Minimizing Seed Set Selection with Probabilistic Coverage Guarantee in a Social Network

Peng Zhang
Purdue University
zhan1456@purdue.edu

Wei Chen
Microsoft
weic@microsoft.com

Xiaoming Sun[*]
Institute of Computing Technology, CAS
sunxiaoming@ict.ac.cn

Yajun Wang
Microsoft
yajunw@microsoft.com

Jialin Zhang
Institute of Computing Technology, CAS
zhangjl2002@gmail.com

ABSTRACT

A topic propagating in a social network reaches its tipping point if the number of users discussing it in the network exceeds a critical threshold such that a wide cascade on the topic is likely to occur. In this paper, we consider the task of selecting initial seed users of a topic with minimum size so that *with a guaranteed probability* the number of users discussing the topic would reach a given threshold. We formulate the task as an optimization problem called *seed minimization with probabilistic coverage guarantee (SM-PCG)*. This problem departs from the previous studies on social influence maximization or seed minimization because it considers influence coverage with *probabilistic* guarantees instead of guarantees on *expected* influence coverage. We show that the problem is not submodular, and thus is harder than previously studied problems based on submodular function optimization. We provide an approximation algorithm and show that it approximates the optimal solution with both a multiplicative ratio and an additive error. The multiplicative ratio is tight while the additive error would be small if influence coverage distributions of certain seed sets are well concentrated. For one-way bipartite graphs we analytically prove the concentration condition and obtain an approximation algorithm with an $O(\log n)$ multiplicative ratio and an $O(\sqrt{n})$ additive error, where n is the total number of nodes in the social graph. Moreover, we empirically verify the concentration condition in real-world networks and experimentally demonstrate the effectiveness of our proposed algorithm comparing to commonly adopted benchmark algorithms.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Algorithms, Experimentation, Performance

Keywords

social networks, influence diffusion, independent cascade model, seed minimization

1. INTRODUCTION

With online social networks such as Facebook and Twitter becoming popular for people to express their thoughts and ideas, or to chat with each other, online social networks provide a platform for triggering a hot topic and then influencing a large population. Different from most traditional media (such as TV and newspapers), information spread on social networks mainly base on the trust relationship between individuals. Consider the following scenario: when someone publishes a topic on the online social network, his/her friends will see this topic on the website. If they think it is interesting or meaningful, they may write some comments to follow it or just forward it on the website as a response. Similarly, the comments or forwarding from these friends will attract their own friends, leading to more and more people on the social network paying attention to that topic. When the number of users discussing about this topic on the online social network reaches certain critical threshold, this topic becomes a *hot topic*, which is likely to be surfaced at the prominent place on the social networking site (e.g. 10 hot topics of today), and is likely to be picked up by traditional media and influential celebrities. In turn this will generate an even wider cascade causing more people to discuss about this topic.

Therefore, making a topic reach the critical threshold (also called the *tipping point* [10]) is the crucial step to generate huge influence on the topic, which is desirable by companies large and small trying to use social networks to promote their products, through the so called *viral marketing campaigns*. Besides making the content of the topic attractive

[*]This work was supported in part by the National Natural Science Foundation of China Grant 61170062, 61222202 and the National Program for support of Top-notch Young Professionals.

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM http://dx.doi.org/10.1145/2623330.2623684.

and viral, another key aspect is to select seed users in the network that initiate the topic discussion effectively to trigger a large cascade on the topic. Due to the cost incurred for engaging seed users (e.g. providing free sample products), it is desirable that the size of seed users is minimized. Moreover, the marketers also need certain probabilistic guarantee on how likely the viral marketing campaign could reach the desired critical threshold in order to trigger an even larger cascade via hot topic listings, traditional media coverages, and celebrity followings. Hence, the problem at hand is how to select a seed set of users of minimum size to trigger a topic cascade such that the cascade size reaches the desired critical threshold with guaranteed probability.

In this paper, we formulate the above problem as the following optimization problem and call it *seed minimization with probabilistic coverage guarantee (SM-PCG)*. A social network is modeled as a directed graph, where nodes represent individuals and directed edges represent the relationships between pairs of individuals. Each edge is associated with an *influence probability*, which means that once a node is activated, it can activate its out-neighbors through the outgoing edges with their associated probabilities at the next step. Our analytical results work for a large class of influence diffusion models that guarantee submodularity (the diminishing marginal return property in terms of seed set size), but for illustration purpose, we adopt the classic *independent cascade (IC) model* [14] as the influence diffusion model. In the IC model, initially all seed nodes are activated while others are inactive, and at each step, nodes activated at the previous step have one chance to activate each of its inactive out-neighbors in the network. The total number of active nodes after the diffusion process ends is referred as the *influence coverage* of the initial seed set. Given such a social network with influence probabilities on edges, given a required coverage threshold η and a probability threshold P, the SM-PCG problem is to find a seed set S^* of minimum size such that the probability that the influence coverage of S^* reaches η or beyond is at least P.

The formulation of the SM-PCG problem significantly departs from previous optimization problems based on social influence diffusion (e.g. [14, 5, 4, 12]) in that it requires the selected seed set to satisfy a *probabilistic* coverage guarantee, while previous research focuses on *expected* coverage guarantee. For the application of generating a hot topic, we believe that it is reasonable to ask for a guarantee on the probability of influence coverage exceeding a given threshold, since this provides direct information on the likelihood of success of the viral marketing campaign, which is very helpful for marketers to gauge their cost and benefit trade-offs for the campaign. Merely saying that the expected influence coverage exceeds the required coverage threshold is not enough in this case. To the best of our knowledge, this is the first work that focuses on probabilistic influence coverage guarantee among existing studies on social network influence optimization problems.

In this paper, we first show that the set functions based on the SM-PCG problem are not submodular, which means that it is more difficult than most of the existing social influence optimization problems that rely on submodular set function optimizations. Next, we investigate two computation tasks related to SM-PCG problem, one is to fix a seed set S and a coverage threshold η and compute the probability of influence coverage of S exceeding η, and the other is to fix a seed set S and a probability threshold P, and compute the maximum coverage threshold η such that the probability of influence coverage of S exceeding η is at least P. We show that the first problem is #P-hard but can be accurately estimated, while the second one is #P-hard to even approximate the value within any nontrivial ratio. These results further demonstrate the hardness of the problem.

We then adapt the greedy approximation algorithm targeted for expected influence coverage problem (which is submodular) to the SM-PCG problem. Although the adapted algorithm still follows the greedy approach, our main contribution is on a detailed analysis, which proves that our algorithm approximates the optimal solution with both a multiplicative ratio and an additive error. The multiplicative ratio is due to the greedy approximation algorithm for expected influence coverage and is tight, while the additive error is determined by the concentration property (in particular the standard deviations) of influence coverage distributions of two specific seed sets. For one-way bipartite graphs where edges are directed from one side to the other side, we analytically show that the influence coverage distributions are well concentrated and we could reach an additive error of $O(\sqrt{n})$ where n is the total number of nodes in the graph.

Finally, using several real-world social networks including a network with influence probability parameters obtained from prior work, we empirically validate our approach by showing that (a) influence coverage distributions of seed sets are well concentrated, and (b) our algorithm selects seed sets with sizes much smaller than commonly adopted benchmark algorithms.

To summarize, our contributions include: (a) we propose the study of seed minimization with probabilistic coverage guarantee (SM-PCG), which is more relevant to hot topic generation in online social networks and has not been studied before; (b) we show that neither of the two versions of set functions related to SM-PCG is submodular, one version is #P-hard to compute but allows accurate estimation while the other version is #P-hard to even approximate to any nontrivial ratio; (c) we adapt the greedy algorithm targeted for expected coverage guarantee to SM-PCG, and analytically show that the adapted algorithm provides an approximation guarantee with a tight multiplicative ratio and an additive error depending on the influence coverage concentrations of certain seed sets; and (d) we empirically demonstrate the effectiveness of our algorithm using real-world datasets.

1.1 Related Work

Influence maximization, as the dual problem of seed minimization, is to find a seed set of at most k nodes to maximize the expected influence coverage of the seed set. Domingos and Richardson are the first to formulate influence maximization problem from an algorithmic perspective [7, 17]. Kempe et al. first model this problem as a discrete optimization problem [14], provide the now classic independent cascade and linear threshold diffusion models, and establish the optimization framework based on submodular set function optimization. A number of studies follow this approach and provide more efficient influence maximization algorithms (e.g. [5, 4, 6, 13]). In [16], Long et al. first study independent cascade and linear threshold diffusion models from a minimization perspective. In [12], Goyal et al. provide a bicriteria approximation algorithm to minimize the size of the seed set with its expected influence coverage reaching a given threshold. Recently, a continuous time diffusion model is proposed and studied in [18] and [8]. All these existing studies focus on expected influence coverage, and rely on the submodularity of expected influence coverage function for the optimization task. In contrast, we are the first to address probabilistic coverage guarantee for the seed minimization problem, which is not submodular.

Seed minimization with non-submodular influence coverage functions under different diffusion models have been studied. Chen [3] studies the seed minimization problem under the fixed threshold model, where a node is activated when its active neighbors exceed its fixed threshold. He shows that the problem cannot be approximated within any polylogarithmic factor (under certain complexity theory assumption). Goldberg and Liu [11] study another variant of fixed threshold model and provide an approximation algorithm based on the linear programming technique. Influence coverage functions in both models are deterministic and non-submodular. However, these models are quite different from the model we study in this paper, and thus their results and techniques are not applicable to our problem.

The rest of this paper is organized as follows. We define the diffusion model and the optimization problem SM-PCG in Section 2, and provide related results and tools in Section 3, including the non-submodularity of the set functions for SM-PCG. In Section 4 we investigate the computation problems related to SM-PCG. In Section 5 we provide our algorithm for general graphs and analyze its approximation guarantee. In Section 6 we provide algorithmic and analytical results for one-way bipartite graphs. We empirically validate our concentration assumption on influence coverage distributions and the effectiveness of our algorithm in Section 7, and conclude the paper in Section 8 with a discussion on potential future directions. Due to the space constraint, some of the technical proofs and empirical results are omitted, and they are included in our full technical report [20].

2. MODEL AND PROBLEM

In our problem, a social network is modeled as a directed *social graph* $G = (V, E)$, where V is the set of n nodes representing individuals in a social network, and E is the set of directed edges representing influence relationships between pairs of individuals. Each edge $(u, v) \in E$ is associated with an *influence probability* $p_{u,v}$. Intuitively, $p_{u,v}$ is the probability that node u activates node v after u is activated. The influence diffusion process in the social graph G follows the independent cascade (IC) model, a randomized process summarized in [14]. Each node has two states, *inactive* or *active*. The influence diffusion proceeds in discrete time steps, and we say that a node u *is activated at time* t if t is the first time step at which u becomes active. At the initial time step $t = 0$, a subset of nodes $S \subseteq V$ is selected as active nodes, defined as the *seed set*, while other nodes are inactive. For any time $t \geq 1$, when a node u is activated at step $t-1$, u is given a single chance to activate each of its inactive out-neighbors v through edge (u, v) independently with probability $p_{u,v}$ at step t. Once activated, a node stays as active in the remaining time steps. The influence diffusion process stops when there is no new activation at a time step.

Given a target set $U \subseteq V$, let $\mathit{Inf}_U(S)$ be the random variable denoting the number of active nodes in U after the diffusion process starting from the seed set S ends. When the context is clear, we usually omit the subscript U and use $\mathit{Inf}(S)$ to represent this random variable, and we refer $\mathit{Inf}(S)$ as the *influence coverage* of seed set S (for target set U). The optimization problem we are trying to solve is to find a seed set S of minimum size such that the influence coverage of S is at least a required threshold with a required probability guarantee. The formal problem is defined below.

Definition 1 (Seed minimization with probabilistic coverage guarantee) *We define the problem of* seed minimization with probabilistic coverage guarantee (SM-PCG) *as follows. The input of the problem includes the social graph $G = (V, E)$, the influence probabilities $p_{u,v}$'s on edges, the target set U, a coverage threshold $\eta < |U|$,*[1] *a probability threshold $P \in (0, 1)$. The problem is to find the minimum size seed set S^* such that S^* can activate at least η nodes in U with probability P, that is,*

$$S^* = \operatorname*{argmin}_{S:\Pr(\mathit{Inf}(S) \geq \eta) \geq P} |S|.$$

The following theorem shows the hardness of the SM-PCG problem.

THEOREM 1. *The problem SM-PCG is NP-hard, and for any $\varepsilon > 0$, it cannot be approximated within a ratio of $(1 - \varepsilon) \ln n$ unless NP has $n^{O(\log \log n)}$-time deterministic algorithms.*

PROOF (SKETCH). We show that the NP-complete problem Set Cover is a special case of SM-PCG, and the lower bound on approximation ratio is due to the result in [9] on the Set Cover problem. □

With the above hardness result, we set our goal as to find algorithms that solve the SM-PCG problem with approximation ratio close to $\ln n$.

3. USEFUL RESULTS AND TOOLS

In this section, we provide some useful results and tools in preparation for our algorithm design.

Almost all previous work on social influence maximization or seed minimization is based on submodular function optimization techniques. Consider a set function $f(\cdot)$ which maps subsets of a finite ground set into real number set $\mathbb{R}$. We say that $f(\cdot)$ is *submodular* if for any subsets $S \subseteq T$ and any element $u \notin T$, $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$. Moreover, we say that $f(\cdot)$ is *monotone* if for any subsets $S \subseteq T$, $f(S) \leq f(T)$.

Consider a monotone and submodular function $f(\cdot)$ on subsets of nodes in the social graph $G = (V, E)$. Suppose that each node $v \in V$ has a cost $c(v)$, given by a cost function $c : V \to \mathbb{R}^+$. The cost of a subset S is defined as $c(S) = \sum_{v \in S} c(v)$. In [12], Goyal et al. investigate the problem of finding a subset $S \subseteq V$ with minimum cost such that $f(S)$ is at least some given threshold η. As in many optimization tasks for submodular functions, the following greedy algorithm is applied to solve the problem: starting from the emptyset $S_0 = \emptyset$, in the i-th iteration with

[1] We believe that $\eta < |U|$ is reasonable for the application scenarios we described since typically it requires only a fraction of the entire target node set to make a topic hot. For the case of $\eta = |U|$, we also worked out a separate solution for one-way bipartite graphs, but due to page limit, we omit it in this paper.

$i = 1, 2, \ldots$, find a node v_i that provides the largest marginal gain on f per-unit cost, that is find

$$v_i = \operatorname*{argmax}_{v \in V \setminus S_{i-1}} \frac{f(S_{i-1} \cup \{v\}) - f(S_{i-1})}{c(v)},$$

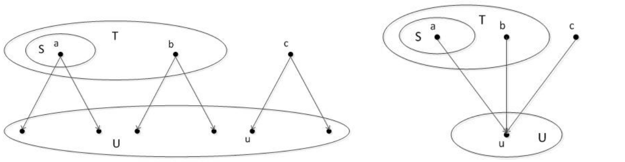
and add v_i to S_{i-1} to obtain S_i; continue this process until iteration j in which $f(S_j) \geq \theta$, where θ is a threshold that could be η or some other value chosen by the algorithm as the stopping criteria, and output S_j as the selected subset S. However, generally computing $f(\cdot)$ exactly is #P-hard, but for most influence spread models, it can be estimated by Monte Carlo simulation as accurately as possible. We say an estimation $\hat{f}(\cdot)$ is a *γ-multiplicative error estimation* of $f(\cdot)$, if for any subset S, $|\hat{f}(S) - f(S)| \leq \gamma f(S)$. Goyal et al. show a bicriteria approximation result for the above greedy algorithm when $\gamma = 0$ [12]. For the case of uniform node cost and $\eta < f(V)$, we slightly improve the above result by removing the bicriteria restriction and generalizing to the case of $\gamma \geq 0$.

THEOREM 2. *Let $G = (V, E)$ be a social graph, and let $f(\cdot)$ be a nonnegative, monotone and submodular set function on the subsets $|V|$. Given a threshold $0 < \eta < f(V)$,let $S^* \subseteq V$ be a subset of minimum size such that $f(S^*) \geq \eta$, and S be the greedy solution using a γ-multiplicative error estimation function $\hat{f}(\cdot)$ with the stopping criteria $\hat{f}(S) \geq (1 + \gamma)\eta$. For any $0 \leq \varepsilon_0 \leq 1$, for any $0 \leq \gamma \leq \frac{\varepsilon_0(f(V)-\eta)}{8|V|(f(V)+\eta|V|)}$, we have $f(S) \geq \eta$, and $|S| \leq \alpha|S^*| + 1$ where $\alpha = \max\{\left\lceil \ln \frac{(1+\varepsilon_0)\eta|V|}{f(V)-\eta} \right\rceil, 0\}$.*

Note that when $\eta = \Theta(f(V))$, we have $\gamma \leq \frac{\varepsilon_0(f(V)-\eta)}{8|V|(f(V)+\eta|V|)} = \Theta(\frac{\varepsilon_0}{|V|^2})$.

Kempe et al. show that set function $E[\mathit{Inf}(S)]$ for *expected influence coverage* is monotone and submodular under the IC model [14]. Therefore, if our problem is to find a seed set of minimum size such that the expected influence coverage is at least a threshold value η, Theorem 2 already provides the approximation guarantee of the greedy algorithm. We call this problem the *seed minimization with expected coverage guarantee (SM-ECG)*, to differentiate with the problem concerned in this paper — seed minimization with *probabilistic* coverage guarantee (SM-PCG).

For the SM-PCG problem, we want the influence coverage to be at least η with a guaranteed probability P. This seemingly minor change from SM-ECG actually alters the nature of the problem. The SM-PCG corresponds to two variants of set functions, but neither of them is submodular. In the first variant, we fix influence threshold η, and define $f_\eta : 2^{|V|} \to \mathbb{R}^+$ where $f_\eta(S) = \Pr(\mathit{Inf}(S) \geq \eta)$. In the second variant, we fix probability P, and define $g_P : 2^{|V|} \to \mathbb{R}^+$ where $g_P(S) = \max_{\eta' : \Pr(\mathit{Inf}(S) \geq \eta') \geq P} \eta'$. Neither $f_\eta(\cdot)$ nor $g_P(\cdot)$ is submodular, as shown by the two examples below. For f_η, see Figure 1(a), G is a bipartite graph where all edges are associated with probability 1, and U contains all the nodes in the lower part. We fix $\eta = 5$. Let $S = \{a\}$ and $T = \{a, b\}$, then $f_\eta(S \cup \{u\}) - f_\eta(S) = 0$, since neither S nor $S \cup \{u\}$ could reach 5 nodes in U. Similarly, $f_\eta(T) = 0$. However, $f_\eta(T \cup \{u\}) = 1$, since 5 nodes are reached by $T \cup \{u\}$. Therefore, $f_\eta(T \cup \{u\}) - f_\eta(T) > f_\eta(S \cup \{u\}) - f_\eta(S)$, and thus $f_\eta(\cdot)$ is not submodular. For g_P, see Figure 1(b), G is a bipartite graph where all edges are associated with probability 0.5, and $U = \{u\}$. We set $P = 0.8$. Let $S = \{a\}$ and $T = \{a, b\}$, then $g_P(S \cup \{c\}) - g_P(S) = 0$ and $g_P(T \cup \{c\}) - g_P(T) = 1$. Since $g_P(S \cup \{c\}) - g_P(S) < g_P(T \cup \{c\}) - g_P(T)$, g_P is not submodular.



(a) f_η (b) g_P

Figure 1: Function f_η and g_P are nonsubmoduar.

Algorithm 1 Function MC-CompProb[R]: R is a tuning parameter controlling the accuracy of the estimate

```
Input:  $G = (V, E), \{p_{u,v}\}_{(u,v) \in E}, U, S, \eta$
Output: estimate of $P = \Pr(\mathit{Inf}(S) \geq \eta)$
$t = 0$
for $i = 1$ to $R$ do
    simulate IC diffusion with seed set $S$
    $N_i$ = number of final active nodes in $U$
    if $N_i \geq \eta$ then
        $t = t + 1$
    end if
end for
return $t/R$
```

Since neither $f_\eta(\cdot)$ nor $g_P(\cdot)$ is submodular, we cannot apply Theorem 2 on $f_\eta(\cdot)$ or $g_P(\cdot)$ to solve the SM-PCG problem. In this paper, we address this non-submodular optimization problem by relating it to the SM-ECG problem through a concentration assumption on random variable $\mathit{Inf}(S)$ for certain seed sets S.

4. INFLUENCE COVERAGE COMPUTATION

Before working on the SM-PCG problem directly, we first address the related computation issue when a seed set S is given. As we mentioned in last section, there are two variants in influence coverage computation. The first variant is that, given a seed set S and a coverage threshold η, we need to compute the probability $f_\eta(S)$ that S can activate at least η nodes in U. Note that we have $E[\mathit{Inf}(S)] = \sum_{i=1}^{n-1}(f_i(S) - f_{i+1}(S)) \cdot i + f_n(S) \cdot n$. Thus the exact computation of $f_\eta(S)$ must be #P-hard in the IC model since computing expected influence coverage $E[\mathit{Inf}(S)]$ of seed set S has shown to be #P-hard in the IC model [4]. However, we can use Monte Carlo simulation to compute an accurate estimate of the probability. Algorithm 1 shows the procedure MC-CompProb[R] for this task, which simulate the diffusion from seed set S for R runs and use the fraction of runs in which the number of active nodes in U reaches η as the estimate of the probability.

The following lemma shows the relationship between the number of simulations R and the accuracy of the estimate.

LEMMA 1. *Let $\hat{P}$ be the estimate of true value $P = \Pr(\mathit{Inf}(S) \geq \eta)$ output by* MC-CompProb[R] *in Algorithm 1. To guarantee an error of at most ε, i.e. $|\hat{P} - P| \leq \varepsilon$*

[2], *with probability at least* $1 - 1/n^{\delta}$, *it is sufficient to set* $R \geq \ln(2n^{\delta})/(2\varepsilon^2)$.

Proof (Sketch). This is a direct application of Hoeffding's Inequality since Monte Carlo simulation runs are mutually independent. □

The second variant is that, given a seed set S and a specified probability P, we need to compute the maximum influence coverage η of S with at least probability P, that is, $\eta = \max_{\eta':\Pr(\mathit{Inf}(S)\geq\eta')\geq P} \eta'$. Unlike the first variant, we show below that this problem is #P-hard to approximate to any non-trivial ratio. We say that an algorithm approximates a true value v for a computing problem with ratio $\alpha > 1$ if the output of algorithm $\hat{v}$ satisfies $v/\alpha \leq \hat{v} \leq \alpha v$. Note that if the range of value v is from 1 to n, then using $\hat{v} = n^{1/2}$ gives a trivial approximation ratio of $\alpha = n^{1/2}$.

Theorem 3. *For any fixed probability* $P \in (0, 1)$*, the problem of computing* $\eta = \max_{\eta':\Pr(\mathit{Inf}(S)\geq\eta')\geq P} \eta'$ *given a directed social graph* $G = (V, E)$*, influence probabilities* $\{p_{u,v} \,|\, (u, v) \in E\}$*, target set* $U = V$*, and a seed set* S *is #P-hard to approximate within a ratio of* $|V|^{1/2-\varepsilon}$ *for any* $\varepsilon > 0$.

Note that we treat P as a fixed parameter of the problem rather than as part of the input to the computation problem, which makes the result stronger.

Proof (Sketch). We prove the theorem by first reducing the #P-complete counting problem of *s*-*t* connectivity in a directed graph [19] to its decision version, and then reducing the decision version to our computation problem. □

5. APPROXIMATION ALGORITHM

In this section, we overcome the nonsubmodularity nature of the SM-PCG problem discussed in Section 3 by connecting it with the submodular problem SM-ECG. We first provide the general algorithm, and then show that the algorithm returns a seed set that approximates the optimal solution with both a multiplicative ratio and an additive error. The multiplicative ratio is due to the connection with the SM-ECG problem. For the additive error term, we show that it would be nontrivial when certain concentration assumption on influence coverages holds.

Algorithm 2 illustrates algorithm MinSeed-PCG for solving the SM-PCG problem. The algorithm builds up a sequence of subsets $S_0, S_1, S_2, \ldots$, where for each $i \geq 1$, S_i contains one more element u than S_{i-1} such that u provides the largest marginal increase in *expected* influence coverage to seed set S_{i-1}. The way of constructing seed sets S_i's is in line with the greedy approach as discussed in Section 3. In our algorithm, $\hat{E}[\mathit{Inf}(\cdot)]$ is a γ-multiplicative error estimation of exact expected influence $E[\mathit{Inf}(\cdot)]$. Every time a new set S_i is constructed, we compute the probability that the influence coverage of S_i is at least η (line 5). The CompProb in line 5 is a generic function computing $\Pr(\mathit{Inf}(S_i) \geq \eta)$, which could be MC-CompProb[R] in Algorithm 1 for general graphs, or Bi-CompProb in Algorithm 3 for one-way bipartite graphs, or some other functions for this purpose. If the probability computed is at least $P + \varepsilon$, where $\varepsilon \in [0, (1 - P)/2)$ is a parameter of the algorithm, we stop and return S_i as the seed set found by the algorithm. Parameter ε is related to the accuracy of the function CompProb. If CompProb accurately computes $\Pr(\mathit{Inf}(S) \geq \eta)$ (e.g. Bi-CompProb for one-way bipartite graphs), we set ε to 0. If CompProb only provides an estimate (e.g. MC-CompProb[R] for general graphs), we set ε to be an appropriate value related to the error term of the estimate given by the function. We will discuss parameter ε with more technical details later.

Algorithm 2 MinSeed-PCG[ε]: $\varepsilon \in [0, (1-P)/2)$ is a control parameter

```
Input: $G = (V, E), \{p_{u,v}\}_{(u,v)\in E}, U, \eta, P$
Output: seed set $S$, which is an approximation to $S^* = \mathrm{argmin}_{S':\Pr(\mathit{Inf}(S')\geq\eta)\geq P}\{|S'|\}$
1: $S_0 = \emptyset$
2: for $i = 1$ to $n$ do
3:   select $u = \mathrm{argmax}_v\{\hat{E}[\mathit{Inf}(S_{i-1} \cup \{v\})] - \hat{E}[\mathit{Inf}(S_{i-1})]\}$
4:   $S_i = S_{i-1} \cup \{u\}$
5:   $prob$ = CompProb$(G, \{p_{u,v}\}_{(u,v)\in E}, U, S_i, \eta)$
6:   if $prob \geq P + \varepsilon$ then
7:     return $S_i$
8:   end if
9: end for
```

Let S^* be the optimal seed set for the SM-PCG problem, that is, $S^* = \mathrm{argmin}_{S:\Pr(\mathit{Inf}(S)\geq\eta)\geq P} |S|$. Let $n = |V|$ and $m = |U|$. Let $\mathcal{S} = \{S_1, S_2, \ldots, S_n = V\}$ be the sequence of greedy seed sets computed by algorithm MinSeed-PCG[ε] (considering the entire sequence even when MinSeed-PCG[ε] actually stops). Let S_a be the output of MinSeed-PCG[ε] and a is its index in sequence $\mathcal{S}$, and thus S_{a-1} is the set in $\mathcal{S}$ just before S_a.

We define $c = \max\{\eta - E[\mathit{Inf}(S^*)], 0\}$ and $c' = \max\{E[\mathit{Inf}(S_{a-1})] - \eta, 0\}$. Intuitively, we know that $\Pr(\mathit{Inf}(S^*) \geq \eta) \geq P$, and c indicates how much $E[\mathit{Inf}(S^*)]$ could be smaller than η. If $\mathit{Inf}(S^*)$ concentrates well, c should be small. Similarly, we also know that $\Pr(\mathit{Inf}(S_{a-1}) \geq \eta) < P + \varepsilon$, since S_a is the first set satisfying $\Pr(\mathit{Inf}(S_a) \geq \eta) \geq P + \varepsilon$. Thus, c' indicates how much $E[\mathit{Inf}(S_{a-1})]$ could be larger than η, and if $\mathit{Inf}(S_{a-1})$ concentrates well, c' should be small.

The following theorem shows that the output S_a of MinSeed-PCG[ε] approximates the optimal solution S^* with c and c' included in the additive error term.

Theorem 4. *For any* $0 \leq \varepsilon_0 \leq 1$ *and any* $0 \leq \gamma \leq \frac{\varepsilon_0(m-(\eta+c'))^2}{8mn(m+\eta n)}$. *If* $\hat{E}[\mathit{Inf}(\cdot)]$ *is a* γ*-multiplicative error estimation of* $E[\mathit{Inf}(\cdot)]$ *for any subset of nodes, the size of the output by algorithm* MinSeed-PCG[ε] *approximates the size of the optimal solution in the following form:*

$$|S_a| \leq \left\lceil \ln \frac{(1 + \varepsilon_0)\eta n}{m - \eta} \right\rceil \cdot |S^*| + \frac{(c + c')n}{m - (\eta + c')} + 3 + \varepsilon_0. \quad (1)$$

First, note that we assume $m > \eta$, so the multiplicative term above is well defined. Moreover, $\eta + c'$ must be less than m, because otherwise $E[\mathit{Inf}(S_{a-1})] = m = |U|$, which implies $\Pr(\mathit{Inf}(S_{a-1}) = m) = 1$, contradicting the fact that $\Pr(\mathit{Inf}(S_{a-1}) \geq \eta) < P + \varepsilon < 1$. Second, for the multiplicative ratio of $\lceil \ln \frac{\eta n}{m-\eta} \rceil$, when η is a constant fraction of m, i.e. $\eta = \beta m$ where β is a constant independent of m and n, it is

[2] This lemma holds when $\varepsilon > P$. However, we usually set ε smaller than P to make the estimate more reasonable.

$\ln n + O(1)$, which is tight, since Theorem 1 already states that the ratio cannot be better than $\ln n$. The additive error term involves c and c', and we will discuss it in more detail after providing the proof to the theorem below. Third, when $\eta+c'$ is a constant fraction of m, $\gamma \le \frac{\varepsilon_0(m-(\eta+c'))^2}{8mn(m+\eta n)} = \Theta(\frac{\varepsilon_0}{n^2})$. Fourth, by Chernoff bound, to achieve a γ-multiplicative error estimation of expected influence with probability $1-1/n$ for all subsets computed in our algorithm, it is sufficient to sample $\Theta(\gamma^{-2} n \log n)$ number of graphs for each set.

Proof. We only prove the case of $\gamma = \varepsilon_0 = 0$ in this paper. Let i be the minimum index such that $S_i \in \mathcal{S}$ and $E[Inf(S_i)] \ge \eta - c$, and S_i^* be the minimum-sized seed set such that $E[Inf(S_i^*)] \ge \eta - c$. Since $E[Inf(S^*)] \ge \eta - c$, we know that $|S_i^*| \le |S^*|$. By Theorem 2, since $\left\lceil \ln \frac{(\eta-c)n}{m-(\eta-c)} \right\rceil > 0$, we have that

$$|S_i| \le \left\lceil \ln \frac{(\eta-c)n}{m-(\eta-c)} \right\rceil \cdot |S_i^*| + 1 \le \left\lceil \ln \frac{\eta n}{m-\eta} \right\rceil \cdot |S^*| + 1.$$

Let j be the minimum index such that $S_j \in \mathcal{S}$ and $E[Inf(S_j)] \ge \eta + c'$. Since $E[Inf(S_{a-1})] \le \eta + c'$, we know that $|S_j| \ge |S_{a-1}|$. To bound the difference between $|S_{a-1}|$ and $|S_i|$, it is sufficient to compute the difference between $|S_j|$ and $|S_i|$.

By the definition of j, we have that $E[Inf(S_{j-1})] < \eta + c'$, thus $E[Inf(S_{j-1})] - E[Inf(S_i)] < c + c'$. Moreover, by the submodularity of $E[Inf(\cdot)]$, we know that for each $i < t < j$,

$$\begin{aligned} E[Inf(S_t)] - E[Inf(S_{t-1})] &\ge \frac{m - E[Inf(S_{t-1})]}{n} \\ &> \frac{m-(\eta+c')}{n}. \end{aligned}$$

Thus, we have that

$$\begin{aligned} |S_{j-1} \setminus S_i| &\le \frac{E[Inf(S_{j-1})] - E[Inf(S_i)]}{\min_{i<t<j}\{E[Inf(S_t)] - E[Inf(S_{t-1})]\}} \\ &< (c+c') \cdot (\frac{m-(\eta+c')}{n})^{-1} \\ &= \frac{(c+c')n}{m-(\eta+c')}. \end{aligned}$$

It means that

$$|S_j \setminus S_i| < \frac{(c+c')n}{m-(\eta+c')} + 1.$$

Since $|S_a| \le |S_j| + 1 = |S_i| + |S_j \setminus S_i| + 1$, we have

$$|S_a| \le \left\lceil \ln \frac{\eta n}{m-\eta} \right\rceil \cdot |S^*| + \frac{(c+c')n}{m-(\eta+c')} + 3.$$

The theorem holds. □

We now discuss the additive term in Inequality (1). To make it nontrivial, we need the additive term to be $o(n)$ as n grows. This means first that the target set size m should be increasing with n, which is reasonable. Then we should have $c + c' = o(m)$ in order to make the additive term $o(n)$. In the following theorem, we bound c and c' by the variances of the influence coverage of S^* and S_{a-1} respectively using Chebyshev's inequality, and thus linking the above requirement on c and c' to the requirement on the variances of influence coverages.

Theorem 5. *For algorithm* MinSeed-PCG[ε] *with any parameter* ε, *we have*

$$c \le \sqrt{\frac{Var(Inf(S^*))}{P}}. \qquad (2)$$

If we use MC-CompProb[R] *for function* CompProb *and set* $R \ge \ln(2n^2)/(2\varepsilon^2)$, *then algorithm* MinSeed-PCG[ε] *finds a seed set* S_a *such that, with probability at least* $1 - 1/n$, $\Pr(Inf(S_a) \ge \eta) \ge P$ *and*

$$c' \le \sqrt{\frac{Var(Inf(S_{a-1}))}{1-P-2\varepsilon}}. \qquad (3)$$

Theorem 5 shows that the variances of influence coverages of seed sets, or more exactly the standard deviations of influence coverages, determine the scale of the additive error term of the algorithm MinSeed-PCG[ε]. If influence coverages concentrate well with small standard deviations, the algorithm would have a good additive error term. Consider the common case where target set size $m = \Theta(n)$, and η is a constant fraction of m, and P is a normal probability requirement not too close to 0 or 1 (e.g. 0.1 or 0.5), if we could have $Var(Inf(S^*)) = O(m)$ and $Var(Inf(S_{a-1})) = O(m)$, then $c + c' = O(\sqrt{m})$, and the additive error term is $O(n/\sqrt{m}) = O(\sqrt{n})$. Together with Theorem 4, we would know that

$$|S_a| \le (\ln n + O(1))|S^*| + O(\sqrt{n}).$$

In the next section, we analytically show that for one-way bipartite graphs indeed $c+c' = O(\sqrt{m})$. We also empirically verify that in real-world graphs the standard deviations of influence coverages are indeed small, close to $\sqrt{m}$. Therefore, our algorithm are likely to perform well in practice.

We remark that our theorems in this section can be applied to a class of models with the following characteristics:

1. the influence coverage function of a seed set (i.e., $Inf(\cdot)$) is nonnegative, monotone and submodular, thus greedy algorithm gives an $O(\log n)$-approximation ratio for SM-ECG (Theorem 2) and provides a tight multiplicative ratio.
2. the influence coverage when choosing the whole set of nodes as seeds is the size of the targeted set (i.e., $Inf(V) = |U|$), which guarantees that the additive error is reasonable.

The above class includes many diffusion models, such as linear threshold model, general threshold model and continuous time diffusion model.

6. RESULTS ON BIPARTITE GRAPHS

In this section, we solve the SM-PCG problem on a one-way bipartite graph $G = (V_1, V_2, E)$, where all edges in E are from V_1 to V_2. For the sake of convenience, we just assume that $U = V_2$ in this section. It is easy to remove this assumption and make U to be any subset of $V_1 \cup V_2$.

One-way bipartite graphs provide two significant advantages over general graphs. First, it allows a dynamic programming method to compute the exact influence coverage distribution given any seed set S. Second, it allows a theoretical analysis on the concentration of influence coverages of seed sets. We illustrate both aspects below.

We first show how to implement exact computation of function CompProb. We assign indices for nodes in V_2:

$v_1, v_2, \ldots, v_m$. Let $A(S,i,j)$ denote the probability that seed set S can activate exactly j nodes in the first i nodes of V_2: $v_1, \ldots, v_i$, where $j \leq i$. Let $p(S,v)$ be the probability that v can be activated by S. When $i = 1$, it is trivial to get $A(S,1,j)$. When $i > 1$, we can use $A(S,i-1,j-1)$ and $A(S,i-1,j)$ to compute $A(S,i,j)$. If $j = 0$, it means $v_1, \ldots, v_{i-1}$ and v_i are all inactive. If $0 < j < i$, there are two cases: j nodes are activated in the first $i-1$ nodes while v_i is not activated; $j-1$ nodes are activated in the first $i-1$ nodes and v_i is activated. If $j = i$, both $v_1, \ldots, v_{i-1}$ and v_i are activated. Thus, we have the following recursion,

$$A(S,1,j) = \begin{cases} p(S,v_1), & j = 1 \\ 1 - p(S,v_1), & j = 0 \end{cases}$$

and

$$A(S,i,j) =$$

$$\begin{cases} A(S,i-1,j) \cdot (1 - p(S,v_i)), & j = 0 \\ A(S,i-1,j) \cdot (1 - p(S,v_i)) & \\ +A(S,i-1,j-1) \cdot p(S,v_i), & 0 < j < i \\ A(S,i-1,j-1) \cdot p(S,v_i), & j = i \end{cases}$$

For IC model, $p(S,v_i) = 1 - \prod_{u \in S}(1 - p_{u,v_i})$; for LT model, $p(S,v_i) = \sum_{u \in S} p_{u,v_i}$. Using the above dynamic programming formulation, we can implement function CompProb as function Bi-CompProb given in Algorithm 3.

Algorithm 3 Function Bi-CompProb for bipartite graphs

```
Input:  $G = (V_1, V_2, E), \{p_{u,v}\}_{(u,v) \in E}, S, \eta$
Output: $P = \Pr(Inf(S) \geq \eta)$
for $i$ from 1 to $n$, and $j$ from 1 to $i$ do
    compute $A(S,i,j)$ via dynamic programming
end for
return $\sum_{j=\eta}^{m} A(S,m,j)$
```

One-way bipartite graphs have an important property that the activation events of nodes in V_2 are mutually independent. This allows us to bound c and c' defined in Section 5 using Hoeffding's Inequality, and get the following result.

Theorem 6. *For one-way bipartite graphs, algorithm* MinSeed-PCG[0] *using function* Bi-CompProb *returns seed set S_a such that $\Pr(Inf(S_a) \geq \eta) \geq P$, and when we consider the probability threshold P as a constant independent of n and m, we have*

$$|S_a| \leq (\ln n + O(1)) \cdot |S^*| + O(\frac{n}{\sqrt{m}}). \tag{4}$$

We note that one-way bipartite graphs are a restricted class of graphs, where the influence cascading is a 1-hop cascading process and cannot be generated to a cascade with greater depth. However, we believe their analytical results can shed lights on more realistic networks when most of node activations in the network are independent.

7. EXPERIMENTS

We conduct experiments on real social networks for the following purposes: (1) test the concentration of influence coverage distributions of seed sets; (2) validate the performance of our algorithm against baseline algorithms.

7.1 Experiment setup

Datasets. We conduct experiments on three real social networks. The first one is wiki-Vote [15], a network relationship graph from Wikipedia community with totally 7,115 nodes and 103,689 edges. In wiki-Vote graph, each node represents a user, and an edge (u,v) represents user u votes for user v, which means that v has an influence on u. Thus, in our experiment, we reverse all edges to express the influence between pairs of nodes. We use weighted cascade (WC) model [14] to assign the influence probabilities on edges. For each edge (u,v), we assign its probability to be $1/d_{in}(v)$, where $d_{in}(v)$ is the in-degree of node v.

The second network is NetHEPT, which is a standard dataset used in [5, 4, 6, 13, 12]. NetHEPT is an academic collaboration network from arXiv (http://www.arXiv.org), with totally 15,233 nodes and 58,891 edges. In NetHEPT graph, each node represents an author, and each edge represents coauthor relationship between two authors. NetHEPT is an undirected graph, and in our experiment we add two directed edges between two nodes if there exists at least one edge between these two nodes in NetHEPT. Similar to wiki-Vote, we use WC model to assign edge influence probabilities. We assign the probability on directed edge (u,v) to be $d(u,v)/d(v)$, where $d(u,v)$ is the number of papers collaborated by u and v, and $d(v)$ is the number of papers published by v.

The last one is Flixster, an American movie rating social site. Each node is a user, and edges describe the friendship between users. In this network, we use a Topic-aware Independent Cascade Model from [1] to learn the real influence probabilities on edges for different topics. We simply use two different topics, say topic 1 and topic 2, and get the edge probabilities that one user influences his/her friend on the specific topic. In both topics, we remove edges with probability 0 and isolated nodes. For topic 1, there are 28,317 nodes and 206,012 edges. The mean of edge probabilities is 0.103, and the standard deviation is 0.160. For topic 2, there are 25,474 nodes and 135,618 edges. The mean of edge probabilities is 0.133, and the standard deviation is 0.205.

Experiment methods. In the experiment, for the sake of convenience, we set $U = V$.

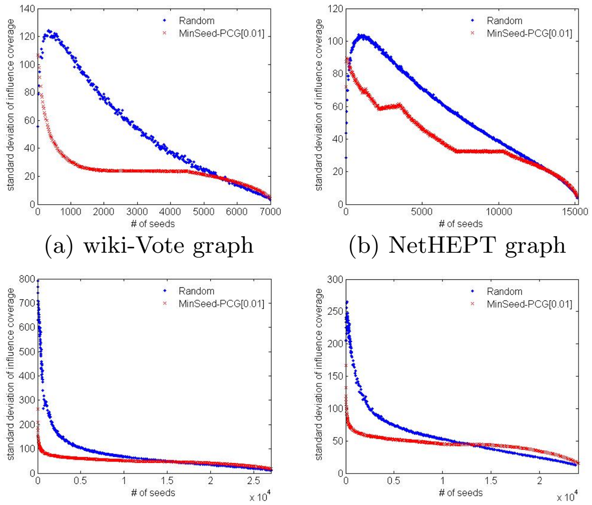
Our first task is to test the concentration of influence coverage distributions of seed sets. To do so, we test the variances (or their square roots, i.e. standard deviations). According to Theorem 5, small standard deviations imply small c and c' and thus small additive errors of the MinSeed-PCG[ε] algorithm output. By Inequality (3), to verify that c' is small, we just need to test the standard deviations of all seed sets generated by the algorithm. For quantity c, we need to test the standard deviation of the influence coverage of the optimal seed set, according to Inequality (2). However, finding the optimal seed set is NP-hard, therefore we cannot fully verify the bound on c. To compensate, we test randomly selected seed sets as follows. For each fixed seed set size k, we independently select 10 seed sets of size k at random, and compute the maximum standard deviations of the influence coverage distributions of these selected seed sets. Although randomly selected seed sets may be far from the optimal seed set, what we hope is that by testing standard deviations on both randomly selected sets and greedily selected sets by algorithm MinSeed-PCG[ε], we have a general understanding of standard deviations of influence coverages of seed sets, which may provide us with hints for other seed sets, such as the optimal seed set. To estimate the standard deviations of influence coverage of a seed set S, we use 10,000 times Monte Carlo simulation and compute the standard deviation.

Our second task is to test the performance of seed selection algorithm $\mathsf{MinSeed\text{-}PCG}[\varepsilon]$. We compare the performance with three baseline algorithms: (a) Random, which generates the seed set sequence in random order; (b) High-degree, which generates the seed set sequence according to the decreasing order of the out-degree of nodes; and (c) PageRank, which is a popular method for website ranking [2]. We use $p_{v,u}/\sum_{(w,u)\in E} p_{w,u}$ as the transition probability for edge (u, v). Higher $p_{v,u}$ means that v is more influential to u, indicating that u ranks v higher. We use 0.15 as the restart probability and use the power method to compute PageRank values. When two consecutive iterations are different for at most 10^{-4} in L_1 norm, we stop. As for our $\mathsf{MinSeed\text{-}PCG}[\varepsilon]$ algorithm, to speed up the algorithm, we use the state-of-the-art PMIA algorithm of [4] to greedily generate the seed set sequence. For all the above algorithms, we use the same $\mathsf{MC\text{-}CompProb}[R]$ algorithm to compare whether a seed set S in the sequence satisfies the condition $\Pr(\mathit{Inf}(S) \geq \eta) \geq P + \varepsilon$. Since the seed set sequence generations in all the above algorithms are fast comparing to the Monte Carlo simulation based $\mathsf{MC\text{-}CompProb}[R]$ algorithm, our implementation actually generates the sequence first and then uses binary search to find the seed set in the sequence satisfying $\Pr(\mathit{Inf}(S) \geq \eta) \geq P + \varepsilon$.

We set parameters $R = 10,000$ and $\varepsilon = 0.01$. One may see that these settings do not satisfy the condition $R \geq \ln(2n^2)/(2\varepsilon^2)$ in Theorem 5 for our datasets: in our datasets, n is around 10^4, and thus $\ln(2n^2)/(2\varepsilon^2)$ is around 9.6×10^4. However, we can justify our choice as follows. First, the condition $R \geq \ln(2n^2)/(2\varepsilon^2)$ is a conservative theoretical condition for obtaining high probability of $1 - 1/n$ for our approximation guarantee. In practice, a smaller R of $10,000$ is good enough for illustrating our results. Second, all algorithms use the same $\mathsf{MC\text{-}CompProb}[R]$ algorithm, so the comparison is fair among them, and is focused on the difference in their generations of seed set sequences, not on the accuracy of the estimate of function CompProb. Third, the seed selections actually depends only on the combined parameter $P' = P + \varepsilon$, and not on P and ε separately. Thus setting $\varepsilon = 0.01$ is only for intuitive understanding and setting it to some other value would not change the results as long as P' remains the same.

7.2 Experiment results

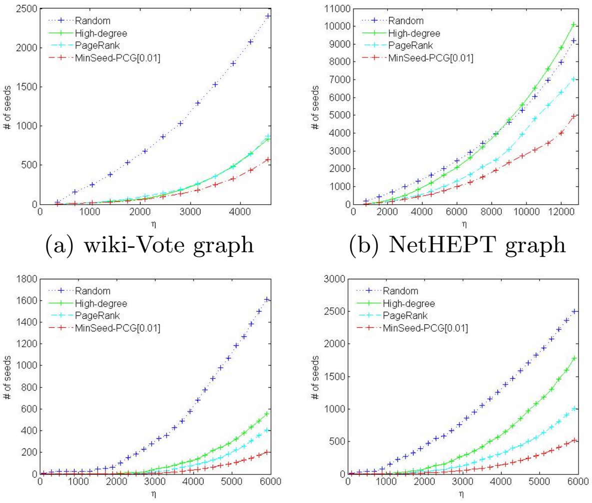
Concentration of influence coverages. Figure 2 shows the standard deviations of influence coverages of randomly selected seed sets and greedily selected seed sets (by algorithm $\mathsf{MinSeed\text{-}PCG}[\varepsilon]$) on wiki-Vote, NetHEPT and Flixster. We can see that in all graphs, standard deviations for greedily selected seed sets quickly drop, while for randomly selected seed sets sometimes it has a small increase when the seed set size is small, and then quickly drop too. The maximum value is about 130 for wiki-Vote ($|V| = 7,115$), 105 for NetHEPT ($|V| = 15,233$), 760 for Flixster with topic 1 ($|V| = 28,317$), and 270 for Flixster with topic 2 ($|V| = 25,474$). Thus by observation the standard deviation is at the order of $\sqrt{|V|}$. As discussed after Theorem 5, this means that the additive error of our algorithm would be $O(\sqrt{|V|})$, a small and satisfactory value.



(a) wiki-Vote graph (b) NetHEPT graph

(c) Flixster graph with topic 1 (d) Flixster graph with topic 2

Figure 2: Standard deviations of influence coverages of seed sets.



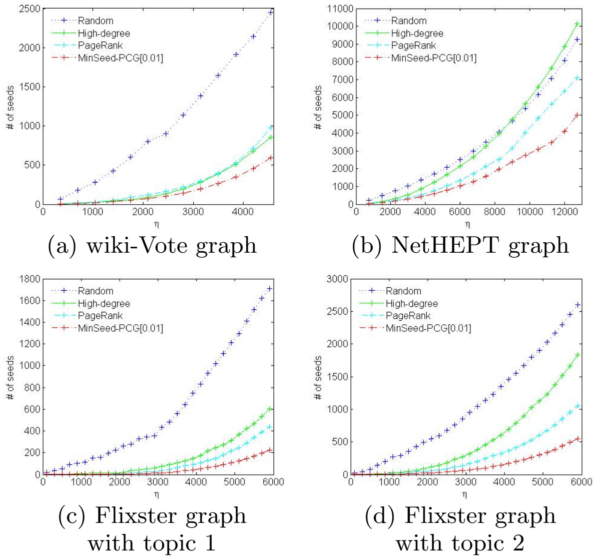
(a) wiki-Vote graph (b) NetHEPT graph

(c) Flixster graph with topic 1 (d) Flixster graph with topic 2

Figure 3: Size of selected seed sets vs. coverage threshold η under a fixed probability threshold $P =$ 0.1.

The standard deviations for wiki-Vote are larger than those for NetHEPT at small seed set size even though the number of nodes of wiki-Vote is smaller. We believe this is because wiki-Vote has more edges (103,689) than NetHEPT (58,891), and thus when the seed set size is small more edges could cause larger variances in influence coverage. This can also explain why in Flixster topic 1 (with 206,012 edges) has larger standard deviations than topic 2 (with 135,618 edges).

Performance of $\mathsf{MinSeed\text{-}PCG}[\varepsilon]$ compared with baselines. We conduct two sets of tests for this purpose. First, we fix the probability threshold P to 0.1 and 0.5, and vary the coverage threshold η to compare the size of seed sets selected by various algorithms. Figure 3 and Figure 4 show



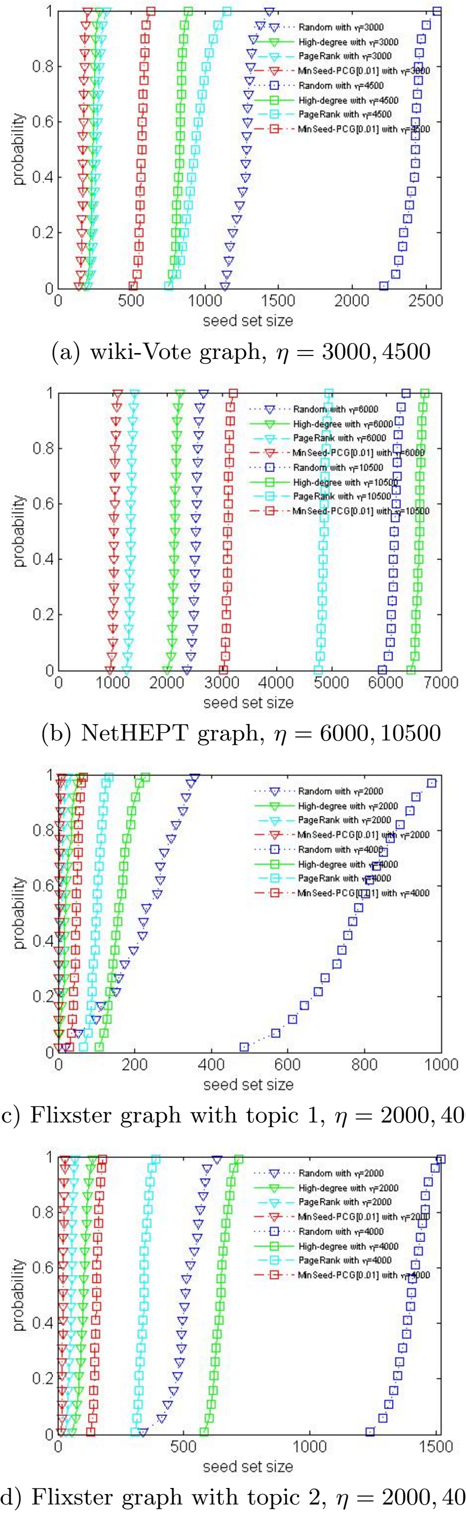
(a) wiki-Vote graph (b) NetHEPT graph

(c) Flixster graph with topic 1 (d) Flixster graph with topic 2

Figure 4: Size of selected seed sets vs. coverage threshold η under a fixed probability threshold $P = 0.5$.

the test results on three datasets. All test results consistently show that our algorithm performances the best, and sometimes with a significant improvement over the Random, High-degree and PageRank heuristics. In particular, for wiki-Vote and $P = 0.1$ (Figure 3(a)), on average our algorithm MinSeed-PCG[ε] selects seed sets with size 88.2% less than those selected by Random, 20.2% less than High-degree, and 30.9% less than PageRank. For NetHEPT and $P = 0.1$ (Figure 3(b)), on average our algorithm selects seed sets with size 56.7% less than Random, 46.0% less than High-degree, and 24.4% less than PageRank. The High-degree heuristic performs close to MinSeed-PCG[ε] in wiki-Vote, but performs badly in NetHEPT, even worse than Random when η is large. This shows that High-degree is not a good and stable heuristic for this task. For Flixster with topic 1 and $P = 0.1$ (Figure 5(c)), on average MinSeed-PCG[ε] selects seed sets with size 94.4% less than Random, 54.0% less than High-degree, and 29.2% less than PageRank. For Fixster with topic 2 and $P = 0.1$ (Figure 5(d)), on average MinSeed-PCG[ε] selects seed sets with size 91.1% less than Random, 73.0% less than High-degree, and 24.4% less than PageRank. Figures 4 show the results for $P = 0.5$. The curves are almost the same as the corresponding ones for $P = 0.1$. This can be explained by the sharp phase transition to be observed in the next set of tests, which is due to concentration of influence coverage, such that typically only a few tens of more seeds would satisfy probability threshold P from 0.1 to 0.5.

Our second set of tests is to fix a coverage threshold η, and observe the change of coverage probability $\Pr(Inf(S) \geq \eta)$ as the seed set S grows as computed by various algorithms. Figure 5 shows the test results. Wiki-Vote, NetHEPT and Flixster with topic 2 (Figure 5(a), (b), (d)) have sharp phase transition: there is a short range of seed set size where the probability increases very fast from 0.01 to very close to 1 (only several nodes are needed to reach a 0.1 increment in probability). While Filxster with topic 1 (Figure 5(c)) has a relatively smooth phase transition. This phase transition



(a) wiki-Vote graph, $\eta = 3000, 4500$

(b) NetHEPT graph, $\eta = 6000, 10500$

(c) Flixster graph with topic 1, $\eta = 2000, 4000$

(d) Flixster graph with topic 2, $\eta = 2000, 4000$

Figure 5: Size of selected seed sets vs. probability threshold P under a fixed coverage threshold η.

phenomenon is clearly due to the concentration of influence coverages of seed sets, as already verified in Figure 2.

In all our tests, MinSeed-PCG[ε] performances the best: its phase transition comes first before the other algorithms, which means it uses less number of seeds to achieve the same probability threshold P. Random performs much worse than MinSeed-PCG[ε], while PageRank and High-degree perform close to MinSeed-PCG[ε] when η is small, but noticeably worse than MinSeed-PCG[ε] when η gets larger. For wiki-Vote, on average MinSeed-PCG[ε] selects a seed set with size 34.1% less than PageRank, 27.7% less than High-degree, and 86.4% less than Random when $\eta = 3,000$. When $\eta = 4,500$, MinSeed-PCG[ε] selects a seed set with size on average 38.8% less than PageRank, 30.8% less than High-degree, and 76.3% less than Random. For NetHEPT, when $\eta = 6,000$, on average MinSeed-PCG[ε] selects a seed set with size 22.8% less than PageRank, 51.8% less than High-degree, and 59.2% less than Random. When $\eta = 10,500$, on average MinSeed-PCG[ε] selects a seed set with size 36.1% less than PageRank, 52.9% less than High-degree, and 49.6% less than Random. For Flixster with topic 1, when $\eta = 2,000$, on average the output number of seeds by MinSeed-PCG[ε] is 44.1% less than PageRank, 78.9% less than High-degree, and 98.3% less than Random. When $\eta = 4,000$, the corresponding results are 53.2%, 70.7% and 93.9%. For topic 2, when $\eta = 2,000$, the output number of seeds by MinSeed-PCG[ε] is 59.0% less than PageRank, 78.6% less than High-degree, and 95.8% less than Random. When $\eta = 4,000$, the corresponding results are 54.9%, 76.2% and 89.0%.

For all these graphs, we do not test the case when η is very close to the number of nodes. Since in this case a large seed set close to the full node set is needed, and greedy-based seed selection loses its advantage comparing to simple random or high-degree heuristics when a large number of seeds are needed. Moreover, we believe that requiring η to be close to the full network size is not a realistic scenario in practice.

As a summary, our experimental results validate that influence coverages of seed sets are concentrated well in real-world networks, and thus support the claim that our algorithm provides good approximation guarantee. Moreover, our algorithm performs much better than simple baseline algorithms, achieving significant savings on seed set size.

8. FUTURE WORK

This study may inspire a number of future directions. One is to study the concentration property of other classes of graphs, especially graphs close to real-world networks such as power-law graphs, to see if we can analytically prove that a large class of graphs have good concentration property on influence coverage distributions. Another direction is to speed up the estimation of $\Pr(\mathit{Inf}(S) \geq \eta)$, which is done by Monte Carlo simulation in this work and is slow. One may also study influence maximization problem where reaching the tipping point is the first step, which is followed by further diffusion steps. Our algorithm and results may be an integral component of such influence maximization tasks.

9. REFERENCES

[1] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 81–90. IEEE, 2012.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[3] N. Chen. On the approximability of influence in social networks. *SIAM Journal on Discrete Mathematics*, 23(3):1400–1415, 2009.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD'10*, pages 1029–1038. ACM, 2010.

[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD'09*, pages 199–208, 2009.

[6] W. Chen, Y. Yuan, and L. Zhang. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. In *ICDM'10*, pages 88–97, 2010.

[7] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD'01*, pages 57–66. ACM, 2001.

[8] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in Neural Information Processing Systems*, pages 3147–3155, 2013.

[9] U. Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[10] M. Gladwell. *The Tipping Point:How Little Things Can Make a Big Difference*. Back Bay Books, 2002.

[11] S. Goldberg and Z. Liu. The Diffusion of Networking Technologies. In *SODA'13*, pages 1577–1594, 2013.

[12] A. Goyal, F. Bonchi, L. V. Lakshmanan, and S. Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social Network Analysis and Mining*, pages 1–14, 2012.

[13] A. Goyal, W. Lu, and L. V. S. Lakshmanan. SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model. In *ICDM'11*, pages 211–220, 2011.

[14] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146. ACM, 2003.

[15] J. Leskovec. Wiki-vote social network. `http://snap.stanford.edu/data/wiki-Vote.html`.

[16] C. Long and R.-W. Wong. Minimizing seed set for viral marketing. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 427–436. IEEE, 2011.

[17] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD'02*, pages 61–70. ACM, 2002.

[18] M. G. Rodriguez and B. Schölkopf. Influence maximization in continuous time diffusion networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 313–320, 2012.

[19] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[20] P. Zhang, W. Chen, X. Sun, Y. Wang, and J. Zhang. Minimizing seed set selection with probabilistic coverage guarantee in a social network. *arXiv preprint arXiv:1402.5516*, 2014.