Influence Spread in Social Networks with Both Positive and Negative Influences

Jing (Selena) He[1], Ying Xie[1], Tianyu Du[3], Shouling Ji[2*] and Zhao Li[3]

[1]Department of Computer Science, Kennesaw State University, Marietta, GA, 30060
[2]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[3]Alibaba Group, Hangzhou, China
{jhe4, yxie2}@kennesaw.edu, {tianyu.dty, lizhao.lz}@alibaba-inc.com, sji@zju.edu.cn
* Shouling Ji is the corresponding author

Abstract. Social networks are important mediums for spreading information, ideas, and influences among individuals. Most of existing research works of social networks focus on understanding the characteristics of social networks and spreading information through the "word of mouth" effect. However, most of them ignore negative influences among individuals and groups. Motivated by alleviating social problems, such as drinking, smoking, gambling, and influence spreading problems such as promoting new products, we take both positive and negative influences into consideration and propose a new optimization problem, named the Minimum-sized Positive Influential Node Set (MPINS) selection, to identify the minimum set of influential nodes, such that every node in the network can be positively influenced by these selected nodes no less than a threshold θ. Our contributions are threefold. First, we prove that, under the independent cascade model considering both positive and negative influences, MPINS is APX-hard. Subsequently, we present a greedy approximation algorithm to address the MPINS selection problem. Finally, to validate the proposed greedy algorithm, extensive simulations and experiments are conducted on random Graphs and seven different real-world data sets representing small, medium, and large scale networks.

Keywords: Influence spread, social networks, positive influential node set, greedy algorithm, positive and negative influences

1 Introduction

A social network (e.g., Facebook, Google+, and MySpace) is composed of a set of nodes that share the similar interest or purpose. The network provides a powerful medium of communication for sharing, exchanging, and disseminating information. With the emergence of social applications (such as Flickr, Wikis, Netflix, and Twitter, *etc.*), there has been tremendous interests in how to effectively utilize social networks to spread ideas or information within a community [1–8]. In a social network, individuals may have both positive and negative influence on each other. For example, within the context of gambling, a gambling insulator has positive influence on his friends/neighbors. Moreover, if many of an individual's friends are gambling insulators, the aggregated positive influence is exacerbated. However, an individual might turn into a gambler, who brings negative impact on his friends/neighbors.

One application of MPINS is described as follows. A community wants to implement a smoking intervention program. To be cost effective and get the maximum effect, the community wishes to select a small number of influential individuals in the community to attend a quit-smoking campaign. The goal is that all other individuals in the community

will be positively influenced by the selected users. Constructing an MPINS is helpful to alleviate the aforementioned social problem, and it is also helpful to promote new products in the social network. Consider the following scenario as another motivation example. A small company wants to market a new product in a community. To be cost effective and get maximum profit, the company would like to distribute sample products to a small number of initially chosen influential users in the community. The company wishes that these initial users would like the product and positively influence their friends in the community. The goal is to have other users in the community be positively influenced by the selected users no less than θ eventually. To sum up, the specific problem we investigate in this work is the following: given a social network and a threshold θ, identify a minimum-sized subset of individuals in the network such that the subset can result in a positive influence on every individual in the network no less than θ.

Hence, we explore the MPINS selection problem under the *independent cascade model* considering both positive and negative influences, where individuals can positively or negatively influence their neighbors with certain probabilities.

In this paper, first we formally define the MPINS problem and then propose a greedy approximation algorithm to solve it. Particularly, the main contributions of this work are summarized as follows:

- Taking both positive and negative influences into consideration, we introduce a new optimization problem, named the Minimum-sized Positive Influential Node Set (MPINS) selection problem, for social networks, which is to identify the minimum-sized set of influential nodes, that could positively influence every node in the network no less than a pre-defined threshold θ. We prove that it is an APX-hard problem under the independent cascade model.
- We define a contribution function, which suggests us a greedy approximation algorithm called MPINS-GREEDY to address the MPINS selection problem. The correctness of the proposed algorithm is analyzed in the paper as well.
- We also conduct extensive simulations and experiments to validate our proposed algorithm. The simulation and experiment results show that the proposed greedy algorithm works well to solve the MPINS selection problem. More importantly, the solutions obtained by the greedy algorithm is very close to the optimal solution of MPINS in small scale networks.

The rest of this paper is organized as follows: in Section 2, we review some related literatures with remarking the difference. In Section 3, we first introduce the network model and then we formally define the MPINS selection problem and prove its APX-hardness. The greedy algorithm and theoretical analysis on the correctness of the algorithm are presented in Section 4. The simulation and experimental results are presented in Section 5 to validate our proposed algorithm. Finally, the paper is concluded in Section 6.

2 Related Work

In this section, we first briefly review the related works of social influence analysis. Subsequently, we summarize some related literatures of the PIDS problem and the influence maximization problem.

Influence maximization, initially proposed by Kempe *et al.* [1], is targeting at selecting a set of users in a social network to maximize the expected number of influenced users through several steps of information propagation [9]. A series of empirical studies have been performed on influence learning [10, 11], algorithm optimizing [12, 13], scalability promoting

[14, 15], and influence of group conformity [16, 4]. Saito *et al.* predicted the information diffusion probabilities in social networks under the independent cascade model in [17]. Tang *et al.* argued that the effect of the social influence from different angles (topics) may be different. Hence, they introduced Topical Affinity Propagation (TAP) to model topic-related social influence on large social networks in [18, 19]. Later, Tang *et al.* [20] proposed a Dynamic Factor Graph (DFG) model to incorporate the time information to analyze dynamic social influences.

Wang *et al.* first proposed the PIDS problem under the deterministic linear threshold model in [21], which is to find a set of nodes D such that every node in the network has at least half of its neighbor nodes in D. Subsequently, Zhu *et al.* proved that PIDS is APX-hard and proposed two greedy algorithms with approximation ratio analysis in [22] and [23]. He *et al.* [24, 25] proposed a new optimization problem named the Minimum-sized Influential Node Set (MINS) selection problem, which is to identify the minimum-sized set of influential nodes. But they neglected the existence of negative influences.

To address the scalability problem of the algorithms in [1, 26], Leskovec *et al.* [27] presented a "lazy-forward" optimization scheme on selecting initial nodes, which greatly reduces the number of influence spread evaluations. Laterly, Chen *et al.* [28] showed that the problem of computing exact influence in social networks under both models are #P-Hard. They also proposed scalable algorithms under both models, which are much faster than the greedy algorithms in [1, 26]. Most recently, consider the data from both cyber-physical world and online social network, [29, 30] proposed methods to solve the problem of influence maximization comprehensively.

However, all the aforementioned works did not consider negative influence when they model the social networks. Besides taking both positive and negative influences into consideration, our work try to find a minimum-sized set of individuals that guarantees the positive influences on every node in the network no less than a threshold θ, while the influence maximization problem focuses on choosing a subset of a pre-defined size k that maximizes the expected number of influenced individuals. Since we study the MPINS selection problem under the independent cascade model and take both positive and negative influences into consideration, our problem is more practical. In addition, PIDS is investigated under the deterministic linear threshold model.

3 Problem Definition and Hardness Analysis

In this section, we first introduce the network model. Subsequently, we formally define the MPINS selection problem and make some remarks on the proposed problem. Finally, we analyze the hardness of the MPINS selection problem.

3.1 Network Model

We model a social network by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, where $\mathcal{V}$ is the set of n nodes, denoted by u_i, and $0 \leq i < n$. i is called the node ID of u_i. An undirected edge $(u_i, u_j) \in \mathcal{E}$ represents a social tie between the pair of nodes. $\mathcal{P}(\mathcal{E}) = \{p_{ij} \mid \text{if } (u_i, u_j) \in \mathcal{E}, 0 < p_{ij} \leq 1, \text{ else } p_{ij} = 0\}$, where p_{ij} indicates the social influence between nodes u_i and u_j[1]. It is worth to mention that the social influence can be categorized into two groups: positive influence and negative influence. For simplicity, we assume the links are undirected (bidirectional), which means two linked nodes have the same social influence (*i.e.*, p_{ij} value) on each other.

[1] This model is reasonable since many empirical studies have analyzed the social influence probabilities between nodes [17, 10, 20].

3.2 Problem Definition

The objective of the MPINS selection problem is to identify a subset of influential nodes as the initialized nodes. Such that, all the other nodes in a social network can be positively influenced by these nodes no less than a threshold θ. For convenient, we call the initial nodes been selected as *active nodes*, otherwise, *inactive nodes*. Therefore, how to define *positive influence* is critical to solve the MPINS selection problem. In the following, we first formally define some terminologies, and then give the definition of the MPINS selection problem.

Definition 1. *Positive Influential Node Set ($\mathcal{I}$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, the positive influential node set is a subset $\mathcal{I} \subseteq \mathcal{V}$, such that all the nodes in $\mathcal{I}$ are initially selected to be the active nodes.

Definition 2. *Neighboring Set ($\mathcal{B}(u_i)$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, $\forall u_i \in \mathcal{V}$, the neighboring set of u_i is defined as: $\mathcal{B}(u_i) = \{u_j \mid (u_i, u_j) \in \mathcal{E}, p_{ij} > 0\}$.

Definition 3. *Active Neighboring Set ($\mathcal{A}^{\mathcal{I}}(u_i)$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, $\forall u_i \in \mathcal{V}$, the active neighboring set of u_i is defined as: $\mathcal{A}^{\mathcal{I}}(u_i) = \{u_j \mid u_j \in \mathcal{B}(u_i), u_j \in \mathcal{I}\}$.

Definition 4. *Non-active Neighboring Set ($\mathcal{N}^{\mathcal{I}}(u_i)$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, $\forall u_i \in \mathcal{V}$, the non-active neighboring set of u_i is defined as: $\mathcal{N}^{\mathcal{I}}(u_i) = \{u_j \mid u_j \in \mathcal{B}(u_i), u_j \notin \mathcal{I}\}$.

Definition 5. *Positive Influence ($p_{u_i}(\mathcal{A}^{\mathcal{I}}(u_i))$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, a node $u_i \in \mathcal{V}$, and a positive influential node set $\mathcal{I}$, we define a joint influence probability of $\mathcal{A}^{\mathcal{I}}(u_i)$ on u_i, denoted by $p_{u_i}(\mathcal{A}^{\mathcal{I}}(u_i))$ as $p_{u_i}(\mathcal{A}^{\mathcal{I}}(u_i)) = 1 - \prod_{u_j \in \mathcal{A}^{\mathcal{I}}(u_i)} (1 - p_{ij})$.

Definition 6. *Negative Influence ($p_{u_i}(\mathcal{N}^{\mathcal{I}}(u_i))$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, a node $u_i \in \mathcal{V}$, and a positive influential node set $\mathcal{I}$, we define a joint influence probability of $\mathcal{N}^{\mathcal{I}}(u_i)$ on u_i, denoted by $p_{u_i}(\mathcal{N}^{\mathcal{I}}(u_i))$ as $p_{u_i}(\mathcal{N}^{\mathcal{I}}(u_i)) = 1 - \prod_{u_j \in \mathcal{N}^{\mathcal{I}}(u_i)} (1 - p_{ij})$.

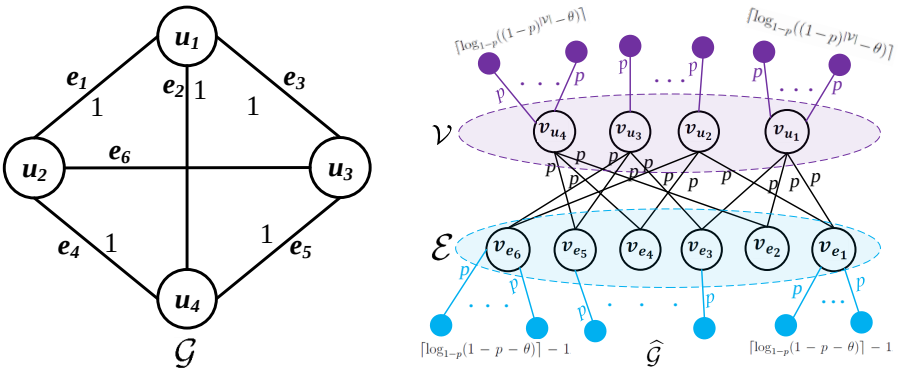
Definition 7. *Ultimate Influence ($\varrho^{\mathcal{I}}(u_i)$).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, a node $u_i \in \mathcal{V}$, and a positive influential node set $\mathcal{I}$, we define an ultimate influence of $\mathcal{B}(u_i)$ on u_i, denoted by $\varrho^{\mathcal{I}}(u_i)$ as $\varrho^{\mathcal{I}}(u_i) = p_{u_i}(\mathcal{A}^{\mathcal{I}}(u_i)) - p_{u_i}(\mathcal{N}^{\mathcal{I}}(u_i))$. Moreover, if $\varrho^{\mathcal{I}}(u_i) < 0$, we set $\varrho^{\mathcal{I}}(u_i) = 0$. If $\varrho^{\mathcal{I}}(u_i) \geq \theta$, where $0 < \theta < 1$ is a pre-defined threshold, then u_i is said been *positively influenced*. Otherwise, u_i is not been positively influenced.

Definition 8. *Minimum-sized Positive Influential Node Set (MPINS).* For social network $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, the MPINS selection problem is to find a minimum-sized positive influential node set $\mathcal{I} \subseteq \mathcal{V}$, such that $\forall u_i \in \mathcal{V} \setminus \mathcal{I}$, u_i is positively influenced, *i.e.*, $\varrho^{\mathcal{I}}(u_i) = p_{u_i}(\mathcal{A}^{\mathcal{I}}(u_i)) - p_{u_i}(\mathcal{N}^{\mathcal{I}}(u_i)) \geq \theta$, where $0 < \theta < 1$.

3.3 Problem Hardness Analysis

In general, given an arbitrary threshold θ, the MPINS selection problem is APX-hard. We prove the APX-hardness of MPINS by constructing a *L-reduction* from Vertex Cover problem in Cubic Graph (denoted by VCCG) to the MPINS selection problem. The decision problem of VCCG is APX-hard which is proven in [31]. A cubic graph is a graph with every vertex's degree of exactly three. Given a cubic graph, VCCG is to find a minimum-sized vertex cover[2].

[2] A vertex cover is defined as a subset of nodes in a graph $\mathcal{G}$ such that each edge of the graph is incident to at least one vertex of the set.



(a) (b)

Fig. 1: Illustration of the construction from $\mathcal{G}$ to $\widehat{\mathcal{G}}$.

First, consider a cubic graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, where $\mathcal{P}(\mathcal{E}) = \{1 \mid (u_i, u_j) \in \mathcal{E}; u_i, u_j \in \mathcal{V}\}$, as an instance of VCCG. we construct a new graph $\widehat{\mathcal{G}}$ as follows:

(1) We create $|\mathcal{V}| + |\mathcal{E}|$ nodes with $|\mathcal{V}|$ nodes $\mathbf{v}_{u_i} = \{v_{u_1}, v_{u_2}, \cdots, v_{u_{|\mathcal{V}|}}\}$ representing the nodes in $\mathcal{G}$ and $|\mathcal{E}|$ nodes $\mathbf{v}_{e_i} = \{v_{e_1}, v_{e_2}, \cdots, v_{e_{|\mathcal{E}|}}\}$ representing the edges in $\mathcal{G}$. (2) We add an edge with influence weight p between nodes v_{u_i} and v_{e_j} if and only if node u_i is an endpoint of edge e_j. (3) We attach additional $\lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta) \rceil$ active nodes to each node v_{u_i}, denoted by set $\mathbf{v}^{\mathcal{A}}_{u_i} = \{v^j_{u_i} \mid 1 \le j \le \lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta) \rceil\}$. Obviously, $|\mathbf{v}^{\mathcal{A}}_{u_i}| = \lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta) \rceil$. (4) We attach additional $\lceil \log_{1-p}(1-p-\theta) \rceil - 1$ active nodes to each node v_{e_j}, denoted by set $\mathbf{v}^{\mathcal{A}}_{e_j} = \{v^j_{e_j} \mid 1 \le j \le \lceil \log_{1-p}(1-p-\theta) \rceil - 1\}$. Obviously, $|\mathbf{v}^{\mathcal{A}}_{e_j}| = \lceil \log_{1-p}(1-p-\theta) \rceil - 1$. (5) $\widehat{\mathcal{G}} = \{\widehat{\mathcal{V}}, \widehat{\mathcal{E}}\}$, where $\widehat{\mathcal{V}} = \{v_{u_1}, \cdots, v_{u_{|\mathcal{V}|}}\} \cup \{v_{e_1}, \cdots, v_{e_{|\mathcal{E}|}}\} \cup \bigcup_{i=1}^{|\mathcal{V}|} \mathbf{v}^{\mathcal{A}}_{u_i} \cup \bigcup_{i=1}^{|\mathcal{E}|} \mathbf{v}^{\mathcal{A}}_{e_i}$, $\widehat{\mathcal{E}}$ is the set of all the edges associated with the nodes in $\widehat{\mathcal{V}}$, and $\mathcal{P}(\widehat{\mathcal{E}}) = \{p \mid \text{for every edge in } \widehat{\mathcal{E}}\}$.

Taking the cubic graph shown in Fig. 1(a) as an example to illustrate the construction procedure from $\mathcal{G}$ to $\widehat{\mathcal{G}}$. There are 4 nodes and 6 edges in $\mathcal{G}$. Therefore, we first create $\{v_{u_i}\}_{i=1}^{4}$ and $\{v_{e_j}\}_{j=1}^{6}$ nodes in $\widehat{\mathcal{G}}$. Then we add edges with influence weight p between nodes v_{u_i} and v_{e_j} based on the topology shown in $\mathcal{G}$. Subsequently, we add additional $\mathbf{v}^{\mathcal{A}}_{u_i} = \{v^j_{u_i} \mid 1 \le j \le \lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta) \rceil\}$ active nodes to each node v_{u_i} (marked by upper shaded nodes in Fig. 1(b)). Similarly, we add additional $\mathbf{v}^{\mathcal{A}}_{e_j} = \{v^j_{e_j} \mid 1 \le j \le \lceil \log_{1-p}(1-p-\theta) \rceil - 1\}$ active nodes to each node v_{e_j} (marked by bottom shaded nodes in Fig. 1(b)). The influence weights on all the additional edges are p. Finally, the new graph $\widehat{\mathcal{G}}$ is constructed as shown in Fig. 1(b).

Lemma 1. $\mathcal{G}$ has a VCCG $\mathcal{D}$ of size at most d if and only if $\widehat{\mathcal{G}}$ has a positive influential node set $\mathcal{I}$ of size at most k by setting $k = |\mathcal{V}| \lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta) \rceil + |\mathcal{E}|(\lceil \log_{1-p}(1-p-\theta) \rceil - 1) + d$.

Due to limited space in this paper, for a comprehensive proof of Lemma 1 we refer the reader to our technical report in [32].

Theorem 1. The MPINS selection problem is APX-hard.

Proof. An immediate conclusion of Lemma 1 is that $\mathcal{G}$ has a minimum-sized vertex cover of size $OPT_{VCCG}(\mathcal{G})$ if and only if $\widehat{\mathcal{G}}$ has a minimum-sized positive influential node set of

size

$$\begin{aligned} &OPT_{MPINS}(\widehat{\mathcal{G}}) \\ &= |\mathcal{V}|\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil + |\mathcal{E}|(\lceil\log_{1-p}(1-p-\theta)\rceil - 1) + OPT_{VCCG}(\mathcal{G}). \end{aligned} \tag{1}$$

Note that in a cubic graph $\mathcal{G}$, $|\mathcal{E}| = \frac{3|\mathcal{V}|}{2}$. Hence, we have

$$\frac{|\mathcal{V}|}{2} = \frac{|\mathcal{E}|}{3} \leq OPT_{VCCG}(\mathcal{G}). \tag{2}$$

Based on Lemma 1, plugging

$$|\mathcal{V}| = \frac{OPT_{MPINS}(\widehat{\mathcal{G}}) - OPT_{VCCG}(\mathcal{G})}{\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil + \frac{3}{2}(\lceil\log_{1-p}(1-p-\theta)\rceil - 1)} \tag{3}$$

into the inequality 2, we have

$$\begin{aligned} &OPT_{MPINS}(\widehat{\mathcal{G}}) \\ &\leq [2\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil + 3\lceil\log_{1-p}(1-p-\theta)\rceil - \tfrac{1}{2}]OPT_{VCCG}(\mathcal{G}). \end{aligned} \tag{4}$$

This means that VCCG is L-reducible to MPINS. In conclusion, we proved that a specific case of the MPINS selection problem is APX-hard, since the VCCG problem is APX-hard. Consequently, the general MPINS selection problem is also at least APX-hard.

Based on Theorem 1, we conclude that MPINS cannot be solved in polynomial time. Therefore, we propose a greedy algorithm to solve the problem in the next section.

4 Greedy algorithm and Performance Analysis

Since MPINS is APX-hard, we propose a greedy algorithm to solve it named MPINS-GREEDY. Before introducing MPINS-GREEDY, we first define a useful contribution function as follows:

Definition 9. *Contribution function ($f(\mathcal{I})$).* For a social network represented by graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$, and a positive influential node set $\mathcal{I}$, the contribution function of $\mathcal{I}$ to $\mathcal{G}$ is defined as: $f(\mathcal{I}) = \sum_{i=1}^{|\mathcal{V}|} \max\{\min(\varrho^{\mathcal{I}}(u_i), \theta), 0\}$.

Based on the defined contribution function, we propose a heuristic algorithm, which has two phases. First, we find the node u_i with the maximum $f(\mathcal{I})$, where $\mathcal{I} = \{u_i\}$; and after that, we select a Maximal Independent Set (MIS)[3] induced by a breadth-first-search (BFS) ordering starting from u_i. Second, employ the pre-selected MIS denoted by $\mathcal{M}$ as the initial active node set to perform the greedy algorithm called MPINS-GREEDY as shown in Algorithm 1. MPINS-GREEDY starts from $\mathcal{I} = \mathcal{M}$. Each time, it adds the node having the maximum $f(\cdot)$ value into $\mathcal{I}$. The algorithm terminates when $f(\mathcal{I}) = |\mathcal{V}|\theta$.

To better understand the proposed algorithm, we use the social network represented by the graph shown in Fig. 2(a) to illustrate the selection procedure as follows. In the

[3] MIS can be defined formally as follows: given a graph $G = (V, E)$, an Independent Set (IS) is a subset $I \subset V$ such that for any two vertex $v_1, v_2 \in I$, they are not adjacent, *i.e.*, $(v_1, v_2) \notin E$. An IS is called an MIS if we add one more arbitrary node to this subset, the new subset will not be an IS any more.

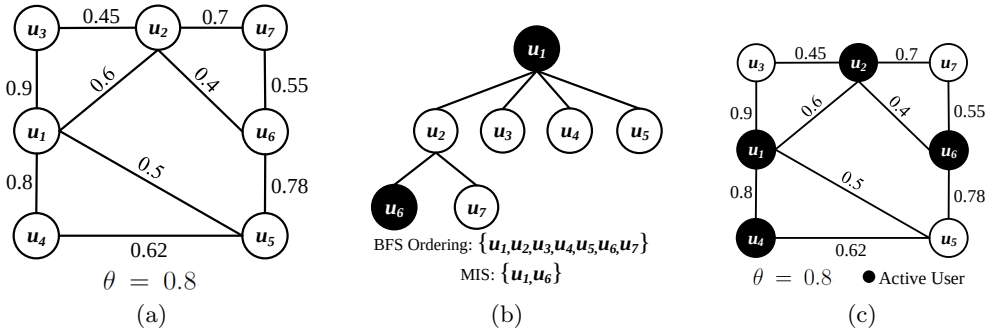


Fig. 2: Illustration of MPINS-Greedy algorithm.

Algorithm 1 MPINS-GREEDY Algorithm

Require: A social network represented by graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$; a pre-defined threshold θ.
1: Initialize $\mathcal{I} = \mathcal{M}$
2: **while** $f(\mathcal{I}) < |\mathcal{V}|\theta$ **do**
3: choose $u \in \mathcal{V} \setminus \mathcal{I}$ to maximize $f(\mathcal{I} \bigcup \{u\})$
4: $\mathcal{I} = \mathcal{I} \bigcup \{u\}$
5: **end while**
6: **return** $\mathcal{I}$

example, $\theta = 0.8$. Since u_1 has the maximum $f(\{u_i\})$ value, we construct a BFS tree rooted at u_1, as shown in Fig. 2(b). with the help of the BFS ordering, we find the MIS set which is $\mathcal{M} = \{u_1, u_6\}$. Next, we go to the second phase to perform Algorithm 1. 1) First round: $\mathcal{I} = \mathcal{M} = \{u_1, u_6\}$. 2) Second round: we first compute $f(\mathcal{I} = \{u_1, u_2, u_6\}) = 4.45$, $f(\mathcal{I} = \{u_1, u_3, u_6\}) = 3.018$, $f(\mathcal{I} = \{u_1, u_4, u_6\}) = 3.65$, $f(\mathcal{I} = \{u_1, u_5, u_6\}) = 3.65$, and $f(\mathcal{I} = \{u_1, u_6, u_7\}) = 3.778$. Therefore, we have $\mathcal{I} = \{u_1, u_2, u_6\}$, which has the maximum $f(\mathcal{I})$ value. However, $f(\mathcal{I} = \{u_1, u_2, u_6\}) = 4.45 < 7*0.8 = 5.6$. Consequently, the selection procedure continues. 3) Third round: we first computer $f(\mathcal{I} = \{u_1, u_2, u_3, u_6\}) = 4.45$, $f(\mathcal{I} = \{u_1, u_2, u_4, u_6\}) = 5.6$, $f(\mathcal{I} = \{u_1, u_2, u_5, u_6\}) = 5.6$, and $f(\mathcal{I} = \{u_1, u_2, u_6, u_7\}) = 4.45$. Therefore, we have $\mathcal{I} = \{u_1, u_2, u_4, u_6\}$[4]. Since $f(\mathcal{I} = \{u_1, u_2, u_4, u_6\}) = 7*0.8 = 5.6$, algorithm terminates and outputs set $\mathcal{I} = \{u_1, u_2, u_4, u_6\}$ as shown in Fig. 2(c), where black nodes represent the selected influential nodes.

Based on Algorithm 1, in each iteration, only one node is selected to be added into the output set $\mathcal{I}$. In the worst case, all nodes are added into $\mathcal{I}$ in the $|\mathcal{V}|$-th iteration. Then, $f(\mathcal{I}) = f(\mathcal{V}) = |\mathcal{V}|\theta$ and Algorithm 1 terminates and outputs $\mathcal{I} = \mathcal{V}$. Therefore, Algorithm 1 terminates for sure. Also, if $f(\mathcal{I}) = |\mathcal{V}|\theta$, then $\forall u_i \in \mathcal{V}$, $\varrho^{\mathcal{I}}(u_i) \geq \theta$ followed by Definition 9. Therefore, all nodes in the network are positively influenced. In another side, if $\forall u_i \in \mathcal{V}, \varrho^{\mathcal{I}}(u_i) \geq \theta$, then we obtain $\forall u_i \in \mathcal{V}, \min(\varrho^{\mathcal{I}}(u_i), \theta) = \theta$. Therefore, Algorithm 1 must produce a feasible solution of the MPINS selection problem.

5 Performance Evaluation

Since there is no existing work studying the MPINS problem under the independent cascade model currently, in the real data experiments, the results of MPINS-GREEDY (MPINS) are compared with the most related work [22] (PIDS), and the optimal solution of MPINS

[4] If there is a tie on the $f(\mathcal{I})$ value, we use the node ID to break the tie.

(OPTIMAL) which is obtained by exhausting searching. To ensure the fairness of comparison, the condition of termination to the algorithm proposed in [22] is changed to find a PIDS, such that every node in the network is positively influenced no less than the same threshold θ in MPINS. All experiments were performed on a desktop computer equipped with Inter(R) Core(TM) 2 Quad CPU 2.83GHz and 6GB RAM.

5.1 Experimental Setting

We also implement experiments run on different kinds of real-world data sets. The first group of data sets are shown in Table 1 come from SNAP [5]. The network statistics are summarized by the number of nodes and edges, and the diameter (*i.e.*, longest shortest path). The data collected in Table 1 is based on the *Customers Who Bought This Item Also Bought* feature of the Amazon website. Four different networks are composed of the data collected from March to May in 2003 in Amazon. In each network, for a pair of nodes (products) i and j, there is an edge between them if and only if a product i is frequently co-purchased with product j [33]. Besides the Amazon product co-purchasing data sets shown in Table 1, we also evaluate our algorithm in the additional real data sets listed as follows:

Table 1: Data Set 1 in Our Experiment

Data	Nodes	Edges	Diameter
A1	262111	1234877	29
A2	400727	3200440	18
A3	410236	3356824	21
A4	403394	3387388	21

1. *WikiVote*: a data set obtained from [34], which contains the vote history data of Wikipedia.The data set includes 7115 vertices and 103689 edges which contains the voting data of Wikipedia from the inception till January 2008. If user i voted on user j for the administrator election, there will be an edge between i to j.
2. *Coauthor*: a data set obtained from [35], which hold the coauthors information maintained by ArnetMiner. We chosen the subset which include 53442 vertices and 127968 edges. When the author i has a relationship with author j, there will be one edge between i to j.
3. *Twitter*: a data set obtained from [36, 37], which stores the information collected from Twitter.We picked the subset with 92180 vertices and 188971 edges, which represent the user account and their relationships.

Moreover, the social influence on each edge (i, j) is calculated by $\frac{1}{deg(j)}$ [38], where $deg(j)$ is the degree of node j. Similarly, if one node is selected as the active node, it has positively influence on all its neighbors. Otherwise, it only has negative influence on its neighbors.

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

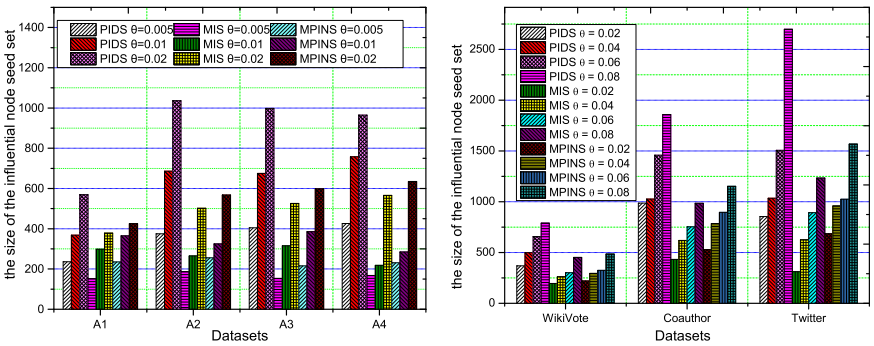


Fig. 3: Size of influential nodes in (a) Amazon, (b) WikiVote, Coauthor, and Twitter.

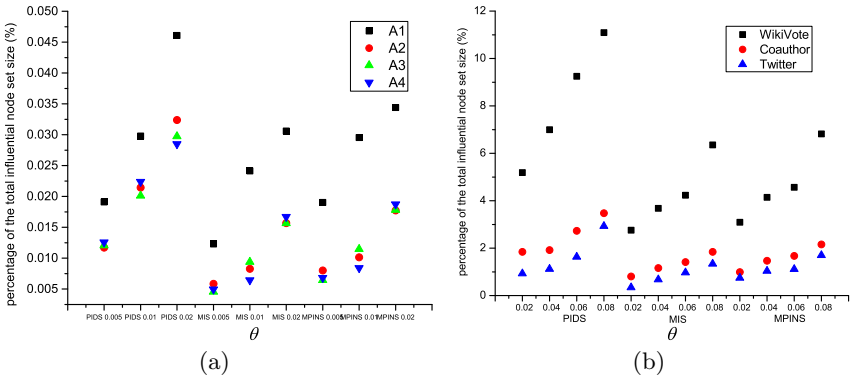


Fig. 4: % of influential nodes in (a) Amazon (b) WikiVote, Coauthor, and Twitter.

Experimental results The impacts of θ on the size of MIS, the solutions of MPINS, and the solution of PIDS on Amazon co-purchase data sets, when θ change from 0.005 to 0.02, are shown in Fig. 3(a). As shown in Fig. 3(a), the solution sizes of PIDS and MINS increase when θ increases. This is because, when the pre-set threshold becomes large, more influential nodes are required to be chosen to influence the whole network. On average, the difference between the size of PIDS and MPINS solutions is 37.23%. This is because that MPINS chooses the most influential node first instead of the node with the largest degree first. Moreover, the growth rate of the solution size of PIDS is higher than that of MPINS.

Similarly, the impacts of θ on the size of MIS, the solutions of MPINS, and the solution of PIDS on WikiVote, Coauthor, and Twitter, when θ change from 0.02 to 0.08, are shown in Fig. 3(b). The solution sizes of PIDS and MINS increase when θ increases as well. For the Twitter data set, MPINS outperforms PIDS significantly, *i.e.*, MPINS selects 45.45% less influential nodes than that of PIDS. On average, the difference between the sizes of PIDS and MPINS solutions is 36.37%.

Fig. 4 shows how many nodes are selected as the influential nodes represented by the ratio over the total number of nodes in the network. Fig. 4 (a) shows the impacts of θ on the ratio of MIS, MPINS, and PIDS on Amazon co-purchase data sets. While, Fig. 4 (b) shows the the impacts of θ on the ratio of MIS, MPINS, and PIDS on WikiVote, Coauthor, and Twitter data sets. One interesting observation here is that much less nodes

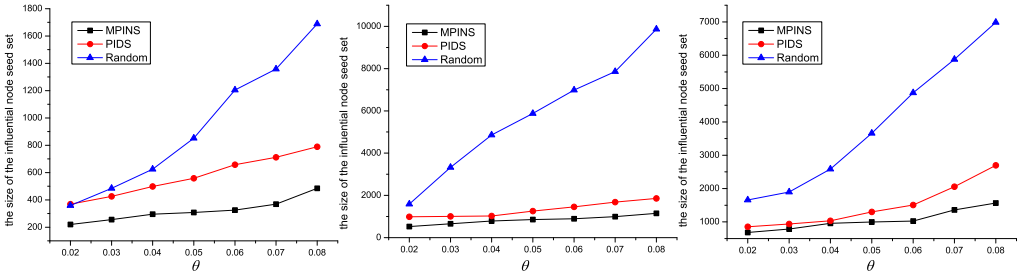


Fig. 5: MPINS VS. PIDS VS. Random in (a) WikiVote (b) Coauthor (c) Twitter

are selected as the influential nodes for Amazon co-purchase data sets compared to the WikiVote, Coauthor, and Twitter data sets.

Finally, we compare the performance of our proposed method MPINS with PIDS and the method denoted by "Random", which randomly chooses a node as the influential node. The impacts of θ on the sizes of the solutions of MPINS, PIDS, and Random, when θ change from 0.02 to 0.08, are shown in Fig. 5 for the WikiVote, Coauthor data set, and Twitter data sets. As shown in Fig. 5, the solution sizes of Random, PIDS and MPINS increase when n increases. Moreover, for a specific θ, MPINS produces a smaller influential node set than PIDS. This is consistent with the simulation results and previous experimental results. Furthermore, both PIDS and MPINS produce much smaller influential node sets than Random for a specific θ. This is because Random picks node randomly without any selection criterion. However, PIDS's selection process is based on degree and our MPINS greedy criterion is based on social influence.

From the results of experiments on real-world data sets, we can conclude that the size of the constructed initial active node set of MPINS is smaller than that of PIDS. Moreover, the solution of MPINS is very close to the optimal solutions in small scale networks.

6 Conclusion

In this paper, we study the Minimum-sized Positive Influential Node Set (MPINS) selection problem in social networks. We show by reduction that MPINS is APX-hard under the Independent Cascade Model. Subsequently, a greedy algorithm is proposed to solve the problem. Furthermore, we validate our proposed algorithm through simulations on random graphs and experiments on seven different real-world data sets. The simulation and experimental results indicate that MPINS-GREEDY can construct smaller sized satisfied initial active node sets than the latest related work PIDS. Moreover, for small scale network, MPINS-GREEDY has very similar performance as the optimal solution of MPINS. Furthermore, the simulation and experimental results indicate that MPINS-GREEDY considerably outperforms PIDS in medium and large scale networks, sparse networks, and for high threshold θ.

Acknowledgment

This research is funded in part by the Kennesaw State University College of Science and Mathematics Interdisciplinary Research Opportunities (IDROP) Program, the Provincial

Key Research and Development Program of Zhejiang, China under No. 2016C01G2010916, the Fundamental Research Funds for the Central Universities, the Alibaba-Zhejiang University Joint Research Institute for Frontier Technologies (A.Z.F.T.) under Program No. XT622017000118, and the CCF-Tencent Open Research Fund under No. AGR20160109.

References

1. 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.
2. K. Saito, M. Kimura, H. Motoda, Discovering influential nodes for sis models in social networks, in: International Conference on Discovery Science, Springer, 2009, pp. 302–316.
3. Y. Li, W. Chen, Y. Wang, Z.-L. Zhang, Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships, in: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013, pp. 657–666.
4. M. Han, M. Yan, Z. Cai, Y. Li, X. Cai, J. Yu, Influence maximization by probing partial communities in dynamic online social networks, Transactions on Emerging Telecommunications Technologies.
5. 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.
6. W. Lu, F. Bonchi, A. Goyal, L. V. Lakshmanan, The bang for the buck: fair competitive viral marketing from the host perspective, in: Proceedings of the 19th SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 928–936.
7. A. S. U. R. B. Shouling Ji, Jing (Selena) He, Y. Li, Cell-based snapshot and continuous data collection in wireless sensor networks, ACM Transactions on Sensor Networks (TOSN) 9 (4).
8. M. Y. Y. P. Jing (Selena) He, Shouling Ji, Y. Li, Genetic-algorithm-based construction of load-balanced cdss in wireless sensor networks, MILCOM 9 (4).
9. H. Albinali, M. Han, J. Wang, H. Gao, Y. Li, The roles of social network mavens, The 12th International Conference on Mobile Ad-hoc and Sensor Networks.
10. A. Goyal, F. Bonchi, L. V. Lakshmanan, Learning influence probabilities in social networks, in: Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, pp. 241–250.
11. M. Han, M. Yan, J. Li, S. Ji, Y. Li, Generating uncertain networks based on historical network snapshots., in: COCOON, Springer, 2013, pp. 747–758.
12. 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.
13. M. Han, M. Yan, Z. Cai, Y. Li, An exploration of broader influence maximization in timeliness networks with opportunistic selection, Journal of Network and Computer Applications 63 (2016) 39–49.
14. C. Wang, W. Chen, Y. Wang, Scalable influence maximization for independent cascade model in large-scale social networks, Data Mining and Knowledge Discovery 25 (3) (2012) 545.
15. M. Han, Z. Duan, C. Ai, F. W. Lybarger, Y. Li, A. G. Bourgeois, Time constraint influence maximization algorithm in the age of big data, International Journal of Computational Science and Engineering.
16. J. Tang, S. Wu, J. Sun, Confluence: Conformity influence in large social networks, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 347–355.
17. K. Saito, R. Nakano, M. Kimura, Prediction of information diffusion probabilities for independent cascade model, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2008, pp. 67–75.

18. J. Tang, J. Sun, C. Wang, Z. Yang, Social influence analysis in large-scale networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 807–816.
19. M. Han, M. Yan, J. Li, S. Ji, Y. Li, Neighborhood-based uncertainty generation in social networks, Journal of Combinatorial Optimization 28 (3) (2014) 561–576.
20. C. Wang, J. Tang, J. Sun, J. Han, Dynamic social influence analysis through time-dependent factor graphs, in: Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, IEEE, 2011, pp. 239–246.
21. F. Wang, E. Camacho, K. Xu, Positive influence dominating set in online social networks, in: International Conference on Combinatorial Optimization and Applications, Springer, 2009, pp. 313–321.
22. F. Wang, H. Du, E. Camacho, K. Xu, W. Lee, Y. Shi, S. Shan, On positive influence dominating sets in social networks, Theoretical Computer Science 412 (3) (2011) 265–269.
23. X. Zhu, J. Yu, W. Lee, D. Kim, S. Shan, D.-Z. Du, New dominating sets in social networks, Journal of Global Optimization 48 (4) (2010) 633–642.
24. J. S. He, S. Ji, R. Beyah, Z. Cai, Minimum-sized influential node set selection for social networks under the independent cascade model, in: Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing, ACM, 2014, pp. 93–102.
25. H. Kaur, J. S. He, Blocking negative influential node set in social networks: From host perspective, Transactions on Emerging Telecommunications Technologies (ETT) 28 (4).
26. D. Kempe, J. Kleinberg, É. Tardos, Influential nodes in a diffusion model for social networks, in: International Colloquium on Automata, Languages, and Programming, Springer, 2005, pp. 1127–1138.
27. 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.
28. W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, 2010, pp. 88–97.
29. M. Han, J. Li, Z. Cai, H. Qilong, Privacy reserved influence maximization in gps-enabled cyber-physical and online social networks, SocialCom 2016 (2016) 284–292.
30. M. Han, Q. Han, L. Li, J. Li, Y. Li, Maximizing influence in sensed heterogenous social network with privacy preservation, International Journal of Sensor Networks.
31. D.-Z. Du, K.-I. Ko, Theory of computational complexity, Vol. 58, John Wiley & Sons, 2011.
32. Technical report.
URL `http://ksuweb.kennesaw.edu/~jhe4/Research/MPINS.`
33. J. Leskovec, L. A. Adamic, B. A. Huberman, The dynamics of viral marketing, ACM Transactions on the Web (TWEB) 1 (1) (2007) 5.
34. J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting positive and negative links in online social networks, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 641–650.
35. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 990–998.
36. J. Hopcroft, T. Lou, J. Tang, Who will follow you back?: reciprocal relationship prediction, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 1137–1146.
37. T. Lou, J. Tang, J. Hopcroft, Z. Fang, X. Ding, Learning to predict reciprocity and triadic closure in social networks, ACM Transactions on Knowledge Discovery from Data (TKDD) 7 (2) (2013) 5.
38. C. Wang, W. Chen, Y. Wang, Scalable influence maximization for independent cascade model in large-scale social networks, Data Mining and Knowledge Discovery 25 (3).