2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)

Detecting Influential Nodes Incrementally and Evolutionarily in Online Social Networks

Jingjing Wang*, Wenjun Jiang*§, Kenli Li* and Keqin Li*
*College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China
Email: wjj_@hnu.edu.cn, jiangwenjun@hnu.edu.cn
§Corresponding Author

***Abstract*—Detecting influential nodes and understanding their evolution patterns are very important for information diffusion in online social networks. Although some work has been done in literature, it is still not clear that: (1) how to measure the influential degree of nodes for information diffusion, and (2) how influential nodes evolve during the diffusion process. To address the two challenges, we identify an incremental approach to measuring users' influential degrees, detecting local and global influential nodes, and analyzing their evolution patterns, for which we propose three methods to partition time window. The three methods are the uniform time window, the non-uniform time window, and the uniform retweets number window, respectively. We apply our model on real data set in Sina weibo and conduct extensive analyses, from which we gain several interesting findings. We also validate the effects of our method, by comparing the influence spread with our detected influential nodes as seeds, to other seed selection algorithms, which shows that our work has better performance.**

***Index Terms*—Evolution patterns, information diffusion, influential nodes, microblogging, online social networks**

I. Introduction

Online social network (OSN) is becoming a popular tool in daily life. Examples include Facebook, Twitter, Sina Weibo and WeChat. There is large amount of information and information diffusion in OSNs. To better understand and even control the diffusion process, it is very necessary to identify influential nodes and understand their evolution patterns. Influential node detection has a directive and practical importance to several applications, including online advertising services [1], recommendation systems [2, 3], influence maximization [4, 5] and so on [6–9].

Although it has attracted many attentions [10–12], two challenges are still open. (1) There is a lack of commonly accepted definition and measurement of the influential degree of nodes. Existing work mainly uses static or quasi-static information, such as the network structure to measure the influential degree (or influence) [13, 14]. (2) It is not clear how influential nodes evolve in the information diffusion process. Some nodes may be influential in some period and be not influential in other periods (we call those local influential nodes). Meanwhile, some nodes may be always influential during all the diffusion process (we call those global influential nodes).

Our work aims to address the two challenges above. Keeping the dynamic characteristic of OSNs in mind, we identify an incremental approach to measuring users' influential degree, detecting local and global influential nodes, and analyzing their evolution patterns. Particularly, we propose three time window partition methods to test their effects. Our contributions are fourfold:

1. We identify an incremental approach to better measuring the influential degree of nodes in dynamic OSNs. To be specific, we provide the exact definitions of local influential nodes, global influential nodes, and incremental influential degree; and we propose an Incremental Influential Nodes Detection (IIND) algorithm according to user behavior in information diffusion.

2. We propose three methods to partition time window in diffusion, so as to explore the evolution patterns of influential nodes. They are the uniform time window, the non-uniform time window, and the uniform retweets number window, respectively.

3. We conduct extensive experiments on real data set in Sina Weibo, from which we find the evolution patterns of influential nodes and we gain some interesting findings.

4. We validate the effects of the proposed IIND algorithm by implementing the independent cascade diffusion model on real data set. Experimental results show that the influential nodes detected by our IIND algorithm leads to a larger influence spread.

The remainder of the paper is organized as follows. Section II introduces related work. Section III states problem definition and solution overview. Time window and the Incremental Influential Nodes Detection (IIND) algorithm are introduced in Section IV in detail. Analysis of experimental results and evaluation are detailed in Section V. Finally, Section VI concludes the paper with future work and directions.

II. Related Work

The study about information diffusion in OSNs is a vital issue, which helps us to control the outbreak of epidemics, understand the dynamics of diffusion, conduct advertisements for e-commercial products and supervise and control of public opinion, and so on. The earliest research on the information diffusion can be traced back to the study of the spread of the disease [15]. With the growing popularity of online social networks, a lot of real data make it easy to study information diffusion. Both academic and industrial have conducted a lot of research and they have made considerable achievements [16–20].

About the study of influential nodes in information diffusion, the researchers make a lot efforts. From the middle of the last century, sociologists have put forward the theory of secondary communication [21], the strength of weak ties [22], the strength of strong ties [23], the structural hole theory [24] and other theories to analyze social influence from different angles. Scientists and academic researchers build models based on these theories to detect the influential nodes of information diffusion [25].

A lot of popular algorithms are mostly based on network topology. The Maximum Degree Heuristic [26] calculates degree of each node in the networks,and chooses the K nodes with the highest degree as the key points, in which the degree of a node is the number of edges incident upon it.

0-7695-6329-5/17/31.00 ©2017 IEEE
DOI 10.1109/ISPA/IUCC.2017.00035

High Clustering Coefficient Heuristic [26] calculates clustering coefficients of the nodes in the networks, and chooses the K nodes with the highest clustering coefficient as the key points, where the clustering coefficient of a node is the ratio of the number of its neighbor nodes connected to each other to the number of its neighbors. Kitsak et al. [27] argued that the location of a node is more significant than its immediate neighbors in evaluating its spreading influence. So, they proposed coreness as a better indicator for a nodes spreading influence, which can be obtained by using the k-core (also called k-shell) decomposition [28] in networks. In addition, the HITs algorithm and the PageRank algorithm are based on iterative refinement centralities of network. These algorithms can detect influential nodes in static or quasi-static network. But they do not consider the user behavior, so they are not applicable dynamic social networks.

Another methods are greedy algorithms. In the references [29, 30], the authors show that the optimization problem of selecting the most influential node is NP-hard. They propose a greedy algorithm, which can get a large influence spread. Till far, there exists a huge number of extensions of the original greedy algorithm by considering the variants of the spreading dynamics [31], by making use of the structural properties of the target networks [32], and so on. However, they are not applicable large social networks, because of their high complexity.

The above algorithms combined with network topology structure, are suitable to the complex networks. However, they are most static or quasi-static, without considering user behavior, and not suitable for this fast-changing online social networks [33]. So, we still lack a dynamic influential nodes detection algorithm based on user behavior and data partition in information diffusion.

From the dynamic angle, the influential nodes are changing in the process of information diffusion. There are evolution patterns of influential nodes. Some nodes are influential in some period and not influential in other periods. But how does the influential nodes change in different time windows and what is the evolution patterns of influential nodes? While these questions has not been studied.

In order to detect well the influential nodes of information diffusion, we give an exact definition of influential nodes, including local influential nodes and global influential nodes. And an Incremental Influential Nodes Detection (IIND) algorithm based on time windows is proposed. Afterwards, numerous experiments are conducted on real Sina Weibo data. We find some interesting findings and evolution patterns about influential nodes from experimental results. Through simulation the diffusion process of information on real Sina Weibo data sets, we demonstrate that the IIND algorithm have good accuracy and performance.

III. Problem Definition and Solution Overview

In this section, we state the problems, and some related definitions. Then, an overview of our solution to solve these problems is introduced.

A. Problem Statement

The tasks of this paper are to detect influential nodes and understand their evolution patterns, which is important for information diffusion in OSNs. There are two major challenges to solve the problem above:

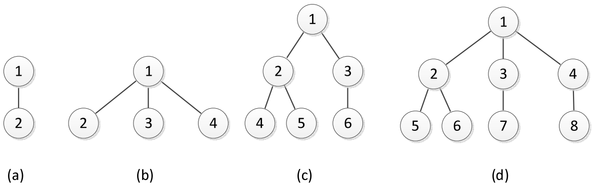


Fig. 1. Examples of propagation tree

(1) How to measure the influential degree of nodes for information diffusion? Some existing methods are based on the network topology [34]. They use the degree or centrality of nodes to measure the influence of nodes. However, these are static and general methods, without user behaviors and not suitable for the fast-changing OSNs.

(2) How influential nodes evolve during the diffusion process? Understanding evolution patterns of influential nodes is necessary for information diffusion trend. But, it has rarely been studied.

B. Related Definitions

Now, we introduce three basic concepts. They are influential nodes, propagation tree, and influential degree.

Definition 1. Influential nodes. In information diffusion, the user who has the potential to spread the information faster and vaster is more influential.

Definition 2. Propagation tree. A propagation tree $T = \langle V,\ E\rangle$ is ordered and directed, which contains $|V|$ vertices and $|E|$ edges. $V = \{0,\ 1,\ ...,\ n\}$ is the set of vertices, in which each $v_i \in V$ represents a user. There is only one distinguished node $v_r \in V$ as the root node. $E\ = \{\langle v_i, v_j\rangle \mid v_i, v_j \in V, v_i \neq v_j\}$ is the set of edges among vertices. Each edge $\langle v_i, v_j\rangle \in E$ is directed, which means v_j retweeted a message from v_i.

It is worth noting that, the propagation tree can well reflect the cascade feature of information diffusion. And it is ordered. The retweet interval of user nodes increases with the increase of height in propagation tree. The children's retweet interval of each node are in a single chronological order from left to right in the propagation tree. Fig. 1 shows the propagation trees of four messages. We take Fig. $1(d)$ for example: the users $2,\ 3,\ 4$ retweeted this message from user 1, and users $5,\ 6$ also retweeted this message from user 2 , and so on. In particular, user 2 retweeted earlier than user $3,\ 4$, and user 3 retweeted earlier than user 4 at the same height in the propagation tree.

Definition 3. Influential degree Influential degree is a measure of influence of user. Specifically, there are centrality degree index, clustering coefficient and so on [11].

C. Solution Overview

In this paper, we propose an Incremental Influential Nodes Detection (IIND) algorithm to detect influential nodes; we also propose three time window partition methods to mine evolution patterns of influential nodes.

There are two steps in IIND algorithm:

1) Detecting local influential nodes all windows from the propagation tree.

2) Detecting global influential nodes for information diffusion.

To be more specific, the IIND algorithm is based on window partition. The user behaviors in information diffusion are also considered in IIND. Combining users' retweeting behaviors, comment behaviors, and like/dislike behaviors, we propose a

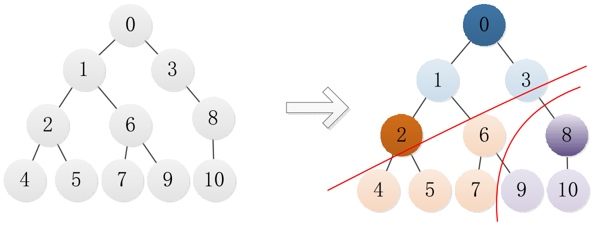


Fig. 2. Process of Incremental Influential Nodes Detection(IIND) algorithm based on window partition in the propagation tree of a microblogging

TABLE I
SYMBOLS TABLE

Symbol	Implication
T	propagation tree
V, E	nodes and edges set of propagation tree T
R, R_i	the retweeted records set in windows, the retweeted records set in ith window
n	the retweeted number of a microblogging
w	the windows number of a microblogging
t, t_i	the time stamp, the retweeting time stamp of node i
Δ, Δ_j	the size of time window, the size of jth time window
η	the size of retweets number for partition window

measurement of influential nodes, that is, incremental influential degree. IIND detects the local influential nodes and global influential nodes based on incremental influential degree. Time window partition methods proposed by this paper can be divided into three categories: the uniform time window, the non-uniform time window, and the uniform retweets number window.

The whole process of IIND based on time window in propagation tree is showed in Fig. 2. Our goal is to detect the influential nodes by changing the propagation tree from the left to the right. In Fig. 2, the propagation tree is divided into three time windows by two red solid lines. Nodes $0, 1, 2, 3$ are newly retweeting nodes in the first time window, nodes $4, 5, 6, 7$ are newly retweeting nodes in the second time window, and nodes $8, 9, 10$ are newly retweeting nodes in the third time window. We distinguish every window with different colors. By computing the incremental influential degree of all nodes in the propagation tree in a time window, we can select the local influential nodes in this time window. In Fig. 2, the dark nodes $0, 2, 8$ are local influential nodes of the 1th, 2th, and 3th time windows respectively.

We will describe the time windows partition methods and the IIND algorithm in detail in Section IV. Besides, the symbols and their implication in this article are shown in Table I.

IV. Time Window and Incremental Influential Nodes Detection (IIND) Algorithm

In this section, three kinds of time window partition methods and the IIND algorithm are be introduced in deteail. In the end, the time complexity analysis of IIND is explained.

A. Time Window

The information diffusion dynamics changes with time and retweeting dynamics in OSNs. We partition time windows from the perspective of time and the number of retweeted respectively. To be specific, we propose three time windows partition methods: the uniform time window, the non-uniform time window, and the uniform retweets number window.

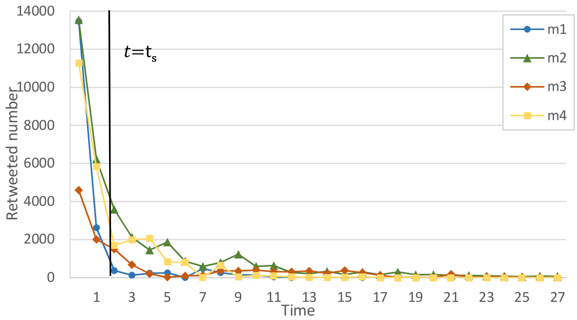


Fig. 3. the retweeting dynamic distribution of information over time

*1) **Method 1:** the uniform time window:* The uniform time window means that the size of time window used to divide the data is fixed.

Therefore, Method 1 can be described as follows. For an original information with n retweeted records, we sort its retweeted records in chronological order. According to the given time window size Δ, $(w+1)$ time points $\{p_0,\ p_1,\ ...,\ p_i,\ ...\ p_w\}$ are generated to split the information data, in which p_0 indicates the original information release time, and the points satisfy the following constraints:

$$\begin{aligned}&p_0 = 0,\\&p_0 < p_1 < ... < p_i < ... < p_w,\\&p_{i+1} - p_i = \Delta(0 \le i \le w),\\&w = \lceil (t_n - t_0)/\Delta \rceil,\end{aligned} \tag{1}$$

where $(t_n - t_0)$ is the retweeting interval of the nth retweeted record. Suppose the records set in the ith time window is R_i, then, the interval of every retweeted record in R_i is between p_i and p_{i+1}. Finally, records sets in all time windows, R_0, R_1,...,R_i,...,R_w, can be obtained.

*2) **Method 2:** the non-uniform time window:* Considering that the diffusion of information is non-uniformly distributed in time, it is challenging to set a proper size of the time window in Method 1. If Δ is too small, the diffusion process will be divided into too many slices, and there may be fewer retweets in some later time windows. This make it difficult to discover the evolution patterns of influential nodes. Conversely, if Δ is too large, the number of time windows will be quite small. The discovery of influential nodes evolution patterns may be skewed. Therefore, we propose two the non-uniform time window methods. The specific process is as follows.

Method 2.1: *the non-uniform time window based on data statistics:* We conduct intensive data analysis and statistics, and we find that generally, there is a turning time point (denoted as t_s) where the retweets number of a piece of information drops rapidly. Fig. 3 shows the change of retweets numbers with time, for four messages m_1, m_2, m_3 and m_4. The information retweeting dynamics is divided into two part by t_s. Therefore, we can set two sizes of time windows, denoted as Δ_1, Δ_2, respectively. The size of the time windows before t_s is Δ_1, and that after t_s is Δ_2. Generally, we have Δ_1<Δ_2. Finally, we apply Method 1 for each part and the time window partition can be completed.

Method 2.2: *the non-uniform time window based on arithmetic sequence:* Because the retweets number of information declines over time, we propose Method 2.2, the non-uniform time window based on arithmetic sequence. Given the original information release time $p_0 = 0$, and tolerance of the arithmetic sequence d, we can generate $(w+1)$ time points, $\{p_0,\ p_1,\ ...,\ p_i,\ ...,\ p_w\}$, to split the information data, where

$$\begin{aligned} & p_{i+1} - p_i = \Delta_i \ (0 \le i \le w-1), \\ & \Delta_0 < \Delta_1 < \Delta_2 < \ldots < \Delta_{(w-1)}, \\ & \Delta_{(i+1)} - \Delta_i = d \ (0 \le i < w-2). \end{aligned} \tag{2}$$

Suppose the records set in the ith time window is R_i, then, the retweeting interval of every retweeted record in R_i is between p_i and p_{i+1}.

*3) **Method 3**: the uniform retweets number window:* The previous window partition methods are proposed based on time. Now we propose a window partition method based on the number of retweets.

Given the size of retweets number η, the retweeted records of a piece of information can be divided into w sets. $|R_i|$ $(0 \le i \le w-1)$ is the records number in the set R_i $(0 \le i \le (w-1))$, then

$$\begin{aligned} & |R_0| = |R_1| = |R_2| = ... = |R_{w-2}| = \eta, \\ & w = \lceil n/\eta \rceil, \\ & |R_{w-1}| \le \eta. \end{aligned} \tag{3}$$

B. Incremental Influential Nodes Detection Algorithm

Firstly, we define the local and global influential nodes. Secondly, according to user behaviors, we model incremental influential degree, an influential measurement method. Finally, the IIND algorithm is introduced in detail. The goal of IIND algorithm is to find influential nodes of information using incremental influential degree, including local influential nodes and global influential nodes.

Definition 4. Local influential nodes. Local influential nodes are users that trigger a large number of diffusion behaviors in a time window.

Definition 5. Global influential nodes. In the set of local influential nodes, some users occur in most windows as local influential nodes, so we call them global influential nodes.

Definition 6. Incremental influential degree. Incremental influential degree of user in a time window refers to its incremental contributions to the information diffusion in this time window.

In OSNs, the influences of users are associated with their attributes and behaviors. For example, the user with certification has a greater influence than other without certification. the user mentioned more time (including being retweeted, being commented and being liked) has a greater influence than other. Based on this fact, we model the Incremental influential degree.

Supposing that, in a propagation tree, the number of new incremental nodes is r^i from ith window to $(i+1)$th window, and a node v have r_v^i new children from ith window to $i+1$th window, the incremental influential degree f_v^i of user v in ith time window is as follow:

$$\begin{aligned} f_v^i = & \alpha Ratio^i_{retweeted}(v) + \beta Ratio^i_{commented}(v) + \\ & \gamma Ratio^i_{liked}(v) + g(v), \end{aligned} \tag{4}$$

where α, β, and γ are the weight of each component, which are given in the experimental section. These functions $Ratio^i_{retweeted}(v)$, $Ratio^i_{commonted}(v)$ and $Ratio^i_{liked}(v)$ are the ratio of the number of retweeting, comment, liking inspired by user v in ith time window to total number of retweeting, comment, liking in ith time window, respectively. And the function g(v) expresses the certification status of user v and other users retweeting his information. Finally, their concrete calculating methods are as follows.

$$Ratio^i_{retweeted} = \frac{r_v^i}{r^i}; \tag{5}$$

$$Ratio^i_{commented} = \frac{\sum_{j=1}^{r_v^i} c_j}{\sum_{j=1}^{r^i} c_j}; \tag{6}$$

$$Ratio^i_{liked} = \frac{\sum_{j=1}^{r_v^i} l_j}{\sum_{j=1}^{r^i} l_j}. \tag{7}$$

Equation (5),(6) and (7) explain the contribution of user v in i time window in the aspects of the retweeting behaviors, comment behaviors and liking behaviors. In Equation (6), c_j is the number of comments generated by the jth retweeting user. In Equation (7), l_j is the number of liking generated by the jth retweeting user. In addition, Equation (8) and (9) explains the impact of user attributes on influence. Intuitively, h(v) constrains the model to bias towards the representative nodes with certification. More specially, a representative node must has certification. $g(v)$ has a maximum of 1, and a minimum of 0. The value of $g(v)$ is depend on certification status of user v and his retweering users, and it effectively avoids prejudice. If user v is certified, and all his retweeting users with certifications, the $g(v)$ is 1. If user v is not certified, or all his retweeting users without certifications, the $g(v)$ is 0.

$$g(v) = h(v) \times \frac{\sum_{j=1}^{r_v^i} h(j)}{\sum_{j=1}^{r_v^i} 1} \tag{8}$$

$$h(v) = \begin{cases} 1, & user\ v\ with\ certification \\ 0, & other \end{cases} \tag{9}$$

Now, the IIND algorithm is introduced In detail. The IIND algorithm have two steps: local influential nodes detection and global nodes detection. The results of detection local influential nodes are the candidate nodes of global influential nodes. We describe these two steps in the following.

The process of detection local influential nodes in all windows from the propagation tree is shown in Algorithm 1. First, the propagation tree T is partitioned by window partition method M, and local influential nodes set L is empty(step 1). Next, this algorithm do the following for each time window i (step 3 to step 9): (i) for every node in propagation tree, calculating its incremental influential degree in ith time window (step 4, 5); (ii) selecting the top k nodes with maximum incremental influential degree, and adding these nodes to set L_i as local influential nodes of ith window (step 6 to step 8). (iv) merging set L_i into set L (step 9). Finally, this algorithm return the local influential nodes set L.

We select global influential nodes from the candidate local influential nodes. The number of global influential nodes depends on the information. That is important, because the number of global influential nodes directly affects the propagation characteristics, which are described in detail later.

Algorithm 2 is to detect global influential nodes from the candidate local influential nodes. Firstly, the global influential nodes set G is initialized empty, which indicates the fact that each candidate node is not added to G (step 1). Next, this algorithm do the following for each node l in the L (step 3 to

Algorithm 1 Local influential nodes detection

```
Input:
    $T$: the propagation tree for a given information;
    $M$: a window partition method to partition the propagation tree;
    $k$: the number of local influential nodes in per time window.
Output:
    $L$: the set of local influential nodes all windows.
 1: partition the propagation tree $T$ to $w$ windows by the method $M$;
 2: $L \leftarrow \emptyset$;
 3: for $i = 1$ to $w$ do
 4:     for user node $v$ in the propagation tree $T$ do
 5:         calculate the incremental influential degree $v.d$ of $v$ in the $i$th window;
 6:     for $j = 1$ to $k$ do
 7:         select node with the $j$th maximum incremental influential degree in propagation tree, denoted $a$.
 8:         add node $a$ to the local influential nodes set $L_i$;
 9:     $L \leftarrow L \cup L_i$;
10: return $L$
```

Algorithm 2 Global influential nodes detection

```
Input:
    $L$: the local influential nodes set in all time windows;
    $w$: the number of windows;
    $\lambda$: a threshold parameter.
Output:
    $G$: global influential nodes set.
1: $G \leftarrow \emptyset$;
2: for each local influential node $l \in L$ do
3:     calculate the number of $l$ occurrences as local influential node all windows $l_n$;
4:     $l_f \leftarrow l_n/w$;
5:     if $l_f \geq \lambda$ then
6:         $G \leftarrow G \bigcup l$;
7: return $G$
```

step 6): (i) calculating the number l_n of l occurrences as local influential nodes all windows; (ii) calculating the frequency of l as local influential nodes l_f; (iii) examining whether the l_f greater than or equal to λ. If the l_f satisfies this condition, then the user node l is added to G. The end conditions of this algorithm is that all user nodes in L are visited and returns a global influential nodes set G.

C. Time Complexity Analysis

The computational complexity of IIND can be analyzed from the section A. Suppose that there are N nodes in the propagation tree T, the number of time windows w and the number of local influential nodes k. In the process of local influential nodes detection, the computational complexities of calculating the incremental influential degree of nodes, and determining the $k$local influential nodes are $O(w \times N)$ and $O(k \times N)$. In general, the number of time windows is larger than the number of local influential nodes in a time window. Hence, the computational complexity of local influential nodes detection is $O(\max\{w \times N, k \times N\}) = O(w \times N)$. In the global influential nodes detection, the computational complexity of counting the occurrences of every local influential node in all time windows is $O((k \times w)^2)$. To sum up, the overall computational complexity of IIND is $O(\max w \times N(k \times w)^2)$ It should be noted that the parameters k and w much less than N in large-scale networks.

TABLE II
SUMMARY DATA INFORMATION OF THE THREE TOPICS EVENTS

Hot topics	Days	Original tweets	Retweeted records	Follower networks
Yang Jiang event	13	40860	1035785	40860
Baidu event	49	21408	205606	21408
H hotel event	12	6035	547102	6035

V. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we carry out experiments on real data set. By analyzing experiment results, we discover the evolution patterns of influential nodes. Finally, we evaluate the efficiency of IIND algorithm with simulation the diffusion process on information networks.

A. Experimental Setup

1) Data Set: To track information diffusion in online social networks, we crawl data from Sina Weibo, the most mainstream and the most popular microblogging platform in China. We selected three the most popular topics in 2016 ,the death of Yang Jiang, Baidu event and H hotel event. To capture tweets related to these three topics, we first identified the set of keywords describing the topics by consulting news web sites and informed individuals. Then we grabbed relevant original tweets by searching for the keywords on Sina Weibo platform. The time period of each topic begin with the moment the event occurrence, until there is no relevant microblogging about this topic in Sina Weibo platform. Afterwards, for each original microblogging, we grabbed its retweeted records and relationships network. Because there were network constraints, interruptions, repetition and other phenomenons in this process, we carry out a data cleansing process to improve the data quality. Finally, the detailed information of data set is shown in Table II.

In data preprocessing phase, we filter out the microbloggings with less than 500 retweeted records. In addition, for each microblogging in data set, we create a propagation tree T with its retweeted records set R.

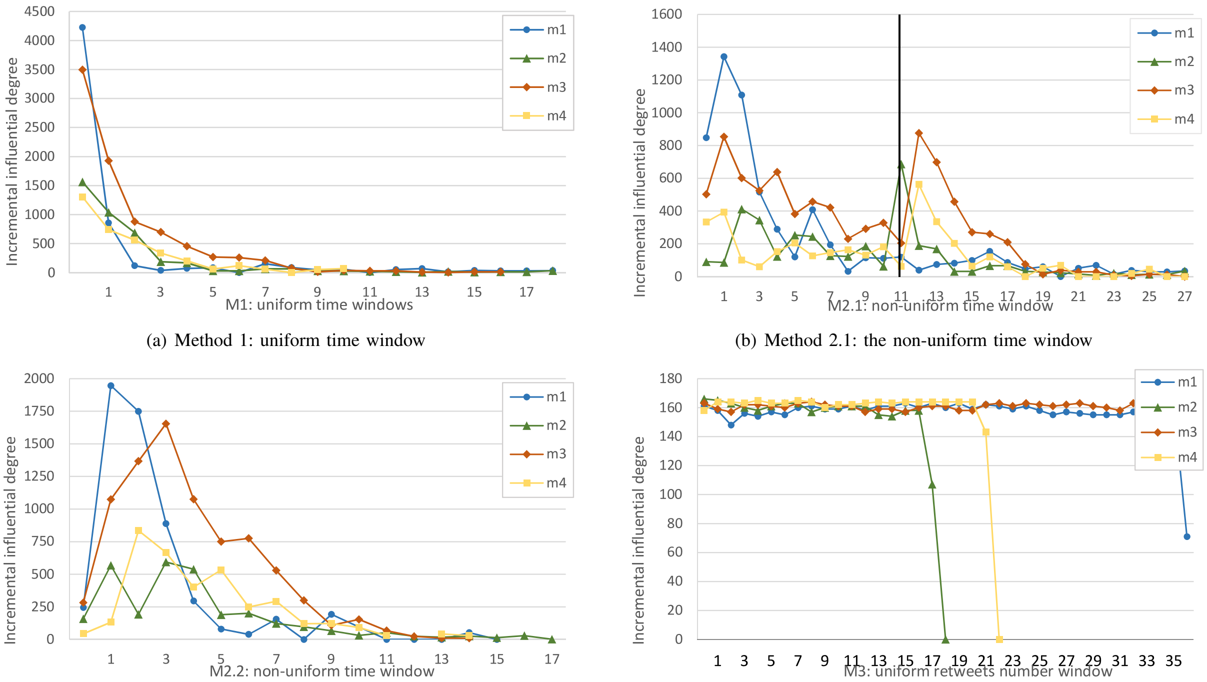
2) Parameters set: During the experiments, the parameters in three window partition methods were set as follows:

Method 1: (uniform time window) by analysis this microblogging data, we find that after 48 hours the retweets numbers of microbloggings are quite small till the end. In this article, we set the size of time window Δ=1 hour.

Method 2.1: (the non-uniform time window based on data statistics) as shown in Fig. 2, the retweets number of microbloggings drop rapidly near $2th$ hour. So, we set t_s=$2th$ hour, $\Delta 1 = 1/6$ hour and $\Delta 2$ =1 hour in the method based on data statistics.

Method 2.2:(the non-uniform time window based on arithmetic sequence) in the method based on arithmetic squence, the tolerance is set $d = 1/6$ hour.

Method 3: (uniform retweets number window) in experiments, we set the size of window $\eta = 100, 500$ successively to partition microblogging data. In addition, through the statistical analysis, we set $\alpha = 0.7$, $\beta = 0.2$ and $\gamma = 0.1$ in incremental influential degree model.



(a) Method 1: uniform time window

(b) Method 2.1: the non-uniform time window

(c) Method 2.2: the non-uniform time window

(d) Method 3: the uniform retweets number window

Fig. 4. The distribution of the average incremental influential degree of influential nodes based on different time windows

B. *Experimental Results*

After window partition, we detect the local influential nodes and global influential nodes. For every window, the average incremental influential degree of top κ local influential nodes was counted. Then we create a line graph for all windows.

Fig. 4 shows the distribution of the average incremental influential degree of influential nodes detected by the IIND algorithm based on different methods over corresponding time windows. The Fig. 4(a)-(d) are Method 1, Method 2.1, Method 2.2 and Method 3, respectively.

From Fig. 4(a), we find that the average incremental influential degree distribution of local influential nodes is a power law distribution, which has a long tail. Further more, this distribution agrees with the microblogging diffusion distribution over time in Fig. 3. In other words, the local influential nodes based on Method 1 have a good predictive ability for microblogging diffusion.

Fig. 4(b) and (c) show the distribution about local influential nodes in non-uniform time windows based on Mehtod 2.1 and Method 2.2 respectively. We find that these distribution are quite different from Fig. 4(a). Specifically, every microblogging has two peaks in the distribution in Fig. 4(b). The window is divided into two parts by the t_s. And peaks occur in the beginning of each part. It is said that no mater which part, the diffusion of information decrease with time. While, in Fig. 4(c), the peaks occur in early time windows, through the sizes of later time windows are on the increase.

Fig. 4(d) corresponds the IIND algorithm experiment based on Method 3. The size of uniform retweets number window is $\eta = 500$, and the number of local influential nodes is in every window $\kappa = 3$. It is obvious that the average incremental degrees of local influential nodes are equal in every window, which fluctuates around a fixed value 160. So the sum of incremental influential degree sum of 3 local influential nodes in every window is about 480, accounting for 96% of the total retweets number in every window. There is a similar phenomenon in other experiment with $\eta = 100$, and $\kappa = 2$. It shows clearly that the influential nodes play leading roles in the microblogging diffusion process.

By comparison, we find that the numbers of windows partitioned by different methods are different. For example, for the microblogging $m2$, the green lines in Fig. 4, it has 27 time windows based on Method 2.1 in Fig. 4(b), while it only has 17 time windows based on Mehtod 3 in Fig. 4(d). So, influential nodes detected by IIND based on different methods are different. The performance of IIND based on the three methods will be evaluated later.

C. *Evolution Patterns Analysis*

For analyzing the evolution of the influential nodes detected by IIND algorithm, we mainly extract six features of influential nodes. They are certification, certification reasons, the number ratio of followers to followees, status, city and retweet interval respectively.

Certification: Sina Weibo celebrity certification system is introduced to protect the interests of celebrities, so this is a sign used to distinguish ordinary people. **Certification reasons:** we can determine the user identity according to microblogging certification reasons. **Twitters:** Twitters of user u is his/her microblogging number so far. Twitters of users can be used to measure their activity and interaction frequency with other users in microblogging platform. **City:** the city of influential nodes can reflect the effect of different economic condition on people social entertainment. **Retweeted interval:** it is time interval between retweeted time stamp and publishing time stamp. **The number ratio of followers to followees:** only

TABLE III
FEATURE ANALYSIS OF INFLUENTIAL NODES

Influential nodes	Certification ratio	Certification reasons	#fowllowers/#followees	Twitters	City	Retweeted interval
global influential nodes	1	news media, actor	56936.5	27850.6	BJ,SH,GD	0.40h
local influential nodes1	17/20	actor, and humorous etc well-known bloger	7789.02	21441	most BJ,SH,GD	3.46h
local influential nodes2	1/3	local TV station, official microblogging of organizations	768	13764	countrywide	11.93h

a large number of followers can not explain that the range of the user is large. Maybe its followees number is larger.

TABLE III shows the features of influential nodes. The influential nodes are divided into 3 groups, global influential nodes, local influential nodes 1, local influential nodes 2 respectively. The local influential nodes 1 are in the first half of windows, while local influential nodes2 are in the second half windows.

There are some evolution patterns from TABLE III. We describe these patterns in the following.

1) The global influential nodes have greater influence spread than local influential nodes, and they are more active to update microblogging. In TABLE III, all of global influential nodes are certified, and these users are celebrities, including new media and actor. The average number ratio of followers to followees and the average status of global influential nodes is the highest of the three groups.

2) From the comparison data in TABLE III, we find that the influential nodes are changed from new media to famous bloggers to official microblogging of organizations. It claims that news media has significant influence in information diffusion in OSNs, which is similar to the ideas in reference [35].

3) The influential nodes change from the large cities to the countryside in information diffusion. All of global influential nodes are in Beijing, Shanghai, Guangdong, three cities with the strongest economic strength in China. And, early local influential nodes are most in these cities, and then local influential nodes evolve to countrywide. Maybe the people in the economically developed cities are sensitive to hot topics.

4) The certification ratio of the influential nodes gradually decreased from 1 in information diffusion. And, early influential nodes trigger large number of users to retweet microlbogging. In other words, the smaller the retweet interval of the influential nodes are, the larger the retweets number triggered by them.

At last, we discuss the relations between the number of global influential nodes and the retweets number of microbogging. From Fig. 5, we see that the number of global influential nodes is positively associated with the retweets number of microblogging. In other words, for a microblogging, the more global influential nodes it has, the larger retweets number it will trigger.

D. Effectiveness Evaluation

The evaluation experiment measures the efficiency of IIND algorithm. The independent cascade model [29] with weights is employed as the information diffusion model to simulate the information diffusion process . We demonstrate the performance of IIND algorithm on Sina WeiBo data sets in comparison with two popular heuristic algorithms [26]: MDH and HCH, which are introduced in related work.

The process of diffusion in this model is described as: in directed microblogging network graph $G(V, E)$, the edge from

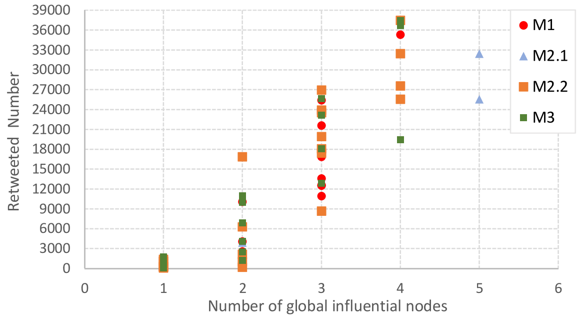


Fig. 5. Retweets number versus the number of global influential nodes of microbloggings

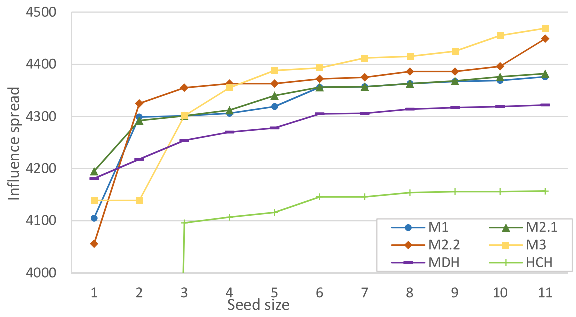


Fig. 6. Influence spreads of an initial seed set selected by the IIND algorithm based on three window partition methods and baselines

u to v has weight $w_{uv} = 1/Id_v$, where the Id_v is in-degree of user v, denoting the followee number of user v. It starts with an seed set of active nodes. In step t, every newly activated node i only has single change to activate each currently inactive neighbor j with probability w_{ij}, which is independent of the history thus far. If j has multiple newly activated neighbors, their attempts are sequenced in an arbitrary order. If i succeeds, then j will become active in step $(t + 1)$; but whether or not i succeeds, it cannot make any further attempts to activate j in subsequent rounds. Again, the process runs until no more activations are possible.

We initialize seed set with influential nodes by IIND and baselines algorithms (MDH and HCH) respectively. Through the simulation procedure of information diffusion and many iterations, we compare the expected number of final active nodes (influence spread) to evaluate efficiency of algorithms [36]. As reported in Fig. 6, the seed set selected by IIND has larger influence spread than MDH and HCH algorithms, no matter what window partition method the IIND algorithm based on. Further more, influential nodes detected by IIND based on M3 have a little larger influence spread.

VI. Conclusion

This paper focuses on information diffusion in OSNs in a fine-grained way. To be specific, we focus on measuring the influential degree of nodes and studying the evolution patterns of influential nodes. Keeping the dynamic features of OSNs in mind, we propose an Incremental Influential Nodes Detection (IIND) algorithm, and three time window partition methods. We conduct extensive experiments on real data set in Sina Weibo and gain several interesting findings on the evolution patterns of influential nodes. We also validate the effects of the proposed algorithm, which shows its advantage. In future work, we are interested in applying the influential nodes to intervene information diffusion in OSNs.

Acknowledgement

This work is supported by NSFC grants 61502161 and 61632009, the Chinese Fundamental Research Funds for the Central Universities 531107040845, ISTCP grant 2013DF-B10070.

References

[1] E. Bakshy, D. Eckles, R. Yan, and I. Rosenn, “Social influence in social advertising: evidence from field experiments,” in *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 2012, pp. 146–161.

[2] W. Jiang, J. Wu, G. Wang, and H. Zheng, “Fluidrating: A time-evolving rating scheme in trust-based recommendation systems using fluid dynamics,” *Proc. IEEE INFOCOM*, pp. 1707–1715, 2014.

[3] W. Jiang, J. Wu, and G. Wang, “On selecting recommenders for trust evaluation in online social networks.” *ACM Trans. Internet Techn.*, vol. 15, no. 4, pp. 14–1, 2015.

[4] F. Morone and H. A. Makse, “Influence maximization in complex networks through optimal percolation,” *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[5] S. Li, Y. Zhu, D. Li, and D. Kim, “Influence maximization in social networks with user attitude modification,” in *IEEE International Conference on Communications*, 2014.

[6] W. Jiang, G. Wang, M. Z. A. Bhuiyan, and J. Wu, “Understanding graph-based trust evaluation in online social networks: Methodologies and challenges,” *ACM Comput. Surv.*, vol. 49, no. 1, pp. 10:1–10:35, May 2016.

[7] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, “Trust evaluation in online social networks using generalized flow,” *IEEE Transactions on Computers (TC)*, vol. 65(3), pp. 952–963, 2016.

[8] W. Jiang and J. Wu, “Active opinion-formation in online social networks,” *Proc. IEEE INFOCOM*, pp. 1440–1448, 2017.

[9] W. Jiang, G. Wang, and J. Wu, “Generating trusted graphs for trust evaluation in online social networks,” *Future Generation Computer Systems*, vol. 31, pp. 48–58, 2014.

[10] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, “Vital nodes identification in complex networks,” *Physics Reports*, vol. 650, pp. 1–63, 2016.

[11] J. Sun and J. Tang, “A survey of models and algorithms for social influence analysis,” *Social network data analytics*, pp. 177–214, 2011.

[12] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, “Information diffusion in online social networks: a survey,” *ACM SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013.

[13] L. L, T. Zhou, Q. M. Zhang, and H. E. Stanley, “The h-index of a network node and its relation to degree and coreness,” *Nature Communications*, vol. 7, p. 10168, 2016.

[14] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, “A model of internet topology using k-shell decomposition,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11 150–11 154, 2007.

[15] H. E. Tillett, *Infectious Diseases of Humans; Dynamics and Control*. Cambridge University Press, 1991.

[16] 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 Transactions on Industrial Informatics*, 2017.

[17] S. Dhamal, K. J. Prabuchandran, and Y. Narahari, “Information diffusion in social networks in two phases,” *IEEE Transactions on Network Science and Engineering*, vol. PP, no. 99, pp. 1–1, 2017.

[18] S. Mahdizadehaghdam, H. Wang, H. Krim, and L. Dai, “Information diffusion of topic propagation in social media,” *IEEE Transactions on Signal and Information Processing Over Networks*, vol. 2, no. 4, pp. 569–581, 2016.

[19] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, “Modeling information diffusion over social networks for temporal dynamic prediction,” *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[20] K. Dedecius and P. M. Djurić, “Sequential estimation and diffusion of information over networks: A bayesian approach with exponential family of distributions,” *IEEE Transactions on Signal Processing*, vol. 65, no. 7, pp. 1795–1809, 2017.

[21] P. F. Lazarsfeld, B. Berelson, and H. Gaudet, “The peoples choice: how the voter makes up his mind in a presidential campaign.” 1968.

[22] M. S. Granovetter, “The strength of weak ties,” *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[23] D. Krackhardt, N. Nohria, and B. Eccles, “The strength of strong ties,” *Networks in the knowledge economy*, p. 82, 2003.

[24] R. S. Burt, “The social structure of competition,” *Explorations in economic sociology*, vol. 65, p. 103, 1993.

[25] T. Lou and J. Tang, “Mining structural hole spanners through information diffusion in social networks,” in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 825–836.

[26] R. Narayanam and Y. Narahari, “A shapley value-based approach to discover influential nodes in social networks,” *IEEE Transactions on Automation Science and Engineering*, vol. 8, no. 1, pp. 130–147, 2011.

[27] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, “Identification of influential spreaders in complex networks,” *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[28] S. N. Dorogovtsev, A. V. Goltsev, and J. F. Mendes, “k-core organization of complex networks,” *Physical Review Letters*, vol. 96, p. 040601, 2006.

[29] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[30] D. Kempe, J. Kleinberg, and E. Tardos, “Influential nodes in a diffusion model for social networks,” in *International Colloquium on Automata, Languages, and Programming*. Springer, 2005, pp. 1127–1138.

[31] M. Kimura and K. Saito, “Tractable models for information diffusion in social networks,” in *Knowledge Discovery in Databases: Pkdd 2006, European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings*, 2006, pp. 259–271.

[32] Y. Wang, G. Cong, G. Song, and K. Xie, “Community-based greedy algorithm for mining top-k influential nodes in mobile social networks,” in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 1039–1048.

[33] W. Jiang, J. Wu, G. Wang, and H. Zheng, “Forming opinions via trusted friends: Time-evolving rating prediction using fluid dynamics,” *IEEE Transactions on Computers (TC)*, vol. 65(4), pp. 1211–1224, 2016.

[34] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[35] L. Changhyun, K. Haewoon, P. Hosung, and M. Sue, “Finding influentials based on the temporal order of information adoption in twitter,” in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1137–1138.

[36] C. Wang, W. Chen, and Y. Wang, “Scalable influence maximization for independent cascade model in large-scale social networks,” *Data Mining and Knowledge Discovery*, vol. 25, no. 3, p. 545, 2012.