Further Sketching for Reverse Sampling in Influence Maximization

Xiaolong CHEN
The Hong Kong University of Science and Technology (Guangzhou)
Nansha, Guangzhou, China
xchen738@connect.hkust-gz.edu.cn

ABSTRACT

Influence Maximization (IM) is an important algorithmic problem in social influence analysis, which aims to select a set of k nodes in a social network as the source of influence spread to maximize the expected number of influenced nodes. There has been a bunch of work putting forward efficient and effective algorithms to give approximate solution to this problem, with theoretical guarantees. However, the time complexity can still be improved. Inspired by a recent work which provides an approximation algorithm for k-cover problem, we propose an algorithm to further improve the time complexity of to give a $(1 - 1/e - \varepsilon)$-approximate solution to the IM problem.

CCS CONCEPTS

• **Mathematics of computing → Graph algorithms**; • **Information systems → Social networks**.

KEYWORDS

Social Networks, Influence Maximization, Set Coverage

ACM Reference Format:
Xiaolong CHEN. 2022. Further Sketching for Reverse Sampling in Influence Maximization. In *Proceedings of (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The prevalence of online social networks (OSNs) during the last decades has prompted much attention on information diffusion, which is powerfull in many applications [10], such as political campaign and viral marketing [7], etc. A key problem related to information diffusion study is the influence maximization (IM) problem, which aims to select k users in an OSN with the maximum influence spread, i.e., the expected number of influenced users after the propagation process. Kempe et al. [11] formulated IM problem as a combinatorial optimization problem and showed that it is NP-hard. After that, a lot of approximation algorithms have been proposed [9, 12, 17, 18], among which the most powerful kind of Such algorithms come with theoretical guarantees. However, despite the marvelous performance of these algorithms, the theoretical guarantee often turns out to be loose and we often overestimate the cost we need to reach a certain accuracy.

Among coverage problems, the most basic one is k-cover problem. Given a ground set $\mathcal{E}$ of $n_{\mathcal{E}}$ elements and a family $\mathcal{S} \subseteq 2^{\mathcal{E}}$ of $n_{\mathcal{S}}$ subsets of the elements. The coverage function is defined as $C(S) = |\bigcup_{U \in S} U|$ for any $S \in \mathcal{S}$. k-cover problem aims to find S with size k in $\mathcal{S}$ that maximize $C(S)$. That is, to find k sets that maximizes the union of covered elements. This setting can be formulated as a bipartite graph, where one type of nodes represents sets and the other represents elements. If an edge connects a set and an element, it means that the set contains the element.

This work basically further reduces the cost of giving an approximate solution to the IM problem, by utilizing the idea of sketching in k-cover approximate algorithms.

The report is organized as follows. Section 2 will introduce some previous works that are closely related to this project. Section 3 illustrates our method and gives proofs for the theoretical guarantee. Our experiment and discussion are given in section 4. Finally, section 5 gives a brief summary of this work.

The notations used in this report and corresponding explanations are given in Table 1.

Notation	Description
G	Network
V	The set of nodes
E	The set of edges
n	Number of nodes, $n = \|V\|$
m	Number of edges, $m = \|E\|$
k	Size of seed set
S	Seed set
$I(S)$	The number of influenced nodes from S
ε_0	Error term when using θ RR sets
ε	Error term introduced by sketching
θ	Number of RR-sets needed to give a $(1 - 1/e - \varepsilon_0)$ solution. (TIM or IMM)
θ_p	Number of RR sets for a fixed p
$\mathcal{R}$	All θ RR-sets
$\mathcal{R}_p$	All θ_p RR-sets
$N_{\mathcal{R}}(S)$	Number of RR sets in $\mathcal{R}$ that intersects with S
$F_{\mathcal{R}}(S)$	The fraction of RR sets in $\mathcal{R}$ that are covered by S
Opt_k	The maximum number of RR sets that can be covered by a seed set of size k
OPT	The influence spread of the optimum solution.
R_i	i-th RR-set

Table 1: Notations

Conference acronym 'XX, ,

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

2.1 Influence Maximization

In 2003, Kempe et al. [11] firstly presented the algorithmic study on the influence maximization problem. They showed that this problem is NP-hard in general and proposed to use a greedy algorithm to approximate the solution with a factor of $1 - 1/e - \varepsilon$. The key idea is to use Monte Carlo simulation [14] to estimate the expected influence spread (i.e. $\mathbb{E}[I(S)]$). With this approach, Kempe's algorithm greedily selects the node in G that gives the largest increment of influence, until k nodes are chosen. However, the time complexity is too high for large graphs. There are a lot of work [3, 5, 6, 9, 12] trying to improve the efficiency of this algorithm. Among these techniques, the most famous ones are CELF [12] and CELF++ [9], which make use of the monotonicity and submodularity of the influence function. Although they have the same worst time complexity as simple greedy algorithm, it turns out that they can provide higher empirical efficiency.

In 2014, Borgs et al. [4] made a great breakthrough by presenting a near-linear time algorithm under the IC model. They proposed the idea of reverse influence sampling (RIS), which significantly improve the time complexity for giving a $(1 - 1/e - \varepsilon)$-approximate solution with high probability. From a high-level perspective, it converts the IM problem to a k-cover problem. After that, Tang et al. [18] pointed out the deficiency of Borgs's algorithm and proposed Tow-phase Influence Maximization algorithm (i.e., TIM), which makes it more efficient in practice. In 2015, they took a martingale approach [17] and further improve the algorithm (i.e., IMM). Later, Tang et al. proposed a more efficient online processing algorithm based on RIS and a novel way to compute the empirical guarantee [16], which is tighter than the theoretical guarantee.

2.2 Coverage Problem

Coverage problem in the context of set arrival models has been studied extensively in previous years [1, 8, 15]. Plenty of algorithms with approximation guarantees have been proposed. In 2017, McGregor et al. [13] and Bateni et al. [2] presented approximation algorithms for k-cover problem in the streaming settings, respectively. Both of their methods have $\tilde{O}(n)$ space complexity.

3 SKETCHING FOR INFLUENCE MAXIMIZATION

3.1 IM as a Coverage Problem

In 2014, Borgs et al. [4] proposed the idea of reverse sampling for the IM problem, which is called Reverse Influence Sampling (RIS) in other following works. To illustrate their method, two important concepts will be introduced first:

Definition 3.1 (Reverse Reachable Set). *Let v be a node in G, and g be the graph obtained by removing each edge in G with probability $1 - p(e)$. The reverse reachable set (RR set) for v is the set of nodes in g that can reach v.*

Definition 3.2 (Random RR Set). *Suppose $\mathcal{G}$ is the distribution of g induced by the randomness in edge removals from G. A random RR set is defined as an RR set generated by a random sample from $\mathcal{G}$, for a node selected uniformly at random from g.*

Definition 3.3. *When we say a RR set is **covered** by a node set S, we mean that the RR set contains at least one node in S.*

The idea is to sample a large number of RR sets. For k iterations, select the node that covers the most RR sets and remove the covered RR sets. Algorithm 2 shows a simple implementation of RIS. By considering every node as a set and every RR set as an element, this process can be formulated as the greedy algorithm for k-cover problem. An example is shown in Figure 1.

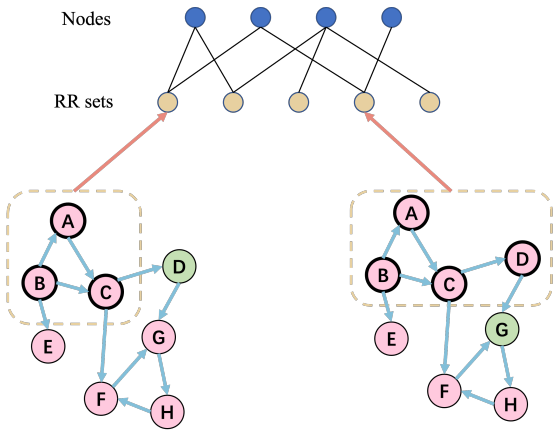


Figure 1: Illustration of connecting RIS and k-cover problem.

Algorithm 1 `nodeselection`

Require: $\mathcal{R}, k$

$S_k \leftarrow \emptyset$
for $i = 1$ to k **do**
  $v \leftarrow \arg\max_u F_{\mathcal{R}}(S \cup \{u\}) - F_{\mathcal{R}}(S)$
  $S_k \leftarrow S_k \cup \{u\}$
end for
return S_k

Algorithm 2 simple implementation of RIS

Require: G

$\mathcal{R} \leftarrow$ Sample θ RR sets in G
return `nodeselection`$(\mathcal{R}, k)$

3.2 Our Methods

Although there have been works successfully reducing the number of RR sets needed to provide a solution with theoretical guarantee [16–18], no previous work consider approximating the solution from the perspective of coverage problem. Combining approximation for k-cover problem and IM, we propose algorithm 3, which further reduces the number of RR sets to provide an approximate solution that is theoretically guaranteed. The argument is formulated as Theorem 3.4.

Algorithm 3 Sketch for IM

Require: δ'', G, ε, k, θ, r
1: $\delta \leftarrow \delta'' + \log\log_{\frac{1}{1-\varepsilon}} \theta$
2: $n_R = \min(\frac{12(\delta + k\log n)}{r\varepsilon^2}, \theta)$
3: $\mathcal{R} \leftarrow \emptyset$
4: **while** Number of RR sets does not reach n_R **do**
5: Select a node u from G uniformly at random
6: $S \leftarrow$ `GetRRSet`(u)
7: $\mathcal{R} \leftarrow \mathcal{R} \cup \{S\}$
8: **end while**
9: **return** `nodeselection`$(\mathcal{R}, k)$

Theorem 3.4. *For any θ that gives $(1-1/e-\varepsilon_0)$-approximate solution to the IM problem, Algorithm 3 will return a $(1-1/e-\varepsilon_0-3\varepsilon)$-approximate solution, with probability at least $1-e^{-\delta}-n^{-l}$.*

The proof of the theorem will be given in the end of this section. Before that, we will introduce the main idea and several helpful lemmas.

Suppose θ RR sets will give a $(1-1/e-\varepsilon_0)$-approximate solution to the IM problem. Consider the following process, we generate a random number from a uniform distribution on $[0,1]$ and set p as the threshold. That is, if $p_i > p$, we will throw away the corresponding RR set R_i. This process will give a subset of $\mathcal{R}$, which is denoted as $\mathcal{R}_p$. First, we show that for a certain p, the number of RR sets in $\mathcal{R}_p$ that are covered by S can be used to estimate that in $\mathcal{R}$.

Lemma 3.5. *For any $\delta' \geq 1$, pick $\frac{6\delta'}{\varepsilon^2 Opt_k} \leq p \leq 1$, and let S be an arbitrary subset of V s.t. $|S| \leq k$. With probability $1-e^{-\delta'}$, we have*

$$\left|\frac{1}{p}N_{\mathcal{R}_p}(S) - N_{\mathcal{R}}(S)\right| \leq \varepsilon Opt_k \tag{1}$$

Proof: Let X_i be a r.v indicating whether R_i is abandoned or not. That is, $X_i = 1$ if $h(R_i) \leq p$, 0 otherwise. Then we have

$$\sum_{i=1}^{N_{\mathcal{R}}(S)} X_i = N_{\mathcal{R}_p}(S) \tag{2}$$

$$\mathbb{E}[\sum_{i=1}^{N_{\mathcal{R}}(S)} X_i] = \sum_{i=1}^{N_{\mathcal{R}}(S)} \mathbb{E}[X_i] = pN_{\mathcal{R}}(S). \tag{3}$$

Inspired by Chernoff bounds,

$$\begin{aligned}
&\Pr\left[\left|\sum_{i=1}^{N_{\mathcal{R}}(S)} X_i - \mathbb{E}[\sum_{i=1}^{N_{\mathcal{R}}(S)} X_i]\right| \geq \varepsilon'\mathbb{E}[\sum_{i=1}^{N_{\mathcal{R}}(S)} X_i]\right] \\
&= \Pr[|N_{\mathcal{R}_p}(S) - pN_{\mathcal{R}}(S)| \geq \varepsilon' pN_{\mathcal{R}}(S)] \\
&\leq 2\exp\left[\frac{-\varepsilon'^2 pN_{\mathcal{R}}(S)}{3}\right]
\end{aligned} \tag{4}$$

Let $\varepsilon' = \varepsilon \cdot \frac{\mathrm{Opt}_k}{N_{\mathcal{R}}(S)}$,

$$\begin{aligned}
&\Pr\left[|N_{\mathcal{R}_p}(S) - pN_{\mathcal{R}}(S)| \geq \varepsilon p\mathrm{Opt}_k\right] \\
&\leq 2\exp\left(\frac{-\varepsilon^2\frac{\mathrm{Opt}_k^2}{N_{\mathcal{R}}^2(S)}pN_{\mathcal{R}}(S)}{3}\right) \\
&= 2\exp\left(-\frac{p\varepsilon^2\mathrm{Opt}_k^2}{3N_{\mathcal{R}}(S)}\right) \\
&= 2\exp\left(-\frac{6\delta'}{\varepsilon^2\mathrm{Opt}_k}\cdot\frac{\varepsilon^2\mathrm{Opt}_k^2}{3N_{\mathcal{R}}(S)}\right) \\
&= 2\exp\left(-\frac{2\delta'\mathrm{Opt}_k}{N_{\mathcal{R}}(S)}\right) \\
&\leq 2\exp\left(-2\delta'\right) \\
&\leq \exp\left(1-2\delta'\right) \\
&\leq \exp\left(-\delta'\right).
\end{aligned} \tag{5}$$

Then,

$$\begin{aligned}
&\Pr\left[\left|N_{\mathcal{R}_p}(S) - pN_{\mathcal{R}}(S)\right| \leq \varepsilon p\mathrm{Opt}_k\right] \\
&\geq 1-\exp\left(-\delta'\right)
\end{aligned} \tag{6}$$

That is,

$$\Pr\left[\left|\frac{1}{p}N_{\mathcal{R}_p}(S) - N_{\mathcal{R}}(S)\right| \leq \varepsilon\mathrm{Opt}_k\right] \geq 1-e^{-\delta'} \tag{7}$$

Then we are able to prove that any α-approximation to the k-cover problem on $\mathcal{R}_p$ is an $(\alpha-2\varepsilon)$-approximate solution to the k-cover problem on $\mathcal{R}$.

Lemma 3.6. *Pick $\frac{6(\delta+k\log(n))}{\varepsilon^2 Opt_k} \leq p \leq 1$. With probability $1-e^{-\delta}$, any α-approximation solution to k-cover for $\mathcal{R}_p$ is an $(\alpha-2\varepsilon)$-approximation solution for that of $\mathcal{R}$. And for any S s.t. $|S| \leq k$, $\left|\frac{1}{p}N_{\mathcal{R}_p}(S) - N_{\mathcal{R}}(S)\right| \leq \varepsilon Opt_k$.*

Proof: Let $\delta' = \delta + k\log(n)$. By Lemma 3.5, with probability $1-e^{-(\delta+k\log(n))}$,

$$\left|\frac{1}{p}N_{\mathcal{R}_p}(S) - N_{\mathcal{R}}(S)\right| \leq \varepsilon\mathrm{Opt}_k.$$

We have $\binom{n}{k}$ different sets of size k. By union bound, with probability $1-\binom{n}{k}e^{-(\delta+k\log(n))} \geq 1-n^k e^{-(\delta+k\log(n))} = 1-e^{-\delta}$, all the choices satisfy inequality (1).

Let S^* be solution to maximize $F_{\mathcal{R}}(S)$, which means $F_{\mathcal{R}}(S*) = \frac{\mathrm{Opt}_k}{\theta}$. And suppose S is an α-approximation solution to maximize $F_{\mathcal{R}_p}(S)$. We simultaneously have

$$\left|\frac{1}{p}N_{\mathcal{R}_p}(S^*) - \mathrm{Opt}_k\right| \leq \varepsilon\mathrm{Opt}_k, \tag{8}$$

$$\left|\frac{1}{p}N_{\mathcal{R}_p}(S) - N_{\mathcal{R}}(S)\right| \leq \varepsilon\mathrm{Opt}_k, \tag{9}$$

$$N_{\mathcal{R}_p}(S) \geq \alpha N_{\mathcal{R}_p}(S_p) \geq \alpha N_{\mathcal{R}_p}(S^*). \tag{10}$$

By inequality (8),

$$N_{\mathcal{R}_p}(S^*) \geq p(1-\varepsilon)\mathrm{Opt}_k. \tag{11}$$

From inequality (9),

$$\begin{aligned} N_{\mathcal{R}}(S) &\geq \frac{1}{p} N_{\mathcal{R}}(S) - \varepsilon \mathrm{Opt}_k \\ &\geq \frac{1}{p} \alpha N_{\mathcal{R}_p}(S^*) - \varepsilon \mathrm{Opt}_k \\ &\geq \alpha(1-\varepsilon)\mathrm{Opt}_k - \varepsilon \mathrm{Opt}_k \\ &\geq (\alpha - 2\varepsilon)\mathrm{Opt}_k \end{aligned} \tag{12}$$

It has been proved that the fraction of RR sets that are covered by S can be used to estimate the expected influence spread.

Lemma 3.7 ([18]). *Given a θ satisfying the corresponding condition [17, 18], for any set S of at most k nodes, the following inequality holds with probability at least $1 - n^{-l}/\binom{n}{k}$.*

$$|nF_{\mathcal{R}}(S) - \mathbb{E}[I(S)]| \leq \frac{\varepsilon_0}{2} OPT \tag{13}$$

The following lemma suggests that $N_{\mathcal{R}_p}(S)$ is bounded by some form that is irrelevant to Opt_k.

Lemma 3.8. *For any arbitrary $C \geq 1$, let $p = \frac{6(\delta + k\log n)}{\varepsilon^2 Opt_k}$. With probability $1 - e^{-\delta}$, we have*

$$\max_{S \in V, |S|=k} N_{R_p}(S) \leq \frac{12(\delta + k\log n)}{\varepsilon^2} \tag{14}$$

Proof: By applying Lemma 3.5 to $\max_{S\in V,|S|=k} N_{R_p}(S)$, with probability $1 - e^{-\delta'}$, we have

$$\begin{aligned} \max_{S \in V, |S|=k} N_{R_p}(S) &\leq pN_R(S) + \varepsilon p \mathrm{Opt}_k \\ &\leq p\mathrm{Opt}_k + \varepsilon p \mathrm{Opt}_k \\ &\leq p(1+\varepsilon)\mathrm{Opt}_k \\ &\leq 2p\mathrm{Opt}_k \\ &= \frac{12(\delta + k\log n)}{\varepsilon^2 \mathrm{Opt}_k} \mathrm{Opt}_k \\ &= \frac{12(\delta + k\log n)}{\varepsilon^2} \end{aligned}$$

But the above result is not enough for us to bound θ_p. So we introduce the following assumption that connects Opt_k and θ_p. This assumption indicates that at least r of the RR sets will be covered by the optimal solution.

Assumption 3.9. *For the set cover problem considered in this report,*

$$Opt_k(p) \geq r\theta_p \tag{15}$$

Now we are able to prove Theorem 3.4.

Proof of Theorem 3.4: Combining Assumption 3.9 and 3.8, we can give a bound for θ_p.

$$\theta_p \leq \frac{\mathrm{Opt}_k(p)}{r} \leq \frac{12(\delta + k\log n)}{r\varepsilon^2} \tag{16}$$

Suppose the probability resulting $\frac{12(\delta+k\log n)}{r\varepsilon^2}$ RR sets is p^*. Then $\mathrm{Opt}_k(p^*) \geq \frac{12(\delta+k\log n)}{\varepsilon^2}$, from which we can deduce that $p^* \geq p$.

Use S_p to denote the set returned by Algorithm 3, S^+ to denote the set that covers the most RR sets in $\mathcal{R}$ (i.e. $F_{\mathcal{R}}(S^+) = \frac{Opt_k}{\theta}$) and S^o to denote the optimal solution for the IM problem. According to Lemma 3.7,

$$\begin{aligned} \mathbb{E}[I(S_p)] &\geq nF_{\mathcal{R}}(S_p) - \frac{\varepsilon_0}{2} OPT \\ &= n\frac{N_{\mathcal{R}(S_p)}}{\theta} - \frac{\varepsilon_0}{2} OPT \\ &\geq n(1 - 1/e - 2\varepsilon)F_{\mathcal{R}}(S^+) - \frac{\varepsilon_0}{2} OPT \\ &\geq n(1 - 1/e - 2\varepsilon)F_{\mathcal{R}}(S^o) - \frac{\varepsilon_0}{2} OPT \\ &\geq (1 - 1/e - 2\varepsilon)(1 - \frac{\varepsilon_0}{2})OPT - \frac{\varepsilon_0}{2} OPT \\ &> (1 - 1/e - \varepsilon_0 - 2\varepsilon)OPT \end{aligned} \tag{17}$$

4 EXPERIMENTS

4.1 Settings

The information of datasets used in the experiment is presented in Table 2. The first four are small datasets and the rest are relatively large datasets. We set $r = 1$ for small datasets and $r = 0.4$ for large datasets. In the experiment, we set $\varepsilon = 0.4$. For smaller datasets, we implement three methods, which are CELF, IMM and our method to illustrate the power of RIS framework. And due to the high time complexity of simulation-based methods, we only implement IMM and our method to see the improvement brought by our method.

Name	n	m
Karate	34	78
Sparrow	52	454
Dolphins	62	159
Email-enron	143	623
weaver	445	1335
Retweet-copen	761	1029
Wiki-votes	889	2914
Email-univ	1133	5451

Table 2: Datasets

4.2 Results

The results are shown in Figure 2 and 3. We can see that our methods can obtain good results with less time cost. With the Email-univ dataset, the effect of ε on the expected spread and time cost for $k = 5$ is shown in Figure 4.

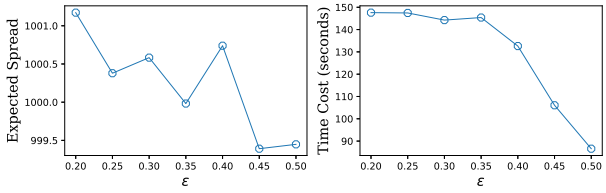
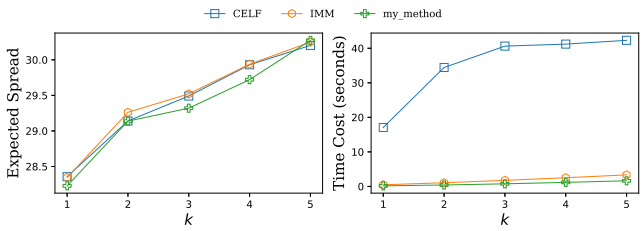
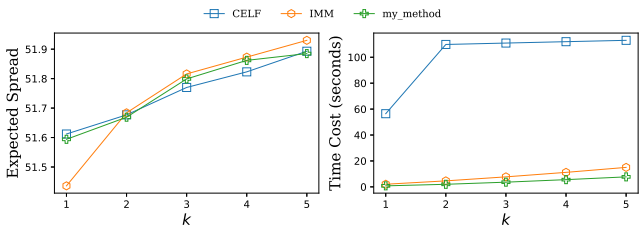


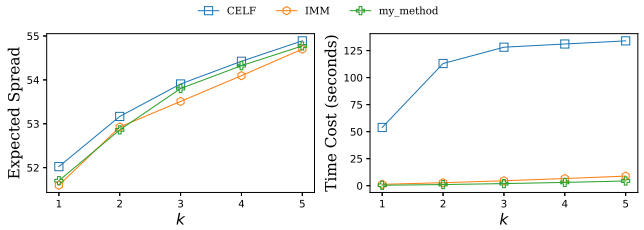
Figure 4: The effect of ε on the expected spread and time cost, with Email-univ dataset.



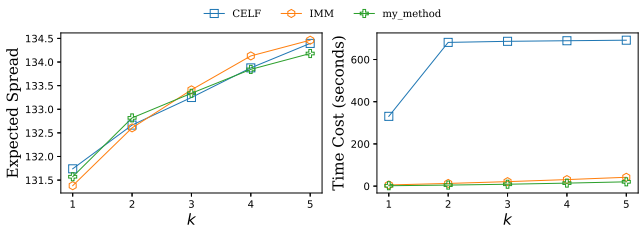
(a) Karate



(b) Sparrow

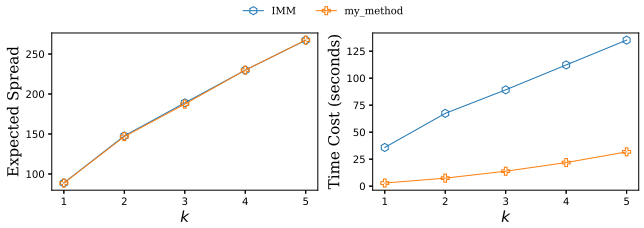


(c) Dolphins

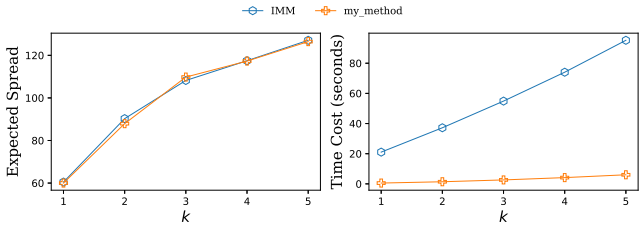


(d) Email-enron

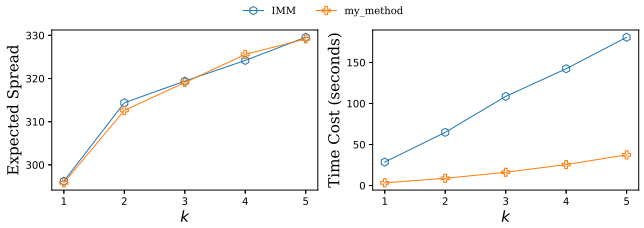
Figure 2: Expected influence spread and time cost of each method for the four small datasets that we look into.



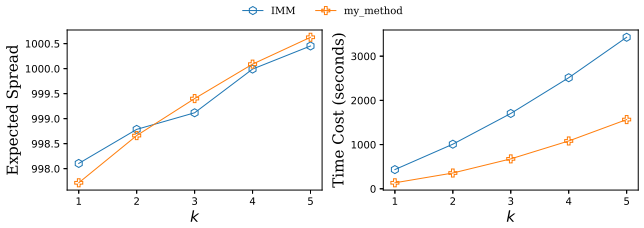
(a) Weaver



(b) Retweet-copen



(c) Wiki-votes



(d) Email-univ

Figure 3: Expected influence spread and time cost of each method for the four larger datasets that we look into.

We can see that with larger ε, the spread of the solution set will decrease. Although the monotonicity is not guaranteed due to the randomness of the influence propagation, the overall tendency can be observed. The effect of ε on time cost is more obvious. If we allow the error to be large, we can use less time to get the result. One may notice that when $\varepsilon \leq 0.35$, the time cost does not change much. The reason is that when ε is small, $p \geq \frac{6(\delta+k\log(n))}{\varepsilon^2 \mathrm{Opt}_k}$ will be larger than 1 and since the probability is bounded by 1, we will simply use θ as the number of RR sets, which means our method does not reduce the time cost in this scenario.

However, we should notice that the experiment setting is actually not fair for our method since the theoretical here is not valid ($1 - 1/e - 2\varepsilon < 0$). But it still gives good results. This is because the theoretical guarantee is naturally loose. The reason why this report does not choose a valid setting is that if we use smaller ε, basically the sketching behavior will not be observed since $p > 1$. Due to the large constant factor in p, this method only takes effect for very large datasets, which, unfortunately, we cannot process due to the programming language issue and limited computation resource.

5 CONCLUSION

To summarize, by utilizing the idea of sketching in approximate algorithms for k-cover problem, we propose a method to further reduce the number of RR sets we need to provide an approximate solution to the IM problem. In our experiment, we show that RIS framework significantly outperforms the simulation-based method and that our method brings slight improvement for IMM.

From a high-level perspective, the essence of this work is rather simple. The RIS framework finally formulates the IM problem as a k-cover problem to give an approximate solution. And it has been proved that solving k-cover problem on the sketch of the original bipartite graph can give an approximate solution to the complete graph. What we do is just use the idea in k-cover approximation in IM problem and further reduce the number of RR sets that we need. But the theoretical proof is not trivial, which is also the key contribution of this report.

REFERENCES

[1] Sepehr Assadi, Sanjeev Khanna, and Yang Li. 2016. Tight bounds for single-pass streaming complexity of the set cover problem. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 698–711.

[2] MohammadHossein Bateni, Hossein Esfandiari, and Vahab Mirrokni. 2017. Almost optimal streaming algorithms for coverage problems. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures*. 13–23.

[3] Shishir Bharathi, David Kempe, and Mahyar Salek. 2007. Competitive influence maximization in social networks. In *International workshop on web and internet economics*. Springer, 306–311.

[4] Christian Borgs, Michael Brautbar, Jennifer Chayes, and Brendan Lucier. 2014. Maximizing social influence in nearly optimal time. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 946–957.

[5] Wei Chen, Chi Wang, and Yajun Wang. 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1029–1038.

[6] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 199–208.

[7] Pedro Domingos and Matt Richardson. 2001. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 57–66.

[8] Yuval Emek and Adi Rosén. 2016. Semi-streaming set cover. *ACM Transactions on Algorithms (TALG)* 13, 1 (2016), 1–22.

[9] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011. Celf++ optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*. 47–48.

[10] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record* 42, 2 (2013), 17–28.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[12] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 420–429.

[13] Andrew McGregor and Hoa T Vu. 2019. Better streaming algorithms for the maximum coverage problem. *Theory of Computing Systems* 63, 7 (2019), 1595–1619.

[14] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.

[15] Barna Saha and Lise Getoor. 2009. On maximum coverage in the streaming model & application to multi-topic blog-watch. In *Proceedings of the 2009 siam international conference on data mining*. SIAM, 697–708.

[16] Jing Tang, Xueyan Tang, Xiaokui Xiao, and Junsong Yuan. 2018. Online processing algorithms for influence maximization. In *Proceedings of the 2018 International Conference on Management of Data*. 991–1005.

[17] Youze Tang, Yanchen Shi, and Xiaokui Xiao. 2015. Influence maximization in near-linear time: A martingale approach. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 1539–1554.

[18] Youze Tang, Xiaokui Xiao, and Yanchen Shi. 2014. Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 75–86.

Received DD MM YYYY; revised DD MM YYYY; accepted DD MM YYYY