] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, "UAV-enabled intelligent transportation systems for the smart city: applications and challenges," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 22–28, 2017.
- [2] W. Huang, J. Peng, and H. Zhang, "User-centric intelligent UAV swarm networks: performance analysis and design insight," *IEEE Access*, vol. 7, pp. 181469–181478, 2019.
- [3] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78–85, 2016.
- [4] R. Tang, J. Zhao, H. Qu et al., "User-centric joint admission control and resource allocation for 5G D2D extreme mobile broadband: a sequential convex programming approach," *IEEE Communications Letters*, vol. 7, 2017.
- [5] P. Zhang and Y. Zhang, "Research on power control algorithm of two-layer Stackelberg game in UUDN," *Computer Engineering*, vol. 46, no. 9, pp. 186–192, 2020, in Chinese.
- [6] H. Zhang and W. Huang, "Tractable mobility model for multi-connectivity in 5G user-centric ultra-dense networks," *IEEE Access*, vol. 6, pp. 43100–43112, 2018.
- [7] S. Bhattacharya and T. Basart, "Game-theoretic analysis of an aerial jamming attack on a UAV communication network," in *Proceeding of the American Control Conference*, pp. 818–823, IEEE, Baltimore, MD, USA, July 2010.
- [8] Q. Li, "Non-cooperative game power control in swarm UAV networks," *Telecommunication Engineering*, vol. 7, 2019, in Chinese.
- [9] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: uplink association, power control and interference coordination," in *Proceedings of the IEEE Global Communications Conference*, Abu Dhabi, UAE, December 2018.
- [10] Y. Shi, M. Peng, and X. Cao, "A game theory approach for joint access selection and resource allocation in UAV assisted IoT communication networks," *IEEE Internet of Things Journal*, vol. 99, 2018.
- [11] L. Li, Q. Cheng, K. Xue, C. Yang, and Z. Han, "Downlink transmit power control in ultra-dense UAV network based on mean field game and deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15594–15605, 2020.
- [12] Z. R. Gajic and S. Koskie, "Newton iteration acceleration of the Nash game algorithm for power control in 3G wireless CDMA networks," in *Proceedings, ITCOM 2003, Conference on Performance and Control of Next Generation of Communication Networks*, no. 5, pp. 115–121, Orlando, FL, USA, August 2003.
- [13] C. M. XU and J. WU, "Non- unified pricing power control based on the Stackelberg game in the ultra-dense network," *Journal of Computer Applications*, vol. 38, no. 08, pp. 2323–2329, 2018, in Chinese.