An Influence Model Based on Heterogeneous Online Social Network for Influence Maximization

Xiaoheng Deng, Fang Long, Bo Li, Dejuan Cao, and Yan Pan

***Abstract*—Influence maximization is an important technique in advertisement post, viral marketing, and public opinion monitoring. Seed set identification is one of the key issues in influence maximization. In reality, there exist heterogeneous nodes, such as user nodes, message nodes in social networks. The complex association relationship among heterogeneous nodes, which are seldom considered, significantly increases the complexity of the seed set identification. In this paper, we propose a Measuring Influence (MIF) model to capture social influence with heterogeneity. MIF considers the interaction among adjacent nodes, the tag of users, the users' social friendships and the similarity in user interests, and studies the interaction based influence, tag based influence, friendship based influence, and topic based influence, respectively. As obtaining the seed set in social networks has been proved to be a NP-hard problem, we propose an algorithm called Influence Maximization Greedy Algorithm (IMGA) to solve this problem by maximizing the marginal influence of selected seed nodes. Series of experiments are designed to evaluate the performance of the proposed model and algorithm. Our results show that MIF model and IMGA algorithm have better influence spread effects and higher quality of the seed set identification comparing to the approaches under IC, LT, CDNF, MIA, and BBA, models.**

***Index Terms*—Social network, influence maximization, heterogeneous information.**

I. Introduction

Online social platforms have become an important communication channel to share information. Moreover, the way of influence spread via the word-of-mouth (WoM) effect and viral marketing in social networks is the same as the cascaded way of information dissemination in online social platforms. As a result, there are challenges and opportunities with how to utilize the feature of information dissemination to identify the influence of users. Once influence is determined, we can send messages to users who have maximum impacts on others when advertising, marketing and so on, which can save time and money. Hence, influence maximization [1] is an important research topic. Extensive recent research focuses on two aspects. One of them is to improve the efficiency of algorithm combining with the structure of networks, and the other is to design heuristic strategy based on IC and LT model to reduce computation cost [14], [15]. In addition, a few methods were proposed to exploit text mining, users' bias of information, common friends, and common tags comprehensively to improve the measuring accuracy of influence [5], [9]. However, these methods do not capture well the complex correlation among heterogeneous nodes in social networks. In influence maximization, we should consider that people tend to have more impact on users who have more similar interests [2], more common friends [3], [25], and more similar bias on topics [4], [26]. Thus, analyses on complex relationship among heterogeneous nodes to find users' preferences, topic bias, and social relationship have great importance [28], [31], [32]. To overcome the limitations of previously proposed methods, we propose a Measuring influence (MIF) model taking the heterogeneity of social networks into consideration, finding influence with interactions, friendships, tags, and topics in social networks. Specifically, interaction based influence among users utilizes behavior records. Friendship based influence takes use of users' social relationships. Tag based influence uses users' tags, and topic based influence depends on the similarities among users' messages in social networks. In general, MIF characterizes the relationship among heterogeneous nodes of social networks from multiple perspectives and uses these complex relationships to calculate influence. The way to compute influence is based on the complicated correlation among heterogeneous nodes in social networks, constructing tags' feature vectors and social friendships' feature vectors for each user, measuring similarities among feature vectors. Moreover, topic based influence is also based on users' interests on topics, aiming at disseminating influence among heterogeneous nodes based on similarity of topic.

However, how to utilize the heterogeneity of users' behavior records, users' tags, and user's social relationships is one of the biggest challenges in our model. At the last step, MIF models the influence as the sum of all kinds of influence. Besides, influence maximization has been proved as a NP-hard problem. In order to find seed set efficiently, we propose an algorithm called Influence maximization greedy algorithm (IMGA). The contributions of our work are summarized as follows:

(1) We design a MIF model that measure social influence among users based on interactions records, social friendships, tags, and topics. MIF characterizes the complex correlation of social networks considering all

Manuscript received January 31, 2018; revised May 6, 2019; accepted May 21, 2019. Date of publication June 3, 2019; date of current version June 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61772553 and 61379058 and in part by the Grant 2017zzts515 from Central South University. Recommended for acceptance by L. Trajkovic. *(Corresponding author: Xiaoheng Deng.)*

X. Deng, F. long, B. Li, D. Cao, and Y. Pan are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: dxh@csu.edu.cn; Longfang@csu.edu.cn; csulibo@csu.edu.cn; caodejuan@csu.edu.cn; panyan@csu.edu.cn).

Digital Object Identifier 10.1109/TNSE.2019.2920371

the information above. Hence, heterogeneity of social networks are fully considered.

(2) We propose an algorithm IMGA that solves how to select seed set in influence maximization by selecting users who have maximum marginal gain as a node of seed set recursively. The marginal gain is the result of simulation of influence spreading.

Our experiments on the Flicker dataset and Ego Network dataset show that the proposed MIF model solving by IMGA algorithm outperforms IC model [8], LT model [8], CDNF model proposed in [5], MIA model [10], and BBA model [22] on the effect of influence spread and on the quality of seed set.

The reminder of this paper is organized as follows. Section II reviews the related work. Section III presents the proposed MIF model and Section IV defines the IMGA algorithm. Section V evaluates the performance of MIF model and IMGA algorithm against several methods, and finally, Section VI concludes the paper.

II. Related Work

Richardson, Domingos *et al.* tried to solve the problem of influence maximization in an optimal way [6], [29], [30]. After that, many different strategies have been introduced to solve the problem of influence maximization, and most of them tried to improve algorithm efficiency by exploiting heuristic algorithm [7], use different function to measure influence [8], and boost the accuracy by greedy algorithms.

A. Influence Maximization With Heuristic Algorithm

PMIA algorithm combines the local tree structure with greedy algorithm to optimize the computation progress [14]. Although algorithms such as PMIA and LDAG [15] can improve the efficiency by using heuristic strategy, the accuracy decreases, comparing to traditional greedy algorithm. Besides, some research focuses on measuring influence with different methods. In reality, influence of one user can be influenced by a lot of factors, which indicates that the measurement of influence is very complex. In [5], authors measure influence from three perspectives, proposing GNF algorithm to solve CDNF model. In [9], Yan pan *et al.* combine the constraint conditions of propagation restrictions in influence, the characteristics of the delay constraints that prevail in the process of propagation, and the credit distribution and user behavior log to construct the propagation of the influence in social networks, proposing CDTC model. Zhou *et al.* extend IC model [8], considering different degree of information preferences of users, proposing EIC model. Moreover, they take the motivation probabilities among users into consideration to dynamic their model. In [10], authors proposed a novel method named as MIA to restrict IC model onto shortest paths. To identify seed set, they optimize the greedy algorithm to reduce the complexity. In [11], authors extend [10], to learn both different degree of information preferences and the topic of information, which improve the results of [10].

B. Influence Maximization With Greedy Algorithm

Kempe *et al.* formulated the issue of influence maximization and proposed IC (Independent Cascade model) and LT (linear Threshold Model). They proved that greedy algorithm can ensure the error rate lower than $1 - 1/e$ when solving the influence maximization in IC and LT model [8]. However, IC and LT need a lot of *Monte-Carlo* simulations to guarantee the precision of influence measurement, which limit their universality in large scale social networks. Therefore, how to find an efficient, universal, and scalable algorithm becomes an urgent problem. In [12], Amit *et al.* extended CELF [13] as CELF++, which greatly improved the accuracy of greedy algorithm. However, all the above models ignore the correlation among heterogeneous nodes in social networks. In [22], authors use belief theory to fuse heterogeneous information. Based on the result of the fusion, influence and marginal gain are figured out. Influence maximization is solved by cost effective lazy-forward algorithm [27].

III. MIF Model

The goal of researches on the influence maximization in social networks is to find specific number of seed set so that the nodes in this set have greatest influence on the nodes outside this set. Thus, the first core task of influence maximization is to measure the influence among users precisely in a social network and the second one is to assess the process of influence propagation. However, both of these two issues have been proved to be NP-hard [16]. MIF model is proposed to solve the first core task in influence maximization measuring social influence by using heterogeneity of social networks so that seed set can be found effectively. MIF divided social influence in heterogeneous social networks into four parts: interaction based influence, social friendship based influence, tag based influence, topic based influence.

A. Problem Definition

In this paper, let $G(U, M, T, E)$ be the graph model of social networks, with U being the set of users in social networks, M the set of messages produced by interactions among users in social networks, T the set of tags, an attribute of users in social networks, and E the social relationships in social networks. $n = |U|$ is the number of users, $|M|$ is the number of messages and $|T|$ is the number of tags. U, M and T are heterogeneous nodes for each other, and all of them together constitute the vertex set V of G. Our MIF model inputs the users' behavior records B, expressed by any type of interactions records such as thumb up, repost, and comment and so on. Based on the behavior records B, for each user i, we can obtain the messages i has interactions with, denoted by $EID(i)$. In addition, we consider one adjacency matrix $A \in \mathbb{R}^{n \times n}$ to present social relationships. Based on adjacency matrix A, for each user i, we can calculate the neighborhoods N_i with his friends. Moreover, we also take one adjacency matrix C to present the relationship among users and tags. Based on the matrix C, for each user i, we can obtain tags belonging to i, denoted by $T(i)$.

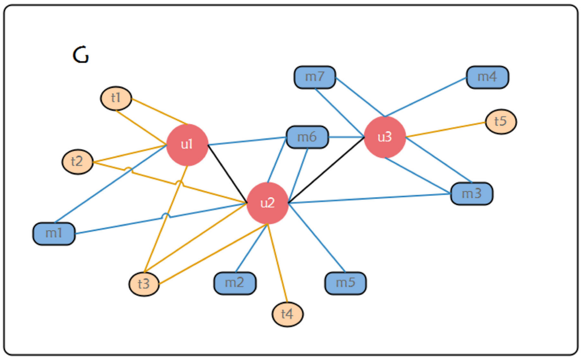


Fig. 1. Graph model of social networks.

For example, in a social network as shown in Fig. 1, red nodes represent users, with orange nodes being the tags and blue nodes the messages. We can see complex correlation among these heterogeneous nodes. What deserves to be mentioned is that one tag can belong to same user more than one times and each user can operate on same message repeatedly as well. Therefore, the degree of tag nodes and message nodes can be larger than one. With the setting above, in Fig. 1, $n = 3$, $|T| = 5$, $|M| = 7$, $V = T \cup M \cup U$. E is the edges in Fig. 1. $A(u1, u2) = 1$, $A(u1, u3) = 0$, $A(u2, u3) = 1$, neighbors of $u1$, denoted by N_{u1}, equals to $u2$. Moreover, $EID(u1) = \{m1, m6\}$ and $T(u1) = \{t1, t2, t3\}$.

In general, users' behaviors records intuitively reflect the route of influence spreading and latency, which can be used to mine influence of users. Moreover, in graph model G, similar nodes have common nodes in a heterogeneous social network, comparing to dissimilar nodes because of the frequent interactions among similar nodes [17], [18]. For instance, if there are a lot of common friends or common tags between two users, it is reasonable that larger influence exists between them. Moreover, since users have different preference on the topics, users prefer to disseminate messages which they are interested in [19]. As a result, users' influence on a particular topic may be propagated to other users who have similar preference on topics, according to which can infer influence exists among users who have similar preference on topics. With these settings, our problem is formally defined as follows:

Defintion 1 (Problem): Given (i) neighborhoods in the adjacency matrix A, (ii) messages in the behavior records B, and (iii) tags in the adjacency matrix C, the goal of the proposed MIF model is to calculate the influence among users in a heterogeneous social network consisting of user nodes, tag nodes, and message nodes.

B. Interaction Based Influence

Generally, users perform a wide variety of behaviors in social networks. {Publish, like, repost, and comment} are actions considered in this paper. Furthermore, MIF divide these actions into two kinds: *executive actions* (*EA*) and *interaction behaviors* (*IB*).

Defintion 2 (Executive Actions, EA): Executive actions denote the actions which are spontaneous. Specifically, MIF defines {publish, like, repost, and comment} as users' executive actions (EA).

Defintion 3 (Interaction Behaviors, IB): Interaction behaviors denote the actions which can be motivated. Specifically, MIF defines {like, repost, comment} as users' interaction behaviors (IB).

As mentioned above, the frequency and latency of *IB* among users are related to the degree of closeness and importance, thus reflecting influence among users. The more frequent *IB* between user u and user v happen, the smaller latency *IB* between user u and user v have, the larger influence exists between user u and user v. On the contrary, the less frequent *IB* between user u and user v happen, the larger latency *IB* between user u and user v have, the smaller influence exists between user u and user v. In social networks, the user's behavior records intuitively reflect the frequency and latency of *IB* among users. Thus, users' behavior records are used to compute the users' influence.

Given graph model $G(U, M, T, E)$, considering user u, we can find another user $v \in N_u$. v may do some *IB* to u on message m after a period time t if u does *EA* on message m. Aiming at mining the interaction based influence between u and v, the $MP_{u,v}$ (Motivate Probability u to v) is defined as follows:

Defintion 4 (Motivate Probability): xThe motivate probability u to v—$MP_{u,v}$, shows the probabilities of u's *EA* motivate v's *EA*.

Let $EID(u)$ be the ID set of messages u has done *EA* with, and $|EID(u)|$ the number of $EID(u)$. According to Definition 4, the $MP_{u,v}$ is computed as follows:

$$MP_{u,v} = \frac{|EID(u) \cap EID(v)|}{|EID(u)|} \tag{1}$$

Meanwhile, the $BO_{u,v}$ (Behavior Occupancy u to v) is defined as follows:

Defintion 5 (Behavior Occupancy): The Behavior Occupancy u to v—$BO_{u,v}$, shows the probabilities of v's responding to u's *EA*.

According to Definition 5, the $BO_{u,v}$ is computed as follows:

$$BO_{u,v} = \frac{|EID(u) \cap EID(v)|}{|EID(v)|} \tag{2}$$

Obviously, the $MP_{u,v}$ is the ratio between the frequency of *IB* between u and v and the frequency of *EA* of u, because one message can not be published by different users. The larger $MP_{u,v}$ is, the more likely u's *EA* motivates v, the more interaction based influence u has on v. The $BO_{u,v}$ is the ratio between the frequency of *IB* between u and v and frequency of *EA* of v. The larger $BO_{u,v}$ is, the more concerned v cares about u's *EA*, the greater direct influence u has on v. However, if u or v is an inactive node, that is, $|EID(u)|$ or $|EID(v)|$ is small. $MP_{u,v}$ or $BO_{u,v}$ may be very high, which reflects influence incorrectly, resulting in a large error. In order to revise this kind of error, we consider $MP_{u,v}$ and $BO_{u,v}$ comprehensively to calculate $II_{u,v}$(interaction based influence between u and v), which is shown as follows:

$$\begin{aligned} II_{u,v} &= \lambda \cdot MP_{u,v} + (1-\lambda)BO_{u,v} \\ &= \lambda \frac{|EID(u) \cap EID(v)|}{|EID(u)|} + (1-\lambda)\frac{|EID(u) \cap EID(v)|}{|EID(v)|} \end{aligned} \tag{3}$$

where $\lambda \in (0, 1)$.

$MP_{u,v}$ and $BO_{u,v}$ base on the frequency of *IB* between u and v. However, latency of *IB* between u and v indicates sensibility between v and u, relating to the influence between u and v as well. In social networks, as the sensibility of users are different, there is need to measure the sensibility among users. We take the following modification:

$$II_{u,v} = \{\lambda \cdot MP_{u,v} + (1-\lambda)BO_{u,v}\}e^{-\frac{t_{v,u}-T_v}{T_v}} \tag{4}$$

where $t_{v,u}$ is average latency of *IB* v to u and T_v is v's average latency to all his neighbors N_v. The second term of Eq. (4) uses exponential function to take the latency into consideration, improving the preciseness in measurement of influence. According to Eq. (4), if latency v to u is very small, $-\frac{t_{v,u}-T_v}{T_v}$ is close to 1. On the contrary, if latency v to u is very large, $-\frac{t_{v,u}-T_v}{T_v}$ is close to negative infinity. As a result, $-\frac{t_{v,u}-T_v}{T_v} \in (-\infty, 1)$ and $e^{-\frac{t_{v,u}-T_v}{T_v}} \in (0, 2.7)$. Multiplying exponential function to capture the correlation between influence and latency of *IB*. By doing so, $II_{u,v}$ becomes larger as the $t_{v,u}$ decreases and $II_{u,v}$ becomes smaller as $t_{v,u}$ increases. According to Eq. (4), we already have a method to measure interaction based influence. But it is not enough to capture the real influence in a heterogeneous social network. As we shown in the next subsection, we consider the friendship based influence.

C. *Friendship Based Influence*

In social networks, some users become estranged from each other while some users become intimate as time passed by. That means the relationship among users are dynamic changed with time [20]. Given two users u and v, if they have a lot of common friends, they are more likely to establish connection and become familiar as time passing by, which means higher influence exists among them. In our graph model $G(U, M, T, E)$, we explore friendship based influence by the similarity of users' common social relationship using *TF-IDF* algorithm.

In a social network, each user has his own friendships, which means that there are different neighbors for different users. If two users u and v have many common neighbors, the influence of u can be strongly applied to v through their common neighbors, and vice versa. Therefore, we construct the neighbor eigenvectors according to the topology structure of networks. Comparing the similarity of this eigenvectors helps to get the friendship based influence. The calculation of friendship based influence summarizes as follows:

(1) Based on graph model $G(U, M, T, E)$ and adjacency matrix A, for $\forall u \in U$, we use N_u to represent the neighbors of u.

(2) Initializing a neighbor eigenvector for u, denoted by $Neigh_u$, whose dimension equals to $|U|$. Specifically:

$$Neigh_u = \underbrace{(0, 0, \ldots, 0)}_{|U|}$$

(3) If $A(u, v) = 1$, the v-th dimension in $Neigh_u$ equals to 1. The v-th dimension in $Neigh_u$ is denoted by $neigh_{v,u}$. Specifically:

$$Neigh_u = \underbrace{(\mathrm{neigh}_{1,u}, \mathrm{neigh}_{2,u}, \ldots, \mathrm{neigh}_{|U|,u})}_{|U|}$$

for instance, in Fig. 1, $Neigh_{u1} = (0, 1, 0)$, $Neigh_{u2} = (1, 0, 1)$.

(4) The *TF-IDF* value of $neigh_{v,u}$ is calculated as follow:

$$TF-IDF(neigh_{v,u}) = \frac{neigh_{v,u}}{\sum_{k=1}^{|U|} neigh_{k,u}} \cdot \leqslant \frac{|U|}{|N_v|} \tag{5}$$

where $|N_v|$ is the number of N_v.

(5) Updating $Neigh_u$:

$$Neigh_u = (TF-IDF(neigh_{1,u})), \ldots, TF-IDF(neigh_{|U|,u})) \tag{6}$$

(6) Calculating friendship based influence: given two users u and v, we can obtain the neighbor eigenvectors $Neigh_u$ and $Neigh_v$. Utilizing the cosine similarity between $Neigh_u$ and $Neigh_v$ to measure the friendship based influence between u and v, details as follow:

$$Neigh\_inf(u, v) = \cos(Neigh_u, Neigh_v) \tag{7}$$

where *cos* means the cosine similarity.

D. *Tag Based Influence*

In social networks, the characteristics of users can be discovered by analyzing users' tags. Given two users u and v, if the similarity of tags between them are high, which means they are more likely have more common personalities, it is more possible that u and v have impact on each other [21]. Accordingly, our strategy to calculate influence based on tags is presented as follows:

(1) Based on graph model $G(U, M, T, E)$, for $\forall u \in U$, the tags belonging to u is denoted by $T(u)$.

(2) Initializing a tag eigenvector for u, denoted by Tag_u, whose dimension equals to $|T|$. Specifically:

$$Tag_u = \underbrace{(0, 0, \ldots, 0)}_{|T|}$$

(3) The i-th dimension in Tag_u is denoted by $tag_{i,u}$ and the i-th tag in T as tag_i. $tag_{i,u}$ indicates the number of times tag_i appears in $T(u)$. Specifically:

$$Tag_u = \underbrace{(tag_{1,u}, tag_{2,u}, \ldots, tag_{|T|,u})}_{|T|}$$

for example, in Fig. 1, $Tag_{u1} = (2, 1, 1, 0, 0)$, $Tag_{u2} = (0, 1, 2, 1, 0)$ and $Tag_{u3} = (0, 0, 0, 0, 1)$.

(4) The *TF-IDF* value of $tag_{i,u}$ is calculated as follow:

$$TF - IDF(tag_{i,u}) = \frac{tag_{i,u}}{\sum_{k=1}^{|T|} tag_{k,u}} \cdot \lessgtr \frac{|U|}{|v : tag_i \in T_v|} \quad (8)$$

where $|v : tag_i \in T(v)|$ is the number of users whose tags contain tag_i.

(5) Updating Tag_u:

$$Tag_u = (TF - IDF(tag_{1,u}), \ldots, TF - IDF(tag_{|T|,u})) \quad (9)$$

(6) Calculating tag based influence: given two users u and v, we can obtain the tag eigenvectors Tag_u and Tag_v. Utilizing the cosine similarity between Tag_u and Tag_v to measure the tag based influence between u and v, details as follow:

$$Tag\_inf(u, v) = cos(Tag_u, Tag_v) \quad (10)$$

where cos means the cosine similarity.

E. Topic Based Influence

In general, users have their own expertise, having different impact on different topics. For example, economists have larger influence in the field of economics while having smaller influence in the field of ecology. If an economist publishes a message about economics, more concern his friendships will pay to this message. While if he publishes a message about ecology, it will be less influential. That is to say, we can excavate users' influence by the topics of messages they have done *EA* with and their expertise. Moreover, in social networks, when one user publishes a message, his friendships may forward, comment or like this message. In other words, his friendships may do some *IB* to this message. However, the frequency of *IB* messages received are different, and different type of *IB* represent different levels of concern. That means every user's influence reflected in messages might be different. Therefore, it is meaningful to model the relationship between user and message. However, in $G(U, M, T, E)$, U and M are heterogeneous. How to model the relationship among heterogeneous nodes is one of the biggest challenges in social networks as well in our model. Before we describe our method to measure topic based influence, it is necessary to discuss it amply:

As stated above, users' influence are related to the topics of messages they have done *EA* with and their expertise. Naturally, if $m1$ and $m2$ are the messages u and v have done *EA* with, the topic of $m1$ can be seen as a portray of $u's$ expertise, so does $m2$. The similarity between $m1$ and $m2$ is the bridge to compute the similarity in expertise of u and v. As a result, considering a possible route of influence disseminates $(u \rightarrow m1 \rightarrow m2 \rightarrow v)$, the influence between user nodes (u or v) and message nodes ($m1$ or $m2$) and the similarity among message nodes ($m1$ and $m2$) are two cores of our topic based influence.

1) Influence Between User Nodes and Message Nodes: Given a user u and a message m u has done *EA* with at time t, the influence between u and m can be calculated by the number of times m been done *IB* by N_u after t. With the setting of *IB*={like,forward,comment}, the influence between user nodes and message nodes details as follows:

(1) Assuming $\zeta_m = \{a_1, a_2, a_3\}$ are the number of times m being forwarded, commented, and liked by N_u after t, respectively, and $\omega = \{m_1, m_2, m_3\}$ are weights represented by forwarding, commenting, and liking where $m_1, m_2, m_3 \in (0, 1)$, accordingly.

(2) Calculating the influence u has on m:

$$U\_m\_inf(u, m) = \frac{1}{1 + e^{-(\zeta_m \cdot \omega)}} \quad (11)$$

According Eq. (11), the more *IB* m received after t, the more influential u is to m. However, the messages also have influence on users, denoted by $M\_u\_inf(m, u)$, which we define equals to $U\_m\_inf(u, m)$.

$$U\_m\_inf(u, m) = M\_u\_inf(m, u) \quad (12)$$

2) Similarity Among Message Nodes: In social networks, if the specialty of users are similar, it's more likely they have similar interests and get in contact with each other. As a result, they will have more influence on each other. According to the possible route of influence propagating, it is meaningful to measure similarity among messages nodes.

(1) Based on graph model $G(U, M, T, E)$, for $\forall m \in M$, the key words of m is denoted by K_m and all the key words as $KEYs$.

(2) Initializing an eigenvector Key_m for m, whose dimension equals to the number of $KEYs$. Specifically:

$$Key_m = \underbrace{(0, 0, \ldots, 0)}_{|KEYs|}$$

(3) The i-th dimension of Key_m is denoted by $key_{i,m}$ and the i-th key word in $KEYs$ as key_i. $key_{i,m}$ indicates the number of times key_i appears in K_m. Specifically:

$$Key_m = \underbrace{(key_{1,m}, key_{2,m}, \ldots, key_{|KEYs|,m})}_{|KEYs|}$$

(4) The *TF-IDF* value of $key_{i,m}$ is calculated as follow:

$$TF - IDF(key_{i,m}) = \frac{key_{i,m}}{\sum_{k=1}^{|KEYs|} key_{k,m}} \cdot \lessgtr \frac{|M|}{|p : key_i \in K_p|} \quad (13)$$

where $|M|$ is the number of messages in G and $|p : key_i \in K_p|$ is the number of messages whose key words contain key_i.

(5) Updating Key_m:

$$Key_m = (TF - IDF(key_{1,m}), \ldots, TF - IDF(key_{|KEYs|,m})) \tag{14}$$

(6) Calculating similarity among message nodes: given two messages m_i and m_j, we can obtain the eigenvectors Key_{m_i} and Key_{m_j}. Utilizing the cosine similarity between Key_{m_i} and Key_{m_j} to measure the similarity between m_i and m_j, details as follow:

$$Sim(m_i, m_j) = cos(Key_{m_i}, Key_{m_j}) \tag{15}$$

where *cos* means the cosine similarity.

As previously stated, the topic based influence can be obtained:

$$Topic\_inf(u,v) = \sum_{m \in M_u, n \in M_v} U\_m\_inf(u,m) \cdot Sim(m,n) \cdot M\_u\_inf(n,v) \tag{16}$$

where M_u is the messages set u has done *EA* with.

F. Comprehensive Influence

In summary, the influence among users are the collective effect of interaction based influence, friendship based influence, tag based influence, and topic influence, which we called as *comprehensive influence*. We formally define *comprehensive influence*, denoted as $Comp\_inf_{u,v}$, as follows:

$$\begin{cases} \mathrm{Attr\_inf(u,v)} = \alpha \cdot \mathrm{Tag\_inf(u,v)} + (1-\alpha) \cdot \mathrm{Neigh\_inf(u,v)} \\ \mathrm{FP}_{u,v} = \mu \cdot \mathrm{Attr\_inf(u,v)} + (1-\mu) \cdot \mathrm{Topic\_inf(u,v)} \\ \mathrm{Comp\_inf}_{u,v} = \mathrm{II}_{u,v} + \mathrm{FP}_{u,v} \end{cases} \tag{17}$$

IV. IMGA Algorithm

IMGA(Influence maximization greedy algorithm) algorithm is designed for identifying seed set based on MIF model. IMGA aims at adding users who have maximum marginal gain into seed set, learning from the idea of greedy algorithm. Before we detail IMGA algorithm, we define marginal gain at first.

A. Marginal Gain of Users

For user u, according to MIF model, we can get users who have *comprehensive influence* on u. These users together form a collection, which is denoted by $I(u)$. From a social network perspective, the influence between two users is the result of influence spreading, while *comprehensive influence* calculated by Eq. (17) is a result from the relationship between u and v, ignoring the propagation of influence. As a result, we calculate the *global comprehensive influence* between user u and v, denoted by $\phi_{u,v}$, considering the spread of *comprehensive influence*, as follows:

$$\phi_{\mathrm{u,v}} = \begin{cases} 0 & if \;\; X \\ Comp\_\mathrm{inf}_{u,v} & if \;\; Y \\ \sum_{w \in I(v)} \phi_{u,w} \cdot Comp\_\mathrm{inf}_{w,v} & if \;\; Z \end{cases} \tag{18}$$

where condition X represents there is no propagation path of *comprehensive influence* between u and v, and Y means there is only one hop path to disseminate *comprehensive influence* between u and v. Z represents there are multi-hop paths to propagate *comprehensive influence* between u and v. Specifically, in Fig. 1, if there is a new user $u4$ who have no social relationship, no behavior records, and no tags. Then $\phi_{u1,u4}$ satisfies condition X. If $u4$ only has a social relationship with $u2$, which means $A(u2,u4) = 1$. As a result, $\phi_{u2,u4}$ satisfies condition Y. According to Eq. (18), if there is no path to spread *comprehensive influence* between u and v, the *global comprehensive influence* between them is 0. If there is one hop path to spread *comprehensive influence* between u and v, for instance, the path is $u \rightarrow v$, the *global comprehensive influence* equals to *comprehensive influence*. If there are multi-hop paths to spread influence, for instance, the paths are $u \rightarrow h1 \rightarrow \cdots \rightarrow b1 \rightarrow w1 \rightarrow v, u \rightarrow h2 \rightarrow \cdots \rightarrow b2 \rightarrow w2 \rightarrow v, \ldots, u \rightarrow h_n \rightarrow \cdots \rightarrow b - n \rightarrow w_n \rightarrow v$, the *global comprehensive influence* between them equals to the *global comprehensive influence* between u and $I(v)$ multiplies *global comprehensive influence* between $I(v)$ and v.

If we regard the $\phi_{u,u} = 1$, Eq. (18) can be summarized as:

$$\phi_{u,v} = \sum_{w \in I(v)} \phi_{u,w} \cdot Comp\_inf_{w,v} \tag{19}$$

Assuming S be the seed set we select, the *global comprehensive influence* of S can be obtained as:

$$\Phi(S) = \sum_{s \in S} \sum_{u \notin S} \phi_{s,u} \tag{20}$$

So the marginal gain of u — $\sigma(u)$ can be obtained as follows:

$$\begin{aligned} \sigma(u) &= \Phi(S+u) - \Phi(S) \\ &= \left(1 - \sum_{s \in S} \phi_{s,u}\right) \cdot \sum_{v \in (U-S)} \phi_{u,v} \end{aligned} \tag{21}$$

Unless stated otherwise, the remaining paper regard *comprehensive influence* as *global comprehensive influence*.

B. Identifying Seed Set

It has been demonstrated that the issue of computing the influence of seed set after propagation in the IC and LT models is NP-hard. Many current heuristic algorithms are proposed to effectively approximate the influence maximization problem. However, they can not guarantee the accuracy and reliability. By contrast, the use of greedy algorithm to select the seed set can get more accurate results [22]. As a result, we propose IMGA algorithm employing greedy algorithm. As previously stated in Section IV-A, we solve the influence maximization problem by selecting users who have maximum marginal gain as seed set. Steps are shown in Algorithm.1.

Algorithm 1: IMGA algorithm.

Input: $k, G(U, M, T, E), \alpha, \lambda, \mu$
Output: Seed set S

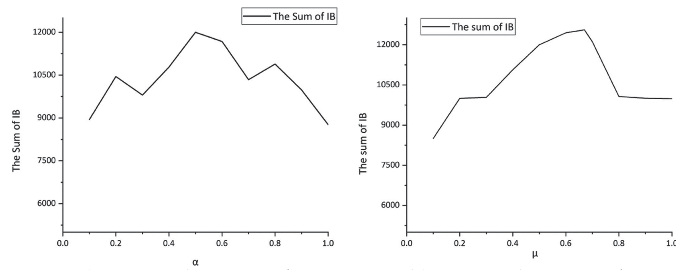
$S \leftarrow \varnothing$;
for $u, v \in U$ **do**
  $calculating\ FD_{u,v}, Neigh\_inf_{u,v}, Tag\_inf_{u,v},$
  $Topic\_inf_{u,v}, Comp\_inf_{u,v}$;
end for
$Array\ q = null$;
$Queue\ re = null$;
for $u \in U$ **do**
  $\sigma(u) = \Phi(S+u) - \Phi(S)$;
  $u.flag \leftarrow 0$;
  $q.append(\sigma(u))$;
end for
$sorted(q, ascending = False)$;
for $i \in q$ **do**
  $re.push(i)$;
end for
while $|S| <= k$ **do**
  $u \leftarrow pop(q)$;
  if $u.flag = |S|$ **then**
    $S \leftarrow S \cup u$;
    $updating\ nodes'\ marginal\ revunue\ in\ U - S$;
    $updating\ q$;
  else
    $Updating\ u's\ marginal\ revenue$;
    $u.flag \leftarrow |S|$;
    $updating\ q$;
  end if
end while
return S;

We initialize seed set S and a queue q which records users' marginal gain at first. Then, for $\forall u \in U$, $v \in N_u$, we calculate interaction based influence—$FD_{u,v}$, friendship based influence—$Neigh\_inf_{u,v}$, tag based influence—$Tag\_inf_{u,v}$, topic based influence—$Topic\_inf_{u,v}$, and comprehensive influence—$Comp\_inf_{u,v}$. For all users in G, according to Eq. (21), we compute their marginal gain, ordering marginal gains in a descending order, and storing them in queue q. IMGA recursively selects a user having maximum marginal gain as a candidate t. Since a user is selected, it is necessary to update the *global comprehensive influence* of seed set and marginal gain of users. As a result, we use *flag* to indicate the number of times marginal gain updating. If t's *flag* equals to the number of S, insert t into the seed set S, and update *global comprehensive influence* on users in $U - S$; otherwise, recalculate the marginal gain of t and insert t into q at corresponding position. After k iteration, seed node S can be obtained.

V. Experiments

Aiming at validating the accuracy of MIF model and the efficiency of IMGA algorithm, we evaluate our experiments on Flickr dataset from SNAP and an ego network dataset from [23]. We extract 4546 photos, 2662 messages and 40808 users from Flickr. However, we use all the information from ego network dataset. Ego network dataset from [23] records the concerns and interactions of 104 employees of a listed company on the internal social platform. Since there are no interaction messages in the Ego dataset, we ignore the topic based influence in this experiment. In our experiments, we try to explain four questions:

- Is the time complexity of MIF and IMGA acceptable ?
- Does heterogeneous information is important in MIF and IMGA ?
- How is the quality of MIF and IMGA ?
- How was the threshold affect our model ?



(a) α and the sum of *IB* (b) μ and the sum of *IB*

Fig. 2. Effect of parameters.

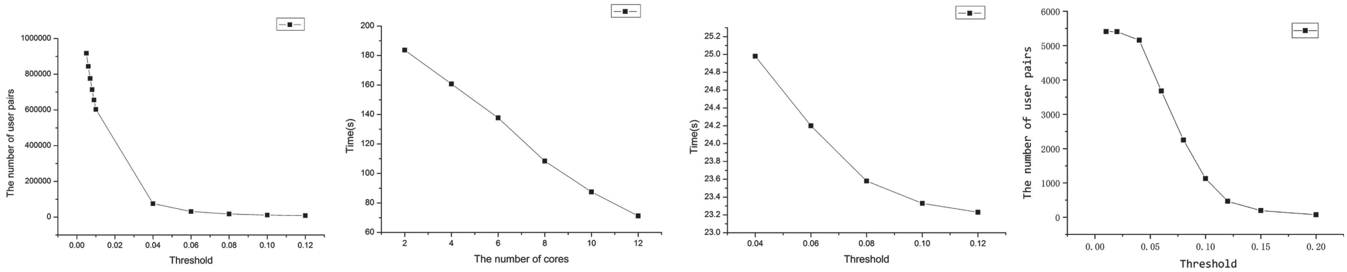
A. Parameters Settings

As above mentioned, there are some parameters in our model. This section presents the effects of those parameters. All of our parameters are determined by experiments based on the Flickr dataset. The parameters in the experiments based on the Ego Network dataset are consistent with them for a better comparison. We regard the sum of *IB* of seed set as a standard to determine our parameters.

- λ . In Eq. (4), λ measures the importance of *Motivate Probability* and *Behavior Occupancy*. The larger the λ, the more important *Motivate Probability* is in interaction based influence. We design $\lambda = 0.5$ to balance the importance of *Motivate Probability* and *Behavior Occupancy*.
- α . In Eq. (17), α measures the importance of tag based influence and friendship based influence. The larger the α, the more important tag based influence is. We conduct an experiment with $\lambda = 0.5$, $\mu = 0.5$ to analyze the correlation between α and the sum of *IB* of seed set. Fig. 2(a) shows the result. We can observe that when $\alpha = 0.5$, the sum of *IB* of seed set is the largest. Therefore, we set $\alpha = 0.5$.
- μ. In Eq. (17), μ measures the importance of topic based influence. The larger the μ, the less important topic based influence is. We also perform an experiment with $\lambda = 0.5$, $\alpha = 0.5$ to analyze the correlation between μ and the sum of *IB* of seed set. Fig. 2(b) shows the result. As a result, we set $\mu = 0.67$.

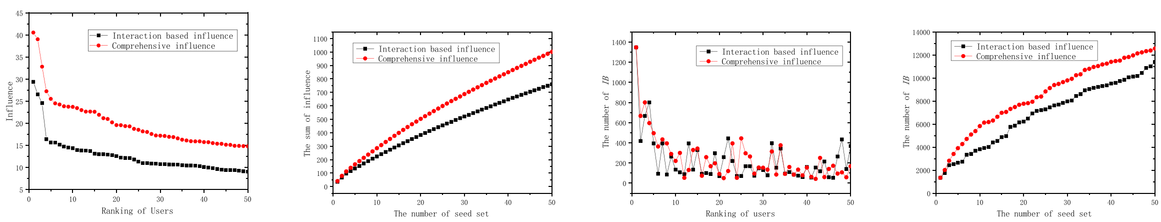
B. Time Complexity

In Flickr, we use Eq. (4) to calculate interaction based influence, obtaining 75269 user pairs have interaction based



(a) Threshold and the number of user pairs of Flickr (b) time and the number of cores (c) time and threshold (d) Threshold and the number of user pairs of Ego Network

Fig. 3. Time complexity and threshold.



(a) Influence distribution (b) The sum of influence (c) The number of *IB* (d) The sum of *IB*

Fig. 4. Flickr experiments.

influence in our result. Based on the similarity of users' social friendships, tags, and topics, we calculate $FP_{u,v}$ using Eq.(17), whose time complexity is $\mathcal{O}(n^2)$, where n is the number of nodes in social networks. However, to reduce the computation complexity of our model, we consider a threshold th when compute $FP_{u,v}$ according to Eq. (17). If $FP_{u,v}$ less than th, we consider the $FP_{u,v}$ equals to 0. To evaluate the reasonable value for th, we report the number of user pairs having $FP_{u,v}$ and the time complexity, and Fig. 3(a)–Fig. 3(c) show the results. Because the interaction data is the source data from Flickr, then we just ignore it when discuss the threshold.

We evaluate the number of user pairs having $FP_{u,v}$ by varying the value of th. In Fig. 3(a), we can clearly see a negative correlation between the threshold and the number of user pairs. When the threshold equals to 0.01, the number of user pairs MIF find is 603652 while when the threshold equals to 0.04, the number of user pairs is 75760.

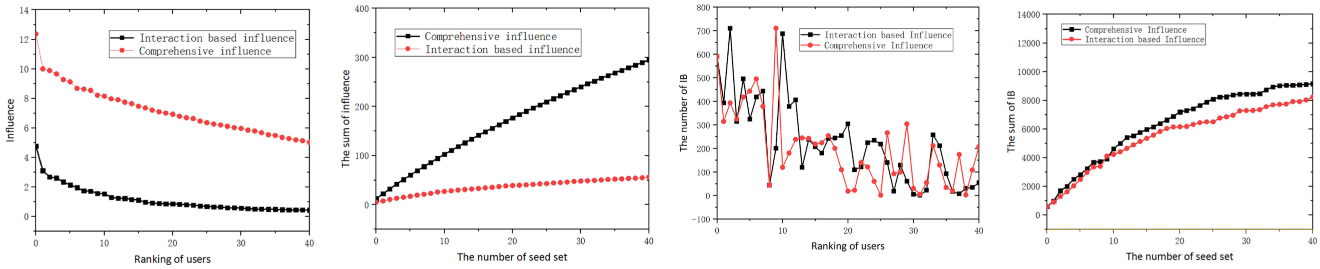
Moreover, we evaluate the time complexity by varying th. Firstly, we discuss the effect on time, with changes in computation ability which can be represented by the number of cores we use. Fig. 3(b) shows the correlation between them. We can observe that when threshold equals to 0.005, we spend 182 seconds to compute $FP_{u,v}$ if we use 2 cores. While we spend 71 seconds if we use 12 cores. As a result, the time complexity are negative affected when more cores are used. Secondly, Fig. 3(c) presents the effect on time, with changes in th. In this set of experiments, we use 12 cores. In particular, the time ranges in 23–25 seconds if we set $th \in [0.04, 0.12]$. Moreover, when $th \in [0.04, 0.12]$, user pairs MIF find ranges in 10000–70000, which are the same magnitude when stored. As a result, we set the threshold as 0.04.

In the experiments of Ego network dataset, we adopt the same strategy to analyze the time complexity. The correlation between the number of user pairs and threshold are depicted as Fig. 3(d). 5408 user pairs have comprehensive influence if no threshold is settled. It is an acceptable magnitude. As a result, threshold are settled as $th = 0$.

C. *Importance of Heterogeneous Information*

When experimenting on Flickr, the $FP_{u,v}$ uses the heterogeneous information to measure influence, and the difference of *comprehensive influence* and interaction based influence is $FP_{u,v}$. As a result, we evaluate the importance of heterogeneous information by comparing the performance of seed set, considering interaction based influence only and *comprehensive influence*. In this set of experiments, we use 12 cores of *Spark* and $\omega = \{m_1, m_2, m_3\} = \{0.3, 0.5, 0.8\}$. Fig. 4(a)–4(b) show the effect on influence and the sum of influence, respectively. In particular, according to the value of influence, we rank users in a descending order. In Fig. 4(a), there is a fast decrease in rank 1–5, but most of users have stable influence. The first 5 users have great influence both in interaction based influence and *comprehensive influence*, revealing the unbalanced influence in a social network, which is consistent with the reality.

In Fig. 4(b), the sum of influence increases with the increment on the number of seed set. However, when we select 50 users as seed set, the sum of influence are 1002 and 760, basing on *comprehensive influence* and interaction based influence, accordingly. This indicates MIF significantly boost the performance of seed set. Moreover, the differences between interaction based influence and *comprehensive influence* increases but the speed of increment decreases with the changes in the number of seed set, which means $FP_{u,v}$ can also distinguish users. It explains the importance of heterogeneous information.



(a) Influence distribution (b) The sum of influence (c) The number of IB (d) The sum of IB

Fig. 5. Ego network experiments.

To validate the real influence of our seed set in social networks, we regard the number of *IB* seed set received as a description of real influence. In Fig. 4(c), we evaluate the distribution of the number of *IB* when interaction based influence and *comprehensive influence* are considered to select seed set. However, Eq. (4) is sensitive to latency and the number of *IB* seed set received is not the only standard to measure real influence. As a result, variability in Fig. 4(c) is reasonable. Clearly, we find out that the first user receive *IB* 1346 times, which is greatly larger than other users. Moreover, there is greater volatility when rank between 1 and 5, which complies with the fast decrease in rank 1–5 in Fig. 4(a). The stable decrease in Fig. 4 (c) shows the effectiveness in selection of seed set.

We also study the sum of *IB* seed set received. In Fig. 4(d), when we select 50 users as seed set, the sum of *IB* seed set received are 12558 and 11381, which indicates better quality of seed set in our model, comparing to models which only consider interaction based influence. This also helps to illustrate the importance of heterogeneous information.

Considering Ego Network dataset, as mentioned above, we compare the interaction based influence and *comprehensive influence* to measure the importance of heterogeneous information. Fig. 5(a)–Fig. 5(b) show the effect on influence and the sum of influence, respectively. In particular, according to the value of influence, we rank users in a descending order. In Fig. 5(a), there is a fast decrease in rank 1–2, but most of users have stable influence. The first 2 users have great influence both in interaction based influence and *comprehensive influence*, which also reveals the unbalanced influence in a social network. In Fig. 5(b), the sum of influence increases with the increment on the number of seed set. However, when we select 40 users as seed set, the sum of influence are 289.7 and 54.8, accordingly, basing on *comprehensive influence* and interaction based influence, respectively. This indicates heterogeneous information significantly benefit the mining of seed set.

To validate the real influence of our seed set in social networks, we discuss the number of *IB* seed set received. In Fig. 5(c), we evaluate the distribution of the number of *IB* when interaction based influence and *comprehensive influence* are considered to select seed set. In Fig. 5(c), we can see that the number of *IB* and the rank of users is not always positive related, which is different from the result of Flickr dataset. Ego network dataset records the interactions of all employees of a listed company. The number of *IB* in a company is not a good standard to measure influence because contacting is an attribute of position. We also compare the sum of *IB* in Fig. 5 (d). The influence gap in Fig. 5(b) is very large, but the number gap in Fig. 5(d) is smaller than Fig. 5(b). It illustrates that *IB* can not measure influence effectively. However, the small gap in Fig. 5(b) indicates the heterogeneous information helps to mine seed sets more efficiently because the sum of *IB* basing on *comprehensive influence* is greater than that basing on interaction based influence. As a result, we can conclude that the heterogeneous information benefit the task of influence maximization both in Flickr and Ego Network dataset.

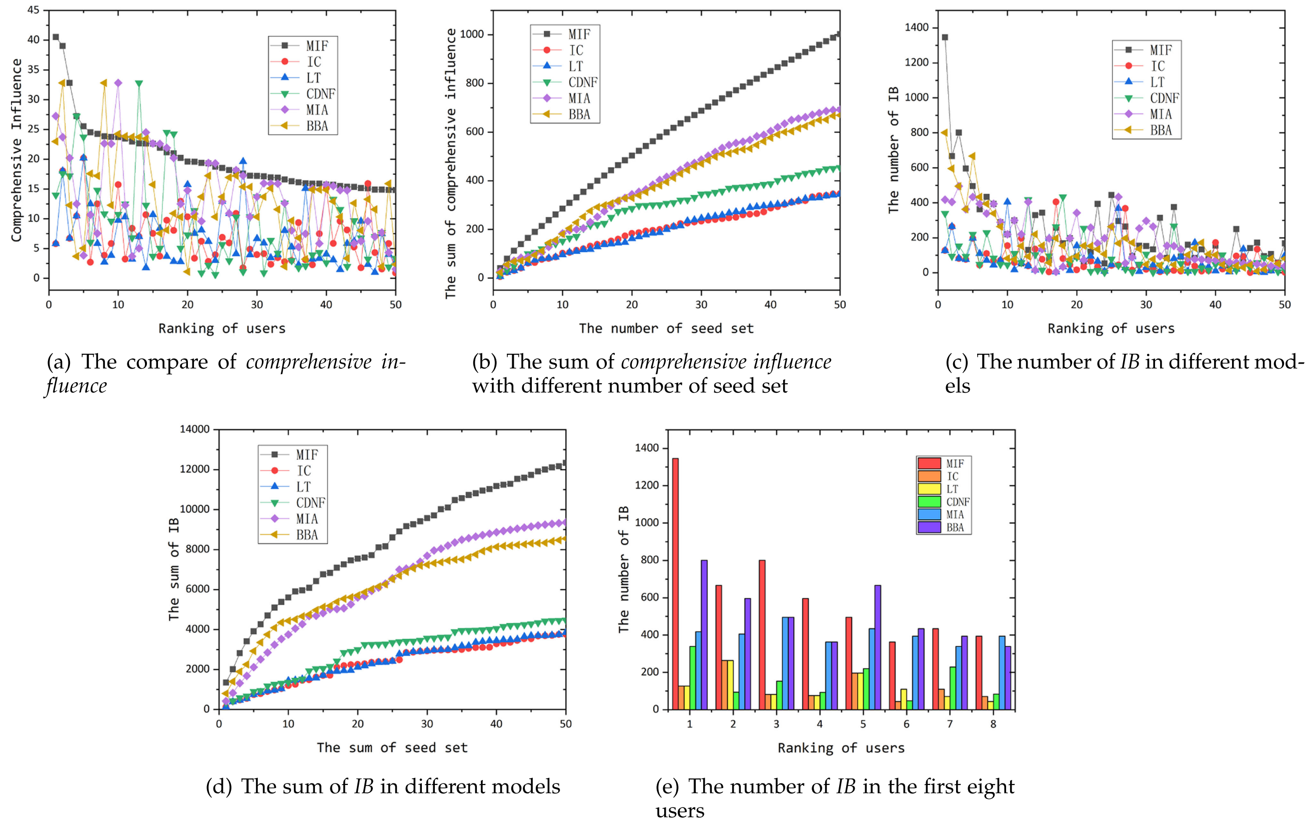
D. Performance Evaluation

we compare the performance of seed set in our model with IC [8], LT [8], CDNF [5], MIA [10] models, and BBA [22], by comparing the *comprehensive influence* and *IB* of seed set in these models. The following is a list of algorithms we compare.

- **IC model**. One baseline model. The edge weights in IC is learned by EM algorithm, while the motivate probability is learned by Eq. (22).

$$p(u,v) = \frac{1}{N(u)} \tag{22}$$

  where $N(u)$ is the in-degree of user u.
- **LT model**. Another typical model in influence maximization. It is also the baseline model.
- **CDNF model**. This model combines the constraint conditions of propagation restrictions in influence, the characteristics of the delay constraints that prevail in the process of propagation, and the credit distribution and user behavior log to construct the propagation of the influence in social networks.
- **MIA model**. This model matches the IC model restricted onto shortest path in a social network. It proposes a more efficient greedy algorithm to solve influence maximization. The motivate probability is the same as IC model for a better comparison.
- **BBA model**. This model uses the belief theory to fuse the heterogeneous information, and it has two variants. One of them is one-level BBA and the other is two-level BBA. The difference between them is that the number of hops considered in the influence propagation. The weights of edges are calculated by the method presented in this paper.



(a) The compare of *comprehensive influence*

(b) The sum of *comprehensive influence* with different number of seed set

(c) The number of *IB* in different models

(d) The sum of *IB* in different models

(e) The number of *IB* in the first eight users

Fig. 6. Evaluation of flickr.

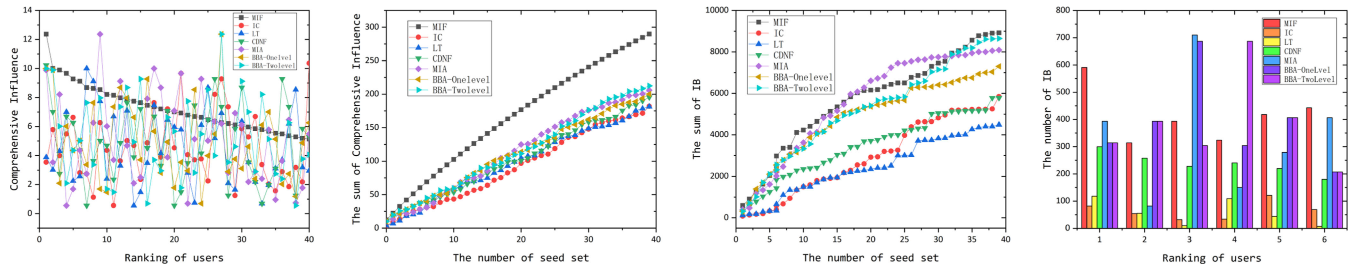
When it comes to Flickr, we use two-level BBA but ignore the one-level BBA model for comparison because of the complexity of social network in Flickr. Fig. 6(a) show the results. *Comprehensive influence* are negative affected in our model, with decrement in rank, while volatility exists in IC, LT, CDNF, MIA, and BBA model, and obviously, the first user in MIF has 40.5 *comprehensive influence*, whereas the first users in the IC, LT, CDNF, MIA, and BBA models are 5.8, 5.8, 14.0, 27.2 and 23.0 respectively, which indicates that MIF has better capabilities to find influencer in social networks.We also evaluate the sum of *comprehensive influence* of seed set by varying the number of seed set. Fig. 6(b) is performed to measure the quality of seed set. Clearly, when we select 50 users as seed set, the sum of *comprehensive influence* are 1002.6, 691.8, 669.2, 453.7, 346.6, and 344.1 in MIF, MIA, BBA, CDNF, IC, and LT models, accordingly. Hence it illustrates the ability to spread influence of seed set in MIF model outperforms other five models. We can see that MIA has better ability to spread influence than other four models.

The same as Section.V-C, we count the number of *IB* seed set received in different models. Fig. 6(c) presents the effect on the number of *IB* with the ranking changes in [0.50]. We can observe that the number of *IB* in MIF model are mostly larger than those in IC, LT, CDNF, MIA, and BBA models. In this paper, we regard the number of *IB* seed set received as a description of real influence. Therefore, larger number of *IB* explains the better ability of mining seed set.

Fig. 6(d) report the effect on the sum of *IB*, when downsizing the number of seed set. When we select 50 users as seed set, the sum of *IB* seed set received are 12336, 9362, 8552, 4474, 3759 and 3853 in MIF, MIA, BBA, CDNF, IC, and LT models, respectively. The sum of *IB* in MIF is almost three times as that in CDNF, IC, and LT. This indicates the brilliant quality of our model on Flickr dataset.

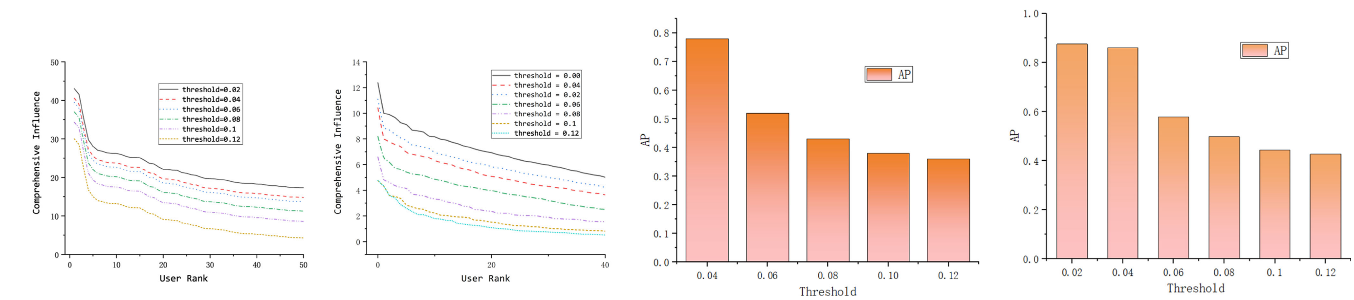
Again, according to the *comprehensive influence*, we order users in a descending order. However, selecting the first eight users in examined six models respectively, we compare the number of *IB* they received. Obviously, in Fig. 6(e), we can find that the number of *IB* in MIF model is superior than other five models in same order. Moreover, MIF is almost three times as IC, LT, CDNF, and MIA when top 1 user is considered and almost 1.5 times as BBA in average. This confirms the effectiveness of our MIF model and IMGA algorithm. However, the number of *IB* MIF acquires is smaller than BBA when ranks in 5–6 but the sum of *IB* is larger than BBA, which shows the defect in ranking users of MIF. We get the improvement of this defeat as a future work.

Taking Ego network dataset into consideration, we compare the performance of seed set in our model with IC, LT, CDNF, MIA, and two variants of BBA models. The compare of *comprehensive influence* is shown in Fig. 7(a). *Comprehensive influence* are negative affected in our model, with decrement in rank, while volatility exists in IC, LT, CDNF, MIA, and BBA model. More clearly, in Fig. 7(a), the maximum influence is ranked 1, 40, 7, 1, 9, 27, and 27 in MIF, IC, LT, CDNF, MIA, one-level BBA, and two-level BBA respectively. The maximum influence in MIF, MIA, one-level BBA and two-level BBA are the same and larger than the IC, LT,



(a) The compare of comprehensive influence (b) The sum of comprehensive influence with different number of seed set (c) The sum of *IB* with different number of seed set (d) The number of IB in the first six users

Fig. 7. Evaluation of ego network.



(a) The comprehensive influence changes with the threshold of Flickr dataset (b) The comprehensive influence changes with the threshold of Ego network dataset (c) The AP of the first eight rank of different threshold of Flickr dataset (d) The AP of the first eight rank of different threshold of Ego network dataset

Fig. 8. Discussion about threshold.

CDNF model. We can conclude that MIF and CDNF can locate the seed set more effectively. We also evaluate the sum of *comprehensive influence* of seed set by varying the number of seed set. Fig. 7(b) is performed to measure the quality of seed set. Clearly, when we select 40 users as seed set, the sum of *comprehensive influence* are 289.7, 182.0, 181.4, 195.3, 205.9, 200.1, and 212.5 in MIF, IC, LT, CDNF, MIA, one-level BBA, and two-level BBA models, accordingly. Hence it illustrates the ability to spread influence of seed set in MIF model outperforms other six models.

Similarly, we also compare the sum of *IB* received by different models. Because the figure comparing the number of *IB* is confusing, we just ignore it. Fig. 7(c) displays the result. When we select 40 users as seed set, the sum of *IB* seed set received are 8921, 5842, 4485, 5782, 8093, 7303, 8647 in MIF, IC, LT, CDNF, MIA, one-level BBA, and two-level BBA, accordingly. MIF get the largest sum of *IB* again. To make a more clearer comparison, we select the number of *IB* received of the first six users in examined seven models for comparison. Fig. 7(d) shows that MIF gets the largest *IB* received in rank 1, 5 and 6. However, when ranking in 2–4, MIF acquires a smaller the number of *IB* than some other models, showing the flaw in ranking users of MIF again. But MIF still gets the maximum sum of *IB*. Therefore, we can still say MIF is an effective model for influence maximization.

E. Discussion About Threshold

As mentioned above, we set a threshold to reduce the complexity of our model. To discuss how the threshold affects our model, we also conduct an experiment. The result is shown in Fig. 8(a)–Fig. 8(b). By changing the threshold, the correlation between the comprehensive influence and user rank is displayed. As we observed, the comprehensive influence decreases with the increment of threshold, which is reasonable because the threshold sets some influence to be zero. Because of the large scale of Flickr dataset, we ignore the experiment with a threshold of zero.

Aiming at discovering the impact of threshold on our rank quality, we select the rank of users with different thresholds to indicate their consistency. To quantify the consistency, we calculate the AP (Average Precision). The AP can be calculated by the following equation [24].

$$AP = \frac{\sum_{j=1}^{n_i} P(j) \cdot y_{i,j}}{\sum_{j=1}^{n_i} y_{i,j}}$$

$$P(j) = \frac{\sum_{k:\pi(k)\leq\pi_i(j)} y_{i,k}}{\pi_i(j)}$$

where $\pi_i(j)$ is the real rank of j and $y_{i,k}$ represents k is relevant or irrelevant. More details can be found in [24].

In experiments based on Flickr dataset, we consider the order of the first eight user with a threshold of 0.02 as the ground truth. Then the AP can measure the accuracy of the new rank. In Fig. 8(c), when the threshold is set to less than 0.04, we can know that the AP is greater than 78%. When the threshold is 0.06, the AP drops to 49%. Therefore, setting the threshold to 0.04 in Section V-B to make a trade-off between complexity and precision is acceptable. In experiments based on Ego network dataset, we consider the order of the first six user with a threshold of zero as the ground truth. In Fig. 8(d), when the threshold is set to less than 0.04, we can know that the AP is greater than 83%. When the threshold is 0.06, the AP drops to 58%. Therefore, 0.04 is also an acceptable threshold in this experiment. The size of Ego network dataset is not very large, we ignore the threshold setting in the experiments in Section V-B. The purpose of this experiment is to supplement the discussion of the impact of threshold.

VI. Conclusion

We present MIF, a model with heterogeneous information in social networks, aiming at mining social influence from interaction based influence, friendship based influence, tag based influence, and topic based influence perspectives. We introduce an algorithm IMGA to simulate the propagation of influence and select seed set effectively. In particular, we validate the effectiveness of MIF and IMGA on real-world dataset Filckr and Ego Network dataset. Our experiments demonstrate the superiority of our MIF model and IMGA algorithm over the IC, LT, CDNF, MIA, and BBA models. Users in social networks tend to express their emotion when commenting, which is another information we can utilize. As future work, we plan to investigate users' sentiment in comments, to describe influence more accurate in social networks. We believe this will help to improve the accuracy of seed set. Moreover, the dynamic changes in networks may affect the users interactions, which is another direction of our work.

References

[1] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun. 2009, pp. 199–208.

[2] S. Liu, L. Ying, and S. Shakkottai, "Influence maximization in social networks: An ising-model-based approach," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2010, pp. 570–576.

[3] D. Li, C. Wang, S. Zhang, G. Zhou, D. Chu, and C. Wu, "Positive influence maximization in signed social networks based on simulated annealing," *Neurocomputing*, vol. 260, pp. 69–78, 2017.

[4] Y. Lin and J. C. Lui, "Analyzing competitive influence maximization problems with partial information: An approximation algorithmic framework," *Perform. Eval.*, vol. 91, pp. 187–204, 2015.

[5] X. Deng, Y. Pan, Y. Wu, and J. Gui, "Credit distribution and influence maximization in online social networks using node features," in *Proc. IEEE 12th Int. Conf. Fuzzy Syst. Knowl. Discovery,* Aug. 2015, pp. 2093–2100.

[6] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,* Jul. 2002, pp. 61–70.

[7] X. Li, D. Yang, X. Liu, and X. M. Wu, "Bridging time series dynamics and complex network theory with application to electrocardiogram analysis," *IEEE Circuits Syst. Mag.*, vol. 12, no. 4, pp. 33–46, Fourthquarter 2012.

[8] D. Kempe, J. Kleinberg, and Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,* Aug. 2003, pp. 137–146.

[9] Y. Pan, X. Deng, and H. Shen, "Credit distribution for influence maximization in online social networks with time constraint," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom*, Dec. 2015, pp. 255–260.

[10] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 1029–1038.

[11] W. Chen, W. Lu, and N. Zhang, "Time-critical influence maximization in social networks with time-delayed diffusion process," in *Proc. 26th AAAI Conf. Artif. Intell.*, Jul. 2012, pp. 1–26.

[12] K. Rahimkhani, A. Aleahmad, M. Rahgozar, and A. Moeini, "A fast algorithm for finding most influential people based on the linear threshold model," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1353–1361, 2015.

[13] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification and shrinking diameters," *ACM Trans. Knowl. Discov.*, vol. 1, no. 1, Mar. 2007. Art. no. 2.

[14] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 88–97.

[15] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jun. 2009, pp. 199–208.

[16] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. Int. Conf. Knowl. Intell. Inf. Eng. Syst.*, Sep. 2008, pp. 67–75.

[17] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Jul. 2009, pp. 203–210.

[18] Y. Q. Zhang, J. Cui, S. M. Zhang, Q. Zhang, and X. Li, "Modelling temporal networks of human face-to-face contacts with public activity and individual reachability," *Eur. Phys. J. B*, vol. 89, no. 2, pp. 1–26, 2016.

[19] H. Ma, M. R. Lyu, and I. King, "Learning to recommend with trust and distrust relationships," in *Proc. 3rd ACM Conf. Recommender Syst.*, Oct. 2009, pp. 189–196.

[20] X. Li, P. Yao, and Y. Pan, "Towards structural controllability of temporal complex networks," in *Proc. Complex Syst. Netw.*, 2016, pp. 341–371.

[21] S. Rendle and L. Schmidt-Thieme, "Pairwise interaction tensor factorization for personalized tag recommendation," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, Feb. 2010, pp. 81–90.

[22] S. Jendoubi, A. Martin, L. Litard, H. B. Hadji, and B. B. Yaghlane, "Two evidential data based models for influence maximization in Twitter," *Knowl. Based Syst.*, vol. 121, pp. 58–70, 2010.

[23] J. Yuan *et al.*, "Promotion and resignation in employee networks," *Physica A, Statistical Mechanics Appl.*, vol. 444, pp. 442–447, 2016.

[24] H. Li, "A short introduction to learning to rank," *IEICE Trans. Inf. Syst.*, vol. 94, no. 10, pp. 1854–1862, 2011.

[25] Y. Gao, R. Ji, P. Cui, Q. Dai, and G. Hua, "Hyperspectral image classification through bilayer graph-based learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2769–2778, Jul. 2014.

[26] Y. Dong *et al.*, "Link prediction and recommendation across heterogeneous social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 181–190.

[27] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 420–429.

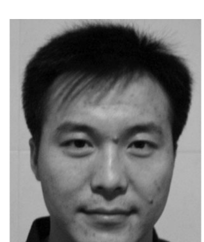
[28] Y. Liu, A. Liu, N. Xiong, T. Wang, and W. Gui, "Content propagation for content-centric networking from location-based social networks," *IEEE Trans. Syst., Man, Cybern., Syst*, vol. 49, no. 10, pp. 1946–1960, Oct. 2019.

[29] Z. Li, Y. Liu, A. Liu, S. Wang, and H. Liu, "Minimizing convergecast time and energy consumption in green internet of things," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2018.2844282.

[30] Y. Liu, M. Ma, X. Liu, N. Xiong, A. Liu, and Y. Zhu, "Design and analysis of probing route to defense sink-hole attacks for Internet of Things security," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 356–372, Jan.-Mar. 2020.

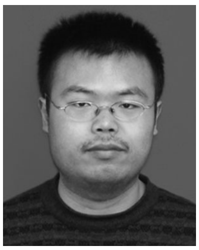
[31] D. Li, Z. Chen, and J. Liu, "Analysis for behavioral economics in social networks: An altruism-based dynamic cooperation model," *Int. J. Parallel Program.*, vol. 47, pp. 686–708, 2018.

[32] J. Liu, W. Gao, D. Li, S. Huang, and H. Liu, "An incentive mechanism combined with anchoring effect and loss aversion to stimulate data offloading in IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4491–4511, Jun. 2019.



Xiaoheng Deng was born in 1974. He received the Ph.D. degree from Central South University, Changsha, China, where he is working as a Professor. His research interests include several areas in wireless communications and networking, especially on network transmission and routing optimization with cross-layer design method. He is a senior member of China Computer Federation.

Fang Long was born in 1994, and is currently student majors in computer science and technologyfrom the Department of Information Science and Engineering, Central South University, Changsha, China. Her research interests focus on recommender systems and social networks.



Bo Li was born in 1988. He received the Ph.D. degree from the School of Information Science and Engineering, Central South University, Changsha, China. His research interests are data mining and social networks.

Dejuan Cao was born in 1992. She received the B.Sc. degree in 2014, and the M.Sc. degree in information and communication engineering from Central South University, Changsha, China. Her research interests are influence maximization in social networks and big data analysis.

Yan Pan was born in 1990. He received the B.Sc. degree in 2013, and the M.Sc. degree from the School of Information Science and Engineering, Central South University, Changsha, China. His research interest is influence maximization in social networks.