Maximizing the Spread of Effective Information in Social Networks

Haonan Zhang, Luoyi Fu, Jiaxin Ding, Feilong Tang, Yao Xiao, Xinbing Wang, Guihai Chen, and Chenghu Zhou

Abstract—Influence maximization through social networks has aroused tremendous interests nowadays. However, people's various expressions or feelings about a same idea often cause ambiguity via word of mouth. Consequently, the problem of how to maximize the spread of "effective information" still remains largely open. In this paper, we consider a practical setting where ideas can deviate from their original version to invalid forms during message passing, and make the first attempt to seek a union of users that maximizes the spread of effective influence, which is formulated as an Influence Maximization with Information Variation (IMIV) problem. To this end, we model the information as a vector, and quantify the difference of two arbitrary vectors as a distance by a matching function. We further establish a process where such distance increases with the propagation and ensure the recipient whose vector distance is less than a threshold can be effectively influenced. Due to the NP-hardness of IMIV, we greedily select users that can approximately maximize the estimation of effective propagation. Especially, for networks of small scales, we derive a condition under which all the users can be effectively influenced. Our models and theoretical findings are further consolidated through extensive experiments on real-world datasets.

Index Terms—Social network, influence maximization, information variation, greedy algorithm

1 Introduction

The paradigms of viral marketing [3] and word of mouth via social networks (SNs) such as Facebook, Twitter and WeChat have gained large popularity during the last few decades. Motivated by these applications, the *influence maximization* (IM) problem [1], [6], [10], [13], [15] has been intensively studied. Conventional IM problem aims at finding a set of seed users to maximize the influence of an idea or a target product over an SN. As an example of the viral marketing, a company wants to market a new product over an SN. They may select some influential individuals as seeds to accept and experience their product, hoping the product can be recommended to a larger population via their social behaviours.

- *Haonan Zhang, Luoyi Fu, Jiaxin Ding, Yao Xiao, Xinbing Wang, and Guihai Chen are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: {zhanghaonan, yiluofu, jiaxinding, 119033910058, xwang8}@sjtu.edu.cn, gchen@cs.sjtu.edu.cn.*
- *Feilong Tang is with Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: tang-fl@cs.sjtu.edu.cn.*
- *Chenghu Zhou is with the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100045, China. E-mail: zhouch@lreis.ac.cn.*

Manuscript received 25 April 2021; revised 16 December 2021; accepted 23 December 2021. Date of publication 28 December 2021; date of current version 7 March 2023.

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1003000, in part by NSF China under Grants 42050105, 62020106005, 62061146002, 61960206002, 61822206, 61832013, and 61829201, in part by Tencent AI Lab RhinoBird Focused Research Program under Grant JR202132, and in part by the Program of Shanghai Academic/Technology Research Leader under Grant 18XD1401800.

(Corresponding author: Xinbing Wang.)

Recommended for acceptance by X. Xiao.

Digital Object Identifier no. 10.1109/TKDE.2021.3138783

Existing literature [6], [13] always implicitly assume that the ideas which are meant to be conveyed by the original information are never changed during the information propagation. That is to say, the influence remains the same as information is being propagated among users. However, in real-world social communications with the effect of word of mouth, the vast majority of users may not convey the ideas uniformly but may express their subjective thoughts based on what they received, which may damage the effectiveness of the original information. For ease of understanding, let us illustrate with an example. Consider that Apple Inc. wants to promote the latest version of smart phone, i.e., *iphone 12* which provides the full-screen with high-resolution via viral marketing. In doing so, Apple Inc. wants to select and persuade some influential seed users in an SN to accept the phone. Suppose Lee is one of the seed users, he may recommend the phone to his friend Young with telling him that "Apple's new full screen mobile phone feels good." According to Young's experience for his last broken full screen phone, he tells Lisa "The new iphone may be good but not durable." Then, Lisa tells Bob "The new iphone is possibly short-life." Naturally, Bob will not recommend the iphone to his friends even if he talks about *iphone 12* with them due to his worry about the durability. In this example, Lee, Young and Lisa are possibly attracted by this smart phone while Bob and his friends are not attracted by the influence from Lee due to the negative comment they received. As a result, the information propagated by Lee and Young which is near from the source (Lee) benefits for marketing and attracts their consumption, and can be treated as effective information for the promotion of the *iphone 12*. In contrast, the information passed by Lisa and

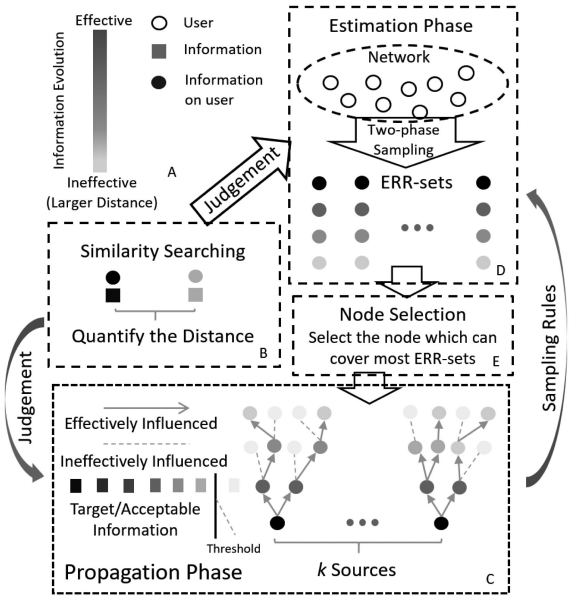


Fig. 1. Procedure of SES.

Bob may damage the reputation of the product, thus being ineffective information for viral marketing. Obviously, there is an effective range of the influence under which the recipients out of the range (friends of Bob, and their friends, etc) can never receive the effective information from the source (Lee). There are numerous analogous examples of information variation in real viral marketing scenarios. Therefore, it is desirable to carefully select the seed users who can maximize the influence of effective information/original ideas. Accordingly, we call this problem *Influence Maximization with Information Variation* (IMIV) which is different from conventional IM problem due to the following reasons:

Conventional IM problem [1], [6], [15], [33], [34] considers that the influence transmission between any two network users is independent and random, leading to all the recipients having the same influence. However, as noted earlier, when the users pass some information or ideas to their friends, they rarely convey the thought in exactly the same way with what they received due to their different expression abilities, experiences, etc. Such information deviation in terms of its original meaning may further cause the different influence received by different users. Even though there are some literatures [2], [20] trying to generalize the "positive - negative" influence and the competition between them, the assunptions and the models they proposed are still hard to truly reflect this kind of generative process of the "nonpositive" influence. Accordingly, in order to completely overcome this kind of problem, we have to capture how such changes happen with the propagation and quantify the differences between the information conveyed by different users to recognize how much the information deviates from the original version. Consequently, we can no longer regard all the influenced individuals in the conventional settings as the users who can receive the information that preserves the attributes in the original version. In this way, the target, instead of simply counting the number of information recipients as prescribed in traditional IM problems, turns to the number of recipients whose information remains effective.

To address above challenges, we propose a framework — *Searching, Estimating and Selecting* (SES) which includes three modules — distance quantification, estimation and node selection to solve IMIV problem. The whole framework is illustrated in Fig. 1. While we defer the details of SES design in later sections (Sections 2 and 3), here we briefly unfold its three major components in addressing the aforementioned challenges in IMIV. First, we build a mapping from an arbitrary piece of information to a multidimensional information vector, each dimension of which represents an attribute of the information (A in Fig. 1). The difference of the attributes included in two different pieces of information is further quantified with a distance function, which can detect the similarity between the elements of two information vectors that can be of different dimensions [17] (B in Fig. 1). Since the distance function increases with the spread of information, we establish a propagation model to ensure that the recipients whose vector distance with the source is less than a given threshold can be effectively influenced (C in Fig. 1). The NP-hardness of IMIV drives the necessity of proposing an approximation algorithm of node selection that can maximize the spread of effective information mentioned above. The essence of the algorithm is to generate an approximate observation for the effective information propagated via the above described way by the technique of Effective Reverse Reachable (ERR) Sets (D in Fig. 1), and then greedily select users based on such ERR sets (E in Fig. 1). The key contributions of our work are summarized as follows.

- *Design:* We first define and analyze the IMIV problem, and further propose a solution with a greedy algorithm framework which mainly contains three components — similarity searching for detecting a matching distance of two pieces of information, estimation for the effective influence propagation and node selection on the estimators to find the users who can approximately maximize the propagation of effective influence.
- *Analysis:* We theoretically analyze the implementation of three modules in our SES framework. Meanwhile, the framework, as a whole, enjoys provable approximation ratio and time complexity. Moreover, we derive a necessary condition when all the network users can be effectively influenced in small-scale networks.
- *Validation:* We find the probability distribution of real-world datasets is consistent with our model, which further indicates that our ideas are effective. Additionally, the results of experiments on the real-world datasets confirm the superiority of our approach for IMIV in the sense of that its capacity of maximizing the spread of effective information signally outperforms the baselines.

The roadmap of this paper is as follows. In Section 2, we describe the process of effective information propagation with a quantified influence evolution model. In Section 3, we design greedy-based algorithms to solve the IMIV problem. Some related theoretical analysis is also proposed in

this section. In Section 4, the effectiveness of our model is verified on real-world dataset. Further, the performance of our algorithm is also evaluated via experiments on real networks. In Section 5, we introduce some related works of our paper. We further conclude this work and point out some future directions in Section 6. Finally, some proofs, if are not in line, are available in our Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2021.3138783.

2 Models and Problem Definition

A social network can be modeled as a directed graph $G = (V, E)$ with N nodes and M edges. Here $V = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes (users), $E = \{e_{ij} | i, j = 1, \ldots, N\} \subseteq V \times V$ is the set of weighted edges (social relationships) among users, on which the information propagation happens. Given a graph, we first propose a model to capture the information variation pattern and quantify the differences of the information. On this basis, the propagation model and the problem formulation are unfolded.

2.1 Information Evolution Model

In order to capture the pattern of information variation, we first represent an arbitrary piece of information as an information vector which is defined as follows.

Definition 1. (Information Vector) *Arbitrary information can be represented as a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ whose elements satisfy $a_i \in (l, u)$ with $l, u \in \mathbb{R}$, $i \in \mathbb{N}^*$ reflecting the features of the attributes in that piece of information.*

Here the attributes of a message[1] may include time, location, activity etc. Based on Definition 1, we use $\boldsymbol{a}^0$ to represent the (original) information carried by the seed nodes, while $\boldsymbol{a}^j$ is the information vector after j steps' variation from $\boldsymbol{a}^0$. (How the propagation affects the information variation (evolution) will be disclosed in Section 2.2.) Note that the value of $\boldsymbol{a}^j$ can be different in any independent process of the variation, which is led by the randomness of the evolution. Definition 2 characterizes how an information vector evolves between different steps.

Definition 2. (Information Vector Evolution Model) *In this model, we represent the evolution of $\boldsymbol{a}^j \rightarrow \boldsymbol{a}^{j+1}$ as a mapping that is divided into 3 cases — deviation of each dimension, mutation and loss for the dimensions of the information vectors, which reflect three different ways of attributes changing for the passed information in the real networks. Denoting the number of elements in $\boldsymbol{a}^j$ as m^j, we have*

- case 1*: Each element of the vector tends to deviate from its value with a (half) Gaussian distribution, i.e., $(a_i^{j+1} - a_i^j) \sim N(\Delta d_i, \sigma_i^2), i \in [1, m^j]$, with probability p_1^i.*[2]
- case 2*: In vector $\boldsymbol{a}^{j+1}$, a new dimension which obeys a (discrete) uniform distribution may appear from vector $\boldsymbol{a}^j$, i.e., $a_{m^j+1}^{j+1} \sim U(l, u)$ with a small probability p_2.*[3]
- case 3*: In vector $\boldsymbol{a}^{j+1}$, a dimension from vector $\boldsymbol{a}^j$ may be dropped with a small probability p_3. The probability of an element in $\boldsymbol{a}^j$ being dropped is $\frac{1}{m^j} \times p_3$.*[4]

The rationality of this model will be empirically disclosed in Section 4. In fact, similar modeling methods are widely appearing in the literatures [2], [27], [28]. Here, we use *case 1* to capture the regular evolution pattern of the elements in a vector. The state space of an element can be regarded as a Markov chain, where the current state of an attribute only depends on the last step. Further, mutations with probability are applied to complete our design of *case 2* and *case 3*. Due to Markov chains and mutation phenomena are always the objective and universal laws of the physical world, we use them here to define our model. Moreover, in literature like [38], the Gaussian distribution is always used to capture the probability distribution of node features and attributes. To keep the consistence for the description of the distribution with considering the data deficiencies in the real networks, the (half) Gaussian distribution is also used in our model. To facilitate the understanding of the 3 cases in Definition 2, let us recall the example in Section 1 first. When Young passes the message to Lisa (the process from evolution step 1 to step 2), he adds an attribute "may be not durable" to the information (corresponds to *case 2*). Lisa forgets "The phone is good." (corresponds to *case 3*) when she further passes the information to Bob (evolution steps 2 to 3). Moreover, Lisa derives the attribute from "may be not durable" to "possibly short-life" (corresponds to *case 1*) in the same step. Note that the parameters Δd, p_1, p_2 and p_3 capture the tendency of whether the variation tends to happen on the propagation of this network, under which a larger Δd, p_1, p_2 or p_3 means the variation is more likely to happen.

2.2 Propagation Model

Upon the modeling of information evolution, we continue to capture how the information evolves with its propagation. In order to measure the changes of the information during its transmission among network users, a key part is to depict the difference of the information variation at two arbitrary evolution steps. However, due to the information vector evolution model, the two information vectors can be faced with the problem of different dimension, which is caused by the element dropping and new dimension generating described in Definition 2. Thus, we are motivated to find a function which can calculate the distance between the vectors of different dimensions. To this end, we use the vector matching distance in Definition 3 to quantify such difference of two information vectors.

Definition 3. (Vector Matching Distance) *Suppose $\boldsymbol{a}^q$ and $\boldsymbol{a}^p$ represent the information vectors that are evolved at two intermediate steps q and p. Then the matching distance function $D(\boldsymbol{a}^q, \boldsymbol{a}^p)$ of these two vectors can be defined as*

$$\begin{aligned} D(\boldsymbol{a}^q, \boldsymbol{a}^p) &= d\big(\|\boldsymbol{a}^q\|_0, \|\boldsymbol{a}^p\|_0\big) \\ &= \big|a_{m^q}^q - a_{m^p}^p\big| + \min\begin{cases} d\big(\|\boldsymbol{a}^q\|_0, \|\boldsymbol{a}^p\|_0 - 1\big) \\ d\big(\|\boldsymbol{a}^q\|_0 - 1, \|\boldsymbol{a}^p\|_0\big) \\ d\big(\|\boldsymbol{a}^q\|_0 - 1, \|\boldsymbol{a}^p\|_0 - 1\big) \end{cases}, \end{aligned} \tag{1}$$

1. In this paper, we use information, messages etc. interchangeably.

2. Δd, σ and p_1 will be used to denote the expectation of the parameters.

3. We will use $a_{m^j+1}^{j+1} = \xi, [\xi \in (l, u)]$ with p_2 to represent this case.

4. We will represent this case as $a_i^{j+1} = 0$ with p_3.

here, l_0 norm is used to represent the effective length (the number of elements which are not 0) of a vector and

$$d(0,0)=0, d(i,0)=d(0,j)=\infty, \quad i=1,2,\ldots,\|\boldsymbol{a}^q\|_0, j=1,2,\ldots,\|\boldsymbol{a}^p\|_0. \tag{2}$$

This design is motivated by [17], [19] which use similar functions to detect the difference and similarity of two sequences. Intuitively, different choices on the computation $\|a^q_{m^q}-a^p_{m^p}\|$ such as $|a^q_{m^q}-a^p_{m^p}|$ or $(a^q_{m^q}-a^p_{m^p})^2$ may influence the final distance result. However, in [17], it has been proved that the selection of function $\|\cdot\|$ does no affect on the relative distance result. That is to say, if the matching distance with $|\cdot|$ between two vectors is greater than the other two vectors, then this relation still holds when $(\cdot)^2$ is used. According to this property, the choices on $\|\cdot\|$ will do no affect on our work. In this paper, in order to facilitate the calculation, absolute value $|\cdot|$ is used as a form of $\|\cdot\|$, i.e., $\|a^q_{m^q}-a^p_{m^p}\|=|a^q_{m^q}-a^p_{m^p}|$.

By leveraging the distance function in Definition 3, we can further measure the difference between an arbitrary piece of information and the original version to distinguish the effective information, as formally defined in Definition 4.

Definition 4. (Effective Information) *For a piece of information which corresponds to the vector $\boldsymbol{a}^p$ in an intermediate step p, it is effective if the vector $\boldsymbol{a}^p$ satisfies $D(\boldsymbol{a}^0,\boldsymbol{a}^p)\le T$.*

Here, T denotes a distance threshold within which the information is effective during the propagation. Meanwhile, the users who have received the effective information are called as being effectively influenced. As we have mentioned earlier, whether the information is effective or not is closely correlated with propagation process. Hence, we need to establish a model that can present the transmission between the users based on the distances between the information they receive. Our model is inspired from a conventional information propagation model called the *Independent Cascade* (IC) model. Before we formally present our model, we need to review the property of IC model first. Note that we adopt discrete time processes, where each time slot represents a segment of a propagation.

Definition 5. (Independent Cascade Model) *[6] In an IC model, the propagation of the influence among users is characterized by probability. Especially, when a user v_i is infected,[5] he has a single chance to infect his neighbor v_j successfully with a probability p_{ij} via their common edge e_{ij}. Such process is independent from history and other nodes.*

However, in the conventional IC model, each recipient possesses the same influence, thus inapplicable to directly capturing the propagation model that should incorporate the evolution of effective information. Based on Definitions 2, 3 and 4, we refine the IC model and come up with the *Effective Independent Cascade* (EIC) model given in Definition 6.

Definition 6. (Effective Independent Cascade Model) *In an EIC model, the effective influence among users is composed of two phases, i.e., IC phase and effective propagation phase. First, a piece of information is propagated from a user v_i to another user v_j as prescribed by the IC model. Then, for the infected user v_j, whether it is effectively influenced, or alternatively, whether the effective information passed from v_i to v_j still remains to be effective, is characterized by the probability $p(D(\boldsymbol{a}^t,\boldsymbol{a}^0)\le T)$, where $\boldsymbol{a}^t$ is the information vector that v_j receives.*

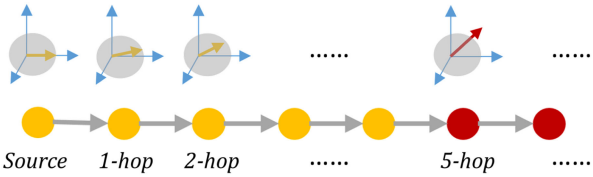


Fig. 2. Vector evolves with information propagation.

To make the EIC model clear, we first explain how to compute $p(D(\boldsymbol{a}^t,\boldsymbol{a}^0)\le T)$. Note that, this probability distribution depends on the 3 cases in Definition 2 and distance function in Definition 3. However, in Definition 2, although the PDF (probability density function) of Case 1 and Case 2 is continuous, it will bring discontinuity to the distance function as $(a_i^{j+1}-a_i^j)$ changes. (This illustration will be disclosed in Section 3.1.) Moreover, Case 3 provides a random function which may bring uncertainty for the distance function. Therefore, it is hard to build an explicit expression for $p(D)\le T$. Thus, we design an advance diffusion process which will be unfolded detailedly in Section 3.1 to estimate the probability $p(D(\boldsymbol{a}^t,\boldsymbol{a}^0)\le T)$ so as to give a prior activation probability for the propagation process like the activation probability defined in IC model. In EIC model, this prior probability only relies on the propagation step t. From this point of view, in the two steps of the EIC model, not only the IC activation probability is provided to capture the propagation relationship between two adjacent nodes, but also the effective activation probability is derived to depict the propagation of effective influence.

We further interprete how EIC model works by an illustration of a graph with linear topology in Fig. 2. Suppose in this graph there is a one dimensional information vector to be propagated from the source node (the leftmost node in Fig. 2). There is an element, i.e., a new dimension generated for the vector in the 3rd and 5th steps of the evolution. In the 5th step, the deviation of the vector exceeds the threshold. Therefore, here, the effective influence spread from a source only happens within the first 4 hops. From the user perspective, a node will be activated iff a piece of effective information is propagated on it. For the example in Fig. 2, once a node receives the information within 4-step variation, it will be activated. Otherwise, it will keep unactivated. Moreover, as long as a node is activated, it will keep this state, which can extend above example in more general and complicated cases.

2.3 Problem Formulation

Now we are ready to formulate the problem. Given a directed graph $G(V,E)$, and a piece of original information which is to be spread in the network with EIC model, the problem is formulated as *Influence Maximization with Influence Variation* (IMIV) which aims to select at most k nodes from V into a seed set S that satisfies

5. The concept of "node is infected/uninfected" is equivalent to "user is influenced/uninfluenced". Note that if a node is infected, it will keep this state during this propagation process.

$$\text{argmax}_{S\subseteq V}\mathbb{E}[|\sigma(S)|], \tag{3}$$

$$\text{s.t. } |S| \leq k, \tag{4}$$

$$\mathbb{E}[|\sigma(S)|] = \sum\nolimits_{X\in\Omega} \Pr[X]\cdot|\sigma(S|X)|. \tag{5}$$

Here, $\sigma(S)$ is a set function representing the node set that is effectively influenced by the seed set S. We use expectation $\mathbb{E}[|\sigma(S)|]$ as our optimization objective to account for the randomness in the propagation. X is a sample of the sampling space Ω which is a collection of all the possible cases of edge existence resulted from EIC model. $\Pr[X]$ is the occurrence probability of sample X while $|\sigma(S|X)|$ is the number of nodes effectively influenced by the seed set S conditioned on the edge existence set X. From the Equations (3), (4) and (5), it can be easily found that IMIV problem shares the same form but different objective function with conventional IM problems. Hence, we need to reconsider the properties of the objective function and consequently redetermine the NP-hardness of the IMIV problem. First, we discuss some properties of the propagation function $\sigma(S)$ in Lemma 1.

Lemma 1. *σ is a monotone submodular function.*

Proof. We suppose there are two node sets $S_1 \subseteq S_2$ with $\sigma(S_1) \subseteq \sigma(S_2)$. When S_2 and S_1 are used as the seed set respectively, the number of the final influenced nodes triggered by S_2 is obviously larger than that from S_1. Therefore, σ is monotone.

Now we demonstrate the submodularity via a reverse method. Suppose that the function is not submodular, i.e., $|\sigma(S_1 \bigcup v)| - |\sigma(S_1)| \geq |\sigma(S_2 \bigcup v)| - |\sigma(S_2)|$ is unsatisfied. Therefore, there exist two sets $S_1^* \subseteq S_2^*$ and a node v^* satisfying $|\sigma(S_1^* \bigcup v^*)| - |\sigma(S_1^*)| < |\sigma(S_2^* \bigcup v^*)| - |\sigma(S_2^*)|$.

$$\begin{aligned}
&|\sigma(S_1^* \bigcup v^*)| - |\sigma(S_1^*)| < |\sigma(S_2^* \bigcup v^*)| - |\sigma(S_2^*)| \\
\Leftrightarrow &|\sigma(v^*)| - |\sigma(v^*)\bigcap\sigma(S_1^*)| < |\sigma(v^*)| - |\sigma(v^*)\bigcap\sigma(S_2^*)| \\
\Leftrightarrow &|\sigma(v^*)\bigcap\sigma(S_1^*)| > |\sigma(v^*)\bigcap\sigma(S_2^*)| \\
\Leftrightarrow &\sigma(v^*)\bigcap\sigma(S_1^*) \supset \sigma(v^*)\bigcap\sigma(S_2^*) \\
\Leftrightarrow &\sigma(S_1^*) \not\subseteq \sigma(S_2^*).
\end{aligned} \tag{6}$$

That means monotone property is violated and assumption fails. Thus, function σ is submodular. □

After the problem IMIV and the properties of the objective function are given, the hardness of IMIV problem should be studied to further find a solution for it. For this purpose, Theorem 1 is derived.

Theorem 1. *The IMIV problem is NP-hard.*

Proof. The NP-hardness is proved through a reduction from an arbitrary set cover problem (SCP) which has been proved to be NP-complete [6] to an IMIV problem. Specifically, as an example, consider $S_1, S_2, S_3, \ldots, S_m$ which are the m subsets of a ground set $\{v_1, v_2, v_3, \ldots, v_n\}$, the SCP is to determine whether there exists a k-set cluster from those m subsets such that the summation of the sets in such cluster is equal to the ground set. We build a bipartite graph in Fig. 3, which is constructed with the principles as follows. First, m and n nodes are generated based on cluster $\{S_j\}$ and ground set $\{v_i\}$ respectively. Second, a directed edge is built from node j to i iff. $v_i \in S_j$. The sample probability $p(e)$ for every directed edge e is set to 1. Further, we set $p(D(\boldsymbol{a}^1, \boldsymbol{a}^0) \leq T) = 1$ by configuring the parameters properly. We can find that if there is a k-element cluster which is a solution for SCP, there must be a k-node set of seeds selected from the top part in Fig. 3 to maximize the function $|\sigma|$ whose optimal value is $k + n$. Therefore, an SCP is transformed to an IMIV problem on the bipartite graph. Obviously, the IMIV problem is at least as hard as SCP, which implies IMIV problem is NP-hard. □

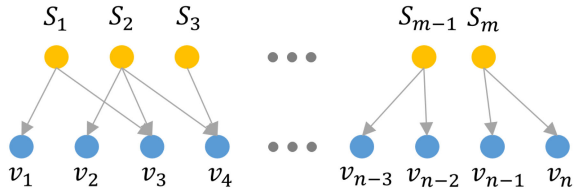


Fig. 3. Bipartite graph constructed according to the set cover problem.

The proved NP-hardness implies that we need to turn to an approximation algorithm that can efficiently select influential users to maximize the spread of effective information. We will introduce the detailed solution in Section 3.

3 Methodology

In this section, we present a framework — *Searching, Estimating and Selecting* (SES) for the efficient seeking of seed users. SES comprises of three important segments, i.e., similarity searching, propagation estimation and node selection. We first present how we algorithmically search the similarity between two information vectors. That supports the estimation for the propagation of effective information. On this basis, we design an estimator and derive an approach to select seed nodes in a greedy manner. Finally, we put all those segments together to complete the design of SES framework.

3.1 Similarity Searching Algorithm

Recall that in Section 2 we have quantified the difference between two information vectors at intermediate steps of information evolution in Definition 3. However, how to algorithmically calculate such distance between the information evolved at any intermediate step and the original version remains to be solved. To this end, we design a similarity searching algorithm and summarize the pseudo-code in Algorithm 1.

The workflow of Algorithm 1 is unfolded as follows. Recalling Definition 2, an information vector tends to evolve to various states with three cases. Especially, in *case 3*, an element, i.e., a dimension may be dropped. Thus, there might be a blank dimension ($a_i \in (l, u)$ is unsatisfied) in a vector after an evolution step. So we use l_0-norm to represent the effective dimension of a vector. It is necessary to screen out these dimensions and make sure that the vector can be calculated by the matching distance function (lines 2-5). However, recalling Definition 2, as the information

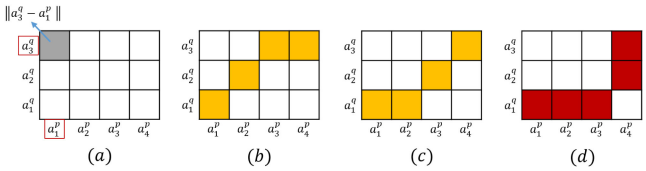


Fig. 4. Graphical representation of vector matching distance.

evolving in the network, some elements of information vector may be dropped and some new content may appear, i.e., description of the semantic evolution. Thus, we need to first make an alignment between the attributes of two vectors and further calculate their matching distance. After processing the information vectors, we calculate the distance between two vectors via all the possible candidate alignments and choose the minimum value as the final distance between two vectors. Meanwhile, the best alignment is determined with the similarity searching finished (lines 6-10). Here $\boldsymbol{e}^p$ is the vector without blank dimension and obtained from $\boldsymbol{a}^p$, d_{ij} is used to represent $\|\boldsymbol{a}_i^0 - \boldsymbol{a}_j^p\|$ and ω_q, Ω are the potential alignment results and corresponding cluster, respectively.

Algorithm 1. Similarity Searching Algorithm

Require: Information vector $\boldsymbol{a}^0$, $\boldsymbol{a}^p$; distance function D
Ensure: Equivalent vector $\boldsymbol{e}^p$ of $\boldsymbol{a}^p$; distance $D(\boldsymbol{a}^0, \boldsymbol{e}^p)$
1: **Initialize:** $\boldsymbol{e}^p = \boldsymbol{0}$, $m^0 = \|\boldsymbol{a}^0\|_0$, $m^p = \|\boldsymbol{e}^p\|_0$, $i = 0$, $min = \infty$
2: **while** $m^p < \|\boldsymbol{a}^p\|_0$ **do**
3: $i = i + 1$;
4: **while** $a_i^p \neq 0$ **do**
5: $m^p = m^p + 1$, $e_{m^p}^p = a_i^p$;
6: Compute $d_{ij}(\|\boldsymbol{a}_i^0 - \boldsymbol{e}_j^p\|)$ for $\forall i \in m^0, j \in m^p$
7: **for** every set of tuple ω_q in cluster Ω **do**
8: $D_q = \Sigma_{(i,j)\in\omega_q} d_{ij}$
9: **if** $D_q \leqq min$ **then**
10: $min = D_q$
11: **return** $D = min, \boldsymbol{e}^p$

To make the computation and alignment processes clearer, we use Fig. 4 as an example to explain. We suppose there are two vectors which have 3 and 4 elements respectively. In order to calculate the matching distance between them, we have to consider all the candidate alignments. In Fig. 4, the abscissa and ordinate represent the attributes of two vectors need to be aligned. Each cell represents the difference between the corresponding elements, i.e., d_{ij}. For example, the gray cell in Fig. 4a represents $\|a_3^q - a_1^p\|$ (d_{31}). In this example, there are 4 candidate alignments (ω_q in Algorithm 1), i.e., $\{(1,1),(1,2),(1,3),(2,4),(3,4)\}$, $\{(1,1),(1,2),(2,3),(3,4)\}$, $\{(1,1),(2,2),(3,3),(3,4)\}$ and $\{(1,1),(2,1),(3,2),(3,3),(3,4)\}$, which constitute the cluster Ω. Thus, what we next need to do is to calculate $d_{11}+d_{12}+d_{13}+d_{24}+d_{34}$, $d_{11}+d_{12}+d_{23}+d_{34}$, $d_{11}+d_{22}+d_{33}+d_{34}$ and $d_{11}+d_{21}+d_{32}+d_{33}+d_{34}$, and find the minimum value among them, from which the distance and alignment can be finally determined.

3.2 Estimation and Node Selection

Algorithm 1 specifies how to calculate the distance between an arbitrary vector with the original version. This can be further used in our design for estimation of the nodes influenced by effective information in propagation. Then, we can select the users greedily based on the estimation.

3.2.1 Estimation for Propagation

We leverage the ideas of the framework in IMM [15], OPIM-C [33] and SUBSIM [34] to design the estimation phase involved in SES. They focus on estimating the spread of the information over the network via the *Reverse Reachable* (RR) set, which provides a most efficient way to resolve the conventional IM problems. Except sharing a same RR set generation process, these works have a same class of contributions, i.e., designing the judgement and termination conditions of the algorithm to obtain different amount of RR sets (usually to reduce the totel number). Thus, to proceed, we first review and define the RR set, as stated in Definition 7.

Definition 7. (Reverse Reachable Set) *[16] Let v be a node in $G(V, E)$. An RR set for the node v is a union of nodes which can reach v via the active edges sampled by the IC model. An RR set is generated as follows. First, a node v is selected and added into the RR set R uniformly and randomly from V. Then, reversely traverse all the incoming neighbors of nodes in R by randomly activating (IC sampling) the corresponding edges and select the nodes connected with the nodes in R into R. Repeating the process until no more nodes are added, we can obtain a RR set.*

Due to the influence estimation phase is strongly dependent on the information propagation model. For example, the generation of RR set is closely dependent on IC model, which, as mentioned, is no longer available for the description of information propagation in IMIV. Recalling the improvement we made on IC model, the *Effective Reverse Reachable* (ERR) set is derived based on EIC model proposed in Section 2.2.

Definition 8. (Effective Reverse Reachable Set) *Let v be a node in $G(V, E)$. An ERR set for v is a union of nodes which can reach v via the active edges sampled by EIC model. An ERR set is generated as follows. First, node v is added into the ERR set R. Then, reversely traverse all the incoming neighbors of nodes in R by first randomly activating (IC sampling) the corresponding edges and select the nodes connected with the nodes in R into a transition set R'. On this basis, calculate the number of steps between the nodes in R' and v, so as to estimate whether the nodes are effectively activated or not based on probability $p(D(\boldsymbol{a}^j, \boldsymbol{a}^0) \leq T)$. Choose the effectively activated nodes from R' into R and empty R'. Repeating the process until no more nodes are added, we can obtain an ERR set.*

ERR set is generated with a given node, and serves as the estimation for the potential information source of this node. In order to estimate for the effective information propagation over the whole network, we endow the ERR set with randomness and derive the *Effective Random Reverse Reachable* (ERRR) set, as formally given in Definition 9.

Definition 9. (Effective Random Reverse Reachable Set) *An ERRR set is an ERR set for a random node of the network, which can be generated as follows. First, sample a node v*

uniformly and randomly from V. Then generate an ERR set R based on v.

Algorithm 2. ERRR Set Generation

Require: Directed Network $G = (V, E)$; threshold T; vector matching distance function D; original vector $\boldsymbol{a}^0$
Ensure: An ERRR set R
1: Sample a node v uniformly and randomly from V
2: Add v to a queue R' and ERR set R
3: **for** every node u in R' **do**
4: Take out u from R'
5: Reversely traverse all unactivated incoming neighbors with an activation probability from IC model.
6: For every node z activated in line 5, calculate its step t far away from v and effectively activate z with probability $p(D(\boldsymbol{a}^t, \boldsymbol{a}^0) \leq T)$.
7: For the nodes activated both in line 5 and line 6, put them into the ERR set R and queue R'.
8: **if** $R' = \emptyset$ **then**
9: Break and return R

An algorithm is designed to realize the two steps involved in the generation of an ERRR set. The pseudo-code of how to realize the two steps is presented in Algorithm 2. The algorithm first selects a random user as the reverse source of ERRR set and finds the potential seeds of the information received by the users from its in-neighbors (lines 1-5). Then, the in-neighbors who are reversely activated both in the first and second stage of EIC model are added into the ERRR set as the potential seeds and regarded as the new reverse sources to seek for more possible seed nodes (lines 5-7). Repeating this process, when there are no more nodes added into the set, we can obtain an ERRR set (lines 8-9). However, from the line 6 of Algorithm 2 and the above illustration, we can find that it is also important to calculate the probability $p(D(\boldsymbol{a}^t, \boldsymbol{a}^0) \leq T)$ so as to serve for the ERR generation and effective information propagation based on the similarity searching method. In order to realize that, we design the Algorithm 3 to specially calculate the activation probability of effective influence. The pseudo-code is shown as follows.The core idea of the algorithm is to simulate the processes of information variation to estimate the probability $p(D(\boldsymbol{a}^i, \boldsymbol{a}^0) \leq T)$ with $i \in [1, t]$ steps in K times (line 1-4). In every time of processing, a i-step variation happens to $\boldsymbol{a}^0$ to generate $\boldsymbol{a}^i$ while the distance between the two vectors is calculated at the same time (lines 5-7). Finally, the frequency of $D(\boldsymbol{a}^l, \boldsymbol{a}^0) \leq T$ appearing is used to approximate $p(D(\boldsymbol{a}^i, \boldsymbol{a}^0) \leq T)$ in every steps during the K-time simulations (lines 8-12).

With the algorithms of similarity searching and calculation of activation probability , an ERRR set can be generated. However, to enhance the persuasiveness of this method, we want to prove the ERRR set generation approach is consistent with the propagation process of effective information as formally given in Lemma 2.

Lemma 2. *The nodes in ERRR set can represent the potential seeds of effective influence. That is, the probability that a node u is in an ERR set R of the node v, is equal to the probability that u will influence v under the information vector evolution model from the path determined by the ERR set generation.*

Proof. First, for every node u in R, there is a geometric distance l between u and v on G. Therefore, the probability that node v can be activated by u from the path determined by the ERR set generation can be calculated as follows, i.e., every node in the path between u and v should be activated.

$$p_u v = \prod_{i=1}^{l} p_C p(D(\boldsymbol{a}^i, \boldsymbol{a}^0) \leq T). \tag{7}$$

Here, p_C and $p(D(\boldsymbol{a}^i, \boldsymbol{a}^0) \leq T)$ represent the probabilities of the two stage activation in EIC model. Due to there are no rings existing in an ERR set, in which each node can only be activated only once. Thus the geometric distance l is fixed on the path determined by the ERR generation whether for forward traversing or reversely traversing. Therefore, the probability that the node u is in an ERR set R of the node v can be also represented as Eqn. (7), where the only difference is the order of traversing. Hence, a node u is in an ERR set R of the node v, is equal to the probability that u will influence v under the information vector evolution model from the path determined by the ERR set generation. □

As mentioned above, Algorithms 1, 2 and 3 constitute the estimation phase of SES framework together. It is worth noting that, different from the previous works, our main idea for designing the ERR set is to use a more precise observation to improve the accuracy of the influence estimation for the information variation phenomenon with preserving the conditions in IMM [15]. That is to say, not only the design for the judgement and termination conditions can balance the accuracy of approximation algorithms and time complexity so as to reduce the number of generated RR set to speed up the algorithm, but the design for the RR set algorithm itself can also achieve this goal. In fact, this design implicitly reduces the number of ERR sets generated in the estimation algorithm while expanding the promulgation scope of effective influence as shown in the Section 4.

Algorithm 3. Calculation of Activation Probability

Require: Threshold T; distance function D; original vector $\boldsymbol{a}^0$; estimation time K; step $i \in [1, t]$]
Ensure: A probability list $[p_1, p_2, \ldots, p_t]$
1: **Initialize:** $i = 1$, $j = 0$, $P = \emptyset$
2: **while** $i \leq t$ **do**
3: **while** $j \leq K$ **do**
4: $l = 0, \boldsymbol{a} = \boldsymbol{a}^0, \eta = 0$
5: **while** $l \leq i$ **do**
6: Generate variation on $\boldsymbol{a}$, obtain $\boldsymbol{a}^l$ according to Definition 2 and update $\boldsymbol{a}, l$ with $\boldsymbol{a} = \boldsymbol{a}^l, l = l + 1$
7: Calculate $D(\boldsymbol{a}^l, \boldsymbol{a}^0)$ with Algorithm 1
8: **if** $D(\boldsymbol{a}^l, \boldsymbol{a}^0) \leq T$ **then**
9: $\eta = \eta + 1$
10: Update j with $j = j + 1$
11: Assign $p_i = \eta / k$, append p_i to P, update $i = i + 1$
12: **return** P

3.2.2 Node Selection

Both the definitions of ERR Set and ERRR set (Definitions 8 and 9) specify how to detect the possible seeds for an

arbitrary node. Hence, with the goal of achieving the information source estimation for all the users, a plenty of ERRR sets need to be generated (The quantitative description will be disclosed in Section 3.3.).

ERRR sets implicitly represent the common potential seeds of effective information from different users. Intuitively, the nodes overlapped by more ERRR sets are more likely to be the seed nodes. On this basis, we greedily select seed nodes which can maximize the number of overlapped ERRR sets to approximately maximize the propagation of the effective information. The pseudo-code is presented in Algorithm 4. Here, $\mathcal{R}$ is a collection of ERRR sets, w is the node selected in each round. $\mathcal{R}\emptyset v$ is defined as the elements of which are the sets including node v from $\mathcal{R}$.

Algorithm 4. Node Selection

Require: $G = (V, E)$; scale of seed set k; collection $\mathcal{R}$
Ensure: Seed set S
1: **Initialize:** $S = \emptyset$, $Y = V$, $j = 0$
2: **while** $j < k$ **do**
3: $w = argmax_{v \in Y} |\mathcal{R}\emptyset v|$;
4: Renew the sets $S = S \bigcup w$, $Y = Y \backslash \{w\}$, $\mathcal{R} \backslash \mathcal{R}\emptyset v$;
5: Renew the number of seed nodes $j = j + 1$;
6: **return** S

3.3 Algorithm and Approximation Guarantee

Now we are ready to combine everything together and finish our design of framework SES.

3.3.1 Greedy Algorithm for IMIV

The pseudo-code of SES is presented in Algorithm 5, which includes the following operations.

During the algorithm, ERRR sets are generated constantly to constitute the collection $\mathcal{R}$ which can be used to estimate the potential seeds of the network (lines 4-5). Then a node selection phase is executed to estimate the accuracy of this $\mathcal{R}$ (lines 6-9). If the accuracy is acceptable, the $\mathcal{R}$ will be updated one last time and the seed node set will be generated (lines 10-12). If not, the ERRR set continues to be generated and added into $\mathcal{R}$. The measurement for the accuracy of the estimator is determined by the parameters $LB, \theta, i, x, \lambda', \lambda^*$ and input parameter ϵ'.

$$\lambda' = \frac{\left(2 + \frac{2}{3}\epsilon'\right) \cdot \left(\log \binom{n}{k} + \ell \cdot \log n + \log \log_2 n\right) \cdot n}{\epsilon'^2}, \tag{8}$$

where $n = |V|$, $\epsilon' = \sqrt{2}\epsilon$. ϵ and ℓ can be configured.

$$\lambda^* = 2n \cdot ((1 - 1/e) \cdot \alpha + \beta)^2 \cdot \epsilon^{-2}, \tag{9}$$

where $\alpha = (\ell \log n + \log 2)^{\frac{1}{2}}$ and $\beta = ((1 - 1/e) \cdot (\log \binom{n}{k} + \ell \log n + \log 2))^{\frac{1}{2}}$. The efficiency of these parameters has been proved in [15]. Note that $\nu(S_i)$ is a function defined as $|\bigcup_{v \in S_i} \mathcal{R}\emptyset v| / |\mathcal{R}|$. It is a monotone submodular function for a collection $\mathcal{R}$.

3.3.2 Guarantee and Complexity Analysis

Next we will introduce the theoretical performance of SES from its approximation ratio and time complexity.

Approximation Guarantee. To analyze the approximation guarantee of SES, we extend the conclusions of the previous IM problems, in which the results of a greedy algorithm with monotone-submodular objective functions on the estimation satisfy the approximation radio of $1 - 1/e - \epsilon$ with at least $1 - 1/n^{\ell}$ probability [15]. Recalling Lemma 1 and Section 3.3, the objective function σ of IMIV and estimation function ν are still submodular and monotone. Thus, SES also satisfies this approximation guarantee.

Algorithm 5. SES

Require: Directed Network $G = (V, E)$
Ensure: Seed set S
1: **Initialize:** $\mathcal{R} = \emptyset$, $LB = \theta = i = x = 0$
2: **while** $i < log_2 n - 1$ **do**
3: Let $\theta_i = \lambda'/x$, $x = |V|/2^i$, λ' is defined in Eqn. (8);
4: **while** $|\mathcal{R}| < \theta_i$ **do**
5: *ERRR Set Generation*; $\mathcal{R} = \mathcal{R} \bigcup \{R\}$;
6: Continue line 12, obtain S_i and renew $j = 0$;
7: **if** $|V| \cdot \nu(S_i) \geq (1 + \epsilon') \cdot x$ **then**
8: $LB = |V| \cdot \nu(S_i)/(1 + \epsilon')$; Break;
9: Renew $i = i + 1$;
10: Let $\theta_i = \lambda^*/LB$, λ^* is defined in Eqn. (9);
11: Continue line 4-5 and obtain $\mathcal{R}$;
12: *Node Selection*;
13: **return** S

Time Complexity. From the algorithms shown above, it is obvious that the time complexity of SES is dominated by the estimation phase, i.e., Algorithms 1, 2 and 3. Thus, we focus on the analysis of the time complexity in these three algorithms. First, for Algorithm 1, the complexity is $O(3^{m^p})$ which can be explained from Fig. 4 and Eqn. (1), in which there are 3 branches for each round of iteration. As a recursive function, we can find that its maximum number of iterations is the length of the vector with more dimensions between the two whose distance should be calculated. Here, we suppose there are at most s dimensions in an information vector. Thus, the time complexity of Algorithm 1 is represented as $O(3^s)$. For Algorithm 2, the time consumption is mainly caused by traversing the nodes and edges in $G(V, E)$, whose time complexity is less than $(|V| + |E|)$ without considering the computation of probability $p(D(\boldsymbol{a}^t, \boldsymbol{a}^0) \leq T)$. Because the complexity of computing this probability in Algorithm 3 has linear relationship with Algorithms 1, 1 and 3 share the some time complexity, i.e., $O(3^s)$. Moreover, as illustrated in Section 2, we can endow a prior activation probability for the calculation of the probability on effective influence, which can be further regarded as a preprocessing. Besides, considering the procedure in Algorithms 4 and 5, we can obtain the result that the final time complexity of our framework is $O(3^s) + O((k + \ell)(|V| + |E|)\log |V|/\epsilon^2)$.

3.4 IMIV in Small-Scale Networks: Full Coverage Case

SES returns a seed set which can maximize the influence of effective information. However, in some situations we hope to spread the effective information to all the users in a network. For example, a district government wants to promote a campaign which need spread to all the people in this

district. This calls for the necessity of modifying our proposed algorithm to make it applicable for such small scale networks. For this reason, we consider a problem which is to effectively influence all the users in a small-scale network called *Full Coverage of Effective Information* (FCEI) problem. If all the network users need to be covered, an effective coverage of each seed user should be first determined.

Definition 10. *(*Effective Coverage*) The effective coverage of a piece of information is the number of nodes that can be effectively influenced through a directed path from a seed node.*

That is to say, if a node is in the effective coverage of a seed node, it will definitely be influenced by the effective information. To find an effective coverage for a seed node, we estimate its farthest neighbor (j_m steps away from the seed) which can receive the effective information from the seed and use j_m as the effective coverage. To this end, Lemma 3 calculates the expected j_m with the distance function, distance threshold and the parameters of the model in Definition 2. We defer the proof to Appendix A, available in the online supplemental material.

Lemma 3. *For networks of small scales, the expectation of the maximum coverage j_m can be represented as*

$$C = \mathbb{E}[\text{argmax}_{j\in\mathbb{N}} D(\boldsymbol{a}^j, \boldsymbol{a}^0)] = \left\lfloor \frac{\sqrt{b^2 + 2p_2p_1\Delta dT} - b}{p_2p_1\Delta d} \right\rfloor \leq \left\lfloor \sqrt{\frac{2T}{p_2p_1\Delta d}} \right\rfloor = O\left(\left(\frac{2T}{p_2p_1\Delta d}\right)^{\frac{1}{2}}\right). \tag{10}$$

Algorithm 6. Greedy Algorithm for FCEI Problem

Require: $G = (V, E)$, T, D, original vector $\boldsymbol{a}^0$, function σ_i^U
Ensure: Seed set S; maximum coverage C
1: **Initialize:** $S = \emptyset$, $U = V$, $i = 0$
2: **while** $D' < T$ **do**
3: Renew the distance $i = i + 1$;
4: Calculate the average of function $D(\boldsymbol{a}^0, \boldsymbol{a}^i)$ as D';
5: $C = i$;
6: **while** $U \neq \emptyset$ **do**
7: $w = argmax_{v\in U}|\sigma_C^U(v)|$;
8: Renew the seed set $S = S \bigcup w$;
9: Renew the uninfected set $U = U \backslash \sigma_C^U(v)$;
10: **return** S, C

Based on the effective coverage provided in Lemma 3, the seed nodes can be selected. Thus, we propose a new algorithm for FCEI, with the pseudo-code listed in Algorithm 6. The working mechanism is illustrated as follows. The expectation of coverage C is first calculated. Then we continuously pick the nodes who can effectively influence the most nodes which are not in the effective coverage of other seed nodes (lines 6-9). Function $\sigma_C(X)$ represents the nodes that can be effectively influenced by the seed set X, i.e., the union of set X and its neighbors within C-hop. The marginal gain function $\sigma_C^U(v)$ is an intersection between set U and $\sigma_C(v)$.

Based on Algorithm 6, Theorem 2 discloses the relationship between effective coverage and the number of seed users. The proof of Theorem 2 is available in Appendix B, available in the online supplemental material.

TABLE 1
Statistics of Feature Dataset

Data Type	Length	Probability of Change	Parameters
Feature0	411749	0.1306	(1.25,0.735)
Feature1	411749	0.1700	(2.12,0.10)
Feature2	411749	0.3546	(-1.00,0.175)
Feature3	411749	0.0648	(0.02,1.50)
Addition	411749	0.0101	(1.00,0.00)

Theorem 2. *An upper bound for the scale of theoretical optimal solution S^* about FCEI problem is $\frac{n}{1+2j}\ln(1+2j)e$ (when $\delta = 2$, δ is the minimum out-degree of a node in the graph) or $\frac{(2-\delta)n}{2-\delta(\delta-1)^j}\ln\frac{(2-\delta(\delta-1)^j)e}{(2-\delta)}$ (when $\delta \geq 3$) which is also a trade-off between C and $|S^*|$. Here $n = N$, $j = j_m = C$. U_x is a set of nodes with degree x.*

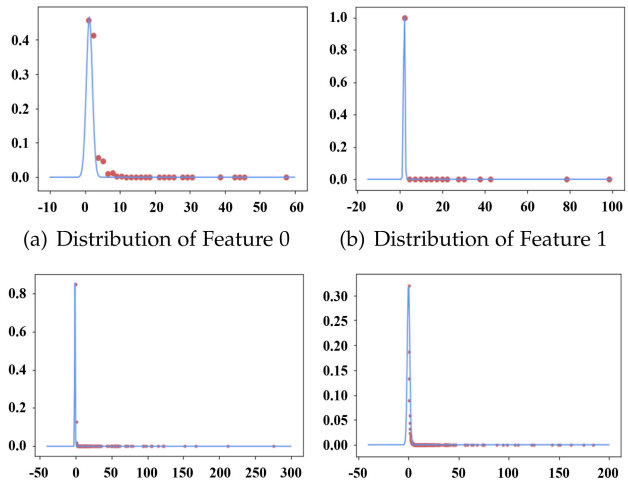
4 Performance Evaluation

In this section, we first verify our model with real-world dataset, and then evaluate our framework for IMIV by extensive experiments in real-world datasets.

4.1 Evaluation of the Models

Dataset. The real network dataset *act-mooc* [23] is applied to verify the effectiveness of *information vector evolution model*. The data in this dataset has multidimensional features (features 0-3) which represent the actions taken by the users, i.e., different attributes of a piece of information on Mooc platform. In detail, the features 0-3 are 4 kinds of high-order but low-dimensional data which is normalized and processed from a bulk of original data to depict the possible consequences (i.e., quit or not) of the user behaviours. Besides, there is an extra binary lable in the dataset, which can be regarded as an additional attribute of the information. Therefore, each data can be rewritten as an information vector. The changed data can be interpreted as a result of an one-step variation from the original. Naturally, the change of features can reflect the effect of the user behaviours on the information attributes, as described by *case 1* of Definition 2. The additional binary lable is a new dimension of the vector described in *case 2* of Definition 2. Our purpose is to evaulate whether our model matches the evolution patterns of the real dataset.

Results. Table 1 and Fig. 5 show the evaluation results. To verify our model whether matches the evolution patterns of the real dataset or not, our evaluation is unfolded with two steps, i.e., the simulation for the probability of occurrence and the verification for the distribution. For the first step, the column *Probability of Change* of Table 1 counts the frenqency (probability) of data variation, from which the p_1^i and p_2 of the information vector evolution model can be determined. On account of that there is no *case 3* happening in this dataset, the p_3 can be set to 0. In the second step, it can be found from the Fig. 5 that the changed data of the features mainly follows the half Gaussian distributions with different parameters, which validates the rationality of the *case 1*. Moreover, the value of the additional dimension follows a discrete uniform distribution, which is also



(a) Distribution of Feature 0 (b) Distribution of Feature 1

(c) Distribution of Feature 2 (d) Distribution of Feature 3

Fig. 5. Evaluation for the distribution in our model.

consistent with our model. The parameters which correspond to the expectation and variance (Δd_i and σ_i) of a complete Gaussian distribution are demonstrated in Table 1.

It can be seen that for every type of information or dataset, our *Information Vector Evolution Model* can provide a parameter combination to match or capture its evolution pattern. Thus, the different parameters can be used to depict the information variation when propagated in an arbitrary network. In Section 4.2, a variety of parameters will be applied to verify the effectiveness of our algorithm.

4.2 Evaluation for SES

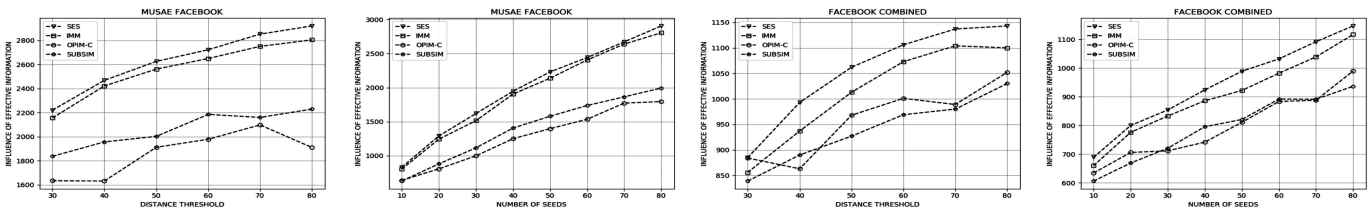
Datasets. Four real directed network datasets are applied in our experiments to verify the superiority of our algorithm.

- *Facebook* [9]: There are 4039 nodes (Facebook users) and 88234 edges (connections) in total.
- *Facebook Large Page-Page* [14]: There are 22,470 nodes (Facebook pages) and 171,002 edges (mutual likes).
- *Brightkite Social Network* [29]: There are 58,228 nodes (Brightkite users) and 214,078 edges (relationships).
- *EU Email Communication Network*[31]: There are 265,214 nodes (addresses) and 420,045 edges (emails).
- *Amazon Product Co-Purchasing Network*[39]: There are 334,863 nodes and 925,872 edges.

Compared Methods. We compare our framework with SOTA baselines IMM [15], OPIM-C [33], SUBSIM-HIST [34].

- *IMM*: IMM is a greedy-based algorithm which uses RR sets to approximate the propagation. However, whether the influence is effective or not is not considered by this approach.
- *OPIM-C*: OPIM-C uses two clusters of RR sets to estimate the influence of the selected seed set, which can control and restrict the generation of RR set.
- *SUBSIM-HIST*: SUBSIM-HIST improves the processes of RR set generation and node selection, and designs a sentinel set to reduce the generated RR sets so as to optimize the time complexity.

Experiment Setup. To evaluate the performance of our framework, we fist make a comparison with three baselines on two facebook networks when setting different seed number k and distance threshold T. Further, to show the scalability of our algorithm, two larger network datasets are added in our contrast experiments. The related experiment results are demonstrated in Fig. 6, Tables 2 and 3 respectively, in which parameters ϵ and ℓ are set to 0.5 and 1 separately. Each solution from the algorithms is evaluated by 500 Monte Carlo simulations with EIC model independently, which can be further used to calculate the average value as the criteria of evaluation. For each directed edge in the network, we set the connectivity probability as $\frac{1}{ind}$, where ind is the in-degree of the node pointed by the edge. These settings have been widely adopted in prior works [15], [16]. The information vector with $|\boldsymbol{a}^0| = 5$, $l = 10000$, $u = 10100$ and model parameters $\Delta d = 0.5$, $p_1 = 1$, $p_2 = p_3 = 0.2$ are adopted to describe the information evolution. Here, l and u are the lower bound and upper bound of the value for each dimension of the information vector in our simulations respectively. However, the value of l and u will not make an obvious effect on the final results because as l and u change, threshold T will follow accordingly. In Figs. 6b



(a) Musae Facebook - T (b) Musae Facebook - k (c) Facebook Combined - T (d) Facebook Combined - k

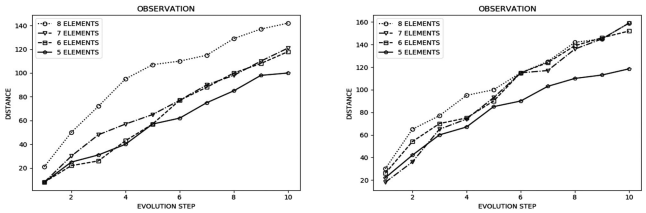
Fig. 6. Evaluation for SES on real datasets.

TABLE 2
EU Email Communication Network: 265214 Nodes and 420045 Edges

Scale/Method	200	400	600	800	1000	2000	3000	4000	5000
SES	**139108**	**172757**	**190140**	**200256**	**206149**	**214801**	**218036**	**220109**	**222484**
IMM	138640	172130	189892	199879	205636	214508	217670	219697	222236
OPIM-C	132978	165777	183279	192600	197829	207669	211545	214351	217157
SUBSIM-HIST	105591	133148	135159	138980	141734	134936	139108	139382	142460

TABLE 3
Amazon Co-Purchasing Network: 334863 Nodes and 925872 Edges

Scale/Method	200	400	600	800
SES	**10159**	**15760**	**20176**	**24113**
IMM	9976	15522	20058	24086
OPIM-C	5867	9950	11423	14359
SUBSIM-HIST	6953	10493	13442	15871



(a) $\Delta d = 0.5, p_2 = p_3 = 0.2, l = 10000, u = 10100$ (b) $\Delta d = 1, p_2 = p_3 = 0.4, l = 1000, u = 1100$

Fig. 7. Observation for distance with information evolution model.

and 6d, the scale of seed set k is set to 10-80 when distance threshold $T = 40$. In Figs. 6a and 6c, T is set to 30-80 when $k = 50$.

To clarify the variation of the distance under the information vector evolution model, we conduct the simulations to observe the expected distance of information propagation under the model in Fig. 7. We simulate the propagation of a piece of information according to Definition 2 from step 1 to 10 with 20 times Monte Carlo simulations and apply two configurations as shown in Figs. 7a and 7b.

In order to evaluate the time complexity of SES and further compare the performance of our framework and our baseline, IMM in Table 4, in which we follow the above mentioned configurations in Tables 2, 3 and Fig. 6 and change the seed scale k and threshold T from 1,000 to 5,000, from 10 to 50 respectively. Note that, the running time of Table 4 includes the processes of ERRR set generation, node selection and propagation with activation probabilities of EIC model calculated in advance.

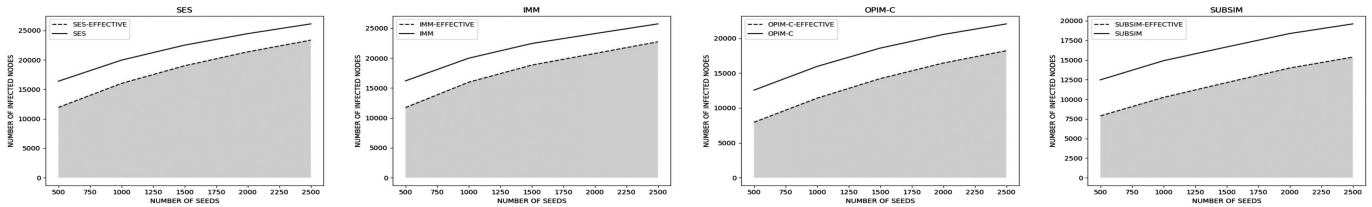
Moreover, we also construct the experiments to see the influence without considering the effectiveness of the information on the seeds selected by SES and draw the figures to observe the fractions of the effectively influenced nodes to the nodes affected without considering effectiveness. As a comparison, we draw the figures of other three baselines as well with k changeing from 500 to 2,500 under the threshold T set as 20, 40 respectively on the Brightkite dataset, and also retain the configurations for the algorithms above.

Comparison of Effective Influence Spread. The basic evaluation results presented in Fig. 6, Tables 2 and 3 prove that our framework can always perform better than baselines. However, when we compare the performance and running between SES and IMM in Table 4, there are some abnormal and unconspicuous evaluation results, which may be caused by the randomness of propagation, ERR/RR set generation that are designed to speed up the algorithm execution process. Even so, our SES still always performs better than IMM in Table 4 with different settings. It worth to mention that, due to the running time shown in our experiment results only including the ERR set generation, node selection and the process of 500 times Monte Carlo simulations for propagation, the time consumption of SES is slightly less than IMM which is lead by that the number of ERR sets generated during the whole SES process is less then RR sets generated in the IMM framework, which can further proves that our design of ERR set generation for IMIV problem is efficient from the perspective of reducing the number of ERR sets by improving estimation accuracy for the propagation of effective information. Besides, the execution time of SES and IMM is dominated by the methods of OPIM-C and SUBSIM-HIST, which accords with their designs, i.e., balance the time complexity and algorithm accuracy. Moreover, except the running time shown above, there is an extra time consumption for the preprocessing of SES, i.e., the calculation for the activation probability of EIC model, which is a little time consuming. In our experiments, the largest data set of Amazon product co-purchasing network is still processible.

The distance variation shown in Fig. 7 demonstrates that with the information propagated step by step, the distance between the pieces of propagated information

TABLE 4
Brightkite Dataset: 58228 Nodes and 214078 Edges

Scale		1000		2000		3000		4000		5000	
Method		SES	IMM	SES	IMM	SES	IMM	SES	IMM	SES	IMM
Threshold=10	Influence	**11490**	11170	**16152**	15527	**19843**	19123	**22902**	22097	**25085**	24638
	Time(s)	**2411**	3828	**2152**	3211	**3069**	4386	**3950**	5511	**2239**	6571
Threshold=20	Influence	**16032**	15926	**21355**	20783	**24681**	24450	**27568**	27299	**29991**	29720
	Time(s)	**1752**	3974	**3087**	3257	**2065**	4414	**2609**	5529	**3135**	6549
Threshold=30	Influence	**18230**	18173	**23341**	22868	**26450**	26221	**29065**	28879	**31189**	31079
	Time(s)	**2094**	3955	3905	**3298**	**2521**	4475	**3206**	5597	**3835**	6588
Threshold=40	Influence	**19238**	19197	**23726**	23694	**26874**	26861	**29415**	29391	**31543**	31502
	Time(s)	**2437**	3946	**2118**	3295	**2903**	4532	**3687**	5603	**4377**	6612
Threshold=50	Influence	**19689**	19636	**24000**	23958	**27059**	27008	**29532**	29499	**31609**	31586
	Time(s)	**2789**	3981	**2341**	3339	**3201**	4516	**4055**	5707	**4826**	6799



(a) SES (b) IMM (c) OPIM-C (d) SUBSIM

Fig. 8. Fractions of effective influence spread with $T = 20$.

and the origianl information always tends to increase. Generally, the distance of the information vector with more elements tends to increase faster which is caused by case 1 of Definition 2. However, because of the element dropping and generation (case 2,3), the variation direction is always complicated, which conforms to the real world settings.

By comparison between the influenced nodes of effective information and the information without considering the effectiveness in Figs. 8 and 9, we can find that our framework not only performs better on the propagation of effective information, but obtain comparable evaluation results on the propagation of information without considering the effectiveness when compared with the three baseline algorithms. Generally, the performance of SES is also better than IMM, OPIM-C and SUBSIM-HIST on the IMIV problem, while SES can keep the performance of IMM on the conventional IM problem from another perspective. In fact, the nodes selected by IMM and SES are not exactly the same, which proves that the seeds of effective information also have the characteristic of influence maximization. Moreover, we can find from the two figures that the ratio between the effective influence and the influence without considering the effectiveness on SES is maximum compared with three baselines, which demonstrates that the seed nodes found by SES tend to have more effective influence. Besides, with the distance threshold T and scale of seed set k increasing, the difference between the two kind of influence is decreasing. The explanation is unfolded as follows. First, distance threshold T increasing means the effective information can be propagated farther, which is closer to the conventional IM (i.e., suppose T is infinite in IMIV) and cause this phenomenon. Then, if more seeds can be used in IMIV problem, the effective information can be propagated more widely. That is to say, some nodes are infected by the information but can not receive the effective information will be effectively influenced when the selected seeds increase, which further demonstrates SES from the baselines. Hence, both the efficiency and effectiveness of our framework have been verified and evaluated with plenty of control experiments on real-world datasets from various perspectives hereto.

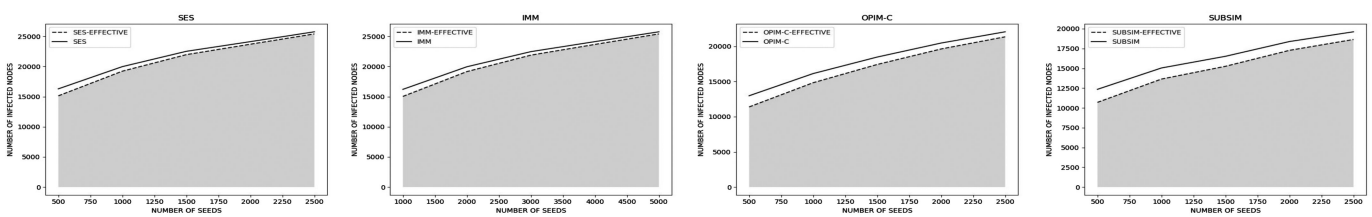
5 Related Work

In this section, some related works about IM and subsequence matching which is the basic idea utilized to calculate the matching distance are introduced.

5.1 Influence Maximization

IM problem is first studied by Domingos and Richardson in [3] and [13]. In [6], IM problem is first formulated as a combinatorial optimization problem under the IC and linear threshold (LT) models, in which a greedy algorithm with $1 - 1/e$ approximation radio is proposed. On this basis, two main types of works about IM problem are derived gradually.

One of the branches is to find the tradeoff between approximation guarantee and practical efficiency and design the algorithm which can work well in large-scale networks. Borgs *et al.* [1] first use the idea of reverse influence sampling (RIS) and derive the greedy algorithm on the estimation of the propagation. Y. Tang *et al.* [16] borrow the ideas from RIS and derive TIM with approximation guarantee. Further, they improve TIM and propose IMM [15] which is the most efficient IM algorithm currently. C. Qian *et al.* propose an iterative and random approach which is a kind of evolutionary algorithm to spend more time finding a better solution [12]. Recently, OPIM-C [33] is proposed by J. Tang *et al.*, whose idea is using two clusters of RR set to estimate the performance of seed set. They prove this framework can further reduce the time complexity from IMM. Later, Q. Guo *et al.* derive a framework, i.e., SUBSIM-HIST [34], which is based on IMM and OPIM-C to uteriorly reduce the complexity of algorithm by first applying a sentinel set selection phase. Both the OPIM-C and SUBSIM-HIST framework show a better experiment result compared with IMM on running time.



(a) SES (b) IMM (c) OPIM-C (d) SUBSIM

Fig. 9. Fractions of effective influence spread with $T = 40$.

Another branch is to study the new application scenarios of IM problem. Considering the negative opinions and humor produced with propagation, [20] and [2] are derived by B. Wang *et al.* and W. Chen *et al.* respectively. Especially, following with [2], there are amount of literatures focusing on the IM problem with propagation of negative opinions. For example, W. Ju *et al.* [35] study the IM problem with the competition of positive and negative influence in signed networks. Y. Chen *et al.* [36] derive an algorithm to study how to prevent the propagation of negative opinions in social networks. In addtion, Y. Li *et al.* [37] study the IM problem with foe users and enemies which can spread negative information exist. However, different from the literatures above, our work focuses on maximizing the original influence with the dynamic evolution process of the information propagation, rather than assumes the negative influence is initially existing. Moreover, the IM problem in the network with communities is considered and solved by L. He *et al.* [4]. J. Li *et al.* [27] use the crowdsourced data to solve the IM problem in the location-based social networks of online and physical world. Further, J. Li *et al.* [26] and X. Wu *et al.* [21] consider the geometric social influence spanning maximization problem and IM problem with location attributes respectively. X. He *et al.* [5] study the influence blocking maximization problem in social networks which takes competitive spread of two opinions into consideration with a competitive LT model. Moreover, recently, G. A. Tong *et al.* [24] find out a hybrid sampling method which improves the principles of RR set to prevent the spread of the misinformation.

As recently updated IM issues, J. Li *et al.* [25] take the diversity of the communities into consideration and derive the CDIM problem, which aims to solve the IM problem from the community perspective. To deal with the condition of access limit, i.e., only a small subset of users can be initially accessible, C. Feng *et al.* [28] use a two-stage model to describe this phenomenon and an adaptive approach to solve this problem. Taking both the cyber and physical user interaction into account, T. Cai *et al.* [32] propose the holistic influence maximization problem in spatial social networks.

5.2 Subsequence Matching

Our work also has tie with subsequence matching, one of the important topics in the field of data stream mining. It has wide applications in many fields. For example, Y. Zhu *et al.* [22] find its application on burst detection which is to find abnormal aggregates in data streams for the sensor network monitoring. Yasuko *et al.* [11] use similar technologies to detect the patterns and trends of the change in web click-steam analysis. Moreover, analysis of network traffic based on the models of data stream is derived in [7].

Meanwhile, the technologies of subsequence matching are also updated and improved continuously. First, Toyoda *et al.* in [18] solve the problem of finding the common local patterns in data streams. Further, a more complete approach — DTW is produced in [17]. Recently, *Cross Match* is proposed to give an algorithm with theoretical guarantee based on DTW in [19]. We use DTW in our work and improve it to be a matching distance function between two vectors (Eqn. (1)).

6 Conclusion

In this paper, the IM problem for the effective information is studied. A matching distance function and an EIC model are first designed and applied to quantify the deviation between the attributes of the information carried by different individuals when the information is being passed. We further prove the problem is NP-hard and design a greedy framework to solve the IMIV problem. For networks of small scales, the FCEI problem is solved and a necessary condition for a full-cover case is derived. Finally, the simulation results and experiments on real-world datasets support our models and algorithms.

Our future work will focus on the problem of how to extract a vector from an any piece of information based on its attributes. Besides, we will improve our information evolution model to cover more possible situations in the real networks.

References

[1] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in *Proc. 25th Annu. ACM-SIAM Symp. Discr. Algorithms*, 2014, pp. 946–957.
[2] W. Chen *et al.*, "Influence maximization in social networks when negative opinions may emerge and propagate," in *Proc. SIAM Int. Conf. Data Mining*, 2011, pp. 379–390.
[3] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2001, pp. 57–66.
[4] L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu, "Joint community and structural hole spanner detection via harmonic modularity," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 875–884.
[5] X. He, G. Song, W. Chen, and Q. Jiang, "Influence blocking maximization in social networks under the competitive linear threshold model," in *Proc. SIAM Int. Conf. Data Mining*, 2012, pp. 463–474.
[6] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 137–146.
[7] F. Korn, S. Muthukrishnan, and Y. Wu, "Modeling skew in data streams," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2006, pp. 181–192.
[8] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 641–650.
[9] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 539–547.
[10] C. Long and R. Chi-Wing Wong, "Minimizing seed set for viral marketing," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 427–436.
[11] Y. Matsubara, Y. Sakurai, C. Faloutsos, T. Iwata, and M. Yoshikawa, "Fast mining and forecasting of complex time-stamped events," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 271–279.
[12] C. Qian, J.-C. Shi, Y. Yu, and K. Tang, "On subset selection with general cost constraints," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, vol. 17, pp. 2613–2619.
[13] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2002, pp. 61–70.
[14] B. Rozemberczki, C. Allen, and R. Sarkar, "Multi-scale attributed node embedding," *J. Complex Netw.*, vol. 9, no. 2, pp. cnab014, 2021.
[15] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1539–1554.

[16] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 75–86.
[17] M. Toyoda and Y. Sakurai, "Discovery of cross-similarity in data streams," in *Proc. IEEE 26th Int. Conf. Data Eng.*, 2010, pp. 101–104.
[18] M. Toyoda, Y. Sakurai, and T. Ichikawa, "Identifying similar subsequences in data streams," in *Proc. Int. Conf. Database Expert Syst. Appl.*, 2008, pp. 210–224.
[19] M. Toyoda, Y. Sakurai, and Y. Ishikawa, "Pattern discovery in data streams under the time warping distance," *Int. J. Very Large Data Bases*, vol. 22, no. 3, pp. 295–318, 2013.
[20] B. Wang, G. Chen, L. Fu, L. Song, and X. Wang, "DRIMUX: Dynamic rumor influence minimization with user experience in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2168–2181, Oct. 2017.
[21] X. Wu, L. Fu, Y. Yao, X. Fu, X. Wang, and G. Chen, "GLP: A novel framework for group-level location promotion in geo-social networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2870–2883, Dec. 2018.
[22] Y. Zhu and D. Shasha, "Efficient elastic burst detection in data streams," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 336–345.
[23] S. Kumar, X. Zhang, and J. Leskovec, "Predicting dynamic embedding trajectory in temporal interaction networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 1269–1278.
[24] G. A. Tong and D. Z. Du, "Beyond uniform reverse sampling: A hybrid sampling technique for misinformation prevention," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 1711–1719.
[25] J. Li, T. Cai, K. Deng, X. Wang, T. Sellis, and F. Xia, "Communitydiversified influence maximization in social networks," *Inf. Syst.*, vol. 92, 2020, Art. no. 101522.
[26] J. Li, T. Sellis, J. S. Culpepper, Z. He, C. Liu, and J. Wang, "Geo-social influence spanning maximization," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1653–1666, Aug. 2017.
[27] J. Li, Z. Cai, M. Yan, and Y. Li, "Using crowdsourced data in locationbased social networks to explore influence maximization," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.
[28] C. Feng *et al.*, "Neighborhood matters: Influence maximization in social networks with limited access," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 10, 2020, doi: 10.1109/TKDE.2020.3015387.
[29] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2011, pp. 1082–1090.
[30] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the Semantic Web," in *Proc. Int. Semantic Web Conf.*, 2003, pp. 351–368.
[31] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution. Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 2–es, 2007.
[32] T. Cai, J. Li, A. Mian, R. Li, T. Sellis, and J. X. Yu, "Target-aware holistic influence maximization in spatial social networks," *IEEE Trans. Knowl. Data Eng.*, early access, Jun. 17, 2020, doi: 10.1109/TKDE.2020.3003047.
[33] J. Tang, X. Tang, X. Xiao, J. Yuan, "Online processing algorithms for influence maximization," in *Proc. Int. Conf. Manage. Data*, 2018, pp. 991–1005.
[34] Q. Guo, S. Wang, Z. Wei, and M. Chen, "Influence maximization revisited: Efficient reverse reachable set generation with bound tightened," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2020, pp. 2167–2181.
[35] W. Ju, L. Chen, B. Li, W. Liu, J. Sheng, and Y. Wang, "A new algorithm for positive influence maximization in signed networks," *Inf. Sci.*, vol. 512, pp. 1571–1591, 2020.
[36] Y. Chen, H. Li, and Q. Qu, "Negative-aware influence maximization on social networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, vol. 32, Art. no. 1023.
[37] Y. Li, W. Chen, Y. Wang, and Z. Zhang, "Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 657–666.
[38] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016, *arXiv:1611.07308*.
[39] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 745–754.

Haonan Zhang received the BE degree from the Department of Electronic Engineering, Xidian University, Xian, China, in 2018. He is currently working toward the PhD degree in the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include property analysis of GNNs and influence propagation of social networks.

Luoyi Fu received the BE degree in electronic engineering and the PhD degree in computer science and engineering from Shanghai Jiao Tong University, China, in 2009 and 2015, respectively. She is currently an assistant professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. Her research of interests are in the area of social networking and big data, scaling laws analysis in wireless networks, connectivity analysis and random graphs. She has been a member of the Technical Program Committees of several conferences including ACM MobiHoc 2018−2020, IEEE INFOCOM 2018−2020.

Jiaxin Ding received the BS degree from Electronics Engineering and Computer Science, Peking University, China, in 2012, and the PhD degree from Computer Science, Stony Brook University, Stony Brook, New York, in 2018. He is currently an assistant professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research of interests are in the area of spatio-temporal data mining, data privacy protection, reinforcement learning, and Internet of Things.

Feilong Tang received the PhD degree in computer science and technology from Shanghai Jiao Tong University, China, in 2005. He was a Japan Society for the Promotion of Science (JSPS) postdoctoral research fellow. Currently, he works with the School of Software, Shanghai Jiao Tong University, China. His research interests include cognitive and sensor networks, protocol design for communication networks, and pervasive and cloud computing. He has published more than 100 papers in journals and international conferences and works as a PI of many projects such as National Natural Science Foundation of China (NSFC) and National High-Tech R&D Program (863 Program) of China.

Yao Xiao received the BE degree from the College of Management and Economics, Tianjin University, Tianjin, China, in 2017. He is currently working toward the master's degree in the School of Computer Science, Shanghai Jiao Tong University, Shanghai, China.

Xinbing Wang received the BS degree (Hons.) from the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 1998, the MS degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001, and the PhD degree, majoring in the electrical and computer engineering and minoring in mathematics, from North Carolina State University, Raleigh, North Carolina, in 2006. He is currently a professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. He has been an associate editor for the *IEEE/ACM Transactions on Networking* and the *IEEE Transactions on Mobile Computing*, and a member of the Technical Program Committees of several conferences including the ACM MobiCom 2012, the ACM MobiHoc 2012−2014, and the IEEE INFOCOM 2009−2017.

Guihai Chen received the BS degree from Nanjing University, China, the ME degree from Southeast University, China, and the PhD degree from The University of Hong Kong, Hong Kong. He visited the Kyushu Institute of Technology, Japan, in 1998, as a research fellow, and the University of Queensland, Australia, in 2000, as a visiting professor. From 2001 to 2003, he was a visiting professor with Wayne State University, Detroit, Michigan. He is currently a distinguished professor and a deputy chair with the Department of Computer Science, Shanghai Jiao Tong University, China. He has published more than 200 papers in peer-reviewed journals and refereed conference proceedings in the areas of wireless sensor networks, high-performance computer architecture, peer-to-peer computing, and performance evaluation. He is a member of the IEEE Computer Society. He has served on technical program committees of numerous international conferences.

Chenghu Zhou received the BS degree in geography from Nanjing University, Nanjing, China, in 1984, and the MS and PhD degrees in geographic information system from the Chinese Academy of Sciences (CAS), Beijing, China, in 1987 and 1992, respectively. He is currently an academician with the Chinese Academy of Sciences, China, where he is also a research professor with the Institute of Geographical Sciences and Natural Resources Research, and a professor with the School of Geography and Ocean Science, Nanjing University, China. His research interests include spatial and temporal data mining, geographic modeling, hydrology and water resources, and geographic information systems and remote sensing applications.