Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Semantics-aware influence maximization in social networks

Yipeng Chen[a,b], Qiang Qu[c,*], Yuanxiang Ying[a,b], Hongyan Li[a,b], Jialie Shen[d]

[a] Key Laboratory of Machine Perception (Peking University), Ministry of Education, China
[b] School of Electronics Engineering and Computer Science, Peking University, China
[c] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[d] School of EEECS, Queen's University Belfast, United Kingdom

ARTICLE INFO

Article history:
Received 17 June 2018
Revised 24 October 2019
Accepted 30 October 2019
Available online xxx

Keywords:
Influence maximization
User semantics
Influence measurement
Social network analysis

ABSTRACT

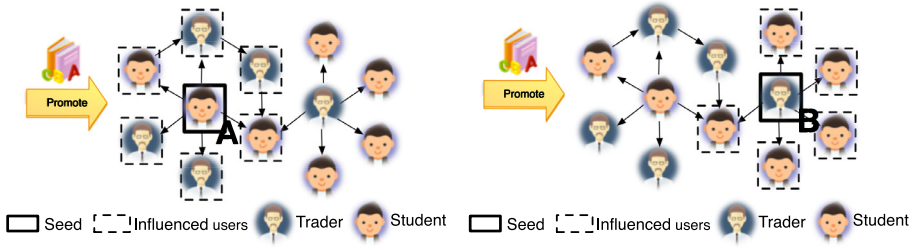
Influence Maximization (*IM*) plays an essential role in various social network applications. One such application is viral marketing to trigger a large cascade of product adoption from a small number of users by utilizing "Word-of-Mouth" effect in social networks. *IM* aims to return a set of users that can influence the largest fraction of a network, such as the early user who demonstrates the good features of a product in marketing. The traditional *IM* algorithms treat all users equally and ignore semantic context associated with the users, though it has been studied previously. To consider the semantics, we introduce a semantics-aware influence maximization (*SIM*) problem. The *SIM* problem integrates semantic information of users with influence maximization by measuring *influence spread* based on semantic values under a given model, and it aims to find a set of users that maximizes the influence spread, shown to be NP-hard. Generalized Reverse Influence Set based framework for SIM problems (GRIS-SIM) is used to solve *SIM* with different semantics, which provides a $(1-1/e-\varepsilon)$-approximation solution for each SIM instance. To our knowledge, the guarantee is state-of-the-art in the *IM* studies. GRIS-SIM enables auto-generation of sampling strategies for various social networks. In this study, we also present three sampling strategies that can be generated to achieve the best approximation guarantee, and one of the three is proved to be the optimal strategy by having the same performance guarantee within the optimal time. Furthermore, in order to show the generality and effectiveness of the proposed *GRIS* technique, we apply it into solving other *IM* problems (e.g., the distance-aware influence maximization, *DAIM*). Extensive experiments on both real-life and synthetic datasets demonstrate the effectiveness, efficiency, and scalability of our methods. The results on large real data show that *GRIS-SIM* is able to achieve 58% improvement on average in expected influence compared with rivals, and the method adopting *GRIS* can achieve 65% improvement on average.

1. Introduction

The proliferation of online users has attracted numerous efforts in studying social networks [1,22,23,25]. Most of the existing studies model the structure of a social network as a graph, where nodes and edges represent users and their interactions [30,31]. Real-world applications often identify and model the importance of the associated user semantics, considered

* Corresponding author.
E-mail addresses: chenyipeng@cis.pku.edu.cn (Y. Chen), qiang@siat.ac.cn (Q. Qu), yingyx@cis.pku.edu.cn (Y. Ying), lihy@cis.pku.edu.cn (H. Li).

https://doi.org/10.1016/j.ins.2019.10.075
0020-0255/



(a) Influenced users cascade from A (b) Influenced users cascade from B

Fig. 1. A social network with social tags as semantics. Children are with student tags, adults are with trader tags.

as an abstract generalization of context information in addition to the structures. For instance, nodes representing users on Facebook with their associated user profiles and social tags can be leveraged to improve performance for recommender systems [28]. In this paper, we introduce semantics into the influence maximization problem [16,21].

As one of the most fundamental problems in social network-related research, influence maximization (*IM*) possesses a wide range of applications, such as viral marketing, rumor control, and information monitoring [21,27,40]. *IM* aims to find a small set of users as seeds, which can eventually influence the maximal number of users in terms of a diffusion model [8,29].

Most of the existing *IM* studies mainly take the structures of given networks into account and largely ignore associated semantics. These methods model the users as equivalent nodes in a graph and then maximize the number of influenced users. However, the semantics may have a significant impact on influence maximization. Specifically, based on the associated semantics, some users may have higher priority to be influenced than others. Considering a real-world example, booksellers intend to perform viral marketing on a social network having user semantics as tags (e.g., student and trader). Students have higher probability of buying reference books than traders. Thus, vendors would prefer to choose users who can influence most valuable (i.e., profitable) users (i.e., users who can reach more students) rather than users who can influence the most number of users in general as in *IM*. Hence, semantic characteristics associated with networks are essential to maximize influence, which cannot be straightforwardly solved by the existing solutions.

Motivated by the above, we propose a novel approach to support effective semantics-aware influence maximization (*SIM*). *SIM* measures influence considering user semantics by a generalized semantic model, and it aims to identify seed users whose activated user set (i.e., influenced users) has the greatest influence. However, the problem poses a challenge to the existing *IM* methods due to the following three reasons:

- **Ignoring semantics**: Traditional *IM* solutions mostly discount semantics in maximizing influence. The influenced users thus may not form the most valuable set. For instance, Fig. 1 illustrates the aforementioned scenario, where booksellers are to promote reference books in a social network. Considering the social tags of users as semantics, users as students are more profitable for booksellers. To get one seed user, the existing *IM* solutions return user *A* because user *A* can eventually influence the most number of users, as in Fig. 1(a). However, user *B* in Fig. 1(b) is a better choice in reality because user *B* can influence the most valuable set of users with semantics indicating students.
- **Breaking structured constraints**: One may propose to divide a social network into sub-networks by discrete semantics (e.g., tags), and then solve sub-problems on the sub-networks by an existing *IM* algorithm in order to obtain the final solution. However, in this case, structured constraints and diffusion properties of the network cannot be sustained, which results in non-optimal solutions. For instance, one can divide the network in Fig. 1 into student and trader sub-networks. The existing *IM* algorithms would return user *A* since user *A* can influence most students in the student sub-network. This result does not provide the same optimalities as picking user *B*, which may be a trader who is not in the student sub-network.
- **Lacking generalization**: Very few recent studies [25,38] have realized the importance of various network semantics in real-world scenarios and consider them in influence maximization problems. Unfortunately, they have failed to generalize the commonality and to provide one model integrating these semantics.

The study formalizes the *SIM* problem by taking semantics into account and it is proven NP-hard. Further, we propose an advanced framework to solve the *SIM* problem with an approximation guarantee. Various theoretical bounds are further provided to show the well-designed properties of the proposed solutions over existing approaches. Both of the theoretic analysis and experimental findings show the superiority of our methods. In summary, our contributions:

- To naturally integrate semantics from broad applications, this paper formalizes a new problem, semantics-aware influence maximization (*SIM*), considering user values under a given semantic model as measurement of influence. The *SIM* problem is proved to be NP-hard.
- To effectively solve *SIM* with the best-known performance guarantee, this paper proposes a generalized *RIS* (*GRIS*) technique that supports different sampling strategies. Based on the *GRIS* technique, a Generalized Reverse Influence Set based framework for Semantics-aware Influence Maximization (*GRIS-SIM*) is proposed with $(1 - 1/e - \varepsilon)$-approximation guar-

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

antee. This paper finds three valid sampling strategies for *GRIS-SIM*, which can be adapted to different semantics settings and achieve the performance guarantee. Besides, the optimal sampling strategy with which *GRIS-SIM* can achieve the same approximation ratio within the optimal time is proposed.
- To evaluate and compare the solutions, the extensive experiments are conducted on six real-world datasets and a series of large synthetic networks against the rivals studying the effectiveness, efficiency, and scalability. Furthermore, we study the proposed techniques in solving relevant problems, e.g., distance aware influence maximization (*DAIM*), to evaluate the performance in different settings.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the problem with preliminaries. The solutions are presented in Section 4. Section 5 shows the experimental evaluation, followed by the conclusion in Section 6.

2. Related work

Discovering key users that have a large impact is an important problem in social network analysis. It has wide applications such as marketing, rumor and disease control, organizational management, operations research [15,21,29,33]. Borgatti [2] proposed the key player problem (*KPP*) that is to find the users maximally disrupting or connecting the network. *KPP* mainly focuses on users' network positions. Unlike *KPP*, the influence maximization (*IM*) problem considers the propagation of information, aims to discover key users in term of users influence spread on a network. *IM* problem is proposed by Domingos and Richardson [6]. The problem is formulated as a discrete optimization problem to prove its NP-hardness [9]. The authors proposed an approximate greedy algorithm based on the maximization of a submodular function. Since then, a series of studies have been proposed to improve the efficiency or accuracy of the greedy algorithm. Generally, these studies can be categorised into four types: (1) submodularity-based approaches; (2) community-based approaches; (3) integer programming-based approaches; and (4) reverse influence sampling-based approaches;

Submodularity-based approaches make use of the submodularity property of the IM problem to improve the time efficiency and use branch-and-bound techniques to reduce the cost of influence calculation [4,15,38]. Some studies [12] prune nodes that are expected to have smaller influence based on predefined metrics, some [4] reduce cost by focusing on partial influence paths, and others [15] reduce the number of spread estimation calls. Khomami [10] solves the minimum positive influence set (MPIS) problem, and uses MPIS to construct seeds for IM. These heuristic methods are often hard to obtain performance guarantee, and may have inferior performance in reality. Community-based approaches often take advantage of community information to reduce calculation [18]. These studies divide a sizeable social networks into different communities, and use the influence of a node in its community to approximate its influence on the overall network, or limit the influence propagation in one community. Similarly, community-based algorithms often fail to give a precise approximation as the network structure and the community detection algorithms usually affect the accuracy of these approaches. Integer programming-based approaches formulate influence maximization problem as the integer programming problem, and these methods aim to obtain exact solutions by optimization [32]. However, these approaches are inefficient for networks with more than a million nodes. Reverse influence sampling is applied in several methods to avoid estimating the expected influence of a large number of node sets in order to speed up the performance [3,36]. To the best of our knowledge, these methods have achieved the $(1 - 1/e - \varepsilon)$-approximation guarantee as the best performance bound [36]. Due to the effectiveness and efficiency of RIS-based algorithm, our solution also applies the reverse influence sampling technique and further generalizes it to support the estimation of influence spread under the SIM problem. Thanks to the similar process of node selection, our proof for the solution approximation is similar to the study [36]. Both proofs take the idea of applying the central limit on martingales. However, the construction of martingales and influence estimate functions are different. Furthermore, most of the RIS-based approaches including IMM [36] only consider structures without semantics in maximizing network influence.

Recently, few studies have been investigating semantics in solving *IM*. For instance, the studies [25] consider community and location in maximizing influence, the authors [19] propose to consider competitors' information, and Li [24] modifies IM propagation model to make use of trajectory data. Although these studies focuses on specific applications, they are not applicable to scenarios with different semantics.

Next, Wang et al. [38] studied the distance-aware influence maximization (*DAIM*) problem, which combined influence spread with the distance between a user and a location. *DAIM* assigns weights to users based on their distances from the location, which yields users with various importance to be influenced. The authors modified the maximum influence arborescence (*MIA*) model and the reverse influence sampling (*RIS*) technique to solve *DAIM*. *DAIM* can be considered as a specific application of *SIM* that generalizes different semantics. Furthermore, the findings in Section 5 show that applying the *GRIS* techniques to the *DAIM* solution, we can dramatically reduce the sample number and significantly improve the performance. Li et al. [20] studied the Keyword-Based Targeted Influence Maximization (*KB-TIM*) problem. In KB-TIM problem, users are associated with weights based on their preference in different topics. They proposed a sampling technique based on weighted reverse influence set to solve this problem. Noticed that their sampling technique is one of the valid sampling strategies for the *GRIS-SIM* framework which is presented in Section 4. *GRIS-SIM* can support more sampling strategies including the strategies defined by users, and thus it applies to more scenarios.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

Table 1
Key notation summary.

Notation	Description
$G = (V, E)$	A social network as a directed graph
$\mathcal{M}$	A semantics translation function
$S \rightsquigarrow v$	Node v can be activated by a node set S in a diffusion process
$H_\theta(S)$	Coverage function, the number of RR sets that overlap the node set S among θ *RR* sets
$F_\theta(S)$	Weighted coverage function, the average weight of a node set S covering among θ *RR* sets
$\mathbb{E}[I(S)]$	The expected influence of a node set S
OPT_k	OPT_k is the maximal expected influence for a given k

Lee et al. [14] focused on maximizing the influence on specific users. They formulated an influence maximization problem as query processing to distinguish specific users. The method applies branch-and-bound method to prune search space for target users. The proposed method can only divide users into three types, i.e., target users, non-target users, and those who are immune from being influenced for an item. Thus, the method is not applicable when users have various types and semantics.

Besides, there are few more *IM* problems considering labels and weights linking with specific semantics. For instance, Li et al. [17] studied the labeled influence maximization, where each user is associated with a label. A Labeled Degree Discount (*LDD*) algorithm was proposed to solve this problem, extending degree discount heuristics by considering labels. In each round, *LDD* computes the users' ability to activate non-activated neighbors, and it measures this ability by estimating the sum of weights of activated neighbors. *LDD* does not have approximation guarantee, and it may perform poorly in some cases. For instance, *LDD* neglects users' ability to activate indirect neighbours, which introduces problems when some users being pruned only have strong influence in later rounds.

Wang et al. [39] studied influence maximization in the weighted IC (WIC) model. Influence Value Tree (*IVT*) and Weight Discount Tree are proposed (*WDT*) to store users' successors and predecessors, respectively. By updating *IVT* and *WDT*, users' influence gain can be estimated and the one with the highest influence gain is picked as a seed. To improve efficiency, the authors proposed a Bounded Weighted Reset algorithm (*BWR*), which uses a threshold θ to bound the sizes of *IVT* and *WDT*. The idea is to prune users with small influence. However, many users with small influence may have significant influence. The above methods were designed based on the independent cascade (IC) model, which cannot be straightforwardly applied into diffusion processes under the linear threshold (LT) model. This study aims a solution on a general diffusion model, i.e. triggering model [8,29,35].

3. Preliminaries

3.1. Problem definition

A social network is modelled as a directed graph $G(V, E)$, where V represents users in the network G and E represents following relationships between users. A user $v \in V$ in G may be associated with various semantics. We introduce a semantics translation function $\mathcal{M}$ on V to generalize the semantics without limitation of user classes or diffusion models. A semantics translation function represents the benefit of activating users. A higher value of the function indicates the corresponding users are more influential to being activated. The semantics translation function $\mathcal{M}$ may have different forms based on applications. Without loss of generalization, this study defines that $\mathcal{M}(v)$ translates the semantics of v into a real number in [0,1] for the simplicity of presentation.

To facilitate a quick lookup, Table 1 presents the essential notations used in the paper.

Next, the Triggering Model [8,35], summarized as follows, is used by the study to characterize a diffusion process on the network G:

- Given a seed set $S \in G$ to diffuse information, and all the nodes in S are initially activated as influenced users.
- Each node $v \in V$ in the network G independently picks a random triggering set $\mathcal{T}_v$ according to some distribution over subsets of its neighbors.
- The diffusion process is to iteratively active inactive nodes. Particularly, an inactive node v becomes active in iteration t if it has a neighbor in its triggering set $\mathcal{T}_v$, which is activated in iteration $t-1$.
- Once a node is activated, it keeps active in all the subsequent iterations. The diffusion process terminates when no more node becomes activated over an iteration.

Since the size of the node-set V is known, the total number of iterations towards termination is bounded by its cardinality $|V|$.

To be noted, existing studies (e.g., [8]) show that many well-known diffusion models including the *Independent Cascade* (*IC*) and the *Linear Threshold* (*LT*) models are special cases of the Triggering Model. Specifically, in the *IC* model, each edge (u, v) is associated with a value $p_{uv} \in [0, 1]$, which denotes the probability to active v by u. In the *LT* model, each edge (u, v)

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

is assigned a value $p_{uv} \in [0, 1]$ and each node v is activated by its neighbors if the following condition holds:

$$\sum_{u \in N_{in}(v)} \mathcal{I}(u \text{ is active}) \cdot p_{uv} \geq \mu_v,$$

where $N_{in}(v)$ is the set of incoming neighbors of v, $\mathcal{I}(\cdot)$ is an indicator function as shown in Table 1, and μ_v drawn from [0,1] uniformly at random is the activation threshold of v.

When the diffusion process starting from S terminates, the sum of $\mathcal{M}(v)$ of the activated nodes, denoted as $I(S)$, is used to measure the effect of the diffusion. Note that the information diffusion process is a random process although all the parameters are known. Thus, we use expected influence of S, denoted as $\mathcal{E}[I(S)]$, as the objective function. Formally, $\mathcal{E}[I(S)]$ is defined by:

$$\mathbb{E}[I(S)] = \sum_{v \in V} \Pr(S \rightsquigarrow v) \cdot \mathcal{M}(v), \tag{1}$$

where $\Pr(S \rightsquigarrow v)$ is the probability that v can be activated by S (i.e., S influences v).

We ultimately propose the *SIM* problem based on the above preliminaries as follows:

Problem Statement. On a social network $G(V, E)$, given a semantics translation function $\mathcal{M}$, the *Semantics-aware Influence Maximization* (*SIM*) problem is to find a size-k seed set S on G (k is a user-specified parameter), that maximizes the expected influence $\mathbb{E}[I(S)]$ in Eq. (1).

Unfortunately, we obtain the following theorem.

Theorem 1. *SIM is NP-hard unless P = NP.*

Proof. The traditional *IM* problem is shown to be NP hard [9]. Given a social network $G(V, E)$ and seed size k, the IM problem is to find a size-k seed set S on G that maximizes the expected influence:

$$S = argmax_{|S|=k} \mathbb{E}[I(S)] \tag{2}$$

$$\mathbb{E}[I(S)] = \sum_{v \in V} \Pr(S \rightsquigarrow v), \tag{3}$$

where $\Pr(S \rightsquigarrow v)$ is the probability that v can be activated by S (i.e., S influences v).

For an instance of the IM problem, we can construct an instance of the SIM problem. Given a social network $G(V, E)$ and a seed size k in the IM instance, we set $\mathcal{M}(v) = 1$, $\forall v \in V$. Therefore, Eq. (1) in the IM instance is equal to the Eq. (3) in the SIM instance. By finding seeds in the SIM instance, we also get the seeds in the IM instance. It means that the IM problem can be reduced to the SIM problem. SIM is as a result of this NP-hard. □

3.2. Reverse influence set review

The Reverse Influence Set (*RIS*) as a technique lies in the core of numerous *IM* solutions [16]. A key notation in *RIS* is reverse reachable (*RR*) set shown as follows:

Definition 1 (Reverse Reachable Set [36].)**.** A reverse reachable (*RR*) set for a node $v \in V$ is generated by the following two steps: i) sample a graph g from G by removing each edge with probability associated with $\mathcal{T}_v$, and ii) form a reverse reachable set R for v by the nodes in g that can reach v, where v is termed as the *root node* of R.

The following lemma reveals the connection between *RR* sets and the diffusion process under the Triggering model.

Lemma 1. *Given a node v, a set of nodes S, and an RR set R generated for v, let ρ_1 be the probability of S overlapping R, and ρ_2 be the probability that S (as a seed set) can activate v in a propagation process on G under the Triggering model. Then, we have $\rho_1 = \rho_2$.*

Based on Lemma 1, the *RIS* technique can be used to estimate the expected number of activated nodes by a node set $S \subseteq V$. The method is outlined by 4 steps:

1. Create a sequence of θ nodes, each of which is sampled from V uniformly at random;
2. Generate an *RR* set for each of the θ nodes;
3. Compute the number of *RR* sets that overlap the node set S, termed $H_\theta(S)$ as a coverage function;
4. Calculate $\frac{|V|}{\theta} \cdot H_\theta(S)$ as the estimated influence of S.

The study [36] shows that the term $\mathbb{E}[\frac{|V|}{\theta} \cdot H_\theta(S)]$ in Step 4 is an unbiased estimator of the objective function in *IM*. Most of the existing *RIS* based algorithms [3,36] utilize this property by applying a greedy approach to find a $(1 - 1/e)$-approximation solution for $\frac{|V|}{\theta} \cdot H_\theta(S)$. Such a solution also guarantees $(1 - 1/e - \epsilon)$-approximation for *IM* with $1 - \delta$ probability where ϵ and δ are two parameters between 0 and 1 if θ is sufficiently large. For θ, Tang et al. [36] give several lower bounds. To the best of our knowledge, the approximation guarantee is the best performance bound of the state-of-the-art IM solutions.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

Algorithm 1 ([36]) An *RIS* based Algorithm (G, k, ϵ, δ).

Require: ~~
G : A social network $G(V, E)$ with identical nodes.
k : Number of nodes to be selected.
ϵ, δ : Controlling parameters.
Ensure: ~~
S_k^* : A size-k seed set.
1: Sampling a set $\mathcal{R}$ of *RR* sets with random root nodes
2: $S_k^* = \emptyset$
3: **for** i := 1 to k **do**
4: v^* := $\arg\max_{v \in V \setminus S_k^*} H_{|\mathcal{R}|}(S_k^* \cup \{v^*\}) - H_{|\mathcal{R}|}(S_k^*)$
5: $S_k^* = S_k^* \cup \{v^*\}$
6: **end for**

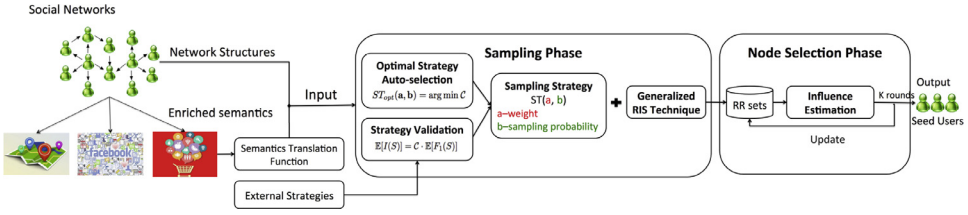


Fig. 2. GRIS-SIM framework overview.

Algorithm 1 shows a traditional *RIS* based algorithm where semantics are ignored [36]. First, a set $\mathcal{R}$ of RR sets are sampled, the root of each RR set is a node selected randomly from V (Line 1). The size of $\mathcal{R}$ is determined by the input parameters ϵ and δ. Then k nodes are found in k rounds (Lines 3–6). In each round, the node v^* with the maximum marginal increase is added into the final result S_k^* (Lines 4-5).

However, Algorithm 1 is designed for the *IM* problem where nodes do not have semantics. As discussed in Section 1, it may not find the most valuable seeds. Hence, we propose a more generalized and robust framework for the SIM problem in next section.

4. GRIS-SIM framework

This section presents our Generalized Reverse Reachable Set based framework for SIM (*GRIS-SIM*). Section 4.1 overviews the *GRIS-SIM framework*. Section 4.2 proposes the *GRIS* technique to sample RR sets in the *GRIS-SIM* framework. Section 4.3 presents methods of sampling strategy validation and influence spread estimation. In Section 4.4, we introduce seeds selection with given RR sets and prove the approximation ratio of the solution. Section 4.5 further discusses auto-generation of valid sampling strategies and the optimal strategy.

4.1. Overview

As illustrated in Fig. 2, *GRIS-SIM* can solve SIM problems with different semantics (e.g., locations, posts, shopping histories). The framework takes social networks and corresponding semantics as input. These networks can be modeled as graph $G(V, E)$ and semantics translation function $\mathcal{M}$ as discussed in Section 3.1. Semantics are integrated into the *GRIS-SIM* framework through a semantics translation function. A semantics translation function is a simple function defined by applications. For example, if vendors want to target female users, a translation function can be defined as:

$$\mathcal{M}(v) = \begin{cases} 1, \text{if v is a female user} \\ 0, \text{otherwise} \end{cases} \tag{4}$$

A translation function shall be able to simply represent the value of users' activation for the given application.

GRIS-SIM aims to efficiently find seed users that can achieve significant influence spread considering semantics. The framework takes the idea of *RIS* based on the *IM* algorithms that uses sampling *RR* sets to estimate influence spread and pick seeds. In order to automatically support different semantics and make sure the solution accuracy for SIM problem, *GRIS-SIM* enhances the sampling phase and bridges RR sets sampling and influence spread estimation under different semantic settings.

GRIS-SIM uses a two-stage process to pick seeds for every instance with given social network structure and semantics. It consists of sampling and node selection phases. In the sampling phase, a series of reverse reachable sets (RR sets) are

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

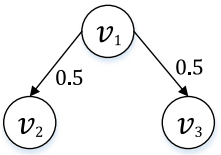


Fig. 3. A running example.

generated, which are used in node selection phase to estimate influence and pick seeds. Different from the traditional RIS-based solutions, *GRIS-SIM* provides a more flexible way to sample *RR* sets. The RR sets are granted with different generation probabilities and weights. External sampling strategies can be set to decide how to generate RR sets. In this case, *GRIS-SIM* can validate predefined strategies and estimate influence spread based on them. Still, *GRIS-SIM* can auto-generate sampling strategies and choose the optimal one for a given social network instance. The optimal strategy samples the smallest RR sets while guaranteeing $(1 - 1/e - \epsilon)$-approximation. In node selection phase, k seed nodes are iteratively selected. In each iteration, *GRIS-SIM* estimates influence spread in terms of the generated RR sets and pick nodes with maximum marginal gains. *GRIS-SIM* can achieve the best influence with the optimal $(1 - 1/e - \epsilon)$-approximation.

4.2. Generalized RIS technique and sampling strategy

The sampling process of *RIS* technique identically generates all RR sets, even for the nodes with different semantics. Accordingly, the *RIS* based solutions are hard to be straightforwardly adopted in *SIM* for two reasons: i) nodes may have different semantics that yields the term $\frac{|V|}{\theta} \cdot H_\theta(S)$ cannot be used to estimate the expected influence spread in *SIM*, and ii) the existing lower bounds on the number of *RR* sets do not hold for *SIM*, which means that we lose the well-proved properties of adopting *RIS* for *IM*. To this end, we propose the Generalized *RIS* (*GRIS*) technique to overcome the limitations of *RIS*.

GRIS technique provides a flexible way to generate RR sets based on sampling strategies. In *RIS*, each node is selected with equal probability to generate an RR set and only the number of RR set is considered in the coverage function $H_\theta(S)$. However, in *GRIS*, RR sets generation is based on the node probabilities, and each RR set R is assigned a weight $w_R \in [0, 1]$ to be considered in the coverage function. The weight and probability assignments are specified by sampling strategies as follow.

Definition 2 (Sampling Strategy)**.** A sampling strategy is represented as $ST(\mathbf{a}, \mathbf{b})$, where $\mathbf{a} \in \mathbb{R}^{|V|}$ is a $|V|$-dimension vector that assigns weight $\mathbf{a}_v \in [0, 1]$ to the RR set generated for node $v \in V$, and $\mathbf{b} \in \mathbb{R}^{|V|}$ is a $|V|$-dimension vector that represents the probabilities of each node being sampled. The vector $\mathbf{b}$ satisfies that $\forall v \in V$, $\mathbf{b}_v \in [0, 1]$ and $\sum_{v \in V} \mathbf{b}_v = 1$.

A sampling strategy characterizes the weight of each RR set and the probability of node sampling. Given a sampling strategy, generating an RR set follows the steps:

1. Select a node $v^* \in V$ by the probability vector $\mathbf{b}$.
2. Generate an *RR* set R for the node v^*.
3. Assign the weight of R to be $\mathbf{a}_{v^*}$.

4.3. Strategy validation

The purpose that we generate RR sets is to estimate the influence spread. The more RR sets a node-set can cover, the greater influence spread this node-set has. To validate if a strategy can be used to estimate influence spread, we firstly introduce weighted coverage function.

Weighted Coverage Function. Given a sequence of θ *RR* sets $R_1, R_2, \ldots, R_\theta$ generated by a sampling strategy $ST(\mathbf{a}, \mathbf{b})$, the weighted coverage function $F_\theta(S)$ is defined as:

$$F_\theta(S) = \frac{1}{\theta} \cdot \sum_{i=1}^{\theta} x_i \cdot w_i, \tag{5}$$

where x_i is 1 if $R_i \cap S \neq \emptyset$ otherwise it is 0, and w_i is the weight of R_i. The weighted coverage function calculates an average weight of a node set S covering among the θ RR sets.

Example. Consider a graph as shown in Fig. 3 with three nodes: v_1, v_2, and v_3. The corresponding semantic function is $\mathcal{M}(v) = 1, v \in \{v_1, v_2, v_3\}$. v_1 can activate v_2 and v_3 each with 0.5 probability. Suppose to estimate the expected influence in the network with a sampling strategy $ST(\mathbf{a}^*, \mathbf{b}^*)$, where $\mathbf{a}^* = [1, 1, 1]$, and $\mathbf{b}^* = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. The sampling strategy indicates that each node is selected with the same probability $\frac{1}{3}$ and the weights of all the generated RR sets are 1. We sample 3 RR sets: $R_1 = \{v_1\}, R_2 = \{v_3, v_1\}$, and $R_3 = \{v_2\}$. Note that in the 3 sets, the first node in each RR set is the root node. By Eq. (5), $F_3(\{v_1\}) = \frac{2}{3}$, $F_3(\{v_2\}) = \frac{1}{3}$ and $F_3(\{v_3\}) = \frac{1}{3}$.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

Next, we find that $F_\theta(S)$ has the following nice property:

Lemma 2. *For a sampling strategy ST*(**a, b**)*, its coverage function satisfies*

$$\mathbb{E}[F_\theta(S)] = \mathbb{E}[F_1(S)],$$

for any $S \subseteq V$ *and any* $\theta \in \mathbb{R}^+$.

Proof. The RR sets generated by a sampling strategy *ST*(**a, b**) follow the same distribution by **b**. It implies that for any $i \in \mathbb{R}^+$, we have

$$\mathbb{E}[x_i \cdot w_i] = \mathbb{E}[x_1 \cdot w_1] = F_1(S).$$

Unfolding $F_\theta(S)$ by definition, we have

$$\begin{aligned}\mathbb{E}[F_\theta(S)] &= \mathbb{E}[\tfrac{1}{\theta} \cdot \sum_{i=1}^{\theta} x_i \cdot w_i] \\ &= \tfrac{1}{\theta} \cdot \sum_{i=1}^{\theta} \mathbb{E}[x_i \cdot w_i] = F_1(S).\end{aligned}$$

The lemma is therefore proved. □

The sampling strategies can be multiple on a social network to generate RR sets. However, $\mathbb{E}[I(S)]$ is unique for any node-set S. We, therefore, define that a sampling strategy is valid only if $\mathbb{E}[I(S)]$ can be calculated by its weighted coverage function $F_\theta(S)$.

Definition 3 (Valid Sampling Strategy)**.** On a social network $G(V, E)$, a sampling strategy *ST*(**a, b**) is called a valid sampling strategy if there exists a coefficient $\mathcal{C} \in \mathbb{R}$ that satisfies

$$\mathbb{E}[I(S)] = \mathcal{C} \cdot \mathbb{E}[F_1(S)] \tag{6}$$

for any node set $S \subseteq V$.

Example. For the running example in Fig. 3, the sampling strategy *ST*($\mathbf{a}^*$, $\mathbf{b}^*$) where $\mathbf{a}^* = [1, 1, 1]$, and $\mathbf{b}^* = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ is a valid sampling strategy because for any $S \subseteq V$, we have

$$\begin{aligned}\mathbb{E}[I(S)] &= \sum_{v \in V} \Pr(v \text{ is activated by } S) \cdot \mathcal{M}(v) \\ &= \sum_{v \in V} \Pr(S \text{ overlaps RR sets of } v) \cdot 1 \\ &= 3 \cdot \sum_{v \in V} \frac{1}{3} \cdot \Pr(S \text{ overlaps RR sets of } v) \\ &= 3 \cdot \mathbb{E}[F_\theta(S)]\end{aligned}$$

Therefore, we can find a coefficient $\mathcal{C} = 3$.

Given a valid sampling strategy, we can estimate the influence spread by its weighted coverage function. Note that a sampling strategy is not directly linked with user semantics, it only tells how to generate RR sets. The semantics impacts coefficient $\mathcal{C}$ in Eq. (6), but not how a sampling strategy generates RR sets.

4.4. Node selection

Algorithm 2 presents how *GRIS-SIM* framework solves a *SIM* instance. With the *GRIS* technique and valid sampling strategy *ST*($\mathbf{a}^*$, $\mathbf{b}^*$) (Lines 1–2), k nodes are selected iteratively to form the result node set S_k^*. In each iteration, the node v^* that can bring the maximum marginal gain with regard to $F_{|\mathcal{R}|}(S_k^*)$ is added into S_k^* (Lines 6–7).

The node selection process can be regarded as an approach to solve the weighted maximum coverage problem [37]. Thus, S_k^* is at least a $(1 - 1/e)$-approximation solution of $\arg\max_{S \subseteq V} F_{|\mathcal{R}|}(S)$. And as the size of $\mathcal{R}$ grows, $F_{|\mathcal{R}|}(S)$ is likely to get closer to $\mathbb{E}[I(S)]/\mathcal{C}$ by Definition 3. We can thus also expect S_k^* getting closer to a $(1 - 1/e)$-approximation solution of $\arg\max_{S \subseteq V} \mathbb{E}[I(S)]$. Hereby, we utilize the following central limit on martingales [5] in Lemma 3 to carry out the theoretical analysis. The proof is similar to the IM algorithm approximation proof [36], the key differences are the construction of martingales and the influence estimated function.

Note that a sequence of random variables $Y_1, Y_2, Y_3, \ldots$ is a martingale, if and only if $\mathbb{E}[|Y_i|] < +\infty$ and $\mathbb{E}[Y_i | Y_1, Y_2, \ldots, Y_{i-1}] = Y_{i-1}$ for any $i > 1$ [36].

Lemma 3. *([5]) Let* $Y_1, Y_2, \ldots$ *be a martingale, such that* $|Y_1| \le a$, $|Y_j - Y_{j-1}| \le a$ *for any* $j \in [1, i]$, *and*

$$Var[Y_1] + \sum_{j=2}^{i} Var[Y_j | Y_1, Y_2, \ldots, Y_{j-1}] \le b,$$

Algorithm 2 GRIS-SIM($G, \mathcal{M}, k, \epsilon, \delta$).

Require: ~~
G : A social network.
$\mathcal{M}$: A semantics translate function.
k : Number of nodes to be selected.
ϵ, δ : Controlling parameters.
Ensure: ~~
S_k^* : A size-k seed set.
1: Construct a valid sampling strategy $ST(\mathbf{a}^*, \mathbf{b}^*)$
2: Generate a set $\mathcal{R}$ of RR sets with $ST(\mathbf{a}^*, \mathbf{b}^*)$
3: /* Node Selection */
4: $S_k^* := \emptyset$
5: **for** i = 1 to k **do**
6: $v^* := \arg\max_{v \in V \setminus S_k^*} F_{|\mathcal{R}|}(S_k^* \cup \{v^*\}) - F_{|\mathcal{R}|}(S_k^*)$
7: $S_k^* = S_k^* \cup \{v^*\}$
8: **end for**

where Var$[\cdot]$ *denotes the variance of a random variable. Then, for any* $\gamma > 0$, *we have*

$$\Pr[(Y_i - \mathbb{E}[Y_i]) \geq \gamma] \leq \exp\left(-\frac{\gamma^2}{\frac{2}{3}a\gamma + 2b}\right). \tag{7}$$

To utilize the central limit on martingales in Lemma 3, we first show that the sampled RR sets in Algorithm 2 can form a martingale in Lemma 4.

Lemma 4. *For a sequence of random RR sets* $R_1, R_2, \ldots, R_\theta$ *generated by a sampling strategy* $ST(\mathbf{a}, \mathbf{b})$ *and any* $S \subseteq V$, *let* $M_i = i \cdot F_i(S) - i \cdot \mu$ *where* $\mu = \mathbb{E}[F_1(S)]$, *the sequence of random variables* $M_1, M_2, \ldots, M_\theta$ *is a martingale.*

Proof. Let x_i and $w_i (i \in [1, \theta])$ be as defined in Eq. (5). As each *RR* set is generated by the same sampling strategy, for any i, $j \in [1, \theta]$, we have

$$\begin{aligned}&\mathbb{E}[x_i \cdot w_i | M_1, M_2, \ldots, M_{i-1}] \\ &\quad = \mathbb{E}[x_i \cdot w_i] = \mathbb{E}[F_1(S)].\end{aligned}$$

Hereby,

$$\begin{aligned}&\mathbb{E}[M_i | M_1, M_2, \ldots, M_{i-1}] \\ &\quad = \mathbb{E}[x_i \cdot w_i - \mu + \sum_{j=1}^{i-1} x_j \cdot w_j - (i-1) \cdot \mu | M_1, M_2, \ldots, M_{i-1}] \\ &\quad = \mathbb{E}[x_i \cdot w_i - \mu + M_{i-1} | M_1, M_2, \ldots, M_{i-1}] \\ &\quad = M_{i-1}.\end{aligned}$$

Therefore, $M_1, M_2, \ldots, M_\theta$ is a martingale by definition. □

Applying Lemma 3 on the martingale $M_1, M_2, \ldots, M_\theta$, we can obtain Lemma 5 disclosing the relationship between the number of sampled random RR sets (i.e., θ) and the accuracy of $\mathbb{E}[I(S)]$ estimation.

Lemma 5. *Let* $ST(\mathbf{a}^*, \mathbf{b}^*)$ *be the valid sampling strategy obtained in Line 1 of Algorithm 2 and* $R_1, R_2, \ldots, R_\theta$ *be a sequence of RR sets generated by* $ST(\mathbf{a}^*, \mathbf{b}^*)$. *Let* $S \subseteq V$ *be any set of nodes and* $\mu = \mathbb{E}[I(S)]/\mathcal{C}^*$ *with* $\mathbb{E}[I(S)] = \mathcal{C}^* \cdot \mathbb{E}[F_1(S)]$. *Then for any* $\epsilon > 0$, *we have*

$$\Pr[F_\theta(S) \geq (1+\epsilon) \cdot \mu] \leq \exp\left(-\frac{\epsilon^2}{2 + \frac{2}{3}\epsilon} \cdot \theta\mu\right) \tag{8}$$

and

$$\Pr[F_\theta(S) \leq (1-\epsilon) \cdot \mu] \leq \exp\left(-\frac{\epsilon^2}{2} \cdot \theta\mu\right). \tag{9}$$

Proof. Note that all the weights are scaled to the interval [0,1] in Algorithm 2, we have $\mu = \mathbb{E}[F_\theta(S)] \in [0, 1]$ and $|M_i - M_{i-1}| = |x_i \cdot w_i| \in [0, 1]$, where $i \in [1, \theta]$. We also have

$$\begin{aligned}&Var[M_1] + \sum_{j=2}^{i} Var[M_j | M_1, M_2, \ldots, M_{j-1}] \\ &= Var[x_1 \cdot w_1 - \mu]\end{aligned}$$

$$+\sum_{j=2}^{i} Var[x_j \cdot w_j - \mu + M_{j-1} | M_1, M_2, \ldots, M_{j-1}]$$
$$= \sum_{j=1}^{i} Var[x_j \cdot w_j] = \sum_{j=1}^{i} \left(\mathbb{E}[(x_j \cdot w_j)^2] - \mathbb{E}[x_j \cdot w_j]^2\right). \tag{10}$$

Note that $\mathbb{E}[(x_j \cdot w_j)^2] \le \mathbb{E}[x_j \cdot w_j]$ as $x_j \cdot w_j \in [0, 1]$ and $\mathbb{E}[x_j \cdot w_j] = \frac{1}{\theta} \cdot \mathbb{E}\left[\sum_i^{\theta} x_i \cdot w_i\right]$, thus we have the right hand side of Eq. 10 to be

$$\begin{aligned}
&\sum_{j=1}^{i} \left(\mathbb{E}[(x_j \cdot w_j)^2] - \mathbb{E}[x_j \cdot w_j]^2\right) \\
&\le \sum_{j=1}^{\theta} \left(\mathbb{E}[x_j \cdot w_j] - \mathbb{E}[x_j \cdot w_j]^2\right) \\
&= \theta\mu \cdot (1 - \mu) \le \theta\mu.
\end{aligned}$$

Thus, according to Lemma 3, we can apply martingale central limit on the martingale $M_1, M_2, \ldots, M_\theta$. By setting $a = 1, b = \theta\mu$, we can obtain Eq. (8). Similarly, for $-M_1, \ldots, -M_\theta$, set $a = 1, b = -\theta\mu$, Eq. (9) holds, which completes the proof. □

Lemma 5 serves as the fundamental of our analysis on Algorithm 2. And then we can derive a formula on the size of $\mathcal{R}$ if we want to ensure S_k^* is at least a $(1 - 1/e - \epsilon)$-approximation solution to *SIM* with $1 - \delta$ probability. Note that this approximation is the best performance of the state-of-the-art *IM* solutions to the best of our knowledge.

First, suppose that S_k° is the optimal solution of *SIM* and $OPT_k = \mathbb{E}[I(S_k^\circ)]$, the following lemma shows that $\mathcal{C} \cdot F_\theta(S_k^\circ)$ gets closer to *OPT* as θ gets larger.

Lemma 6. *On a social network $G(V, E)$ with a semantics translate function $\mathcal{M}$, let ST(**a, b**) be a valid sampling strategy with $\mathbb{E}[I(S)] = \mathcal{C} \cdot \mathbb{E}[F_1(S)]$ for any $S \subseteq V$ on G. Let $\epsilon_1, \delta_1 \in (0, 1)$ and $\theta = |\mathcal{R}|$. Then*

$$\Pr[\mathcal{C} \cdot F_\theta(S_k^\circ) \ge (1 - \epsilon_1) \cdot OPT_k] \ge 1 - \delta_1 \tag{11}$$

holds if $\theta > \theta_1$, where

$$\theta_1 = \frac{2\mathcal{C} \cdot \log(1/\delta_1)}{OPT_k \cdot \epsilon_1^2}. \tag{12}$$

Proof. Let $\mu = \mathbb{E}[F_1(S_k^\circ)] = OPT_k/\mathcal{C}$, by Inequality 9, we have

$$\begin{aligned}
&\Pr[\mathcal{C} \cdot F_\theta(S_k^\circ) \le (1 - \epsilon_1) \cdot OPT_k] \\
&= \Pr[\mathcal{C} \cdot F_\theta(S_k^\circ) \le (1 - \epsilon_1) \cdot \mathcal{C}\mu] \\
&= \Pr[F_\theta(S_k^\circ) \le (1 - \epsilon_1) \cdot \mu] \\
&\le \exp\left(-\tfrac{\epsilon_1^2}{2} \cdot \theta\mu\right) \le \exp\left(-\tfrac{\epsilon_1^2}{2} \cdot \theta_1\mu\right) \\
&= \delta_1.
\end{aligned}$$

Thus, the lemma is proved. □

Next, the size-k node set S_k^* returned by Algorithm 2 is likely $(1 - 1/e - \epsilon)$-approximation of S_k° as Lemma 7.

Lemma 7. *On a social network $G(V, E)$ with a semantics translate function $\mathcal{M}$, let ST(**a, b**) be a valid sampling strategy with $\mathbb{E}[I(S)] = \mathcal{C} \cdot \mathbb{E}[F_1(S)]$ for any $S \subseteq V$, $\delta_2 \in (0, 1)$, $\epsilon > \epsilon_1$, and $\theta = |\mathcal{R}|$, then*

$$\Pr[\mathbb{E}[I(S_k^*)] \ge (1 - 1/e - \epsilon) \cdot OPT_k] \ge 1 - \delta_2 \tag{13}$$

holds if $\mathcal{C} \cdot F_\theta(S_k^\circ) \ge (1 - \epsilon_1) \cdot OPT_k$ and $\theta > \theta_2$, where

$$\theta_2 = \frac{(2 - 2/e) \cdot \mathcal{C} \cdot \log(\binom{|V|}{k}/\delta_2)}{OPT_k \cdot (\epsilon - (1 - 1/e) \cdot \epsilon_1)^2}. \tag{14}$$

Proof. The node set S_k^* returned by Algorithm 2 is at least a $(1 - 1/e)$-approximation of $argmax_{|S|=k} F_{|\mathcal{R}|}(S)$. Thus,

$$\begin{aligned}
\mathcal{C} \cdot F_\theta(S_k^*) &\ge (1 - 1/e) \cdot \mathcal{C} \cdot F_\theta(S_k^\circ) \\
&\ge (1 - 1/e) \cdot (1 - \epsilon_1) \cdot OPT_k.
\end{aligned} \tag{15}$$

Let $\mu = \mathbb{E}[F_1(S_k^*)] = \mathbb{E}[I(S_k^*)]/\mathcal{C}$ and $\epsilon_2 = \epsilon - (1 - 1/e) \cdot \epsilon_1$, we have

$$\begin{aligned}
&\Pr[\mathcal{C} \cdot F_\theta(S_k^*) - \mathbb{E}[I(S_k^*)] \ge \epsilon_2 \cdot OPT_k] \\
&= \Pr[\theta \cdot F_\theta(S_k^*) - \theta\mu \ge \tfrac{\epsilon_2 \cdot OPT_k}{\mathcal{C}\mu} \cdot \theta\mu] \\
&= \Pr[F_\theta(S_k^*) \ge \left(1 + \tfrac{\epsilon_2 \cdot OPT_k}{\mathcal{C}\mu}\right) \cdot \mu].
\end{aligned} \tag{16}$$

Let $\zeta = \frac{\epsilon_2 \cdot OPT_k}{\mathcal{C}\mu}$. By Inequality (8) and Eq. (15), we have the right hand side of Eq. (16) satisfying

$$\begin{aligned}
&\text{r.h.s of Eqn. 16}\\
&\leq \exp\left(-\frac{\zeta^2}{2+\frac{2}{3}\zeta}\cdot\theta\mu\right)\\
&= \exp\left(-\frac{\epsilon_2^2\cdot OPT_k^2}{2\mathcal{C}^2\mu+\frac{2}{3}\epsilon_2\mathcal{C}\cdot OPT_k}\cdot\theta\right)\\
&< \exp\left(-\frac{\epsilon_2^2\cdot OPT_k^2}{2\mathcal{C}(1-1/e-\epsilon)\cdot OPT_k+\frac{2}{3}\epsilon_2\mathcal{C}\cdot OPT_k}\theta\right)\\
&< \exp\left(-\frac{(\epsilon-(1-1/e)\epsilon_1)^2\cdot OPT_k}{(2-2/e)\cdot\mathcal{C}}\cdot\theta\right)\\
&\leq \exp\left(-\frac{(\epsilon-(1-1/e)\epsilon_1)^2\cdot OPT_k}{(2-2/e)\cdot\mathcal{C}}\cdot\theta_2\right)\\
&= \delta_2/\binom{|V|}{k}.
\end{aligned}$$

As there are $\binom{|V|}{k}$ combinations for S_k^*, Inequality 13 does not hold with at most δ_2 probability. The lemma is hereby proved. □

Putting together Lemmas 6 and 7, we finally derive the number of RR sets Algorithm 2 needs to ensure a $(1-1/e-\epsilon)$-approximation solution with high probability as Theorem 2.

Theorem 2. *For any ϵ, $\delta > 0$, Algorithm 2 returns a $(1-1/e-\epsilon)$-approximation solution with at least $1-\delta$ probability if the size of $\mathcal{R}$ in Line 2 is at least λ/OPT_k, where*

$$\lambda = 2\mathcal{C}\cdot\left((1-1/e)\sqrt{\log(2/\delta)}+\alpha\right)^2\cdot\epsilon^{-2}, \tag{17}$$

and

$$\alpha = \sqrt{(1-1/e)\log(2\binom{|V|}{k}/\delta)}. \tag{18}$$

Proof. Let $\delta_1 = \delta_2 = \delta/2$ and $\epsilon_1 = \epsilon\cdot\frac{\sqrt{\log(2/\delta)}}{(1-1/e)\cdot\sqrt{\log(2/\delta)}+\alpha}$. It can be derived that

1. By Lemma 6, if $\mathcal{R} \geq \lambda/OPT$, then $\mathcal{C}\cdot F_{|\mathcal{R}|}(S_k^\circ) \geq (1-\epsilon_1)\cdot OPT_k$ holds with at least $1-\delta_1$ probability.
2. By Lemma 7, $\mathbb{E}[I(S_k^*)] \geq (1-1/e-\epsilon)\cdot OPT_k$ holds with at least $1-\delta_2$ probability if Lemma 6 holds and $\mathcal{R} \geq \lambda/OPT$.

Thus by union bound, if $|\mathcal{R}| \geq \lambda/OPT$, then

$$\begin{aligned}
&\Pr[\mathbb{E}[I(S_k^*)] \geq (1-1/e-\epsilon)\cdot OPT_k]\\
&\geq (1-\delta_1)\cdot(1-\delta_2) \geq 1-\delta_1-\delta_2 = 1-\delta
\end{aligned}$$

which proved the theorem. □

Theorem 2 can obtain the sampling size by knowing how to construct a valid sampling strategy and how to determine OPT_k. Note that OPT_k is the maximum weighted spread in expectation, which is NP-hard to obtain. Alternatively, there are a bunch of techniques [36,38] that can estimate a lower bound LB of OPT_k. As $\lambda/LB \geq \lambda/OPT_k$, it is sufficiently large for the sampling size.

4.5. Optimal strategy auto-selection

In this section, we discuss how to construct a valid sampling strategy automatically for different semantics settings. Moreover, we find an optimal strategy that can achieve optimal time complexity. We leverage the techniques in the study [38] to estimate lower bound LB. In the study [38], following sampling strategy is adopted:

Proposition 1. *On a social network $G(V, E)$ with semantic translate function $\mathcal{M}$, the sampling strategy $ST_1 = ST(\mathbf{a}, \mathbf{b})$ is a valid sampling strategy, where*

$\mathbf{a}_v = \mathcal{M}(v), \forall v \in V$, *and*
$\mathbf{b}_v = 1/|V|, \forall v \in V$.

Proof. By Eq. (1) and Lemma 1, we have

$$\begin{aligned}
\mathbb{E}[I(S)] &= \sum_{v\in V}\Pr(S \rightsquigarrow v)\cdot\mathcal{M}(v)\\
&= \sum_{v\in V}\Pr(RR \text{ sets of } v \text{ overlapping } S)\cdot\mathcal{M}(v)\\
&= |V|\cdot\sum_{v\in V}\frac{1}{|V|}\Pr(RR \text{ sets of } v \text{ overlapping } S)\cdot\mathcal{M}(v)
\end{aligned}$$

$$= |V| \cdot \mathbb{E}[F_1(S)].$$

Thus by Definition 3, ST_1 is a valid sampling strategy with $\mathcal{C} = |V|$. □

The expected running time of the *GRIS-SIM* that adopts ST_1 is $O((k - \log\delta/\log|V|)(|V| + |E|)\log|V|/\varepsilon^2)$ [13,36]. ST_1 defines such a sampling strategy that at first each node v is selected with the same probability of $1/|V|$ and then the RR set generated for v is assigned weight $\mathcal{M}(v)$. However, the strategy is not very efficient. If there are nodes with 0 weight in the social network, many RR sets with 0 weights may be generated as each node is selected to generate RR sets with equivalent probability. By Eq. (5), these 0-weight RR sets have no contribution in influence maximization and thus can be eliminated. Therefore, we propose the following sampling strategy.

Proposition 2. *On a social network $G(V, E)$ with a semantics translate function $\mathcal{M}$, let $V^+ = \{v \in V | \mathcal{M}(v) > 0\}$. The sampling strategy $ST_2 = ST(\mathbf{a}, \mathbf{b})$ is a valid sampling strategy, where*

$$\mathbf{a}_v = \mathcal{M}(v), \forall v \in V, \text{ and}$$
$$\mathbf{b}_v = \begin{cases} 1/|V^+| & , v \in V^+ \\ 0 & , v \notin V^+. \end{cases}$$

Proof. By Eq. (1) and Lemma 1, we have

$$\begin{aligned} \mathbb{E}[I(S)] &= \sum_{v \in V} \Pr(S \rightsquigarrow v) \cdot \mathcal{M}(v) \\ &= \sum_{v \in V^+} \Pr(S \rightsquigarrow v) \cdot \mathcal{M}(v) + 0 \\ &= \sum_{v \in V^+} \Pr(RR \text{ sets of } v \text{ overlapping } S) \cdot \mathcal{M}(v) \\ &= |V^+| \cdot \sum_{v \in V^+} \frac{\Pr(RR \text{ sets of } v \text{ overlapping } S)}{|V^+|} \cdot \mathcal{M}(v) \\ &= |V^+| \cdot \mathbb{E}[F_1(S)]. \end{aligned}$$

Thus by Definition 3, ST_2 is a valid sampling strategy with $\mathcal{C} = |V^+|$. □

Compared with ST_1, ST_2 only selects nodes with positive weights and thus it reduces the number of RR sets. The expected running time of the *GRIS-SIM* that adopts ST_2 is $O(|V^+| \cdot (k - \log\delta/\log|V|)(|V| + |E|)\log|V|/|V| \cdot \varepsilon^2)$. ST_2 reduces the expected time complexity by a factor of $O(\frac{|V|}{|V^+|})$ compared with ST_1. It further rises a question whether there exists an optimal strategy in terms of the required RR sets. To address this problem, we have the following lemma that can compare two valid sample strategies.

Lemma 8. *Given ST' and ST'' as two valid sampling strategies on the social network $G(V, E)$, $F_1'(S)$ and $F_1''(S)$ are their corresponding coverage functions. For any $S \subseteq V$, if $\mathbb{E}[F_1'(S)] \geq \mathbb{E}[F_1''(S)]$, then using ST' in Algorithm 2 requires less RR sets comparing with ST''.*

Proof. By Definition 3, for any $S \subseteq V$, there are coefficients $\mathcal{C}'$ and $\mathcal{C}''$ satisfying

$$\mathbb{E}[I(S)] = \mathcal{C}' \cdot \mathbb{E}[F_1'(S)] = \mathcal{C}'' \cdot \mathbb{E}[F_1''(S)].$$

Thus, we have $\mathcal{C}' \leq \mathcal{C}''$ given $\mathbb{E}[F_1'(S)] \geq \mathbb{E}[F_1''(S)]$. On the other hand, the number of RR sets is monotonically increasing with regard to $\mathcal{C}$ by Eq. (17). Thus, ST' requires less RR sets, which proves the lemma. □

Furthermore, Lemma 9 provides a sufficient and necessary condition to construct a valid sampling strategy.

Lemma 9. *On a network $G(V, E, W)$, a sampling strategy $ST(\mathbf{a}, \mathbf{b})$ satisfies $\mathbb{E}[I(S)] = \mathcal{C} \cdot \mathbb{E}[F_1(S)]$ for some $\mathcal{C}$ if and only if*

$$\mathcal{M}(v) = \mathcal{C} \cdot a_v \cdot b_v, \forall v \in V. \tag{19}$$

Proof. $\mathbb{E}[I(S)]$ can be written as

$$\mathbb{E}[I(S)] = \sum_{v \in V} \Pr(S \rightsquigarrow v) \cdot \mathcal{M}(v).$$

And $\mathbb{E}[F_1(S)]$ can be unfolded as

$$\mathbb{E}[F_1(S)] = \sum_{v \in V} a_v \cdot b_v \cdot \Pr(\text{RR sets of } v \text{ overlapping } S)$$

If Eq. (19) holds, then by Lemma 1, we have

$$\begin{aligned} \mathbb{E}[F_1(S)] &= \sum_{v \in V} a_v \cdot b_v \cdot \Pr(\text{RR sets of } v \text{ overlapping } S) \\ &= \sum_{v \in V} \mathcal{M}(v)/\mathcal{C} \cdot \Pr(S \rightsquigarrow v) = \mathbb{E}[I(S)]/\mathcal{C}. \end{aligned}$$

Thus, Eq. (19) is a sufficient condition.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

On the other hand, if $\mathbb{E}[I(S)] = \mathcal{C} \cdot \mathbb{E}[F_\theta(S)]$ for any $S \subseteq V$, then by Eq. (1) and Lemma 1, we have

$$\sum_{v \in V} \Pr(S \rightsquigarrow v) \cdot \mathcal{M}(v) = \mathcal{C} \cdot \sum_{v \in V} a_v \cdot b_v \cdot \Pr(S \rightsquigarrow v) \tag{20}$$

for any $\Pr(S \rightsquigarrow v) \in [0, 1]$, where $v \in V$. We can obtain $\mathcal{M}(v) = \mathcal{C} \cdot a_v \cdot b_v$ $(\forall v \in V)$ by setting $\Pr(S \rightsquigarrow u) = 0, \forall u \neq v$. Thus, Eq. (19) is a necessary condition.

Combing the two parts, the lemma is therefore proved. □

By Definition 2 and Lemma 9, we can conclude that the optimal sampling strategy $ST(\mathbf{a}, \mathbf{b})$ shall be the solution of the following optimization problem:

$$ST_{opt}(\mathbf{a}, \mathbf{b}) = \arg\min \mathcal{C}$$

$$\text{s.t.} \quad \mathcal{M}(v) = \mathbf{C} \cdot a_v \cdot b_v, \forall v \in V, \tag{21}$$

$$\sum_{v \in V} b_v = 1, \tag{22}$$

$$\mathbf{0} \leq \mathbf{a} \leq \mathbf{1}, \tag{23}$$

$$\mathbf{b} \geq \mathbf{0}, \mathcal{C} \geq 0. \tag{24}$$

To address the problem, we reform Eq. (21) as:

$$b_v = \frac{\mathcal{M}(v)}{\mathcal{C} \cdot a_v}. \tag{25}$$

Then using Eq. (25) to replace b_v in Eq. (22), we have

$$\mathcal{C} = \sum_{v \in V} \frac{\mathcal{M}(v)}{a_v}.$$

By Eq. (23), we have the minimum $\mathcal{C}$ is $\sum_{v \in V} m_{z_v}$ by setting $a_v = 1$, $\forall v \in V$, which draws Theorem 3.

Theorem 3. *For an instance of SIM on a network $G(V, E)$ with a semantics translate function $\mathcal{M}$, the sampling strategy that achieves the optimal time complexity $O(\mathcal{C} \cdot (k - \log \delta / \log |V|)(|V| + |E|) \log |V| / |V| \cdot \varepsilon^2)$ in the GRIS-SIM framework is $ST_{opt} = ST(\mathbf{a}, \mathbf{b})$ where*

$$\mathcal{C} = \sum_{v \in V} \mathcal{M}(v),$$
$$a_v = 1, \forall v \in V,$$
$$b_v = \tfrac{\mathcal{M}(v)}{\mathcal{C}}, \forall v \in V.$$

5. Experimental evaluation

5.1. Settings and experimental methodology

[Datasets]. We employ six real public datasets[1] with various associated network semantics, including four social networks with users' label information and 2 location-based social networks. We also use a series of synthetic datasets for scalability evaluation. Here are the brief descriptions of the real-world datasets. Hamsterster [11] is a friendship network among users on hamsterster.com. The users are labelled by a binary gender, i.e., male and female. BlogCatalog [42] includes both contact network and selected interests. Users are labeled by their interests. Flickr [42] is a network representing both friendship and group information. Users are labeled by their joined groups. YouTube [42] is a contact network with user-group information. Users are labeled by their joined groups. Brightkite and Gowalla [34,41] are two friendship social networks where users share their locations by checking in.

Table 2 presents the properties of the real datasets. The synthetic data is generated by the tool [7], which is scale-free and large, up to **10 million** nodes.

[Competitors]. We compare our solutions (i.e., *GRIS-SIM* with three different sampling strategies) against the following algorithms in terms of effectiveness, efficiency, and scalability. Note that we use *GRIS-SIM*$_1$, *GRIS-SIM*$_2$, and *GRIS-SIM*$_{opt}$ to represent solutions that adopt the strategies ST_1, ST_2, and ST_{opt}, respectively.

- **BWR** [39] is a heuristic algorithm proposed under the Independent Cascade (IC) model, which maintains users' influence path by tree structures and bounds the volume of trees to improve the efficiency.

[1] http://socialcomputing.asu.edu/pages/datasets

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

Table 2
Properties of the real datasets.

Datasets	# of nodes	# of labels	# of edges
Hamsterster	1858	2	12,534
BlogCatalog	10,312	39	333,983
Flickr	80,513	195	5,899,882
YouTube	1,138,499	48	2,990,443
Brightkite	58,228	–	214,078
Gowalla	196,591	–	950,327

- **LDD** [17] is heuristic by modifying the degree discount heuristic strategy to solve labeled influence maximization problem. It can be transformed into the *SIM* problem under the Independent Cascade (IC) model.
- **HighWeightDeg** is a simple heuristic algorithm that picks k nodes with the highest weighted degree. Since both *BWR* and *LDD* can only be applied to the IC model, we use *HighWeightDeg* as the baseline under the Linear Threshold (LT) model.

To show wide uses of the proposed *GRIS* techniques, we apply *GRIS* to the following algorithm designed for the *IM* problem with specific semantics.

- **RIS-DA** [38] is a reverse influence sampling-based method for the Distance-Aware Influence Maximization (DAIM). In *DAIM*, given a promoted location q, a user's distance from q is transformed into a weight. To show the generality and effectiveness of our *GRIS* technique, we use it as a component in the original algorithm as *GRIS-DA* for evaluation.

Note that *BWR* and *LDD* are heuristic, and we have not noticed any given approximation ratio for theoretical guarantee. We implemented our solutions in C++ and we adopted the C++ implementations of the other algorithms with recommended parameter settings. For *BWR*, we have to bound the length of influence path to 4, because *BWR* is intolerable costly (24 hours+) under general influence probability setting for IC model.

[Semantics translate function]. Our proposed framework supports general semantics translate function. We consider the typical application of *SIM* (i.e., viral marketing). For viral marketing, vendors usually measure users' values based on their types, just as the example in Fig. 1. Only parts of the users are targeted, some users are highly influential while others are less. Hence, we grant users semantic values based on their labels for the first four datasets. Furthermore, we select 50% types of users as target users. Among targeted users, half types of them are chosen as highly influential users while other users are chosen as less influential users. Since the semantic values can be normalized into range [0, 1], we randomly select semantic values in [0.9, 1] for highly influential users and those in [0, 0.1] for less influential users. As for the last two location-based social networks, we use user's weights computed in the study [38] as their semantic values for fair comparison.

[Diffusion Models and Cascading]. Our proposed framework supports instances of triggering model. However, triggering model is an abstract model that needs to be instantiated. In our experiments, two well-adopted instances of triggering model are considered. Namely the Independent Cascade (IC) and the Linear Threshold (LT) models [8,26]. For the IC model, we set the propagation probability of each edge (u, v) to be $1/d_v$, where d_v is the in-degree of node v. This setting is well-used by the existing studies [26]. For the LT model, we assign each incoming neighbour u of node v a value that is randomly drawn from [0,1], and then we normalize the assigned values in order to have the sum less than 1. Assigned value is the probability of a node to be sampled from the triggering set of node v.

[Environment]. All the experiments were conducted on an Intel i5 3.2GHz CPU machine with 32GB RAM.

We estimated the expected influence of a seed set by taking its average influence in 10,000 simulations of the diffusion process.

5.2. Result discussion for SIM

This section discusses the findings for the *SIM* problem with various datasets from the aspects of effectiveness, efficiency, scalability, and parameter sensitivity. All the results are averaged over 5 runs.

5.2.1. Effectiveness

The effectiveness of SIM solutions means the picked seeds can eventually have great influence among the social networks. Thus, we can use the estimated influence of chosen seeds to measure the effectiveness of the SIM solutions. The greater estimated influence the picked seeds can achieve, the more effective the solution is. We evaluate the effectiveness of our solutions against two rivals on the 4 datasets under both IC and LT models, with k varying from 10 to 90. The results are shown in Figs. 4.

Fig. 4 shows that under the IC model, our *GRIS* based solutions ((i.e., *GRIS-SIM*$_2$, *GRIS-SIM*$_2$, and *GRIS-SIM*$_{opt}$) achieve higher expected influence than the competitors (i.e., *BWR* and *LDD*) on all the 4 real datasets. The expected influences of the *GRIS* based solutions are close due to the same theoretical approximation ratios. Similar results can be observed from Fig. 5.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

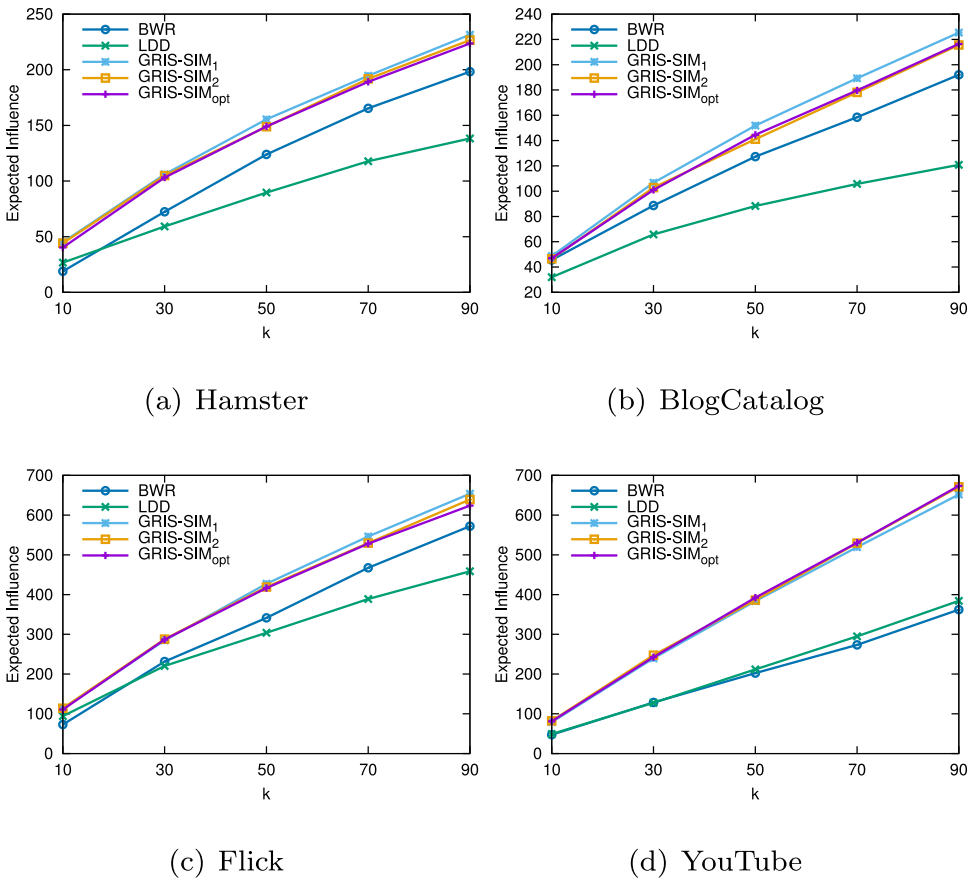


Fig. 4. Expected influence varying k under the IC model ($\varepsilon = 0.5$ and $l = 1$ for *GRIS*).

Table 3
Number and ratio of different users.

Sources	Low semantic values	High semantic values
Hamsterster	968 (0.521)	889 (0.479)
BlogCatalog	7799 (0.756)	2513 (0.244)
Flickr	52,690 (0.654)	27,823 (0.346)
YouTube	20,838 (0.018)	1,117,661 (0.982)

A higher ratio of activated users with high semantic values means that the solution is more effective in distinguishing the high from those who have low semantic values. Table 3 reports the ratio of users with low and high semantic values, respectively. Note that for YouTube dataset, 98% of users have high semantic values, which means almost all users have high semantic values in the dataset (i.e., densely clustered and the values are similarly high). Meanwhile, Fig. 4 shows that *BWR* and *LDD* perform worse for the YouTube dataset than the others. This is because the heuristic strategies of *BWR* and *LDD* can work well when the semantic values are not densely clustered. However, the findings show that our solutions can perform well for all cases. Fig. 6 shows the percentage of activated users with high semantic value under the IC model. The higher the percentage is, the more effective the solution can distinguish the high from those who have low semantic values. It can be observed that *GRIS* based solutions have better results than *BWR* and *LDD*. It means our solution can find seeds that are able to activate more users with high semantic values and thus achieve stronger influence.

5.2.2. Efficiency

The efficiency of getting seeds in real applications is significant. A solution is more efficient if it costs less time picking seeds. We evaluate the efficiency of our solutions by comparing the total running time of picking seeds with the existing. The results are presented in Fig. 7.

In Fig. 7, *LDD* is the fastest benefiting from simple heuristics. *BWR* is slower than *GRIS-SIM*$_{opt}$ for all the datasets. Another observation is that in both Figs. 7 and 8, *GRIS-SIM*$_2$ is consistently faster than *GRIS-SIM*$_1$, which supports the theoretical analysis in Section 4.5. It shows that the optimization toward targeted users is able to improve the efficiency of the *GRIS-SIM*$_1$

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

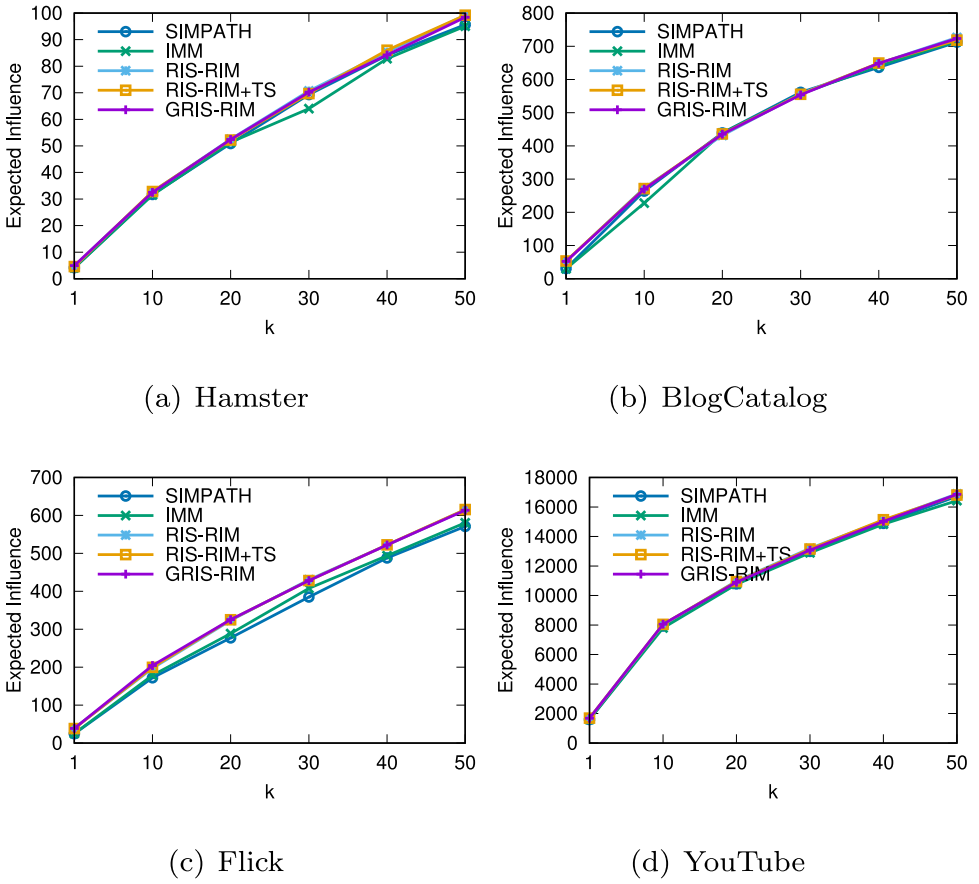


Fig. 5. Expected influence varying k under the LT model ($\varepsilon = 0.5$ and $l = 1$ for *GRIS*).

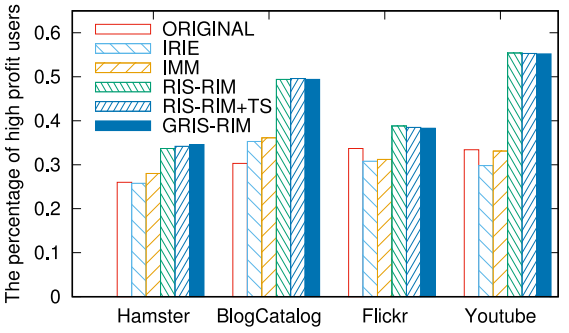


Fig. 6. Percentage of activated users with high semantic value ($\varepsilon = 0.5$ and $l = 1$, $k = 50$).

algorithm. The reason is that *GRIS-SIM*$_2$ ignores non-targeted users in the sampling phase avoiding significant redundant computation in *GRIS-SIM*$_1$.

GRIS-SIM$_{opt}$ generally performs faster than the other *GRIS* based methods. Note that the difference between the three *GRIS* based methods is small for the YouTube dataset. This is because the semantic values of users in this dataset are close, which makes the performance of the sampling strategies ST_1 and ST_2 similar to the optimal sampling strategy ST_{opt}.

The running time of all the *GRIS* methods is proportional to the sampling sizes (i.e. $|\mathcal{R}|$). Fig. 9 shows the sampling sizes when $k = 50$. The results show that *GRIS-SIM*$_{opt}$ has the smallest sampling size, which explains the reason why *GRIS-SIM*$_{opt}$ is the most efficient.

5.2.3. Scalability

Since different kinds and scales of social networks exist in practice, the proposed solutions shall efficiently work in different settings. We conducted experiments to evaluate the scalability of the proposed solutions by observing running

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

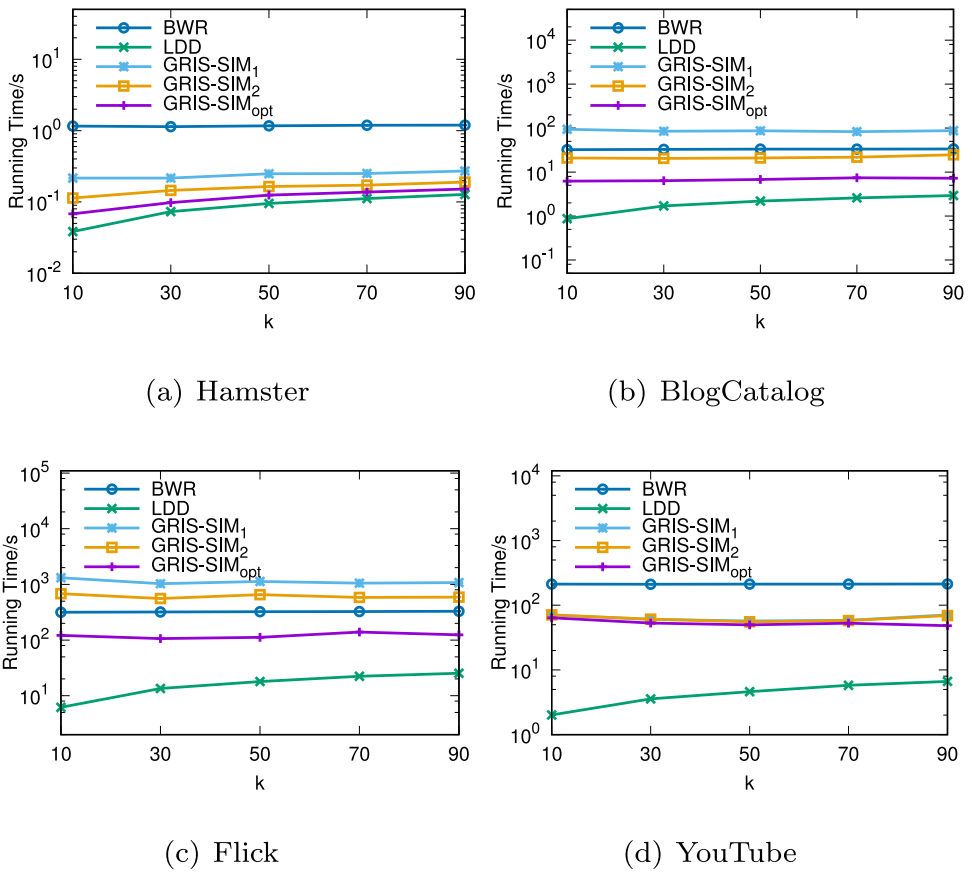


Fig. 7. Running time by varying k under the IC model ($\varepsilon = 0.5$ and $l = 1$ for *GRIS*).

time under different scales of target users, graph sizes, and user types. Note that all the *GRIS-SIM* solutions have similar results in the last two sets of experiments, we only report the results of *GRIS-SIM*$_{opt}$.

First, we vary the proportion of the target users for performance evaluation. In this set of experiments, we use the most extensive dataset, YouTube. The findings in Fig. 10 (a) show that with the decrease of target users the runtime of *GRIS-SIM*$_1$ increases while decreasing for the other two. This is because the sampling phase of *GRIS-SIM*$_2$ and *GRIS-SIM*$_{opt}$ ignore non-targeted users. When the number of target users decreases, they sample fewer RR sets. With more non-targeted users, *GRIS-SIM*$_1$ returns a lower estimator of OPT_k and it results in a larger sampling size according to Lemma 2. *GRIS-SIM*$_{opt}$ outperforms *GRIS-SIM*$_2$ in all the cases because of the lower C as discussed in Theorem 3. Therefore, *GRIS-SIM*$_{opt}$ has good scalability in terms of target users.

Second, we evaluate scalability by graph sizes. We generate five synthetic datasets from 2 to 10 million nodes with the tool [7] to test the scalability. Fig. 10 (b) shows that the running time of *GRIS-SIM*$_{opt}$ grows linearly with the increase of the graph size. The reason is that when graph size increases, the number of the required *RR* sets also increases, which is expensive for computation. The increase agrees with the time complexity provided in Section 4.5.

Third, we vary the total number of user types on YouTube to test the scalability. We increase the number of user types by dividing original types into subtypes. The findings in Fig. 10 (c) shows that when the size of type set increases, the running time of *GRIS-SIM*$_{opt}$ keeps stable, which demonstrates that *GRIS-SIM*$_{opt}$ is scalable with user types.

5.2.4. Parameters sensitivity

In this set of experiments, we study the sensitivity of the parameters of the proposed solutions to see how these parameters affect the effectiveness and efficiency of the proposed solutions. The findings of *GRIS-SIM*$_{opt}$ are shown, which are similar to the other *GRIS* based solutions. We evaluated the effect of all the parameters (ε, ℓ, and k) on the performance of *GRIS-SIM*$_{opt}$.

Fig. 11 (a-b) presents the expected influence and the running time of *GRIS-SIM*$_{opt}$ while varying ε. The parameter ε affects the approximation guarantee (accuracy) and the running time (efficiency) of *GRIS-SIM*$_{opt}$. The results show that the expected influence is insensitive to ε, while the running time increases at ε. The findings indicate that *GRIS-SIM*$_{opt}$ can achieve strong expected influence even without high accuracy, which suggests both effectiveness and efficiency of *GRIS-SIM*$_{opt}$.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

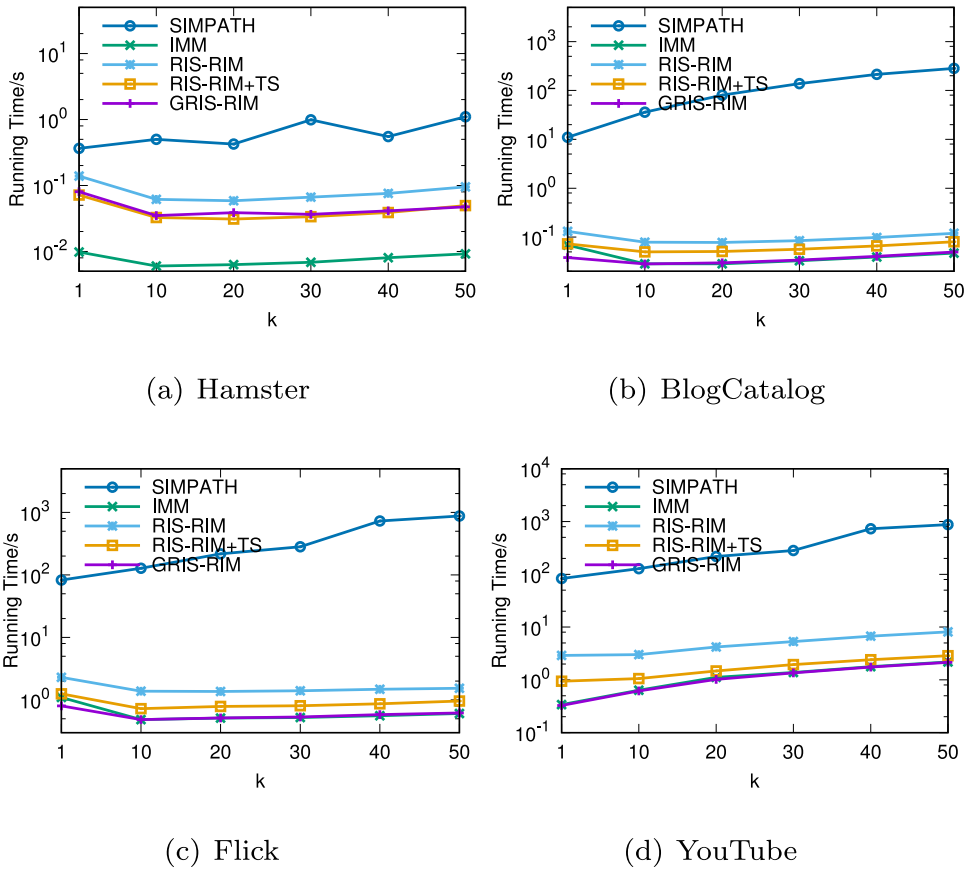


Fig. 8. Running time by varying k under the LT model ($\varepsilon = 0.5$ and $l = 1$ for *GRIS*).

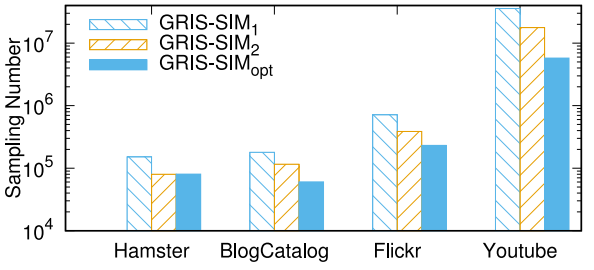


Fig. 9. Sampling size for various datasets ($\varepsilon = 0.5$ and $l = 1$, $k = 50$).

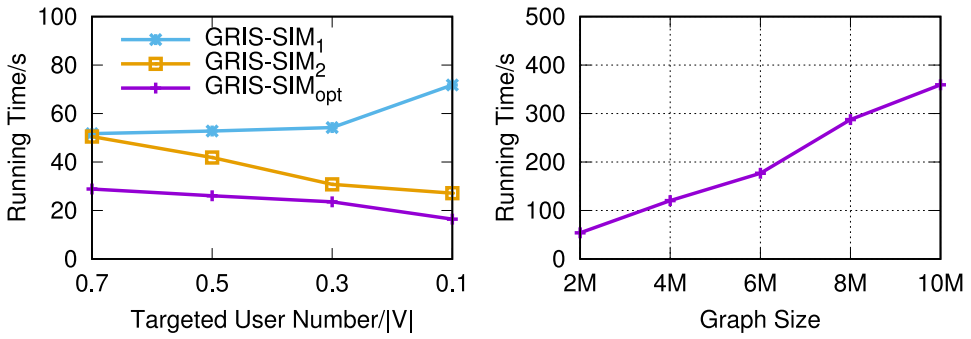
Fig. 11 (c-d) shows the expected influence and the runtime of *GRIS-SIM* while varying ℓ on YouTube. The parameter ℓ controls the probability with which *GRIS-SIM* guarantees its approximation ratio (accuracy) and time complexity. It can be observed that the expected influence of *GRIS-SIM*$_{opt}$ is insensitive to ℓ, and the running time of *GRIS-SIM*$_{opt}$ increases slightly with the growth of ℓ. The results show that the setting of ℓ is not crucial to the performance of *GRIS-SIM*$_{opt}$.

GRIS-SIM$_{opt}$'s sensitivity to k can be observed from Figs. 4 and 7. k corresponds to the budget of real applications. The expected influence of *GRIS-SIM*$_{opt}$ increases with an increase of k. This is because with more seeds users are more likely to be activated. When k increases, on the one hand, the decrease of sampling size reduces the sampling time, and on the other hand, the increase of seeds requires more node selection time. As a result, the running time of *GRIS-SIM* is relatively stable with an increase of k.

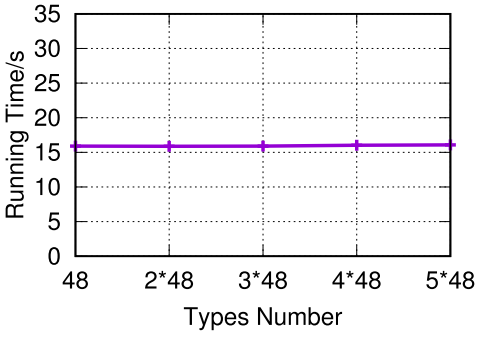
5.3. Result discussion for DAIM

GRIS technique can be applied to other RIS-based methods designed for specific applications to improve their efficiency. This section gives a brief introduction of *RIS-DA* designed for *DAIM* and of how we apply *GRIS* to the problem. Then, we evaluate the performance of the new algorithm (*GRIS-DA*) with the two location-based datasets that are used in the study [38].

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075



(a) The number of targeted users (b) Graph size



(c) The number of user types

Fig. 10. Running time under different settings ($\varepsilon = 0.5$ and $l = 1$ for *GRIS*).

5.3.1. RIS-DA

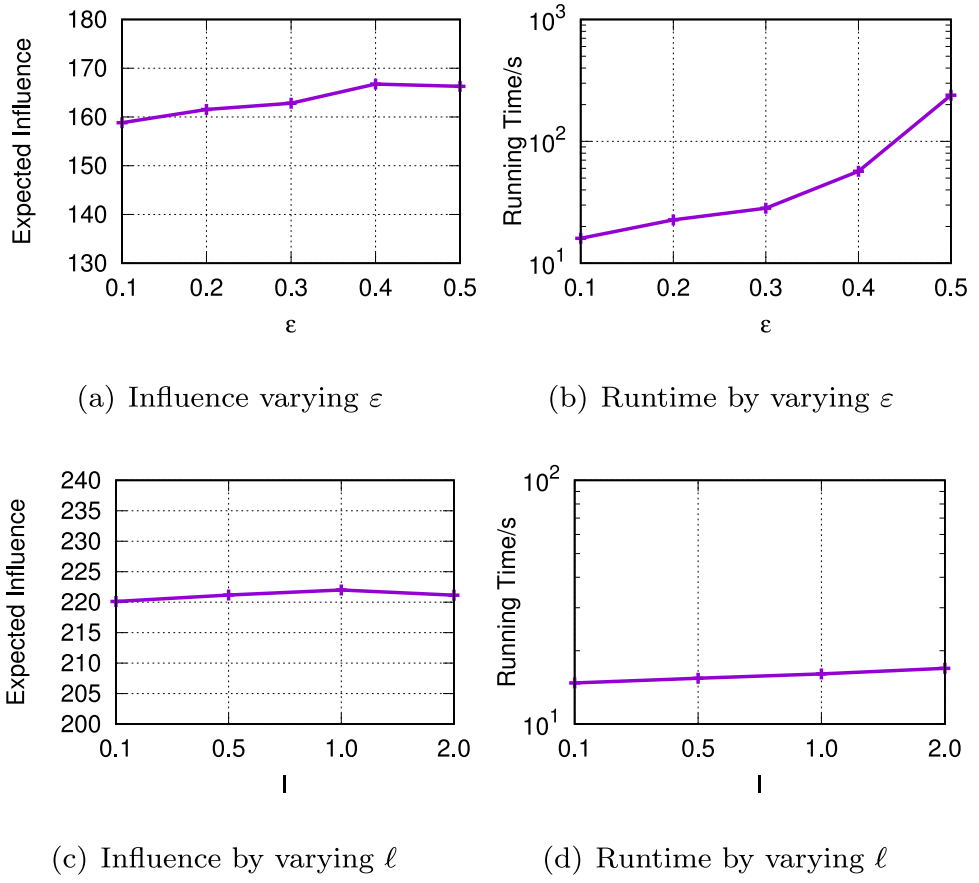
RIS-DA is the advanced method for *DAIM* proposed in the study [38]. *RIS-DA* uses the *RIS* technique to solve *DAIM*. The key point is how to decide the sampling size for each query. *RIS-DA* samples a series of nodes offline as pivots, then it uses pivots as queries and computes the expected influence. Next, *RIS-DA* divides the map into cells based on the distance to pivots and computes the maximum required sampling size for nodes in each cell. For online querying, *RIS-DA* computes the required sampling size of the cell in which the query locates, then it selects nodes in the sampling sets just as normal *RIS* based methods do. To improve *RIS-DA*, we apply *GRIS* to corresponding computation for pivots, i.e., the expected influence and the required sampling size. The following experiments show that *GRIS-DA* uses less sampling sets, which can improve efficiency, meanwhile it maintains the effectiveness.

5.3.2. Effectiveness

We evaluate the effectiveness of *GRIS-DA* against *RIS-DA* on the Brightkite and Gowalla datasets. Similar to Section 5.2.1 we use estimated influence of seeds to measure effectiveness. We use the default settings in the study [38]. The number of pivots is set to 2000. For the parameters of offline computing, $\epsilon_0 = 0.1$ and $\delta_0 = 1/10n$. For the parameters of online querying, $\epsilon_1 = 0.5$ and $\delta_1 = 1/n$. The locations of queries are randomly selected from the entire space. Fig. 12 (a-b) shows the averaged excepted influence spread of *RIS-DA* and *GRIS-DA*. These two lines coincide with each other, which shows two methods have almost the same excepted influence on both datasets. This is because *GRIS-DA* has the same theoretical approximation ratio as *RIS-DA*. The results show that *GRIS-DA* is as effective as *RIS-DA*, even if *GRIS-DA* improves the efficiency of *RIS-DA*.

5.3.3. Efficiency

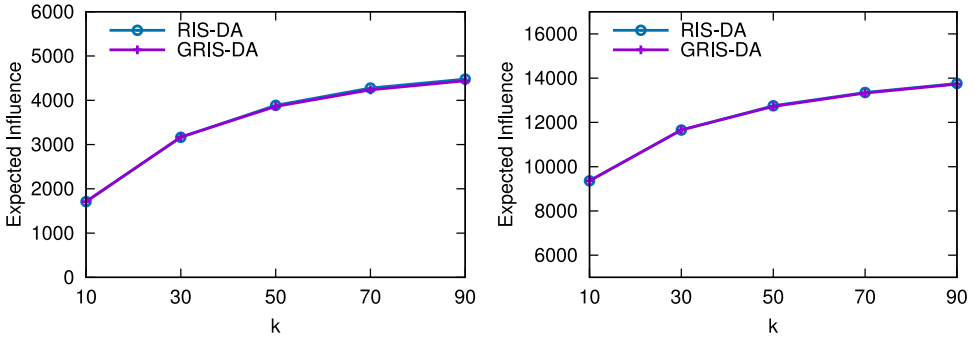
We evaluate the efficiency of *GRIS-DA* and *RIS-DA* by comparing their running time of online processing. Fig. 13 (a-b) reports the average running time of online processing under different seed size. The findings show that *GRIS-DA* is faster than *RIS-DA* by roughly an order of magnitude with various datasets and seed size k. It shows that *GRIS-DA* are more efficient than *RIS-DA*. This is because by applying *GRIS-DA*, the sampling size of each pivot is optimal to maintain the given approximation ratio, which reduces the required sampling size of online processing. Fig. 13 (c-d) reports the averaged sampling



(a) Influence varying ε (b) Runtime by varying ε

(c) Influence by varying ℓ (d) Runtime by varying ℓ

Fig. 11. Expected influence and runtime by varying ε ($l = 1$, $k = 50$) and ℓ ($\varepsilon = 0.5$, $k = 50$) on YouTube under IC model.



(a) Brightkite (b) Gowalla

Fig. 12. Expected influence by varying k under the IC model.

number of both methods. The results show that *GRIS-DA* can dramatically reduce the sampling number at least one order of magnitude, which improves efficiency.

6. Conclusion

This paper introduces the semantics-aware influence maximization problem (*SIM*) in social networks. Compared with the existing studies that mostly focus on networks, *SIM* takes into account both structural and semantic information of social networks, and it is proved to be NP-hard. We propose the *GRIS-SIM* framework to efficiently solve *SIM* problems with $(1 - 1/e - \varepsilon)$ approximation bound for accuracy. A generalized *RIS* technique is proposed to support different sampling

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

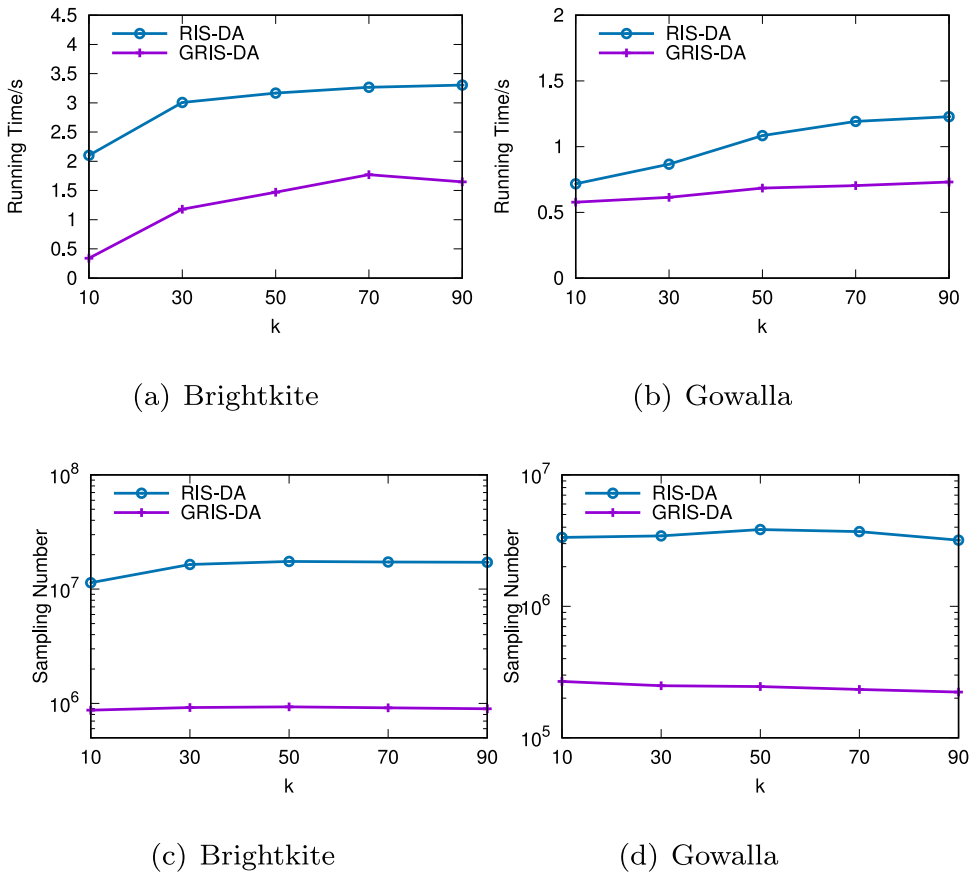


Fig. 13. Running time and online sampling size by varying k under the IC model.

strategies in *GRIS-SIM*. *GRIS* uses two vectors to present the sampling probability and weight of each sampling sets, and provides a flexible way to generate RR sets. Next, we find two valid sampling strategies that can be auto-generated for different settings of semantics unlike the models in the related work that are designed for special scenarios, and we further show the optimal sampling strategy with which the *GRIS-SIM* solution achieves the best time complexity for the same approximation performance. The extensive experiments demonstrate the efficiency, effectiveness, scalability and wide uses of the proposed scheme.

GRIS-SIM still has some limitations that deserve future study. First, *GRIS-SIM* is designed for the setting of static networks. In the future, the study on the support of dynamic networks would be important since a real-life social network is changing. Second, *GRIS-SIM* is able to deal with semantics that can be expressed as a semantics translation function. For the other situations, e.g., constraints of multiple optimization goals, *GRIS-SIM* would be not a good choice. Therefore, other alternatives of semantics would be interesting for investigation. Finally, *GRIS-SIM* is proposed for the IM problem on a triggering model. In the future we would study how to apply *GRIS-SIM* to variants of diffusion models such as models that consider competitors or time delays.

Declaration of Competing Interest

We confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgments

The work was partially supported by the CAS Pioneer Hundred Talents Program with grant number 2017-063 and the National Natural Science Foundation of China with grant number 61902385.

References

[1] S.M.H. Bamakan, I. Nurgaliev, Q. Qu, Opinion leader detection: a methodological review, Expert Syst. Appl. 115 (2019) 200–222.

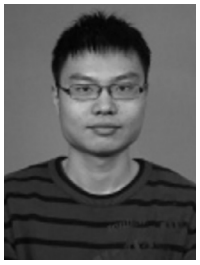
Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

[2] S. Borgatti, The key player problem, 2008.
[3] C. Borgs, M. Brautbar, J. Chayes, B. Lucier, Maximizing social influence in nearly optimal time, in: Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms, SIAM, 2014, pp. 946–957.
[4] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: IEEE ICDM, 2010, pp. 88–97.
[5] F. Chung, L. Lu, Concentration inequalities and martingale inequalities: a survey, Internet. Math. 3 (1) (2006) 79–127.
[6] P. Domingos, M. Richardson, Mining the network value of customers, in: ACM SIGKDD, 2001, pp. 57–66.
[7] A. Hadian, S. Nobari, B. Minaei-Bidgoli, Q. Qu, ROLL: fast in-memory generation of gigantic scale-free networks, in: ACM SIGMOD, 2016, pp. 1829–1842.
[8] S. Jendoubi, A. Martin, L. Liétard, H.B. Hadji, B.B. Yaghlane, Two evidential data based models for influence maximization in twitter, Knowl. Based Syst. 121 (2017) 58–70.
[9] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network (2003) 137–146.
[10] M.M.D. Khomami, A. Rezvanian, N. Bagherpour, M.R. Meybodi, Minimum positive influence dominating set and its application in influence maximization: a learning automata approach, Appl. Intell. 48 (3) (2017) 1–24.
[11] KONECT, Hamsterster friendships network dataset, 2016, (http://konect.uni-koblenz.de/networks/petster-friendships-hamster).
[12] S. Kundu, C.A. Murthy, S.K. Pal, A new centrality measure for influence maximization in social networks, in: Pattern Recognition and Machine Intelligence-international Conference, 2011.
[13] D. Lacker, Mean field games via controlled martingale problems: existence of markovian equilibria, Stoch. Process. Appl. 125 (7) (2015) 2856–2894.
[14] J.-R. Lee, C.-W. Chung, A query approach for influence maximization on specific users in social networks, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 340–353.
[15] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: ACM SIGKDD, 2007, pp. 420–429.
[16] D. Li, C. Wang, S. Zhang, G. Zhou, D. Chu, C. Wu, Positive influence maximization in signed social networks based on simulated annealing, Neurocomputing 260 (2017) 69–78.
[17] F.-H. Li, C.-T. Li, M.-K. Shan, Labeled influence maximization in social networks for target marketing, in: SocialCom, 2011, pp. 560–563.
[18] H. Li, S.S. Bhowmick, A. Sun, J. Cui, Conformity-aware influence maximization in online social networks, VLDB J. 24 (1) (2015) 117–141.
[19] H. Li, L. Pan, P. Wu, Dominated competitive influence maximization with time-critical and time-delayed diffusion in social networks, J. Comput. Sci.
[20] Y. Li, D. Zhang, K. Tan, Real-time targeted influence maximization for online advertisements, PVLDB 8 (10) (2015) 1070–1081.
[21] S.-C. Lin, S.-D. Lin, M.-S. Chen, A learning-based framework to handle multi-round multi-party influence maximization on social networks, in: ACM SIGKDD, 2015, pp. 695–704.
[22] S. Liu, Q. Qu, Dynamic collective routing using crowdsourcing data, Transp. Res. Part B 93 (2016) 450–469.
[23] S. Liu, Q. Qu, S. Wang, Heterogeneous anomaly detection in social diffusion with discriminative feature discovery, Inf. Sci. 439–440 (2018) 1–18.
[24] G. Long, D. Zhang, W. Wei, C. Gao, K.L. Tan, Influence maximization in trajectory databases, IEEE Trans. Knowl. Data Eng. PP (99) (2017). 1–1
[25] M.G. Lozano, J. Schreiber, J. Brynielsson, Tracking geographical locations using a geo-aware topic model for analyzing social media data, Decis. Support Syst. 99 (2017) 18–29.
[26] F. Lu, W. Zhang, L. Shao, X. Jiang, P. Xu, H. Jin, Scalable influence maximization under independent cascade model, J. Network Comput. Appl. 86 (2017) 15–23.
[27] W. Lu, W. Chen, L.V. Lakshmanan, From competition to complementarity: comparative influence diffusion and maximization, Proc. VLDB Endowment 9 (2) (2015) 60–71.
[28] H. Ma, M.R.-T. Lyu, I. King, Mining web graphs for recommendations, IEEE Trans. Knowl. Data Eng. 24 (2011) 1051–1064.
[29] J.S. More, C. Lingam, A si model for social media influencer maximization, Appl. Comput. Inf. (2017).
[30] Q. Qu, S. Liu, C.S. Jensen, F. Zhu, C. Faloutsos, Interestingness-driven diffusion process summarization in dynamic networks, in: PKDD, 2014, pp. 597–613.
[31] Q. Qu, S. Liu, F. Zhu, C.S. Jensen, Efficient online summarization of large-scale dynamic networks, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3231–3245.
[32] S. Raghavan, R. Zhang, Weighted target set selection on social networks, The Robert H. smith school of business and institute for systems research, University of Maryland Maryland, USA, Tech. Rep., 2015.
[33] D.M. Schwartz, T. Rouselle, Using social network analysis to target criminal networks, Trends Org. Crime 12 (2) (2009) 188–207.
[34] Y. Si, F. Zhang, W. Liu, Ctf-ara: an adaptive method for poi recommendation based on check-in and temporal features, Knowl. Based Syst. 128 (2017) 59–70.
[35] D. Silvestre, P. Rosa, J.P. Hespanha, C. Silvestre, Self-triggered and event-triggered set-valued observers, Inf. Sci. 426 (2018) 61–86.
[36] Y. Tang, Y. Shi, X. Xiao, Influence maximization in near-linear time: amartingale approach, in: ACM SIGMOD, 2015, pp. 1539–1554.
[37] V.V. Vazirani, Approximation Algorithms, Springer, 2001.
[38] X. Wang, Y. Zhang, W. Zhang, X. Lin, Efficient distance-aware influence maximization in geo-social networks., IEEE Trans. Knowl. Data Eng. 29 (3) (2017) 599–612.
[39] Y. Wang, H. Wang, J. Li, H. Gao, Efficient influence maximization in weighted independent cascade model, Int. Conf. Database Syst.Adv. Appl. (2016) 49–64.
[40] Y. Yang, Y. Xu, E. Wang, K. Lou, D. Luan, Exploring influence maximization in online and offline double-layer propagation scheme, Inf. Sci. 450 (2018) 182–199.
[41] Y. Yang, Y. Xu, E. Wang, K. Lou, D. Luan, Exploring influence maximization in online and offline double-layer propagation scheme, Inf. Sci. 450 (2018) 182–199.
[42] R. Zafarani, H. Liu, Social computing data repository at ASU, 2009.

Yipeng Chen graduated from Peking University, and he is currently a researcher in Infrastructure Group of Media Intelligent Department at Sohu China. His current research interests include social network analysis, data mining, and personalization recommendation.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075

Qiang Qu is a professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), and the director of Guangdong Provincial R&D Center of Blockchain and Distributed IoT Security. He received his Ph.D. degree from Aarhus University. His current research interests are in data-intensive applications and systems, focusing on efficient and scalable algorithm design, blockchain, data sense-making, and mobility intelligence. His recent research has been published in leading journals and international conferences, including ACM SIGMOD, VLDB, AAAI, the IEEE transactions on Data Engineering, the IEEE Transactions on Intelligent Transportation Systems, and Information Sciences. He was TPC member of several prestige conferences, and he chaired workshops in VLDB 2018, VLDB 2017, ICDM 2015, and APWEB-WAIM 2017 on mobility analysis.



Yuanxiang Ying graduated from Peking University and he is a Software Development Engineer at Microsoft in Beijing. His research interests are in social networks and databases.

Hongyan Li is a professor in the Department of Intelligence Science, School of Electronics Engineering and Computer Science. She obtained her B.Eng. and Ph.D. from Northwestern Polytechnical University in 1993 and 1999 respectively. Her research interests include data management, business process control, data mining and knowledge discovery, with a focus on big data management, large-scale data analysis, state perception of complex systems,and deep learning. She has published more than 100 research papers, an academic monograph, a teaching material and a translation. She has seven patents for the invention of the country. She has served in Database Specialized Committee of China Computer Federation and Cloud Computing & SaaS Specialized Committee of China Institute of Communications.

Jialie Shen is a reader at School of Electronics, Electrical Engineering, and Computer Science, Queen's University Belfast, U.K.. He received his Ph.D. in Computer Science from the University of New South Wales (UNSW), Australia. He worked as a faculty member at UNSW, Sydney and researcher at information retrieval research group, the University of Glasgow for a few years. His main research interests include information retrieval, economic-aware media analysis, and statistical machine learning.

Please cite this article as: Y. Chen, Q. Qu and Y. Ying et al., Semantics-aware influence maximization in social networks, Information Sciences, https://doi.org/10.1016/j.ins.2019.10.075