Identification of Key Nodes in Microblog Networks

Jing Lu and Wanggen Wan

A microblog is a service typically offered by online social networks, such as Twitter and Facebook. From the perspective of information dissemination, we define the concept behind a spreading matrix. A new WeiboRank algorithm for identification of key nodes in microblog networks is proposed, taking into account parameters such as a user's direct appeal, a user's influence region, and a user's global influence power. To investigate how measures for ranking influential users in a network correlate, we compare the relative influence ranks of the top 20 microblog users of a university network. The proposed algorithm is compared with other algorithms — PageRank, Betweeness Centrality, Closeness Centrality, Out-degree — using a new tweets propagation model — the Ignorants-Spreaders-Rejecters model. Comparison results show that key nodes obtained from the WeiboRank algorithm have a wider transmission range and better influence.

Keywords: WeiboRank algorithm, centrality measures, ISR model, key nodes, complex networks.

I. Introduction

Key nodes in complex networks are those nodes that have more influence on network structure and features compared to other nodes. For example, in a malicious attack on a scale-free network, attacks on even a small number of key nodes will likely result in the collapse of the whole network.

The mining of key nodes is an important research topic and one that is deeply rooted in the field of information science; together with link prediction and information recommendation, they form the three key research areas in the field of network information mining. In recent years, as studies on network science move from the macro-level to the micro-level, the ordering and mining of key nodes become hot topics. There are many methods designed to assess the importance of network nodes; however, they are all derived from graph theory. In particular, data mining is based on graph theory in the form of finding the most critical nodes and edges within a network.

In research into online social networks (OSNs), nodes that are found to have the most influence in such networks are called opinion leaders. Consequently, influence analysis of these so-called opinion leaders is an important area in OSN research. With respect to the spread of viruses, rumors, or opinions, it is very important to find these relatively more influential nodes.

A microblog is a relatively new type of service typically offered by OSNs. Utilizing WEB 2.0 technologies, a microblog's nodes are formed by individual beings and information between nodes is exchanged in a point-to-point transmission fashion on the web. It is a product designed primarily for the exchanging and spreading of information. A microblog is an Internet application with a strong property of "We media." It changes people's habits of information

Manuscript received Aug. 14, 2015; revised Oct. 5, 2015; accepted Oct. 14, 2015.

This work was supported by the National Nature Science Foundation of China (No. 61373084).

Jing Lu (corresponding author, lujingshiep@163.com) is with the Department of Communication and Information Engineering, Shanghai University, also with the school of Electronics and Information Engineering, Shanghai University of Electric Power, China.

Wanggen Wan (wanwg@staff.shu.edu.cn) is with the Department of Communication and Information Engineering, Shanghai University, China.

acquisition, and becomes an information sharing platform that combines many Internet properties such as instantaneity, convenience, and openness. Within a microblog platform, every user is a broadcaster, listener, and messenger. Using computers, mobile devices, and so on, individual users are able to broadcast whatever they see and feel, at whatever time and place, in the form of a brief message of no more than 140 words [1], a sentence, a photo, a short video, or even a sign of emotion.

In microblog networks, key nodes have the power to influence. According to the Merriam-Webster Dictionary, influence is defined as "the power or capacity of causing an effect in indirect or intangible ways... ." In sociology, its definition is still vague [2]. In this paper, *influence* refers to a user's function on information spread.

Key nodes perform major roles in terms of a network's structure and functions. Therefore, the mining and ranking of key nodes are important research topics — ones that can provide useful guidance for e-commerce and that can affect public sentiment.

Prior to the process of passing information within a microblog network (for example, when there is need to broadcast a message, such as in product marketing or event planning), the choice of a set of appropriate nodes helps both a given piece of information spread and prolongs the duration of such a spread. On the other hand, when there is a need to stop the spread of information, selective immunization of key nodes as well as monitoring and control can prove effective in this regard.

How to identify the key nodes of OSNs has become an interesting problem for researchers. To identify influential nodes on Twitter, Kwak and others [1] ranked users according to the number of followers and found that the result of this was similar to that obtained by using the PageRank algorithm. They also studied the topological characteristics of Twitter and its power as a new medium of information sharing.

Cha and others [2] presented an in-depth comparison of three measures of influence — *indegree*, retweets, and mentions — using a large amount of data collected from Twitter, and observed that popular users who have a high degree of indegree are not necessarily influential in terms of spawning retweets or mentions. Computational models are presented for quantifying social influence, and the corresponding algorithms of social influence analysis are applied to different social network data, such as Twitter, Weibo, and Slashdot [3].

Zhang [4] provided a formal definition for the notion of *social influence locality* and developed two instantiation functions based on pairwise influence and structural diversity. Users' behaviors are mainly influenced by close friends in their ego networks.

Walisa and Wichian [5] defined a number of factors for leadership in social networks and showed that those individuals who are central to social networks serve as opinion leaders. A fuzzy data mining algorithm to find association rules was proposed for analyzing and dividing message posts into quantitative values from which some interesting sequential patterns were obtained.

A new approach based on social network analysis and text mining has been presented [6], [7]; it is capable of detecting opinion leaders in online communities. Cho and others [8] used a social network approach and threshold model to find effective opinion leaders in the diffusion of technological innovation. Li and Du [9] proposed an ontology-based opinion-leader identification framework for word-of-mouth marketing in online social blogs using the information retrieved from blog content, authors, readers, and their relationships. Unfortunately, all of these identification methods do not consider information diffusion rules in OSNs.

Kitsak and others [10] researched a method for ranking influential nodes within a network from a new perspective and found that the most efficient *spreaders* are those located in the inner core of a network, as identified by *k*-shell decomposition analysis. However, it should be noticed that the *k*-shell method assigns many nodes with the same *k*-core value even though they perform entirely differently in the spreading process, and the method itself did not serve well the tree diagram, regular network, and BA scale-free network.

Liu and others [11] presented an improved parameterless method to generate a ranking list to evaluate the spreading influence of any given node; the list could identify the spreading influence of nodes more accurately than those generated from degree, closeness centrality, *k*-shell, or mixed degree decomposition methods. However, this improved method only considered the impact of excess degree when decomposing a network and supposed nodes of the same level to have the same number of neighbors in the outer layer — a supposition that ultimately affects the accuracy of any ranking.

Kang and others [12] proposed a new notion of diffusion centrality in which semantic aspects of a graph as well as a diffusion model of how a diffusive property is spreading are used to characterize the centrality of vertices. The experiments, using real YouTube data, showed the diffusion centrality produced better quality results. However, the implementation method had higher computational complexity. Gao and others [13] proposed a local structural centrality measure to rank the spreading ability of nodes in complex networks, which considers both the number and the topological connections of the neighbors of a node. But, it was found to be only suitable for undirected networks and had some limitations for directed

networks, such as microblog networks.

In this study, we select microblog networks from a number of OSNs as our research object and propose a new WeiboRank algorithm to rank the importance of each given node within these networks. It combines local metrics with global metrics based on information dissemination characteristics in microblog networks. The algorithm can effectively identify key nodes in microblog networks. To evaluate the effectiveness of the proposed algorithm, a new Ignorants-Spreaders-Rejecters (ISR) model is proposed, which is more consistent with the actual model of information propagation dynamics.

This paper is organized as follows. In Sections II and III, we briefly review traditional identification algorithms of key nodes, including the centrality algorithm based on social network analysis and the PageRank algorithm based on WEB search. In Section IV, we introduce the WeiboRank algorithm and the transmission mechanism based on the newly proposed ISR model. The model is used to evaluate and compare the performance of our method with other ranking algorithms, in Section V. Experimental results and discussions are also presented. We conclude this paper in Section VI.

II. Centrality Metrics

Centrality theory [14], in statistics, is used to estimate the statistical parameters of the centralization degree of samples. Within the scope of computer networks, it is given a new purpose; that is, it can be applied to complex networks, such as microblog networks. In our method, a social network is represented by a graph. Entities are represented as network nodes and interactions between entities are represented as edges. In an abstract network structure, in terms of nodes and edges, the research on centrality is of both theoretical and practical value.

1. Degree Centrality (DC)

The DC [15] of a vertex is a simple local measure based on the notion of neighborhood. This measure is useful in the case of static graphs, for situations when we are interested in finding nodes that have the most direct connections to other nodes. Given a vertex, k, the DC of k is calculated as follows:

$$\mathrm{DC}(k) = \frac{d(k)}{N-1}, \tag{1}$$

where N is the total number of network vertices and $d(k)$ is the degree value of vertex k.

A high degree of centrality would indicate a node has a large number of connections with other nodes. Because users within a microblog network participate in two-directional relationships, in terms of representing this in a directed network, the degree of centrality can be expressed in two different ways [16] — in-degree (IDC) and out-degree (ODC). In terms of our method, in-degree is considered to represent the number of followers a user has, and out-degree the number of friends.

2. Closeness Centrality (CC)

CC [15] is a global index that measures the steps it takes for a vertex to contact all other vertices in a network. For instance, assuming a network with N nodes, the CC of node k is defined as

$$\mathrm{CC}(k) = \frac{N-1}{\sum_{i=1}^{N} d_{ki}} \qquad k \neq i, \tag{2}$$

where d_{ki} refers to the length of the shortest path from node k to node i in the same network. Nodes with a smaller total distance $\sum_{i=1}^{N} d_{ki}$ are considered relatively more important. The shorter the average distance of a node to other nodes, the more direct and efficient is the node's influence on other nodes. This is because there exist fewer intermediate nodes with which to relay its messages. We normally think of nodes with high CC scores as being well-positioned to obtain timely information as it arrives.

3. Betweenness Centrality (BC)

BC [15] is a graph analysis technique based on shortest-path enumeration for identifying key individuals in large-scale interaction networks. BC is defined as the share of times that a node, i, needs a different node, k, to reach a further different node, j, via the shortest path. Specifically, let g_{ij} be the total number of shortest paths between vertices i and j, and $g_{ij}(k)$ the total number of these shortest paths that pass through node k. Then the BC of node k is given by

$$\mathrm{BC}^{*}(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{g_{ij}(k)}{g_{ij}} \qquad i \neq j \neq k, \tag{3}$$

where N is the total number of vertices within the network. The normalized index of BC is defined as

$$\mathrm{BC}(k) = \frac{\mathrm{BC}^{*}(k)}{(N-1)(N-2)/2}. \tag{4}$$

The BC of a vertex measures the control that the vertex has over its communication links and its influence on the other nodes. As such, the BC index can be used to identify critical vertices. A higher BC value indicates that the vertex in question can reach other vertices on relatively shorter paths, or that the vertex lies on a significant fraction of shortest paths connecting

pairs of other vertices.

III. PageRank Algorithm

The PageRank algorithm was originally proposed by Larry Page, one of the founders of Google and was used exclusively by Google's search engine. It measures the importance of website pages relative to other website pages. The number of links to a given website page plays an important role.

The main idea behind the PageRank algorithm is that "more important websites are likely to receive more links from other websites" [17]. The PageRank algorithm has two underlying assumptions — a number assumption and a quality assumption. In this paper, the PageRank algorithm is used for microblog networks. Microblog users follow others or are followed. For a given user in the network, A, the user's PageRank value is calculated based on the following two assumptions [17]: (a) the number assumption (that is, the more followers a user receives, the more important that user is) and (b) the quality assumption (that is, important users who follow a user, will pass on more weights to that user). Under the two assumptions above, the steps for the calculation using the PageRank algorithm are as follows:

1) Initialization step: form a relationship network through "follower–friend" by users — every user has the same PageRank value.
2) After certain rounds of recursive calculations, every user will get a final PageRank value. All users will receive their *renew* PageRank values after each round of calculations.

The details of each round of PageRank calculations are as follows:

1) Each user distributes its PageRank value evenly to the outbound links that it follows. In this way, each outbound link will receive the same weight.
2) Each user updates its PageRank value by adding all the weights from inbound links of followers that it receives. After every user in the network updates its PageRank value, that round of calculations is completed.

Mathematically, the PageRank algorithm can be expressed as follows:

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_j)} \frac{\mathrm{PR}(p_j)}{L(p_j)}, \tag{5}$$

where $p_1, p_2, \ldots, p_N$ are the network users, N is the total number of users, $M(p_i)$ is the set of users that link to p_i, $L(p_j)$ is the number of outbound links of user p_j, and d is the damping factor, which is usually set to 0.85.

Note that in the PageRank algorithm, a PageRank value is evenly distributed to outbound links; thus, the algorithm ignores the importance of the user itself.

IV. WeiboRank Algorithm

A flow chart for the WeiboRank algorithm is shown in Fig. 1. After constructing a spreading matrix, we will calculate the WR value of nodes in the microblog networks. Then we will evaluate the performance of the WeiboRank algorithm by ranking nodes based on ISR model.

1. Spreading Matrix

In a computer-based analysis of microblog networks, the first problem encountered is how to represent a network. A microblog user-relationship network represents the *following* and *followed* relationships among users. Because of such a network, information can then spread across the network.

In Fig. 2, the directions of the arrows indicate the relationships between users. That is to say that user C follows user B, and user B follows user A; equally, user C is a follower of user B, and user B is a follower of user A. After user A initiates and transmits a message, M, the message (called tweet) will appear on the microblog homepage of user B who is a follower of user A. After user B reads the tweet and retweets it, the tweet will appear again on the microblog homepage of user C. As such, the corresponding tweet-spread link is given by $M(\mathrm{A}) \rightarrow M(\mathrm{B}) \rightarrow M(\mathrm{C})$. It can be seen that the path of microblog information is in the opposite direction to that which depicts the sense of *following*.

To characterize the relationships between the nodes in

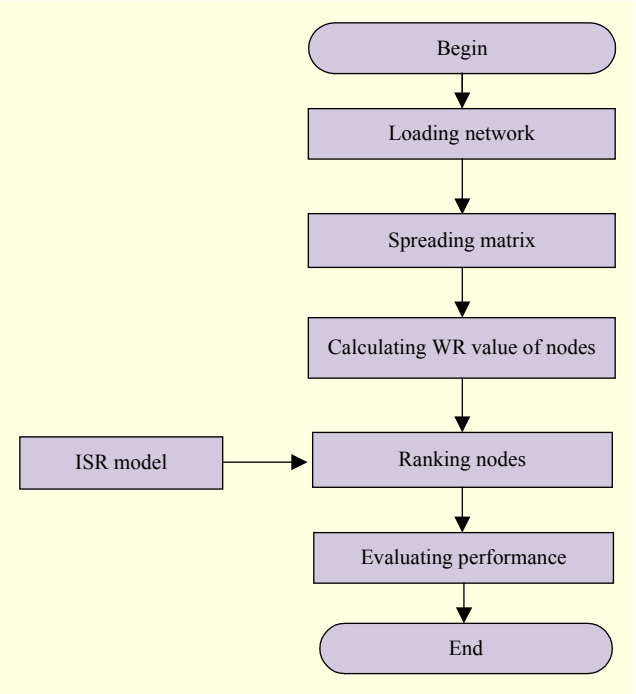

Fig. 1. Flow chart of WeiboRank algorithm.

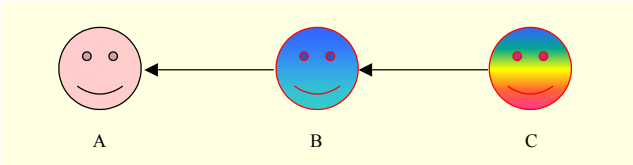

Fig. 2. Simple directed relationship between users.

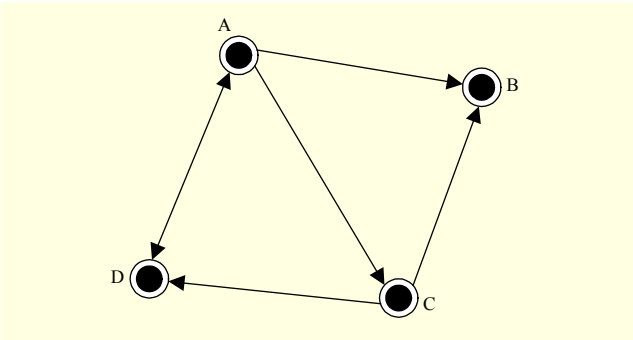

Fig. 3. Simple example of relationship network.

microblog networks, let **B** be an $N \times N$ matrix whose (i, j)th element, b_{ij}, is defined as follows:

$$b_{ij} = \begin{cases} 1 & \text{message flow from vertex } i \text{ to } j, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

So, the spreading matrix corresponding to Fig. 3 is given as

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

In summary, an information-spread matrix for a network can be obtained by taking the transpose of a corresponding relationship matrix for the same network. The actual spread of information is a subset of the information-spread matrix based on some transmission mechanism.

2. ISR Model

Assume there are N nodes in a microblog network and each node represents one registered microblog user. Depending on how users deal with microblog information, users can be divided into *ignorants*, *spreaders*, or *rejecters*. Ignorants are those who do not know of a piece of information; that is, they have not read a specific tweet. Spreaders are those who initiated the tweet, or who have read the tweet and then retweeted the read message. Rejecters are those who are not interested in the tweet; that is, who have read and not retweet the message. Spreaders and rejecters combined can be called *readers*. The initial condition in this model is to set all users to ignorants; that is, to grant them "*I*" status. At some time, a node

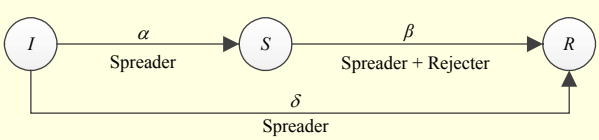

Fig. 4. ISR spreading model.

posts a tweet and changes its status to that of a spreader; that is, they now have "*S*" status. According to the model shown in Fig. 4, this tweet will spread within the microblog network until the network has no spreaders. This status is called terminal status.

Spread-Mechanism Analysis. When a spreader posts a tweet, all of its followers can receive it in real time. Upon reading the tweet, those followers who find it interesting and want to share it with others retweet the same message. This process can be interpreted as that where the ignorant node meets the spreader node and becomes a spreader at a probability of α.

Those followers who read the tweet through following the tweet initiator and find it uninteresting will choose not to retweet it. This process can be interpreted as that where the ignorant node meets the spreader node and becomes a rejecter at a probability of δ. On the other hand, there are those followers that have not yet logged into their accounts and thus will have missed the tweet. As such, we have $\alpha + \delta \leq 1$.

After a user posts or retweets a tweet that has been browsed, the user is considered to have lost their associated retweet value. Because of this loss upon retweeting a post, the node is then deemed a rejector. This process is considered as follows: when a spreader posts or retweets a tweet which has been read by its followers, it becomes a rejector at a probability of β.

3. WeiboRank Value of Node

Before our description of the WeiboRank (WR) algorithm, we define the following parameters:

1) Direct capability of influence, $F_1(v)$: Figure 5 shows the complete topology of a microblog network. In fact, it represents the following relationships amongst users. User v is at the center of the network. According to the distance of information travel, we collect all nodes that are a distance of one or less from the center and form a concentric circle; namely, the N_1 layer. It is clear that the number of nodes represents the number of followers that user v has, which is denoted by $F_1(v)$. The value of $F_1(v)$ indicates the direct capability of influence of user v from the local metric point of view. For instance, if user v has three followers, then the N_1 layer will contain three nodes, and its value of direct capability of influence $F_1(v)$ is equal to three.
2) Region of influence, R: In general, the more nodes connected directly or indirectly to v, the wider the region of

influence along the information path that includes node v. Using the same method as in 1), we collect all nodes that are within a distance of two, and form an N_2 layer. It can be seen that all the nodes in the N_2 layer form the collection of all the followers of the users in $F_1(v)$, denoted by $F_2(v)$. This process continues until the biggest concentric circle, N_M layer, is formed, in which all nodes are *leaf* nodes. The further the information-spread path travels, the more profound the influence of node v. Supposing that a tweet has spread across the whole of a network and all follower nodes have had a chance to read the sent or retweeted tweet from user v, we may define the region of influence of user v by $R = \sum_{i=1}^{M} F_i(v)$. The value R represents the number of nodes that the tweet information sent from user v has reached. In Fig. 5, there are eight nodes in layer N_2; hence, the value of $F_2(v)$ is equal to eight. The region of influence of user v in this instance is thus $R = 3 + 8 + 15 = 26$.

3) A user's WeiboRank value is defined as the product of the value of direct capability of influence and the average information load. Assuming a tweet is sent from user v, its WR value is as follows:

$$\mathrm{WR}(v) = F_1(v) \cdot \sum_{j=1}^{R} d_{vj} / R, \tag{7}$$

where node j is the node that the tweet can reach according to network-spreading matrix **B**, and d_{vj} is the distance between nodes v and j. Note that $\sum_{j=1}^{R} d_{vj} / R$ can be considered as the average information load, which measures a node's power of influence from the global perspective of the whole network. The higher the WeiboRank value is, the greater is the user's influence; hence, its position is deemed to be more critical.

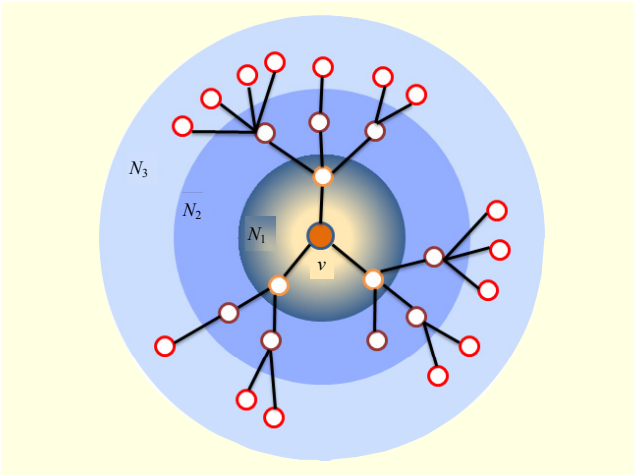

Fig. 5. Complete topology of tweet spread network.

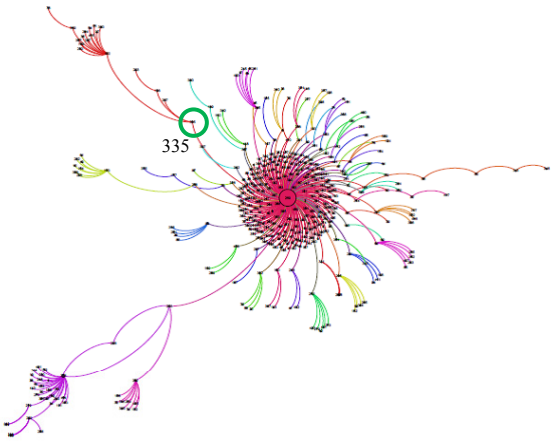

Fig. 6. Transmission network for one tweet.

Below is a simple example illustrating how a WeiboRank value is calculated. Figure 6 shows the transmission network for one tweet. The observed node is the node number 335. The value for direct capability of influence is the number of its followers, which is three, as shown in the figure. Its region of influence R is 17. The average information load for node 335 is $(1\times3+2\times13+3\times1)/17=1.88$. Therefore, node 335 has a WR value of 5.65.

V. Experiment and Results Analysis

1. Measurement of Centrality

To investigate the effectiveness of the proposed WeiboRank algorithm, we need to access a complete, actual data set of some particular community of microblog users; for example, at a university. A Weibo search tool is limited in terms of the number of users obtained and the presented data might not be necessarily true. As such, a Weibo API is used in conjunction with the search tool.

The Sina Weibo application programming interface (API) is an open interface, just like Twitter API. We take the following steps to obtain a university user community at both the institutional and the individual level. Firstly, we select ten institutional users relevant to the university by means of the Sina Weibo search tool. Secondly, we find all of the followers of the ten institutional users by Sina API functions. Finally, amongst these followers, we simultaneously select those users who belong to more than three fan groups. Experiments (to be presented below) show that the choice of three fan groups is an appropriate number considering the total number of available users.

We use Python language to acquire 3,131 valid Sina Weibo users from Fudan University (FDU) and 2,098 users from

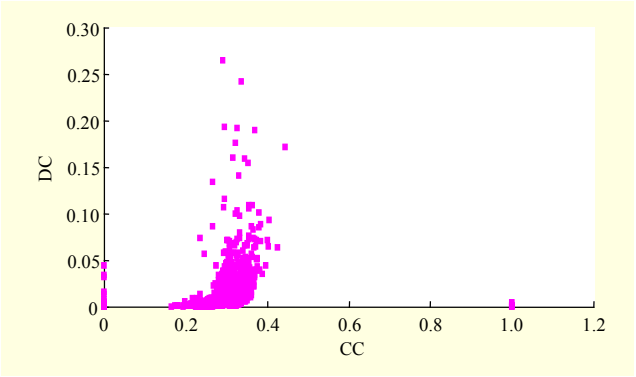

Fig. 7. Corresponding relationship between CC and DC.

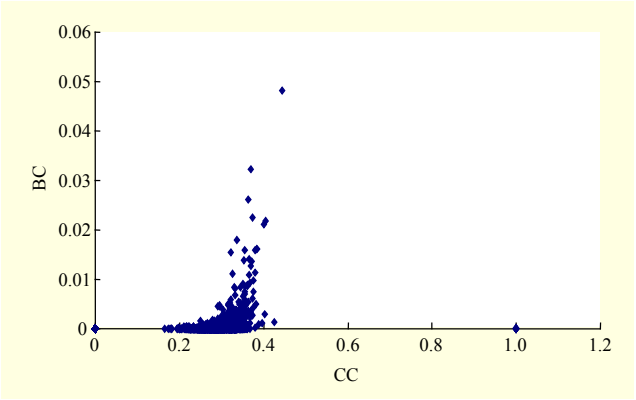

Fig. 8. Corresponding relationship between CC and BC.

Shanghai International Studies University (SISU) through Sina Weibo API functions. These users include both institutional and individual users. The complete user data sample was collected in December 2012. From the graph of FDU (SISU) microblog network, the node number is 3,131 (2,098) and the edge number is 47,376 (46,729). The centrality metrics of all 3,131 nodes of the FDU microblog network are calculated.

Figure 7 reveals the correspondence between CC and DC. The CC value of nodes with large DC value is relatively large, while the distribution of CC of nodes with a small DC value is wide. This indicates that those nodes with a small DC value do not always have a small influence on other nodes in the microblog network. The corresponding relationship between CC and BC, as shown in Fig. 8, is similar to the relationship between CC and DC.

2. Top Ranking by WeiboRank Method

For each user in the microblog network, we calculate its WeiboRank value and rank the results in terms of influence. Table 1 gives the user screen names of the top 15 most important key nodes in both the FDU and the SISU microblog networks. The corresponding analysis of these ranked lists is as follows.

Table 1. Screen name of top 15 users by WeiboRank value.

Ranking	Screen name in FDU	Screen name in SISU
1	Fudan library official microblog	SISU employment guidance center
2	FDU star forum	SISU BBS
3	FDU student union	SISU alumni
4	Fudan UTV	SISU thinking forum
5	Fudan BBS	SISU GSU
6	Foreign affairs office	SISU student union
7	FDU press	SISU young alumni community
8	Shanghai FDU alumni association	Jiang Zhu Zi
9	FDU germanic school	SISU MBA education center
10	FDU youth league committee	SISU young news center
11	FDU MTA project	Happy Wu Ying
12	FDU international politics department	ISUS
13	FDU MPAcc project	SISU JC express
14	FDU alumni association	SISU overseas cooperation school
15	FDU graduate student enrollment	SISU students' association union

Firstly, the institutional microblog accounts within higher education communities have the dominant influence. Specifically, these institutions include the library; student communities; enrollment platforms; university alumni; Open Lectures Room program; and University BBS, MBA Education, Arts departments and colleges — all of which whose tweets have overwhelming fan bases and play key roles in the spreading of information. It is rather ironic that the universities' own official microblogs (for example, FDU microblog account), which have been officially verified, do not have the same level of influence among the students and teaching staff as the previously mentioned institutions, although the official microblogs are perhaps extremely popular among fans outside of the respective universities. As such, the roles of the official microblogs are diminished significantly.

Secondly, smaller-scale, more specialized colleges should pay more attention to the influence of their own microblog accounts, which often represent some of their more well-known "star" teaching staff.

Amongst the Sina Weibo users from the Shanghai University of Foreign Language, a user named 'Jiang Zhu Zi' is a verified user, whose real identity is the deputy dean of the Law School of the university; the user named 'Happy Wu Ying' is also a

Sina Weibo–verified user, whose real identity is an associate professor of the College of Journalism and Communication of the same university. It can be seen that the smaller, more specialized colleges have the more frequent interactions amongst their college staff and students. Stories and wisdom from "star" teaching staff are more likely to spread within communities that share a common background (for example, student communities and classes). In this way, the accumulation of reputation by "star" teaching staff might produce a so-called Matthew effect. In contrast, national and comprehensive universities are of a larger scale; thus, departmental differences are more apparent. This limits, to certain degree, the scope of influence by the "star" staff at such institutions. Consequently, their power of influence is significantly limited.

3. Comparison of Relative Influence Ranks

To find the effectiveness of the WeiboRank algorithm, we compare it with existing well-known measures in terms of relative influence ranks. Rather than comparing different relative influence ranks directly, we use Kendall's coefficient, τ, [18] to show the degree of difference between them. The closer the value of τ is to +1 or –1, the stronger is the likely correlation. Table 2 shows the obtained τ values for the correlation between WR and BC, CC, ODC, or PageRank measure using the user data set as described previously. The results show that the lists of top 5 user ranking by the WeiboRank and PageRank algorithm have a strong correlation and a weak correlation by the WeiboRank and CC measure, respectively. The performance in Table 2 shows that the τ-value for the best-case scenario is 0.6.

Table 2. Kendall τ rank correlation coefficients.

Correlation	Top 5	Top 10	Top 15	Top 20
WR vs. PageRank	0.6000	0.1111	0.2488	0.3219
WR vs. BC	–0.4000	–0.2444	0.1429	0.1263
WR vs. CC	–0.4000	–0.4222	0.0286	–0.0737
WR vs. ODC	–0.4000	–0.5556	–0.1238	–0.1053

4. Comparison Based on ISR Model

In microblog networks, we consider the scope of spread as the criterion for key nodes. Specifically, after a user posts a tweet, how many times the tweet was read in the process of information spread is a feasible criterion. In a microblog network, it is of interest to know how many users, in the end, have read the tweet. In the whole process of tweet spreading, the number of spreaders first increases, then decreases, and then reaches zero when the tweet no longer continues to be spread. At this time, the microblog network reaches a terminal state and has only ignorants and rejectors. The density of the propagation range can be represented by the density of rejectors in the ISR model. Taking a density value of 0.6, as an example, it means that 60% of microblog users have read the tweet in the end.

Table 3 shows the userIDs of the top 20 microblogs using five key-node mining algorithms — PageRank, Betweeness Certrality, Closeness Centrality, Out-degree, and WeiboRank. The values in Table 3 represent the userIDs of microblog users selected from different ranking algorithms. Once a user has successfully registered for a microblog service, their registered account is assigned an ID unique to the whole network; that is, every microblog account has its own unique ID identifying the owner/user. A user's screen name can be changed, but not its ID number associated with the account.

Table 3. Top 20 userIDs using five mining algorithms.

Ranking	PageRank	BC	CC	ODC	WR
1	2094660043	1700066312	2009286791	1700066312	2094660043
2	2062166941	2052377882	2092671335	1729332983	1961500343
3	1979720530	1997205417	2806360034	1666928707	1949791100
4	1944846581	2091852760	2381807622	2005733113	1928683063
5	2052377882	1802587971	2253168707	1997205417	1869954915
6	1961500343	1949418640	2032492724	1802587971	2062166941
7	2499036843	1961500343	2125816775	1847229792	1748944243
8	1896361524	2430380332	1933329405	1949418640	2052377882
9	1869954915	1727640434	1780917364	2761453384	1984915141
10	1690054162	2005733113	2042719623	2870088300	1979720530
11	1874428952	1874428952	2938303623	1441630795	2113303782
12	1949791100	1869954915	1700066312	2586501737	1944846581
13	1802587971	2016427591	1729332983	1884015017	1976281511
14	1764257785	1979720530	1802587971	2605300910	1700066312
15	1960631320	1668524297	1666928707	1760040537	2168710354
16	1833155991	1760040537	1949418640	1704767831	1895633571
17	1925326447	1704767831	2761453384	1727640434	1941191692
18	1732449142	2062166941	2605300910	2091852760	1690054162
19	1621865890	1915525114	2430380332	1901940004	1833155991
20	1987326232	1885375684	1727640434	1006856393	2271643410

Based on an ISR microblog spread model with identical parameters (α = 0.3, β = 0.8, δ = 0.5), Fig. 9 shows the corresponding curves of the density of propagation range, selecting the top k users who initiated tweets. Because the curves for Out-Degree and PageRank are almost on top of each

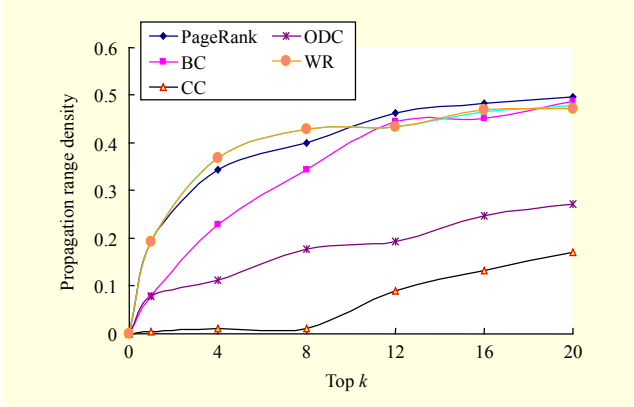

Fig. 9. Comparing performance of five methods.

other for the top k users under consideration, we omit one of them. From Fig. 9, overall, it can be seen that the WR, BC, and PageRank algorithms are clearly superior to the CC and ODC algorithms in terms of the density of scope of spread. The value calculated from the former group of algorithms is much bigger than that from the latter group. For the top ten users, the WeiboRank algorithm is the best, for it produces the largest density value. From the top ten to top twenty users, the Weibo Rank, Betweeness Centrality, and PageRank algorithms produce similar results. In actual applications, k-values are typically not big, $k \leq 10$ in general. Therefore, the WeiboRank algorithm has some very good benefits.

The WeiboRank method is very suitable when it is necessary to reach the largest amount of nodes in a short time. Figure 9 shows that the inclusion of the top 10 nodes is generally enough in terms of propagation scope density. This is because a further increase in the considered number of top nodes does not increase the total number of nodes reached in a significant manner. The PageRank method only considers the first level of followers expressed by the in-degree. The WeiboRank algorithm extends the analysis to include more levels, trying to reach other significant nodes; that is, the PageRank algorithm focuses only on the first set of connections. The WeiboRank algorithm takes into account all the nodes of the network, providing a larger vision of all connections.

VI. Conclusion

This paper presented a new algorithm to identify key nodes in microblog networks. The proposed WeiboRank algorithm combines local metrics with global metrics based on tweet propagation characteristics. To simulate the actual spreading process of tweets, a new transmission dynamics model known as an ISR model was presented. Using an actual data set from the Sina Weibo network, we compared the WeiboRank method with existing well-known methods for ranking a user's relative influence. The results show that there is a strong correlation between the top five users by the WeiboRank and PageRank algorithms. Comparison results included the spreading ability of the top k nodes based on the ISR model. For up to the top 10 users, the WeiboRank algorithm produced the largest propagation range density value.

In microblog networks, the mining of key nodes, which are typically institutional microblog accounts, field experts, and well-known public figures with the capability of detonating "an information explosion," can be applied to the design of strategies for advertisement campaigns and marketing. This is in line with the "individual rule" — one of the three key factors stipulated by Gladwell in the setting off of a fashion explosion.

References

[1] H. Kwak et al., "What is Twitter, a Social Network or a News Media?" *Int. Conf. World Wide Web*, Raleigh, NC, USA, Apr. 26–30, 2010, pp. 591–600.

[2] M. Cha et al., "Measuring User Influence in Twitter: The Million Follower Fallacy," *Int. AAAI Conf. Weblogs Soc. Media*, Washington, DC, USA, May 23–26, 2010, pp. 10–17.

[3] J. Sun and J. Tang, "Models and Algorithms for Social Influence Analysis," *ACM Int. Conf. Web Search Data Mining*, Rome, Italy, Feb. 4–8, 2013, pp. 775–776.

[4] J. Zhang et al., "Social Influence Locality for Modeling Retweeting Behaviors," *Int. Joint Conf. Artif. Intell.*, Beijing, China, Aug. 3–9, 2013, pp. 2761–2767.

[5] R. Walisa and P. Wichian, "Applying Mining Fuzzy Sequential Patterns Technique to Predict the Leadership in Social Networks," *Int. Conf. ICT Knowl. Eng.*, Bangkok, Thailand, Jan. 12–13, 2012, pp. 134–137.

[6] K. Song et al., "Detecting Opinion Leader Dynamically in Chinese News Comments," *Lecture Notes Comput. Sci.*, Wuhan, China, Sept. 14–16, 2011, pp. 197–209.

[7] B. Freimut and K. Carolin, "Detecting Opinion Leaders and Trends in Online Communities," *Int. Conf. Digit. Soc.*, St. Maarten, Netherlands, Feb. 10–16, 2010, pp. 124–129.

[8] Y. Cho, J. Hwang, and D. Lee, "Identification of Effective Opinion Leaders in the Diffusion of Technological Innovation: A Social Network Approach," *Technol. Forecasting Soc. Change*, vol. 79, no. 1, Jan. 2012, pp. 97 –106.

[9] F. Li and T.C. Du, "Who is Talking? An Ontology-Based Opinion Leader Identification Framework for Word-of-Mouth Marketing in Online Social Blogs," *Decision Support Syst.*, vol. 51, no. 1, Apr. 2011, pp. 190–197.

[10] M. Kitsak et al., "Identification of Influential Spreaders in Complex Networks," *Nature Physics*, vol. 6, no. 11, Aug. 2010, pp. 888–893.

[11] J.-G. Liu, Z.-M. Ren, and Q. Guo, "Ranking the Spreading

Influence in Complex Networks," *Physica A: Statistical Mechanics Appl*., vol. 392, no. 18, Sept. 2013, pp. 4154–4159.
[12] C. Kang et al., "Diffusion Centrality in Social Networks," *IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, Istanbul, Turkey, Aug. 26–29, 2012, pp. 558–564.
[13] S. Gao et al., "Ranking the Spreading Ability of Nodes in Complex Networks Based on Local Structure," *Physica A: Statistical Mechanics Appl*., vol. 403, June 2014, pp. 130–147.
[14] N.E. Friedkin, "Theoretical Foundations for Centrality Measures," *American J. Sociology*, vol. 96, no. 6, May 1991, pp. 1478–1504.
[15] S.P. Borgatti, "Centrality and Network Flow," *Soc. Netw*., vol. 27, no. 1, Jan. 2005, pp. 55–71.
[16] A. Mislove et al., "Measurement and Analysis of Online Social Networks," *ACM SIGCOMM Internet Meas. Conf*., San Diego, CA, USA, Oct. 24–26, 2007, pp. 29–42.
[17] V. Kandiah and D.L. Shepelyansky, "PageRank Model of Opinion Formation on Social Networks," *Physica A: Statistical Mechanics Appl*., vol. 391, no. 22, Nov. 2012, pp. 5779–5793.
[18] M.G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1–2, June 1938, pp. 81–93.

Jing Lu received both her BS degree in electrical engineering and her MS degree in information and communication engineering from China University of Mining and Technology, Xuzhou, China, in 2001 and 2004, respectively. She is currently a PhD candidate at the School of Communication and Information Engineering, Shanghai University, China. She has been an associate professor at the Shanghai University of Electric Power. Her current research interests include complex networks, data mining, and machine learning.

Wanggen Wan received his PhD degree in information engineering from Xidian University, China, in 1992. From 1991 to 1992, he was a visiting scholar with the Computer Engineering Department, former Minsk Radioengineering Institute, Belarus, the USSR. He was a postdoctoral research fellow with the Information and Control Engineering Department of Xi'an Jiao-Tong University, China, from 1993 to 1995. From 1995 to 1997, he was an associate professor with the Electronic and Information Engineering Department of Shanghai University china, and was promoted to professor in 1998. He was a visiting scholar with the Electrical and Electronic Engineering Department of Hong Kong University of Science and Technology, Clear Water Bay, China, from 1998 to 1999. He was a visiting professor and section head at the Multimedia Innovation Center of Hong Kong Polytechnic University, Hung Hom, China, from 2000 to 2004. Since 2004, he has been with the School of Communication and Information Engineering, Shanghai University, China, where he is currently a professor, dean of the International Office, director of the Institute of Smart City, and program leader of the Circuit and Systems Department. He is a fellow of the IET and an IEEE senior member. He has been co-chairman for many international conferences since 2008. His research interests include data mining, signal processing, and computer graphics. He is a coauthor of approximately 200 academic papers.