Sealing Technology
ISSN: 1350-4789

FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis

Nagam Aanjaneyulu, Lankala Mounika, Mr. Merugu Anand Kumar,

Dr. G. Samba Siva Rao

[1,2,3] Assistant Professor, [4] Professor

anji.amrexamcell@gmail.com, lankala.mounikareddy@gmail.com

meruguanand502@gmail.com, profgssrao@gmail.com

Department of CSE, A M REDY MEMORIAL COLLEGE OF ENGINEERING AND TECHNOLOGY,

PETLUVARI PALEM, ANDHRA PRADESH-522601

Abstract:

Media like as social networks play a crucial role in the dissemination of knowledge, ideas, and sway among people. Understanding the properties of social networks, learning how information spreads via the "word-of-mouth" impact of social networks, and learning about the social effects among individuals are the primary areas of study in the extant literature. Persons and communities alike. However, most studies don't account for the presence of destructive influences between people. To combat social ills like excessive drinking, smoking, and gambling, as well as influence-spreading issues like the promotion of new products, we take both positive and negative influences into account and propose a new optimization problem called the Minimum-sized Positive Influential Node Set (MPINS) selection problem to find the smallest group of nodes from which every other node in the network can benefit. Our help here is threefold. In the first place, we show that MPINS is APX-hard when seen as an independent cascade model with both positive and negative impacts. The MPINS selection issue is then addressed by a greedy approximation approach that we provide. Finally, we run extensive simulations and experiments on random graphs and seven different real-world data sets that represent small-, medium-, and large-scale networks to verify the efficacy of the proposed greedy algorithm.

1 Introduction

Like Facebook, Google+, and MySpace, social networks are made up of "nodes," or entities, that all have something in common. The social network is an effective means of communication for distributing information and gaining followers outside of one's immediate social circle. Since their inception, social networks have greatly widened our spheres of influence and served as a conduit between our offline lives and the online world. Massive attention has been paid to how social networks may be used efficiently to disseminate ideas or information within a community since the advent of social apps (such as Flickr, Wikis, Netflix, and Twitter, etc.) [1-6]. Understanding the positive and negative social impacts resulting from interactions between individuals and between groups is essential to solving The challenging challenge of capturing the dynamics of a social network. It's possible for members of a social network to have both good and negative effects on one another. A gaming insulator, for instance, would have a beneficial impact on his social circle and community as a whole. The favorable effect is compounded if a large number of a person's friends are also battling against the want to gamble. But a person runs the risk of becoming an addicted gambler who harms his social circle. In Fig. 1's social network, for instance, edge weights indicate the social impacts present in the network. Jack and Bob (represented by the individual with the red tie) may be good neighbors if they prevent their friends and family from being addicted to gambling. To be more precise, Jack is 60% likely to have a favorable impact on Chris. Mary's negative impact on Tony is 90% more likely given that she is a gambler. And among the people shown in Fig. 1, only Tony is completely

immune to the gambling culture. In order to reduce harmful social behaviors including excessive drinking, smoking, and gambling, this study seeks to identify a set of positively influential nodes (called an MPINS) that can reach every member of a network and have an effect on them of at least Among the many possible uses for MPINS are: Take the case of a town that plans to launch a smoking cessation initiative. The community hopes to pick a limited number of powerful members of the community who will attend a quit-smoking campaign in order to assure cost-effectiveness and acquire the greatest impact. The objective is for the chosen users to have a beneficial impact on the rest of the community. The aforementioned social issue may be mitigated and new items promoted in the social network if an MPINS is built.

Another situation is provided as a source of inspiration: One tiny business's goal is to promote its latest merchandise in a group setting. The goal of the company's sample product distribution to a select group of customers is to minimize costs while maximizing earnings.

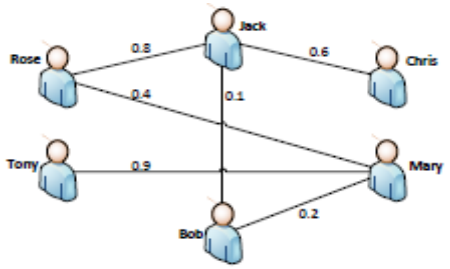


Figure 1: An example of a social network with peer pressure along the vertices.

The corporation is banking on the fact that these Users will have a pleasant experience and encourage others to buy the product. No less than _ of the people in the community should be able to have a lasting, positive impact on the lives of the people who utilize the community's services. In conclusion, the following narrow issue is what we look into: Given a social network and a threshold of _, find the smallest subset of its members that may have a net positive effect on no more than _ other members of the network. To ensure that every other node has at least half of its neighbors in D, researchers in a previous study [7] determined a minimum size for the Positive Influence Dominating Set (PIDS). In that study, we solely looked at the beneficial effects of having close neighbors and completely disregarded the drawbacks. The authors in Ref. [7] also looked at the PIDS selection problem in the context of the deterministic linear threshold model, where the weight between two nodes represents the influence between them, and an individual is positively influenced when the sum of the weights exceeds a given threshold. According to the authors in Ref. [7], the influence between any two nodes is always assumed to be 1, and a node is considered to be positively impacted if at least half of its neighbors are also in D. Since the strength of social influence between different pairs of nodes may vary and is actually a probabilistic value in the physical world, the deterministic linear threshold model is unable to comprehensively characterize the social influence between each pair of nodes in an actual social network[9-13]. Therefore, we investigate the MPINS selection problem in the context of the independent cascade model, where individuals can have both positive and negative influences on their neighbors with varying probabilities.

2 Related Works

First, we provide a quick overview of the literature on social influence analysis. We next provide a brief overview of the research around the PIDS issue and the challenge of maximizing one's impact, followed by commentary.

2.1 Social influence analysis

Kempe et al. [1] first suggested the concept of influence maximization, which seeks to pick a group of users in a social network so as to maximize the predicted number of a given outcome. Impacted people through many intermediate stages of knowledge dissemination [14]. Influence learning [10, 15], algorithm optimization [16-18], scalability promotion [19-21], and the impact of group conformity [4, 22] have all been the subject of empirical research. Information diffusion probabilities in social networks were predicted by Saito et al. [23] using the independent cascade model. After explicitly defining the likelihood maximization issue, they used an EM method to find the optimal solution. It has been stated by Tang et al. [9, 24, and 25] that looking at social impact from various perspectives (subjects) may provide varying results. So, they came up with TAP (Topic Affinity Propagation) to simulate the spread of information in massive social networks based on shared interests. In order to account for the passage of time in the examination of shifting social impacts, Wang et al. [11] devised a Dynamic Factor Graph (DFG) model. Learning impact probabilities from past node activities is an issue that Goyal et al. [10] also investigated.

2.2 Positive influence dominating set problem

Under the deterministic linear threshold model, Wang et al. [26] first proposed the PIDS problem, which is to locate a set of nodes D such that each node in D is connected to all the other nodes in D. At least half of the nodes in a network's neighbors are located in D. A selection method was created, and its efficacy was evaluated using data from actual social networks. Subsequently, Wang et al. [7, 27] used approximation ratio analysis to demonstrate that PIDS is APX-hard, and they proposed two greedy algorithms. Using the term "Minimum-sized Influential Node Set" (MINS), he and his colleagues [28] developed a novel optimization issue. The purpose of this task is to find the smallest collection of influential nodes such that all other nodes may be impacted by them by at most some fixed threshold. However, they failed to account for the fact that adverse factors do exist.

2.3 Influence maximization problem

The node selection challenge in social network information dissemination was initially highlighted by Domingos and Richardson [29, 30]. Taking into account people' social connections, they offered a probabilistic information transmission model and many heuristic approaches to the issue. The impact maximization issue was subsequently articulated by Kempe et al. [1, 31], who went on to investigate it in the context of two models—the linear threshold model and the independent cascade model. In both cases, they examined the suggested greedy algorithms and found that their performance ratios were 1 1e. Leskovec et al.[32] proposed a "lazy-forward" optimization strategy of picking beginning nodes, which drastically cut down on the amount of impact spread assessments, thereby solving the scalability issue of the algorithms in Ref. [1, 31]. Both models of #P-Hard were presented by Chen et al. [33, 34], along with their proposed scalable algorithms that are significantly faster than the greedy algorithms in Refs. [1, 31]. Recently, Refs [35-37] suggested approaches to give a holistic solution to the issue of influence maximization by taking into account data from both the cyber-physical environment and online social networks.

However, the influence maximization problem was looked at by Goyal et al.[38] from a statistical point of view. Credit distribution is a novel model that directly uses existing propagation traces to understand how power is distributed in a network and to make predictions about that distribution. The authors also developed an approximation approach and demonstrated that the influence maximization issue under the credit distribution model is APX-hard. The rapid information propagation issue was introduced by Zou et al.[39], who were the first to add the latency restriction to the influence maximization problem under the linear threshold model. Fast information propagation was also shown to be APX-hard in Ref. [40]. In addition, two heuristic methods are provided, and their relative performance is discussed. In contrast to prior research on maximizing or minimizing social impact, Zhang et al. [41] investigated influence coverage with probabilistic assurances rather than predicted influence coverage guarantees. In Ref. [41], the authors propose a novel optimization problem, dubbed Seed Minimization with Probabilistic Coverage Guarantee (SM-PCG), provide a thorough theoretical analysis, and provide practical findings that support the efficacy of the corresponding method.

2.4 Remarks

The aforementioned canonical works may be divided into three classes: the study of the features and qualities of social networks, including but not limited to social influences; investigating the lately popular influence maximization issue (with or without a time limitation) and solving the PIDS conundrum. However, when modeling social networks, none of the aforementioned works took negative influence into account. Our work differs from the influence maximization problem not only because we consider both positive and negative influences, but also because we find a subset of individuals of minimum size k that maximizes the expected number of influenced individuals while still guaranteeing positive influences on every node in the network with no less than a threshold of In addition, we have a unique approach to the PIDS issue that sets us apart.

3 Problem Definitions and Hardness Analysis

The network model is presented first in this chapter. The MPINS selection issue is then clearly defined, and some commentary is included on the suggested solution problem. Finally, we investigate how challenging the MPINS selection issue is.

The Network Model, Version 3.1

We use the undirected graph G.V; E; P.E// to represent a social network, where V is the collection of n nodes represented by ui and 06 in. i is referred to

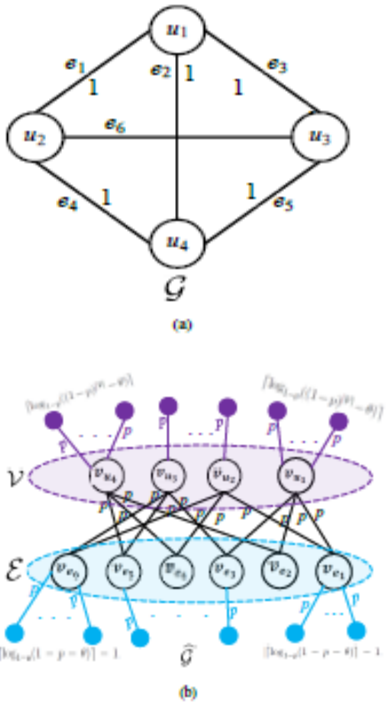
as ui's node ID. A line with no destination. A social connection between the it and jth nodes is denoted by ui; uj / 2 E. Where pij represents the social influence between nodes ui and uj _, the formula reads: P.E/ D fpij j if.ui; uj / 2 E; 0 pij 6 1; else pij D 0g. It's important to note that there are two types of social influence: constructive and maladaptive. For the smoking intervention program, for instance, a neighbor who decides to attend a quit-smoking campaign has a positive effect on all of their other neighbors, while smokers have a negative effect on theirs. Definitions 5 and 6 of Section 3.2 provide the formal meanings of positive and negative impact, respectively. For the sake of brevity, let's suppose that all of the linkages are undirected (bidirectional), meaning that any two nodes that are connected by a link have the same level of social influence (pij value).

3.2 Identifying the Issue

Finding an initial set of nodes in a social network that may favorably impact all the other nodes with a threshold of _ is the goal of the Maximum Possible impact Node Set (MPINS) selection issue. The first nodes that were chosen are referred to as active nodes for simplicity. Therefore, understanding how to define beneficial influence is essential for resolving the MPINS selection issue. We begin by formally defining certain terms, and then we define the MPINS selection issue.

3.3 Evaluation of Problem Difficulty

The MPINS selection issue is APX-hard in general, for any _. By creating an L-reduction from the Vertex Cover problem in a Cubic Graph (VCCG) to the MPINS selection problem, we show that MPINS is APXhard. Proof that the VCCG decision issue is APX-hard may be found in Ref. [42]. A cubic graph is a graph in which the degree of each vertex is 3. Finding the smallest possible vertex cover for a given cubic graph is the goal of VCCG. Let's start with a simple example of VCCG, a cubic graph with the formula P.E/ D fl j.ui ; uj / 2 EI ui ; uj 2 Vg. Here's how we make the new graph bG:



The transition from G to BG G is seen in Fig. 2.

4 Proof of Lemma 1

The proof is as follows: if G has a Vertex Cover (VC) D of size at most d, then there exists an extra set I in bG which consists of: _

$|\mathcal{V}|\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil$ Active nodes, $\bigcup_{i=1}^{|\mathcal{V}|} v_{u_i}^{\mathcal{A}}$

_ All the additional $|\mathcal{E}|\lceil\log_{1-p}(1-P-\theta)\rceil-1$, $\bigcup_{j=1}^{|\mathcal{E}|} v_{e_j}^{\mathcal{A}}$

In the bottom shading of Fig. 2b, all the nodes vui stand in for the nodes ui;

VC $\mathcal{D}$ in $\mathcal{G}$, i.e., $\{v_{u_i} \mid u_i \in \mathcal{D} \text{ in } \mathcal{G}\}$.

Therefore, we have $|\mathcal{I}| = |\mathcal{V}|\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil + |\mathcal{E}|(\lceil\log_{1-p}\theta\rceil-1)+d \leqslant k.$

Now, we need to check whether I satisfy

$$\forall v_k \in \widehat{\mathcal{G}}, \varrho^{\mathcal{I}}(v_k) = p_{v_k}(\mathcal{A}^{\mathcal{I}}(v_k)) - p_{v_k}(\mathcal{N}^{\mathcal{I}}(v_k)) \geqslant \theta.$$

For an inactive node vui 2 vui , because it connects to $\lceil\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)\rceil$ Active nodes

$v_{u_i}^A = \{v_{u_i}^j \mid 1 \leqslant j \leqslant \lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta)\rceil\}$, we have $\varrho^{\mathcal{I}}(v_{u_i}) = p_{v_{u_i}}(\mathcal{A}^{\mathcal{I}}(v_{u_i})) - p_{v_{u_i}}(\mathcal{N}^{\mathcal{I}}(v_{u_i})) = [1-(1-p)^{\log_{1-p}((1-p)^{|\mathcal{V}|}-\theta)}] - [1-(1-p)^{d_i}] \geqslant (1-p)^{d_i} - (1-p)^{|\mathcal{V}|} + \theta \geqslant \theta$,

Where di is the degree of a node and vui is the resulting graph

5 Proof of Theorem 1

Proof The first lemma proves right away that G has at least an OPTVCCG-size vertex cover. If bG has some minimal positive influence size, then G/. Set size of nodes

$$\mathrm{OPT}_{\mathrm{MPINS}}(\widehat{\mathcal{G}}) = |\mathcal{V}|\lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta)\rceil + |\mathcal{E}|(\lceil \log_{1-p}(1-p-\theta)\rceil - 1) + \mathrm{OPT}_{\mathrm{VCCG}}(\mathcal{G}) \quad (1)$$

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

$$\frac{|\mathcal{V}|}{2} = \frac{|\mathcal{E}|}{3} \leqslant \mathrm{OPT}_{\mathrm{VCCG}}(\mathcal{G}) \quad (2)$$

On the basis of Lemma 1, plugging

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

Into Formula (2), we have

$$\mathrm{OPT}_{\mathrm{MPINS}}(\widehat{\mathcal{G}}) \leqslant [2\lceil \log_{1-p}((1-p)^{|\mathcal{V}|} - \theta)\rceil + 3\lceil \log_{1-p}(1-p-\theta)\rceil - \frac{1}{\gamma}]\mathrm{OPT}_{\mathrm{VCCG}}(\mathcal{G}) \quad (4)$$

This proves that MPINS is an L-reduction of VCCG. To sum up, we demonstrated that a subset of the MPINS selection issue is APX-hard. VCCG is an APX-hard issue. Since this is the case, we can conclude that the general MPINS selection problem is at least APXhard. Our analysis leads us to the conclusion that MPINS cannot be solved in polynomial time, as shown by Theorem 1. Thus, in the following section, we propose a greedy algorithm to address this issue.

6 Greedy Algorithms and Performance

We present a greedy technique to address the fact that Analysis MPINS is APX-hard. For the planned MPINS-GREEDY is the name of the algorithm. We define a practical contribution function as follows before introducing MPINS-GREEDY: Function of contribution (.I/) (f). The contribution function of a collection of influential nodes I to a social network G represented by graph G.V; E; P.E// is defined as

$$f(\mathcal{I}) = \sum_{i=1}^{|\mathcal{V}|} \max\{\min(\varrho^{\mathcal{I}}(u_i), \theta), 0\}.$$

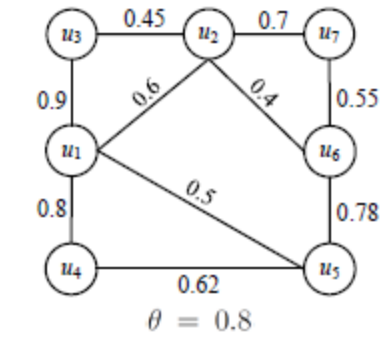
We present a two-stage heuristic approach based on the specified contribution function. To begin, we identify ui, the node with the highest f.I/. In which D fui g I am. After that, we use a Breadth-First-Search (BFS) ordering that begins with up to choose a Maximal Independent Set (MIS). Second, in Algorithm 1, the set of active nodes for MPINS-GREEDY is initially comprised of the pre-selected MIS, designated by M. MPINS-GREEDY originates in the I DM.
Every time, it incorporates into I the node whose f._/ value is highest. When f.I/ D jVj_, the algorithm ends.

Algorithm 1 MPINS-GREEDY Algorithm

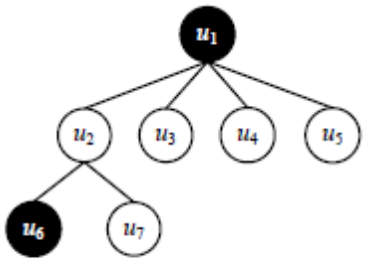
Require: A social network represented by graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{P}(\mathcal{E}))$; a pre-defined threshold θ.

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** $\mathcal{I}$

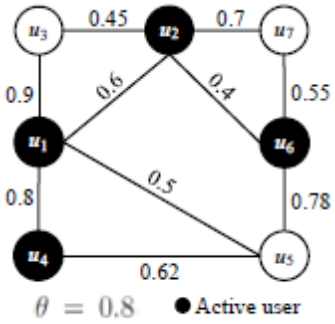
The formal definition of MIS is as follows: An Independent Set (IS), for a graph G D.V; E/, is a subset I _ V in which no two vertices v1; v2 are neighbors. Managing Information System The set is no longer an IS if we insert a single more node at random into it.



(a)



(b)



(c)

The MPINS-Greedy algorithm is shown in Fig. 3.

It's simple to verify that values for u3, u5, and u7 have been altered for the better. So, the artificial me is a feasible MPINS selection issue solution. In order to reduce convergence time, the suggested approach begins its search from the MIS set (M) rather than the empty set. The following theorem then demonstrates the soundness of Algorithm 1 theoretically. Conjecture 2 The MPINS selection issue is effectively solved by the first algorithm. To be more precise, (1) Algorithm 1 is guaranteed to end. Only if I is a collection of positively influencing nodes—that is, if every node (i.e., 8ui 2 V) is positively impacted by nodes in I by more than _—will the condition (2) hold.

7 Proof of Theorem 2

Proof In (1), one node is picked at random to be added to the final output set I based on Algorithm 1. Adding every node is the worst-case scenario. At the jVj-th repetition, into I. After that, I D V is returned as the final output of Algorithm 1.
Therefore, it is guaranteed that Algorithm 1 will end.

$$\text{For (2), } \Rightarrow \text{: if } f(\mathcal{I}) = |\mathcal{V}|\theta, \text{then } \forall u_i \in \mathcal{V}, \varrho^{\mathcal{I}}(u_i) \geqslant \theta$$

Followed by Definition 9. Therefore, all nodes in the network are positively influenced.

$$\Leftarrow \text{: if } \forall u_i \in \mathcal{V}, \varrho^{\mathcal{I}}(u_i) \geqslant \theta,$$

Then we obtain $\forall u_i \in \mathcal{V}, \min(\varrho^{\mathcal{I}}(u_i), \theta) = \theta.$

$$f(\mathcal{I}) = \sum_{i=1}^{|\mathcal{V}|} \max\{\min(\varrho^{\mathcal{I}}(u_i), \theta), 0\} = |\mathcal{V}|\theta.$$

Algorithm 1 must provide a workable answer to the MPINS selection issue based on these two criteria.

8 Performance Evaluations

Considering that no other study has investigated the MPINS selection issue using an autonomous cascade model, the simulation and experimental findings of We evaluate MPINS-GREEDY (denoted by MPINS), the most closely related work [7] (denoted by PIDS), and the ideal solution of MPINS (denoted by ideal) by conducting an exhaustive search. To make sure that all nodes in the network are favorably impacted by at least the same threshold of _ in MPINS, we modify the termination condition of the technique given in Ref. [7] to discover such a PIDS. We test our model and algorithm on both synthetic and real-world data to see how well it performs.
The Intel(R) Core(TM) 2 Quad CPU 2.83 GHz desktop PC with 6GB RAM was used for all simulations and tests.

8.1 Simulation results

8.1.1 Simulation setting

Based on the random graph model G.n; p/ D FG jG, we devised our own simulator to produce random graphs, each of which contains n nodes and an edge connecting any pair of nodes. Has a chance of pg of being created. The related social influence, 0 pij 6 1,

for the graph G D.V; E/ 2 G.n; p/; ui; uj 2 V and.ui; uj / 2 E is produced at random. It's important to note that there are two types of social influence: constructive and maladaptive. When one node is chosen to be the active node, it influences its neighbors for the better. Otherwise, it can have nothing but a destructive effect on its surrounding areas. One hundred examples are created for each configuration. The outcomes are an average of these 100 separate events. The simulation results for many cases are shown below.

8.1.2 Simulation results on random graphs

Both MPINS and PIDS aim to reduce to a minimum the size of the resulting subsets. Here, we verify the MPINS, PIDS, and PIDS solution sizes. Optimum in a wide variety of contexts and for nondeterministic graphs. The influence threshold _, the probability that an edge may be created in the random graph model G.n; p/, and the size of the network n are all variables in this simulation. Since we use exhaustive searching to locate the BEST MPINS solution, it is impossible to test on truly massive networks. Therefore, we begin by simulating smaller networks, ranging in size from 10 to 20 nodes. Figure 4 displays the obtained data. Figures 4a-c show the effects of n, p, and _ on the solution sizes of MPINS, PIDS, and OPTIMAL. All three methods provide solutions that grow in size as n rises, as seen in Figure 4a. This is because a larger network requires more effort to influence. Furthermore, PIDS generates a larger sized solution than MPINS does for a given network size. This is because, in each iteration, PIDS prioritizes the node with the greatest degree, but in MPINS, it is the node with the biggest f._/ value that is added after finding the most influential Maximal Independent Set (MIS) in the network. Some neighbors may have strong detrimental effects on the individuals in a social network, so a high degree is not always indicative of a strong ultimate influence. More nodes need to be added to the subset before they can exert influence over all the nodes in the entire network, but MPINS avoids the node selection bias in some specific regions by choosing a MIS first. The MPINS solution comes very near in size to the OPTIMAL outcome. Specifically, PIDS generates 3.75 times as many nodes as the OPTIMAL solution, while MPINS generates only 1.07 times as many. Results suggest that in small-scale networks, the proposed greedy algorithm MPINS-GREEDY may yield a solution that is extremely near to the OPTIMAL solution.

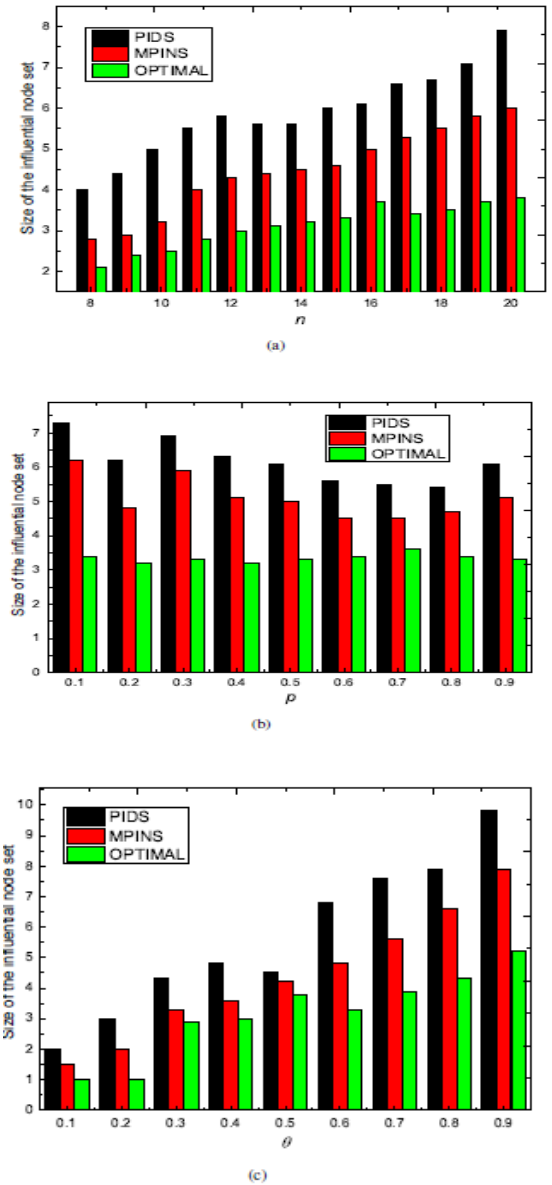


Figure 4: The Relative Size of Solutions on Local Area Networks. n = 15, p = 0.5, and _ = 0.5 are the default values.

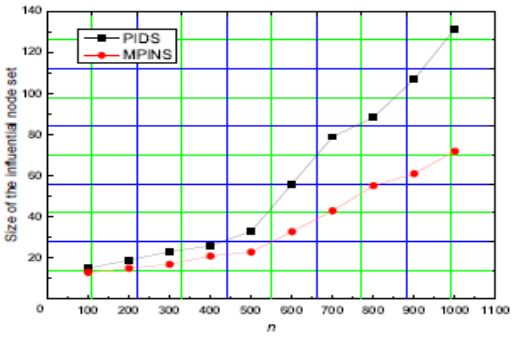
Fig. 4b shows that there is no discernible pattern. Because more edges in the network means that a given node may have more negative or positive neighbors, the solution sizes of all three algorithms grow as p grows. Distinguishing the typical size distribution of sets of chosen influential nodes in a dense network is challenging. Since the goal of PIDS is not to obtain the most influential and no-regionally-biased nodes in the network, for a given p, PIDS produces larger sized solutions. Once again, MPINS can build a solution that is as compact as the optimum. While PIDS generates an average of 3:16 more nodes than the OPTIMAL solution, MPINS generates an average of only 1:6 more nodes. When

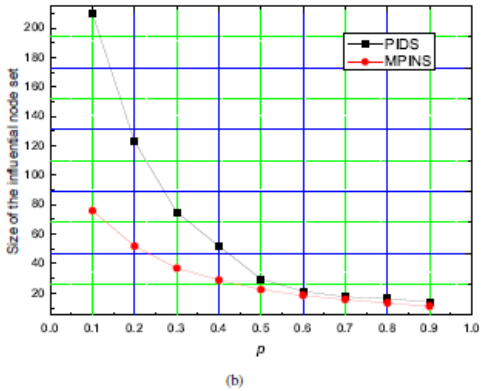
is large, more nodes must be included in the initial active node set in order to exert influence over all the other nodes, as shown in Figure 4c. This causes all the solutions to grow in size.

Since PIDS's greedy criterion prioritizes nodes with the highest degree first, MPINS performs similarly to OPTIMAL and better than PIDS. In general, MPINS generates 1–3 times as many nodes as the OPTIMAL solutions, while PIDS solutions are much smaller. Compared to the OPTIMAL method, PIDS generates an average of 3–7 more nodes. The same explanation as previously applies. We also conduct a series of simulations on medium-sized networks, varying the network size from 100 to 1000. In Fig. 5, we can see how n, p, and _ affect MPINS and PIDS. Larger social networks need more dynamic influential nodes, which are shown in Figure 5a's solution sizes for MPINS and PIDS. In addition, the gap between MPINS and PIDS sizes widens with increasing n. In a small-scale network (i.e., n 500), the size of the initial active node set is small (no more than 30 from Fig. 5), allowing MPINS to find a positive influential node set that is smaller than that of PIDS at a specific n. As a result, it might be difficult to distinguish between the two approaches. However, our proposed MPINS significantly expands the initial active node set compared to PIDS in a medium-sized network, where n 1000. The same explanation applies to this case as the one we gave before. MPINS generates a positive influential node set that is 22.5% less in size than PIDS on average. Figure 5b shows that when p grows, both the PIDS and MPINS solution sizes decrease. Increasing p suggests that more edges are present in the

Network grows, it follows that the average number of neighbors for each node also grows.

So, a single influential active node can affect the behavior of many





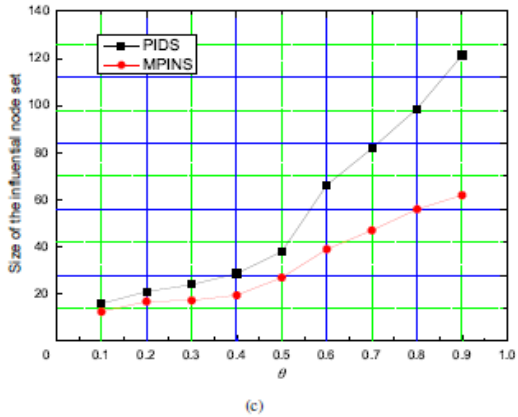


Figure 5: Solutions' Typical Size in Very Large Networks the default parameters are (n = 15), (p = 0.5), and (_ = 0.5).

Repeat PIDS for a given p. yields a bigger solution size than MPINS. Small solution sizes make it hard to tell which approach is superior. In sparse networks, however, such as p D 0:1, MPINS is demonstrably superior to PIDS. Because the degrees of all nodes are small when p is small, the decreasing trend of PIDS is very rapid when p is increased. Therefore, PIDS may iterate until a solution is found that guarantees a positive influence on every node in the network with a threshold of at least _. A positive influencing node set of modest size may be contributed to the solution after bigger degree nodes are added when p is large, which may cause PIDS to end sooner. Compared to MPINS, PIDS generates an average of 31.52% more nodes. Similarly to what was found for Fig. 4c, larger values of _ result in larger PIDS and MPINS solution sizes (as seen in Fig. 5c). In addition, when _ grows, PIDS produces more nodes than MPINS. When comparing PIDS with MPINS, the former generates an average of 23.2% more nodes. The greedy searching in MPINS begins on a hand-picked selection of highly important MIS nodes, while in PIDS, the set is initially empty.

In addition, PIDS's greedy search criteria, node degree, may to locating some regionally biased nodes in order to expand the solution overall. Our suggested MPINS approach initially chooses a MIS, thus it never faces the aforementioned conundrum. Exhibits 6-8

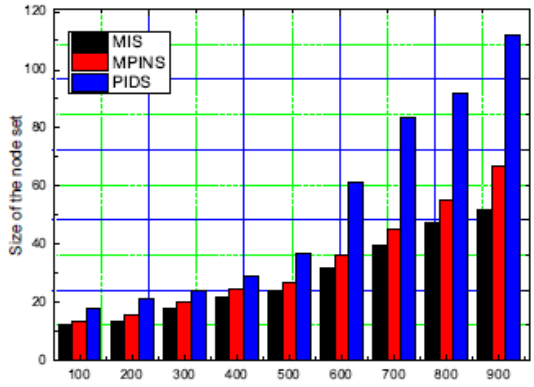
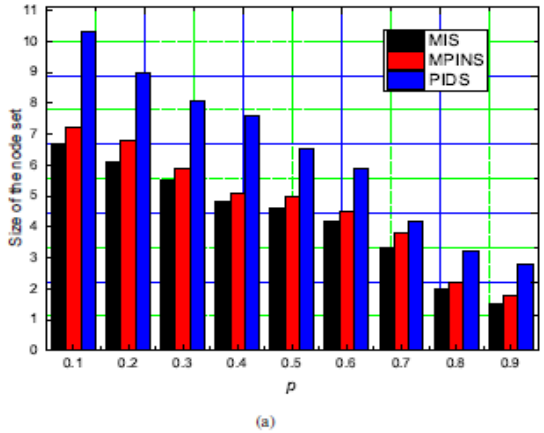
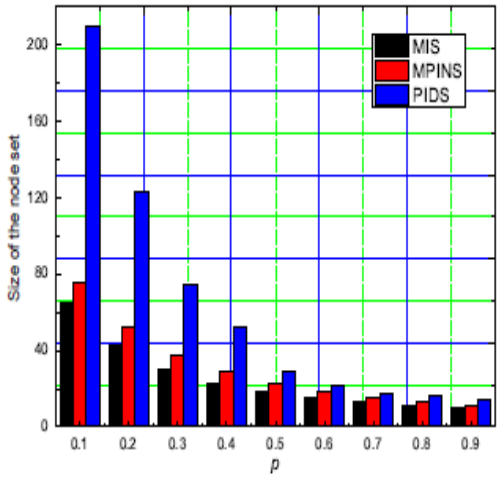


Fig. 6 Size of the node set: The default settings are p = 0.5 and _ = 0.5.

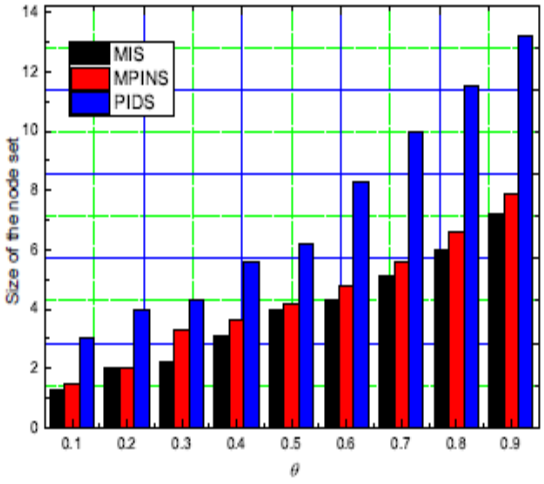


(a)

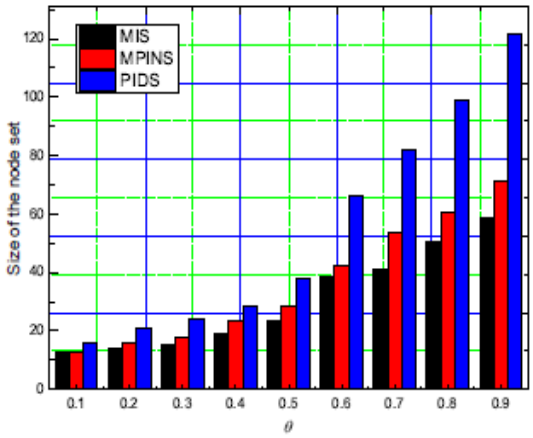


(b)

Fig. 7 Size of the node set: (a) n=20 and _ = 0.5; (b) n = 500 and _ = 0.5.



(a)



(b)

Comparison of network sizes in Fig. 8: (a) n = 20 and p = 0.5; (b) n = 500 and p = 0.5.

Changes in n, p, and _. Based on these findings, more MPINS-GREEDY iterations are likely not necessary. Sprint to discover an answer to MPINS after deciding on a powerful MIS. The suggested greedy method for solving PIDS, on the other hand, requires much more iterations than MPINS-GREEDY.

8.1.3 The findings of a simulation of a very big network

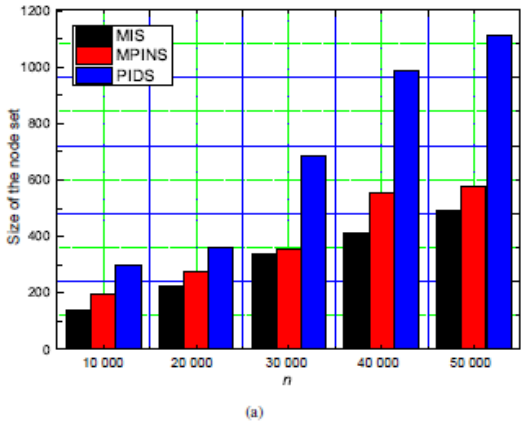
The growth of social media's user base has been meteoric. As a result, we do a number of simulations on very extensive systems. From 10,000 nodes, the network now supports 50,000. Figure 9 displays the effects of n, p, and _ on MPINS and PIDS. Figure 9a

demonstrates that as n grows, so do the sizes of MPINS and PIDS solutions. This growth arises because bigger social networks need a greater number of dynamic influential nodes. As n grows larger, there becomes a larger disparity between MPINS and PIDS. As can be shown in Fig. 9a, our suggested MPINS significantly improves the size of the first active node set compared with PIDS in a large-scale network, where n D 50,000 is used as an example. MPINS generates a positive influential node set that is 42:1 smaller on average than PIDS. The solution sizes of PIDS and MPINS grow as _ grows, as shown in Fig. 9b for the same reasons discussed for Fig. 5. In addition, when _ grows, PIDS produces more nodes than MPINS. There are 4182% more nodes generated by PIDS than by MPINS on average.

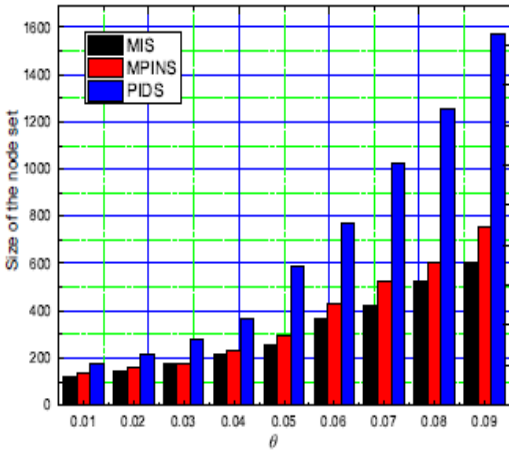
8.2 Experimental results on real data sets

8.2.1 Experimental setting

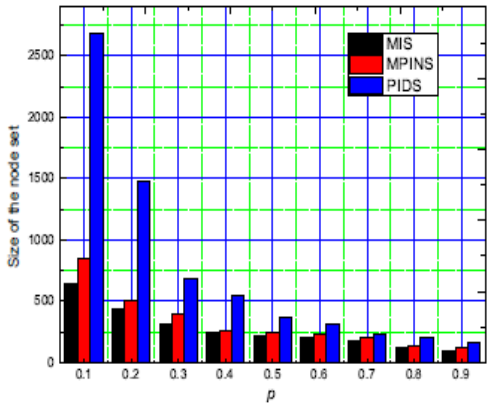
We also include trials conducted on various types of real-world data. The primary data sets,



(a)



(b)



(c)

Figure 9: Node set size for (a) _ = 0.02 and (b) p = 0.2 and (c) n = 50,000 and (p = 0.2).

The source of which, as stated in Table 1, is The Stanford University Large Network Dataset Collection (SNAP) is a repository for publicly available network datasets. Linked system

Table 1 Data set 1 in our experiment.

Data set	Number of nodes	Number of edges	LWCC(N)	LWCC(E)	LSCC(N)	LSCC(E)	Diameter
A1	262 111	1 234 877	262 111	1 234 877	241 761 (0.922)	1 131 217 (0.916)	29
A2	400 727	3 200 440	400 727	3 200 440	380 167 (0.949)	3 069 889 (0.959)	18
A3	410 236	3 356 824	410 236	3 356 824	390 304 (0.951)	3 255 816 (0.970)	21
A4	403 394	3 387 388	403 364	3 387 224	395 234 (0.980)	3 301 092 (0.975)	21

Note: N stands for nodes, E stands for edges.

Summaries of statistics include the number of nodes and edges, the size of the largest weakly connected component (LWCC), and the number of nodes and edges in the smallest strongly connected component. The number of edges, and the diameter (i.e., the longest and shortest route) of the LSCC. The information in Table 1 was compiled using the Amazon.com tool "Customers Who Bought This Item Also Bought." Information gathered in the Amazon from March to May of 2003 was used to construct four distinct networks. If product i am commonly bought alongside product j, then there will be an edge between the two goods in the network [43]. We also test our system on the following real-world datasets, in addition to the Amazon product co-purchasing datasets provided in Table 1:

One such dataset is Wiki Vote, which can be found in Ref. [44] and provides information on past votes on Wikipedia. Voting information for Wikipedia from its beginning in 2001 to January 2008 is included in a data collection with 7115 vertices and 103 689 edges. There will be a connection between users i and j if i voted for j in the administrative election. For (2) Coauthor, see Ref. [45], which provides access to the

data collection containing the coauthors' information as kept by ArnetMiner__.The set we settled on has exactly 53 442 vertices and exactly 127 968 edges. If author i is connected to author j, then there will be one edge between them.

Figure 10 displays the average degree of each data set, which may be used to get insight into the data features of the real-world datasets. Figure 11 provides a concise overview of the social impact distribution between every pair of nodes in the datasets. Figure 11a demonstrates that the majority of the vertices are influenced socially during the interval. Time stamps of 0:005; 0:05 in the Amazon A1-A4 co-purchase datasets. Based on this finding, we allowed _ to vary in the experiment data sets including Amazon co-purchases from 0:005 to 0:02. Most social impacts around the boundaries lie inside the range, as shown in Figure 11b.Data sets from Wiki Vote (0:02), Coauthor (0:10), and Twitter (0:10). In the same way, we varied _ in the experiments from 0:02 to 0:08 for these three datasets.

8.2.2 Experimental results

The effects of changing from 0:005 to 0:02 on Amazon co purchase data set _ sizes, MPINS solutions, and PIDS solutions in Fig. 12a for your perusal. Because more influential nodes must be chosen when the pre-set threshold is high, the size of the PIDS and MINS solutions grows as _ increases, as shown in Fig. 12a. MPINS generates more compact sets of influential nodes than PIDS does for a given _. MPINS's solution size is also somewhat similar to that of MIS. Those findings line up with

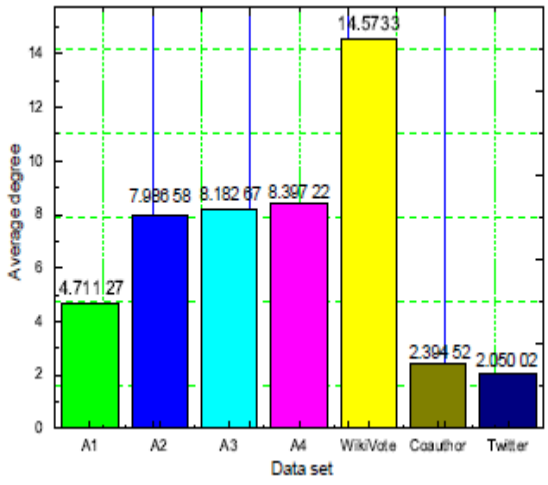


Fig. 10 Average degree of each real-world data sets.

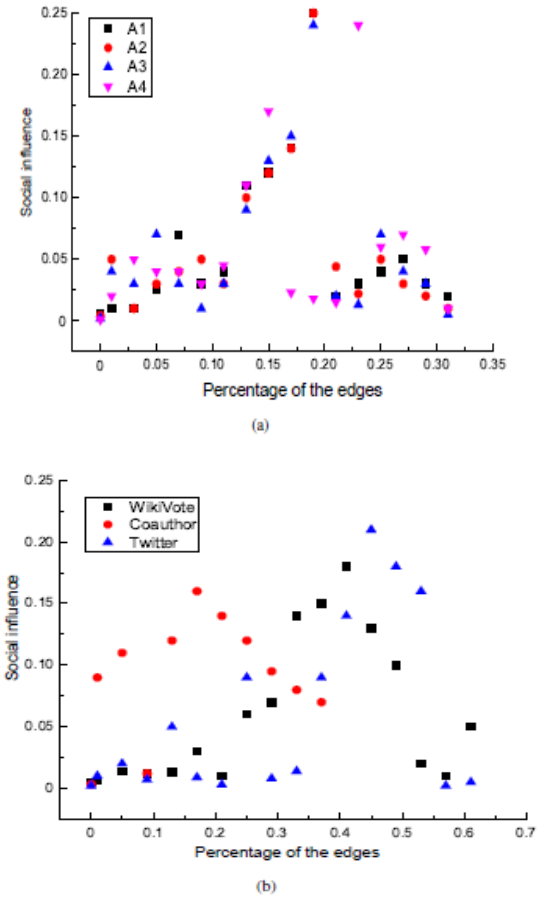
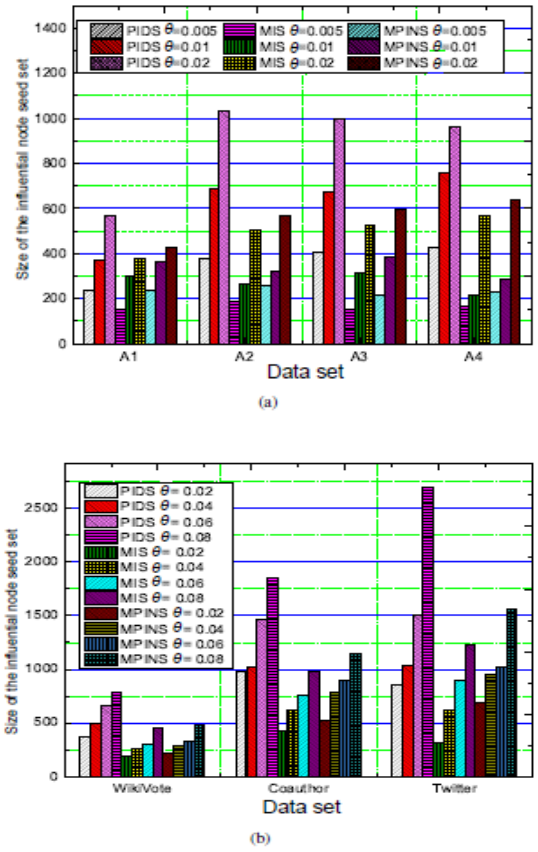


FIGURE 11: The distribution of possible outcomes for the Wiki Vote, Coauthor, and Twitter datasets, as well as the Amazon co-purchase dataset (a).

The outcomes of the simulation. Our recommended analysis of data set A2 MPINS is a much superior technique than PIDS. When compared to PIDS, MPINS chooses nodes with 31% less influence on average. The ratio of PIDS to MPINS solution size is around 37:23.

Because MPINS prioritizes the node with the most influence, rather than the node with the largest degree, this outcome occurs. In addition, PIDS has a faster rate of solution size expansion than MPINS. In particular, the typical rates of increase for PIDS and MPINS solution sizes are 62 and 38%, respectively. Once again, the findings demonstrate that more degree does not equal greater sway in a social network. Changes in _ from 0:02 to 0:08 have similar effects on the MIS size, MPINS solutions, and PIDS solution on the WikiVote, Coauthor, and Twitter data sets, as shown in Fig. 12b. Sizes of PIDS and MINS solutions grow as _ grows, as seen in Fig. 12b. MPINS generates a smaller number of influence nodes than does



Datasets for (a) Amazon copurchases and (b) WikiVotes, Coauthorship, and Twitter are combined in Figure 12 to illustrate the overall number of significant nodes in each. PIDS. The MPINS solution is also around the same size as the MIS. When applied to the Twitter dataset, MPINS picks 45–45% less influential nodes than PIDS. The mean absolute difference between PIDS and MPINS results is 36.37 percentage points. Further, PIDS's rate of solution size expansion is higher than that of MPINS. In instance, the average increase rate of the size of the solution is 54:1% for PIDS and 43:6% for MPINS. Figure 13 displays the percentage of the network's nodes that play a significant role. Figure 13a shows the results of _ for the Amazon co-purchase data sets on the ratio of MIS, MPINS, and PIDS, while Figure 13b shows the results of _ for the WikiVote, Coauthor, and Twitter data sets.

The consequences of the past don't repeat themselves. Nonetheless, a single idea stands out:

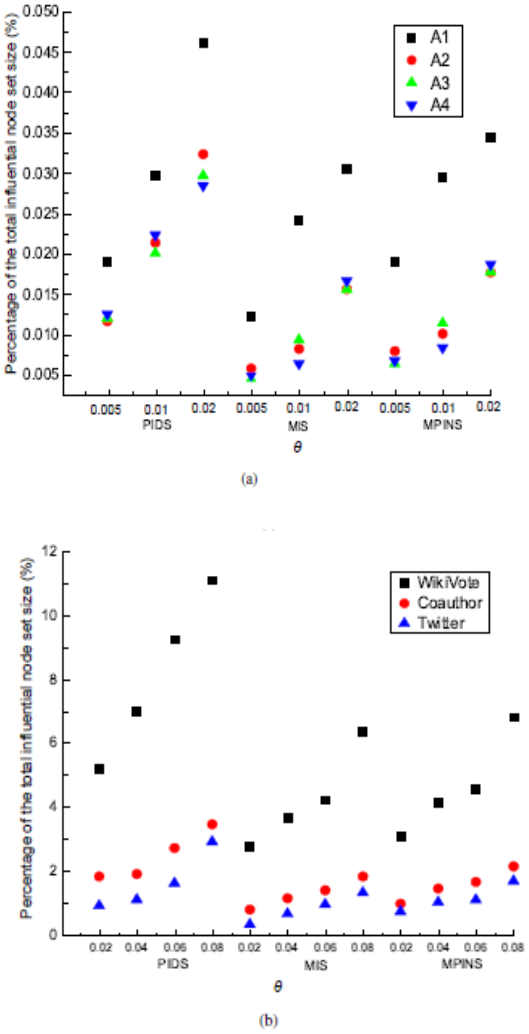


Figure 13: In (a) Amazon co-purchase data sets and (b) WikiVote, Coauthor, and Twitter data sets, the percentage of the total size of prominent node sets. That compared to WikiVote, Coauthor, and Twitter data sets, Amazon co-purchase data sets have a significantly smaller number of nodes identified as influential nodes. To be more precise, PIDS and MPINS choose nodes with influence scores of 0.047 and 0.035, respectively, under the worst case scenario (for the data set A1 with _ D = 0.02). Both PIDS and MPINS select 11.2 percent of nodes as influential for the Wiki Vote dataset with _ D = 0.08. Compared to the WikiVote, Coauthor, and Twitter data sets, the Amazon copurchase data sets seem to be more conducive to the spread of social impacts. Users' past Amazon purchases are used to make comparable product recommendations, which speeds up the influence spread.

Finally, we evaluate MPINS against PIDS and a "Random" technique that selects a node at random to serve as the influential node. Figure 14 displays the

effects of _ on the solution sizes of MPINS, PIDS, and Random when _ varies from 0:02 to 0:08 for the WikiVote, Coauthor, and Twitter datasets. Figure 14 illustrates how the sizes of Random, PIDS, and MPINS solutions grow with n. In addition, MPINS yields a less significant effect for a given _.

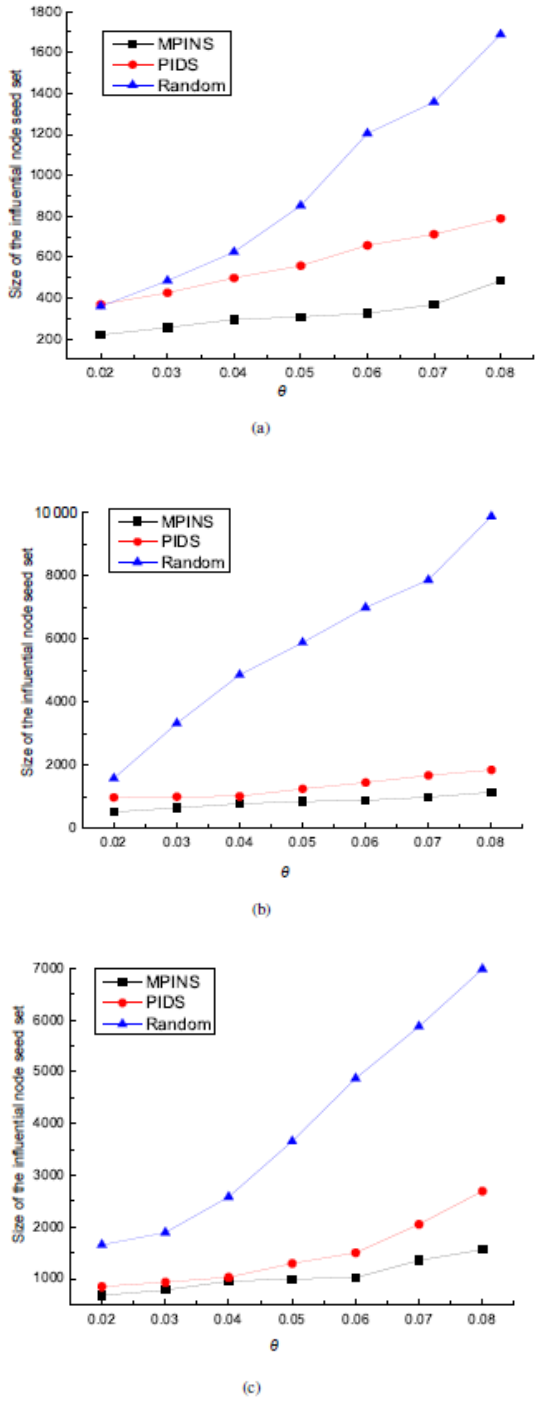


Figure 14: Comparison of MPINS, PIDS, and Random in (a) WikiVote, (b) Coauthor, and (c) Twitter.

cluster than PIDS does. This finding accords with findings from past experiments and results from simulations. For a given _, Random selects a node at random without any selection criteria, whereas PIDS and MPINS generate significantly narrower collections of influential nodes. In contrast to our MPINS' greedy criteria, PIDS's selection method is degree-based, while it is influence-based. Both PIDS and MPINS seem like they ought to do better than Random, at least intuitively. On average, MPINS chooses 48.33% fewer important nodes than PIDS for theWikiVote dataset (shown in Fig. 14a). When compared to Random, MPINS chooses nodes with 61% fewer influence 69% of the time. Compared to PIDS, MPINS picks, on average, 15-32% less important nodes from the Coauthor data set (shown in Fig. 14b). On average, MPINS chooses 13–77% less influential nodes than Random. Compared to PIDS, MPINS picks, on average, 23.121% less important nodes from the Twitter data set (Fig. 14c). When compared to Random, MPINS chooses nodes with 66% fewer influence on average.

9 Conclusions

This research investigates the commercially-relevant MPINS selection issue in social networks. By use of simplification, we prove that MPINS according to the independent cascade model is challenging for APX. As a result, we propose a greedy algorithm, MPINSGREEDY, to address the issue. We test our proposed technique on seven distinct real-world datasets and verify it through simulations on random networks. Evidence from simulation and testing suggests that MPINS-GREEDY can build satisfied initial active node sets that are smaller than the most recent related study, PIDS. Furthermore, MPINS-GREEDY performs similarly to the optimal MPINS solution for small-scale networks. In addition, MPINSGREEDY significantly outperforms PIDS in sparse networks, large-scale networks, and with a high threshold.

References

*[1] D. Kempe, J. Kleinberg, and ´ E. Tardos, Maximizing the spread of influence through a social network, in Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 137–
146.*

[2] K. Saito, M. Kimura, and H. Motoda, Discovering influential nodes for sis models in social networks, in Proc. International Conference on Discovery Science, 2009, pp. 302–316.
[3] Y. Li,W. Chen, Y.Wang, and Z. Zhang, Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships, in Proc. of the Sixth ACM International Conference on Web Search and Data Mining, 2013, pp. 657–666.
[4] M. Han, M. Yan, Z. Cai, Y. Li, X. Cai, and J. Yu, Influence maximization by probing partial communities in dynamic online social networks, Transactions on Emerging Telecommunications Technologies, vol. 28, no. 4, p. e3054, 2016.
[5] X. He, G. Song,W. Chen, and Q. Jiang, Influence blocking maximization in social networks under the competitive linear threshold model, in Proc. of the 2012 SIAM International Conference on Data Mining, 2012, pp. 463– 474.
[6] W. Lu, F. Bonchi, A. Goyal, and L. V. Lakshmanan, The bang for the buck: Fair competitive viral marketing from the host perspective, in Proc. of the 19th SIGKDD International Conference on Knowledge Discovery and
Data Mining, 2013, pp. 928–936.
[7] F. Wang, H. Du, E. Camacho, K. Xu, W. Lee, Y. Shi, and S. Shan, On positive influence dominating sets in social networks, Theoretical Computer Science, vol. 412, no. 3, pp. 265–269, 2011.
[8] M. Han and Y. Li, Influence analysis: A survey of the stateof- the-art, in International Symposium on Bioinformatics Research and Applications, 2018, pp. 259–264.
[9] J. Tang, J. Sun, C. Wang, and Z. Yang, Social influence analysis in large-scale networks, in Proc. of the 15th
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 807–816.
[10] A. Goyal, F. Bonchi, and L. V. Lakshmanan, Learning influence probabilities in social networks, in Proc. of the
Third ACM International Conference on Web Search and Data Mining, 2010, pp. 241–250.