Theoretical Computer Science 857 (2021) 71–86

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Matching influence maximization in social networks ☆

Guoyao Rao [a], Yongcai Wang [a], Wenping Chen [a], Deying Li [a,*], Weili Wu [b]

[a] *School of Information, Renmin University of China, Beijing, 100872, China*
[b] *Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA*

ARTICLE INFO

Article history:
Received 12 November 2020
Received in revised form 10 December 2020
Accepted 22 December 2020
Available online 29 December 2020

Keywords:
Influence maximization
Online matching
Offline matching

ABSTRACT

Influence maximization (IM) is a widely studied problem in social networks, which aims at finding a seed set with limited size that can maximize the expected number of influenced users. However, existing studies haven't considered the matching relationship, which refers to such scenarios that influenced users seek matched partners among the influenced users, such as time matching with friends to watch movie, or matching for opposite sex in the blind date. In this paper, we investigate different matching scenarios and propose *online-matching* (*offline-matching*), in which the matching and influence propagation are simultaneous (asynchronous). For the matching result, we introduce two matched types '*s-matched*', i.e., $i \to j$ and '*d-matched*', i.e., $i \leftrightarrow j$. Then, we formulate the matching influence maximization (MM) problem to optimize a limited seed set that maximizes the expected number of matched users. We prove that the MM problem is NP-hard and the computation of the matching influence is #P-hard. Next, we analyze the submodularity of the matching influence. To address the problem, we propose efficient methods OPMM (SAMM) to solve the MM in online-matching (offline-matching) with $(1 - 1/e - \epsilon)$-approximation ($\beta(1 - 1/e - \epsilon)$-approximation) guarantee. Experiments on the real-world datasets show our algorithms outperform state of the art algorithms in terms of more accurate matching propagation results.

1. Introduction

The online social network has a profound impact on our daily life. Specially in viral marketing, many business activities are promoted through social network platforms such as Facebook, Twitter, Wechat and Weibo. They usually choose a few customers to experience the product firstly, and then let them spread the positive information by using the effect of social network such as word-of-mouth. Then this creates the problem of influence maximization (IM) [2]. Kemp et al. [3] model the diffusion process as IC and LT, and formulate the basic IM problem at first. Since then, based on the basic diffusion models (IC, LT) and the concept of IM, a lot of studies have been presented. On one hand, some researchers find more efficient algorithms to solve the basic IM, such as the work in [4–10]. On the other hands, researchers extend the basic IM problem to consider more practical factors based on the reality scenes, such as [11–20]. However, to the best of our

☆ This is an enhanced and extended version of a paper [1] presented in The 14th Annual International Conference on Combinatorial Optimization and Applications (COCOA 2020). This work is supported by the National Natural Science Foundation of China (Grant NO. 12071478, 61972404), and partially by NSF 1907472.

* Corresponding author.
E-mail addresses: gyr@ruc.edu.cn (G. Rao), ycw@ruc.edu.cn (Y. Wang), wenpingchen@ruc.edu.cn (W. Chen), deyingli@ruc.edu.cn (D. Li), weiliwu@utdallas.edu (W. Wu).

https://doi.org/10.1016/j.tcs.2020.12.040
0304-3975/

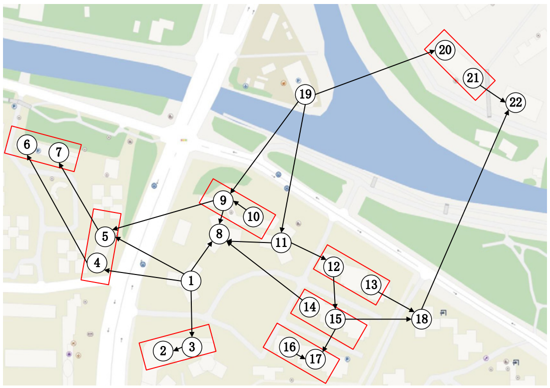


Fig. 1. The example of location-based matching on social network.

knowledge, a commonly seen factor *matching relationship*, i.e., whether influenced people can find any other influenced people matched them as partners, hasn't been investigated in literature.

Matching relationships are commonly seen in social networks and more viral market strategies are paying attention to this factor, specially in the online group-buying markets [21,22] such as Pinduoduo, a famous online shopping company gives discounts to group buyers matched with purchase intention. We also give two following examples to show how necessary matching relationships are in many viral markets.

Example 1.1. When a new movie is propagated through social network, people that are interested commonly will introduce the movie to their friends and go to see the movie with the influenced friends in their common free time, because people usually don't want to see a movie alone. For two interested friends, whether they have common time can be seen as the matching relationship.

Example 1.2. The business often proposes group-buying, in which the members of a group must have similar address to reduce the transportation cost. For two participants, whether they are in common address can be seen as the matching relationship.

With limited budget, business often selects a few users as seeds to propagate their products or activities through the social network. A common strategy of existing studies in IM is to choose seeds that can maximize the diffusion influence. However, considering the matching relationship, this may be a bad strategy. We concrete the Example 1.2 in Fig. 1. The red rectangular box represents a common address and the black arrow represents the definite influence from one user to another through social network. We suppose that the business with the limited budget can only choose one user to propagate the group-buying activity. Node v_{19} is the best choice in tradition IM since it influences the most nodes $\{v_9, v_8, v_5, v_7, v_{11}, v_{12}, v_{15}, v_{17}, v_{18}, v_{20}, v_{22}\}$. But note that none of the 11 people can join the activity as they can't group with any other for none of them has common address. If we choose node v_1 instead, it influences $\{v_4, v_5, v_6, v_7, v_2, v_3\}$. All these six nodes can join the activity as the grouping result is $\{(v_4, v_5), (v_6, v_7), (v_2, v_3)\}$. Obviously v_1 is a better choice.

In this paper, we consider the IM problem with matching relationships. The key contributions are as follows:

1. We firstly investigate the general matching model in social networks. We propose two diffusion-matching processes, i.e., *online-matching* and *offline-matching*. Online-matching requires matching happens only between influence sending person and those influenced by him/her. The matching of two persons follows the online diffusion between them closely. Offline-matching allows matching among all influenced people, two influenced persons can match each other at any time. For the matching result of a node, *s-matched* and *d-matched* are proposed. *s-matched* means being matched unilaterally, while *d-matched* means being matched bilaterally.

2. We then propose the matching influence maximization problem (MM) which is to find k seeds such that the number of people matched on the social network is maximized. We prove that the MM problem is NP-hard and the computation of influence is #P-hard. We also prove that the matching influence is submodular for online-matching but not submodular for offline-matching.

3. For the online-matching problem, we propose efficient methods OPMM^{o,s} for *s-matched* type problem and OPMM^{o,d} for *d-matched* type problem respectively. Both methods lead to approximate solution with $1 - 1/e - \epsilon$ ratio based on the reverse reachable set.

4. For the offline-matching problem, we present efficient methods SAMM^{f,s} for *s-matched* type problem, and SAMM^{f,d} for *d-matched* type problem respectively. Both methods lead to approximate solutions with $\beta(1 - 1/e - \epsilon)$ ratio based on sandwich idea.

5. At last, we model the matching relationship from the real-word labeled datasets. The results of experiments on real-world datasets demonstrate that the proposed methods provide significant improvement, in terms of more accurate

matching influence and less ineffective influence in which the node is influenced but not matched, relative to the traditional influence maximization strategy.

2. Related work

Soical influence problems:

Kemp et al. [3] firstly formulated the basic problem of influence maximization (IM) which is to find k seeds such that the number of influenced nodes is maximized, when the influence spreads from these seeds through social network, based on two basic spread models which they proposed as IC and LT. After that researchers extended the basic IM problem by considering more practical factors based on the reality scenes from different perspectives, such as time-constrained [11,12], topic-aware [13,14], competition [15,16], multi-round [20], rumor-control [17], location-target [18,19], companies-balance [23], group-fairness [24]. Recently, the work [25] proposed the group influence maximization problem in social networks, and some scenes they considered are similar to our paper such as group buying. But in their model, they firstly supposed that there are many definite groups and then they aim to solve how to activate these groups as more as possible, where a group is said to be activated if a certain proportion of nodes in this group is activated. However many times, it's hard to get the group in advance. But in our paper, we can consider the groups are dynamically formed according to the matching relationships instead.

Related algorithms:

There are many works in designing and improving algorithms to solve the IM problem. Kempe [3] proved that the basic IM problem is NP-hard and computing the influence is #P-hard. They also proved the good property of submodularity for the target function. As the NP-hardness, we hope to find efficient algorithms with good approximation-guaranteed. Nemhauser, Wolsey, and Fisher [26] show that the general greedy hill-climbing algorithm (start with the empty set, and repeatedly add an node with the maximum marginal gain for the target function) approximates the optimum to within a factor of $(1-1/e)$, when the target set function $f: 2^V \to R$ satisfied following conditions:

- **Non-negative:** $f \geq 0$;
- **Non-decreasing:** $f(S_1) \geq f(S_2)$, $\forall\ S_2 \subseteq S_1$;
- **Submodular:** $f(S_1 \cap \{v\}) - f(S_1) \leq (S_2 \cap \{v\}) - f(S_2)$, $\forall\ S_2 \subseteq S_1 \subset V$, $v \in V/S_2$;

As the #P-hardness of the target function, it costs too much time using the heavy Monte Carlo simulations to estimate the marginal gain of node's influence. So although the greedy hill-climbing method can guarantee the $(1-1/e-\epsilon)$-approximation where ϵ is the loss of influence estimation, and there are many improvements such as the works in [4,27], and CELF [5], CELF++ [6], but such methods are still inefficient in the large scale network. Tang et al. [8] and Borgs et al. [7] proposed the reverse influence set (RIS) sampling method to estimate the influence. The RIS-based methods are efficient for the large scale network, but there is a key problem that how to sample the RIS sets as less as possible to reduce the time cost and guarantee the $(1-1/e-\epsilon)$-approximation with high confidence. So later, there are many RIS-based extensions and improvements such as the Influence Maximization via Martingales (IMM) [9], Stop-and-Stare (SSA) and Dynamic Stop-and-Stare (D-SSA) [10,28]. Recently, as far as we have known, the Online Processing Information Maximization (OPIM) [29] has superior empirical effectiveness.

3. The social network matching model

3.1. Review of IC model as diffusion process

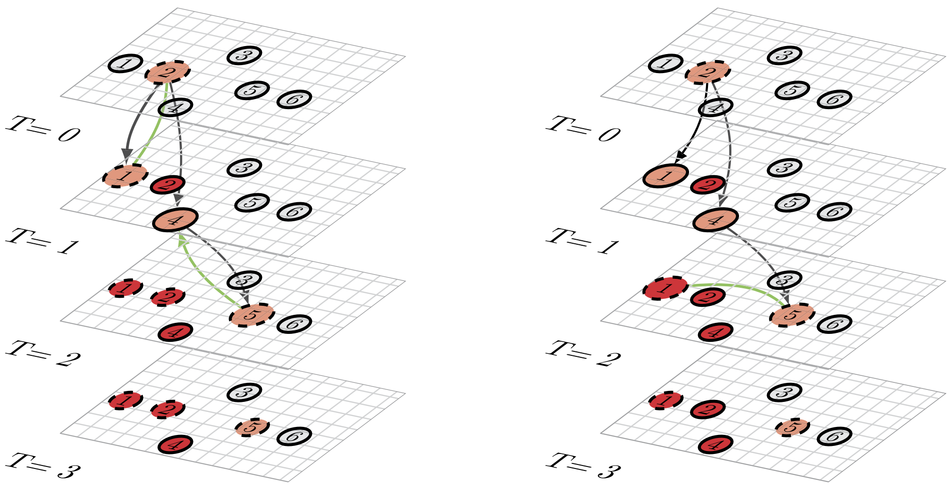
In this paper, we consider the information diffusion process follows the IC model [3]. Let's denote $G^i(V^i, E^i, P^i)$ the influence propagation graph with the influence probability P^i for each direct edge. The influence spreads from a seed set S in rounds. Initially, only the seeds are $active$ and other nodes are $inactive$. In each round, every node that became $active$ in the previous round has one chance to influence its out-neighbors following the influence probability. The process terminates when no $inactive$ nodes can be influenced to become $active$.

3.2. The matching relationship

Matching is common in daily life, such as matching to see the movie, go shopping, travel and have dinner with friends who have common free time, or matching to organize blind date for single people according to the interest and age. Whether the matching is successful or not may often be uncertain in reality, and so without losing generality, we model the uncertain matching relationship between two users in a weighted direct graph defined as follows.

Definition 1. The matching graph $G^m(V^m, E^m, P^m)$ is a direct graph. For each direct edge $\vec{e}_{uv} \in E^m$, the weight $p^m_{uv} \in [0, 1]$ called matching probability represents the probability that v matches u successfully.

(a) The social influence graph G^i with $P^i \equiv 0.5$. (b) The matching graph G^m with $P^m \equiv 0.5$.



(c) An online-matching instance (d) An offline-matching instance

Fig. 2. The examples for the online-matching and offline-matching. In (c) and (d), the dashed-red nodes represent the nodes have been matched and red nodes represent the *active* nodes. The directed-black line such as the line from v_2 to v_1 represents that the v_2 has influenced v_1. The green-dashed-direct line such as the line from v_5 to v_4 in (c) represents that v_5 has matched v_4, and the green-dashed-undirect line such as the line between v_5 to v_1 represents that v_5 and v_1 has matched each other.

3.3. The diffusion-matching process

Note that the premise of matching between two nodes is that they have been influenced to be *active* firstly. Similar to influencing process, we consider there is at most one chance to match each other for two nodes. Considering realistic scenarios, we introduce two different influencing and matching processes (diffusion-matching), i.e., online-matching and offline-matching.

3.3.1. Online-matching

Two nodes u and v in G^m will try to match each other online once v has influenced u or u has influenced v. That is, if there is no online influence process between two *active* nodes, they won't match. We formulate the dynamic process in rounds as follows.

Let S_t be the set of new *active* nodes in tth round. Initialize the seed set as S_0. At tth round, every node $u \in S_{t-1}$ will try to influence its out-neighbor v by the probability p^i_{uv}. If v has been influenced by u, u and v will try to match each other by the probability p^m_{uv} and p^m_{vu} respectively if they haven't matched before.

3.3.2. Offline-matching

Two nodes u and v can match offline as long as both are *active*, although they didn't influence each other directly. During the influence diffusion process, any pair of two *active* node u and v can try to match each other offline, respectively by the probability p^m_{uv} and p^m_{vu} whenever as long as they haven't matched before.

All processes terminate when no more nodes can become *active*, whatever in online-matching or in offline-matching. Fig. 2 illustrates the differences of these two processes from the seed node 2. Fig. 2 (a) gives the influence graph and Fig. 2 (b) is the matching graph.

Fig. 2 (c) shows an instance of online-matching, where the influence and matching are simultaneously. In first round, seed node 2 influenced its out-neighbor nodes 1, 4 and matched with node 1 bilaterally. In second round, node 4 continued to influence its unique neighbor node 5 and matched with it unilaterally. In the third round, the node 5 tried to influence and match its neighbor node 6 but failed and the process terminated.

Fig. 2 (d) shows an instance of offline-matching, in which the influence and matching are asynchronous. In first round, seed node 2 influenced its out-neighbor nodes 1, 4. In second round, node 4 influenced its unique neighbor node 5 but failed to match with it, and then the influenced node 1 got chance and matched with non-neighbor node 5 bilaterally. In the next round, the node 5 tried to influence its neighbor node 6 but got failed and the process terminated.

According to the result of the diffusion-matching process, we define two types for a matched node:

Definition 2. After the diffusion-matching process, in the matching graph G^m, node u is **s-matched** if there is at least one node having matched u, and node u is **d-matched** if there is at least one $active$ node v such that u has matched v and v has matched u.

According to the definition, a *d-matched* node must be also *s-matched*.

4. The matching influence maximization formulation

General influence maximization problem (IM) is to find the best seed set of limited size to maximize the diffusion influence of information. But as we mentioned above, IM is not suitable if the matching relationship is considered. We therefore propose the matching influence maximization problem. Considering that there are two matched types ($s-matched, d-matched$) and two diffusion-matching processes (online-matching, offline-matching), we define the matching influence maximization problem as follows.

Definition 3. Matching Influence Maximization (MM): Given the influence graph G^i, the matching graph G^m, a diffusion-matching process $P \in$ {online-matching, offline-matching}, the matched type of node $M \in$ {d-matched, s-matched} and a set size $k \in Z_+$, let the matching influence $\sigma_{P,M}(S)$ be the expected number of nodes matched as M after the stochastic diffusion-matching process P from the seeds S in G^i and G^m. The MM problem is to find a seed set $S^*_{P,M}$ with size k to maximize the matching influence $\sigma_{P,M}$:

$$S^*_{P,M} := \underset{S\subseteq V^i, |S|=k}{argmax}\ \sigma_{P,M}(S)$$

In order to express convenience, we use symbol 's' ('d') to replace 'M' when we talk *s-matched* (*d-matched*), and use symbol 'o' ('f') to replace 'P' when we talk online-matching (offline-matching), e.g., the mark $\sigma_{o,s}$ represents the matching influence of *s-macthed* in *online-maching*. We denote the expected number of $active$ nodes as the diffusion influence $\eta(S)$ when the influence spreads from a seed set S in G^i. η is nonnegative, monotonically increasing and submodular according to [3]. For the matching influence, we have similar properties easily proved by the definition.

Theorem 1. *The matching influence $\sigma_{P,M}$ is nonnegative and monotonically increasing.*

Note that *d-matched* node must also be *s-matched*, and any matched node must be $active$ first and there are more matching chances for a node in offline-matching than online-matching. Then given the G^i, G^m, S, we have following theorem.

Theorem 2. *The following properties hold: (1) $\sigma_{f,d} \le \sigma_{f,s}$; (2) $\sigma_{o,d} \le \sigma_{o,s}$; (3) $\sigma_{o,s} \le \sigma_{f,s} \le \eta$; (4) $\sigma_{o,d} \le \sigma_{f,d} \le \eta$.*

Next we show the hardness to compute $\sigma_{P,M}$ and the hardness to solve the MM problem.

Theorem 3. *The computation problem of $\sigma_{P,M}$ is #P-hard and the MM problem is NP-hard.*

Proof. Given an arbitrary instance of IM problem in IC, let $G(V, E, P)$ be the influence graph and ϕ be the diffusion influence in G. Then we construct another influence information graph $G^i(V \cup \{a\}, E^i, P^i)$ by adding a node a into V as follows. For each node $v \in V$, we add a direct edge $\vec{e}_{va}$ into E with influence probability 1. We also construct a matching graph $G^m(V \cup \{a\}, E^m, P^m)$. There is no matching edge for any two nodes in V. But for each node v in V, there is a bidirectional[1] edge between a and v in E^m, with the matching probability $P^m \equiv 1$. Considering the MM problem on G^i and G^m, we denote $p_{G,S}(u)$ as the probability of u being $active$ in G with seed set S. For any set $S \subseteq V \cup \{a\}$ with k ($k \in Z_+$) size, node a must be influenced that is $p_{G^i,S}(a) = 1$, because any node in V has the probability 1 to influence a. For any $v \in V$, a can't influence any other node and a can't change the influenced probability of v with the same seed set in G^i and G, then we have $p_{G,S/\{a\}}(v) = p_{G^i,S}(v)$. $\eta(S) = \sum_{v\in V} p_{G^i,S}(v) + p_{G^i,S}(a)$ and $\phi(S/\{a\}) = \sum_{v\in V} p_{G,S/\{a\}}(v)$, so we have

$$\eta(S) = \phi(S/\{a\}) + 1$$

[1] A bidirectional edge denoted as e^d_{uv} between node u and v is that there must be $\vec{e}_{uv}$ and $\vec{e}_{vu}$.

Table 1
Frequently used computation rules on graph.

Notation	Description
$\bar{g}$	the graph which consists of edges reversed in g.
g_S	the subgraph of g which consists of all nodes in the node set S and all paths starting from S.
$\mathcal{V}^s(g)$	the set of nodes which have outgoing edges in g.
$\mathcal{V}^d(g)$	the set of nodes which have bidirectional edges in g.
$\mathcal{N}_u^s(g)$	the set of node u's outgoing neighbors in g.
$\mathcal{N}_u^d(g)$	the set of node u's neighbors in g, which have bidirectional edges with u.
$g_1 \triangleright g_2$	the edge-induced subgraph of g_2, which consists of all such edges in g_2, that the 2 endpoints of each edge are both in g_1.
$g_1 \trianglerighteq g_2$	the edge-induced subgraph of g_2 which consists of all such edges in g_2 that the 2 endpoints of each edge also have an edge in g_1.
$g_1 \sqcup g_2$	the graph which consists of the all edges in g_1 and g_2.
$g_1 \sqsubseteq g_2$	g_1 is a subgraph of g_2.
$I_u(g)$	the subgraph which consists of the ingoing edges of node u in g.
$O_u(g)$	the subgraph which consists of the outgoing edges of node u in g.

In fact, if $S \neq \{a\}$, any *active* node $v \in V$ must match with a because $p^m_{ua} = 1$, $p^m_{au} = 1$ and $p^i_{ua} = 1$ whenever online-matching or offline-matching, that is, any *active* node in G^i must be *d-matched* and *s-matched* in G^m, so we have $\sigma_{o,s}(S) = \sigma_{o,d}(S) = \sigma_{f,s}(S) = \sigma_{f,d}(S) = \eta(S)$ and then

$$\sigma_{P,M}(S) = \phi(S/\{a\}) + 1$$

Since ϕ is #P-hard [30], so $\sigma_{P,M}(S)$ is #P-hard.
In the following, we prove that the MM problem is NP-hard. Suppose we can get the optimum solution S^*_{MM} for the MM problem within polynomial time, we construct a k size set $S^*_{IM} \subseteq V$ as follows.

$$S^*_{IM} = \begin{cases} S^*_{MM} & \text{if } a \notin S^*_{MM} \\ S^*_{MM}/\{a\} \cup \{v\}, \forall v \in V/S^*_{MM} & \text{if } a \in S^*_{MM} \end{cases}$$

for any k size set $S \subseteq V$, we have $\phi(S^*_{IM}) \geq \phi(S^*_{MM}/\{a\}) \geq \phi(S/\{a\}) = \phi(S)$, so we get the optimum solution S^*_{IM} in the IM problem. As IM is NP-hard [3], so is MM. Then this theorem is proved. $\square$

Next we analyze the submodularity of the matching influence through an equivalent generating model.

4.1. The equivalent generating model

To formulate better, we give notations as shown in Table 1, in which g, g_1, g_2 are direct graphs. We will give an equivalent view for the diffusion-matching process.

We may regard that node u attempts to influence its neighbor v as flipping a coin with the probability p^i_{uv}, and that v matches u as flipping a coin with probability p^m_{uv}. All the flipping is independent for different edges. So, before the matching process begins, we can determine the influence and matching relationship through reserving all edges as flipping a coin independently in G^i, G^m, that is, we get a definite influence graph g^i and a definite matching graph g^m in which each edge represents the definite influence or matching. We denote the probability distribution of g^i, g^m as $g^i, g^m \sim G^i, G^m$. Given g^i, g^m and a seed set S, we show the instance generated in different kinds of diffusion-matching processes as follows.

(1) Online-matching: We can get the definite online-matching with the result $g^i_S \trianglerighteq g^m$ in which edge $\vec{e}_{uv}$ represents that u has matched v successfully. That u is *s-matched* is equivalent to that there exists a node v satisfying following two conditions: (1) $\vec{e}_{uv} \in g^i_S$ or $\vec{e}_{vu} \in g^i_S$; (2) $\vec{e}_{uv} \in g^m$. That is, all the nodes in $\mathcal{V}^s(g^i_S \trianglerighteq g^m)$ are the *s-matched* nodes. u is *d-matched* is equivalent to that there exists a node v satisfying following two conditions: (1) $\vec{e}_{uv} \in g^i_S$ or $\vec{e}_{vu} \in g^i_S$; (2) $\vec{e}_{uv} \in g^m$ and $\vec{e}_{vu} \in g^m$. So the set $\mathcal{V}^d(g^i_S \trianglerighteq g^m)$ is the set of *d-matched* nodes.

Computing the expectation number of the all instances weighted with the probabilities, the matching influence in online-matching is as follows.

$$\sigma_{o,s}(S) = \sum_{g^i, g^m \sim G^i, G^m} Pr(g^i, g^m)|\mathcal{V}^s(g^i_S \trianglerighteq g^m)| \tag{1}$$

$$\sigma_{o,d}(S) = \sum_{g^i, g^m \sim G^i, G^m} Pr(g^i, g^m)|\mathcal{V}^d(g^i_S \trianglerighteq g^m)| \tag{2}$$

(2) Offline-matching: We can also get the definite offline-matching with the result as $g^i_S \triangleright g^m$. That v has matched u is equivalent to that u and v satisfy following two conditions: (1) $u \in g^i_S$ and $v \in g^i_S$; (2) $\vec{e}_{uv} \in g^m$. So the set $\mathcal{V}^s(g^i_S \triangleright g^m)$

is the set of *s-matched* nodes. And u and v has matched each other is equivalent to that u and v satisfy following two conditions: (1) $u \in g^i_S$ and $v \in g^i_S$; (2) $\vec{e}_{uv} \in g^m$ and $\vec{e}_{uv} \in g^m$. So the set $\mathcal{V}^d(g^i_S \triangleright g^m)$ is the set of *d-matched* nodes.

Computing the expectation number of the all instances weighted with the probabilities, the matching influence in offline-matching is as follows.

$$\sigma_{f,s}(S) = \sum_{g^i, g^m \sim G^i, G^m} Pr(g^i, g^m) |\mathcal{V}^s(g^i_S \triangleright g^m)| \tag{3}$$

$$\sigma_{f,d}(S) = \sum_{g^i, g^m \sim G^i, G^m} Pr(g^i, g^m) |\mathcal{V}^d(g^i_S \triangleright g^m)| \tag{4}$$

4.2. The submodularity for the matching influence

To analyze the submodularity of the matching influence, we first give the following lemma.

Lemma 1. *We have the following properties:*

1. $\mathcal{V}^s(g_1) \subseteq \mathcal{V}^s(g_2)$ *and* $\mathcal{V}^d(g_1) \subseteq \mathcal{V}^d(g_2)$ *if* $g_1 \sqsubseteq g_2$.
2. $g_{S_1 \cup S_2} = g_{S_1} \sqcup g_{S_2}$.
3. $g_1 \trianglerighteq g_3 \sqsubset g_2 \trianglerighteq g_3$ *if* $g_1 \sqsubseteq g_2$.
4. $(g_1 \sqcup g_2) \trianglerighteq g_3 = (g_1 \trianglerighteq g_3) \sqcup (g_2 \trianglerighteq g_3)$
5. $\mathcal{V}^s(g_1 \sqcup g_2)/\mathcal{V}^s(g_2) = \mathcal{V}^s(g_1)/\mathcal{V}^s(g_2)$.
6. $(g_1 \trianglerighteq g_3) \trianglerighteq (g_2 \trianglerighteq g_3) = (g_2 \trianglerighteq g_3) \trianglerighteq (g_1 \trianglerighteq g_3)$
7. $\mathcal{V}^d(g_1 \sqcup g_2)/\mathcal{V}^d(g_2) = \mathcal{V}^d(g_1)/\mathcal{V}^d(g_2)$, *if* $g_1 \trianglerighteq g_2 = g_2 \trianglerighteq g_1$

Proof. It's easy to have properties 1, 2, 3, 4, and we prove property 5 as follows. $\mathcal{V}^s(g_1)/\mathcal{V}^s(g_2) \subseteq \mathcal{V}^s(g_1 \sqcup g_2)/\mathcal{V}^s(g_2)$ is obvious according to (3). We need to prove $\mathcal{V}^s(g_1 \sqcup g_2)/\mathcal{V}^s(g_2) \subseteq \mathcal{V}^s(g_1)/\mathcal{V}^s(g_2)$. For any node $u \in \mathcal{V}^s(g_1 \sqcup g_2)/\mathcal{V}^s(g_2)$, u has an outgoing edge $\vec{e}_{uv}$ in $g_1 \sqcup g_2$ but no any outgoing edge in g_2, and then $\vec{e}_{uv}$ must be in g_1. So $u \in \mathcal{V}^s(g_1)/\mathcal{V}^s(g_2)$. We have proved 5.

We prove property 6 as follows. For any edge $\vec{e}_{uv}$ in $(g_1 \trianglerighteq g_3) \trianglerighteq (g_2 \trianglerighteq g_3)$, we can infer that: (1) there must be an edge $\vec{e}_{uv}$ or $\vec{e}_{vu}$ in $g_1 \trianglerighteq g_3$, then there must be an edge $\vec{e}_{uv}$ or $\vec{e}_{vu}$ in g_1 and g_3; (2) $\vec{e}_{uv} \in g_2 \trianglerighteq g_3$, so there must be an edge $\vec{e}_{uv}$ or $\vec{e}_{vu}$ in g_2 and $\vec{e}_{uv} \in g_3$; $\vec{e}_{uv} \in g_1 \trianglerighteq g_3$, and there must be an edge $\vec{e}_{uv}$ or $\vec{e}_{vu}$ in $g_2 \trianglerighteq g_3$, so $\vec{e}_{uv} \in (g_2 \trianglerighteq g_3) \trianglerighteq (g_1 \trianglerighteq g_3)$. Hence $(g_1 \trianglerighteq g_3) \trianglerighteq (g_2 \trianglerighteq g_3) \sqsubseteq (g_2 \trianglerighteq g_3) \trianglerighteq (g_1 \trianglerighteq g_3$. It's similar to prove $(g_2 \trianglerighteq g_3) \trianglerighteq (g_1 \trianglerighteq g_3) \sqsubseteq (g_1 \trianglerighteq g_3) \trianglerighteq (g_2 \trianglerighteq g_3)$. We have proved 6.

We prove property 7 as follows. It's obvious that $\mathcal{V}^d(g_1)/\mathcal{V}^d(g_2) \subseteq \mathcal{V}^d(g_1 \sqcup g_2)/\mathcal{V}^d(g_2)$. We just need to prove that $\mathcal{V}^d(g_1 \sqcup g_2)/\mathcal{V}^d(g_2) \subseteq \mathcal{V}^d(g_1)/\mathcal{V}^d(g_2)$. For $\forall u \in \mathcal{V}^d(g_1 \sqcup g_2)/\mathcal{V}^d(g_2)$, u has bidirectional edge in $g_1 \sqcup g_2$ but no bidirectional edge in g_2. For any bidirectional edge with an endpoint u in $g_1 \sqcup g_2$, both direct edges of this bidirectional edge must be in g_1, and otherwise if there exists one in g_2, it will conflict with $g_1 \trianglerighteq g_2 = g_2 \trianglerighteq g_1$, so u must has bidirectional edge in g_1 and $u \in \mathcal{V}^d(g_1)/\mathcal{V}^d(g_2)$. We have proved 7. □

Theorem 4. *$\sigma_{o,M}$ is submodular in online-matching.*

Proof. Let $\Delta_v \sigma_{o,M}(S)$ be the gain after adding v into the seed set S. Suppose seed set $S_1 \subset S_2$, and according to Equation (1), (2), Lemma 1, we have

$$\begin{aligned}
\Delta_v \sigma_{o,M}(S_1) &= \sum Pr(g^i, g^m)(|\mathcal{V}^M(g^i_{S_1 \cup \{v\}} \trianglerighteq g^m)| - |\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)|) \\
&= \sum Pr(g^i, g^m)|\mathcal{V}^M(g^i_{S_1 \cup \{v\}} \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)| \\
&= \sum Pr(g^i, g^m)|(\mathcal{V}^M((g^i_{S_1} \sqcup g^i_{\{v\}}) \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)| \\
&= \sum Pr(g^i, g^m)|\mathcal{V}^M((g^i_{S_1} \trianglerighteq g^m) \sqcup (g^i_{\{v\}} \trianglerighteq g^m)) \\
&\quad /\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)| \\
&= \sum Pr(g^i, g^m)|\mathcal{V}^M(g^i_{\{v\}} \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)|
\end{aligned}$$

and we also have

$$\Delta_v \sigma_{o,M}(S_2) = \sum Pr(g^i, g^m)|\mathcal{V}^M(g^i_{\{v\}} \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_2} \trianglerighteq g^m)|.$$

As $\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m) \subseteq \mathcal{V}^M(g^i_{S_2} \trianglerighteq g^m)$, so we have

$$|\mathcal{V}^M(g^i_{\{v\}} \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_2} \trianglerighteq g^m)| \le |\mathcal{V}^M(g^i_{\{v\}} \trianglerighteq g^m)/\mathcal{V}^M(g^i_{S_1} \trianglerighteq g^m)|.$$

Hence $\Delta_v \sigma_{o,M}(S_2) \le \Delta_v \sigma_{o,M}(S_1)$. The theorem is proved. □

For the offline-matching, we give an example with the graphs in Fig. 2 (a), 2 (b). We have $\Delta_{v_5}\sigma_{f,s}(\{v_3, v_4\}) = 0.75$, $\Delta_{v_5}\sigma_{f,s}(\{v_3\}) = 0$, $\Delta_{v_5}\sigma_{f,d}(\{v_3, v_4\}) = 0.375$ and $\Delta_{v_5}\sigma_{f,d}(\{v_3\}) = 0$. $\sigma_{f,s}$ isn't submodular because $\Delta_{v_5}\sigma_{f,s}(\{v_3, v_4\}) > \Delta_{v_5}\sigma_{f,s}(\{v_3\})$, and $\sigma_{f,d}$ also isn't submodular because $\Delta_{v_5}\sigma_{f,d}(\{v_3, v_4\}) > \Delta_{v_5}\sigma_{f,d}(\{v_3\})$. So we have the theorem for offline-matching as follows.

Theorem 5. *$\sigma_{f,s}$ and $\sigma_{f,d}$ can't be guaranteed to be submodular in offline-matching.*

5. The algorithms for MM problem

Note that the computation problems of diffusion influence and matching influence both are #P-hard. The General greedy method uses the heavy simulations of Monte Carlo to estimate the influence. To estimate the diffusion influence more efficiently in IM, RIS [7] generates the Reverse Reachable sets (RRS) through searching reversely and randomly in the influence graph G^i from a random node. Then RIS provides the solution of IM with good approximation guarantee by selecting k nodes greedily that cover[2] the most RRS sets. In this section, we will solve MM with similar idea of RRS.

5.1. The reverse reachable set in MM

We denote $\mathbb{I}$ as the indicator function for an expression denoted *exp*. If the *exp* is true, $\mathbb{I}\{exp\} = 1$, otherwise $\mathbb{I}\{exp\} = 0$. Denote the independent joint distribution as $u, g^i, g^m \sim V^m_M, G^i, G^m$, in which u is selected randomly from the node set $V^m_M := \mathcal{V}^M(G^m)$, and g^i, g^m is induced respectively by reserving edge from G^i with probability P^i and G^m with probability P^m.

To get the RRS in MM problem, we firstly analyze the seed set S which leads u to be matched for the definite u, g^i, g^m as follows.

In online-matching, that u being matched as type M is equivalent to $u \in \mathcal{V}^M(g^i_S \trianglerighteq g^m)$. At the same time, S must contain one node s which satisfies at least one of following conditions: (1) u can reach s in $\bar{g}^i$ when u has an out-neighbor v in g^i such that $e^M_{uv} \in g^m$; (2) s can be reached from u's out-neighbor v in $\bar{g}^i$ such that $e^M_{uv} \in g^m$. Then we denote the set of nodes like s which satisfy (1) or (2) as $R^{o,s}_{ug^ig^m}$ for *s-matched* and $R^{o,d}_{ug^ig^m}$ for *d-matched*, and we have

$$R^{o,M}_{ug^ig^m} = \begin{cases} \mathcal{V}(\bar{g}^i_{\{u\}}) & \text{if } u \in \mathcal{V}^M(O_u(g^i) \trianglerighteq g^m) \\ \mathcal{V}(\bar{g}^i_{\mathcal{N}^M_u(I_u(g^i) \trianglerighteq g^m)}) & \text{otherwise} \end{cases} \tag{5}$$

In offline-matching, that u being matched as M is equivalent to $u \in \mathcal{V}^M(g^i_S \triangleright g^m)$. At the same time, S must contain two kinds of nodes as follows: (1) the node can be reached from u in $\bar{g}^i$; (2) the node can be reached from any other node v in $\bar{g}^i$ such that $e^M_{uv} \in g^m$. We construct a set pair

$$PR^M_{ug^ig^m} = \{\mathcal{V}(\bar{g}^i_{\{u\}}), \mathcal{V}(\bar{g}^i_{\mathcal{N}^M_u(g^m)})\} \tag{6}$$

where the first set is the set of all nodes as type (1) and the second is the set of all nodes as type (2).

Then we propose two following definitions.

Definition 4. A random reverse reachable set $\mathbb{RRS}_{o,M}$ in online-matching with matched type M is a set as $R^{o,M}_{ug^ig^m}$ where u, g^i, g^m are sampled from the independent joint distribution $(u, g^i, g^m) \sim (V^m_M, G^i, G^m)$.

Definition 5. A random reverse reachable set pair $\mathbb{PRRS}_{f,M}$ in offline-matching with matched type M is a set pair as $PR^M_{ug^ig^m}$ where u, g^i, g^m are sampled from the independent joint distribution $(u, g^i, g^m) \sim (V^m_M, G^i, G^m)$.

Next we prove that we can estimate the matching influence through the random set $\mathbb{RRS}_{o,M}$ and random set pair $\mathbb{PRRS}_{f,M}$ as shown in Theorem 6, 7. Before proving the theorems, we introduce two following lemmas.

Lemma 2. $\mathbb{I}\{u \in \mathcal{V}^M(g^i_S \trianglerighteq g^m)\} = \mathbb{I}\{S \cap R^{o,M}_{ug^ig^m} \neq \emptyset\}$.

[2] We say that a node v covers a set V is that $v \in V$, and specially we say that a node v covers a set pair of V and U is that $v \in V$ and $v \in U$.

Proof. When $u \in \mathcal{V}^M(g_S^i \trianglerighteq g^m)$, u must have such neighbor v in g_S^i that $v \in \mathcal{N}_u^M(g^m)$, so both $\mathcal{V}(\bar{g}^i_{\{v\}}) \cap S \neq \emptyset$ and $\mathcal{V}(\bar{g}^i_{\{u\}}) \cap S \neq \emptyset$. In fact, if $\vec{e}_{uv} \in g_S^i$, $R^{o,M}_{ug^ig^m} = \mathcal{V}^M(\bar{g}^i_{\{u\}})$ as $u \in \mathcal{V}^M(O_u(g^i) \trianglerighteq g^m)$. Or if $\vec{e}_{vu} \in g_S^i$, we have $\vec{e}_{uv} \in \bar{g}^i$, then $\mathcal{V}(\bar{g}^i_{\{v\}}) \subseteq R^{o,M}_{ug^ig^m}$ as $v \in \mathcal{N}_u^M(I_u(g^i) \trianglerighteq g^m)$ and $v \in \bar{g}^i_{\{u\}}$. So $S \cap R^{o,M}_{ug^ig^m} \neq \emptyset$.

When $u \notin \mathcal{V}^M(g_S^i \trianglerighteq g^m)$, u has no such neighbor v in g_S^i that $v \in \mathcal{N}_u^M(g^m)$, then $R^{o,M}_{ug^ig^m} = \emptyset$ as we have that $u \notin \mathcal{V}^M(O_u(g^i) \trianglerighteq g^m)$ and $\mathcal{N}_u^M(I_u(g^i)) \trianglerighteq g^m) = \emptyset$. So we have $S \cap R^{o,M}_{ug^ig^m} = \emptyset$. $\square$

Lemma 3. $\mathbb{I}\{u \in \mathcal{V}^M(g_S^i \triangleright g^m)\} = \mathbb{I}\{S \cap \mathcal{V}(\bar{g}^i_{\{u\}}) \neq \emptyset\}\mathbb{I}\{S \cap \mathcal{V}(\bar{g}^i_{\mathcal{N}_u^M(g^m)}) \neq \emptyset\}$.

Proof. When $u \in \mathcal{V}^M(g_S^i \triangleright g^m)$, there exists a node v both in g_S^i and $\mathcal{N}_u^M(g^m)$. $S \cap \mathcal{V}(\bar{g}^i_{\{u\}}) \neq \emptyset$ as u in g_S^i. $S \cap \mathcal{V}(\bar{g}^i_{\{v\}}) \neq \emptyset$ as v in g_S^i. $S \cap \mathcal{V}(\bar{g}^i_{\mathcal{N}_u^M(g^m)}) \neq \emptyset$ as $\mathcal{V}(\bar{g}^i_{\{v\}}) \subset \mathcal{V}(\bar{g}^i_{\mathcal{N}_u^M(g^m)})$.

When $S \cap \mathcal{V}(\bar{g}^i_{\{u\}}) \neq \emptyset$ and $S \cap \mathcal{V}(\bar{g}^i_{\mathcal{N}_u^M(\bar{g^m})}) \neq \emptyset$. So $u \in \mathcal{V}(g_S^i)$ and $\mathcal{N}_u^M(g^m) \cap \mathcal{V}(g_S^i) \neq \emptyset$, then $u \in \mathcal{V}^M(g_S^i \triangleright g^m)$. $\square$

Theorem 6. $\sigma_{o,M}(S) = |V_M^m| Pr\{S \cap \mathbb{RRS}_{o,M} \neq \emptyset\}$.

Proof. According to the Equation (1), (2) and Lemma 2, we have

$$\begin{aligned}
\sigma_{o,M}(S) &= \sum_{g^i,g^m \sim G^i,G^m} Pr(g^i,g^m)|\mathcal{V}^M(g_S^i \trianglerighteq g^m)| \\
&= \sum_{g^i,g^m \sim G^i,G^m} Pr(g^i,g^m) \sum_{u \in V_M^m} \mathbb{I}\{u \in \mathcal{V}^M(g_S^i \trianglerighteq g^m)\} \\
&= \sum_{u \in V_M^m} \sum_{g^i,g^m \sim G^i,G^m} Pr(g^i,g^m)\mathbb{I}\{u \in \mathcal{V}^M(g_S^i \trianglerighteq g^m)\} \\
&= |V_M^m| \sum_{u \in V_M^m} \sum_{g^i,g^m \sim G^i,G^m} \frac{1}{|V_M^m|} Pr(g^i,g^m)\mathbb{I}\{u \in \mathcal{V}^M(g_S^i \trianglerighteq g^m)\} \\
&= |V_M^m| \sum_{u,g^i,g^m \sim V_M^m,G^i,G^m} Pr(u,g^i,g^m)\mathbb{I}\{u \in \mathcal{V}^M(g_S^i \trianglerighteq g^m)\} \\
&= |V_M^m| \sum_{u,g^i,g^m \sim V_M^m,G^i,G^m} Pr(u,g^i,g^m)\mathbb{I}\{S \cap R^{o,M}_{ug^ig^m} \neq \emptyset\} \\
&= |V_M^m| Pr\{S \cap \mathbb{RRS}_{o,M} \neq \emptyset\} \quad \square
\end{aligned}$$

Theorem 7. $\sigma_{f,M}(S) = |V_M^m| Pr\{S \cap \mathbb{PRRS}_{f,M} \neq \emptyset\}$.[3]

Proof. According to Equation (3), (4) and Lemma 3, similarly to the proof in Theorem 6, we have

$$\begin{aligned}
\sigma_{f,M}(S) &= \sum_{g^i,g^m \sim G^i,G^m} Pr(g^i,g^m)|\mathcal{V}^M(g_S^i \triangleright g^m)| \\
&= |V_M^m| \sum_{u,g^i,g^m \sim V_M^m,G^i,G^m} Pr(v,g^i,g^m)\mathbb{I}\{u \in \mathcal{V}^M(g_S^i \triangleright g^m)\} \\
&= |V_M^m| \sum_{u,g^i,g^m \sim V_M^m,G^i,G^m} Pr(u,g^i,g^m)\mathbb{I}\{S \cap \mathcal{V}(\bar{g}^i_{\{u\}}) \neq \emptyset\}\mathbb{I}\{S \cap \mathcal{V}(\bar{g}^i_{\mathcal{N}_u^M(g^m)}) \neq \emptyset\} \\
&= |V_M^m| Pr\{S \cap \mathbb{PRRS}_{f,M} \neq \emptyset\} \quad \square
\end{aligned}$$

To construct the sample of $\mathbb{RRS}_{o,M}$ and $\mathbb{PRRS}_{f,M}$, we first need to get definite g^i, g^m by reserving all edges in G^i, G^m with the edge probability. Then construct the set $R^{o,M}_{ug^ig^m}$ or set pair $PR^M_{ug^ig^m}$ as the Equation (5), (6). We will check and

[3] Note that given a set pair $A = \{A_1, A_2\}$ and a set B, $A \cap B \neq \emptyset$ is that $A_1 \cap B \neq \emptyset$ and $A_2 \cap B \neq \emptyset$.

search some local edges associated with u in g^i and g^m and get nodes which will be used as origin nodes for graph reverse searching in g^i. It may not search and check all of the edges in G^i and G^m, so we just randomly reserve the edge we may check and search during the construction to reduce the sampling cost. Based on such idea of random breadth-first searching (BFS) on reverse G^i and G^m, we propose the time efficiency sampling algorithm S-RRS^{o,M} for $\mathbb{RRS}_{o,M}$ and S-RRSP^{f,M} for $\mathbb{PRRS}_{f,M}$, the details of which are shown in Algorithm 1 and 2 respectively.

Algorithm 1: S-RRS^{o,M} (G^i, G^m).

Create an empty FIFO queue Q and empty sets C, R;
Choose a node u randomly from V^m_M;
for *each edge* $e^M_{uv} \in G^m$ **do**
  if *reserve* e^M_{uv} *in* G^m *by* $p^{m,M}_{uv}$ *successfully.* **then**
    $C \leftarrow C \cup \{v\}$;
for *each node* v *in* C *which has* $\vec{e}_{uv}$ *in* G^i **do**
  if *Reserve* $\vec{e}_{uv}$ *in* G^i *by* p^i_{uv} *successfully* **then**
    $Q.enqueue(u)$;
    goto *reverse search in* G^i
for *each node* v *in* C *which has* $\vec{e}_{vu}$ *in* G^i **do**
  if *Reserve* $\vec{e}_{vu}$ *in* G^i *by* p^i_{vu} *successfully .* **then**
    $Q.enqueue(u)$;
reverse search in G^i:
while Q *is not empty* **do**
  $v \leftarrow Q.dequeue()$;
  $R \leftarrow R \cup \{v\}$;
  for $w \in \mathcal{N}^s_v(\bar{G}^i)$ *and* w *is not visited* **do**
    if *Reserve* $\vec{e}_{vw}$ *in* G^i *by* p^i_{vw} *successfully* **then**
      $Q.equeue(w)$ and mark w visited;
return *R as a sample of the* $\mathbb{RRS}_{o,M}$

Algorithm 2: S-RRSP^{f,M} (G^i, G^m).

Create an empty FIFO queue Q and empty sets N, R_1, R_2;
Choose a node u randomly from V^m_M ; /* Remark e^s_{uv} as $\vec{e}_{uv}$, $p^{m,s}_{uv}$ as p^m_{uv}, and $p^{m,d}_{uv}$ as $p^m_{uv} \cdot p^m_{vu}$. */
for *each edge* e^M_{uv} *in* G^m **do**
  if *Reserve* e^M_{uv} *in* G^m *by* $p^{m,M}_{uv}$ *successfully .* **then**
    $N \leftarrow N \cup \{v\}$;
if N *is not empty* **then**
  $R_1 \leftarrow R_1 \cup \{u\}$;
  $Q.enqueue(N \cup \{u\})$;
  while Q *is not empty* **do**
    $v \leftarrow Q.dequeue()$;
    for $w \in \mathcal{N}^s_v(\bar{G}^i)$ **do**
      if *Reserve* $\vec{e}_{vw}$ *in* G^i *by* p^i_{vw} *successfully* **then**
        if w *wasn't visited before* **then**
          $Q.enqueue(w)$;
          Mark w visited;
      $R_1 \leftarrow R_1 \cup \{w\}$ if w can be visited from u;
      $R_2 \leftarrow R_2 \cup \{w\}$ if w can be visited from N_u;
return *$\{R_1, R_2\}$ as a sample of* $\mathbb{PRRS}_{f,M}$

5.2. OPIM to solve MM in online-matching

Sample θ times for $\mathbb{RRS}_{o,M}$ independently, and denote the results as $X^M_\theta = \{x_1, x_2, ..., x_\theta\}$. Function $F_{X^M_\theta}(S) = \frac{1}{\theta}\sum_{i=1}^{\theta}\mathbb{I}\{S \cap x_i \neq \emptyset\}$ which is the proportion of sets in X^M_θ covered by S, is a statistical estimation for $Pr\{\mathbb{RRS}_{o,M} \cap S \neq \emptyset\}$. Then we have $\sigma_{o,M} \approx |\mathcal{V}^m_M| F_{X^M_\theta}(S)$. The greedy Max-Coverage (Algorithm 3) provides a $(1-1/e)$-approximation solution for the maximum coverage problem [31] to choose k nodes that cover the maximum number of sets in X^M_θ. So at the same time we may get same approximation solution for our neighbor matching maximization problem ignoring the estimation error, and suppose that $S^{k,\theta}_{o,M}$ is the solution given by the Algorithm 3. As the good properties such as the submodularity of $\sigma_{o,M}$ in Theorem 1, 4, according to the theory analysis in [9], we have following theorem.

Algorithm 3: Max-Coverage(X^M_θ, V, k).

$S^{k,\theta}_{o,M} \leftarrow \emptyset$;
for $i = 1$ *to* k **do**
  Get the node $s_i \in V / S^{k,\theta}_{o,M}$ which *covers* most sets in X^M_θ;
  Remove all sets from X^M_θ, which is *covered* by s_i;
  Let $S^{k,\theta}_{o,M} \leftarrow S^{k,\theta}_{o,M} \cup \{s_i\}$;
return $S^{k,\theta}_{o,M}$ *as the selected seeds*

Theorem 8. *Given parameters* $\epsilon \in (0, 1]$ *and* $l > 0$, $\sigma_{o,M}(S^{k,\theta}_{o,M}) \geq (1 - 1/e - \epsilon)\sigma_{o,M}(S^{opt_k}_{\sigma_{o,M}})$ *holds with probability at least* $1 - |V^m_M|^{-l}$, *when*

$$\theta \geq (8 + 2\epsilon)|V^m_M| \cdot \frac{l \log |V^m_M| + \log \binom{|V^m_M|}{k} + \log(2)}{S^{opt_k}_{\sigma_{o,M}} \cdot \epsilon^2}, \tag{7}$$

where $S^{opt_k}_{\sigma_{o,M}}$ *is the optimum solution for the MM problem with matched type* M *in online-matching.*

As shown in Theorem 8, the number of samples θ must be large enough to ensure the approximation, because the more samples lead to less error between the statistical estimation and the truth, but more sampling cost. Note that it's difficult to set θ directly from Equation (7), since $S^{opt_k}_{\sigma_{o,M}}$ is unknown. Many algorithms to solve such parameters estimation in sampling have been proposed such as TIM [8], IMM [9], SSA [10], OPIM [29]. We adapt the Algorithm 4 from the OPIM which provides an algorithm framework to solve the above set sampling problem when the target function can be estimated by the statistical method of covering a random set. And we can get following formulations.

Theorem 9. *The adapted OPMM can guarantee: The output* $S^{k,\theta}_{o,M}$ *is an* $(1 - 1/e - \epsilon)$*-approximation solution with probability at least* $1 - \delta (0 < \delta < 1)$; *When* $\delta \leq 1/2$, *the expected sampling number of MRS sets is* $O\left(\left(k \ln |V| - \ln(\delta)\right)|V^m_M|\epsilon^{-2}/S^{opt_k}_{\sigma_{o,M}}\right)$.

Hence, by this theorem, we have the expected time cost is $O\left(\frac{\mathrm{ESS}\left(k \ln |V| - \ln(\delta)\right)|V^m_M|\epsilon^{-2}}{S^{opt_k}_{\sigma_{o,M}}}\right)$, where ESS the expected number of the nodes searched in sampling a $\mathbb{RRS}_{o,M}$ set.

Algorithm 4: OPMM^{o,M}(G^i, G^m, k, ϵ, l).

Estimate the samples number θ by the algorithm in OPIM [29] with parameters setting (ϵ, l);
Sample θ sets of $\mathbb{RRS}_{o,M}$ as X_θ by Algorithm 1;
$\mathcal{S} \leftarrow \emptyset$;
for $q = 1$ *to* k **do**
  Get the node s_q which *covers* most sets in X^M_θ;
  Remove all sets from X^M_θ, which is *covered* by s_q;
  Let $S \leftarrow S \cup \{s_q\}$;
return $\mathcal{S}$ *as the seed set*

5.3. Greedy method to solve MM in offline-matching

Sample θ times for $\mathbb{PRRS}_{f,M}$ independently, and denote the results as $XY^M_\theta = \{(x_1, y_1), (x_2, y_2), ..., (x_\theta, y_\theta)\}$. Function $F_{XY^M_\theta}(S) = \frac{1}{\theta}\sum_{i=1}^{\theta} \mathbb{I}\{S \cap x_i \neq \emptyset\}\mathbb{I}\{S \cap y_i \neq \emptyset\}$ is a statistical estimation for the probability $Pr\{\mathbb{PRRS}_{f,M}[1] \cap S \neq \emptyset \wedge \mathbb{PRRS}_{f,M}[2] \cap S \neq \emptyset\}$. We have $\sigma_{f,M} \approx |\mathcal{V}^M(G^m)|F_{XY^M_\theta}(S)$ according to Theorem 7. By selecting nodes greedily to solve the maximization problem of $F_{XY^M_\theta}$ with k size limited, we propose a heuristic Algorithm 4 G-RRSP^{f,M} for the MM problem with matched type M in offline-matching.

Algorithm 5: G-RRSP$^{f,M}(G^i, G^m, k, \theta)$.

Sample θ set pairs of $\mathbb{PRRS}_{f,M}$ as XY_θ^M by Algorithm 2.;
$S \leftarrow \emptyset$;
for q *in* $1, \ldots, k$ **do**
  Get the node s_q which *covers* most set pairs.;
  if *Got no such node* **then**
    Get the node s_q which *covers* most set;
  Let $S \leftarrow S \cup \{s_q\}$;
  Remove all sets *coverd* by s_q;
  Remove all sets pairs in which both sets are empty.;
return *S as the seed set*

5.4. Sandwich algorithm for MM in offline-matching

As $F_{XY_\theta^M}$ is not submodular, the greedy algorithm G-RRSP^{f,M} can't provide approximation guarantee. Sandwich Approximation (SA) [16] provides an approximation analysis if there are submodular upper and lower bounds for the objective function. According to Theorem 2, η is a submodular upper bound of $\sigma_{f,M}$ and $\sigma_{o,M}$ is a submodular lower bound of $\sigma_{f,M}$. Then Given G^i, G^m, k, we propose algorithm SAMM^{f,M} to solve the offline-matching as follows. Run OPIM, OPMM^{o,M} with same parameter setting (ϵ, l) and run G-RRSP^{f,M} with a suitable θ. Then get the corresponding solutions as S_η^k, $S_{\sigma_{o,M}}^k$, $S_{\sigma_{f,M}}^k$. We have the solution $S_{sa}^k := argmax_{S\in\{S_\eta^k, S_{\sigma_{o,M}}^k, S_{\sigma_{f,M}}^k\}} \sigma_{f,M}(S)$, which provides the approximation guarantee as following theorem.

Theorem 10. *The solution given by SAMM^{f,M} satisfies*

$$\sigma_{f,M}(S_{sa}^k) \geq \beta(1 - 1/e - \epsilon).\sigma_{f,M}(S_{f,M}^{opt_k}) \tag{8}$$

with probability at least $1 - 2|V_M^m|^{-l}$, where $S_{\sigma_{f,M}}^{opt_k}$ is the optimum solution for the $MM^{f,M}$ problem, and $\beta = max\{\frac{\sigma_{f,M}(S_\eta^k)}{\eta(S_\eta^k)}, \frac{\sigma_{o,M}(S_{\sigma_{f,M}}^{opt_k})}{\sigma_{f,M}(S_{\sigma_{f,M}}^{opt_k})}\}$.

Proof. According to the approximation analysis in OPIM [9] and OPMM in Theorem 7, we have

$$\eta(S_\eta^k) \geq (1 - 1/e - \epsilon) \cdot \eta(S_\eta^{opt_k})$$

holding with probability at least $1 - |V|^{-l}$ where $S_\eta^{opt_k}$ is the optimum solution for IM and

$$\sigma_{o,M}(S_{\sigma_{o,M}}^k) \geq (1 - 1/e - \epsilon) \cdot \sigma_{o,M}(S_{\sigma_{o,M}}^{opt_k})$$

holding with probability at least $1 - |V_M^m|^{-l}$. We denote the event that the following inequality holds as E_1

$$\begin{aligned}\sigma_{f,M}(S_\eta^k) =& \frac{\sigma_{f,M}(S_\eta^k)}{\eta(S_\eta^k)} \cdot \eta(S_\eta^k) \geq \frac{\sigma_{f,M}(S_\eta^k)}{\eta(S_\eta^k)} \cdot (1 - 1/e - \epsilon) \cdot \eta(S_\eta^{opt_k}) \\ \geq& \frac{\sigma_{f,M}(S_\eta^k)}{\eta(S_\eta^k)} \cdot (1 - 1/e - \epsilon) \cdot \sigma_{f,M}(S_{\sigma_{f,M}}^{opt_k})\end{aligned}$$

and we denote the event that following inequality holds as E_2

$$\begin{aligned}\sigma_{f,M}(S_{\sigma_{o,M}}^k) \geq& \sigma_{o,M}(S_{\sigma_{o,M}}^k) \geq (1 - 1/e - \epsilon) \cdot \sigma_{o,M}(S_{\sigma_{o,M}}^{opt_k}) \\ \geq& (1 - 1/e - \epsilon) \cdot \sigma_{o,M}(S_{\sigma_{f,M}}^{opt_k}) \\ \geq& \frac{\sigma_{o,M}(S_{\sigma_{f,M}}^{opt_k})}{\sigma_{f,M}(S_{\sigma_{f,M}}^{opt_k})} \cdot (1 - 1/e - \epsilon) \cdot \sigma_{f,M}(S_{\sigma_{f,M}}^{opt_k})\end{aligned}$$

Then we have $Pr(E_1) \geq 1 - |V|^{-l}$ and $Pr(E_1) \geq 1 - |V_M^m|^{-l}$. We denote the event that the Inequality (8) holds as E_0. To get the lower bound of E_0's probability, considering the probability that E_1 and E_2 both hold, we have

Table 2
Origin labeled datasets.

Network	# nodes	# edges	direct
Facebook	4049	88234	False
Twitter	81306	1768149	True
Gplus	107614	13673453	True

$$\begin{aligned} Pr(E_0) &\geq Pr\{E_1 \wedge E_2\} = Pr\{E_1\} - Pr\{E_1 \wedge \bar{E}_2\} \\ &\geq Pr\{E_1\} - Pr\{\bar{E}_2\} = Pr\{E_1\} + Pr\{E_2\} - 1 \\ &= 1 - |V_M^m|^{-l} - |V|^{-l} \geq 1 - 2|V_M^m|^{-l} \end{aligned}$$

The theorem has been proved. □

6. Experiments

In this section, our experiments aim to evaluate the performance of the methods proposed, based on 3 real-world labeled datasets[4] (facebook, twitter, gplus) as shown in Table 2. All codes of the experiments are written in c++ with parallel optimization, and we run the experiments on a linux machine with 6 cores, 12 threads, 3.6 hz, CPU and 16G RAM. At first, we model the G^i and G^m from original datasets as follows.

G^i: For facebook, we construct two influence edges $\vec{e}_{uv}, \vec{e}_{vu}$ if there is the friendship between two users u and v. For twitter and gplus, we construct an influence edge $\vec{e}_{uv}$ if user v follows user u. As the general setting, for each edge $\vec{e}_{uv}$ we let the influence probability p^i_{uv} to be $\frac{1}{d_i(v)}$ in which $d_i(v)$ is the indegree of v.

G^m: We construct the matching probability by a natural idea that more common features provide higher matching probability. For two nodes u, v with feature set L_u, L_v, the probability of v matching u is with correlation to $\frac{|L_u \cap L_v|}{|L_u|}$. Specially we set a threshold λ and then construct the matching to be definite as follows:

$$p^m_{uv} = \begin{cases} 1 & \text{if } \frac{|L_u \cap L_v|}{|L_u|} \geq \lambda \\ 0 & \text{if } \frac{|L_u \cap L_v|}{|L_u|} < \lambda \end{cases}$$

Given G^i, G^m, we compare our algorithms with the following algorithms as baselines. OPIM [9] is the method to solve the IM problem in G^i. And HighDegree is the method choosing the seeds from top k nodes with high outdegree in G^i.

6.1. Parameters setting

We set $\lambda = 0.1$ for facebook and $\lambda = 0.2$ for gplus and $\lambda = 0.3$ for twitter. With lower ϵ and higher l, the algorithms OPMM provide a higher approximation with higher probability, but cost more running time, so we set the parameters $\epsilon = 0.1$ and $l = 1$ in OPIM and SAMM. We set the number of samples $\theta = 10|V_m^M|$ for G-RRSP. To estimate the matching influence for given solutions, we use Monte Carlo with simulation times $r = 10000$. By setting the seeds size $k = 1$ and $k = 10$ to 200 with a step of 10, we run the experiments to evaluate the performance with changed seed size.

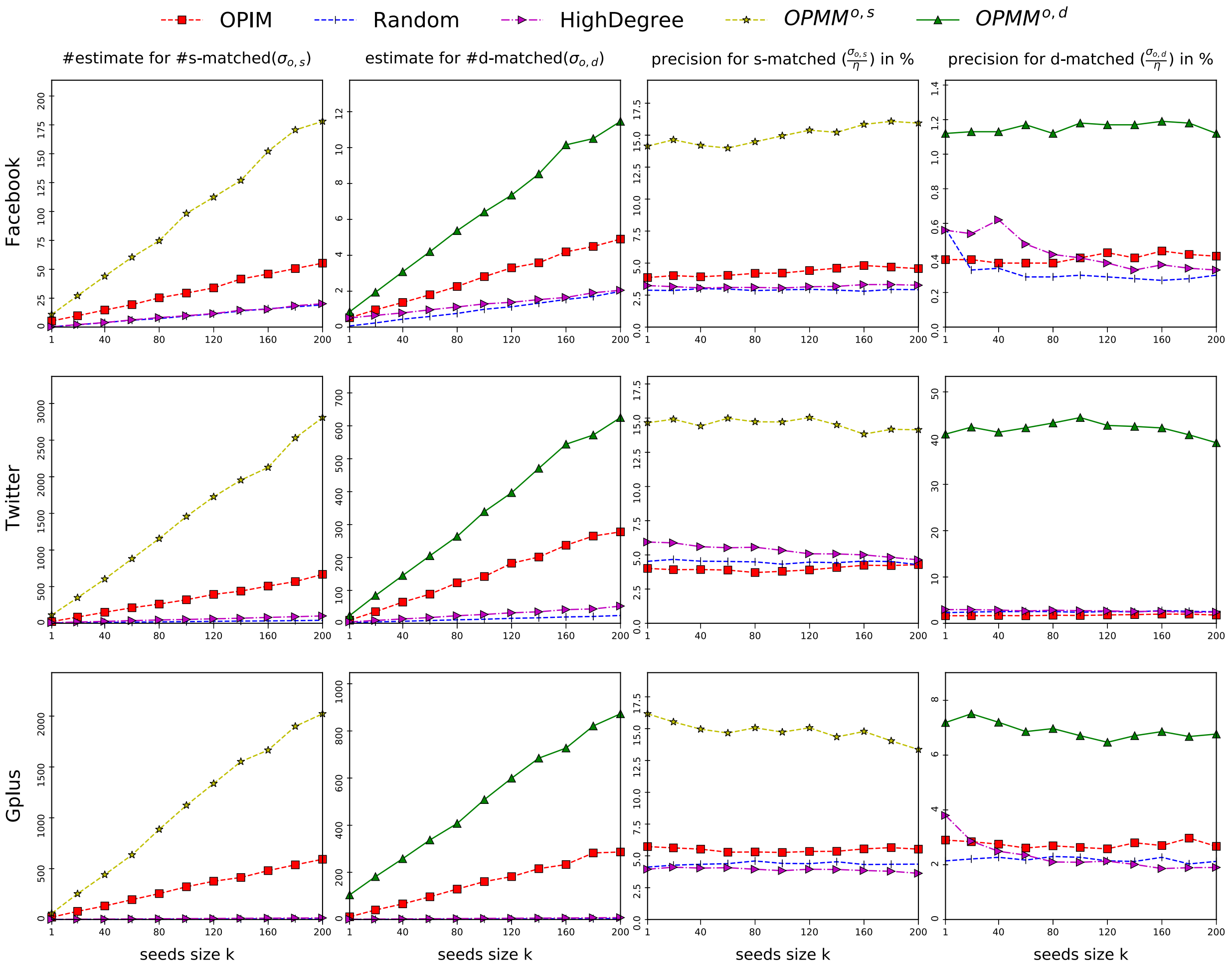
6.2. Performance comparison

We evaluate the performance of the algorithms from not only the matching influence but also matching precision which is the percentage of matched nodes in all influenced nodes.

Online-matching: We compare the solution of our algorithm OPMM with OPIM, Random, HighDegree in online-matching. As shown in Fig. 3(a), in each dataset, OPMM is always outperformed than the baselines both in the matching influence and matching precision. We have the solutions provide the matching influence with following comparisons: OPMM>OPIM>HighDegree≈Random. OPIM provides higher matching influence than Random and HighDegree as the advantage in influence spread, but OPIM only aims to maximize the number of influenced nodes but can't distinguish the node with low matching probability, and so OPIM has no obvious difference with Random and HighDegree in terms of matching precision.

Offline-matching: For the offline-matching, we compare the solution of our algorithm SAMM with OPIM, Random, HighDegree. As shown in Fig. 3(b), similar to the result in online-matching, the solutions of SAMM we proposed also have the best performance in all three datasets, but note that SAMM will choose the best solution among OPMM, OPIM, G-RRSP, so it

[4] http://snap.stanford.edu/data/.



(a) Online-matching

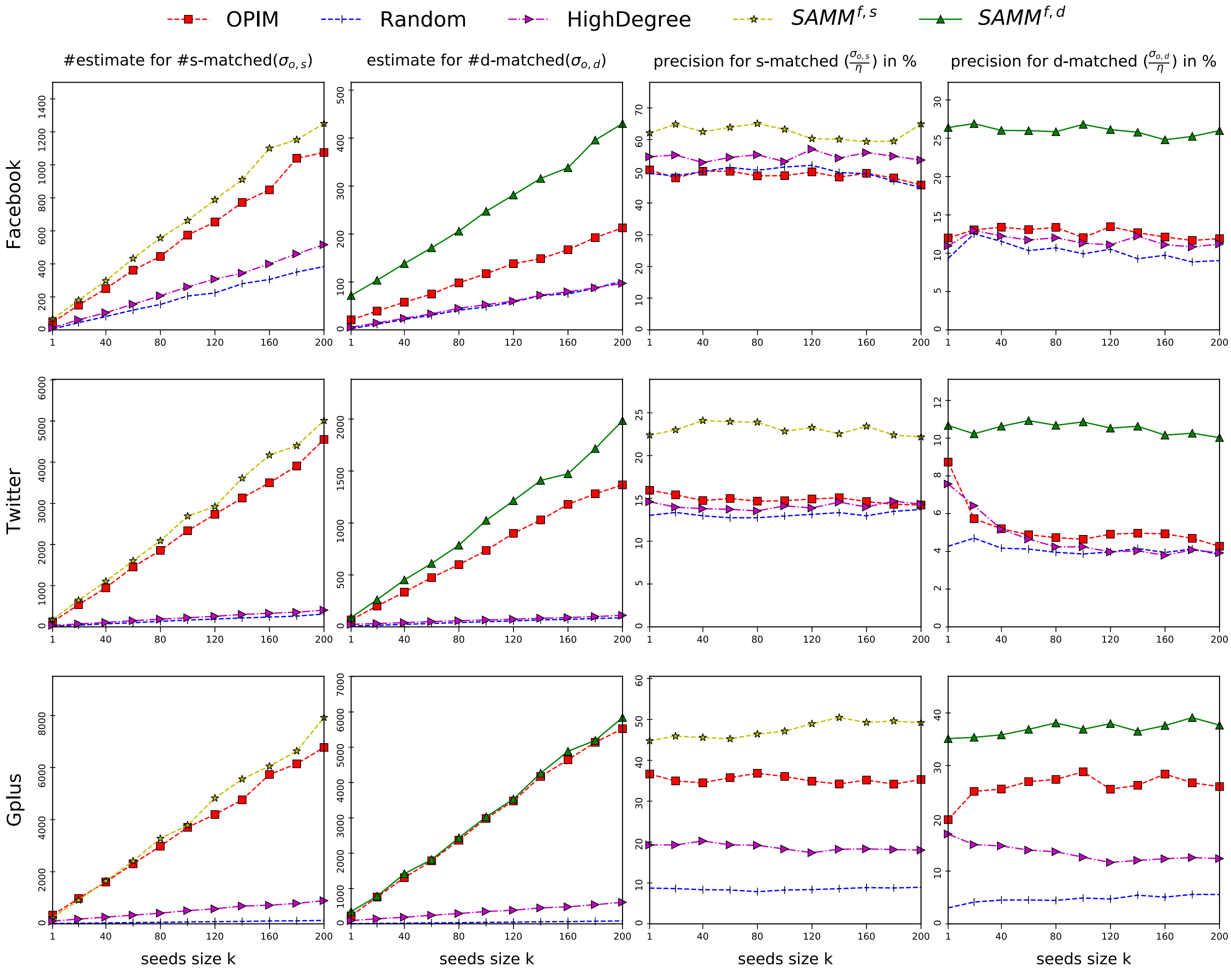
Fig. 3. Performance comparisons with different matching models in Facebook, Twitter, Gplus.

can't be worse than OPIM in offline-matching. We can see that the solutions of SAMM are always given by G-RRSP or OPMM in our experiments as the curves of SAMM are above the curves of OPIM. In fact, offline-matching for a influenced node is easier than the online-matching as more matching chances, so we can see that there is no significant improvement from OPIM to SAMM in Gplus and Twitter in terms of the matching influence, but SAMM still improve the matching precision by about 10%.

Running time: As shown in Fig. 4, we compare the running time for OPIM, OPMM, and G-RRSP by running each algorithm repeatedly with 20 times. The average running time increases nearly linearly with the increase of the network scale from facebook to gplus. The OPMM can estimate the sample number for the reverse reachable sets which is high correlation to k. So we compare the running time of OPMM with the change of seeds size k. The experiments show that the running time also increases nearly linearly with the increase of the seed size. Our experiments show that the running time of OPMM and G-RRSP are smaller than OPIM, as OPIM may cost more time to sample a RIS set than the RRS set or RRSP pair we proposed. Note that, in offline-matching, the SA-MM needs to run three algorithms TIM-MM, TIM-IM, G-RRSP, so it costs at least 3 times running time than OPMM or G-RRSP. Note that we ignore the HighDegree and Random method as they cost almost no time.

7. Conclusions

In this paper, we point out that the tradition IM can't offer a good solution in the viral marketing with matching requirements. We introduce two diffusion-matching models as online-matching and offline-matching, and model two matching types by a node being matched unilaterally or bilaterally. We are the first one to formulate the MM problems, the goal of which is to find a small size seed set such that the expected number of matched nodes is maximized. To solve the MM prob-



(b) Offline-matching

Fig. 3. (*continued*)

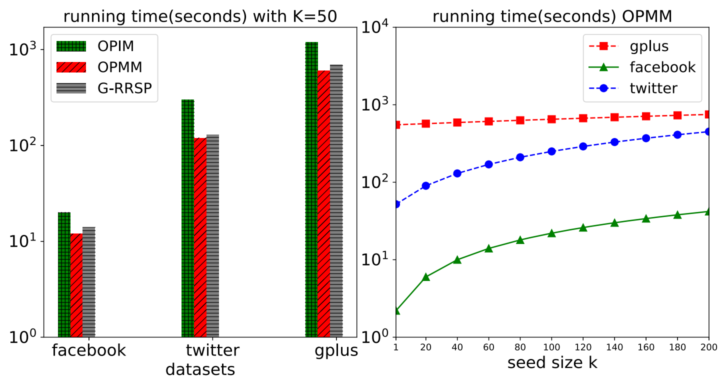


Fig. 4. Time efficiency in facebook, twitter, gplus.

lems, we design the efficient algorithm OPMM (SAMM) with the $(1-1/e-\epsilon)$-approximation $(\beta(1-1/e-\epsilon)$-approximation) guarantee for online-matching (offline-matching). At last, a lot of experiments have been conducted on real-world datasets showing that the method we proposed outperforms other comparison methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Rao, Y. Wang, W. Chen, D. Li, W. Wu, Matched participants maximization based on social spread, in: International Conference on Combinatorial Optimization and Applications, Springer, 2020, pp. 214–229.
[2] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 57–66.
[3] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.
[4] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 199–208.
[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2007, pp. 420–429.
[6] A. Goyal, W. Lu, L.V. Lakshmanan, Celf++: optimizing the greedy algorithm for influence maximization in social networks, in: Proceedings of the 20th International Conference Companion on World Wide Web, ACM, 2011, pp. 47–48.
[7] 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.
[8] Y. Tang, X. Xiao, Y. Shi, Influence maximization: near-optimal time complexity meets practical efficiency, in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ACM, 2014, pp. 75–86.
[9] Y. Tang, Y. Shi, X. Xiao, Influence maximization in near-linear time: a martingale approach, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1539–1554.
[10] H.T. Nguyen, M.T. Thai, T.N. Dinh, Stop-and-stare: optimal sampling algorithms for viral marketing in billion-scale networks, in: Proceedings of the 2016 International Conference on Management of Data, ACM, 2016, pp. 695–710.
[11] B. Liu, G. Cong, D. Xu, Y. Zeng, Time constrained influence maximization in social networks, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 439–448.
[12] E. Cohen, D. Delling, T. Pajor, R.F. Werneck, Timed influence: computation and maximization, arXiv preprint, arXiv:1410.6976, 2014.
[13] N. Barbieri, F. Bonchi, G. Manco, Topic-aware social influence propagation models, Knowl. Inf. Syst. 37 (3) (2013) 555–584.
[14] S. Chen, J. Fan, G. Li, J. Feng, K.-l. Tan, J. Tang, Online topic-aware influence maximization, Proc. VLDB Endow. 8 (6) (2015) 666–677.
[15] S. Bharathi, D. Kempe, M. Salek, Competitive influence maximization in social networks, in: International Workshop on Web and Internet Economics, Springer, 2007, pp. 306–311.
[16] W. Lu, W. Chen, L.V. Lakshmanan, From competition to complementarity: comparative influence diffusion and maximization, Proc. VLDB Endow. 9 (2) (2015) 60–71.
[17] X. He, G. Song, W. Chen, Q. Jiang, Influence blocking maximization in social networks under the competitive linear threshold model, in: Proceedings of the 2012 Siam International Conference on Data Mining, SIAM, 2012, pp. 463–474.
[18] G. Li, S. Chen, J. Feng, K.-l. Tan, W.-s. Li, Efficient location-aware influence maximization, in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ACM, 2014, pp. 87–98.
[19] C. Song, W. Hsu, M.L. Lee, Targeted influence maximization in social networks, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, 2016, pp. 1683–1692.
[20] L. Sun, W. Huang, P.S. Yu, W. Chen, Multi-round influence maximization, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 2249–2258.
[21] K.S. Anand, R. Aron, Group buying on the web: a comparison of price-discovery mechanisms, Manag. Sci. 49 (11) (2003) 1546–1562.
[22] X. Jing, J. Xie, Group buying: a new mechanism for selling through social interactions, Manag. Sci. 57 (8) (2011) 1354–1372.
[23] R. Becker, F. Coro, G. D'Angelo, H. Gilbert, Balancing spreads of influence in a social network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 3–10.
[24] A. Tsang, B. Wilder, E. Rice, M. Tambe, Y. Zick, Group-fairness in influence maximization, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 5997–6005.
[25] J. Zhu, S. Ghosh, W. Wu, Group influence maximization problem in social networks, IEEE Trans. Comput. Soc. Syst. 6 (6) (2019) 1156–1164.
[26] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, An analysis of approximations for maximizing submodular set functions—i, Math. Program. 14 (1) (1978) 265–294.
[27] N. Ohsaka, T. Akiba, Y. Yoshida, K.-i. Kawarabayashi, Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[28] K. Huang, S. Wang, G. Bevilacqua, X. Xiao, L.V. Lakshmanan, Revisiting the stop-and-stare algorithms for influence maximization, Proc. VLDB Endow. 10 (9) (2017) 913–924.
[29] J. Tang, X. Tang, X. Xiao, J. Yuan, Online processing algorithms for influence maximization, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 991–1005.
[30] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1029–1038.
[31] V.V. Vazirani, Approximation Algorithms, Springer Science & Business Media, 2013.