TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 10/16 pp729–742
DOI: 10.26599/TST.2022.9010021
Volume 28, Number 4, August 2023

Node Search Contributions Based Long-Term Follow-Up Specific Individual Searching Model

Yayong Shi†, Fei Chang†, Yetao Sun, Guangcheng Yang, Rui Wang*, and Yuan Yao*

Abstract: In this paper, we introduce a long-term follow-up specific individual searching (SIS) model. This model introduces the concept of node search contributions by considering the characteristics of the network structure. A node search contribution indicates the ability of a certain node to correctly guide the search path and successfully complete an SIS. The influencing factors of node search contributions have three components: the individual influence index, attribute similarity, and node search willingness. On the basis of node search contributions and the PeopleRank idea, this paper proposes an SIS model based on node search contribution values and conducts comparison experiments with several mainstream SIS algorithms in three aspects: the search failure rate, the minimum number of search hops, and the search size. The experimental results verify the advanced nature and operability of the model proposed in this paper, which presents theoretical and practical significance to the quantitative study of the SIS process.

Key words: specific individual search; complex networks; propagation dynamics; long-term follow-up

1 Introduction

In recent years, the number of online social network activities has greatly expanded in scope and volume, opening new opportunities for public exposure[1–4]. A specific individual search (SIS) is a type of group behavior initiated on the internet based on a popular social event, in which multiple internet users collect and screen relevant pieces of information to find a specific individual[5–7]. Hundreds of SIS incidents are launched every day, but most of these incidents cannot evoke enough responses. Only a few SIS incidents can attract the attention and participation of most netizens[8]. When a social event arouses wide attention on the internet, it may lead to the emergence of SIS events[9]. Public opinion is divided regarding SIS, terming it a double-edged sword[10, 11]. Some people laud it for providing a channel to seek justice and realize civil superintendence, while others frown upon it as "becoming out of control", with punitive consequences such as detainment by police, loss of employment, cyber violence, and even suicide[12–14]. The frequent occurrence of SIS in recent years has attracted the research interest of scholars from several fields who have made substantial research progress[15–18].

To understand the multivariate search mechanism of a SIS, scholars have proposed different models of search information to describe and simulate the propagation process of search information. From the perspective of modeling methods, it can be divided into structure-based modeling and non-structured-based modeling. Structure-based modeling is mainly used in search path prediction, differential search, and cascade search[18–22]. This method selects a group of active nodes as initiating nodes and simulates the propagation

• Yayong Shi, Fei Chang, Yetao Sun, Guangcheng Yang, and Rui Wang are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100183, China. E-mail: yayongshi16@163.com; changfeifei@outlook.com; 2531888497@qq.com; 2606253647@qq.com; wangrui@ustb.edu.cn.
• Yuan Yao is with Institute for Hospital Management Research, Chinese PLA General Hospital, Beijing 100853, China. E-mail: yaoyuan301@sina.cn.
• Yayong Shi and Fei Chang contributed equally to this work.
∗ To whom correspondence should be addressed.
Manuscript received: 2022-01-22; accepted: 2022-06-16

process of searching for information synchronously with discrete time[23, 24]. In terms of the path prediction of a SIS, most scholars[25–27] mainly establish similarity indicators based on the structural characteristics of the network, such as the nearest neighbor, the degree of nodes, and the degree of community contribution, and judge whether a link relationship exists between nodes by ranking similarity. Heng et al.[28] established three contagious mechanisms driving SIS behavior (i.e., fueled emotionalism, blind acceptance, and collective amnesia) and two noncontagious mechanisms (i.e., isolated dissension and sluggish update), which also provide insights into the technology affordances and constraints. Critical theoretical contributions and important practical implications for SIS initiators, facilitators, and designers of forums have been discussed[28]. These methods are simple in calculation and low in complexity, but they are mostly based on static network structures, ignoring the influence of dynamic changes in netizens' attributes in social networks[1]. On this basis, scholars further introduced multisource information to study dynamic link prediction[29]. For example, Wen et al.[30] proposed a model that explicitly integrates temporal point process theory with the construction of a networked community to describe the dynamics of collective action propagation with seasonal fluctuation. Zhu et al.[31] proposed a propagation link prediction model in dynamic social networks, which can predict propagation links based on the network structure of different time series.

Although these methods can dynamically predict the probability of the search information search between nodes, they only study the overall search trend from a macro perspective, ignore the searchability difference of different participants, and lack in-depth analysis and an important assessment of individual search behavior. Non-structured-based modeling is a process of information dissemination that attracts more users by selecting influential users. On the basis of the traditional independent influence model, scholars proposed a continuous influence node discovery algorithm, which substantially reduced the algorithm complexity. Gao et al.[15] believe that the existing models rely on discovering the implicit structures of diffusion from user behaviors without considering the impact of different diffused contents. To address this issue, they propose a novel model, which was based on information-dependent embedding, to map the users in an observed diffusion process into a latent embedding space, and then the temporal order of users with the timestamps in the cascade can be preserved by the embedding distance of users. However, although the unstructured model method can quantify individual search ability, it ignores specific communication needs, fails to analyze and establish internal connections between searchers, and leads to a selection bias of the search path. To summarize, although the existing research ideas and theoretical tools have a positive reference role in solving the problem of searching for specific groups, some challenges remain. The existing search models only evaluate the search ability of the transit node to the target node through the network structure characteristics when selecting the information transit node. These models consider single factors and lack comprehensive evaluation of transit nodes, resulting in high cost, low efficiency, and unpredictable search results of existing SIS algorithms. To summarize, the innovations of this paper are as follows.

(1) The key factors affecting individual search ability are analyzed, and specific quantitative methods are given.

(2) A node search contribution function based on key factors is proposed to evaluate the search ability of nodes in social networks to the target node.

(3) A SIS model based on node search contribution value is proposed to optimize the search path and improve the search efficiency.

The remainder of this paper is organized as follows. Section 2 describes the theoretical basis and research status of SIS. Section 3 proposes an optimization algorithm of the SIS information search path based on individual communication contribution values. Section 4 verifies the advancement of the algorithm through comparative experiments with several indices. The paper ends with its conclusions in Section 5.

2 Literature Review

2.1 Individual searching and social networks

As the carriers of SIS, social networks are an abstraction that captures the interactions between people relying on internet-based infrastructure[32]. People join online social networks with different goals, such as socializing and staying connected with friends as well as reading and/or sharing news[22]. The ability of every user to propagate information is an important benefit of online social networks, but it also has an adverse effect. Alongside legitimate information, searching for misinformation may have a disruptive effect, including the distrust and unreliability of information[33]. An

online social network is modeled as a graph[34]. The nodes and edges indicate users and the relationships between them, respectively. In this paper, a social graph is denoted as $G = (V, E)$, where $V = \{v_1, v_2, v_3, \ldots, v_{|V|}\}$ and $E \subseteq V \times V$ represent nodes and edges of the graph, respectively. If $e_{ij} \in E$, then a relationship exists between nodes v_i and v_j, and these nodes are called neighbors. For any individual, an ego network[35] with v_i as the central node is obtained by considering only the central node, its friends, and the relationships between them. The network model emphasizes the study and understanding of the structure and function of social networks from the perspective of the individual because individuals understand and participate in social activities essentially by interacting with their direct contacts. If a certain node wants to obtain information fragments related to the target node, it will evaluate one or several individuals most likely to know the information fragments in the egocentric network and send search messages to them, using them as search information relay nodes. This neighbor node can be used as a new ego node, looking for the next round of transit nodes in its ego network, and after multiple ego network interactions, the search is repeated until the SIS process is completed or abandoned.

2.2 PeopleRank algorithm

The structure characteristics of an ego network, position in the network, relationship with neighboring nodes, and the attributes of the node determine the differences in search ability. The amount and accuracy of information obtained by nodes differ between locations, and the search effect on the search information varies greatly. Existing information search models focus on the impact of node attributes on the speed and scale of information dissemination but lack effective guidance on the search direction. Thus, the core of the SIS model studied in this paper is to effectively and accurately evaluate the ability of nodes to search for the target node and then choose the appropriate search direction to end the SIS process as soon as possible. PageRank[36], a Google's classic web ranking algorithm, is based on two main criteria for evaluating the role of a webpage: the number and importance of pages citing the page. The PeopleRank algorithm presented at the 2010 Infocom conference first used the PageRank algorithm in social networks to evaluate the importance of nodes in a social network[37]. The PeopleRank algorithm mainly regards the number of web links as the number of neighbor nodes and evaluates the importance of nodes in the search process by the two attributes of node exit degree and node importance degree. The specific expression is as follows.

$$PeR(Ni) = (1-d) + d \sum_{N_j \in F(N_i)}^{|F(Ni)|} \frac{PeR(N_j)}{\left|F(N_j)\right|} \quad (1)$$

$F(N_i)$ indicates the total number of neighboring nodes connected to the current node. Parameter d is the probability that the neighboring node is willing to contribute the PeopleRank value to the current node.

However, the PeopleRank algorithm has two defects in node search contribution evaluation. First, the PeopleRank algorithm severely discriminates against newly added nodes because the outgoing and incoming connections of newly added nodes in the network are usually very few, and this algorithm only relies on the node outgoing degree as a single measurement standard; thus, the PeopleRank value is very low. In addition, the PeopleRank algorithm relies only on the number and importance of external connections for ranking and ignores the correlation between nodes such that some irrelevant nodes (with target nodes) obtain larger PeopleRank values, which affects the accuracy of the search results. Therefore, a major research point of this paper is to establish an individual communication contribution evaluation system that meets the research objectives based on the research purpose of finding a specific population.

3 Methodology

3.1 Calculation method of key factors

Essentially, a SIS is the process of information dissemination in social networks. The initial node transmits information to the target node. In the process of information transmission, many transfer nodes exist to receive and transfer information. The initiating node, the transfer node, and the target node constitute one or more information transmission paths of a SIS. Many different search paths exist based on different transit strategies. These paths have different lengths, forwarding times, and probabilities of success. To reduce the search path length and forwarding time and improve the probability of search success, we propose a search path optimization method based on searching for specific populations. The method is based on node search contributions to evaluate the search ability in the SIS process. SIS ability refers to the ability to spread information to the target node as soon as possible to optimize the transmission path.

Node v_i is considered more capable of helping find the target node if its node search contribution is higher than those of other transit nodes. Therefore, the challenge of this study is how to quantify the communication contribution of each node and optimize the information transmission path based on this process. In this paper, the key factors affecting individual communication contributions are studied, and the evaluation function of individual communication contributions is constructed by calculating the entropy value.

3.1.1 Node influence index

On the basis of social network topology, this paper maps individuals as nodes and the relationships between individuals as edges to obtain the social network topology graph $G = (V, E)$, where the default is a directed graph, V is the set of nodes in the network, and E is the set of edges between nodes. The node weight is the contribution of this search to finding the target node. As the number of search hops increases, the size of the network grows, and graph G has scale-free network characteristics[37]. In particular, each node in V and each of its neighboring nodes can be considered an independent, and self-centred network. $g = (v, e)$, among which $v \in V, e \in E$. Before the dissemination of the search information, each node in the network is considered an independent potential dissemination participant as well as an existing target node. The network topology is completely unknown to us before the search. In terms of network structure, the influence of a node in that network depends to some extent on the ability of neighboring nodes to propagate in the social network. The search ability of a node is related to its out-degree. The output of a node represents the number of nodes that a node can influence in a social network. In other words, a node is considered to have a strong influence in a network when its neighbor nodes have a high out-degree. Therefore, the influence of nodes in the network is defined as follows.

Definition 1. Node influence index.

$$I_{v_i} = \frac{k_{v_i}}{\sum_{j=1}^{m} k_{v_j}} \tag{2}$$

where k_{v_i} is the out-degree of node v_i, v_j is all the nodes in the central network n_{vk}, and $\sum_{j=1}^{m} k_{v_j}$ denotes the total number of alternative nodes in the central network with ego node v_i. This definition means that among all candidate nodes, the neighboring node with more links should be selected as the next-hop node for an information search.

3.1.2 Attribute similarity

When we search the target node, most of the information we have is related to its attribute description. In this paper, we will refer to the idea of using Jaccard's similarity coefficient to calculate node similarity. In particular, in the Jaccard's similarity coefficient method, the difference between the intersection and the concatenation of two node attributes is used as the similarity value. In contrast, we will focus on finding the node that is near to a specific disconnected person based on its proximity, so we mainly rely on the ratio of the common attributes of two nodes to the attributes of the disconnected person rather than the concurrent attributes. From this approach, the attribute similarity can be specified as the ratio of the intersection of attributes between nodes to the number of attributes of the missing linkers. The definition is as follows.

Definition 2. Attribute similarity.

$$S_{v_i} = \frac{|att\ (v_i) \cap att\ (v_{tar})|}{|att\ (v_{tar})|} \tag{3}$$

S_{v_i} is attribute similarity of node v_i and target node v_{tar}, and $att(\cdot)$ denotes the set of attributes of the nodes.

3.1.3 Node search willingness

Obviously, when a node receives a SIS request, it has two choices: (A) accept and forward the message and (B) receive but do not forward the message. To complete the SIS process, we prefer to disseminate the information for the individual who chooses A. The probability of a node's willingness to forward information depends on the strength of its relationship with another node, which in turn is positively correlated with the frequency of interaction between the two nodes. That is, the higher the frequency of interaction between the two nodes is, the higher the probability that the two nodes are willing to forward information when they receive a SIS request from each other. Therefore, the following definition of node search willingness is given.

Definition 3. Node search willingness.

$$R_{v_i} = \frac{b_{v_i, v_j}}{\sum_{t=1}^{m} b_{t, v_j}} \tag{4}$$

where b_{v_i, v_j} is the number of interactions between nodes v_i and v_j, and t is the neighbor node of v_i. b_{t, v_j} indicates the total number of interactions between all alternative nodes and v_i in the current ego network n_{vk}. On the basis of their definition, the above key factors can be further quantified as the node search contribution of transit nodes and the people-seeking search ability of

transit nodes can be assessed.

3.2 Contribution evaluation function

3.2.1 Construction of the contribution evaluation function based on the PageRank algorithm

In this paper, we borrow the algorithm idea of PageRank and improve parameter d as follows: the similarity and relationship strength can optimize parameter d, and both attributes can form a new parameter d^*. $F(N_i)$ indicates the node search influence, which is still expressed as the connection out-degree of the nodes. In this way, not only is the singularity of relying only on the out-degree of nodes as an assessment factor solved but the potential relationship between the searching nodes and the disconnected individual is assessed to improve the accuracy of finding the disconnected individual. Thus, the evaluation function of individual communication contributions can be expressed as follows.

$$Con(v_i) = \left(1 - d^*\right) + d \times \sum_{v_j \in F(v_j)}^{|F(v_i)|} \frac{Con(v_j)}{\left|F(v_j)\right|} \quad (5)$$

where the parameter d^* consists of the node search willingness and attribute similarity and indicates the probability that a node is willing to participate in the SIS process. $F(v_i)$ denotes the size of the set consisting of all the direct neighbor nodes of node v_i, and v_j is an item in this set. Thus, we evaluate the search ability of nodes in terms of two dimensions: node out-degree $|F(v_i)|$ and node contribution probability d^*. In a two-dimensional coordinate system, the distance from the current point to the origin can be used to describe the distance between the current node and the target node to be found (the reflection of relationship intimacy); the larger this value is, the smaller the distance between the nodes and the higher the probability of finding the target node. The crux of the problem lies in solving the parameter d^*. In Section 3, we define the node search willingness R_{v_i} and the attribute similarity S_{v_i}, whereupon the mathematical function of parameter d^* can be constructed as follows.

$$f(d^*) = \alpha R_{v_i} + (1 - \alpha) S_{v_i} \quad (6)$$

Here, α and $1 - \alpha$ denote the weights of node search willingness and node similarity to specific node attributes, respectively. In the following subsection, the weight values of these two factors are solved.

3.2.2 Solving for parameter values using objective information entropy

The current methods for determining factor weights are divided into three categories: the subjective assignment method, objective assignment method, and combined assignment method. The subjective assignment method is generally used by experts to determine the factor weights based on their experience and the subjective importance of each factor, such as the expert survey method, hierarchical analysis method, and least-squares method. However, these methods have strong subjective arbitrariness and poor objectivity, so the accuracy is unstable, and the application scenario is limited. Given the shortcomings of the subjective assignment method, most of the existing studies focus on the objective assignment method, which mainly determines the weights through certain mathematical methods based on the relationship between the original data. Its judgment results do not depend on specific individual subjective judgment and have a strong mathematical and theoretical basis. Commonly used methods include the entropy method, principal component analysis, and multi-objective planning method. The entropy method is a commonly used weight calculation method whose concept is derived from information theory, which is a measure of uncertainty. The greater the amount of information is, the smaller the uncertainty and the lower the entropy; the smaller the amount of information is, the greater the uncertainty and the higher the entropy. According to the characteristics of entropy, the entropy value can be calculated to judge the randomness and disorder of an event and the dispersion of a certain index: the greater the dispersion of the index is, the greater the influence (weight) of the index on the comprehensive evaluation and the smaller the entropy value. In this paper, on the basis of research data characteristics, we use the objective information entropy method to solve the weight value corresponding to the search intention of nodes and the similarity of node attributes, which is mainly the value of parameter α, and then solve the value of parameter d^*. The specific solution steps are as follows.

(1) Assume that the node number stored in our dataset is m and includes the related information. The two factors (node search willingness R_{v_i} and attribute similarity S_{v_i}) and the corresponding m values are used as a comprehensive evaluation system.

(2) An initial matrix is constructed based on the information related to the m nodes, and because two factors are involved here, an initial matrix $m \times 2$ of R_{ij} can be constructed. Based on this result, the matrix is normalized using by the following equation.

$$R_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}},\ 1 \leqslant j \leqslant 2 \tag{7}$$

where x_{ij} is the value of the element in the i-th row of the j-th column of the matrix. R_{ij} is the ratio of x_{ij} to the sum of the elements in its column, i.e., each matrix element is divided by the sum of the elements in its column to obtain the normalized matrix.

(3) Next, the entropy value of each factor can be solved using the following equation.

$$e_j = -k \sum_{i=1}^{m} R_{ij} \ln R_{ij},\ k = (\ln m)^{-1} \tag{8}$$

(4) Calculating information redundancy. The amount of information provided by the j-th factor is determined by $1-e_j$ (1 the information entropy of the corresponding factor), and the ratio of the redundancy of each factor to the sum of the redundancy of all factors is the value of its weight W_j.

$$W_j = \frac{1-e_j}{\sum_{j=1}^{2}(1-e_j)} \tag{9}$$

From this equation, the values of W_1 and W_2 can be calculated as the corresponding weight values α and $1-\alpha$ for the two factors.

This paper uses the real and public Sina Weibo dataset to calculate the actual weight values of the two factors. This dataset has the information on 636 413 Sina Weibo users, including the user IDs, user nicknames, user names, user locations, user homepage URLs, user genders, user followers, user followings, user tweets, user favorites, and user creation times; there are 1 391 718 user friend relations (connection side), and each record consists of two fields, suid and tuid. The m value of each record consists of two fields: suid and tuid, indicating that suid follows tuid; 27 759 microblog retweet relationships, each record consisting of two fields, smid and tmid, indicating that the smid microblog retweets the tmid microblog; and 84 168 collected microblogs on 12 topics, etc. Thus, the corresponding m value is 63 641, and the final factor weights are calculated as shown in Table 1 (three decimal places are retained).

Table 1 Calculation of the weights of the two factors

	Node search willingness	Attribute similarity
e_j	0.942	0.956
$1-e_j$	0.058	0.044
W_j	0.568	0.432

Then, the value of the parameter d^* can be obtained from Eq. (6) as

$$d^* = f(d^*) = \alpha R_{v_i} + (1-\alpha) S_{v_i} = 0.568 R_{v_i} + 0.432 S_{v_i} \tag{10}$$

Further, based on Eq. (5), the function for evaluating the node search contribution can be obtained as

$$Con(v_i) = (1 - 0.568 R_{v_i} - 0.432 S_{v_i}) + (0.568 R_{v_i} + 0.432 S_{v_i}) \sum_{v_j \in F(v_i)}^{|F(v_i)|} \frac{Con(v_j)}{|F(v_j)|} \tag{11}$$

where $F(v_i)$, R_{v_i}, and S_{v_i} denote the node's outreach, search intention, and attribute relevance, respectively, which correspond to the three key factors affecting individual search behavior previously summarized.

Thus, the $Con(v_i)$ value can be used to evaluate the node search contribution. In other words, it refers to the ability of the nodes involved in a search to find the target node. It quantifies the search capability of each candidate search node in the search process so that the search path can be optimized and the search efficiency can be improved.

3.2.3 A SIS model based on node search contribution values

In the process of a SIS, the search path used to find the target node cannot strictly rely on a single-hop search between two nodes. A single-hop search is a single search between two nodes. The search strategy should be chosen flexibly considering the various factors. These factors include the search failure rate, the minimum number of search hops, and the search scale. After considering the advantages and disadvantages of existing search strategies, this paper proposes a node search contributions based long-term follow-up specific individual searching model (SISM) and verifies the advancement of the algorithm by comparing experiments with several metrics in the following contents. The details of the model are shown below.

(1) Randomly select a finite number of initial nodes to initiate the SIS process.

(2) For each initial node, an independent search contribution threshold, which is the node search contribution value of that node, is set as an ego node of an ego network.

(3) For each ego node, the node whose node search contribution value is greater than the current search contribution threshold is selected as the information search node for the next hop in the list of available

neighbor nodes and set to a new ego node.

(4) Repeat Step 3 until the search path is terminated if one of the following conditions is met.

(a) No neighbor node can be selected, or all neighbor nodes have been involved in the search process.

(b) The node search contribution values of all the neighbor nodes are less than the current search contribution threshold.

(5) If a neighbor node exists in any search path that matches the attribute characteristics of the target node, all search paths are terminated, and the search path is committed.

As shown in Fig. 1, node v_1 is the initial node for finding the target node.

Among all neighbors of node v_1, only nodes v_2 and node v_3 have individual search contribution values greater than that of node v_1, so node v_1 sends the search request to nodes v_2 and node v_3.

Nodes v_2 and node v_3 derive two different search paths: Search path 1 and Search path 2. For all neighbors of node v_2, only node v_5 has a greater individual search contribution greater than node v_2, so the next-hop node is node v_5. Among all the neighboring nodes of node v_5, only node v_{10} and node v_{11} has a greater individual search contribution than node v_5.

After node v_5 sends the search request to v_{10} and v_{11}, no node among all the neighbors of v_{10} and v_{11} can satisfy the required individual search contribution threshold for the next hop, so this search path is aborted. In addition, on Search path 1, v_3 sends a search request to v_4, v_4 sends a search request to v_7, v_8 and v_9 that satisfies the individual search contribution value condition, where all neighbors of v_8 and v_9 do not satisfy the condition, so only v_7 can continue to search the target node. Additionally, among all neighbors of v_7, the attributes of v_{20} match all the characteristics of the

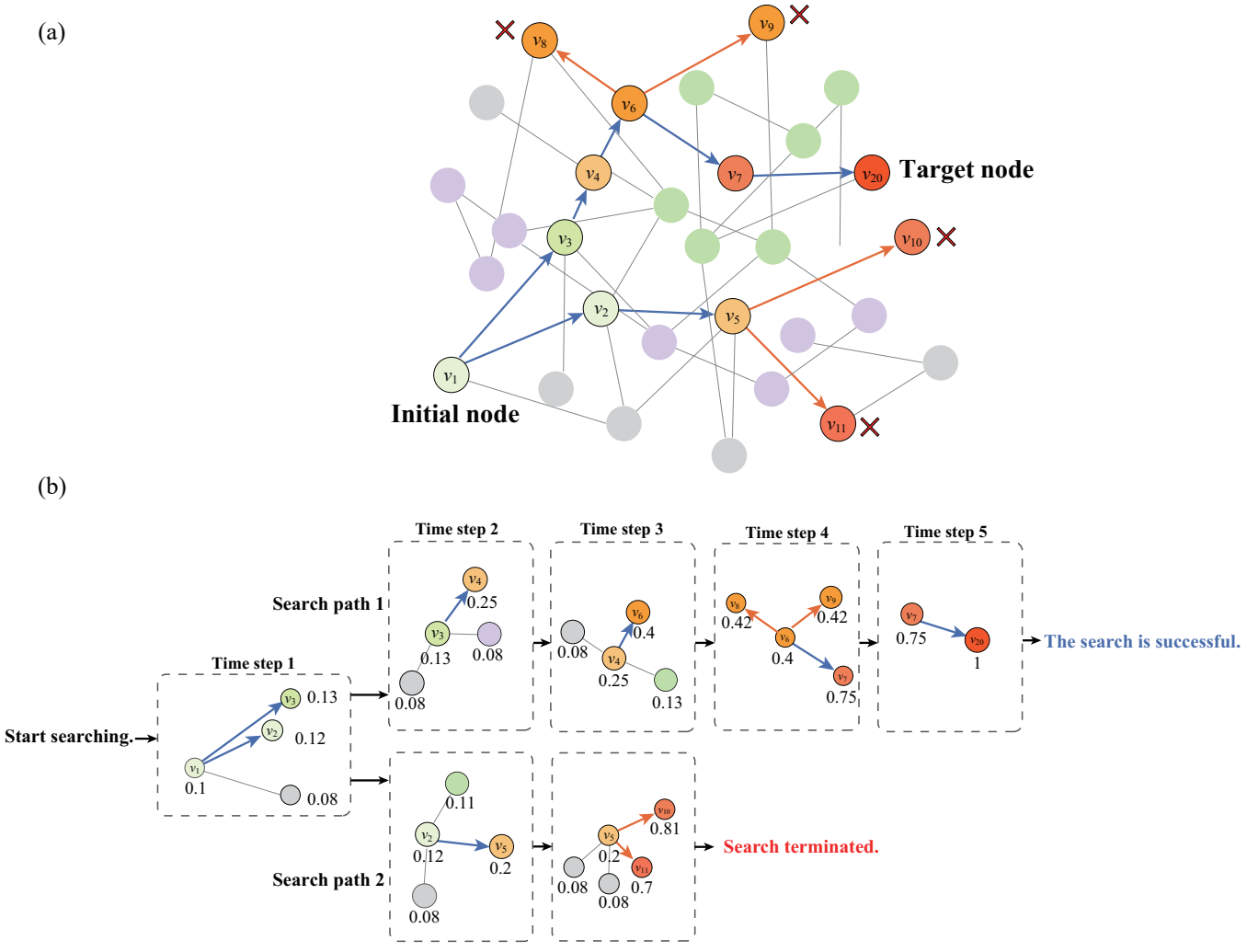


Fig. 1 (a) Search process of the tar node based on search contribution values; (b) Relationship between the transit node selection in ego networks and individual search contributions during the search process.

target node, so the search succeeds and returns the target node location.

4 Experiment

4.1 Experimental environment and dataset

4.1.1 Experimental environment

Programming language: Python 2.7.

Programming tools: Sublime Text3 and Jupyter Notebook.

4.1.2 Experimental dataset

We mainly analyze the publicly available, real social network datasets of Facebook and Google+ provided by Stanford University, which contain the number of connections, attribute characteristics, and the total number of nodes. In the Facebook dataset, the entire social network consists of 10 ego networks, and in Google+, 132 networks, which are connected by anonymous user IDs and their relationships. Table 2 shows the properties of the Facebook and Google+ datasets.

The graph describes the overall structure of the network. An ego network is a set of networks based on direct membership relationships, where the nodes in the network have common attributes (e.g., family, friends, etc.). In fact, the dataset in Google+ contains nodes that have 0 connectivity, which means that they receive information but do not forward it, and we call them inactive nodes. To further investigate the performance of our search model, we analyze all nodes, including inactive nodes.

4.2 Indicators and details of the experiment

In this section, we choose flooding search, random search, word of mouth (WOM) search, and independent cascade (IC) search as the comparison experiments with the model presented above. The following two algorithms represent two typical, realistic ways of searching: a flooding search searches the search information of all friends of each person and conducts a wide search. Obviously, this approach maximizes the success rate of all search methods, and the total number of communication hops needed to find a person must be the smallest, but this approach wastes considerable social resources. In the process of a random search, the participants randomly select neighboring nodes as staging nodes instead of sending the search message to all neighbor nodes. Compared with flooding search, random search is more consistent with the realistic characteristics.

To fully evaluate the search effectiveness of an algorithm, the following performance evaluation metrics are set.

(1) Search failure rate. The search failure probability is the ratio of the number of times the target node is not found to the total number of experiments after multiple trials by a particular SIS algorithm.

(2) Minimum search hops. The number of search hops usually indicates the efficiency of the entire search process. The larger the hop count is, the longer the search time and the higher the search cost. Usually, multiple search paths exist for a single SIS, and each search path has a different search hop count. Therefore, the minimum search hop count is generally used to evaluate algorithm performance.

(3) Propagation scale. Propagation size refers to the number of nodes involved in the SIS process and is also used to indicate the social resources occupied in the SIS process. Apparently, propagation size and search failure rate have a negative correlation under the same search strategy. A flooding search is a case where the search size is the largest. This search strategy is fast but uses considerable social resources and is not relevant for discussion. There should exist a search strategy that can effectively control the propagation size and the minimum number of search hops while ensuring a low search failure rate. In the following content, this paper experimentally verifies the above proposed SIS algorithm. Specifically, we randomly selected different initial search nodes and target nodes and conducted 1000 sets of experiments. Each group of experiments was repeated one hundred times, and we recorded the average of 100 tests as the final result of a group. The comparative results are shown below.

4.3 Analysis of experimental results

4.3.1 Comparison and analysis of the search failure rate

In this experiment, the search failure rate of four

Table 2 Experimental dataset.

Dataset	Node number	Edge number	Ego network number	Statue	Non-active nodes number
Facebook	4039	88 234	10	nondirectional graph	0
Google+	107 614	573 453	132	directional graph	5343

strategies in the SIS process was compared.

To ensure the reliability of experimental results, 1000 groups of experiments were set up. Different initial parameters were set for each group, and each group was repeated 100 times. In Fig. 2, the horizontal axis represents the serial number of each group of experiments, and the vertical axis represents the probability of failing to find the target node in each group of experiments. As shown in Fig. 2a, based on the Facebook dataset, among 1000 groups of experiments, 863 groups and 0 groups of SISM and random strategies successfully completed the SIS, respectively, while 0 groups and 5 groups of experiments completely failed.

Figure 2b shows the experimental results on the Google+ dataset. In the independent 100 tests of all groups, SISM and random found the target node in 735 groups and 0 groups, respectively. Table 3 lists the means and variances of the experimental results of the three search modes (three decimal places are retained). Figure 2c and 2d describes the total number of failures in 100 tests of SISM, WOM, and IC in 1000 groups. As shown, based on the Facebook dataset, in 1000 groups of experiments, SISM, WOM, and IC have 917 groups, 4 groups, and 2 groups, respectively, that found target nodes in 100 tests of each group, while a SIS has been completely successful in 100 tests of 0 groups, 924 groups, and 938 groups.

Table 3 shows that the search failure rate of Random is 25.579 times higher than that of SISM in the Facebook dataset and 13.311 times higher than that of SISM in the Google+ dataset. Table 4 shows that the search failure rate of IC is 13.984 times higher than that of SISM, and the search failure rate of WOM is 13.728 times higher than that of SISM in the Facebook dataset.

The results show that SISM has a much higher search success rate than random, WOM, and IC. However, SISM also failed to find all target nodes. For example, 137 and 265 groups of experiments were declared failures in the Facebook dataset and Google+ dataset, respectively. This failure has two causes: the target node is not closely connected to the neighbor node, and the attributes are not similar, so the node search contribution of the neighbor node is lower than the search threshold, and the search process is terminated. The transit node is an inactive node, which will also terminate the search process and lead to the failure of SIS. Therefore, the SISM search model can be further optimized, such as by using a larger multi-class dataset to calculate the weight of each search influence factor and make it more universal. Once the search capability of nodes is well quantified, appropriate search nodes can be selected to further improve the success rate of a search.

Table 3 Mean and variance of search failure rates based on Facebook and Google+ datasets (SISM, random, flooding).

Strategy	Facebook		Google+	
	Mean	Variance	Mean	Variance
SISM	1.743	28.728	4.030	105.16
Random	51.551	196.027	57.672	382.505
Flooding	0	0	0.006	0.007

Table 4 Mean and variance of search failure rates based on the Facebook dataset (SISM, WOM, IC).

Strategy	Mean	Variance
SISM	6.498	61.6232
WOM	95.700	54.164
IC	97.363	33.456

4.3.2 Comparison and analysis of minimum search hops

The minimum number of search hops is an important index for evaluating algorithm performance. The number of search hops represents the time required to complete a SIS process. To ensure the search success rate, shortening the search time can greatly affect the development trend of relevant social time and avoid greater loss. The experiment compares the minimum number of search hops of five search strategies. In Fig. 3, the horizontal axis represents the serial number of each experiment group, and the vertical axis represents the average minimum number of search hops of each experiment group. As shown in Fig. 3, based on the Facebook dataset, the minimum number of search hops of SISM, random, and flooding was 4–7, 9–13, and 3–6, respectively. Based on the Google+ dataset, the minimum number of search hops of SISM, random, and flooding was 5–10, 7–12, and 4–8, respectively. Figure 4 shows that, based on the Facebook dataset, the minimum number of search hops was 14–16 for WOM and IC. In addition, the mean and variance of the experimental results are shown in Tables 5 and 6, respectively (three decimal places are retained). In the Facebook dataset, the minimum number of search hops of SISM is smaller than that of the random strategy by 78.892%; in the Google+ dataset, 34.487%, and the search efficiency is improved by 44.1%. These results indicate that the transit node selection strategy of SISM is more helpful than that of the random strategy for finding the target node. This finding shows the advanced nature of the algorithm proposed in this paper.

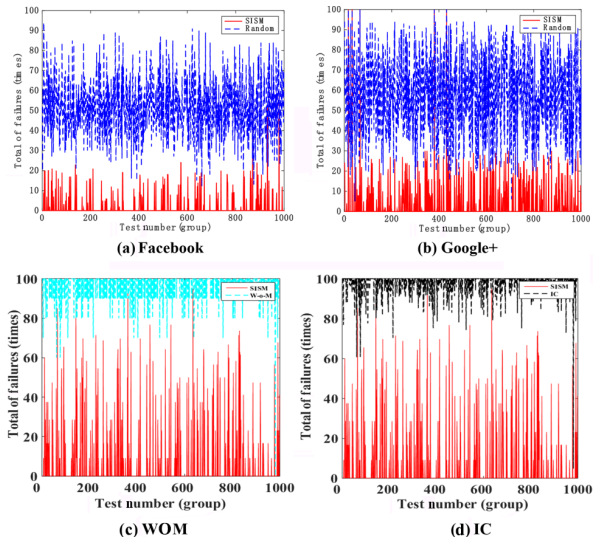


Fig. 2 Comparison of the search failure rates of four search strategies ((a) comparison of the search failure rates of SISM and random strategies based on Facebook dataset; (b) comparison of the search failure rates of SISM and ramdom strategies based on Google+ dataset; (c) comparison of the search failure rates of SISM and WOM based on Facebook dataset; (d) Comparison of search failure rates of SISM and IC strategies based on Facebook dataset.)

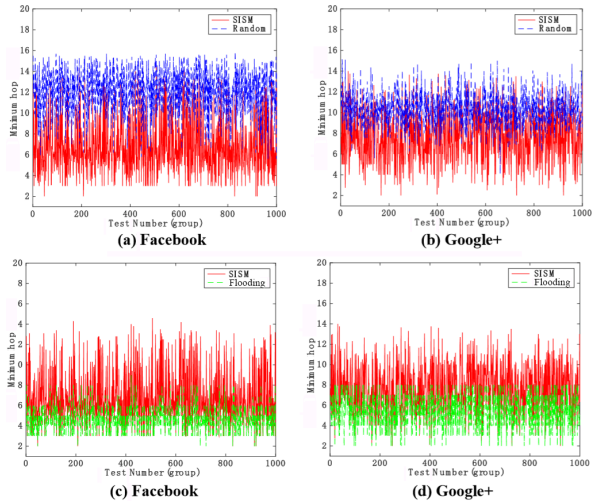


Fig. 3 Comparison of SISM, random, and flooding strategies regarding the minimum number of search hops based on Facebook and Google+ datasets.

Table 5 Mean and variance of the minimum number of search hops based on Facebook and Google+ datasets (SISM, random, and flooding).

Strategy	Facebook		Google+	
	Mean	Variance	Mean	Variance
SISM	6.784	5.831	7.568	5.702
Random	12.136	4.482	10.299	2.672
Flooding	4.665	1.386	5.444	2.585

Table 6 Mean and variance of the minimum number of search hops based on Facebook dataset (SISM, WOM, and IC).

Strategy	Mean	Variance
SISM	5.468	2.254
WOM	15.365	1.038
IC	15.499	0.748

4.3.3 Search scale comparison and analysis

The search scale is an important index for evaluating algorithm performance. The search scale represents the number of nodes participating in a SIS. To a certain extent, the search scale represents the difficulty of completing a SIS and the social resources occupied. As is well known, most SIS fail. This failure is mainly due to the inability to effectively expand the search scope, and a large search scale also means that more social resources need to be occupied. If we can reduce the search scale under the premise of ensuring the success rate, we can improve the success rate to a certain extent, reduce the search cost, and facilitate completing the SIS process.

Figure 5 shows that the search range of SISM is 300–1200 and 500–10 000 based on Facebook and Google+ datasets, respectively, and that of random is 1000–1800 and 3000–15 000 under these datasets, respectively. As shown in Fig. 6, based on the Facebook dataset, SISM's search range is 1000–2500, and WOM and IC's search range is 100–200. This smaller range is due to the activation thresholds of WOM and IC being directly related to the number of edge connections of candidate nodes and the number of connected active nodes. The number of connected active nodes is usually less than 5 in the initial state. Therefore, its activation threshold is barely greater than its random threshold. Therefore, nodes are not easily activated, and, accordingly, the search scale is low. Generally speaking, this type of independent cascade transmission model is not suitable

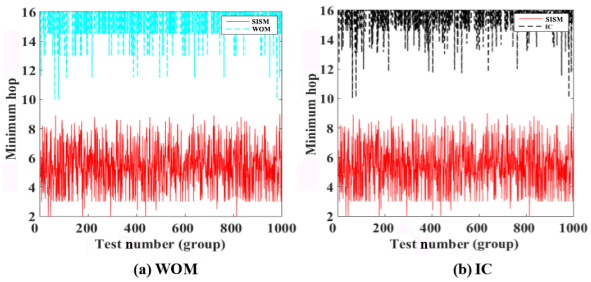


Fig. 4 Comparison of the minimum number of search hops of SISM, WOM, and IC based on the Facebook dataset.

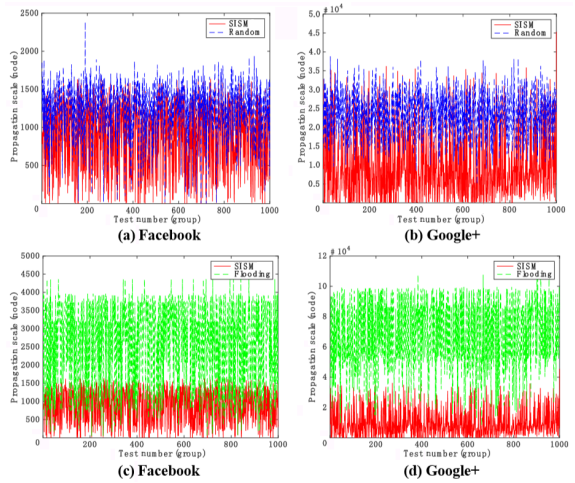


Fig. 5 SISM, random, and flooding spread size comparison based on Facebook and Google+ datasets.

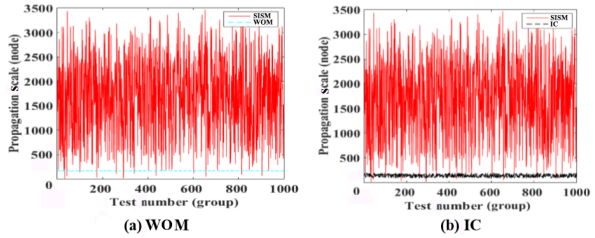


Fig. 6 Comparison of SISM, WOM, and IC search sizes based on the Facebook dataset.

for searching for a specific lost population. Tables 7 and 8 show the mean and variance of the three search modes and the experimental results of the model, respectively (three decimal places are retained).

The statistics in Table 7 and Table 8 show that in the Facebook dataset, SISM policy has a 97.712% and 80.082% better spread scale than the flooding and random policy experiments, respectively. In the Google+ dataset, the SISM policy propagated 136.719% and 679.252% more than the random and flooding policies, respectively. Therefore, the experimental results clearly show the advantages of the SISM strategy. Of course, the search scale was larger than that of the other two experiments (WOM and IC). This difference emerged because, during the SISM strategy search, non-active nodes were encountered in multiple tests because of the changes in the nodes' willingness to participate, but random and flooding simply avoided these inactive nodes. In this case, SISM can lead to unnecessary transmission. The increase in search hops correspondingly makes the search scale expand. In general, SISM outperforms the random and flooding policies in terms of search efficiency, search cost, and search failure rate.

Table 7 Mean and variance of search size based on Facebook dataset (SISM, WOM, IC).

Strategy	Mean	Variance
SISM	1713.83	573.143
WOM	169.373	17.305
IC	138.737	8261.099

Table 8 Mean and variance of the search size based on Facebook and Google+ datasets (SISM, random, and flooding).

Strategy	Facebook		Google+	
	Mean	Variance	Mean	Variance
SISM	871	2013	9943	85 201
Random	1241	14 261	22 538	354 683
Flooding	2643	12 090	77 480	5 193 918

4.4 Validating the model in a real society

This study was derived from the National Key Research and Development Medical Research Priority Special Breast Cancer Cohort Study. Previously, we showed through comparative experiments that our model significantly facilitates SIS in social networks. To verify the search capability of the model in real society, this paper selected data from some of the lost targets based on the breast cancer cohort follow-up dataset constructed in this project and combined it with the SISM model proposed in this paper to develop a search plan and conduct a target-specific search task in two county-level cities (Taixing, Jiangsu Province, and Zoucheng, Shandong Province). The known information of the lost target individual included the name, birth date, height, family members, home address, age, weight, marital status, and phone number. We randomly selected 34 and 22 lost subjects in the Zoucheng and Taixing sites, respectively, for the experiment, and the experimental results are shown in Table 9. We can see that 17 and 16 people were recovered in Zoucheng and Taixing, respectively, with a combined recovery rate of 58.9% (95% CI 45.0%–71.9%) of the lost subjects. These results indicate that our model has a good search effect for finding a lost target individual.

Table 9 Experimental data and result statistics.

City	Lost target individual	Number of successfully located	Rate of successfully located (%, 95%CI)
Zoucheng	34	17	50 (32.4∼67.6)
Taixing	22	16	72.7 (49.8∼89.3)
Total	56	33	58.9 (45.0∼71.9)

As the analysis shows, differences exist between the actual search results and the simulation experimental results for two main reasons. One reason is that in the real social experiment, we obtained very little information about the lost target individual, while the simulation experiment can obtain a large amount of user information, so it is more favorable to find the lost individual. The second reason is the limitation of the grassroots conditions. The search process can only be recorded by qualification materials and then qualitative analysis, while the simulation experiment is a completely quantitative calculation; thus, the effect of the model is better than the real-world social experiment results.

5 Conclusion

We first identified the key factors affecting individual search behavior and borrowed the idea of the PeopleRank algorithm for evaluating the node search contributions. On this basis, a SIS model based on node search contribution values is developed, and the search process is described in detail. Using the publicly available real social network datasets of Facebook and Google+, the proposed SIS model in this paper is compared with other existing search models in three metrics: search failure rate, minimum number of search hops, and search size. The results show that the SIS model proposed in this paper can improve search efficiency by optimizing the search direction. Finally, we also validated the proposed model in real society, and the results show that this model has important practical significance for guiding a real-world search for specific individual.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2016YFC0901303), and the National Natural Science Foundation of China (Nos. 72004147 and 62173158).

References

[1] P. Yang, G. Yang, J. Liu, J. Qi, Y. Yang, X. Wang, and T. Wang, DUAPM: An effective dynamic micro-blogging user activity prediction model towards cyber-physical-social systems, *IEEE Trans. Ind. Inf.*, vol. 16, no. 8, pp. 5317–5326, 2020.

[2] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, Leveraging analysis of user behavior to identify malicious activities in large-scale social networks, *IEEE Trans. Ind. Inf.*, vol. 14, no. 2, pp. 799–813, 2018.

[3] B. Wang, D. Shan, A. Fan, L. Liu, and J. Gao, A sentiment classification method of web social media based on multidimensional and multilevel modeling, *IEEE Trans. Ind. Inf.*, vol. 18, no. 2, pp. 1240–1249, 2022.

[4] Z. Guo and H. Wang, A deep graph neural network-based mechanism for social recommendations, *IEEE Trans. Ind. Inf.*, vol. 17, no. 4, pp. 2776–2783, 2021.

[5] F. Nian and H. Diao, A human flesh search model based on multiple effects, *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1394–1405, 2020.

[6] P. H. Cheong and J. Gong, Cyber vigilantism, transmedia collective intelligence, and civic participation, *Chin. J. Commun.*, vol. 3, no. 4, pp. 471–487, 2010.

[7] K. Zaamout and K. Barker, Structure of crowdsourcing community networks, *IEEE Trans. Computat. Soc. Syst.*, vol. 5, no. 1, pp. 144–155, 2018.

[8] H. Zhu and B. Hu, Agent based simulation on the process of human flesh search—from perspective of knowledge and emotion, *Phys. A Stat. Mech. Appl.*, vol. 469, pp. 71–80, 2017.

[9] G. Li, Y. Liu, B. Ribeiro, and H. Ding, On new group popularity prediction in event-based social networks, *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1239–1250, 2020.

[10] Y. Zhang and H. Gao, Human flesh search engine and online privacy, *Sci. Eng. Ethics*, vol. 22, no. 2, pp. 601–604, 2016.

[11] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri, Leveraging analysis of user behavior to identify malicious activities in large-scale social networks, *IEEE Trans. Ind. Inf.*, vol. 14, no. 2, pp. 799–813, 2018.

[12] W. Craig, M. Boniel-Nissim, N. King, S. D. Walsh, M. Boer, P. D. Donnelly, Y. Harel-Fisch, M. Malinowska-Cieslik, M. G. de Matos, A. Cosma, et al., Social media use and cyber-bullying: A cross-national analysis of young people in 42 countries, *J. Adolesc. Health*, vol. 66, no. 6S, pp. S100–S108, 2020.

[13] K. S. Choi and J. R. Lee, Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory, *Comput. Human Behav.*, vol. 73, pp. 394–402, 2017.

[14] J. Peterson and J. Densley, Cyber violence: What do we know and where do we go from here? *Aggress. Violent Behav.*, vol. 34, pp. 193–200, 2017.

[15] S. Gao, H. Pang, P. Gallinari, J. Guo, and N. Kato, A novel embedding method for information diffusion prediction in social network big data, *IEEE Trans. Ind. Inf.*, vol. 13, no. 4, pp. 2097–2105, 2017.

[16] L. Gao, The emergence of the human flesh search engine and political protest in china: Exploring the internet and online collective action, *Media Cult. Soc.*, vol. 38, no. 3, pp. 349–364, 2016.

[17] L. Y. C. Chang and J. Zhu, Taking justice into their own hands: Predictors of netilantism among cyber citizens in Hong Kong, *Front. Psychol.*, vol. 11, p. 556903, 2020.

[18] N. Kolli and B. Narayanaswamy, Influence maximization from cascade information traces in complex networks in the absence of network structure, *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 6, pp. 1147–1155, 2019.

[19] A. Castiglione, G. Cozzolino, F. Moscato, and V. Moscato, Cognitive analysis in social networks for viral marketing, *IEEE Trans. Ind. Inf.*, vol. 17, no. 9, pp. 6162–6169, 2021.

[20] J. Gamble, H. Chintakunta, A. Wilkerson, and H. Krim, Node dominance: Revealing community and core-periphery structure in social networks, *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 2, pp. 186–199, 2016.

[21] N. Foroutan and A. Hamzeh, Discovering the hidden structure of a social network: A semi supervised approach, *IEEE Trans. Comput. Soc. Syst.*, vol. 4, no. 1, pp. 14–25, 2017.

[22] S. S. Zhang, X. Liang, Y. D. Wei, and X. Zhang, On structural features, user social behavior, and kinship discrimination in communication social networks, *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 2, pp. 425–436, 2020.

[23] G. Ghoshal and A. L. Barabási, Ranking stability and super-stable nodes in complex networks, *Nat. Commun.*, vol. 2, no. 1, p. 394, 2011.

[24] D. Guilbeault and D. Centola, Topological measures for identifying and predicting the spread of complex contagions, *Nat. Commun.*, vol. 12, no. 1, p. 4430, 2021.

[25] P. Basaras, G. Iosifidis, D. Katsaros, and L. Tassiulas, Identifying influential spreaders in complex multilayer networks: A centrality perspective, *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 1, pp. 31–45, 2019.

[26] A. Godoy-Lorite and N. S. Jones, Inference and influence of network structure using snapshot social behavior without network data, *Sci. Adv.*, vol. 7, no. 23, p. eabb8762, 2021.

[27] F. Nian, Y. Shi, and J. Cao, Modeling information propagation in high-order networks based on explicit–implicit relationship, *J. Comput. Sci.*, vol. 55, p. 101438, 2021.

[28] C. S. Heng, Z. Lin, X. Xu, Y. Zhang, and Y. Zhao, Human flesh search: What did we find? *Inf. Manag.*, vol. 56, no. 4, pp. 476–492, 2019.

[29] L. Hu and K. C. C. Chan, Fuzzy clustering in a complex network based on content relevance and link structures, *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 2, pp. 456–470, 2016.

[30] Q. Wen, C. Zhan, Y. Gao, X. Hu, E. Ngai, and B. Hu, Modeling human activity with seasonality bursty dynamics, *IEEE Trans. Ind. Inf.*, vol. 16, no. 2, pp. 1130–1139, 2020.

[31] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan, Scalable temporal latent space inference for link prediction in dynamic social networks, *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2765–2777, 2016.

[32] V. Raghavan, G. Ver Steeg, A. Galstyan, and A. G. Tartakovsky, Modeling temporal activity patterns in dynamic social networks, *IEEE Trans. Comput. Soc. Syst.*, vol. 1, no. 1, pp. 89–107, 2014.

[33] A. Shrestha and F. Spezzano, Online misinformation: From the deceiver to the victim, in *Proc. 2019 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining* (*ASONAM*), Vancouver, Canada, 2019, pp. 847–850.

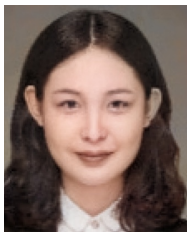
[34] F. Tudisco and D. J. Higham, Node and edge nonlinear eigenvector centrality for hypergraphs, *Commun. Phys.*, vol. 4, p. 201, 2021.

[35] A. Kumar, D. Chhabra, B. Mendiratta, and A. Sinha, Analyzing information diffusion in ego-centric twitter social network, in *Proc. 2020 6^{th} Int. Conf. on Signal Processing and Communication* (*ICSC*), Noida, India, 2020, pp. 363–368.

[36] H. Zhang, B. Liu, H. Susanto, G. Xue, and T. Sun, Incentive mechanism for proximity-based mobile crowd service systems, in *Proc. 35^{th} Annu. IEEE Int. Conf. on Computer Communications*, San Francisco, CA, USA, 2016, pp. 1–9.

[37] A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

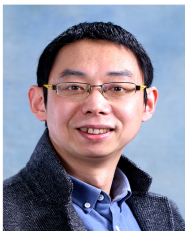
Yayong Shi received the BS degree in engineering from Qilu University of Technology, Jinan, China, in 2017. He received the MS degree in software engineering at Lanzhou University of Technology, Lanzhou, China in 2021. He is currently pursuing the PhD degree in computer science and technology with the University of Science and Technology Beijing. His research interests include edge intelligence, battlefield modeling, medical intelligence, and medical big data analysis.



Fei Chang received the BS degree in computer science and technology from Henan University, Henan, China, in 2016. She received the MS degree in software engineering at University of Science and Technology Beijing in 2019. Her research interests include social networks, and mobile and ubiquitous computing.

Yuan Yao received the PhD degree in Health Management from the Academy of Military Medical Sciences, Beijing, China, in 2015. She is currently a chief technician, Institute for Hospital Management Research, Chinese PLA General Hospital. Her research interests include health management, medical intelligence, mobile health based on social networks and medical big data analysis.

Yetao Sun received the BS degree in computer science and technology from Tianjin Normal University, Tianjin, China, in 2021. She is currently pursuing the MS degree in computer technology with the University of Science and Technology Beijing. Her research interests include the Internet of things and federated learning.



Rui Wang received the PhD degree in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xian, China, in 2007. He is currently a professor with the Department of Computer Science and Technology, School of Computer and Communication Engineering, University of Science and Technology Beijing, China. His research interests include social networks, mobile and ubiquitous computing, and distributed systems.

Guangcheng Yang received the BS degree in in formation and computer science from the University of Science and Technology Beijing, China, in 2018. He received the MS degree in computer science and technology with the University of Science and Technology Beijing in 2021. His research interests include social networks and SIS.