IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2019, accepted February 8, 2019, date of publication February 12, 2019, date of current version March 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2898895

Effective Data Transmission Strategy Based on Node Socialization in Opportunistic Social Networks

YEQING YAN, ZHIGANG CHEN, (Member, IEEE), JIA WU, (Member, IEEE), LEILEI WANG, KANGHUAI LIU, AND YUZHOU WU

School of Software, Central South University, Changsha 410075, China
"Mobile Health" Ministry of Education-China Mobile Joint Laboratory, Changsha 410075, China

Corresponding authors: Zhigang Chen (czg@csu.edu.cn) and Jia Wu (jiawu5110@163.com)

This work was supported in part by the Major Program of National Natural Science Foundation of China under Grant 61672540, in part by The National Natural Science Foundation of China under Grant 71633006, in part by the China Postdoctoral Science Foundation funded project under Grant 2017M612586, in part by The Postdoctoral Science Foundation of Central South University under Grant 185684, and in part by the "Mobile Health" Ministry of Education—China Mobile Joint Laboratory.

ABSTRACT With the development of big data and high-speed communication networks, traditional end-to-end transmission mechanisms in social networks are difficult to achieve large amounts of data communication between mobile devices. Therefore, the implementation of effective data transmission in social networks requires "opportunity". Opportunistic social networks suggest choosing the most appropriate next hop nodes for effective data transmission. Most existing routing algorithms attempt to use the interest points of nodes and the social relationships between them to choose optimal relay nodes among neighbors. However, most community-based algorithms take node attributes and social relations into account but fail to consider the energy consumption of inefficient nodes which accounts for a large part of routing cost. To improve the transmission strategy, this paper proposes an effective transmission strategy based on node socialization (ETNS), which divides nodes in the network into several different communities. The proposed scheme also involves a community reduction method that removes some inefficient nodes according to the attributes of optimal relay nodes. The simulation results show that the packet delivery ratio of ETNS is 13% higher than the epidemic algorithm, and ETNS also has lower transmission delay and routing overhead.

INDEX TERMS Opportunistic social networks, node socialization, trust evaluation, opportunity, community.

I. INTRODUCTION

With the popularization of network and the development of social informatization, message transmission based on various online social platforms has become extremely important [1]. Many social platforms, such as Facebook, Instagram and Twitter, have enough power to support billions of users to participate in information transfer process [2]–[4]. Through online platforms, people can attract more public attention by sharing interesting things in their lives. As users communicate and get online with mobile devices, they can post photos or videos wherever and whenever they want [5]–[10].

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad S. Khan.

Nowadays, networks often have the ability to communicate at high speed, and it is very important in social interactions. The history of data exchange activities can be recorded and analyzed by assessing human communication activities and their interest preferences [11]. With the development of online communication platforms, individual commodity recommendations have become effective [12]. However, the process of retrieving large amounts of structured data from human activities is very complex and requires a large of storage and computing resources. This characteristic makes some traditional wireless sensor network approaches no longer applicable.

When we face this problem, it is necessary and important to establish a suitable environment in wireless networks to ensure the stability of data transmission. Opportunistic

network is a kind of working structure suitable for wireless communication research. The biggest feature of this scheme is that the information transmission between nodes needs to look for ''opportunity''. This way of information transmission can provide communication services by node movements and cooperation among them [13]–[16]. The method of opportunistic network is more and more applicable to the scenarios of social network because people contact each other by chance using mobile wireless devices. In online social networks, ''opportunity'' [17] can be provided by reliably neighbors which have enough resources and cache room to save what we want to share, such as images and videos, or have similar points of interests to share our experiences. Since nodes in online social networks need to wait for suitable neighbors to appear, the state of ''storage'' and ''carry'' are suitable for online social networks. The ''forwarding'' state in social networks may represent a valid data transmission process. ''Opportunity'' in the studies of social network means that it is possible to determine whether useful information can be transmitted. In social networks, only reliable neighbors can participate in the selection of the optimal relay nodes when establishing node communication.

Methods based on the characteristics of opportunistic network structure have become more and more popular in the establishment of message transmission model in the study of online social networks. Researchers believe that networks that perform message routing and data sharing using human social characteristics such as similarity, daily transactions, mobility patterns and interests are opportunistic mobile social networks. In this kind of network, the social network is regarded as the application scenario, and the work structure of opportunistic network is used to solve problems [18]–[22].

In opportunistic social networks, each node has a large number of social attributes which represent the relationship between different users. Therefore, different algorithms [23], [24], [42] can divide nodes in the network into communities through various methods and different attributes of nodes to improve algorithm performance. Unsupervised learning [25], [26] is a branch of machine learning that learns from test data that has not been labeled, classified or categorized. Clustering algorithm is an important type of unsupervised learning. Unsupervised learning plays an important role in data mining [27]–[30], and have a wide spread use in our real-life world, such as community division [31].

Despite the emergence of community-based opportunistic network transmission strategies, how to divide efficient communities is still a hot issue. This is not because the existing community partition method is not feasible, but because the community partition does not consider whether all nodes in the community can meet the transmission requirements. In fact, most of processes of community partitioning consider the interest points and social relations of nodes in the real scene, but not every node in the community are suitable for transmission. These kind of nodes in communities are inefficient and consume transmission resources, but do not improve transmission performance, so it is necessary to propose a reduction method to cut down the cost. Social network communications with large amounts of data transmitted simultaneously can result in excessive energy consumption, low delivery rates, and transmission delays. Thus, it is necessary to propose a community reduction method to improve the performance of community-based algorithms.

To solve this problem, an effective transmission strategy based on node socialization is proposed in this work, which divides nodes in the network into several different communities using clustering method. A community reduction method that reduces inefficient nodes according to attributes of optimal delay nodes is presented in this strategy. The established Effective Data Transmission Strategy based on Node Socialization (ETNS) enables efficient transmission of messages between source node, communities and target node. After analyzing the data information required by the algorithm proposed in this paper, we comprehensively considered and compared the existing real data sets, and selected the most suitable four datasets for simulation experiments. Excellent performance is achieved by enhancing the aggregation relationship between nodes and reducing the energy consumption of inefficient nodes in social communication environments.

The contributions of this paper are listed as follows:

1) By analysing the characteristics of nodes in the network comprehensively, the concept of community reduction is proposed and an effective method is established to reduce inefficient and unapplicable nodes in communities.
2) An effective data transmission strategy based on node socialization is proposed with analysing the clustering features of nodes in social communications. Nodes in networks are divided into several non-overlapping communities of similar size according to the similarity between nodes.
3) According to the conditions for selecting optimal next hop nodes in networks, this article proposes four unique features for nodes. The active index in this paper is an unique concept originally proposed by authors for the mobility characteristics in opportunistic networks.
4) Using the simulation tool ONE (Opportunistic Network Environment) [32], this work analyzes the performance of ETNS algorithm and compared it with other four algorithms. Simulation results prove the algorithm proposed in this paper can improve the performance of network communication in terms of packet delivery ratio and routing overhead.

The rest of this paper is structured as follows. In Section 2, we briefly describe and analyze related works. The model of ETNS algorithm is proposed and analyzed in section 3. In the section 4, the complexity analysis of system model are provided. The simulation results are shown in Section 5. The last section concludes this paper. The symbols and notations used in this work is shown in table 1.

TABLE 1. List of notations.

Notations	Definitions
H_i^n	The number of nominations accumulated by node V_i during the $n-th$ iteration
$\tilde{H}_i^n$	Normalized nomination value of node V_i
C_i	The clustering coefficient of the node V_i
$D(x_i, x_j)$	The Euclid distance between node V_i and V_j
$S(x_i, x_j)$	The similarity between node V_i and V_j
$Q_c(x)$	The modularity function
LR_{ij}	The trust evaluation of the sender node V_j when the node V_i is the receiver
LS_{ij}	The trust evaluation of the receiver node V_j when the node V_i is the sender
$N_{Rh}(i,j)$	The number of honest transactions with node V_j when V_i is the receiver node
$N_{Rm}(i,j)$	The number of malicious transactions with node V_j when V_i is the receiver node
$N_{Sh}(i,j)$	The number of honest transactions with node V_j when V_i is the sender
$N_{Sm}(i,j)$	The number of malicious transactions with node V_j when V_i is the sender
N_{pun}	The penalty coefficient
GR_i	The global trust value of the node as the sender
GS_i	The global trust value of the node as the receiver
$r_n(t)$	The amount of data received by node V_n during a time interval t
B_S	The cache occupied by a node collection unit data
B_T	Cache consumption rate
$B_n^{total}(t)$	The total cache of node V_n in time slot t
$B_n^{remain}(t)$	The remaining cache of the node n in time slot t
$\widehat{\tau}$	The seconds of the mapping time
$\widehat{T}_i$	The mapping time
$Dist_i$	The active index of node V_i
$E(F_i)$	The information entropy function
E_j	Corresponding entropy value of each feature index
W_j	Corresponding weight for each feature indicator j

II. RELATED WORK

With rapid growth in the number of human using portable communication devices recently, it becomes more necessary and urgent to evaluate opportunistic networks formed by various devices carried by users. At present, since research around the routing algorithms have been a hot issue, various methods have been proposed for different application scenarios. According to the relevance to communities and social attributes, existing routing algorithms of opportunistic social networks can be roughly classified into two categories: community-ignorant routing algorithms, and community-based routing algorithms.

A. THE PROPOSED COMMUNITY-IGNORANT ROUTING ALGORITHMS

In existing social-ignorant routing algorithms, transmission strategies are proposed to improve the performance of opportunistic networks which do not consider social relationships. Epidemic algorithm [33] is a flooding routing algorithm which takes the source node as the pathogen and other nodes in the network as vulnerable populations. The source node transmits messages to all nodes it encounters, so epidemic algorithm has the highest packet delivery ratio and routing overhead theoretically. Direct transmission algorithm [34] has the lowest successful rate of packet delivery because the source node only forwards messages to the target node, which also means that the algorithm has the lowest routing overhead. Sisodiya *et al.* [35] is a controlled replication-based opportunistic routing algorithm which has heavy routing overhead. Borah *et al.* [37] proposed an efficient probabilistic routing protocol based on encounter and transmission history, which controls flooding in opportunistic networks. However, due to the social attribute features of opportunistic social network, it is necessary to consider the social attributes of nodes when calculating the encounter probability based on historical records. Kanghuai *et al.* [48] explored the correlations of social attributes of nodes in opportunistic networks, and proposed an algorithm which take mobile similarity into account. The transmission of nodes in real life is not completed by single node, but by a group of nodes, which is not considered by the author.

Some complex mathematical methods are have been applied to improve the performance of opportunistic transmission models, such as decision trees, Markov chains, etc. In literature [38], Dhurandher *et al.* proposed a forwarding strategy to predict location of nodes, which utilize the model of markov chain to calculate the encounter probabilities for nodes. But in real scenarios, the performance of this algorithm is affected by the movement of devices, because the movement data is difficult to update immediately. Sharma *et al.* [39] proposed a machine leaning-based protocol for effective opportunistic routings. This algorithm uses the concept of decision tree and neural networks to predict the successful ratio of packet deliveries. Li *et al.* [40] proposed a cross-layer opportunistic routing algorithm, which adopts the fuzzy logic and topology technologies to control the scheme. The algorithm has the characteristics of high packet transmission rate and low computational complexity. Considering the social attributes of nodes, these two routing algorithms [39], [40] will have better performance in real urban scenarios. Wang *et al.* [41] proposed an efficient data transmission scheme which discussed three different

scenarios and considered mutually exclusive requirements in transmission process. This work reduces propagation latency while causing a small routing overhead, but in our scenario there is a high cost.

B. THE PROPOSED COMMUNITY-BASED ROUTING ALGORITHMS

As online social network is the main social media and advertisement platform nowadays, routing algorithms that take social relationships and community into account are necessary and suitable for different opportunistic social application scenarios. Research work [42] presented a weight distribution and community reconstitution based on communities communications in social opportunistic networks. Wu *et al.* mentioned that many routing algorithms deliver messages depending on one or two nodes, and thus may cause huge transmission delay. So they proposed a strategy to reduce transmission delay, but the packet delivery ratio is not improved much. Fu *et al.* [43] proposed an interest-driven community mobility strategy, which not only considered the social attributes of human behavior, but also considered the location preference and time variance of each node. This routing scheme can improve the transmission ratio, but there is no significant improvement in routing overhead and transmission delay. Dragan *et al.* [44] proposed that nodes could be divided into several communities according to their closeness and time spent together, and leaders could be selected in each community. This community division approach can improve the performance of opportunistic network routing algorithms.

In research work [45], dynamic social characteristics and their enhancement are introduced to obtain the historical contact behavior of nodes. Considering more social relations among nodes, community structure is adopted to select the next hop node in the scheme of multicast comparison and segmentation to improve the efficiency of information transmission. Zhang *et al.* [46] proposed a routing protocol that use fine-grained contact characteristics in communities to design a mobile opportunistic network. It used the sliding window mechanism to characterize the contact history and timing information in a fine-grained manner, and predicted future contacts based on fine-grained contact information, thus improving the accuracy of contact prediction. Zeng *et al.* [23] proposed a community-based routing strategy, in which each node selects the nodes with close social relationships to form communities. Tao *et al.* [47] expressed their opinion that exploiting community structure for opportunistic forwarding in social mobile networks is promising to improve routing performance. However, there are some inefficient nodes in the communities, which will affect the transmission efficiency of the transmission strategy.

One application of the social-aware routing algorithms is the vehicle networks. Wang *et al.* [49] addressed a social-based encounter utility rank router in cooperative vehicular sensor networks, which can enhance the delivery ratio by relaying the messages via remaining lifetime. Leilei *et al.* [50] proposed a social-based strategy to predict vehicle trajectory. The advantage of this algorithm is that the calculation process considers both movement prediction and trajectories distribution. The application of routing algorithms [49] and [50] in the vehicular sensor networks are good, but they don't perform well in the opportunistic social network scenarios.

FIGURE 1. In real life, human beings have complex social relationships as carriers of mobile communication devices. Therefore, when a mobile device communicates as a node in a network, it has social attributes, and this social relationship affects the transmission performance of the node.

Different to existing researches, we take the advantage of exploiting community structure for opportunistic social networks and presents a community reduction strategy. This scheme pays attention to activities and attributes of nodes, using the concept of information entropy to evaluate the three features of nodes. Three feature propose in this paper are the global trust of nodes, the caching room of mobile devices and active index of human beings. Through considering the importance of social attributes in opportunistic networks, we can judge the behavior of nodes in the communities and establish a community reduction strategy to decrease the number of inefficient nodes.

III. SYSTEM MODEL DESIGN

The development of social interaction has gone from face-to-face interpersonal communication to telephone and fixed networks, and eventually broke out to various forms of web-based interaction enabled by the ubiquity of the internet. These social interactions no longer require physical presence amongst participants and are enabled by many software applications, which make relationships among people complicated but convenient as shown in figure 1.

As we discussed above, the applications of social network's theory have been largely used to build opportunistic network models. A key concept in social network analysis is ''community'' which is a group of people with social relationships. In opportunistic social networks, the distribution of nodes has the characteristic of aggregation in a certain period of time as shown in figure 2. Since mobile users with social connections usually have long-term collaboration and are less error-prone than individual nodes, community-based algorithms outperform traditional routings in opportunistic social networks. Next, we will give a detail introduction of our algorithm, which takes advantages of existing community-based routings and improve it.

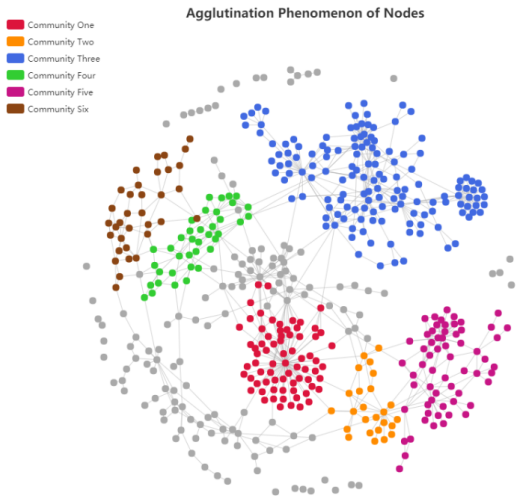


FIGURE 2. In opportunistic social networks, nodes in the network have an aggregation phenomenon in a certain period of time, which is very similar to human communities.

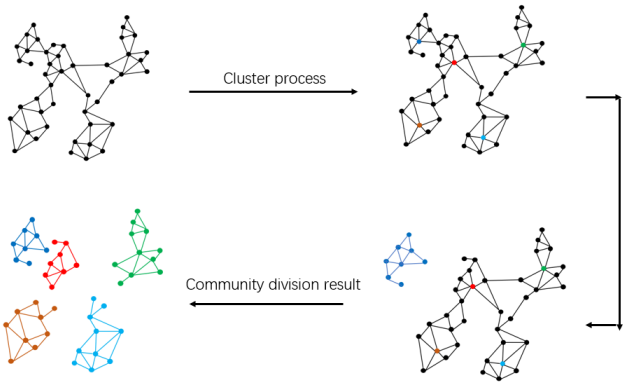


FIGURE 3. The community division process in opportunistic social network. Through the clustering process, the nodes in the network can be divided into several different communities.

A. CLUSTER-BASED COMMUNITY DIVISION

The method of community division is cluster based and needs to be divided into two stages. In the first stage, we need to get the clustering times. In the second stage, the characteristics of nodes are analyzed and then the community will be divided by clustering method. The number of clusters can be obtained according to the number of important nodes in the network, and the clustering process is performed by comparing the similarities between the nodes. Through clustering process, nodes in the network can be divided into several communities as shown in the figure 3.

1) DETERMINATION OF CLUSTERING TIMES

Nodes in networks have different levels of activity when transmitting data. There are always some active nodes in a network structure, which help other nodes to send data through their continuously motion. In general, we believe that nodes with higher activity are more important in the network. Correspondingly, there are always some nodes in the networks that do not actively participate in data transmission process. If such inactive nodes are selected as relay nodes during data forwarding, messages are usually not successfully transmitted or the transmission cost is high. The clustering process is actually the process of neighboring nodes grouping around the nodes with high activity, which likes human communities.

In the case of unsupervised clustering, the similarity between the clustering center and its neighbor nodes is adopted to determine whether the node can cluster with the clustering center. Therefore, the number of clustering center nodes is very important for community division process, and the critical condition to judge whether the node can become the clustering center is the importance level of nodes. There is a fact we need to know is that clustering center nodes must be very important in the network, but not all important nodes can be clustering centers. In other words, the number of important nodes must be larger than the number of clustering centers. Therefore, if we get the number of important nodes and use it as the clustering number, we can ensure that the clustering times is sufficient and appropriate, so as to ensure the good performance of the clustering process.

The importance of nodes in a network can be defined as reputation value which can be accumulated by a nomination mechanism, and depends on how many times that node successfully pass messages. Each of node is assigned an initial nomination value when estimating the importance of them. In the subsequent information transmission process, each time the data is successfully forwarded by node i, the nomination value H_i^n of the node is increased by one. The reputation of a node should not depend only on the number of times it participates in data forwarding, but also on the performance of its neighbors during data transmission. In this case, the reputation of a node is ultimately defined as the sum of the nomination value of the node itself plus the nominations of all nodes connected to it in each iteration. If a node transmits and forwards information stably in the network, the number of nominations of it will be a constant after multiple rounds of interaction. Nodes with a high reputation value are considered to be important nodes in the network. The nomination accumulation process is expressed as:

$$H_i^{n+1} = H_i^n + \sum_j^n a_{ij} H_j^n \tag{1}$$

where H_i^n represents the number of nominations accumulated by node V_i during the $n-th$ iteration, and the initial nomination value is 1. In each iteration, the information needed for calculation is only selected from the information provided by neighbor nodes in the network structure at the current time t. a_{ij} represents the element of the $i-th$ row and the $j-th$ column in the adjacency matrix A of the network. In order to compare the reputation value of nodes in the network intuitively, this work normalized the reputation value H_i^n. The normalized nomination value $\tilde{H}_i^n$ after n iterations can be expressed as:

$$\tilde{H}_i^n = \frac{H_i^n}{\sum_j^n H_j^n} \tag{2}$$

According to equation (1) and equation (2), if normalization process is performed after each iterative process, the normalized reputation of node V_i can be expressed as:

$$\tilde{H}_i^{n+1} = \frac{\tilde{H}_i^n + \sum_j^n a_{ij}\tilde{H}_j^n}{\sum_{k=1}^n [\tilde{H}_k^n + \sum_j^n a_{kj}\tilde{H}_j^n]} \tag{3}$$

among

$$\sum_i^n \tilde{H}_j^n = 1 \quad and \quad \tilde{H}_i^n = 1 \tag{4}$$

The size of reputation determines the number of important nodes in the network, which can be used to estimate the approximate proportion of important nodes. If the reputation value of a node is greater than the average reputation value of all nodes in the network, the node is considered to be an important node. Therefore, we can get the number of important nodes and use it as the clustering times.

2) CLUSTER COMMUNITIES OF SIMILAR SIZE

In an opportunistic social network scenario, mobile nodes represent people equipped with communication devices, and these nodes can belong to one or more dynamic networks. If the cluster community is too large or too small, it will affect the performance of routing algorithm. In order to avoid this problem, a clustering algorithm is established to ensure that the size of each community is similar by limiting the number of nodes near the cluster center. We define a neighborhood parameter named *Eps* and a minimum data node parameter *MinPts* to control the clustering range.

Definition 1: Limited domain. Limited domain represents an annular region between circles with a radius of *Eps* and 2*Eps*, centered on a given object.

Definition 2: Expansion node. Expansion nodes mean nodes located in the ring area which centered on the core node and have a radius of (*Eps*, 2*Eps*).

Definition 3: Modularity function. It can be used to quantitatively measure the quality of community partitioning. Community with high modularity value have dense connections between the nodes within Clusters.

The aggregation coefficient is an index that indicates the degree of aggregation of nodes in a network. After determining the number of clusters by the reputation value of nodes, we can select the cluster center by the clustering coefficient. The clustering coefficient C of node V_i is defined as:

$$C_i = \frac{E_i}{T_i}, C_i \in [0, 1] \tag{5}$$

$$T_i = \frac{k_i * (k_i - 1)}{2} \tag{6}$$

where the degree of the node V_i is k_i, representing the number of all adjacent nodes of node V_i. E_i represents the actual number of connected edges between the $k - th$ neighbor nodes of node V_i. T_i represents the maximum number of connections that the $k - th$ neighbor node of node V_i may form. When $C_i = 1$, it represents that all neighbor nodes of node V_i are connected, and means that node V_i is absolutely located central. In the clustering process, the nodes with the largest clustering coefficient among nodes that have not been divided into the community are selected them as the clustering center in order.

When clustering is used for community partitioning, the size of clusters is not consistent. In order to reduce the irrationality of community partitioning caused by inconsistent cluster size, we limit the clustering range of each center nodes by setting an *Eps* domain. If the number of nodes in the *Eps* domain of current node has reached *MinPts*, current community no longer receives other nodes. We will compare the aggregation coefficients of the nodes in the node set, and select the node with the second largest aggregation coefficient as the next cluster center.

After selecting the cluster center, the nodes are clustered according to the similarity between the comparison nodes, which is measured based on the distance between pair of nodes. If the shortest path sizes of node *i* and node *j* reaching other nodes are similar, the two nodes are very similar. This is why euclid distance can be used to measure the similarity of nodes. Suppose that G is an undirected graph with M nodes, $G = \{x_1, x_2, \ldots, x_m\}$, x_{ik}, x_{jk} respectively represent the shortest path from node V_i, V_j to node V_k. When each community substructure is divided, the distance information between nodes in the substructure at the current time t is used to calculate the euclid distance. The euclid distance between node V_i and V_j is expressed as:

$$D(x_i, x_j) = \sqrt{\sum_{i=1}^M (x_{ik} - x_{jk})^2} \tag{7}$$

The euclid distance and similarity are inversely proportional in value. The smaller the euclid distance, the greater is the similarity between these two nodes will be. In order to ensure that the similarity range of any two nodes belongs to (0,1), we define the similarity $S(x_i, x_j)$ as:

$$S(x_i, x_j) = \frac{1}{1 + D(x_i x_j)} = \frac{1}{1 + \sqrt{\sum_{i=1}^M (x_{ik} - x_{jk})^2}} \tag{8}$$

According to equation (8), we calculate the similarity between all nodes and store the results in corresponding dissimilarity matrix. The average similarity value of all nodes can be used as a threshold to determine whether the node meets the requirements in the clustering process. When the number of nodes in the *Eps* domain has not reached *Minpts*, all nodes which has higher similarity than average $\bar{S}$ are added to the current community. In the process of community division, nodes may be located in multiple communities. In order to avoid the occurrence of overlapping communities, we put forward the concept of modularity. Modularity is often used in optimization schemes for detecting community structure in networks [51], and a method utilizing modularity to avoid overlapping community is proposed in this paper. The modularity function is defined as:

$$Q_c(x) = \sum_{x=1}^m [\frac{l_x^{in}}{L} - (\frac{d_x}{2L})^2] \tag{9}$$

where $l_x^i n$ represents the number of connected edges within the community; d_x is the degree of nodes contained within the community; L is the total number of connected edges. According to formula (9), it is easy to find that subcommunities with higher modularity have closer internal structure and thus has higher information transmission efficiency. The calculation of modularity is based on the relevant information of peripheral data at a certain time t, and the required information is the information of the current subcommunity. If a node can be partitioned into two different communities repeatedly during the clustering process, the node will be divided into the community with larger values of the modularity $Q_c(x)$ rather than communities with lower modularity. Such community division can avoid overlapping communities and improve the quality of community division.

The flow-chart of the clustering process is shown as figure 4. The community division method adopted in this paper utilizes the idea of unsupervised clustering, but the clustering times are obtained by preprocessing before clustering. In addition, domain constraints are given in clustering process so that the final community structure size is similar. The improvement proposed in this paper on traditional clustering can reduce the consumption of node resources in the process of community partition.

B. REDUCTION STRATEGY OF INEFFICIENT NODE BASED ON MULTIPLE ATTRIBUTES

Community is combined by several nodes with close internal connections, but not every node in community has the ability to become a relay node. Nodes in the community that do not meet the transmission conditions can be considered as inefficient nodes. Removing these inefficient nodes in communities can reduce the transmission overhead and make the community more efficient. We analyze the characteristics of conditions that optimal relay nodes need to meet, and propose four indicators to measure nodes within communities. Inefficient nodes in communities can be identified and deleted according to the methods described below so as to achieve the goal of improving transmission success rate and reducing overhead.

1) RECEIVE TRUST AND SEND TRUST

In social life, social interactions and trades between people need to be based on trust. Only people who are honest can successfully trade with others, and so does the process of information transfer in opportunistic social networks. Trust can be used to evaluate the performance of a node and determine whether this node is trustworthy. The performance of nodes as different roles in the process of information transmission can be used to measure the reputation of nodes, this paper adopts a distributed trust mechanism which uses local trust and global trust to directly evaluate the integrity of the node.

Each nodes evaluate the local trust value of other nodes according to the historical transmission records and social relationships. Since social relationships are difficult to quantify in real-time, this paper only evaluates the integrity based on the historical transaction records. We define an evaluation mechanism for information exchange. Nodes give good or bad evaluations after each message transmission process, and local trust values are quantified according to these evaluations. The calculation model is defined as:

$$\begin{cases} LR_{ij} = \dfrac{N_{Rh}(i,j)}{N_{Rh}(i,j) + N_{Rm}(i,j) \times N_{pun}} \\ LS_{ij} = \dfrac{N_{Sh}(i,j)}{N_{Sh}(i,j) + N_{Sm}(i,j) \times N_{pun}} \end{cases} \tag{10}$$

where LR_{ij} represents the trust evaluation of the sender node V_j when the node V_i is the receiver, and LS_{ij} represents the trust evaluation of the receiver node V_j when the node V_i is the sender. The value range of LR_{ij} and LS_{ij} is $[0, 1]$, and the initial value of them are 0.5. When the value of LR_{ij} and LS_{ij} are higher than 0.5, the node is considered to be honest, and the closer the value of the node is to one, the more reliable the node is. Conversely, the node is considered not trustworthy. $N_{Rh}(i,j)$ and $N_{Rm}(i,j)$ respectively indicate the number of honest and malicious transactions with node V_j when V_i is the receiver node. $N_{Sh}(i,j)$ and $N_{Sm}(i,j)$ respectively indicate the number of honest and malicious transactions with node V_j when node V_i is the sender. N_{pun} represents the penalty coefficient, which makes reputation values fall faster than they rise, reflecting the punishment for malicious transactions.

The local trust assessment of the node can intuitively evaluate the integrity of a node. However, within a trusted mechanism, the global trust value of node i should be evaluated by all nodes that had transactions with current node, whenever the node i is act as the sender (GR_i) or the receiver (GS_i). For the evaluation content of current node, the opinions of nodes with higher integrity are more important than nodes with low integrity. Similarly, if a node has multiple times of stable connections and data transfers with current node, the evaluation between the two nodes is more credible. Therefore, the node's global trust GR_i and GS_i should be a comprehensive measurement that consider local reputation value, number of transactions and trust evaluation for all connected nodes. A weighted average approach is utilized in this work to express global receive trust GR_i and send trust GS_i as following:

$$\begin{aligned} GR_i &= \frac{\sum\limits_{j \in Vs_i} GS_i^{\omega} \times \left(1 - e^{\frac{N_S(j,i)}{5}}\right) \times LS_{ji}}{\sum\limits_{j \in Vs_i} GS_i^{\omega} \times \left(1 - e^{\frac{N_S(j,i)}{5}}\right)} \\ GS_i &= \frac{\sum\limits_{j \in Vr_i} GR_j^{\omega} \times \left(1 - e^{\frac{N_R(j,i)}{5}}\right) \times LR_{ji}}{\sum\limits_{j \in Vs_i} GR_j^{\omega} \times \left(1 - e^{\frac{N_R(j,i)}{5}}\right)} \end{aligned} \tag{11}$$

Among, $N_R(j,i) = N_{Rh}(j,i) + N_{Rm}(j,i)$ and $N_S(j,i) = N_{Sh}(j,i) + N_{Sm}(j,i)$. Vr_i represents a collection of nodes that have transactions with node i and are at the receiving end; Vs_i

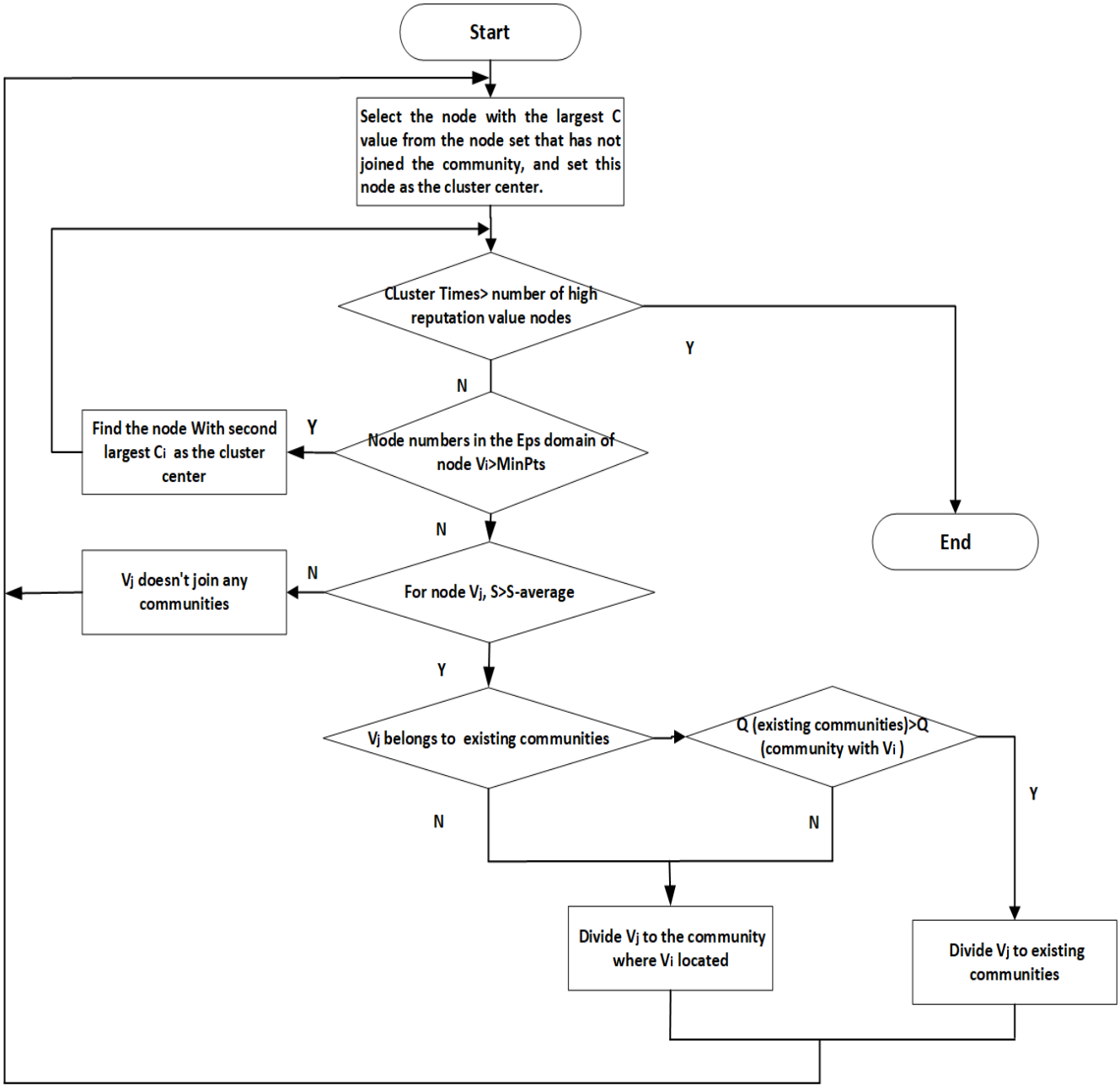


FIGURE 4. The flow-chart of the clustering process.

represents a collection of nodes that have transactions with node i and are at the sending end. The weighting function $1-\exp(-N(j,i)/5)$ has a negative exponential growth trend with the changes of $N(j,i)$, indicating that the more times of a node connected to current node stably, the more important its evaluation contents are. ω represents the weighting coefficient of node trust value in the calculation. When $\omega = 0$, the credit value is the average of the local evaluation value of participating transaction nodes. If there are many malicious nodes in the networks and there exist collusive deception, setting $\omega = 0$, which can make the trust value fairer and more justice. Experiments have shown that when the number of malicious nodes is greater than 40% of the total number of nodes, setting $\omega = 0$ can avoid conspiracy deception. But when the number of malicious nodes is less than 40% of the total number of nodes, $\omega = 1.6$ works best.

If a node passes through multiple rounds of time slice transactions, it can be considered that the global trust value of this node as the sender can be calculated based on the previous round of global trust value of the node as the receiver. This approach expressed in equation (12) reduces the number of iterations of equation (11), which can save communication

and computational overhead:

$$GR_i(k+1) = \frac{\sum_{j \in Vs_i} (GS_i(k))^{\omega} \times \left(1 - e^{\frac{Ns(j,i)}{5}}\right) \times LS_{ji}}{\sum_{j \in Vs_i} (GS_i(k))^{\omega} \times \left(1 - e^{\frac{Ns(j,i)}{5}}\right)}$$
$$GS_i(k+1) = \frac{\sum_{j \in Vr_i} (GR_i(k))^{\omega} \times \left(1 - e^{\frac{Nr(j,i)}{5}}\right) \times LR_{ji}}{\sum_{j \in Vs_i} (GR_i(k))^{\omega} \times \left(1 - e^{\frac{Nr(j,i)}{5}}\right)} \quad (12)$$

When transmitting information, node first measures the global reputation value of the next hop node which acts as the sender (receiver). If the reputation value of the node as sender and receiver are both high, the data can be sent to the node for transferring. If the node has a lower reputation value when it act as the sender, it is only used as an information storage node when the network load is large and do not forward messages.

2) NODE CACHING

The remaining cache of nodes is an important factor which is required to be considered when transmitting information. Only nodes with sufficient cache space have the ability to forward data. During a time interval t, the amount of data received by the node V_n is expressed as $r_n(t)$. The cache used for data reception is linear with the amount of data $r_n(t)$ collected, expressed as $B_S * r_n(t)$, where B_S is the cache occupied by the node collecting unit data. If the sink node allocates channels to node V_n, then node V_n sends data with cache consumption rate B_T. Therefore, the total cache B_n^{total} of node V_n in time slot t can be expressed as:

$$B_n^{total}(t) = B_S * r_n(t) + \sum_{k \in \kappa} J_{n,k}(t) B_T, \quad \forall n \in N \quad (13)$$

The amount of received data $r_n(t) \le r_{\max}$ and the number of allocated channels $\sum_{k \in \kappa} J_{n,k}(t) \le 1$, we can get the upper limit of the cache of nodes in a time interval, expressed as $B_{\max} = B_s r_{\max} + B_T$. The remaining cache of the node in time slot t is:

$$\begin{aligned} B_n^{remain}(t) &= B_{\max} - B_n^{total}(t) \\ &= B_{\max} - B_S * r_n(t) - \sum_{k \in \kappa} J_{n,k}(t) B_T, \quad \forall n \in N \end{aligned} \quad (14)$$

3) ACTIVE INDEX

In real urban scenarios, due to the fixed working hours, people's mobile mode is relatively stable in terms of time and geography. For the information transmission process, only the periods when human beings are working are useful because of the characteristic of human activities. The regularity of human activities makes the time which information can be transmitted are generally continuous. By analyzing the life of an normal human being, it is easy to find that the activity of the nodes are generally extremely low at night time. Therefore, the concept of node active index is presented which adopts the mapping time of social network information transmission nodes, rather than ordinary daily time.

In order to convert daily time into mapping time, we define the seconds of the mapping time and the daily time as $\widehat{\tau}$ and τ respectively, and they all take [0, 86400) as their value range. The difference between these two parameters is that the daily time can only be a positive integer but the mapping time can be a decimal. The average number and the total number of messages transmitted in per second can be obtained by analyzing the data set, presented as M^* and M_τ respectively. The value of M^* does not change. But the value of M_τ changes with the change of parameter τ. It is very obvious that M_τ in working periods is significantly higher than it in rest time. We define the time mapping function as equation (15):

$$\widehat{\tau} = g(\tau) = \sum_{j=0}^{\tau} M_j / M^* \quad (15)$$

For the original timestamp T of a dataset, we define its corresponding mapping timestamp as $\widehat{T}$. In order to convert the original timestamp into a mapping timestamp, we firstly need to convert T to τ_* by the following formula:

$$\tau_* = \mathrm{mod}(T + N_{zone} * N_{\mathrm{sec}\,onds}, Max_\tau) \quad (16)$$

where mod (a, b) returns the modulus after division of a by b. $N_{\mathrm{sec}\,onds}$ is the number of seconds in an hour. The time zone is represented as N_{zone}, and Max_τ is the maximum value of τ. The time zone is related to the location's district of nodes, so we should add $N_{zone} * N_{\mathrm{sec}\,onds}$ for each T when calculating. After integrating Equations (15) and (16), the mapping time is expressed as:

$$\widehat{T}_i = T_i - \tau_* + \widehat{\tau}_* = T_i - \tau_* + g(\tau_*) \quad (17)$$

In the opportunistic social networks, the movement of nodes brings communication opportunities among devices. We define the active index as the sum distances of the node in the mapping time period. It is possible to determine whether node V_i is active according to the active index of node, as shown in equation (18). v represents the average speed of node i according to history records.

$$Dist_i = v * \widehat{T}_i = v * [T_i - \tau_* + g(\tau_*)] \quad (18)$$

4) REDUCTION STRATEGY FOR NODES WITHIN THE COMMUNITY

For each community in networks, we analyzed the various social attributes of the nodes and measure the impact of various social attributes on the information transmission. By weighting and combining these attributes, we can get a comprehensive index which can be used to evaluated nodes in communities and decreased the number of inefficient nodes. In this paper, we use the concept of information entropy to assign weights to each attributes. Information entropy is a variable that can describe the degree of disorder of information. The larger the entropy value is, the higher the degree of information disorder will be. Therefore, through the method

of information entropy, we can measure the entropy value of each feature and assign appropriate weights to these attributes according to their relevance in information transmission. Information entropy is defined as:

$$E(F_i) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i), \quad (i = 1, 2, \ldots, n) \tag{19}$$

where $E(F_i)$ denotes the entropy of each feature F_i, and $p(x_i)$ denotes the probability function of F_i.

According to the nature of weight, it can be known that function E needs to have the characteristic of symmetry, monotony, continuity and additivity. When using entropy for weight analysis, the order of each feature does not change the weight of them. The function also has the continuity and monotonous change when the number of features is determined. We construct the function $E(F_i)$ based on these principles above:

$$E(x_1 \ldots x_n) = -c\sum_{i=1}^{n} x_i \ln x_i \tag{20}$$

Obviously, function $E(F_i)$ has the characteristic of symmetry, monotonicity and continuity. Next the additivity of function $E(F_i)$ will be proved.

Theorem 1: For function $E(x_1 \ldots x_n) = -c\sum_{i=1}^{n} x_i \ln x_i$, if $x_n = y_1 + y_2$,then $E(x_1 \ldots x_{n-1}, y_1, y_2) = E(x_1 \ldots x_n) + x_n E(\frac{y_1}{x_n}, \frac{y_2}{x_n})$.

Proof 1:

$$\begin{aligned}
&E(x_1 \ldots x_{n-1}, y_1, y_2)\\
&= -c(\sum_{i=1}^{n-1} x_i \ln x_i + y_1 \ln y_1 + y_2 \ln y_2)\\
&= -c\sum_{i=1}^{n-1} x_i \ln x_i - cx_n \ln x_n + cx_n(\ln x_n - \frac{y_1}{x_n}\ln y_1\\
&\quad - \frac{y_2}{x_n}\ln y_2)\\
&= -c\sum_{i=1}^{n} x_i \ln x_i + cx_n(\ln x_n - \frac{y_1}{x_n}\ln y_1 - \frac{y_2}{x_n}\ln y_2)\\
&= E(x_1 \ldots x_n) + cx_n(\ln x_n - \frac{y_1}{x_n}\ln y_1 - \frac{y_2}{x_n}\ln y_2)\\
&= E(x_1 \ldots x_n) + c(y_1 + y_2)\begin{bmatrix} \ln(y_1 + y_2) \\ -\frac{y_1}{(y_1+y_2)}\left(\ln\frac{y_1}{x_n} + \ln x_n\right) \\ -\frac{y_2}{(y_1+y_2)}\left(\ln\frac{y_2}{x_n} + \ln x_n\right)\end{bmatrix}\\
&= E(x_1 \ldots x_n) + c(y_1 + y_2)\ln(y_1 + y_2)\\
&\quad - c\left(y_1 \ln\frac{y_1}{x_n} + y_2 \ln\frac{y_2}{x_n}\right) - cx_n \ln x_n\\
&= E(x_1 \ldots x_n) + cx_n \ln x_n - cx_n\begin{pmatrix}\frac{y_1}{x_n}\ln\frac{y_1}{x_n}\\ +\frac{y_2}{x_n}\ln\frac{y_2}{x_n}\end{pmatrix} - cx_n \ln x_n\\
&= E(x_1 \ldots x_n) + x_n E(\frac{y_1}{x_n}, \frac{y_2}{x_n})
\end{aligned}$$

We analyze receive-trust (GR_i), send-trust (GS_i), the remaining cache of node ($B_i^{remain}(t)$) and the active index ($Dist_i$) in this paper, and then establishing an evaluation matrix for these features:

$$X = \begin{matrix} R-trust & S-Trust & N-caching & A-index \end{matrix}\\ \begin{bmatrix} GR_1 & GS_1 & B_1^{remain}(t) & Dist_1 \\ GR_2 & GS_2 & B_2^{remain}(t) & Dist_2 \\ \ldots & \ldots & \ldots & \ldots \\ GR_{n-1} & GS_{n-1} & B_{n-1}^{remain}(t) & Dist_{n-1} \\ GR_n & GS_n & B_n^{remain}(t) & Dist_n \end{bmatrix} \tag{21}$$

According to equation (22), we can normalize the data matrix X to get the calculation matrix Y, where $\max x_{ij}$, $\min x_{ij}$ and $\bar{x}_{ij}$ represent the maximum, minimum and average values of the $j-th$ column elements in the matrix X, respectively.

$$y_{ij} = \frac{x_{ij} - \bar{x}_{ij}}{\max x_{ij} - \min x_{ij}} \tag{22}$$

Combining with formula (20), we can calculate the corresponding entropy value of each feature index. The entropy function is taken a negative sign to make sure that the value of it always stay positive. The normalization coefficient is defined as $c = \frac{1}{\ln n}$.

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} a_{ij} \ln a_{ij} \tag{23}$$

For each feature index, we can get the corresponding weight as:

$$W_j = \frac{1 - E_j}{n - \sum_{j=1}^{m} E_j} \tag{24}$$

For all nodes in communities, we can filter them according to these characteristics of nodes. The comprehensive feature $\kappa_i = W_1 * GR_i + W_2 * GS_i + W_3 * B_i^{remain}(t) + W_4 * Dist_i$. Thus, we set a threshold κ to determine whether node i can meet the needs of the relay nodes and then delete nodes that are not available. Through this scheme of reducing nodes within the community, we can filter the nodes and delete some of which do not meet the requirements of transmission in the community. After reducing a small number of nodes that do not meet the transmission conditions, the nodes in communities are closely connected and all of them has transmission capacity. After many experiments, we can get the best performance when the threshold $\kappa = 23.265$.

C. COMMUNITY-BASED TRANSMISSION SCHEME

As described above, we can get several communities with close social connections in the network. The nodes in these communities have high credibility, activity, and sufficient cache space. The ETNS algorithm can effectively improve the successful ratio of data delivery,and reduce the routing cost greatly by taking full advantage of the close relationship

among nodes in communities. In this paper, data transmission occurs only in the cluster community where the target node is located.

Suppose that node V_i encounters node V_j, and node V_i has the message to be forward. If node V_j is the target node, node V_i will transmit the information to node V_j and delete the message from its sending queue. If the encountering node V_j is not the target node, the transmission mode of node V_i can be divided into the following two types.

Case 1 (Intra-Community Message Transmission):

If the target node is in the community where node V_i is allocated, and node V_j is also a member of this community, node V_i will transmits a copy of messages to node V_j. Otherwise, node V_i does not transmit messages to node V_j.

Case 2 (Inter-Community Message Transmission):

If the target node is not in the community which node V_i belongs to, node V_i will send a ''request frame'' to node V_j to ask whether the target node is in the community which node V_j belongs to. After receiving the ''request frame'', node V_j will check the current community, confirming whether the target node is in the current community, and sending a ''response frame'' to node V_i. If the target node is in the community where V_j is located, node V_i will send the messages to node V_j. Otherwise, node V_i does not forward the message to node V_j.

The messages that need to be sent are stored in the cache space of nodes for a period of time before node V_i meets the node in the community where the target node is located. If node V_i encounters a node that meets the above rules, the messages will be forwarded, otherwise, the data in buffer will be dropped according to the cache management rules. There exists a possibility that messages can not be transmitted if node V_i meets none of the nodes in the community where the target node is located, and enough buffer capacity will be helpful for message delivery.

IV. COMPLEXITY ANALYSIS OF SYSTEM MODEL

In summary, an effective data transmission strategy based on node socialization is proposed in this paper. ETNS algorithm is applied in many real life scenarios, such as vehicle-to-vehicle communication. In the Internet of vehicles, vehicles can be viewed as several mobile nodes, which can also be divided into several communities. Four attribute indexes proposed here can be used to measure the communication quality of each vehicle. After comprehensive consideration, the vehicles in each community that do not meet the transmission conditions can be removed. The compact community structure that eliminates inefficient vehicles can provide better performance in message transmission. Meanwhile, in order to make the structure of the whole algorithm clear to understand, specific steps of the algorithm are listed as follows:

STEP 1: Nodes in the opportunistic social networks can be divided into several communities according to the gathering characteristics of people in social life. The clustering method is utilized to divide the nodes in networks into several community structures.

Algorithm 1 An Data Transmission Strategy Based on Node Socialization

Input: A graph $G(V, E)$, a source node S, a destination node D;

Output: optimal path and effective communities;

```
Begin
//Clustering-based community division method
For each node
 Judge the condition of node V_i;
   If (node V_i is suitable for dividing into community C_j)
      Divide node V_i into community C_j;
   End If
End For
//Community reduction strategy
For each community
   Calculate GR_i, GS_i, B_n^remain(t) and Dist_i of node V_i;
   Get the weight W_j of each feature;
   Calculate the comprehensive feature κ_i;
   Compare the κ_i with threshold κ;
      If (κ_i < κ)
       Delete V_i from community;
      End If
End For
//Effective transmission process
Forwarding messages from S;
   If D is in the community S allocated
      Forwarding messages to the community S allocated;
   Else
      Send ''request frame'' to encountered node V_j until
find D;
   End If
End
```

STEP 2: The transmission process of information has some requirements for relay nodes. The concept of sending trust, receiving trust, remain cache, and active index is presented in this paper to measure the availability of nodes. In opportunistic social networks, if nodes satisfy these characteristics at the same time, it is more likely that they are optimal relay nodes.

STEP 3: According to the entropy value of each feature, multiple features are considered comprehensively in this scheme to determine the availability of nodes, which can reduce inefficient nodes in the community effectively.

STEP 4: Data packets are transmitted to the destination node through efficient communities, which is benefit to keep the continually, stability and high efficiency of data transmission process.

Figure 5 and algorithm 1 is constructed to introduce the ETNS algorithm in details for better readability. Specifically, during the clustering process, cluster center compares similarity with neighbor nodes according to the sequence, so the time complexity of this phase is $O(\log_2 n)$. Four features of optimal nodes is analysed in this work based on the theory of information entropy. Inefficient nodes in communi-

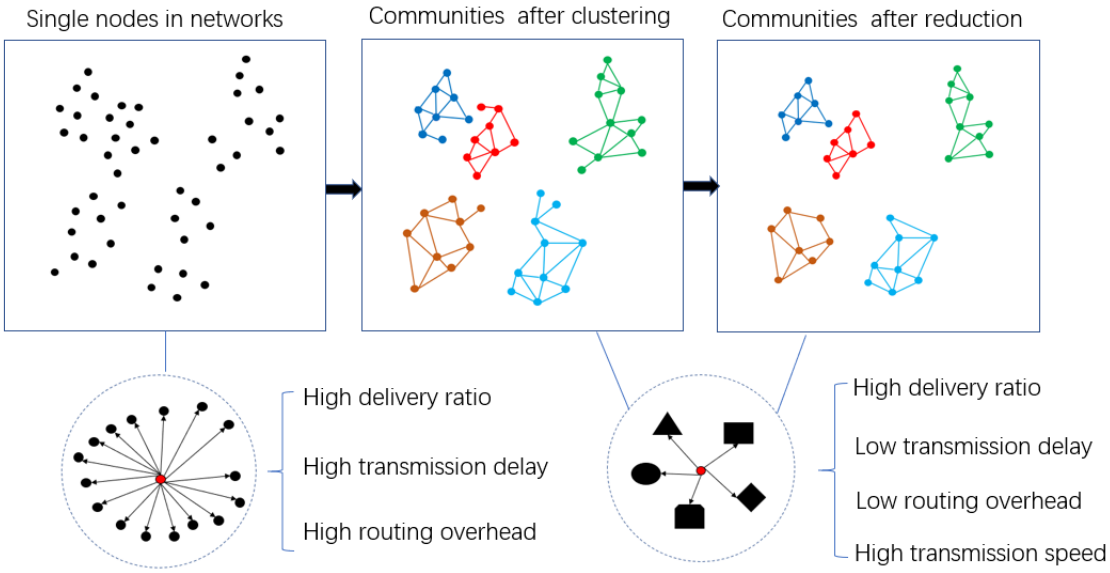


FIGURE 5. This work divides the nodes in the network into several communities, and removes some inefficient nodes to improve the performance of algorithms. Compared with the original algorithm of single nodes, it not only improves the transmission success rate and the transmission speed, but also reduces the transmission delay and routing cost.

TABLE 2. Characteristics of the four experimental data sets.

Dataset	Infocom5	Infocom6	Cambridge	Intel
Device	iMote	iMote	iMote	iMote
Duration(days)	3.5	4	11.5	4
Number of experimental devices	41	98	52	9
Number of internal contacts iMote	22459	170601	10873	1364

TABLE 3. Simulation parameters of four experimental datasets in ONE.

Dataset	Infocom5	Infocom6	Cambridge	Intel
Number of nodes	41	98	52	9
Buffer size	5M	5M	5M	5M
TTL	60min	60min	2 days	0.5 days

ties are reduced according to these features, thus, the time complexity of this process is $O(n)$. Finally, the time complexity of transmission process is $O(n)$ because packet is forwarded to effective communities. On the whole, through rigorous mathematical analysis, the time complexity of the ETNS algorithm can be computed as $O(\log_2 n + n + n) = O(n)$. In retrospect, the time complexity of EWDCR is $O(n^2)$ and the time complexity in the Epidemic routing algorithm is $O(n)$.

V. SIMULATION AND ANALYSIS

The simulation adopts the opportunistic network environment (ONE) to evaluate ETNS by performance comparison with ESR [52] (effective algorithm based on social relationships), FCNS [48] (fuzzy routing-forwarding algorithm), EWDCR [42] (Effective weight distribution and communities reconstitution algorithm) and epidemic algorithm [33]. ESR, FCNS and EWDCR are the latest routing algorithms for opportunistic social networks, while epidemic algorithm is a typical and traditional method. At present, some real datasets can be used for simulation experiments, but the algorithm proposed in this paper requires a lot of data information, and some data sets cannot meet the requirements. After considering the data information required by the algorithm proposed in this paper, the real datasets Infocom 5, Infocom 6, Cambridge and Intel is most suitable to drive node activity. The datasets can be download from CRAWDAD, and the detailed information is shown in table 2. In this experiment, the cache size of node is set to 5M, and the node number and TTL are different to the original four datasets, thus the changes are shown in table 3.

The ETNS routing algorithm is compared with the above routings in the same simulation environment to evaluate its performance. The simulation results mainly concentrate on the following metrics [53]:

1) **Delivery ratio**: This parameter refers to the probability of selecting a relay node during the transmission process within a certain time interval.

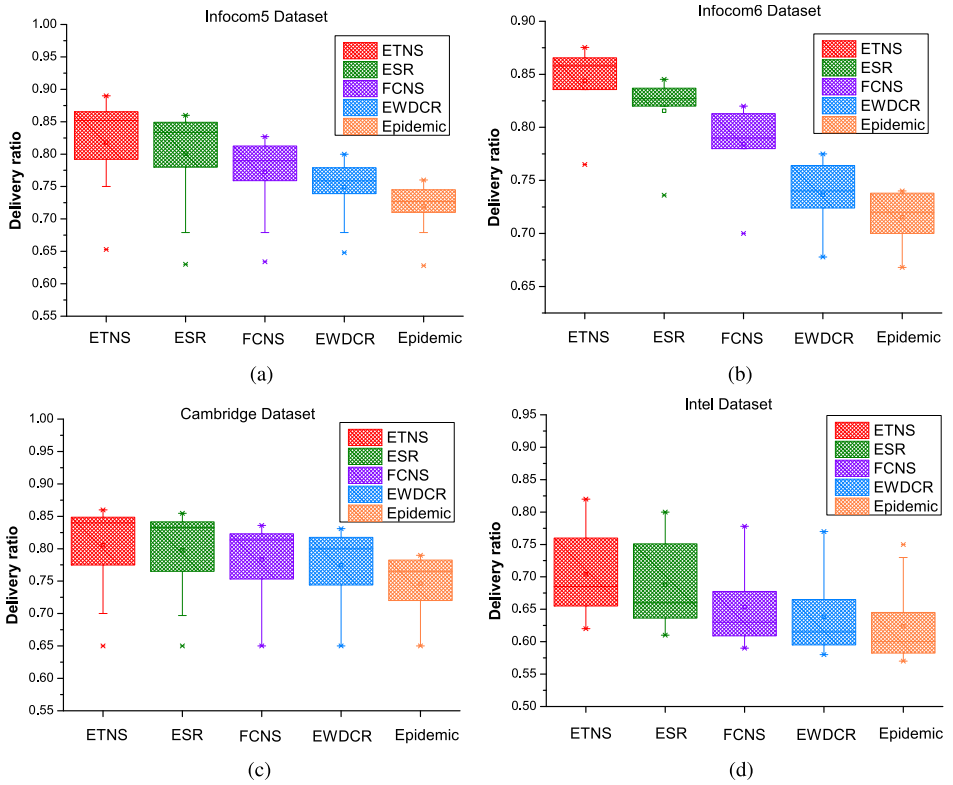


FIGURE 6. Quartiles of packet delivery ratio.

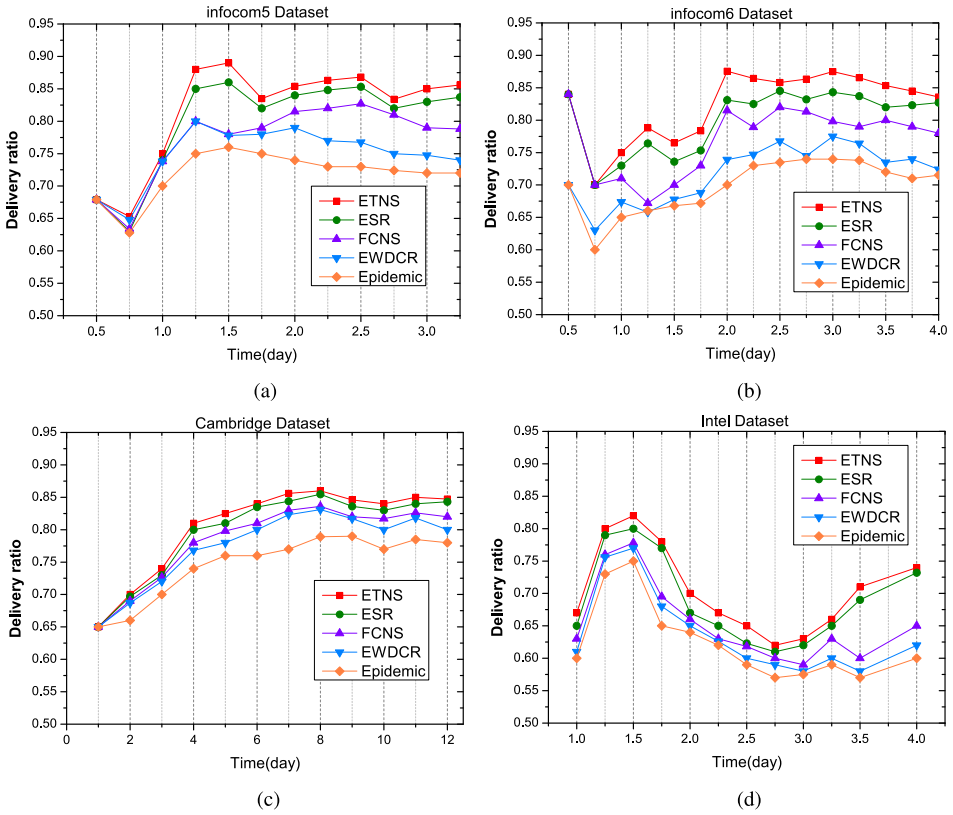


FIGURE 7. Packet delivery ratio comparisons.

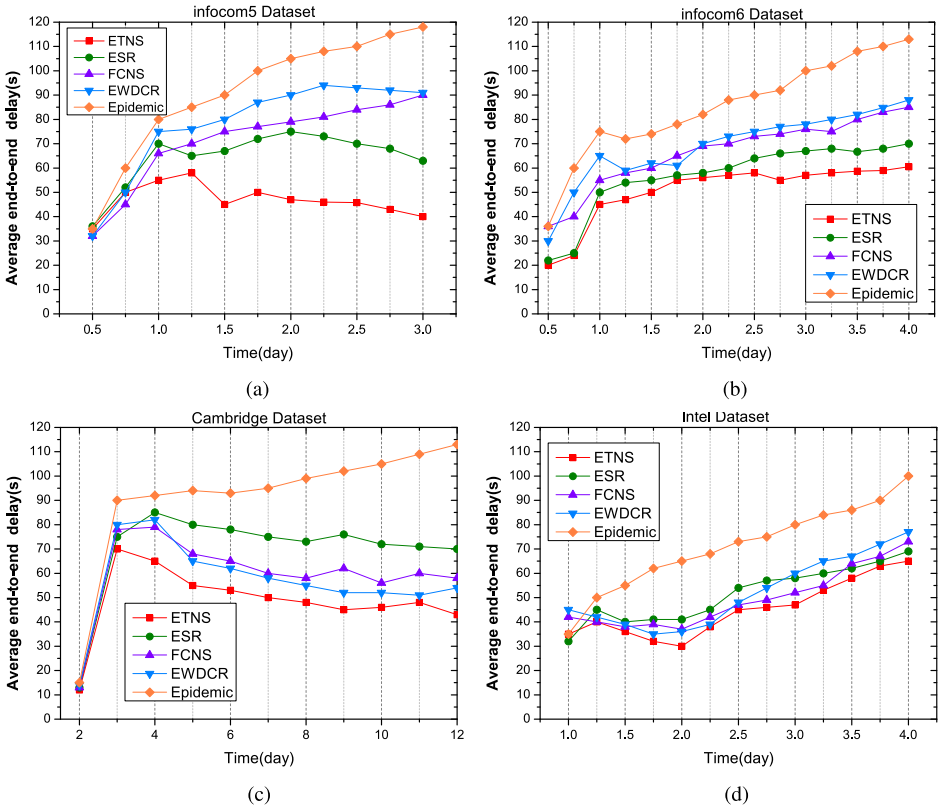


FIGURE 8. Average end-to-end delay comparisons.

2) **Average end-to-end delay**: This parameter comprehensively evaluates the delay of routing selections, relay nodes waiting for messages, and message forwarding.
3) **Routing overhead**: This parameter shows the overhead between a pair of nodes when message is transmitted. Routing overhead can be shown as equation (25), where N_{total_time} is the total number of transmission time and $N_{succeed\_time}$ is the number of succeed transmission time.

$$O_{node} = \frac{N_{total\_time} - N_{Succeed\_time}}{N_{total\_time}} \tag{25}$$

A. THE PERFORMANCE OF COMMUNITY DIVISION

From the simulation, we can get the result that ETNS algorithm has good performance in community division process. In the four data sets used in the experiment, the performance of the algorithm in Infocom6 dataset is the best. This is the only dataset that gets actual 6 clusters, which is closest to the real result of human beings. The community division results in Infocom5 and Cambridge datasets are also good, but slightly worse than the algorithms performance in Infocom6 dataset. The performance in Intel dataset is significantly worse than the other three datasets, because the number of experimental copies in this dataset is small or nodes in this group is random located, which does not conform to the characteristics of human clustering. However, the results of simulation in these datasets are all good, which proves that the process of community division in real datasets by ETNS algorithm is feasible and effective.

B. DELIVERY RATIO

In the simulation, ETNS, ESR, FCNS, EWDCR and Epidemic algorithm run in the four datasets respectively, and the simulation environment is shown in table 2. We evaluate the delivery ratio of each algorithm, and the comparison results between these five algorithms are shown in figure 6. We use quartiles to analyse the experimental results. There are five signs (min, first quartile, median, third quartile and max) for the result of each algorithm. In figure 6, the quartiles can express the distribution center, concentration and spread range of delivery ratio. Experimental results show that ETNS has a higher distribution center of delivery ratio than the other algorithms, with a small spread range and a concentrate better range.

Figure 7 shows the packet delivery ratio of the ETNS, ESR, FCNS, EWDCR and epidemic algorithm with the simulation time varying. When simulation time is less than one day, the advantage of algorithm ETNS is not obvious, and the performance of ETNS is similar to the other four algorithms. As the simulation time increases, we can find that the delivery ratio of ETNS is always the highest among these algorithms, because successful data transmission can be implemented

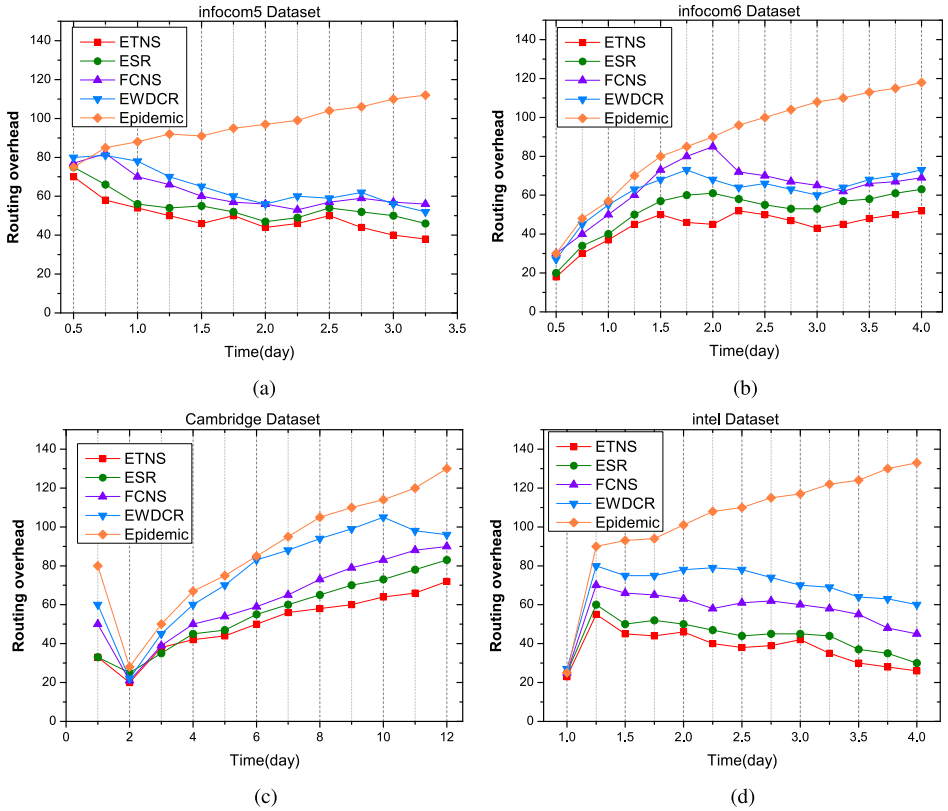


FIGURE 9. Routing overhead comparisons.

by filtering effective nodes in communities. In the ETNS algorithm, nodes in the network are divided into several communities, and each pair of nodes in the community with high similarity may frequently communicate with each others. Meanwhile, ETNS algorithm proposes a reduction strategy of nodes based on multi attributes, which can decrease large number of inappropriate and inefficient nodes, so the high availability of nodes in communities could bring the highest delivery ratio. ESR algorithm is a community based routing algorithm but its reduction method of nodes does not consider social attributes, so ETNS has better performance. As for the FCNS and EWDCR algorithms, similarity degree do not consider the credibility and available room of nodes, which will cause the unavailability of selected relay nodes. Moreover, the epidemic algorithm has a large number of message copies, thus the delivery ratio of the flooding method is relatively low compared with other four algorithms.

C. AVERAGE END-TO-END DELAY

The average end-to-end delay of each algorithm is shown in figure 8. ETNS has the lowest average end-to-end delay compared with the other four algorithms. Since ETNS propose a community reduced strategy by analysing the comprehensive characteristic of nodes, it can reduce inefficient nodes that are not helpful for the transmission process, and thereby reduce the average end-to-end delay. Comparatively speaking, the epidemic algorithm has no requirements for the next hop node, and transmits messages blindly, which causes a sharp increase in the routing and forwarding delays. The ESR algorithm effectively limits the number of copies, so the transmission delay is lower than epidemic algorithm. Besides, the FCNS algorithm analyses the transmission references before data transmission. Data deliver through neighbors and relevance nodes in EWDCR algorithm. Therefore, the average end-to-end delay in FCNS and EWDCR algorithms are lower than traditional routing algorithms. Generally, the average end-to-end delay of ETNS is optimal among these five algorithms.

D. ROUTING OVERHEAD

The comparisons of routing overhead among these five different algorithms are demonstrated in figure 9. The average overhead in the ETNS algorithm always keeps the lowest, because it adopts a community-aware strategy which considered comprehensive characteristics in transmission. In ETNS algorithm, nodes are divided into several close-knit communities, and there is high probability of successful transmission among nodes. Therefore, the routing scheme of ETNS take a small amount of time and resources, which sharply decreasing the overhead on average. The ESR algorithm only considers the influence of nodes on information flow and ignores the current availability of the next hop node,

which causes waiting overhead. In the epidemic algorithm, redundant message copies need mass of time and resources, which is the main reason of huge routing overhead. As for the FCNS and EWDCR algorithms, the routing overhead can be effectively reduced by the similarity of nodes, but the routing overhead is still can be optimized because the consumption of resources by some unavailable nodes can be decreased. In conclusion, the ETNS always performs best among these five algorithms in terms of routing overhead.

VI. CONCLUSION

An effective data transmission strategy based on node socialization is proposed in this work. The algorithm is based on the fact that nodes within the community have higher tightness than those outside. But not all nodes in communities have the necessary conditions for the forwarding process, such as the degree of trust, the remaining cache of nodes, and the active index of nodes. Unlike other algorithms that only simply divide the community for data transmission, an effective reduction strategy of nodes is proposed in this strategy. This reduction method analyzes various attributes of the optimal relay node and combines these attributes to evaluate and remove inefficient nodes. Data packets are transmitted to the destination node through efficient communities, which is benefit to keep the continually, stability and high efficiency of data transmission process. As long as the computing capacity and cache spaces of mobile devices in opportunistic social networks will be further improved, the ETNS algorithm can be applied to the transmission environment of 5G and big data networks. In future works, we will improve the performance of our algorithm and further study security and privacy problems in opportunistic social networks.

REFERENCES

[1] S. Liu, L. Zhang, and Z. Yan, "Predict pairwise trust based on machine learning in online social networks: A survey," *IEEE Access*, vol. 6, pp. 51297–51318, 2018. doi: 10.1109/ACCESS.2018.2869699.

[2] C. Montag *et al.*, "Facebook usage on smartphones and gray matter volume of the nucleus accumbens," *Behav. Brain Res.*, vol. 329, pp. 221–228, Jun. 2017.

[3] P. Giridhar, S. Wang, and T. Abdelzaher, R. Ganti, L. Kaplan, and J. George, "On localizing urban events with instagram," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[4] Y. Abdelsadek, K. Chelghoum, F. Herrmann, I. Kacem, and B. Otjacques, "Community extraction and visualization in social networks applied to Twitter," *Inf. Sci.*, vol. 424, pp. 204–223, Jan. 2017.

[5] X. Lai and H. Wang, "RNOB: Receiver negotiation opportunity broadcast protocol for trustworthy data dissemination in wireless sensor networks," *IEEE Access*, vol. 6, pp. 53235–53242, 2018. doi: 10.1109/ACCESS.2018.2871082.

[6] L. Chancay-García, E. Hernández-Orallo, P. Manzoni, C. T. Calafate, and J. Cano, "Evaluating and enhancing information dissemination in urban areas of interest using opportunistic networks," *IEEE Access*, vol. 6, pp. 32514–32531, 2018. doi: 10.1109/ACCESS.2018.2846201.

[7] A. Thakur, R. Sathiyanarayanan, and C. Hota, "STEEP: Speed and time-based energy efficient neighbor discovery in opportunistic networks," *Wireless Netw.*, no. 2, pp. 1–22, 2018.

[8] Y. He *et al.*, "Deep reinforcement learning-based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10433–10445, Nov. 2017.

[9] M. Emara, H. E. Sawy, S. Sorour, S. Al-Ghadhban, and M.-S. Alouini, "Optimal caching in 5G networks with opportunistic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4447–4461, Jul. 2018.

[10] Y. Deng, Z. Chen, D. Zhang, and M. Zhao, "Workload scheduling toward worst-case delay and optimal utility for single-hop fog-IoT architecture," *IET Commun.*, vol. 12, no. 17, pp. 2164–2173, Jul. 2018. doi: 10.1049/iet-com.2018.5077.

[11] Y. Yang, H. Zhao, J. Ma, and X. Han, "Social-aware data dissemination in opportunistic mobile social networks," *Int. J. Mod. Phys. C*, vol. 28, no. 9, 2017, Art. no. 1750115.

[12] D. Liu, K. Chen, Y. Chou, and J. Lee, "An online activity recommendation approach based on the dynamic adjustment of recommendation lists," in *Proc. 6th IIAI Int. Congr. Adv. Appl. Inform. (IIAI-AAI)*, 2017, pp. 407–412. doi: 10.1109/IIAI-AAI.2017.60.

[13] H.-L. Chiu and S.-H. Wu, "Cross-layer performance analysis of cooperative ARQ with opportunistic multi-point relaying in mobile networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4191–4205, Jun. 2018.

[14] B. Chen, C. Yang, and G. Wang, "High throughput opportunistic cooperative device-to-device communications with caching," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7527–7539, Aug. 2017.

[15] P. Ostovari, J. Wu, and A. Khreishah, "Cooperative Internet access using helper nodes and opportunistic scheduling," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6439–6448, Jul. 2017.

[16] J. Wu, Z. Chen, and M. Zhao, "Effective information transmission based on socialization nodes in opportunistic networks," *Comput. Netw.*, vol. 129, pp. 297–305, Dec. 2017. doi: 10.1016/j.comnet.2017.10.005.

[17] J. Wu, Z. Chen, and M. Zhao, "Information cache management and data transmission algorithm in opportunistic social networks," *Wireless Netw.*, Feb. 2018, pp. 1–18.

[18] N. Wang and J. Wu, "Optimal data partitioning and forwarding in opportunistic mobile networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (IEEE WCNC)*, Apr. 2018, pp. 1–6.

[19] H. Zhou, J. Wu, H. Zhao, S. Tang, C. Chen, and J. Chen, "Incentive-driven and freshness-aware content dissemination in selfish opportunistic mobile networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 9, pp. 2493–2505, Sep. 2015.

[20] E. K. Wang, Y. Li, Y. Ye, S.-M. Yiu, and L. C. K. Hui, "A dynamic trust framework for opportunistic mobile social networks," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 319–329, Mar. 2018.

[21] R. Wang *et al.*, "Social identity–aware opportunistic routing in mobile social networks," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 5, p. e3297, 2018.

[22] K. Zhu, W. Li, X. Fu, and Z. Lin, "VIRO: A virtual routing method for eliminating dead end in opportunistic mobile social network," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 3234–3239.

[23] F. Zeng, N. Zhao, and W. Li, "Effective social relationship measurement and cluster based routing in mobile opportunistic networks," *Sensors*, vol. 17, no. 5, p. 1109, 2017.

[24] M. A. Alim, X. Li, N. P. Nguyen, M. T. Thai, and A. Helal, "Structural vulnerability assessment of community-based routing in opportunistic networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 12, pp. 3156–3170, Dec. 2016.

[25] K. Ch'ng, N. Vazquez, and E. Khatami, "Unsupervised machine learning account of magnetic transitions in the Hubbard model," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 1, 2018, Art. no. 013306.

[26] E. López-Rubio, E. J. Palomo, and F. Ortega-Zamorano, "Unsupervised learning by cluster quality optimization," *Inf. Sci.*, vol. 436, pp. 31–55, Apr. 2018.

[27] C. Miller, Z. Nagy, and A. Schlueter, "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings," *Renew. Sustain. Energy Rev.*, vol. 81, no. 1, pp. 1365–1377, 2017.

[28] S. Mohamad, D. Arifoglu, and C. Mansouri, "Deep online hierarchical unsupervised learning for pattern mining from utility usage data," in *Proc. UK Workshop Comput. Intell.* Cham, Switzerland: Springer, 2018, pp. 276–290.

[29] N. P. Desai, C. Lehman, B. Munson, and M. Wilson, "Supervised and unsupervised machine learning approaches to classifying chimpanzee vocalizations," *J. Acoust. Soc. Amer.*, vol. 143, no. 3, p. 1786, 2018.

[30] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018.

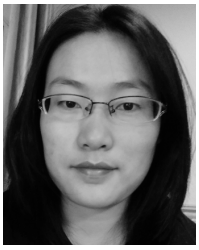
[31] Y. Xu, Z. Zhuang, W. Li, and X. Zhou, "Effective community division based on improved spectral clustering," *Neurocomputing*, vol. 279, pp. 54–62, Mar. 2018. doi: 10.1016/j.neucom.2017.06.085.

[32] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE simulator for DTN protocol evaluation," in *Proc. Int. Conf. Simulation Tools Techn.*, 2009, p. 55.
[33] M. Chitra and S. S. Sathya, "Selective epidemic broadcast algorithm to suppress broadcast storm in vehicular ad hoc networks," *Egyptian Inform. J.*, vol. 19, no. 1, pp. 1–9, 2017.
[34] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002.
[35] S. Sisodiya, P. Sharma, and S. K. Tiwari, "A new modified spray and wait routing algorithm for heterogeneous delay tolerant network," in *Proc. Int. Conf. I-SMAC*, Feb. 2017, pp. 843–848.
[36] S. Jain, M. Chawla, V. N. G. J. Soares, and J. J. Rodrigues, "Enhanced fuzzy logic-based spray and wait routing protocol for delay tolerant networks," *Int. J. Commun. Syst.*, vol. 29, no. 12, pp. 1820–1843, 2016.
[37] S. J. Borah, S. K. Dhurandher, S. Tibarewala, I. Woungang, and M. S. Obaidat, "Energy-efficient prophet-PRoWait-EDR protocols for opportunistic networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.
[38] S. K. Dhurandher, S. J. Borah, I. Woungang, A. Bansal, and A. Gupta, "A location prediction-based routing scheme for opportunistic networks in an IoT scenario," *J. Parallel Distrib. Comput.*, vol. 118, pp. 369–378, Aug. 2018.
[39] D. K. Sharma, S. K. Dhurandher, I. Woungang, A. Mohananey, and J. J. Rodrigues, "A machine learning-based protocol for efficient routing in opportunistic networks," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2207–2213, Sep. 2018.
[40] N. Li, J. F. Martínez-Ortega, and V. H. Diaz, "Cross-layer and reliable opportunistic routing algorithm for mobile ad hoc networks," *IEEE Sensors J.*, vol. 18, no. 3, pp. 5595–5609, Jul. 2018.
[41] N. Wang and J. Wu, "Rethink data dissemination in opportunistic mobile networks with mutually exclusive requirement," *J. Parallel Distrib. Comput.*, vol. 119, pp. 50–63, Sep. 2018.
[42] J. Wu, Z. Chen, and M. Zhao, "Weight distribution and community reconstitution based on communities communications in social opportunistic networks," *Peer-Peer Netw. Appl.*, vol. 12, no. 1, pp. 158–166, 2018. doi: 10.1007/s12083-018-0649-x.
[43] X. Fu, W. Li, and G. Fortino, "A utility-oriented routing algorithm for community based opportunistic networks," in *Proc. IEEE Int. Conf. Comput. Supported Cooperat. Work Design*, Jun. 2013, pp. 675–680.
[44] R. Dragan, R. I. Ciobanu, and C. Dobre, "Leader election in opportunistic networks," in *Proc. Int. Symp. Parallel Distrib. Comput.*, Jul. 2017, pp. 157–164.
[45] X. Chen, C. Shang, B. Wong, W. Li, and S. Oh, "Efficient multicast algorithms in opportunistic mobile social networks using community and social features," *Comput. Netw.*, vol. 111, pp. 71–81, Dec. 2016.
[46] H. Zhang, L. Wan, Y. Chen, L. T. Yang, and L. Peng, "Adaptive message routing and replication in mobile opportunistic networks for connected communities," *ACM Trans. Internet Technol.*, vol. 18, no. 1, pp. 1–22, 2017.
[47] J. Tao, H. Wu, S. Shi, J. Hu, and Y. Gao, "Contacts-aware opportunistic forwarding in mobile social networks: A community perspective," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Barcelona, Spain, Apr. 2018, pp. 1–6. doi: 10.1109/WCNC.2018.8377216.
[48] 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, p. 338, 2018. doi: 10.3390/sym10080338.
[49] T. Wang, Y. Zhou, X. Wang, and Y. Cao, "A social-based DTN routing in cooperative vehicular sensor networks," *Int. J. Cooperat. Inf. Syst.*, vol. 27, no. 1, 2018, Art. no. 1741003.
[50] L. Wang, Z. Chen, and J. Wu, "Vehicle trajectory prediction algorithm in vehicular network," *Wireless Netw.*, pp. 1–14, Jul. 2018. doi: 10.1007/s11276-018-1803-3.
[51] D. Mehrle, A. Strosser, and A. Harkin, "Walk-modularity and community structure in networks," *Netw. Sci.*, vol. 3, no. 3, pp. 348–360, 2015.
[52] Y. Yan, Z. Chen, J. Wu, and L. Wang, "An effective data transmission algorithm based on social relationships in opportunistic mobile social networks," *Algorithms*, vol. 11, no. 8, p. 125, 2018.
[53] J. Wu and Z. Chen, "Sensor communication area and node extend routing algorithm in opportunistic networks," *Peer-Peer Netw. Appl.*, vol. 11, no. 1, pp. 90–100, Jan. 2018. doi: 10.1007/s12083-016-0526-4.

YEQING YAN is currently pursuing the master's degree with the School of Software, Central South University. She is also a Researcher with the "Mobile Health" Ministry of Education-China Mobile Joint Laboratory. Her main research interests include communications and networking, complex networks, and opportunistic networks.

ZHIGANG CHEN was born in 1964. He received the B.E., M.S., and Ph.D. degrees from Central South University, in China, in 1984, 1987, and 1998, respectively, where he is currently a Professor, a Supervisor of Ph.D., and the Dean of the School of Software. He is also the Director and an Advanced Member of the China Computer Federation (CCF), and a member of the Pervasive Computing Committee of CCF. His research interests include cluster computing, parallel and distributed systems, computer security, and wireless networks.

JIA WU (M'–) was born in 1983. He received the Ph.D. degree in software engineering from Central South University, Changsha, Hunan, China, in 2016, where he is currently a Special Associate Professor. He is also an Engineer with the "Mobile Health" Ministry of Education-China Mobile Joint Laboratory and a Postdoctor with the School of Information Science and Engineering, Central South University. Since 2010, he has been an Algorithm Engineer with IBM Company, in Seoul, South Korea and in Shanghai, China. His research interests include wireless communications and networking, wireless networks, big data research, and mobile health in network communication. He is a Senior Member of China Computer Federation (CCF), and a member of ACM.



LEILEI WANG is currently pursuing the master's degree with the School of Software, Central South University. She is also a Researcher with the "Mobile Health" Ministry of Education-China Mobile Joint Laboratory. Her research interests include opportunistic networks and vehicular networks.

KANGHUAI LIU is currently pursuing the master's degree with the School of Software, Central South University. He is also a Researcher with the "Mobile Health" Ministry of Education-China Mobile Joint Laboratory. His research interests include wireless communications and networking, wireless networks, opportunistic networks, medical decision-making, big data, machine learning, deep learning, and data mining.

YUZHOU WU is currently pursuing the Ph.D. degree with the School of software, Central South University. She is also a Researcher with the "Mobile Health" Ministry of Education-China Mobile Joint Laboratory. Her works focus specifically on machine learning, medical image analysis, and opportunistic networks.