WILEY | Hindawi

Research Article

Exploiting Temporal and Spatial Regularities for Content Dissemination in Opportunistic Social Network

Zhiyuan Li [ID],[1,2] Junlei Bi,[1] and Carlos Borrego[3]

[1]*School of Computer Science and Communication Engineering, Jiangsu University, China*
[2]*Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Jiangsu, China*
[3]*Department of Information and Communications Engineering, Autonomous University of Barcelona, Spain*

Correspondence should be addressed to Zhiyuan Li; lizhiyuan81@126.com

Recently, content dissemination has become more and more important for opportunistic social networks. The challenges of opportunistic content dissemination result from random movement of nodes and uncertain positions of a destination, which seriously affect the efficiency of content dissemination. In this paper, we firstly construct time-varying interest communities based on the temporal and spatial regularities of users. Next, we design a content dissemination algorithm on the basis of time-varying interest communities. Our proposed content dissemination algorithm can run in $O(n \log n)$ time. Finally, the comparisons between the proposed content dissemination algorithm and state-of-the-art content dissemination algorithms show that our proposed content dissemination algorithm can (a) keep high query success rate, (b) reduce the average query latency, (c) reduce the hop count of a query, and (d) maintain low system overhead.

1. Introduction

Opportunistic social networking (OSN) emerges as a new paradigm of mobile social networking [1, 2]. Different from online social networking, the mobile users have to use wireless short distance communication technologies to opportunistically communicate with others in OSN. There are new challenges in OSN, and one of the most challenging ones is content dissemination [3]. The goal of content dissemination in OSN is to successfully propagate a data message to a destination node. Nowadays, the existing content dissemination has been classified into three categories, such as the epidemic content dissemination [4, 5], the probability-based content dissemination [6–17], and the interest community-based content dissemination [18–23].

Epidemic content dissemination originates from flooding data messages in intermittently connected mobile networks [4]. For the epidemic content dissemination, each node maintains a list of messages it carries. Whenever it encounters another node, the two nodes exchange the messages that they do not have with each other. If the connection lasts long enough, the data messages may traverse all nodes in OSN. Although the epidemic content dissemination is an optimal solving method under no contention, it is wasteful for network resources due to amounts of data message copies. In the probability-based content dissemination [6, 7], every node records historical contact information with its neighbors. And then, the nodes can create frequent contact lists according to the historical contact information. The data message is forwarded to a close node who can possibly deliver this query to resource holders. The probability-based content dissemination is a locally optimal method, and hence the methods make the data messages stay in a small search scope. For the interest community-based content dissemination [19], each device is associated with an interest vector extracted from user profiles, and the devices with same interests form an interest community. Each data message is also described by an interest vector. The data messages are disseminated within one interest community or multiple interest communities. However, the drawbacks of content dissemination based on interest community lie in two aspects. One aspect is topology mismatched due to not considering the physical connections of nodes. The other aspect is inaccurate interest representation since nodes provide little privacy information.

The two aspects lead to the performance degradation of content dissemination based on interest community.

To cope with the aforementioned limitations of content dissemination, we propose a content dissemination based on time-varying interest community in OSN. The main contributions of this study are shown as follows.

(a) We use temporal and spatial correlations to construct a time-varying interest community model.

(b) Based on the time-varying interest community model, we propose a content dissemination which runs in $O(n \log n)$ time.

(c) We evaluate the performances of proposed content dissemination in terms of query success rate, average query latency, hop count of a query, and system overhead.

The remainder of this paper is organized as follows: Section 2 describes the related work and motivation. In. Section 3 presents a content dissemination algorithm based on time-varying interest communities. In Section 4, the performances of our proposed method are evaluated and analyzed. Finally, Section 5 concludes our study and gives the future directions of this research.

2. Related Work and Motivation

In this section, the advantages and drawbacks of existing content dissemination methods are analyzed by classifying them into three categories: the epidemic content dissemination, the probability-based content dissemination, and the interest community-based content dissemination.

2.1. Epidemic Content Dissemination. For uncertain positions of destinations in OSN, Vahdat et al. [4] proposed an epidemic-based message dissemination for partially connected ad hoc networks. The epidemic-based content dissemination is for sending messages to destinations through those moving nodes. When the moving nodes meet, they exchange their summary vectors to determine which messages remotely stored have not been seen by the local node. In turn, each node then requests copies of messages that it has not yet seen. In this way, the data messages can arrive at every node in the network. The epidemic content dissemination outperforms two other content dissemination methods in terms of successful rate of message delivery. However, the epidemic content dissemination can generate multicopies of a data message during a task execution, and hence the method may incur high system overhead, such as network bandwidth and battery depletion. To control the system overhead, Spyropoulos et al. [5] presented the spray and wait content dissemination including basic spray and wait and binary spray and wait. In the spray and wait content dissemination, the number of copies of a message is set to reduce the system overhead. However, how to automatically adjust the parameter L to achieve the tradeoff between the successful rate of message delivery and the system overhead is still a puzzle.

2.2. Probability-Based Content Dissemination. In OSN, probability-based content dissemination methods [3, 6, 7, 16]

have been proposed to improve the efficiency of content dissemination. Saha and Misra define a function for probabilistic replication of any message m in [3]. When the function value is greater than or equal to a random number $r \in [0, 1]$, the message m is replicated and forwarded. Daly and Haahr [6] proposed a content dissemination protocol based on social network analysis, called SimBet. The SimBet content dissemination uses betweenness centrality and similarity metrics to identify some "bridge" nodes (with high values on these metrics) in networks. To control the system overhead, node n calculates the betweenness centrality Bet_n and similarity Sim_n with its local neighborhood. Then, the node n can compute its $SimBet$ value, which is a weighted combination of the betweenness utility and the similarity utility. If $SimBetUtil_m(d)$ is greater than $SimBetUtil_n(d)$, node n forwards the data message to node m. Otherwise, node n continues to hold the message. But, it is still possible that the node with high $SimBet$ utility fails to deliver the message to the destination due to the uncertainty of future encounters and underlying social graph. To construct an opportunistic ordered service chaining in [7], Saha and Misra also compute the delivery probability values by PRoPHET. However, unlike PRoPHET, they select a node who has the highest delivery probability as the intermediate node and uses the unicast way to deliver the data message to the best service providers. Reference [16] studied the message replication for content dissemination in opportunistic mobile networks. In this study, authors in [16] considered multiple attributes for message replication decision, such as interest hits, the number of entries in forwarding information base, and the average time interval between two successive contacts. And then, they designed an interests replication algorithm from node i to node j according to the attributes mentioned above. The interests replication algorithm can help the node i to make a decision about whether sending the interest message to other nodes or not.

2.3. Interest Community-Based Content Dissemination. Interest community is an overlay network to enhance the scalability of content dissemination [17]. In ecology, a community is an assemblage of two or more populations of different species occupying the same geographical area. In sociology, a community is usually defined as a group of interacting people living in a common location. Community ecologists and sociologists study the interactions between species and people in communities at many temporal and spatial scales [18–22]. Once interest communities are constructed, the content dissemination is efficiently implemented on top of the overlay network. Li and Wu [20] proposed a mobile community-based pub/submethod to promote the efficiency of message delivery. In [20], communities are constructed using encounter frequencies of nodes. However, people with long-term social ties do not necessarily have similar interests. For example, the graduate students and their advisors usually work and talk in the same office from 9 a.m. to 5 p.m., but their interests may be different. Chen et al. [18] leverage the interest similarity to study content dissemination in disconnected wireless networks, called SPOON. In SPOON, an
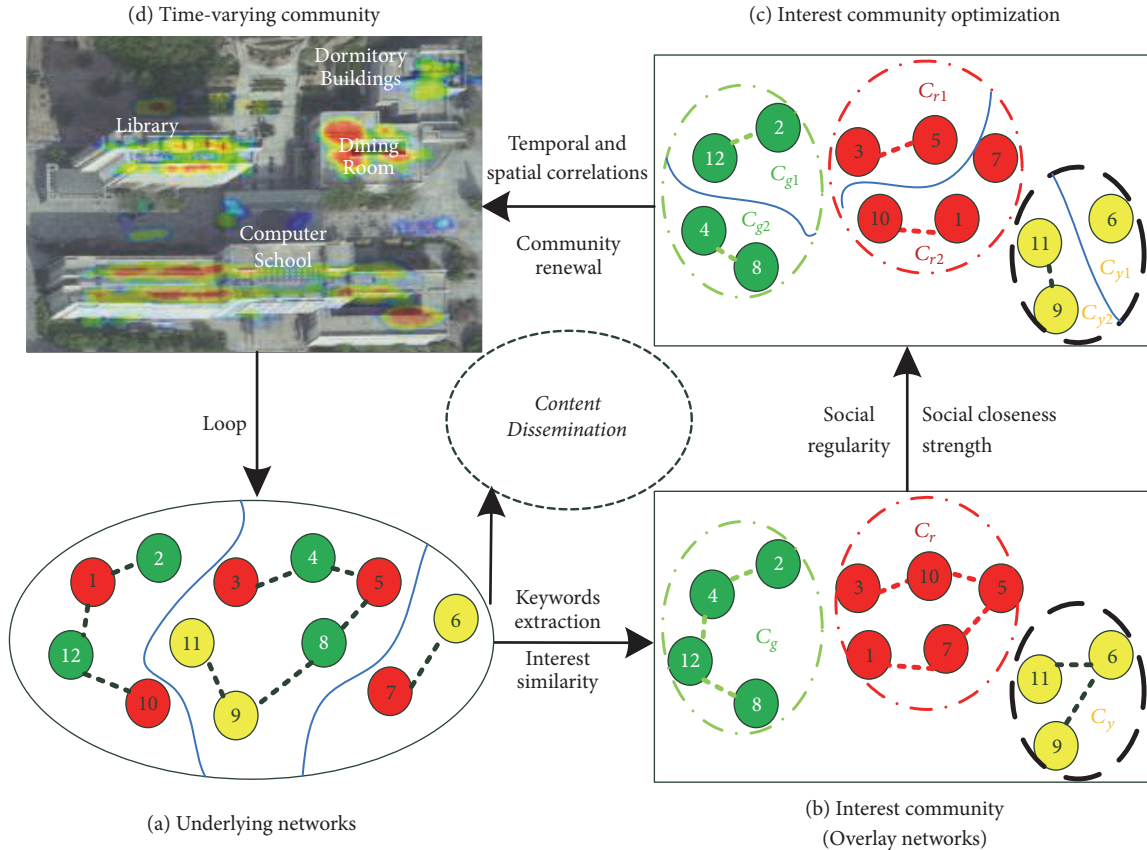
Figure 1: Relationship between components in system model.

interest extraction is used to collect a node's interests from its file resources. The nodes with common interests are classified into the same interest community. Then, the epidemic content dissemination is used within a community. Through the social network analysis, the nodes frequently visiting other communities are assigned as the community ambassadors for intercommunity content dissemination. However, topology mismatch affects the performances of SPOON. The topology mismatch is defined as the difference between the interest community and the underlying networks. The interest community only shows the logical connections among nodes. For example, node i is a neighbor of node j in the interest community. However, node i is far from node j over the physical networks. Hence, the topology mismatch seriously affects the efficiency of content dissemination in OSN.

To cope with the aforementioned limitations of content dissemination, we have extracted interests and preferences of users and constructed time-varying interest communities by the factor of temporal and spatial correlation. The detailed descriptions are shown in Section 3.

3. System Model and Content Dissemination Implementation

In this section, we firstly describe the overview of system model. Next, both the formation of interest communities and the structure optimization of interest communities are given. Then, the time-varying community model based on temporal and spatial correlations is presented. Based on the time-varying communities, we propose a content dissemination algorithm.

3.1. Overview. Figure 1 illustrates the principle of system model. In the figure, different colored circles represent nodes with different interests.

Figure 1(a) shows the underlying networks. As shown in the subfigure, different interest nodes are connected through wireless access points. Interest communities are constructed as overlay networks on top of underlying networks, as shown in Figure 1(b). The nodes in interest communities have similar interests and show the consistent behavior patterns. To overcome the deficiencies of constructed interest communities, we study the social relations of users and present the definition of social closeness strength. Then, the metric of social closeness strength is used to optimize the topological structure of interest communities. Figure 1(c) shows that interest communities are divided into closer subcommunities by the metric of social closeness strength. After that, time-varying community model based on temporal and spatial correlations is proposed to reconstruct interest communities. The reconstruction process for interest communities can be achieved according to steps Figures 1(a)–1(c). Finally, we

TABLE 1: Summary of presented symbols.

Name	Description
k_h	The h-th keyword
f_h	The frequency of the h-th keyword
$f_{k(\alpha)}(n)$	The frequency of keyword $k(\alpha)$ for node n
$w_n^{k(\alpha)}$	The weight of keyword $k(\alpha)$ for node n
T_{sim}	Preset threshold for interest community construction
T_{ij}	Average inter-contact time between nodes i and j
λ	Social closeness strength
$f(t)$	Time remaining to the next contact between nodes after time t

design and implement a content dissemination based on the interest communities.

In order to enhance the readability of this paper, the summary of presented symbols is shown in Table 1.

3.2. Interest Community Formation. Users with common interests tend to join the same community, and their shared resources are similar. The resources owned by users are used to show their interests, such as frequently used documents, photos, music, and videos, but not limited to these items. Hence, we can extract interests of nodes from profiles of users or metadata of shared resources. Next, we describe the steps of interest community formation.

Step 1 (interest representation of a node). We firstly use spaCy package to scan the metadata of multimedia contents and the titles of documents in a tablet, e.g., surface pro 4. After some natural language processing steps including removing stop words (the most common words), such as *the, is, at, which,* and *on,* stemming and lemmatization, the keywords of a node are extracted. The combination of multiple keywords can show interests of a node.

Interests of a node i are represented using an interest vector $s(i)$, as shown in

$$s(i) = \{(k_1, f_1), (k_2, f_2), \cdots (k_n, f_n)\} \tag{1}$$

where the symbol k_h refers to the h-th keyword, the symbol f_h is the frequency of the h-th keyword, the key-value pairs (k_h, f_h), $h \in [1, n]$ are sorted by the frequency f_h in descending order, and the subscript n is self-defined.

The weight of a keyword $k(\tau)$ for an interest vector $s(i)$ is denoted as

$$w_i^{k(\tau)} = \frac{f_{k(\tau)}(i)}{\sum_{\beta=1}^n f_\beta(i)} \tag{2}$$

where the symbol $w_i^{k(\tau)}$ refers to the weight of a keyword $k(\tau)$ for an interest vector $s(i)$ of node i.

Step 2 (interest similarity between nodes). Formula (3) is used to calculate the interest similarity $interest\_sim(i, j)$ between nodes i and j.

$$
\begin{aligned}
interest\_sim(i, j) &= \frac{\sum_{\alpha=1}^m w_i^{k(\alpha)} \times w_j^{k(\alpha)}}{\sqrt{\sum_{\alpha=1}^m \left(w_i^{k(\alpha)}\right)^2} \times \sqrt{\sum_{\alpha=1}^m \left(w_j^{k(\alpha)}\right)^2}} \\
&= \frac{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(i) / \sum_{\beta=1}^n f_\beta(i) \times f_{k(\alpha)}(j) / \sum_{\beta=1}^n f_\beta(j)\right)}{\sqrt{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(i) / \sum_{\beta=1}^n f_\beta(i)\right)^2} \times \sqrt{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(j) / \sum_{\beta=1}^n f_\beta(j)\right)^2}} \\
&= \frac{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(i) \times f_{k(\alpha)}(j) / \sum_{\beta=1}^n f_\beta(i) \times \sum_{\beta=1}^n f_\beta(j)\right)}{\sqrt{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(i) / \sum_{\beta=1}^n f_\beta(i)\right)^2} \times \sqrt{\sum_{\alpha=1}^m \left(f_{k(\alpha)}(j) / \sum_{\beta=1}^n f_\beta(j)\right)^2}}
\end{aligned} \tag{3}
$$

where m is the number of same keywords between nodes i and j, $k(\alpha), \alpha \in [1, m]$ denotes the same keyword of nodes i and j, and $w_i^{k(\alpha)}$ and $w_j^{k(\alpha)}$ represent the weights of the keyword $k(\alpha)$ for nodes i and j, respectively. $f_{k(\alpha)}(i)$ and $f_{k(\alpha)}(j)$ are the frequencies of the keyword $k(\alpha)$ for nodes i and j, respectively.

Each node can calculate interest similarity with others using formula (3). Then, we can construct different interest communities for nodes in OSN.

Step 3 (interest community construction). The nodes can get their interest vectors using formula (1). When the nodes in network are connected, node i will broadcast its interest vector to its neighbors. After that the neighboring nodes will calculate the interest similarity with node i using formula (3). If their interest similarities are greater than or equal to the preset threshold T_{sim}, the nodes will join the same community. Otherwise, a new community is created. When the nodes in network are not connected, it is necessary to wait for the next contact between them. The procedure above is repeated until the constructed interest communities may cover all nodes in OSN. It is noted that a node only belongs to an interest community. In other words, if a node has already joined in an interest community, the node will not join another interest community. The algorithm of interest community construction is shown in Algorithm 1.

```
Input: Interest vector of node n_x, x ∈ [1, N] and the total number of nodes N
Output: Interest communities C_i, i ∈ [0, N]
(1) Initiation: Create the first interest community C_i = ∅, i = 0
(2) Begin
(3) C_i = {n_0}
(4) do
(5) {
(6)      if connections among nodes exist
(7)      {
(8)          for each neighbor n_e of node n_x do
(9)          {
(10)              n_x broadcastTo its neighborn_e
(11)              n_e calculate interest_sim(x, e)
(12)              if interest_sim(x, e) ≥ T_sim
(13)              then     {
(14)                          C_i = C_i ∪ n_e
(15)                      }
(16)              else {
(17)                      creatNewCommunity C_{i++}
(18)                      }
(19)          }
(20) }
(21)      else
(22)      {
(23)          Wait the next contact
(24)          Goto (6):
(25)      }
(26) }
(27) While (Union C of interest community set C_i == N)
(28) Return C
(29) End
```

ALGORITHM 1: Interest community construction.

Next, we give an example to explain the algorithm. Suppose that node n_0 firstly joins the first created interest community C_0. After that, when node n_0 meets with node n_2 and their interest similarity is greater than or equal to T_{sim}, node n_2 sends a request to node n_0 for joining the interest community C_0. Otherwise, node n_2 can create a new interest community, called C_1. If both node n_5 and nodes n_0 and n_2 have higher interest similarity than the preset threshold T_{sim}, node n_5 and its first contact node are in the same interest community. The running of this algorithm will terminate until all nodes have joined an interest community.

3.3. Interest Community Optimization Based on Social Closeness Strength. In this section, we firstly give the metric definition of social closeness strength. Then, we use the social closeness strength to represent social relationship between nodes. The social relationship can help us to optimize the topological structure of interest community. For example, we use the nodes with high social closeness strength instead of current neighbors of a node in an interest community.

The social closeness strength contains three social features, such as contact time, intercontact time, and contact frequency. Figure 2 shows the contacts between nodes i and
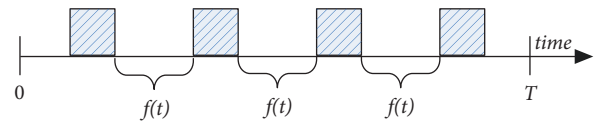
FIGURE 2: Contacts between nodes i and j.

j during the period $[0, T]$. In this figure, the width of a box represents the contact time, the time interval between two adjacent boxes represents the inter-contact time $f(t)$, and the frequency of contacts is four during the period $[0, T]$.

Next, we give the formal definition of social closeness strength. The average intercontact time between nodes i and j is denoted as

$$T_{ij} = \frac{\int_{t=0}^{T} f_{ij}(t) \, dt}{\gamma_{ij}} \quad (4)$$

where T_{ij} represents the average intercontact time between nodes i and j, $f(t)$ represents the time remaining to the next contact between nodes i and j after time t, and γ is the frequency of contacts between them during the period $[0, T]$.
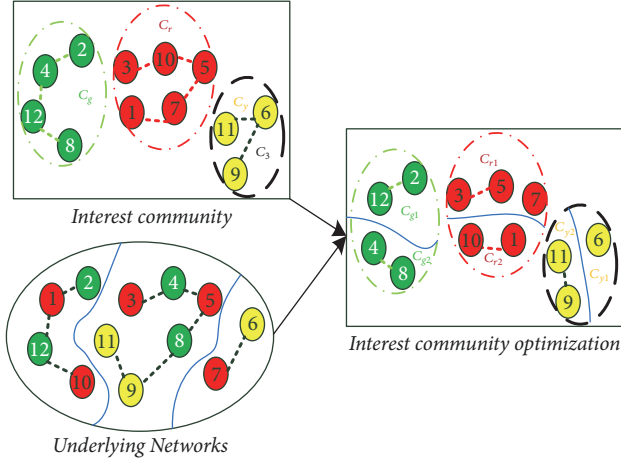
FIGURE 3: Community optimization using social relationship.

Definition 1 (direct social closeness strength). Direct social closeness strength between nodes i and j is the inverse of average intercontact time between the two nodes, as shown in formula (5). The smaller the value T_{ij} is, the closer the friendship between nodes i and j is.

$$\lambda_{ij} = \frac{1}{T_{ij}} = \frac{\gamma_{ij}}{\int_{t=0}^{T} f_{ij}(t)\, dt} \tag{5}$$

In OSN, it is possible for some nodes not to contact with others ever before. It is necessary to define the indirect social closeness strength.

Definition 2 (indirect social closeness strength). Indirect social closeness strength is defined as

$$\lambda_{ij}^{*} = \prod_{\substack{w \in relay}}^{w \neq i,j} \lambda_{iw} \times \lambda_{wj}$$

$$= \prod_{\substack{w \in relay}}^{w \neq i,j} \frac{\gamma_{iw} \times \gamma_{wj}}{\int_{t=0}^{T} f_{iw}(t)\, dt \times \int_{t=0}^{T} f_{wj}(t)\, dt} \tag{6}$$

where w represents the relay nodes between nodes i and j, γ_{iw} is the frequency of contacts between nodes i and w, γ_{wj} is the frequency of contacts between nodes w and j, $f_{iw}(t)$ represents the time remaining to the next contact between nodes i and w after time t, and $f_{wj}(t)$ represents the time remaining to the next contact between nodes w and j after time t.

Figure 3 depicts the process of interest community optimization. After the interest communities C_g, C_r, and C_y are constructed, the social closeness strength between nodes in the C_i is calculated using formulas (5)-(6). Let us take the interest community C_y as an example. Node n_6 tries to calculate the social closeness strength with nodes n_9 and n_{11} using formula (4), but node n_6 finds it to be impossible to calculate this value. And then, node n_6 tries to calculate the social closeness strength with nodes n_9 and n_{11} using formula

(6) again. Node n_6 finds that it is still no result. That is because node n_6 never contacts nodes n_9 and n_{11}. The edges between node n_6 and nodes n_9 and n_{11} are cut off. After that node n_9 tries to calculate the social closeness strength with node n_{11} using formula (5). As a result, node n_9 can get the direct social closeness strength with node n_{11}. The edge between nodes n_9 and n_{11} is added into the interest community C_{y2}. For the interest community C_g, the social closeness strength between nodes n_2 and n_{12} is calculated via the relay node n_1 by formula (6). Also, the social closeness strength between nodes n_4 and n_8 is calculated via the relay node n_5 by formula (6). The edge between nodes n_2 and n_{12} is added into the community C_{g1}. Besides, the edge between nodes n_4 and n_8 is also added into the community C_{g2}.

3.4. Time-Varying Community Model Based on Temporal-Spatial Correlations. The community structure depends on social behaviors of people and hence the constructed interest community has its life cycle. The life cycle of an interest community contains three states: born state, active state, and dead state. The three states depend on time and places of social activities. Next, we present time-varying community model. As shown in formula (7), we define two matrices T and S. T is the time matrix of state transition for a community, and the element t_{ij} in matrix T represents the transition time from state i to state j for a community. S is the probability matrix for a community stay at state i, and the element $s_{i\varepsilon}$ in matrix S represents the probability for the sojourn time of a community stay at state i less than ε time units.

$$T = \begin{pmatrix} 0 & t_{12} & \cdots & t_{1j} \\ t_{21} & 0 & \cdots & t_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \cdots & 0 \end{pmatrix},$$

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1\varepsilon} \\ s_{21} & s_{22} & \cdots & s_{2\varepsilon} \\ \vdots & \vdots & \ddots & \vdots \\ s_{i1} & s_{i2} & \cdots & s_{i\varepsilon} \end{pmatrix} \tag{7}$$

Then, the transition probability of a community from state i to state j after σ time units is defined as

$$L_{ij}(\sigma) = P(X_{\sigma} = j \mid X_0 = i) = p_{ij}s_{(i)(\sigma - t_{ij})} \tag{8}$$

where p_{ij} is the transition probability of a community from state i to state j. The value p_{ij} is generated by the social regularity of users. For example, most people work in their offices between 9 a.m. and 5 p.m. during the weekdays. After that, they will leave their working places and go home or attend parties of their friends.
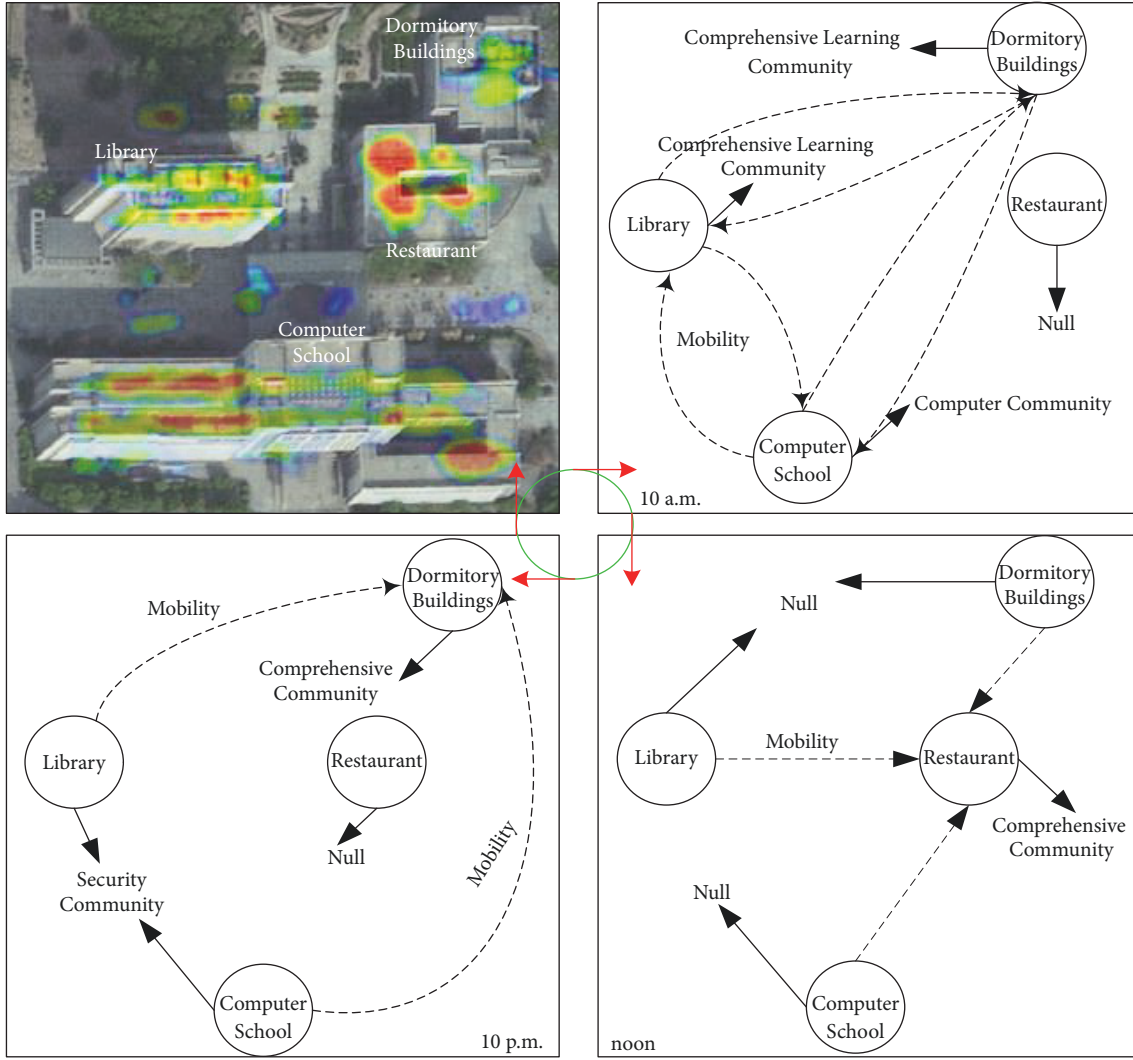
FIGURE 4: Temporal and spatial correlations-based community transformation.

When a community transits from state i into state j via intermediate state r, the transition probability of a community is defined as

$$L_{irj}(\sigma) = \prod_{r \neq i,j} \prod_{\tau=1}^{\sigma-1} L_{ir}(\tau) L_{rj}(\sigma - \tau) \qquad (9)$$

Hence, the general state transition probability of a community is shown in

$$L_{ij}(\sigma) = \begin{cases} p_{ij} s_{(i)(\sigma - t_{ij})}, & r = 0 \\ \prod_{r \neq i,j} \prod_{\tau=1}^{\sigma-1} p_{ir} p_{rj} s_{(i)(\tau - t_{ir})} s_{(r)(\sigma - \tau - t_{rj})}, & r \neq 0 \end{cases} \qquad (10)$$

We take an example to illustrate the transformation of a community, as shown in Figure 4. This figure shows that time-varying communities change with social activities of nodes. At 10 a.m., the computer community is constructed at computer school; the comprehensive learning community is

constructed at library and dormitory buildings. Few people are in restaurant, and hence no community is constructed. At noon, students and faculties enter restaurants. The comprehensive community is constructed at the restaurants. The constructed interest communities at computer school, library, and dormitory buildings are dismissed. At 10 p.m., students and faculties go back to their dormitory buildings. The comprehensive community is constructed at dormitory buildings. The constructed communities at computer school and library have been transformed into security community to protect security of public properties. After communities have transformed, some old communities are dismissed, and some new communities are constructed. The new communities are constructed and updated by Sections 3.2 and 3.3.

3.5. Content Dissemination Based on Interest Community. We take a programming interest community as an example to illustrate content dissemination based on interest community. From 9 a.m. to 5 p.m., the programming interest community is constructed at computer school and library
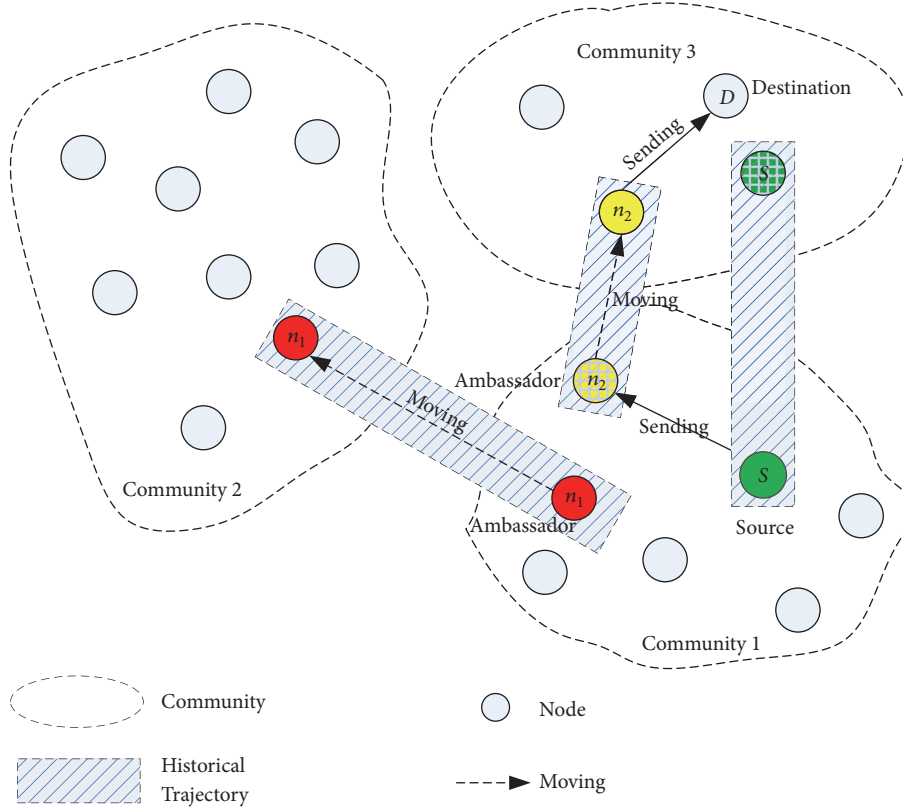
FIGURE 5: Content dissemination out of an interest community using mobility similarity.

buildings. From the sunset to the sunrise, the programming interest community is reconstructed at dormitory buildings. When the time-varying communities keep active, content dissemination includes two aspects. The first aspect is the content dissemination within an interest community. Another aspect is the destination node which is not in an interest community; the data message is sent out of the interest community. The content dissemination is described in the next part.

(1) Content Dissemination within an Interest Community. When there exist connections among nodes in an interest community, data messages will be directly exchanged with each other. Otherwise, the data messages will be delivered by binary spray and wait. The binary spray and wait content dissemination has the advantage of epidemic content dissemination, such as the high delivery ratio. It is worth noting that the binary spray and wait content dissemination efficiently control the number of data message copies. A source node initially generates L copies of a data message. When the source node meets another node, it will hand over half number of data message copies to the node and keep another half for itself. The process of content dissemination will continue until a node only has one data message copy. In this way, data messages will be quickly spread to nodes in the interest community.

(2) Content Dissemination Out of an Interest Community. When we cannot find a destination node within an interest

community, the data message is immediately sent out of the community by an ambassador. Figure 5 shows an example about the content dissemination out of the interest community by the comparison of mobility similarity.

The source node S is in community C_1, and nodes n_1 and n_2 are ambassadors. The ambassadors are a kind of special nodes who frequently switch between different interests communities, e.g., C_1 in Figure 5. The ambassador nodes have multiple interests and they are active most of the time. As shown in the figure, both source node S and node n_2 went to community C_3 ever before, which shows both of them have similar mobility trajectories. Hence, node n_2 has higher successful delivery probability than node n_1. Node S chooses ambassador n_2 to help forwarding the data message. After the data message is carried to the community C_3, the content dissemination within a community is enabled. Hereafter, the data message is successfully delivered to the destination D. The process for how to choose an ambassador is described as below.

Definition 3 (position of a node at time t). Position of a node at time t is defined as the vector (x_i, y_i, t_i). The symbols x_i and y_i represent the latitude and longitude of node i. The symbol t_i represents the current time.

Definition 4 (mobility area of a node during a period). t_s represents the time of a node who arrives at this area and t_l represents the time of the node who leaves this area. Assume that $p(x_i, y_i, t_i)$ is the position of a node at time $t_s \leq t_i \leq t_l$,

```
Input: Source node S, GPS traces of nodes, number of nodes in a community n, number of places in a map m
Output: The best ambassador
(1) Initiation: set variables as zero
(2) Begin
(3) for (i=0, i <n, i++)
(4)     for (j=0, j <m, j++)
(5)     {
(6)             calculate the mobility areas A_j using GPS traces for node n_i
(7)             if A_j is in the place P_j
(8)         {
(9)                 k_j+ = 1
(10)                add A_j(m_j, δ_j, ρ_j) into mobility trajectory T_r(i) for n_i
(11)         }
(12)     }
(13)    for (i=0, i <n, i++)
(14)    {
(15)        for (j=0, j <m, j++)
(16)        {
(17)                n_i(ρ)+ = A_j[ρ_j]
(18)        }
(19)    if n_i(ρ) ≥ ⌈m/2⌉
(20)    node n_i is an ambassador node
(21)    }
(22)    for (i=0, i < ambass_array.length, i++)
(23)    {
(24)    temp_array.add(traceSimilarityComparison(S, ambass_array [i]))
(25)    }
(26) Return max_value(temp_array)
(27) End
```

ALGORITHM 2: Choice of the best ambassador.

and $q(x_j, y_j, t_j)$ is the position of a node at time $t_s \leq t_j \leq t_l$. The time period is set as T_p equal to $(t_l - t_s)$. Then, we can calculate the maximum Euclidean distance $\max\{d(p, q)\}$ between the two positions p and q during the period T_p. Hence, the mobility area of a node during the period is defined as the geographical area A with its diameter equal to $\max\{d(p, q)\}$.

Definition 5 (mobility trajectory of a node). The mobility trajectory of a node is defined as geographical areas where the node entered ever before, as shown in

$$T_r = \left[A_1 \left(T_{p(1)}, \delta_1, \rho_1 \right), A_2 \left(T_{p(2)}, \delta_2, \rho_2 \right), \cdots, \right.$$
$$\left. A_n \left(T_{p(n)}, \delta_n, \rho_n \right) \right] \tag{11}$$

$$\delta_i = \frac{T_{p(i)}}{\sum_{i=1}^{n} T_{p(i)}} \tag{12}$$

where T_r represents the mobility trajectory of a node, A_i is the geographical area where a node entered ever before, $T_{p(i)}$ is the sojourn time for a node to stay at area A_i, δ_i denotes the weight of area A_i where a node has visited, and ρ_i is the times for a node to visit area A_i.

Next, we can get the mobility trajectory similarity between nodes i and j using

$$L_A = T_r(i) \cap T_r(j) \tag{13}$$

where $T_r(i)$ denotes the mobility trajectory of node i, $T_r(j)$ denotes the mobility trajectory of node j, and L_A represents the intersection of mobility trajectories between nodes i and j.

The source node S calculates the similarity of mobility trajectory with candidate ambassadors and then chooses the most similar ambassador as best ambassador to deliver the data message. Next, the pseudocode for choice of the best ambassador is shown in Algorithm 2. Finally, we use big-Oh notation to analyze the time complexity of Algorithm 2.

Content dissemination algorithm analysis: assume that there are m places for nodes in OSN and a node initially generates L message copies. Binary spray and wait algorithm is used as the content dissemination within a community. The binary spray and wait algorithm runs in $O(n+\log L)$ time. The value L is a constant value and hence the binary spray and wait algorithm runs in $O(n)$ time.

Next, the content dissemination out of a community is analyzed, as shown in Algorithm 2. The function between lines (3) and (12) is about calculating the mobility area and the mobility trajectory of a node. The function between lines (13) and (21) is about how to choose candidate ambassadors. Both code segments use $O(mn)$ time. The function between lines (22) and (26) is about the choice of the best ambassador. This code segment guarantees the worst-case running time of $O(n \log n)$ time.

TABLE 2: Simulation parameters.

Parameters	Description	Default values
WorldSize	Scenario area	$600 * 800 \ (m^2)$
EndTime	Simulation time	7(days)
NrofHosts	Number of nodes	90
Interface Type	Communication	SimpleBroadcast
Transmit Speed	Bandwidth	2 Mbps
Transmit Range	Transmission range	10~20m
Movement Model	Movement model of a node	ShortestPathMapBasedMovement
Mobility Speed	Walking speed of nodes	0.5~1.5(m/s)
Message Size	Size of a message	500~1024(KB)
Message Interval	Generation interval of data messages	25~35(sec)
Message TTL	The time to live value	300 (min)
NrofCopies L	Number of copies for binay spray and wait	6
Number of keywords	Nodes contain the number of keywords	5
ε	The minimal size of the community, containing the number of nodes	1
m	Places where nodes can move into or out	5
T_{sim}	Preset threshold for constructing interest community	0.3

The running time of the proposed content dissemination is given by the sum of these functions. Hence, the running time of the content dissemination is $O(2mn + n \log n + n)$ time. Note that $2mn + n \log n + n \leq 2(m + 1)n \log n$, and the symbol m is also a constant value. This indicates the running time of content dissemination algorithm is $O(n \log n)$ time.

4. Performance Evaluation

We use four performance metrics, success rate of data message delivery, average latency of data message delivery, average hop count of data message delivery, and system overhead, to evaluate the performance of proposed content dissemination. The success rate of data message delivery is the number of successfully delivered data messages divided by the number of generated data messages during the simulation. The average latency of data message delivery is the average latency from the source node to the destination node. The average hop count of data message delivery is the number of hops from the source node to the destination node. The system overhead is the amount of data message replicas to perform a successful data delivery. Our comparisons include the epidemic-based content dissemination, called epidemic [4], the spray, and wait-based content dissemination, called SW [5], and the interest community-based content dissemination, called SPOON [19].

4.1. Experimental Design and Parameter Settings. We implement and evaluate the proposed content dissemination on the simulation platform of Opportunistic network environment (ONE). The simulation parameters are shown in Table 2.

Figure 6 shows a snapshot of simulation running. In the experimental configuration file, we define the initial positions of 80 nodes on the four places of this map, such
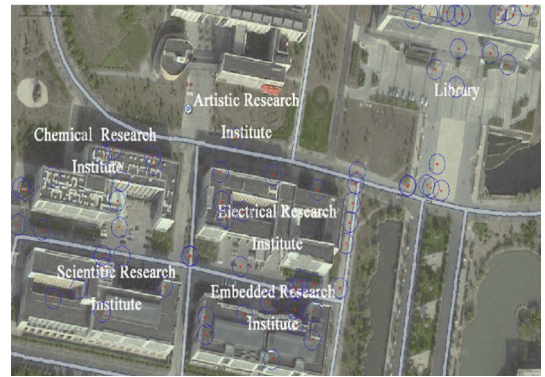

FIGURE 6: Snapshot of simulation running.

as embedded research institute, electrical research institute, chemical research institute, and scientific research institute. Then, the movement scope of 80 nodes is defined as the embedded research institute, the electrical research institute, the chemical research institute, the scientific research institute, the library, and the roads among the five places. The nodes in the five places construct their interest community according to Algorithm 1. Moreover, we define 10 nodes moving along the road of this map. Then, according to Algorithm 2, we can get the best ambassador. Finally, we use the real statistical data to determine the state transition probabilities among the five places. The statistical data is from a weekly survey about the movement of 80 students in the embedded research institute, the electrical research institute, the chemical research institute, and the scientific research institute. Table 3 shows the state transition probabilities among five places, where P_1 denotes embedded research institute, P_2 denotes electrical research institute, P_3 denotes chemical research institute, P_4 denotes scientific research institute, and P_5 denotes library.

TABLE 3: State transition probabilities among places.

	P_1	P_2	P_3	P_4	P_5
P_1	0.0%	8.1%	0.0%	10.8%	81.1%
P_2	10.7%	0.0%	3.6%	21.4%	64.3%
P_3	0.0%	7.7%	0.0%	46.1%	46.2%
P_4	10.0%	15.0%	15%	0.0%	60.0%
P_5	36.6%	22.0%	7.3%	34.1%	0.0%

(a) Number of interactions

(b) Simulation duration
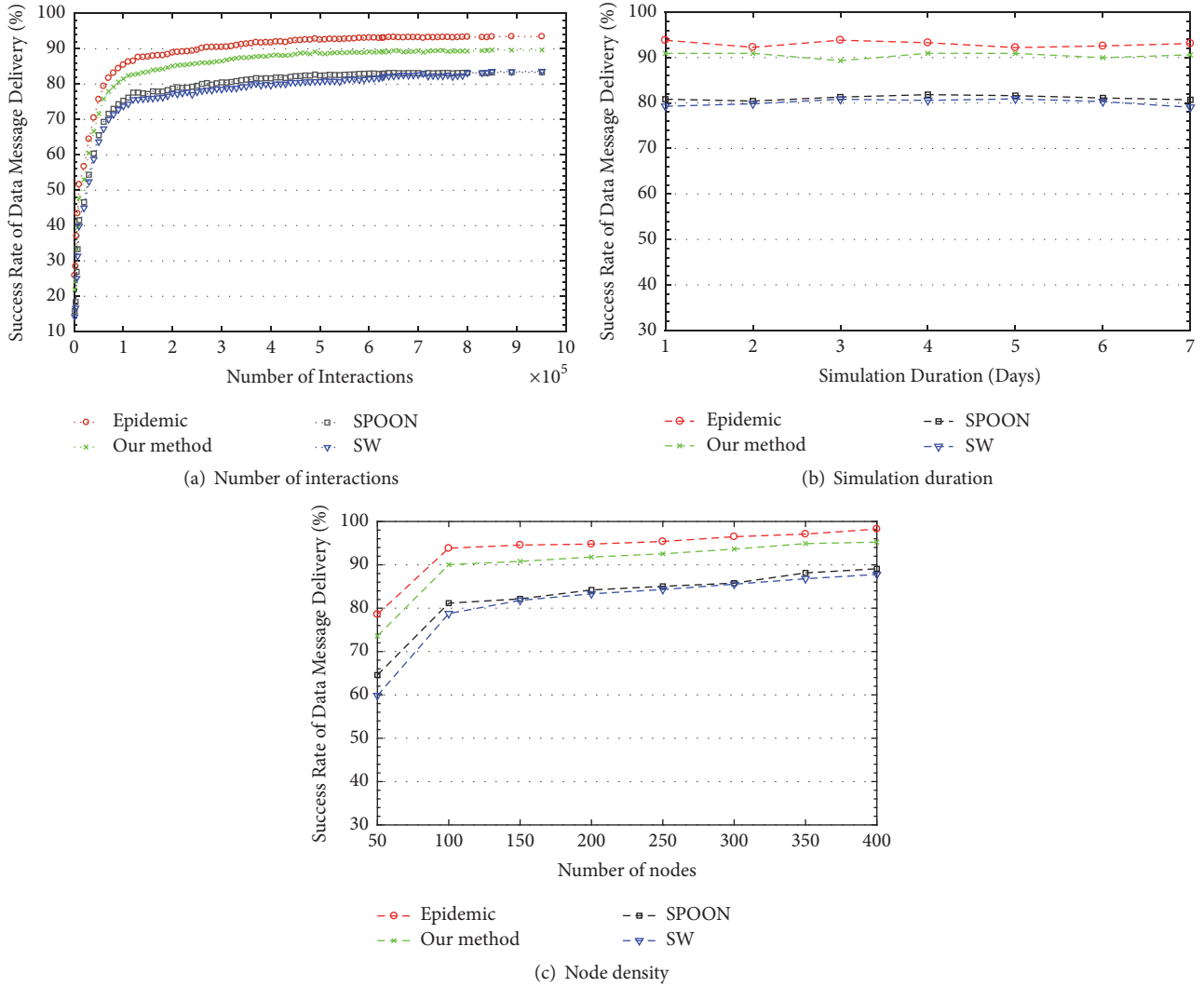
(c) Node density

FIGURE 7: Success rate of data message delivery.

4.2. Experimental Results and Analysis. In this experimental evaluation, *x*-axis is the interactions between nodes, simulation duration, and node density. The interactions are defined as the number of established connections between nodes during the simulation.

(1) Success Rate of Message Delivery. Figure 7 shows the success rate of data message delivery between epidemic, our method, SW, and SPOON. Figures 7(a), 7(b), and 7(c) show the success rate of data message delivery with the number of interactions, simulation duration, and node density, respectively.

There is no doubt that epidemic has the biggest success rate of data message delivery. Its success rate of data message delivery can reach up to 93%. With the increasing number of interactions, simulation duration, and node density, the success rates of data message delivery for the four methods gradually increase and converge at a specific percentage. It is due to the fact that more interactions between nodes can create more delivery opportunities. Hence, the success

(a) Number of interactions
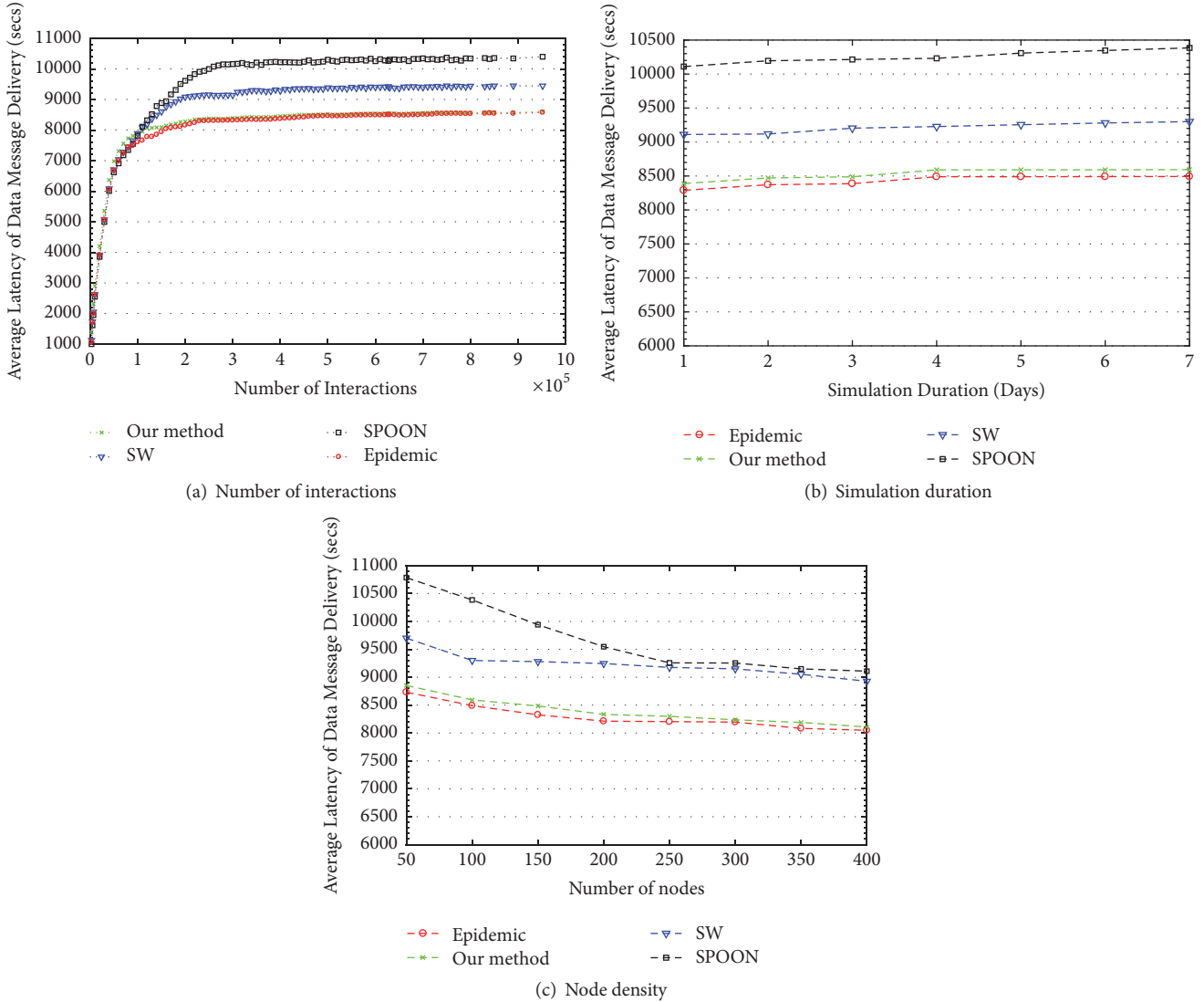

(b) Simulation duration

(c) Node density

FIGURE 8: Average latency of data message delivery.

rate of data message delivery gradually increases. Although the methods of SW and SPOON use different data message delivery strategies, the success rates of data message delivery for them can reach up to 83%. However, their experimental results still have a big difference with the epidemic method. Our proposed method is an improvement for the methods of SW and SPOON. The proposed content dissemination runs on the top of the optimized community structure. The results indicate that our method can greatly improve the success rate of data message delivery.

(2) Average Latency of Data Message Delivery. The average latency of data message delivery is a key performance for content dissemination. Long message latency means the message must occupy the valuable buffer space for a long time, and hence the low delivery latency is necessary. When the scale of interactions is big enough, the epidemic method can get the minimal query latency. At this experiment,

epidemic is chosen as the optimal baseline for delivery latency. Figures 8(a), 8(b), and 8(c) show the average latency of data message delivery with the number of interactions, simulation duration, and node density, respectively.

When the number of interactions is less than 100000, the methods of SW and SPOON outperform the epidemic method. Otherwise, the two methods spend more delivery latency to reach the destination nodes. The SPOON method needs more latency to transmit messages to destinations. It is due to the fact that the constructed interest communities by SPOON cause the topology mismatch. When a data message is delivered to one hop neighbor, the message is actually delivered more than once. Our method optimizes the topological structure of interest communities. The delivered message hop in interest communities is approximately equal to the physical hop on underlying networks. The average latency of data message delivery in Figure 8(b) is stable in one day, which shows the regularities of social activities for

(a) Number of interactions
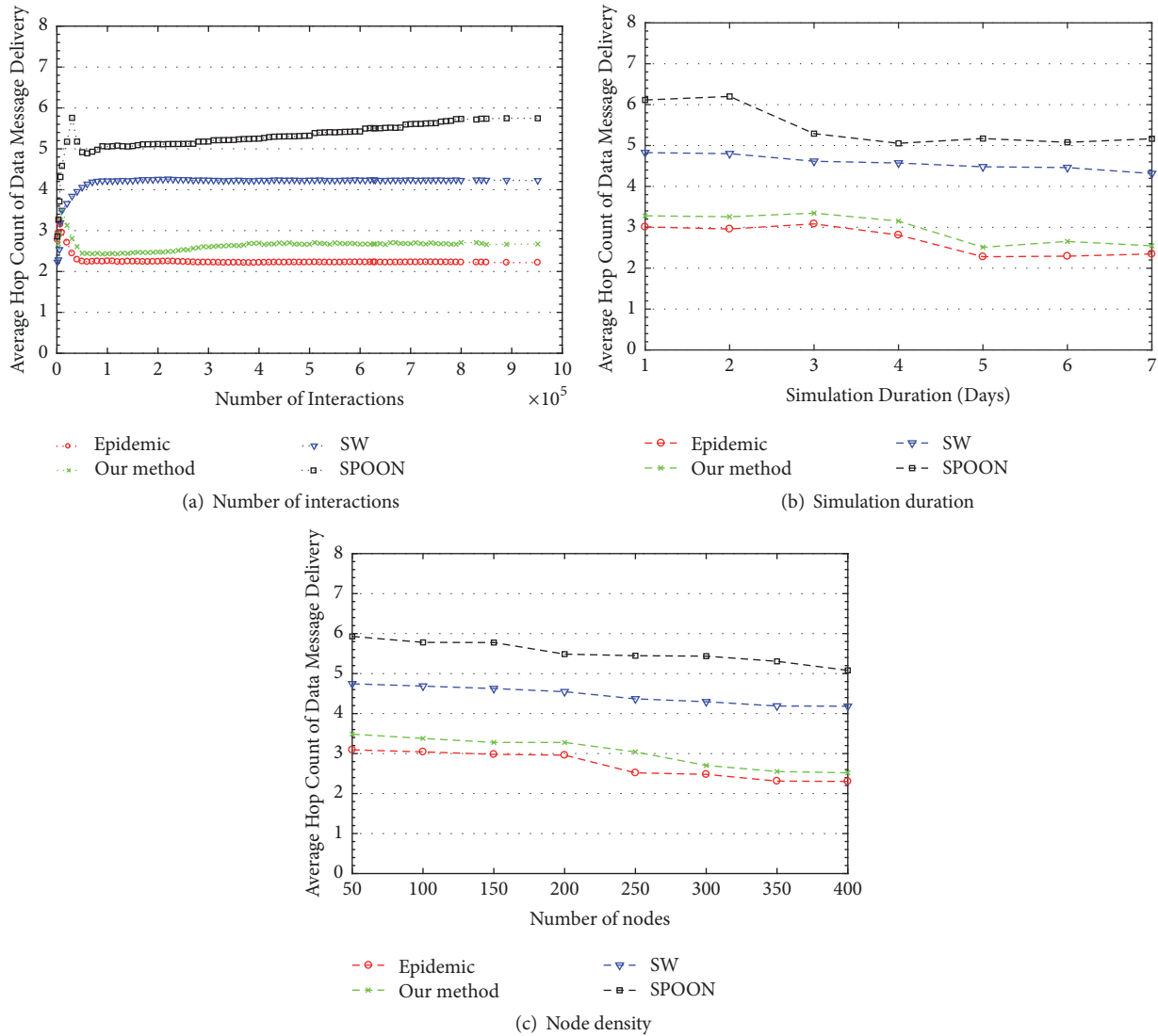

(b) Simulation duration

(c) Node density

FIGURE 9: Average hop count of data message delivery.

users in one week. Figure 8(c) shows the average latency of data message delivery becomes lower with the increasing of node density. The experimental results can also verify the correctness of our method. As shown in this figure, our method can get the lower delivery latency than the methods of SW and SPOON.

(3) Average Hop Count of Data Message Delivery. Figure 9 shows the average hop count of data message delivery. Epidemic is still considered as the benchmark. Epidemic is a multicopy content dissemination. When the data message replicas quickly spread, the data message easily reaches the destinations. As shown in Figure 9(a), the average hop count of the epidemic method firstly increases and then gradually decreases. Our method, SW and SPOON are the optimized multiple-copy content dissemination. However, our method outperforms other two methods. For our method, the average number of hops per a successful data message delivery is

2.68, which is very close to that of the epidemic method equal to 2.28. For SPOON method, the average number of hops per a successful data message delivery is 5.24. Figures 9(b) and 9(c) show the average hop count of data message delivery gradually decreases with the increasing of simulation duration time and node density. It is due to the fact that there will be more connections among nodes with the increasing number of nodes and simulation duration time.

(4) System Overhead. Figures 10(a), 10(b), and 10(c) show the system overhead comparisons among the methods in terms of the number of interactions, the simulation duration, and the node density, respectively. These figures indicate that the number of data message replicas generated by the epidemic method is the highest. The system overhead of our method is the lowest. As shown in this figure, the system overhead for epidemic is average 23 data message replicas for each successful data delivery. With the increasing of

(a) Number of interactions
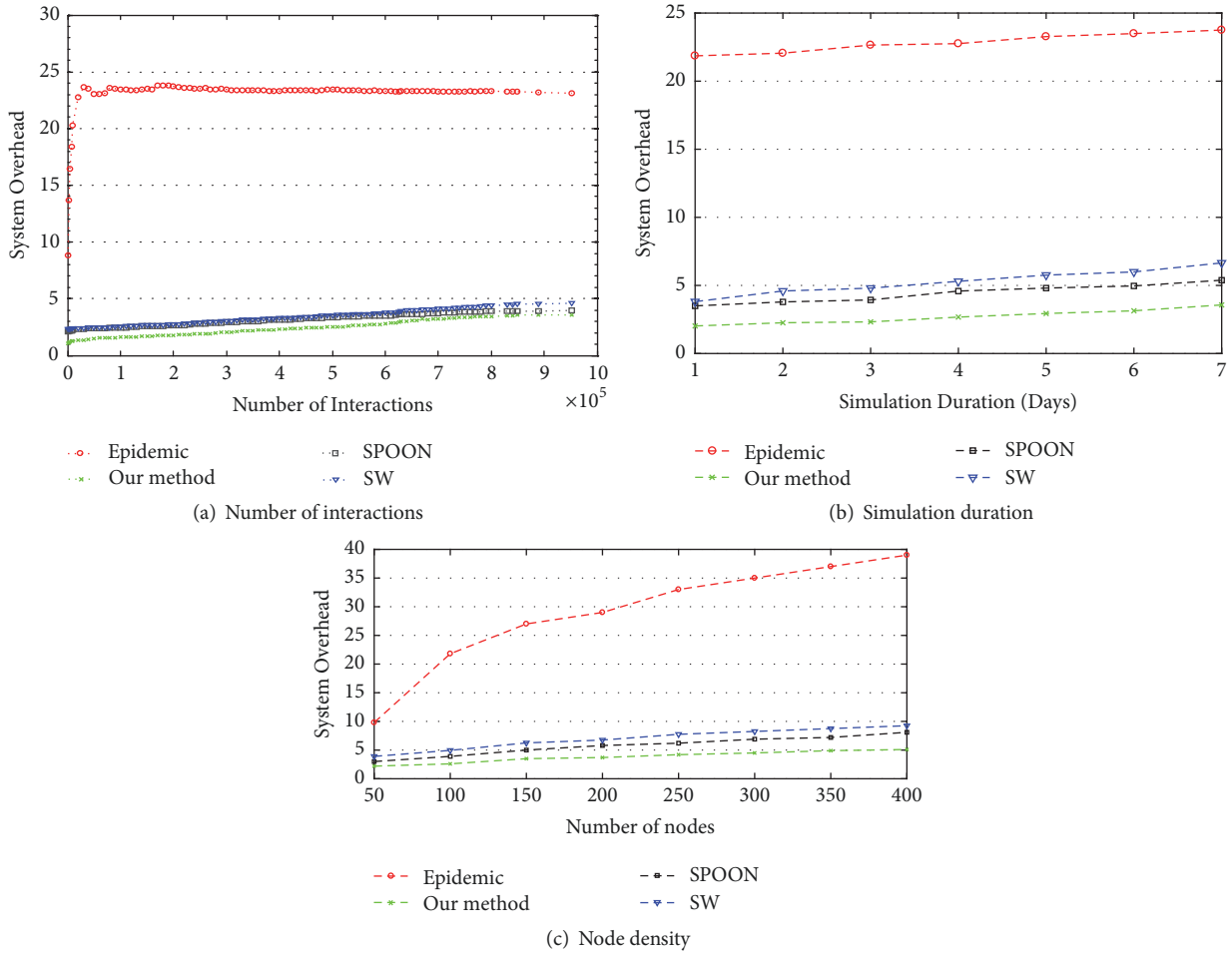

(b) Simulation duration

(c) Node density

FIGURE 10: System overhead.

node density, the system overhead of epidemic continues to increase. This means the message replicas generated by the epidemic method take up more bandwidth and buffer space than other methods. The advantage of our method is the combinatorial optimization between single-copy and multiple-copy content dissemination.

The experimental results show our method can reduce the system overhead compared with other content dissemination methods.

(5) Interest Communities Evolvement. Figure 11 shows the interest communities evolve over time. The period of time is every two hours. Communities 1, 2, 3, 4, and 5 represent the embedded research interest community, the electrical research interest community, the chemical research interest community, the scientific research interest community, and the comprehensive community, respectively.

The embedded research interest community is in embedded research institute. The electrical research interest community is in the electrical research institute. The chemical research interest community is in the chemical research institute. The scientific research interest community is in the scientific research institute. And the comprehensive
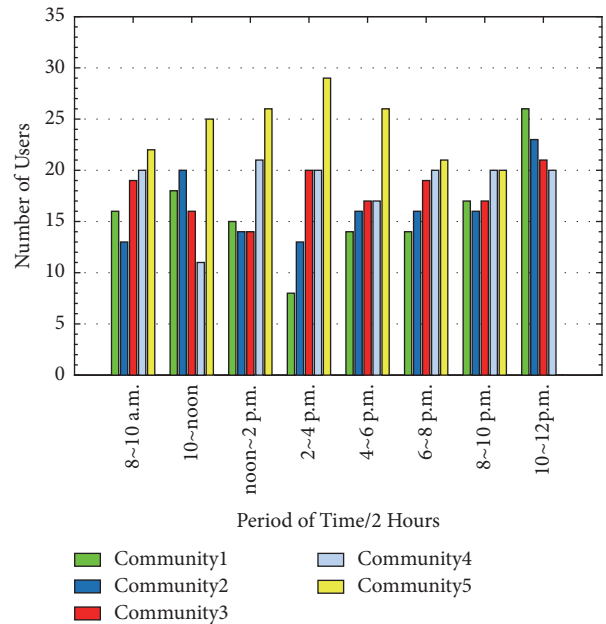

FIGURE 11: Interest communities evolvement.

community is in the library. The number of users in the library increases before 4 p.m. in one day. After 10 p.m., the comprehensive community dismisses in the library due to the door closing. The users in the comprehensive community return back their research institutes, respectively. Hence, the number of users in communities 1, 2, and 3 increases. The experimental results are consistent with the social activity regularities of uses in a university.

5. Conclusion

In this paper, we propose a content dissemination based on interest community for OSN. We exploit the temporal and spatial correlations to construct the time-varying community model and optimize the structure of interest communities. Then, we implement the content dissemination within an interest community and out of an interest community. Our proposed content dissemination algorithm can run in $O(n \log n)$ time. After the performance comparisons between our method and other classical methods, we conclude that our method comes close to the benchmark method epidemic in terms of success rate of data message delivery, average latency of data message delivery, average hop count of data message delivery, and system overhead. More importantly, our method may generate the low number of data message replicas. Next, we study and extend our method under the condition of selfish nodes in OSN.

Data Availability

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

Disclosure

This paper is an extension of the published conference in GLOBECOM 2017. An author in the conference paper Yue Song has left Jiangsu University when he received his M.S. degree in computer science from Jiangsu University, China, in 2017. So we removed his name from the author lists in this journal paper after his consent.

Conflicts of Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work; there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, the manuscript.

Acknowledgments

References

[1] N. Vastardis and K. Yang, "Mobile social networks: architectures, social properties, and key research challenges," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1355–1371, 2013.

[2] A. K. Pietilänen and C. Diot, "Dissemination in opportunistic social networks: the role of temporal communities," in *Proceedings of the 13th ACM International Symposium on Mobile Ad Hoc Networking and Computing, ACM*, pp. 165–174, June 2012.

[3] B. K. Saha and S. Misra, "D2D opportunistic local content dissemination sans location sharing," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6461–6468, 2018.

[4] A. Vahdat and D. Becker, "Epidemic content dissemination for partially connected ad hoc networks," Tech. Rep. CS-200006, Duke University, 2000.

[5] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: an efficient routing scheme for intermittently connected mobile networks," in *Proceedings of the ACM SIGCOMM Workshop on Delay-Tolerant Networking (WDTN '05)*, pp. 252–259, ACM, August 2005.

[6] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant MANETs," in *Proceedings of the 8th ACM International symposium on mobile ad hoc networking and computing (MobiHoc '07)*, pp. 32–40, September 2007.

[7] B. K. Saha and S. Misra, "Evaluation of opportunistic service provisioning with ordered chaining," *IEEE Transactions on Services Computing*, pp. 1–12, 2018.

[8] D. Yang, C. Shen, W. Lee, and M. Chen, "On socio-spatial group query for location-based social networks," in *Proceedings of the 18th ACM KDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp. 949–957, August 2012.

[9] E. Bulut and B. K. Szymanski, "Exploiting friendship relations for efficient routing in mobile social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 23, pp. 2254–2265, 2012.

[10] M. Girolami, P. Barsocchi, and S. Chessa, "A social-based service discovery protocol for mobile Ad Hoc networks," in *Proceedings of the 12th Annual Int. Conf. on Ad Hoc Networking Workshop (MED-HOC-NET)*, pp. 103–110, 2013.

[11] Z. Zheng, T. Wang, L. Song, Z. Han, and J. Wu, "Social-aware multi-file dissmination in device-to-device overlay networks," in *Proceedings of the 33rd IEEE Int. Conf. on INFOCOM, Computer Communications Workshops*, pp. 219-220, IEEE Press, April 2014.

[12] Y. Li, L. Qiu, D. Jin, L. Su, P. Hui, and L. Zeng, "Contact duration aware evaluation for content dissemination delay in mobile social network," *Wireless Communications and Mobile Computing*, vol. 15, no. 3, pp. 527–537, 2015.

[13] A. T. Mansilla and J. L. Esteva, "Question waves: a multicast query routing algorithm for social search," *Information Sciences*, vol. 253, pp. 1–25, 2013.

[14] L. You, J. Li, C. Wei, and L. Hu, "MPAR: a movement pattern-aware optimal routing for social delay tolerant networks," *Ad Hoc Networks*, vol. 24, pp. 228–249, 2015.

[15] E. Pagani, L. Valerio, and G. P. Rossi, "Weak social ties improve content delivery in behavior-aware opportunistic networks," *Ad Hoc Networks*, vol. 25, pp. 314–329, 2015.

[16] B. K. Saha and S. Misra, "Named content searching in opportunistic mobile networks," *IEEE Communications Letters*, vol. 20, no. 10, pp. 2067–2070, 2016.

[17] Y. Zhu, B. Xu, X. Shi et al., "A survey of social-based content dissemination in delay tolerant networks: Positive and negative social effects," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 387–401, 2013.

[18] S. Misra, R. Barthwal, and M. S. Obaidat, "Community detection in an integrated internet of things and social network architecture," in *Proceedings of the 55th IEEE Global Communications Conf. (GLOBECOM'12)*, pp. 1647–1652, IEEE Press, December 2012.

[19] K. Chen, H. Shen, and H. Zhang, "Leveraging social networks for P2P content-based file sharing in disconnected MANETs," *IEEE Transactions on Mobile Computing*, vol. 13, no. 2, pp. 235–249, 2014.

[20] Y. Li, D. Wu, J. Xua, B. Choi, and W. Su, "Spatial-aware interest group queries in location-based social networks," *Data & Knowledge Engineering*, vol. 92, no. 2014, pp. 20–38, 2014.

[21] F. Li and J. Wu, "MOPS: providing content-based service in disruption-tolerant networks," in *Proceedings of the 29th IEEE International Conference on Distributed Computing Systems (ICDCS '09)*, pp. 526–533, IEEE Press, Montreal, QC, Canada, June 2009.

[22] Y. Liu, Y. Han, Z. Yang, and H. Wu, "Efficient data query in intermittently-connected mobile ad hoc social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1301–1312, 2015.

[23] S. Chen, G. Wang, and W. Jia, "Cluster-group based trusted computing for mobile social networks using implicit social behavioral graph," *Future Generation Computer Systems*, vol. 55, pp. 391–400, 2016.