Soft Computing (2019) 23:6267–6285
https://doi.org/10.1007/s00500-018-3283-x

METHODOLOGIES AND APPLICATION

Social community detection and message propagation scheme based on personal willingness in social network

Ke Gu[1,2] · Linyu Wang[1] · Bo Yin[1]

Published online: 5 June 2018

Abstract
Personal willingness is one of the most important factors influencing the construction of social community and the message propagation in social network. Personal willingness is used to describe the subjective initiative of node (user) to communicate information with outside world. The personal willingness is greater, the corresponding user is more willing to make communication with outside world, then the user is more likely to join the corresponding community. So, personal willingness may reduce the probability of generating large-scale communities so as to improve the accuracy and reliability of community detection and increase the stability of community structure. This paper proposes a social community detection and message propagation scheme based on personal willingness in social network. In the proposed scheme, the social community detection algorithm extracts node attributes and then uses modularity degree, interest degree and personal willingness to sophisticatedly detect social communities; also, the message propagation method is based on the exponential model, which constructs feature vector by edge feature and node feature, willingness vector by personal willingness and community willingness, and related basic relationship by propagation probability and propagation delay. Based on the Weibo, YouTube and Digg data, the experiments show that our proposed scheme can ensure the stability and reliability of social community detection and add the initiative and effectiveness of message propagation among users.

Keywords Social network · Personal willingness · Community detection · Message propagation

1 Introduction

With the wide development of online social network, it has become the carrier of social relationship and message propagation. Currently, more and more people participate in information propagation and information access through online social network. So, social software has been an important tool in online social network, where more and more people are using social software to chat with each other and express their individual viewpoints, then the size of users and exchanged messages increases sharply. For example, the social software, such as Facebook, Twitter and Weibo[1], can provide many chances to exchange individual information (node attribute), which include friend records, individual hobbies and location information. Then, the online social network has become a research hot spot with its huge scale, complex structure and huge amount of information. Many scholars focus on the relationship of online social network, off-line society and economics. For example, through analyzing hierarchical structure, social related attributes and message propagation method, we may know how social network influences people's living conditions, lifestyle and social relationship, such as public opinion propagation, false message propagation and network crime .

Currently, online social network provides many service applications, and it can transmit different information contents to users timely. In general, because of the relevance of personal attributes and information content, the interactions among users can dynamically form, change and influence the

Communicated by V. Loia.

✉ Ke Gu
gk4572@163.com

[1] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China

[2] School of Information Science and Engineering, Central South University, Changsha 410083, China

[1] In China, Weibo is a kind of social software provided by the SINA corporation, which is widely used by young people.

structure of social network. As online social network consists of a number of online social organizations and individuals, if there is more number of nodes (users) in one community, then we may consider that the community is more social and active, which is seen as one of online social organizations. In online social network, online community is gradually integrated into people's daily life and plays an important role, which provides a fixed social circle for information propagation and sharing to express personal requirements and relationships. Currently, the researches of overlapping and detecting online community are of great significance to the analysis of structure, function and characteristics of online social network (Newman and Clauset 2015). The related researches are mainly carried out in different directions, such as member influence, user interest degree and structure model. Also, with the rapid development of Internet, the ways of information propagation are greatly changed. Information propagation can be seen as people's interactive behaviors, which have many characteristics such as many to many relationships, real time and quickness. In online social network, information propagation is influenced by a variety of factors, such as network structure, node characteristics and information content. So, the related researches about information propagation have great significance to locate target groups, protect user privacy, monitor internet public opinion and so on Li et al. (2015).

In online social network, node attributes and node wishes (or personal willingness) are important factors to influence community structure and information propagation. Personal willingness can be used to describe the subjective initiative of node (user) to communicate information with outside world. Personal willingness takes full account of related node attributes. The personal willingness is greater, the corresponding user is more willing to make communication with outside world, then the user is more likely to join the corresponding community. So, personal willingness may reduce the probability of generating large-scale communities so as to improve the accuracy and reliability of community detection and increase the stability of community structure.

2 Our contributions

In this paper, we propose a social community detection and message propagation scheme based on personal willingness in social network. In the proposed scheme, the social community detection algorithm uses modularity degree to detect social communities with interest degree and personal willingness, and the message propagation method introduces edge feature and node feature and then constructs willingness vector. The main contributions of this paper are as follows:

(1) The social community detection algorithm extracts node attributes and then uses modularity degree, interest degree and personal willingness to sophisticatedly detect social communities. The algorithm is based on community willingness and personal willingness to detect community structure. The algorithm gives priority to the initiative of nodes in the community detection, and thus, it reduces the probability of generating large-scale communities.
(2) The message propagation method is based on the exponential model, which constructs feature vector by edge feature and node feature, willingness vector by personal willingness and community willingness, and related basic relationship by propagation probability and propagation delay. Because the model constructs willingness vector related to the initiative of nodes, the model is conducive to the stability and continuity of information propagation.
(3) We make the experiments mainly from two aspects of community detection and message propagation to demonstrate the effectiveness of joining personal willingness: (a) in the community detection, the experiments show that personal willingness is greater, the corresponding user is more willing to make communication with outside world, then the user is more likely to join the corresponding community, and thus, the proposed algorithm may reduce the probability of generating large-scale communities so as to improve the accuracy and reliability of community detection and increase the stability of community structure; (b) in the message propagation, the experiments show that personal willingness may influence the construction of message propagation model, where the personal willingness is greater, the corresponding node is more willing to propagate messages.

3 Related work

At present, with the development and evolution of social network, social network has become a research hot spot. The related researches of social network mainly include overlapping community discovery (Devi and Poovammal 2016), community detection (Shang et al. 2016), privacy protection (Buccafurri et al. 2016), message propagation (Liu and Li 2016) and social networking (Guo et al. 2016). In this paper, we focus on community detection and message propagation.

3.1 Community detection

Currently, the community detection algorithms (Shen et al. 2009; Blondel et al. 2008; Ahn et al. 2010; Shi et al. 2013; Von Luxburg 2007; Whang et al. 2013; Lancichinetti et al. 2009; Pons and Latapy 2006; Raghavan et al. 2007; Zhao et al.

2016) mainly use the different methods to divide the online social network structure, which include (1) graph-based partitioning algorithm (Shen et al. 2009); (2) module degree algorithm (Blondel et al. 2008); (3) edge clustering algorithm (Ahn et al. 2010; Shi et al. 2013); (4) hierarchical clustering algorithm (Von Luxburg 2007); (5) seed dispersal method (Whang et al. 2013; Lancichinetti et al. 2009); (6) random walk algorithm (Pons and Latapy 2006) and (7) label propagation method (Raghavan et al. 2007; Zhao et al. 2016). Also, as the social network researches are deepened, many scholars consider that node attributes should be added to community detection. Steinhaeuser and Chawla (2010) proposed a community detection algorithm to classify communities, which is based on weight-based edge and node attribute's similarity. They recommended implementing their method with the random walk approach. Dang and Viennet (2012) proposed a community detection method based on Luovain algorithm, which uses module degree and node attribute's similarity to make weighted sum. Kewalramani (2011) used the similarity of metadata (based on correlation) to detect communities by clustering means in Twitter. Deitrick and Hu (2013) made emotional analysis based on the Twitter content that the users send in a period of time, and they improved the efficiency of the related community detection method. Xu et al. (2016) analyzed community detection and friendship prediction in mobile social network and proposed a method of constructing community structure based on combination entropy. Guo et al. (2016) proposed a relation-weight-clustering model to construct twitter users' network, where their model takes twitter users' "@" and "RT@" behaviors into account. Tagarelli et al. (2017) proposed a novel modularity-driven ensemble-based approach to multilayer community detection, where it may find consensus community structures that not only capture prototypical community memberships of nodes, but also preserve the multilayer topology information and optimize the edge connectivity in the consensus via modularity analysis. Amelio and Pizzuti (2016) proposed a framework for community discovery in temporal multiplex networks by extending the evolutionary clustering approach to encompass both time and multiple dimensions. In their extended framework, the problem of finding community structures for time-evolving networks with multiple types of ties is reformulated by adding the concept of dimensional smoothness. Sun and Lin (2013) proposed a probabilistic generative model to detect latent topical communities among users, where social tags and resource contents are leveraged to model user interest in two similar and correlated ways. Their primary goal is to capture user tagging behavior and interest and to discover the emergent topical community structure. Jaho et al. (2011) proposed a framework for interest similarity-based community detection in social networks, where nodes are clustered according to common interests. Their proposed framework detects communities over weighted graphs, where graph edge weights are defined based on measures of similarity between nodes' interests in certain thematic areas. Hutair et al. (2017) proposed a novel algorithm that clusters the nodes in social networks into communities based on their geodesic location and the similarity between their interests. Yang et al. (2011) proposed a node interest similarity method-based P2P trust model, which takes both node interest bias and reputations in each interest domain into consideration and uses interest domain reputation vector to maintain the behaviors of node in specific interest domain. Their proposed method uses interest similarity between nodes to weight domain local trust recommendation. Currently, the above-mentioned papers only use node interest as a measure to construct the algorithms or models of detecting communities. As the related works do not analyze user's behaviors to divide behavior attributes of user in more precise degree, personal willingness is not considered into the existing related works. Compared with node interest, personal willingness may be used to describe the subjective initiative of user to communicate information with outside world. The personal willingness is greater, the corresponding user is more willing to make communication with outside world. So, we focus on what personal willingness can influence in community detection and message propagation in this paper.

3.2 Message propagation

The researches of message propagation mainly focus on modeling online social network. The early models include the independent cascade model (Goldenberg et al. 2001) and the linear threshold model (Young 2000). Based on the independent cascade model, Kempe et al. (2003) proposed a decreasing cascade model. Centola (2010) analyzed the impact of the healthy network structure on message propagation, and then, they considered that the information spreads faster and farther in the good cluster network. Liu and Li (2016) proposed a novel data propagation scheme to maximize data propagating rate under the limiting overhead of propagating messages. In their scheme, a time- consistent Markov model is designed to analyze the interest transformation of each neighbor node to decide when to forward the message and select which node to forward the message further. Also, two utility functions are studied to evaluate the service ability of nodes with different interests in messages. Yang and Counts (2010) constructed another model to capture three main characteristics of information propagation: speed, scale and scope. They analyzed that the factors affecting the three characteristics are user's related attributes and information propagation labels, and they also gave a quantitative measure of the three characteristics. Lagnier et al. (2013) proposed a linear threshold model based on decaying reinforced user-centric. The model can predict how information spreads in online social network, whose influence factors

include interest of users, influence of adjacent users and propagation willingness of users. Based on the asynchronous independent cascade (AsIC) model, Saito et al. (2011) constructed the target function on propagation probability and attribute vector of neighboring nodes and then built a maximum likelihood model to solve propagation rate and delay parameter. Spiro et al. (2012) proposed a time-based model in the Twitter platform, which gives a statistical analysis of message propagation. They considered that the influence factors of delay are: (1) relevant attributes of user, such as user's authority and active degree; (2) relevant attributes of information, such as message label, URL and hot searching events. Ouadrhiri et al. (2017) proposed a message propagation control model under epidemic routing protocol in delay-tolerant networks. They model the messages propagation under the epidemic routing protocol by an ordinary differential equation, and they derive the optimal retention of a message by a node while taking into account that all nodes are infected. Their simulation results show that the proposed model can reach the same performances of epidemic routing while minimizing the resource consumption. Kim and Yoo (2012) examined the role of sentiment in information propagation. They make use of political communication in the Twitter space and relate emotion expressions in a message to the degrees of responses generated by the message. They also compare differences between user reply and retweet behavior with respect to sentiment variables. Itakura and Sonehara (2013) proposed the importance of Twitter's mention function as another method of message propagation. In their work, the graphs constructed from Twitter's retweet, mention and reply functions show structural differences, which suggest that the mention function is the most efficient method of reaching the mass audience.

4 Preliminaries

Given a weighted and directed graph $G(V, E)$ representing the relationship of users in online social network, the set of nodes is represented by V, the set of edges is represented by E denoting the set of messages transferred by users, the number of nodes is represented by $k = |V(G)|$ and the number of edges is represented by $n = |E(G)|$, then a social network with k nodes may be represented by an adjacency matrix $B_{k\times k}$, where $B_{i,j}$ denotes

$$B_{i,j} = \begin{cases} 1, & \text{if the node } i \text{ and the node } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases}$$

and the number of edges is

$$n = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} B_{i,j}.$$

Definition 1 *Node degree (user degree)* It denotes the influence range of the user i in social network; namely, the number of all edges associated with the node (user) i is the node degree, represented as follows:

$$m_i^{\text{in}} = \sum_{j=1}^{k_{\text{in}}} B_{j,i}, \quad m_i^{\text{out}} = \sum_{j=1}^{k_{\text{out}}} B_{i,j}$$

where k_{in} is the number of all edges associated with the node i and k_{out} is the number of all edges associated from the node i. So, m_i^{in} is the in-degree of the node i and m_i^{out} is the out-degree of the node i.

Definition 2 *Edge weight* $w_{i,j}$ The adjacency matrix $W_{k,k}$ is used to represent the edge weights in social network, where the edge weight between the node i and the node j is expressed by the matrix element $w_{i,j} \in W_{k,k}$. The node i and the node j are connected more closely, representing that the messages are transferred more frequently between the node i and the node j; thus, the possibility of message propagation is greater, the value of $w_{i,j}$ is greater; on the contrary, if the connection between the node i and the node j is sparse, then the possibility of message propagation is smaller.

Definition 3 *Module degree* D It is the ratio of the edge density in the community and the edge density among the associated communities, whose formula Newman and Girvan (2004) is as follows:

$$D = \frac{1}{n} \cdot \sum_{i=1}^{z} \sum_{j=1}^{z} \left[B_{i,j} - \frac{m_i^{\text{out}} \cdot m_j^{\text{in}}}{n} \right] \cdot \sigma(c_i, c_j)$$

where n is the number of edges, z is the number of nodes, c_i denotes the community that the node i belongs to, c_j denotes the community that the node j belongs to and σ is the function used to compute the max value of edge number between c_i and c_j. From the above formula, we can know that when the module degree of network is the minimum value of 0, the initialization of each node is independent to form a single community; when the module degree of network is the maximum value of 1, all nodes are detected into a community. The process of community detection and classification is based on the modularity maximization principles: (1) the terms of negative value should be excluded; (2) the nodes with greater interest similarity should be partitioned to the same communities.

Definition 4 *Personal willingness* ε_u **of user (node)** It is the willingness degree chosen by a user when the user joins a community, where $\varepsilon_u \in [0, 1]$ and the user may choose his own willingness value according to his requirements. When $\varepsilon_u = 1$, the user is completely open to outside world, and the information from outside world is unconditional accepted;

$\varepsilon_u = 0$, the user is completely hidden to outside world; namely, the user does not accept any information from outside world.

Definition 5 *Node willingness* $\varepsilon_{i,j}$ It is a measure to control the number of connection (or the edge intimacy) between the node i and the node j where $\varepsilon_{i,j} \in [0, 1]$. Namely, node willingness is used to control the number of transferred messages between the node i and the node j, which is influenced by the personal willingness of the node i and the node j. Node willingness $\varepsilon_{i,j}$ is proportional to the number of transferred message between the node i and the node j, which may dynamically be adjusted. When $\varepsilon_{i,j} = 0$, the communication willingness between the node i and the node j is 0; namely, the node i and the node j are not willing to make communication each other. When $\varepsilon_{i,j} = 1$, the node i and the node j are the most willing to make communication each other.

Definition 6 *Community willingness* ε_c It is a measure to set the communication degree for the community with respect to other communities, where $\varepsilon_c \in [0, 1]$. Community willingness may be set by every community according to the different requirements, where every member from community is required to satisfy the condition of community willingness.

Definition 7 *Edge intimacy degree* $l^{\varepsilon}_{i,j}$ It indicates the frequent degree of interaction or the frequent degree of message transmission between the node i and the node j, which is influenced by node willingness $\varepsilon_{i,j}$. Edge intimacy degree $l^{\varepsilon}_{i,j}$ is computed as follows:

$$r^{\varepsilon}_{i,j} = r_{i,j} \cdot \varepsilon_{i,j},$$
$$r^{\varepsilon}_{i} = \sum_{x=1}^{m^{\text{out}}} r^{\varepsilon}_{i,x}, \; r^{\varepsilon}_{j} = \sum_{x=1}^{m^{\text{in}}} r^{\varepsilon}_{x,j},$$
$$l^{\varepsilon}_{i,j} = \frac{r^{\varepsilon}_{i,j}}{\sqrt{r^{\varepsilon}_{i} \cdot r^{\varepsilon}_{j}}}$$

where we set that $r_{i,j}$ is the original number of message transmission from the node i to the node j without node willingness $\varepsilon_{i,j}$, $r^{\varepsilon}_{i,j}$ is the number of message transmission controlled by node willingness $\varepsilon_{i,j}$, m^{out} is the out-degree of the node i and m^{in} is the in-degree of the node j.

Definition 8 *Information content feature* It is defined as $\text{Cnt} = \{\text{cnt}^1, \text{cnt}^2, \ldots \text{cnt}^N\}$, where cnt^k is the feature of the kth piece of message content.

Definition 9 *User feature set* It is defined as $U = \{u^1, u^2, \ldots u^N\}$, where u^k is the attribute feature of the kth user.

Definition 10 *Manhattan distance* It is the sum of the distance from the multiple dimensions, whose formula is as follows:

$$\text{dist}(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$$

where $x_i \in X$ and $y_i \in Y$.

5 Community detection

In the section, we propose a community detection scheme based on the attributes of node and edge, including edge intimacy degree, node willingness and node interest degree. Our proposed scheme makes weights to the edges of social network according to node degree and edge intimacy degree and then constructs a method for module degree detection with node interest degree and node willingness. Our proposed scheme uses a hierarchical structure approach to detect communities.

5.1 Edge weight

According to the related definitions, the computation of edge weight is based on edge intimacy degree and node degree (Yao et al. 2015). As node willingness $\varepsilon_{i,j}$ is added to the computation of $l^{\varepsilon}_{i,j}$, the computation of edge weight is also related to node willingness. Edge weight is defined as:

$$w_{i,j} = \eta \cdot l^{\varepsilon}_{i,j} + (1 - \eta) \cdot \sqrt{\frac{B_{i,j}}{m^{\text{out}}_i \cdot m^{\text{in}}_j}}$$

where $\eta \in [0, 1]$ is the impact factor. In online social network, because the attributes of node are different, the impact factors of community division are also different. We consider that edge intimacy degree has greater influence to community division when η is greater; on the contrary, it has smaller influence. When $\eta = 0$, community division is only based on topological structure or community structure. When $\eta = 1$, edge weight is only edge intimacy degree, where edge intimacy degree indicates the frequent degree of interaction or message transmission between two nodes. When edge intimacy degree is greater, the possibility of information transmission is higher. Therefore, the impact factor η is introduced to the above formula, which makes balance to decrease the influence of the attributes of node.

5.2 Weighted module degree

Module degree is the most commonly used method for community partition. It is also a general method to measure the quality of community structure. Based on module degree, the community structure may be clearly detected, and the

partitioned community structure is more stable. Therefore, our proposed scheme uses module degree to detect community structure with interest degree and personal willingness, which can increase the stability of community structure and reduce the probability of generating large-scale communities. The work of Yao et al. (2015) proposed a novel measure to compute weighted module degree and its increment. According to Definition 3 and the formula of edge weight, weighted module degree is defined as follows:

$$D^* = \frac{1}{w} \cdot \sum_{i=1}^{z} \sum_{j=1}^{z} \left[w_{i,j} - \frac{w_i^{\text{out}} \cdot w_j^{\text{in}}}{w} \right] \cdot \sigma(c_i, c_j)$$

where $w_{i,j}$ is the edge weight, w_i^{out} is the out-weight of the node i, w_j^{in} is the in-weight of the node j and w is the sum of all the edge weights. Based on the above formula, when the node i joins the community c_j that the adjacent node j belongs to, the increment ΔD^* of module degree is defined as follows:

$$\Delta D^* = \left[\frac{w_{c_j} + w_{i,c_j}}{w} - \frac{(w_{c_j}^{\text{in}} + w_i^{\text{in}})(w_{c_j}^{\text{out}} + w_i^{\text{out}})}{w^2} \right] - \left[\frac{w_{c_j}}{w} - \frac{w_{c_j}^{\text{out}} \cdot w_{c_j}^{\text{in}}}{w^2} - \frac{w_i^{\text{out}} \cdot w_i^{\text{in}}}{w^2} \right]$$

where w_{c_j} is the sum of the internal edge weights of the community c_j, w_{i,c_j} is the sum of the edge weights of the node i and other associated nodes from the community c_j, $w_{c_j}^{\text{in}}$ is the sum of the in-weights of the community c_j, $w_{c_j}^{\text{out}}$ is the sum of the out-weights of the community c_j, w_i^{in} is the in-weight of the node i and w_i^{out} is the out-weight of the node i.

5.3 Node interest degree

Users who share a common interest are more likely to join the same community, and they will share messages of mutual interest. When users join one community, users can define the labels of their own interests as the form of $< keyword, weight >$ representing user's interest. Therefore, the clustering approach to the message vectors of nodes is used to extract the interest similarity of nodes (Rao and Raju 2016).

For a message cnt_i, it can be expressed as

$$\text{cnt}_i = \{(n_1, w_1); (n_2, w_2); \ldots (n_N, w_N)\}$$

where the ith keyword in the message cnt_i is represented by n_i, the weight of the ith keyword is represented by w_i, and they are ranked in descending order.

Therefore, based on the structure of message cnt_i, the interest similarity between the node i and the node j can be defined as:

$$\text{Int}(i, j) = \text{sim}(i, j) = \frac{\sum_{n=1}^{m} w_{i,n} \cdot w_{j,n}}{\sqrt{\sum_{n=1}^{m} w_{i,n}^2} \cdot \sqrt{\sum_{n=1}^{m} w_{j,n}^2}} \cdot \varepsilon_{i,j}$$

where m denotes the number of public keywords, $w_{i,n}$ denotes the weight of the nth public keyword of the message from the node i, $w_{j,n}$ denotes the weight of the nth public keyword of the message from the node j and $\varepsilon_{i,j}$ is node willingness. If the value $\text{Int}(i, j)$ is greater than a preset threshold T, then the probability that the node i and the node j are detected into the same community is greater; on the contrary, the probability is less.

5.4 The proposed detection algorithm

In the section, we propose a community detection algorithm based on module degree detection (Yang and Counts 2010), combined with interest degree and personal willingness. Compared with the previous module measure, our proposed algorithm introduces node interest degree and node willingness to guarantee the structural stability of the community members. The proposed algorithm can not only provide hierarchical community structure detection, but also adjust node willingness according to the realistic requirement of community after a period of time so as to further screen the community members to make the community more stable and the message propagation more fluent. The proposed algorithm is described as follows (shown in Fig. 1), which consists of three subalgorithms:

- *Step 1* input the adjacent network $G(V, E)$ and the total number k of nodes, then measure the increment of module degree, and output a set C_1 of communities (see Algorithm 1 for details).
- *Step 2* input the set C_1 of communities and a given threshold T, compute the interest degrees of nodes, then repeatedly screen the set C_1 according to the interest degrees, where the interest degrees are compared with the given threshold T, and finally output a set C_2 of communities and a set Φ of community willingness (see Algorithm 2 for details).
- *Step 3* input the set C_2 of communities and the set Φ of community willingness, then screen the set C_2 according to the condition whether the corresponding node willingness can satisfy a given requirement or not, and finally output a set C of communities and the total number of communities (see Algorithm 3 for details).

1.**The description of Algorithm 1**

Algorithm 1 filters the communities according to module degree: (1) in the adjacent social network $G(V, E)$ with the total number k of nodes, the algorithm initializes each node willingness as the mean value of personal willingness of the corresponding nodes, then gets the new weight value of each edge by the formula of edge weight and initializes each node to form a community, where the total number of communities is k; (2) for each community $i(i \in G)$, the algorithm calculates all the increments ΔD^* of module degree, where we assume the community i tries to join all the adjacent communities; (3) the algorithm looks up the adjacent community with the maximum value of the increments ($\Delta D^* > 0$), and then, the community i joins the corresponding adjacent community; (4) as long as the values of ΔD^* are changing, the process of merging communities will continue by cycle iteration until the communities cannot be partitioned into the communities of higher level; (5) after the end of the partition, the algorithm returns a set C_1 of communities.

Algorithm 1 Community detection algorithm based on module degree

Input: the adjacent social network $G(V, E)$, the total number k of nodes
Output: a set C_1 of communities
Begin
$C_1 \leftarrow \{\{\}, ...\{\}\}$; *//initialize* C_1
for the node $i \in V$ **do**
for the node j that belongs to the adjacent node set $N(i)$ of the node i **do**
$\varepsilon_{i,j} = \frac{\varepsilon_{u_i} + \varepsilon_{u_j}}{2}$; *//initialize each node willingness*
End for
End for
Initialize each node from $G(V, E)$ to form a community and save the structure to C_1;
while being partitioned into the communities of higher level **do**
$Q \leftarrow \{\}$; *//being emptied*
while the values of ΔD^* are changing **do**
for the community $i \in G$ **do**
for the community j that belongs to the adjacent community set $N(i)$ of the community i **do**
Calculate all the increments ΔD^* of module degree; *//where we assume the community i tries to join all the adjacent communities*
if $\Delta D^* > 0$ **then**
$Q \leftarrow$**push**$(\Delta D^*, index)$; *//save* ΔD^* *and the corresponding index*
End if
End for
$Max\_Community \leftarrow$**argmax**(Q); *//locate the adjacent community with the maximum value of the increments*
The community i joins the corresponding adjacent community $Max\_Community$;
Update C_1;
End for
End while
End while
return C_1;
End

2.**The description of Algorithm 2**

Algorithm 2 again filters the communities returned from Algorithm 1 according to node interest degree: (1) based on the set C_1 of communities and the given interest threshold T, for every node i_k from the community $i(i \in C_1)$, the algorithm calculates the node's interest degrees between the node i_k and all the associated nodes from the adjacent community j according to Sect. 5.3; if all the interest degrees of the node i_k are more than T, then the node i_k is deleted from the community i and joins the adjacent community j (shown in Fig. 1); (2) similarly, as long as the values of $Int(i_k, j_k)$ are changing, the process of filtering communities will continue by cycle iteration; (3) after the end of the filter, the algorithm gets a set C_2 of communities; (4) based on the set C_2 of communities, the algorithm calculates the corresponding community willingness for each community, which is the mean value of all personal willingness of the nodes from the corresponding community, and then, the corresponding community willingness is saved to the set Φ; (5) the algorithm returns the sets C_2 and Φ.

Algorithm 2 Community detection algorithm based on interest degree

Input: the set C_1 of communities, the given interest threshold T
Output: a set C_2 of communities, a community willingness set Φ
Begin
$C_2 \leftarrow \{\{\}, ...\{\}\}$; *//initialize* C_2
$C_2 \leftarrow C_1$; *//save* C_1 *to* C_2
while the values of $Int(i_k, j_k)$ are changing **do**
for the community $i \in C_2$ **do**
for the community j that belongs to the adjacent community set $N(i)$ of the community i **do**
for the node $i_k \in i$ **do**
for the node $j_k \in j$ **do**
Calculate $Int(i_k, j_k)$;
Save $Int(i_k, j_k)$;
End for
if all $Int(i_k, j_k) > T$ **then**
The node i_k is deleted from the community i;
The node i_k joins the community j;
End if
Update C_2;
End for
End for
End for
End while
$\Phi \leftarrow \{\}$; *//initialize* Φ
for the community $i \in C_2$ **do**
$sum_{\varepsilon_c} \leftarrow 0$;
for the node $i_k \in i$ **do**
$sum_{\varepsilon_c} \leftarrow sum_{\varepsilon_c} + \varepsilon_{u_{i_k}}$;
End for
$\varepsilon_{c_i} = \frac{sum_{\varepsilon_c}}{|i|}$; *//*$|i|$ *denotes the node number of the community i,* ε_{c_i} *denotes the community willingness*
$\Phi \leftarrow$**push**(ε_{c_i}, i); *//save community willingness*
End for
return C_2 and Φ;
End

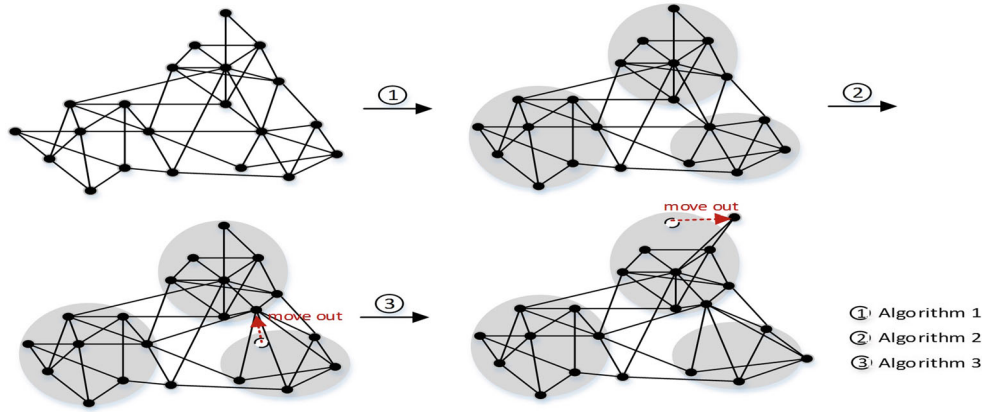


Fig. 1 Community detection

3.**The description of Algorithm 3**

Algorithm 3 adjusts the detection of the communities according to the willingness: (1) based on the set C_2 of communities, for every community c_i from C_2, the algorithm looks up the corresponding community willingness ε_{c_i} from the set $\Phi(\varepsilon_{c_i} \in \Phi)$; (2) for every node from c_i with all $i \in 1, 2, \ldots |C_2|$, the algorithm computes the mean value of all the node willingness between the node and its adjacent nodes; if the mean value is less than $\frac{\varepsilon_{c_i}}{\varpi}$($\varpi$ is the parameter of adjusting willingness, where we may set the parameter value according to our requirement), then it shows the communication willingness of the node is lower than those of other nodes from the same community c_i; thus, the node is removed from c_i (shown in Fig. 1); (3) the algorithm finally returns a set C of communities and the number of all the communities.

Algorithm 3 Community detection algorithm based on willingness

Input: the set C_2 of communities, the community willingness set Φ
Output: a set C of communities and its number $|C|$
Begin
$C \leftarrow \{\{\}, \ldots\ldots\{\}\}$; *//initialize C*
$C \leftarrow C_2$; *//save C_2 to C*
for the community $i \in C$ **do**
for the node $i_k \in i$ **do**
$\varepsilon_{c_i} \leftarrow Find(\Phi, i)$;*//look up the corresponding community willingness ε_{c_i} from the set Φ*
$sum_\varepsilon \leftarrow 0$;
for the node $j_k \in i$ that belongs to the adjacent community set $N(i_k)$ of the node i_k **do**
$sum_\varepsilon \leftarrow sum_\varepsilon + \varepsilon_{i_k, j_k}$;
End for
$average_{i_k} = \frac{sum_\varepsilon}{|N(i_k)|}$; *//$|N(i_k)|$ denotes the node number of $N(i_k)$, $average_{i_k}$ denotes the mean value of all the node willingness between the node i_k and its adjacent nodes*
if $average_{i_k} < \frac{\varepsilon_{c_i}}{\varpi}$ **then**
The node i_k is deleted from the community i;
End if
Update C;
End for
End for
return C and $|C|$;
End

6 Message propagation model

In the section, we show a message propagation model based on personal willingness for social network. In the actual message propagation process, the message transfer is not necessarily synchronous, and the message propagation delay may be generated. Therefore, the message propagation model is commonly constructed on the propagation probability and delay, where the messages are asynchronously transferred and the propagation delays are different. The asynchronous independent cascade (AsIC) model (Saito et al. 2009) is a message diffusion model based on continuous time delay, which can handle and monitor the message asynchronous and cascaded propagation process according to the propagation probability and delay. So, based on the AsIC model, we introduce the features of node attribute and message content to the proposed model and build the model by the exponential mechanism. Also, the proposed model uses the propagation probability $p(i, j, \text{cnt}, \varepsilon)$ and the propagation delay $\tau(i, j, \text{cnt}, \varepsilon)$ as the basic valuated functions. In the proposed model, we introduce and quantify most of the factors that influence message propagation, so that the proposed model can guarantee the security and reliability of message propagation.

6.1 Feature extraction

We need to make feature extraction from social network[2]. In this paper, the extracted features are divided into node fea-

[2] To extract some features whose values are not the interval [0, 1], we may use the min-max standardization method to process the features. Also, other features may be directly got according to the related definitions or formulas.

tures and edge features. Node features include characteristics of communication subjects, characteristics of communication objects and message characteristics. Edge features include relationship characteristics between communication subjects and communication objects, and relationship characteristics between communication objects and propagated information contents. As feature extraction may depict related attributes, related attributes can correspond to relevant features (Zhou et al. 2015).

(1) Node Feature

(1) Characteristics Ψ_s of communication subjects

(a) Node influence: it is the sum of nodes that a node associates with over a period of time, where we set that it is the node's out-degree m_i^{out}; thus,

$$\text{Inf}(i) = m_i^{\text{out}} = \sum_{j=1}^{k_{\text{out}}} B_{i,j},$$

$\text{Inf}(i)$ denotes the node influence of i.

(b) Node authority: it is the difference between in-degree and out-degree of a node. The greater node authority is, the greater the likelihood that messages will be transmitted is; thus,

$$\text{Aut}(i) = \begin{cases} x \in [0,1], & \text{if } m^{\text{in}} - m^{\text{out}} < 0 \\ m^{\text{in}} - m^{\text{out}}, & \text{otherwise} \end{cases}$$

$\text{Aut}(i)$ denotes the node authority of i and x is randomly picked.

(c) Node activity: it is the average number of messages that a node releases, which is computed in days. The propagated messages through highly active nodes are more likely to be released and easier to be received or forward by other nodes (users); thus,

$$Act(i) = r_i^{\varepsilon} = \sum_{x=1}^{m^{\text{out}}} r_{i,x}^{\varepsilon},$$

$Act(i)$ denotes the node activity of i.

(2) Characteristics Ψ_r of communication objects

(a) Node propagation willingness: it is user's willingness to propagate received message, whose definition is shown as the following formula. In the definition, we compute the logarithmic value of the ratio of the number m_{transmit} of messages forwarded by a user to the number m_{original} of original messages, and then, the logarithmic value multiplies the user's willingness ε_u. The stronger the user's willingness to propagate message is, the greater its value is.

$$\text{Wil}(j) = \log\left(\frac{m_{\text{transmit}}}{m_{\text{original}}} + 1\right) \cdot \varepsilon_u,$$

$\text{Wil}(j)$ denotes the node propagation willingness.

(b) Node propagation characteristic: it is defined as the product of the propagation characteristic of the messages from the node j to the node i multiplying with the node willingness $\varepsilon_{i,j}$, shown as the following formula.

$$\zeta_{i \leftarrow j} = \frac{\log(1 + \text{Inf}(j))}{\log\left((1 + \text{Inf}(i)) \cdot (1 + \text{Inf}(j))\right)} \cdot \varepsilon_{i,j},$$

$\zeta_{i \leftarrow j}$ is the node propagation characteristic. From the above formula, we may know the node propagation characteristic is related to the influences of the node i and the node j, where we need to remark that (1) $\zeta_{i \leftarrow j} \neq \zeta_{j \leftarrow i}$ normally; (2) if $\text{Inf}(i) \ll \text{Inf}(j)$, then $\zeta_{i \leftarrow j} \approx \varepsilon_{i,j}$; it shows that the node i is easy to accept the messages transmitted by the node j; thus, the node j is more influential; (3) conversely, if $\text{Inf}(i) \gg \text{Inf}(j)$, then $\zeta_{i \leftarrow j} \approx 0$; it shows that the node i is difficult to accept the messages transmitted by the node j; thus, the node i is more influential.

(3) Characteristics Ψ_{cnt} of propagated messages

(a) For the situation whether message contains URL link[3], we may set the following expression:

$$\text{Url(cnt)} = \begin{cases} 1, & \text{if URL} \subset \text{cnt} \\ 0, & \text{otherwise} \end{cases}$$

(b) For the situation whether message contains label (such as the symbol #label content#)[4], we may set the following expression:

$$\text{Lab(cnt)} = \begin{cases} 1, & if\ \#content\# \subset \text{cnt} \\ 0, & otherwise \end{cases}$$

(2) Edge Feature

(1) Characteristic relation $\Psi_{s,r}$ between the communication subject and the communication object

(a) Interest similarity: the users with the same or similar interests are more likely to propagate the same kind of messages. It is derived from the extraction of node interest in the community detection. The complete formula is shown as

[3] The URL links to a network page that allows more detailed or comprehensive interpretation of the message content.

[4] The content of the label information can attract more readers' attention or generate interest similarity.

the formula of interest similarity from Sect. 5.3, where

$$\mathrm{Int}(i, j) = \mathrm{sim}(i, j).$$

(b) Directed propagation: if message sender directly propagates related messages to receiver according to receiver's ID (such as " receiver's ID"), then it indicates a close connection between the two users and a greater intimacy degree.

$$\mathrm{Dit}(i, j) = \begin{cases} 1, & \text{if } i \text{ and } j \text{ were associated} \\ 0, & \text{otherwise} \end{cases}$$

(2) Characteristic relation $\Psi_{r,\mathrm{cnt}}$ between the communication object and the communication content
The characteristic relation $\Psi_{r,\mathrm{cnt}}$ is represented by propagation interest. Propagation interest is to measure the difference between user's interest and propagated content, which evaluates whether the user is interested in the content. So, it is defined as the similarity between the communication object and the communication content according to the calculation of the Manhattan distance, and the formula is as follows:

$$\mathrm{Spr}(U, C) = \mathrm{dist}(U, C) = \sum_{k=1}^{n} |u_k - c_k|,$$

where $U = (u_1, u_2, \ldots u_n)$ is the issued document vector of a user and $C = (c_1, c_2, \ldots c_n)$ is the propagated document vector of the user[5].

To extract the features, we need to use the min-max standardization method to process some of the features, where all the values of the features are standardized and mapped to the interval [0, 1]. For example, the out-degrees of a node are {10,30,5,25,15,35,7} in a period of time; its node influences are {10,30,5,25,15,35,7} according to the corresponding definition; then, we may standardize the value of 25 by the min-max standardization method as follows:

$$\frac{25-5}{35-5} = \frac{20}{30} \approx 0.667.$$

6.2 Model construction of message propagation

The model construction of message propagation (Saito et al. 2011) is mainly related to the propagation probability function and the propagation delay function. The model construction needs to build up the relationship between them according to the extracted feature set from Sect. 5.1. In this paper, we use the node feature and the edge feature to build the feature vectors, where the vectors' dimensions are related to node feature number k, edge feature number n and resource (content) number r. Also, we build a personal willingness vector Ψ_ε. The construction of all the vectors is shown as follows.

$$\Psi_k = \begin{Bmatrix} \Psi_s \\ \Psi_r \\ \Psi_{\mathrm{cnt}} \end{Bmatrix},$$

$$\Psi_n = \begin{Bmatrix} \Psi_{s,r} \\ \Psi_{r,\mathrm{cnt}} \end{Bmatrix},$$

$$\Psi_\varepsilon = \begin{Bmatrix} \varepsilon_u \\ \varepsilon_{i,j} \\ \varepsilon_c \end{Bmatrix}.$$

In message propagation, the vector Ψ_ε takes full account of the willingness value ε_u of a single node, the willingness value $\varepsilon_{i,j}$ between nodes and the community willingness value ε_c. The vector can show the executive effectiveness of personal willingness to influence message propagation.

In this paper, a basic function $f(i, j, \mathrm{cnt}, \varepsilon)$ is used to linearly represent the correlation characteristics Ψ_k, Ψ_n and Ψ_ε, thus

$$f(i, j, \mathrm{cnt}, \varepsilon) = \alpha_0 + \alpha_1^T \cdot \Psi_k + \alpha_2^T \cdot \Psi_n + \alpha_3^T \cdot \Psi_\varepsilon,$$

where i and j are the node's indexes, cnt is the feature of message content, ε is the personal willingness, α_0 represents a constant value, α_1 represents the weight of node feature, α_2 represents the weight of edge feature and α_3 represents the weight of personal willingness. So, the greater the weight is, the greater the impact on the propagation probability is. Also, the greater the value of personal willingness is, the faster the flow of messages is. Then, the Bayesian logistic function can indicate the propagation probability function $p(i, j, \mathrm{cnt}, \varepsilon)$ as follows:

$$p(i, j, \mathrm{cnt}, \varepsilon) = \frac{1}{1 + \exp\{-f(i, j, \mathrm{cnt}, \varepsilon)\}}.$$

Also, the propagation delay function $\tau(i, j, \mathrm{cnt}, \varepsilon)$ is represented by the linear combination of Ψ_k, Ψ_n and Ψ_ε, thus

$$\tau(i, j, \mathrm{cnt}, \varepsilon) = \beta_0 + \beta_1^T \cdot \Psi_k + \beta_2^T \cdot \Psi_n + \beta_3^T \cdot \Psi_\varepsilon,$$

where β_0 represents a constant value, β_1 represents the weight of node feature, β_2 represents the weight of edge feature and β_3 represents the weight of personal willingness. Similarly, the greater the weight is, the greater the impact on the propagation delay is. Also, the greater the value of personal willingness is, the faster the flow of messages is.

[5] In this paper, we use lexical items to construct document vector.

Then, we build the message propagation model according to the propagation probability function and the propagation delay function, and we introduce time attenuation factor to the model. In Goyal et al. (2010), Liu et al. (2017) and Fang et al. (2018) the related studies pointed out that the message propagation process is asynchronous and random, and the ability and influence of message propagation between nodes will decay as time interval increases, which is consistent with the exponential decay rule. In order to verify the conclusion, the related studies collected and analyzed a large number of cascaded messages distributed in social network (such as Flickr data set) and then found that the message propagation process is closest to the exponential distribution with time interval changing. Therefore, we choose the exponential model to establish the message propagation mechanism in this paper. The formula of the propagation probability density is described as follows:

1) when $t_j > t_i$,

$$\begin{aligned} &y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right) \\ &\quad = p(i,j,\mathrm{cnt},\varepsilon)\cdot\tau\left(i,j,\mathrm{cnt},\varepsilon\right) \\ &\qquad \cdot\exp\left\{-\tau\left(i,j,\mathrm{cnt},\varepsilon\right)\cdot\left(t_i-t_j\right)\right\}, \end{aligned}$$

2) when $t_j \le t_i$,

$$y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)=0,$$

where $y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)$ is the propagation probability density, which represents the probability that the node i will pass the message to the node j during this time between t_i and t_j, $\alpha=(\alpha_0,\alpha_1^T,\alpha_2^T,\alpha_3^T)$ and $\beta=(\beta_0,\beta_1^T,\beta_2^T,\beta_3^T)$.

In message propagation, we introduce the personal willingness vector Ψ_ε to the message propagation model, whose values are the interval [0, 1]. So, when $y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)$ is the propagation probability that the node i will pass the message cnt to the node j during this time between t_i and t_j, the propagation probability will decay as the time interval $\Delta t=t_j-t_i$ increases, which is consistent with the exponential decay rule. Then, we may compute the integral value of the propagation probability density function during this time between t_i and t_j as follows:

$$\begin{aligned} &Y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right) \\ &\quad=\int_{t_i}^{t_j} y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)dt, \end{aligned}$$

where $Y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)$ is the cumulative function of the propagation probability during this time between t_i and t_j. So, the survival probability

$$E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)=1-Y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right),$$

where $E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)$ denotes the probability that the node j does not receive the message cnt from its adjacent node i until the time t_j. Then, the probability that the node j does not receive the message cnt from its adjacent nodes before the time t_j except for its adjacent node ω is described as $\prod_{i\in R\_Node}^{i\neq\omega} E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)$, where R_Node denotes a set whose elements are the adjacent nodes that received the message cnt before the time t_j. Therefore, the probability that the node j only receives the message cnt from the adjacent node ω at the time t_j is described as

$$y\left(\left(j,t_j\right)|\left(\omega,t_\omega\right),\alpha,\beta\right)\cdot\prod_{i\in R\_Node}^{i\neq\omega} E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right).$$

For a message content set $C=\{\mathrm{cnt}_1,\mathrm{cnt}_2,\ldots\mathrm{cnt}_N\}$ where cnt_l is the lth message content, we assume that the message set is propagated in a social network graph $G(V,E)$ (the number of nodes is represented by $k=|V(G)|$). Then, all the nodes that received the lth message content cnt_l and their corresponding times may be denoted as a set $S_l=\{(v_{l,1},t_{l,1}),(v_{l,2},t_{l,2})\ldots(v_{l,k},t_{l,k})\}$, where $v_{l,x}$ is the index of node and $t_{l,x}$ is the corresponding time with $x\in\{1,2,\ldots k\}$. So, for the lth message content cnt_l, we may get its propagation probability in the $G(V,E)$ as follows:

$$\begin{aligned} \hbar(\mathrm{cnt}_l|\alpha,\beta)=\prod_{(j,t_j)\in S_l}&\left[y\left(\left(j,t_j\right)|\left(\omega,t_\omega\right),\alpha,\beta\right)\cdot\right. \\ &\left.\prod_{i\in R\_Node}^{i\neq\omega} E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)\right]. \end{aligned}$$

So, for the message content set C, we may get its propagation probability in the $G(V,E)$ as follows:

$$H(C|\alpha,\beta)=\prod_{\mathrm{cnt}_l\in C}\hbar(\mathrm{cnt}_l|\alpha,\beta).$$

Therefore, we may solve the maximum natural estimating values of α and β according to the works of Saito et al. (2011); Zhou et al. (2015), which are the solutions of the constructed message propagation model; namely, $(\hat{\alpha},\hat{\beta})$ must satisfy the function

$$\mathbf{min}\{-\mathbf{lg}H(C|\alpha,\beta)\}.$$

Further, to solve the values of α and β, we first set the objective function $F(C|\alpha,\beta)=-\mathbf{lg}H(C|\alpha,\beta)$ and solve the partial derivatives $\frac{\partial F(C|\alpha,\beta)}{\partial\alpha}$ and $\frac{\partial F(C|\alpha,\beta)}{\partial\beta}$[6], and then, we can get that

[6] We may prove the function $F(C|\alpha,\beta)$ has the continuous first-order partial derivatives on α and β.

$$
\begin{aligned}
\frac{\partial F(C|\alpha,\beta)}{\partial\alpha} &= -\frac{\partial \mathbf{lg} H(C|\alpha,\beta)}{\partial\alpha} \\
&= -\frac{\partial \mathbf{lg} \prod\limits_{\mathrm{cnt}_l \in C} \hbar(\mathrm{cnt}_l|\alpha,\beta)}{\partial\alpha} \\
&= -\sum_{\mathrm{cnt}_l \in C} \frac{\partial \mathbf{lg}\hbar(\mathrm{cnt}_l|\alpha,\beta)}{\partial\alpha} \\
&= -\sum_{\mathrm{cnt}_l \in C} \\
&\quad \frac{\partial \mathbf{lg} \prod_{(j,t_j)\in S_l}[y\left(\left(j,t_j\right)|\left(\omega,t_\omega\right),\alpha,\beta\right)\cdot\prod_{i\in R\_\mathrm{Node}}^{i\neq\omega} E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)]}{\partial\alpha} \\
&= -\sum_{\mathrm{cnt}_l \in C}\sum_{(j,t_j)\in S_l} \\
&\quad \frac{\partial \mathbf{lg}[y\left(\left(j,t_j\right)|\left(\omega,t_\omega\right),\alpha,\beta\right)\cdot\prod_{i\in R\_\mathrm{Node}}^{i\neq\omega} E\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)]}{\partial\alpha} \\
&= -\sum_{\mathrm{cnt}_l \in C}\sum_{(j,t_j)\in S_l} \\
&\quad \frac{\partial \mathbf{lg}[y\left(\left(j,t_j\right)|\left(\omega,t_\omega\right),\alpha,\beta\right)\cdot\prod_{i\in R\_\mathrm{Node}}^{i\neq\omega}(1-\int_{t_i}^{t_j} y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right)dt)]}{\partial\alpha},
\end{aligned}
$$

where

$$
\begin{aligned}
&y\left(\left(j,t_j\right)|\left(i,t_i\right),\alpha,\beta\right) \\
&\quad = p(i,j,\mathrm{cnt},\varepsilon)\cdot\tau\left(i,j,\mathrm{cnt},\varepsilon\right) \\
&\qquad \cdot\exp\left\{-\tau\left(i,j,\mathrm{cnt},\varepsilon\right)\cdot\left(t_i-t_j\right)\right\}.
\end{aligned}
$$

Similarly, we may solve the partial derivative $\frac{\partial F(C|\alpha,\beta)}{\partial\beta}$.

Then, we may use the stochastic gradient descent algorithm (Vorontsov 1998; Roux et al. 2013; Johnson and Zhang 2013) to solve the maximum natural estimating values $(\hat{\alpha},\hat{\beta})$. The main idea of the algorithm is that 1) it starts to train the objective data set from the initial values, and then, it declines a step (generally set to a small value) along the gradient of the objective function and updates the data set at every iteration; 2) until the objective function is converged, it may find the optimal solutions in a global or local optima. So, according to the stochastic gradient descent algorithm, we may solve the optimal values of α and β. For example, we may compute the optimal value of α by the following formulas:

$$
\begin{cases}
\alpha^{(t)} = \alpha^{(t-1)} - \lambda\cdot\dfrac{\partial F(C|\alpha^{(t-1)},\beta^{(t-1)})}{\partial\alpha^{(t-1)}} \\
\quad\cdots \\
\alpha^{(2)} = \alpha^{(1)} - \lambda\cdot\dfrac{\partial F(C|\alpha^{(1)},\beta^{(1)})}{\partial\alpha^{(1)}} \\
\alpha^{(1)} = \alpha^{(0)} - \lambda\cdot\dfrac{\partial F(C|\alpha^{(0)},\beta^{(0)})}{\partial\alpha^{(0)}}
\end{cases}
\tag{1}
$$

where $\alpha^{(0)}$ is the initial value of α and λ is a step value. Similarly, we may solve the optimal value of β. Finally, based on α and β, we may get a message propagation model with regard to the features mentioned in Sect. 6.1.

7 Experiments and analysis of the proposed scheme

7.1 Data set

Based on the open API interfaces of the social networking software called as Weibo from the Sina company[7], we use the crawler tool to grab some data sets from the Weibo and then process the data sets as our experimental data sets. Additionally, we download other data sets to test our proposed scheme, such as YouTube (Dang and Viennet 2012) and Digg (Lin et al. 2011).

(1) We process the data sets in advance according to the years of data accumulation and then construct different scale networks, as shown in Table 1. In Table 1, the size of the 4 networks is increasing with the steady growth of time; thus, the efficiency of the proposed community detection algorithm can be detected, where the algorithm reduces its randomness and contingency by introducing the interest similarity of network nodes.
(2) We randomly choose some network nodes as the initial nodes to grab the users' personal information and

[7] The Sina company is a big network company, which provides many functions of Web site portal.

Table 1 Different scale network data

Network	Date (partly)	Node number	Edge number
I1	2010	812	5290
I2	2012	1608	12,960
I3	2014	3120	15,870
I4	2015	6029	29,620

Table 2 Preprocessed network data

Node number	1294
Edge number	18,155
Message number	20,000
duration	2012.07.01–2012.11.01

Table 3 Preprocessed network data

Name of data set	YouTube	Digg
Node number	6213	6521
Edge number	19,260	19561

the corresponding contents in a period of time, including original number, forwarding number, total number, attention number, mutual powder number, fan number, information content (including tags and links) and hot topic. Then, according to the attention of the Weibo users, we extract the corresponding data from the original data sets, and then, the preprocessed data are used as the experimental data set, shown in Table 2.

(3) To test the performance of the community detection algorithm on the different data sets, we download the YouTube and Digg data sets, where the preprocessed data are used as the experimental data sets, shown in Table 3.

7.2 Community detection experiment

In the experiment, the network data shown in Tables 1 and 3 are processed and divided to the different communities; thus, we may get the size distribution of the communities. In order to show the compared results clearly, based on the different internal factor value η, the threshold value T and the control parameter ϖ, we can get the more efficient results of community detection on the different influence of the parameters, where T and ϖ are changed at the same time. When we set the personal willingness as 0, 1/2, 1 (maximum), the numbers of the communities and the corresponding community members detected by the proposed algorithm are shown in Figs. 2 and 3.

As we can see from Fig. 2a, the greater personal willingness is, the less the number of detected communities is. It

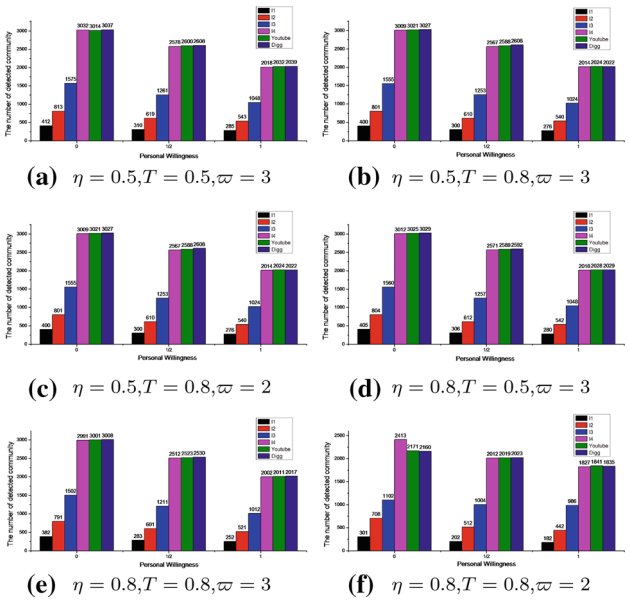


Fig. 2 The number of detected communities

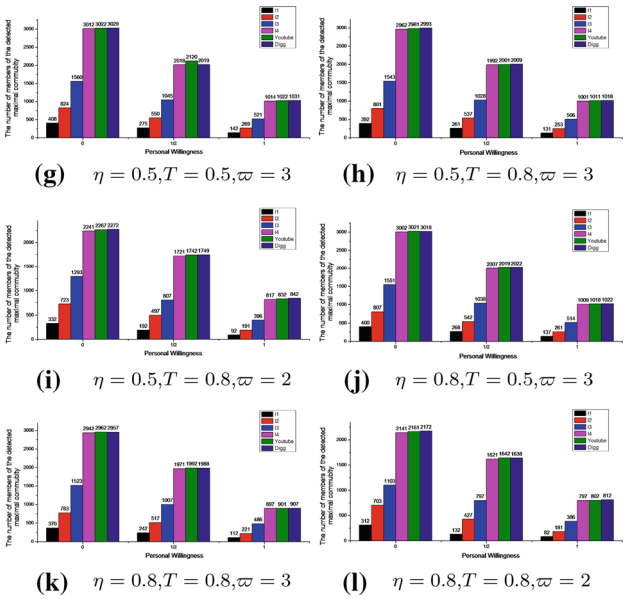


Fig. 3 The number of members of the detected maximal community

indicates that the different personal willingness is influential to the results of detected communities, and detected communities are more stable. As shown in Fig. 2a, d, when the value η becomes larger, the number of detected communities becomes slightly smaller. It indicates that η is lightly influential to the results of detected communities. As shown in Fig. 2a–c, the different values T and ϖ are influential to the results of detected communities. When the values T and ϖ are changed, the number of detected communities becomes smaller. It indicates that T and ϖ are obviously influential to the results of detected communities. Therefore, node inter-

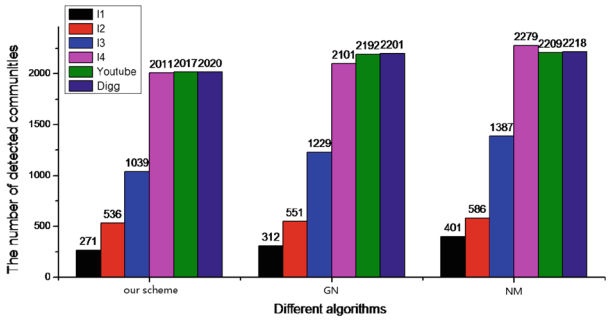
est and personal willingness are influential and important to the results of detected communities. As shown in whole Fig. 2, the different parameters are differently influential to the results of detected communities. When the values η, T and ϖ are changed, the number of detected communities is also changed. It indicates that communities are strictly detected, and detected communities are more stable. Additionally, for the other data sets YouTube and Digg, we can get the similar results as Weibo, and the numbers of communities detected on YouTube and Digg are similar to that of the I4 data set.

As shown in Fig. 3.g, the greater personal willingness is, the less the number of the members of the detected max-size community is. It indicates that when personal willingness becomes larger, the opportunity for generating large communities becomes fewer. As shown in Fig. 3g, j, the value η is changed to incur that the number of the members of the detected max-size community becomes slightly smaller. It indicates that η is lightly influential to the number of the members of the detected max-size community. As shown in Fig. 3g–i, the different values η, T and ϖ are influential to the number of the members of the detected max-size community. When the values η, T and ϖ are changed, as the number of the members of the detected max-size community becomes smaller, the opportunity for generating large communities becomes fewer. Similarly, for the other data sets YouTube and Digg, we can get the similar results as Weibo. Therefore, the community detection algorithm can make community detection more comprehensive and reasonable by introducing node interest and personal willingness. The proposed algorithm may reduce the probability of generating large-scale communities so as to improve the accuracy and reliability of community detection and increase the stability of community structure.

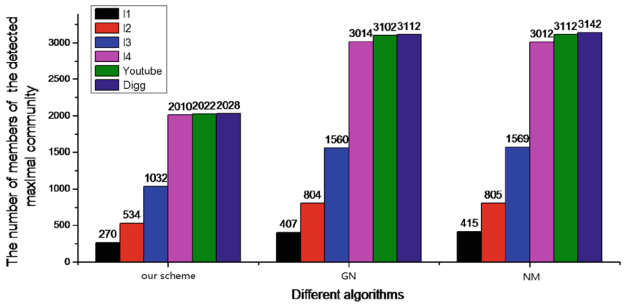
In the experiment, we also compare the GN (Girvan and Newman 2002) and NM (Newman 2004) algorithms with our proposed community detection algorithm by the numbers of detected communities and members of the detected maximal community. The GN algorithm is a community structure discovery algorithm based on edge clustering. The NM algorithm is a community discovery algorithm based on network module degree, which is a greedy algorithm.

As shown in Fig. 4, because our proposed algorithm strictly screens the community members to form the communities, the numbers of communities and members of the maximal community detected by our proposed algorithm are both less than those of the GN and NM algorithms. So, compared with the other two algorithms, our proposed algorithm may reduce the probability of generating large-scale communities and increase the stability of community structure.

According to Table 1, we make experiments for the 4 different scale network data sets. Given $\eta = 0.5$, $T = 0.5$ and $\varpi = 3$, we set personal willingness to the values of 0, 1/2 and



(m) number of detected communities



(n) number of members of detected maximal community

Fig. 4 The performance of different algorithms

1 and then, respectively, calculate the efficiency of the proposed algorithm, where the average degree of the networks is 6, and the number of edges increases continuously from 5000 to 30,000. As shown in Fig. 5, with the increase in the number of edges, the greater personal willingness is, the less the execution time of the proposed algorithm is. Additionally, under the condition of the same number of edges, the greater personal willingness is, the less the execution time of the proposed algorithm is. As shown from the curve of personal willingness to be 1/2, while the number of edges increases, the curve increases relatively quickly at the beginning, and when the number of edges reaches a certain number (2.5×10^4), the increase becomes relatively slow. Then, we may know that when personal willingness is being increased, the increment of the running time of the proposed algorithm decreases with the increase in the number of edges. So, the personal willingness is greater, the efficiency of the proposed algorithm relatively becomes higher.

In summary, whenever we make the same experiments on the data sets Weibo, YouTube and Digg, the experimental results show that the community detection algorithm can make community detection more comprehensive and reasonable because the proposed algorithm introduces node interest and personal willingness to strictly detect community members. Also, compared with the other algorithms (such as the GN and NM algorithms), the numbers of communities and members of the maximal community detected by the proposed algorithm are the fewest. Therefore, our pro-

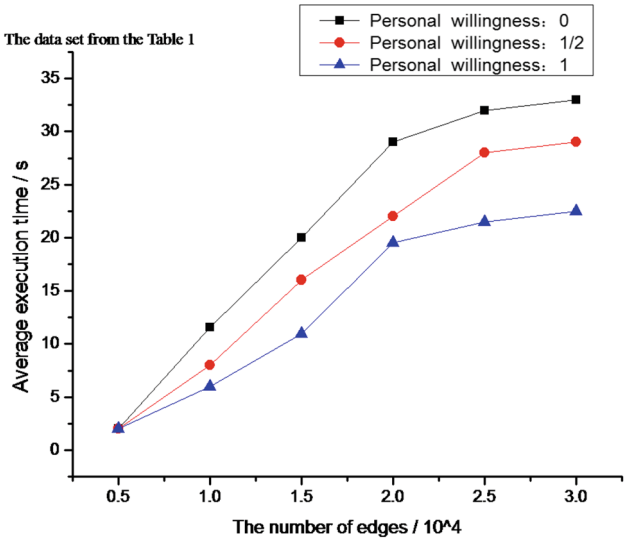


Fig. 5 The computation time of the proposed algorithm with different personal willingness values

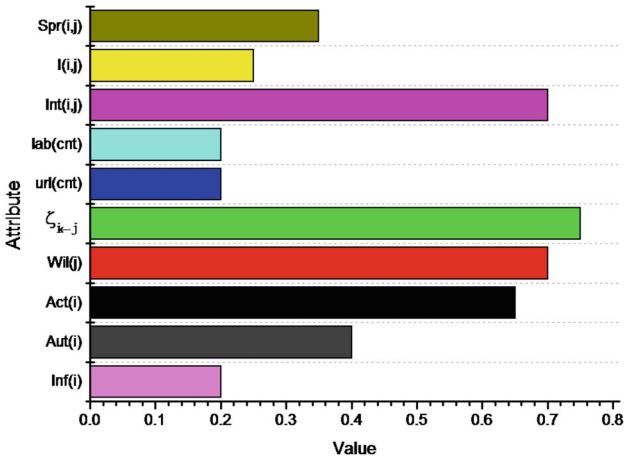
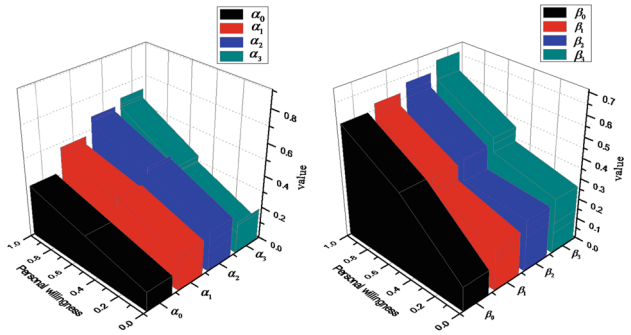


Fig. 6 The comparison of attribute weights in message propagation model



(o) the values of α **(p)** the values of β

Fig. 7 The solutions changing of α and β

posed algorithm may reduce the probability of generating large-scale communities so as to improve the accuracy and reliability of community detection and increase the stability of community structure.

7.3 Message propagation experiment

In the experiment, the message propagation model is tested on the network data from Table 2, where we firstly need to solve the values of α and β by the stochastic gradient descent algorithm. Additionally, because the original data set lacks the attribute of personal willingness, we need to preprocess the data set by randomly setting the values of personal willingness to the original data set. Thus, in order to show the contrast results more clearly, the related weights of the extracted attributes can be got from the experiment, and the values are standardized, shown in Fig. 6. From Fig. 6, we may know that the attributes of larger weight proportion are node propagation characteristic, interest similarity, node propagation willingness and node activity. In the social network of Weibo, the user's interest degree, active degree and propagation willingness will largely determine whether the user can continue to spread the messages. Also propagation characteristics define the characteristics of propagated message, which further show that the users with common interests and larger propagation willingness spread the related messages more possibly.

Figure 7 shows the values of α and β solved by the stochastic gradient descent algorithm when we set the personal willingness to the different values, where $\alpha = (\alpha_0, \alpha_1^T, \alpha_2^T, \alpha_3^T)$ and $\beta = (\beta_0, \beta_1^T, \beta_2^T, \beta_3^T)$. In Fig. 7o, the values of α are increasing with the increasing values of personal willingness, and it denotes that personal willingness may influence the solutions of the message propagation model (where the personal willingness is greater, the objective function of the stochastic gradient descent algorithm converges faster). Additionally, compared with α_0, α_1 and α_2, the α_3 related to personal willingness is changed faster. Figure 7p also shows the similar results. So, from the Fig. 7, we may know personal willingness can influence the construction of message propagation model.

Also, in the tested network data, a certain number of nodes are selected to make the experiment. We randomly select three groups of related nodes (the sizes, respectively, are 5, 10 and 20) in the experiment and number them as the indexes of 1–5, 1–10 and 1–20. According to the relationship between the attributes and the nodes, we extract the attributes of the corresponding nodes and then analyze node activity, node propagation willingness and node propagation characteristic by Figs. 8, 9 and 10. Additionally, we randomly select 6 related nodes to analyze interest similarity between any two nodes by Fig. 11.

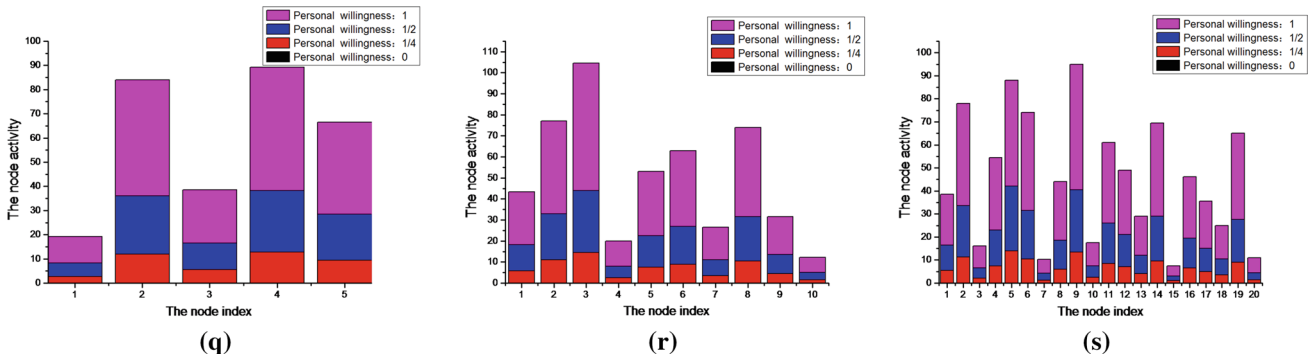


Fig. 8 Node activities with different personal willingness values

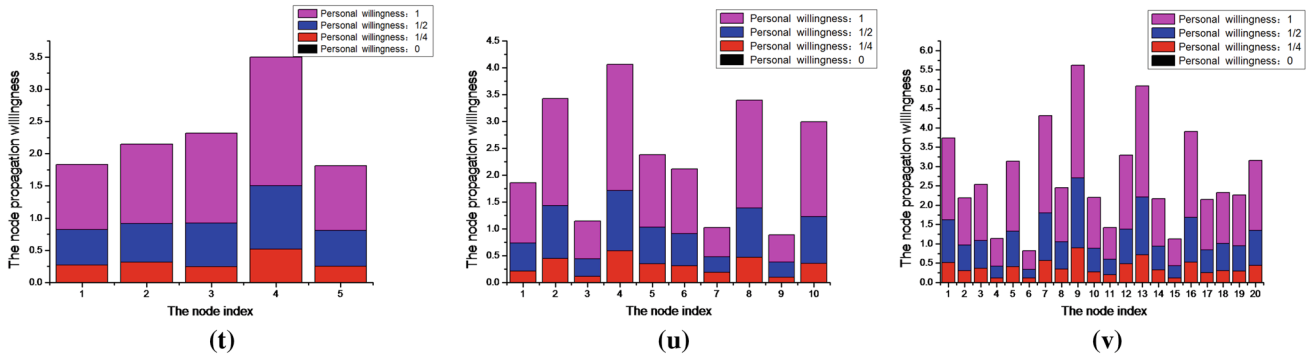


Fig. 9 Node propagation willingness with different personal willingness

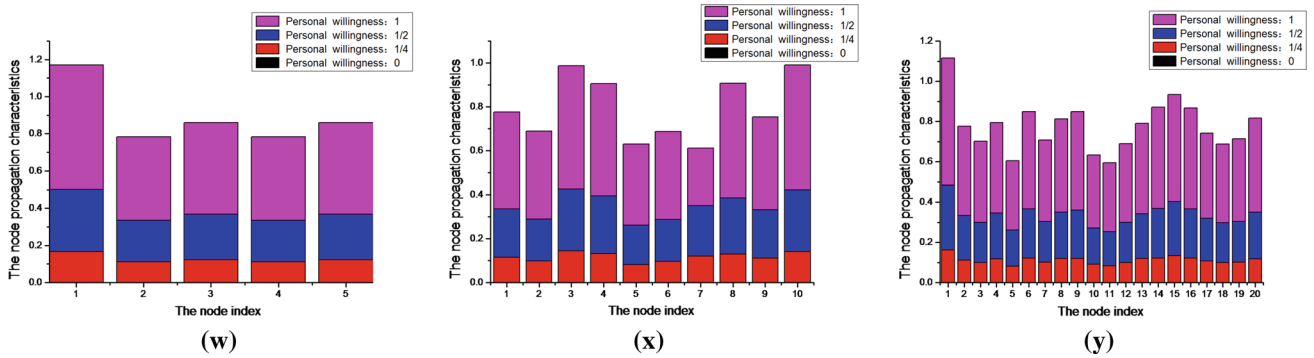


Fig. 10 Node propagation characteristics with different personal willingness

(1) Node activity
As shown in Fig. 8, the black column block with a personal willingness value of 0 is not shown in the diagram, because the node is inactive when the personal willingness has a value of 0. As a whole, regardless of the group sizes being 5, 10 and 20, the greater the value of personal willingness is, the greater the corresponding columnar area is. It shows that personal willingness can affect node activity, where the value of personal willingness is greater, the node is more active. So, personal willingness is influential in the message propagation model.

(2) Node propagation willingness
As shown in Fig. 9, the black column block with a personal willingness value of 0 is not shown in the diagram, because the node refuses to make communication. As a whole, regardless of the group sizes being 5, 10 and 20, the greater the value of personal willingness is, the greater the corresponding columnar area is. It shows that personal willingness can affect node propagation willingness, where the value of personal willingness is greater, the node is more willing to propagate messages.

(3) Node propagation characteristic
As shown in Fig. 10, the black column block with the value of 0 is not shown as the previous figures. From

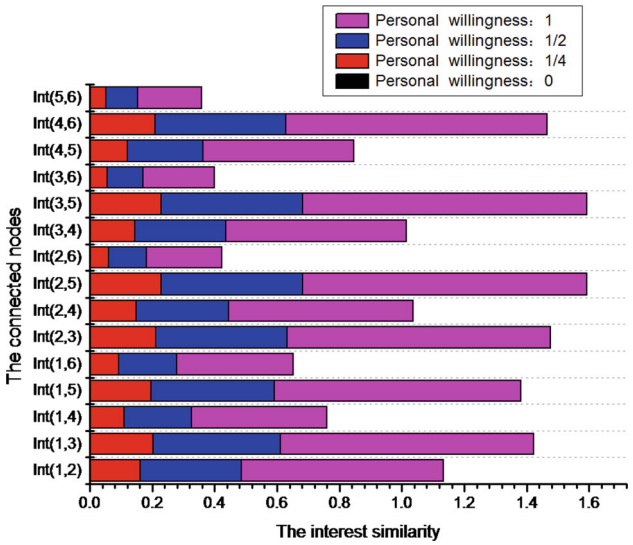


Fig. 11 Interest similarities with different personal willingness

the whole, regardless of the group sizes being 5, 10 and 20, the greater the value of personal willingness is, the greater the corresponding columnar area is. It shows that personal willingness can affect node propagation characteristic, where the value of personal willingness is greater, the value of node propagation characteristic is greater. So, the node with the greater value is more willing to propagate messages.

(4) Interest similarity
According to the definition of interest similarity, we may know that the interest similarity $\mathrm{Int}(i, j)$ denotes the interest similarity between the node i and the node j. From Fig. 11, we may know when the value of personal willingness is 0, the black column block is not shown as the previous figures. Because the two nodes with the personal willingness value of 0 are in a same state refused to outside interest, the interest similarity between the two nodes is no practical significance. As a whole, the greater the value of personal willingness is, the greater the corresponding columnar area is. It shows that personal willingness can affect interest similarity, where the value of personal willingness is greater, the value of interest similarity is greater. So, the nodes with the greater values are more willing to propagate some similar messages.

8 Conclusions

Mining community structure in complex social network, researching on message propagation model, exploring their correlation with personal willingness and deeply analyzing related impact factors are important to the prevention of network crime and the network monitoring of public opinion. Therefore, we propose the social network community detection and message propagation scheme based on personal willingness in this paper: (1) we present the community detection algorithm, which uses module degree with interest degree and personal willingness to make community detection; additionally, based on personal willingness, we use edge intimacy degree and node interest degree as the referred relationship to divide community structure, so as to improve the quality of community detection and reduce the size of large community; the experiment shows that the proposed algorithm based on personal willingness can improve the stability and reliability of online social communities; (2) we present the message propagation model based on personal willingness, in which we extract the characteristics of user attribute and propagated message content to build the model according to propagation probability and propagation delay; the proposed model can take full account of the initiative and effectiveness of users and ensure the quality of message propagation; the experiment shows that personal willingness may influence the construction of message propagation model, where the personal willingness is greater, the objective function of the stochastic gradient descent algorithm converges faster.

Acknowledgements Funding was provided by National Natural Science Foundations of China (No. 61402055, No. 61504013), Hunan Provincial Natural Science Foundation of China (No. 2018JJ2445, No. 2016JJ3012) and Scientific Research Project of Hunan Provincial Education Department (No. 15C0041, No. 12C0010).

Compliance with ethical standards

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

References

Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466(7307):761–4

Amelio A, Pizzuti C (2016) Evolutionary clustering for mining and tracking dynamic multilayer networks. Comput Intell 33(2):181–209

Blondel VD, Guillaume JL, Lambiotte R et al (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10-008

Buccafurri F, Fotia L, Saraswat V et al (2016) Analysis-preserving protection of user privacy against information leakage of social-network Likes. Inf Sci Int J 328(C):340–358

Centola D (2010) The spread of behavior in an online social network experiment. Science 329(5996):1194–1197

Dang T, Viennet E (2012) Community detection based on structural and attribute similarities. In: The sixth international conference on digital society (ICDS), pp 7–12. ISBN:978-1-61208-176-2

Deitrick W, Hu W (2013) Mutually enhancing community detection and sentiment analysis on twitter networks. J Data Anal Inf Process 1:19

Devi JC, Poovammal E (2016) An analysis of overlapping community detection algorithms in social networks. Proc Comput Sci 89:349–358

Fang M, Shi P, Shang W et al (2018) Locating the source of asynchronous diffusion process in online social networks. IEEE Access PP(99):1-1

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Nat Acad Sci 99(12):7821–7826

Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Market Lett 12(3):211–223

Goyal A, Bonchi F, Lakshmanan LVS (2010 February) Learning influence probabilities in social networks. In: International conference on web search and web data mining, WSDM 2010, New York, NY, USA, pp 241–250

Guo L, Ding Z, Wang H (2016) Behavior-based Twitter overlapping community detection [M] database systems for advanced applications. Springer, Berlin

Guo C, Li B, Tian X (2016) Flickr group recommendation using rich social media information. Neurocomputing 204:8–16

Hutair MB, Aghbari ZA, Kamel I (2017) Social community detection based on node distance and interest. In: IEEE/ACM international conference on big data computing applications and technologies. IEEE, pp 274–289

Itakura KY, Sonehara N (2013) Using Twitter's mentions for efficient emergency message propagation. In: International conference on availability, reliability and security. IEEE Computer Society, pp. 530–537

Jaho E, Karaliopoulos M, Stavrakakis I (2011) ISCoDe: a framework for interest similarity-based community detection in social networks. In: Computer communications workshops. IEEE Xplore, pp 912–917

Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: International conference on neural information processing systems. Curran Associates Inc., pp 315–323

Kempe D, Kleinberg J, Tardos (2003) Maximizing the spread of influence through a social network. In: ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 137–146

Kewalramani MN (2011) Community detection in Twitter. University of Maryland, Baltimore, pp 231–300

Kim J, Yoo J (2012) Role of sentiment in message propagation: reply vs. retweet behavior in political communication. In: International conference on social informatics. IEEE computer society, pp 131–136

Lagnier C, Denoyer L, Gaussier E et al (2013) Predicting information diffusion in social networks using content and users profiles. In: European conference on information retrieval. Springer, Berlin, pp 74–85

Lancichinetti A, Fortunato S, Kertsz J (2009) detecting the overlapping and hierarchical community structure in complex networks. N J Phys 11(3):033015

Li K, Lin Z, Wang X (2015) An empirical analysis of users' privacy disclosure behaviors on social network sites. Elsevier Science Publishers B.V, Amsterdam

Lin YR, Sun J, Sundaram H et al (2011) Community discovery via metagraph factorization. ACM Trans Knowl Discov Data 5(3):1–44

Liu G, Li Y (2016) Social-aware data dissemination service in mobile social network with controlled overhead. Pervasive Mobile Comput 33:127–139

Liu L, Chen B, Qu B et al (2017) Data driven modeling of continuous time information diffusion in social networks. In: IEEE second international conference on data science in cyberspace. IEEE, pp 655–660

Newman MEJ, Clauset A (2015) Structure and inference in annotated networks. arXiv preprint arXiv:1507.04001

Newman MEJ (2004) Analysis of weighted networks. Phys Rev E Stat Nonlinear Soft Matter Phys 70(5 Pt 2):056131

Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E Stat Nonlinear Soft Matter Phys 69(2 Pt 2):026113–026113

Ouadrhiri AE, Kamili ME, Rahmouni I (2017) Messages propagation control in delay tolerant networks under epidemic routing protocol. In: International wireless communications and mobile computing conference, pp 1552–1557

Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithms Appl 10(2):191–218

Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106

Rao CS, Raju SV (2016) Similarity analysis between chromosomes of Homo sapiens and monkeys with correlation coefficient, rank correlation coefficient and cosine similarity measures. Genom Data 7(C):202–209

Roux NL, Schmidt M, Bach F (2013) A stochastic gradient method with an exponential convergence rate for finite training sets. Adv Neural Inf Process Syst 4:2663–2671

Saito K, Ohara K, Ohara K et al (2009) Learning continuous-time information diffusion model for social behavioral data analysis. In: Asian conference on machine learning: advances in machine learning. Springer, pp 322–337

Saito K, Ohara K, Yamagishi Y et al (2011) Learning diffusion probability based on node attributes in social networks. In: International symposium on methodologies for intelligent systems. Springer, Berlin, pp 153–162

Shang R, Luo S, Zhang W et al (2016) A multiobjective evolutionary algorithm to find community structures based on affinity propagation. Phys Stat Mech Appl 453:203–227

Shen H, Cheng X, Cai K et al (2009) Detect overlapping and hierarchical community structure in networks. Phys Stat Mech Appl 388(8):1706–1712

Shi C, Cai Y, Fu D et al (2013) A link clustering based overlapping community detection algorithm. Data Knowl Eng 87:394–404

Spiro E, Irvine C, DuBois C et al (2012) Waiting for a retweet: modeling waiting times in information propagation. In: 2012 NIPS workshop of social networks and social media conference. http://snap.stanford.edu/social2012/papers/spiro-dubois-butts.pdf. Accessed 12

Steinhaeuser K, Chawla NV (2010) Identifying and evaluating community structure in complex networks. Pattern Recognit Lett 31(5):413–421

Sun X, Lin H (2013) Topical community detection from mining user tagging behavior and interest. J Assoc Inf Sci Technol 64(2):321–333

Tagarelli A, Amelio A, Gullo F (2017) Ensemble-based community detection in multilayer networks. Data Min Knowl Discov 3:1–38

Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Vorontsov MA (1998) Stochastic parallel-gradient-descent technique for high-resolution wave-front phase-distortion correction. J Opt Soc Am A 15(10):2745–2758

Whang JJ, Gleich DF, Dhillon IS (2013) Overlapping community detection using seed set expansion. In: Proceedings of the 22nd ACM international conference on conference on information and knowledge management, ACM, pp 2099–2108

Xu K, Zou K, Huang Y et al (2016) Mining community and inferring friendship in mobile social networks. Neurocomputing 174(PB):605–616

Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in Twitter. ICWSM 10:355–358

Yang L, Zhang Y, Xing C et al (2011) A node interest similarity based P2P trust model. In: IEEE international conference on communication technology. IEEE, pp 572–575

Yao L, Xiaohui K, Hong G et al (2015) A community detecting method based on the node intimacy and degree in social network. J Comput Res Dev 52(10):2363–2372 (in Chinese)

Young HP (2000) The diffusion of innovations in social networks. Gen Inf 413(1):2329–2334

Zhao Y, Li S, Jin F (2016) Identification of influential nodes in social networks with community structure based on label propagation. Neurocomputing 210:34–44

Zhou D, Han W, Wang Y (2015) A fine-grained information diffusion model based on node attributes and content features. J Comput Res Dev 52(1):156–166 (in Chinese)