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

Peng Zhang¹, Wei Chen², Xiaoming Sun³,
Yajun Wang², Jialin Zhang³

1. Purdue University
2. Microsoft
3. Institute of Computing Technology, CAS
Background
Background
How to select most “influential” people in social network?
Common Framework

Most of the work is based on optimization of submodular functions. E.g., Influence Maximization [Kempe et al., KDD’03], Seed Minimization [Long et al., ICML’11, Goyal et al., SNAM’12].
Common Framework

Most of the work is based on optimization of submodular functions. E.g., Influence Maximization [Kempe et al., KDD’03], Seed Minimization [Long et al., ICML’11, Goyal et al., SNAM’12].

\[ f(\cdot) : 2^V \rightarrow R \text{ is submodular if for any } S \subseteq T \subseteq V \text{ and any } u \in V \setminus T, f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T). \]
Common Framework

Most of the work is based on optimization of submodular functions. E.g., Influence Maximization [Kempe et al., KDD’03], Seed Minimization [Long et al., ICML’11, Goyal et al., SNAM’12].

\[ f(\cdot) : 2^V \to R \] is submodular if for any \( S \subseteq T \subseteq V \) and any \( u \in V \setminus T \),
\[ f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T). \]

Greedy algorithm

- \((1 - \frac{1}{e})\) - approximation for influence maximization;
- \(\ln n\) - approximation for seed minimization.
What about *nonsubmodular* influence maximization/seed minimization?
Hot Topic

Hot Topic


- To become a “hot topic”
  - # of people discussing the topic reaches a *threshold*;
  - require certain *probabilistic* guarantee.
Problem Definition

Seed Minimization with Probabilistic Coverage Guarantee (SM-PCG)

*Input:* graph $G = (V, E)$ with $|V| = n$, target set $U$ with $|U| = m$, influence diffusion model, coverage threshold $\eta < |U|$, probability threshold $P \in (0, 1)$.

*Output:*

$$S^* = \arg\min_{S: \Pr(\Inf_U(S) \geq \eta) \geq P} |S|.$$  

$\Inf_U(S)$: # of nodes in $U$ activated by seed set $S$ under the specific influence diffusion model.
An Example: Independent Cascade Model

[Kempe, Kleinberg and Tardos, KDD’03]
Nodes in seed set $S$ are active, while others are inactive.
An Example: Independent Cascade Model

[Kempe, Kleinberg and Tardos, KDD’03]
Once $u$ is activated, $u$ has a single chance to activate inactive $v$ successfully w. p. on edge $(u, v)$. 
An Example: Independent Cascade Model

[Kempe, Kleinberg and Tardos, KDD’03]

$\text{Inf}_U(S)$ is the $\#$ of active nodes in set $U$. 
Nonsubmodularity of SM-PCG

\[ f_{\eta}(S) = \Pr(\inf U(S) \geq \eta) \]
\[ S^* = \arg\min_{\sum(S) \geq P \mid S} \]

\[ g_P(S) = \max_{\eta': \Pr(\inf U(S) \geq \eta') \geq P \eta'} \]
\[ S^* = \arg\min_{\sum(S) \geq \eta} \mid S \mid \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\text{Inf}_U(S) \geq \eta) \]
\[ S^* = \arg\min_{f_\eta(S) \geq P} |S| \]

\[ g_P(S) = \max_{\eta': \Pr(\text{Inf}_U(S) \geq \eta') \geq P} \eta' \]
\[ S^* = \arg\min_{g_P(S) \geq \eta} |S| \]

Edge probabilities are 1, \( \eta = 5 \).
\[ f_\eta(S \cup \{u\}) - f_\eta(S) = 0 \]
\[ f_\eta(T \cup \{u\}) - f_\eta(T) = 1 \]

Edge probabilities are 0.5, \( P = 0.8 \).
\[ g_P(S \cup \{c\}) - g_P(S) = 0 \]
\[ g_P(T \cup \{c\}) - g_P(T) = 1 \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\text{Inf}_U(S) \geq \eta) \]
\[ S^* = \arg\min_{f_\eta(S) \geq P} |S| \]

\[ g_P(S) = \max_{\eta': \Pr(\text{Inf}_U(S) \geq \eta') \geq P} \eta' \]
\[ S^* = \arg\min_{g_P(S) \geq \eta} |S| \]

edge probabilities are 1,
\[ \eta = 5. \]
\[ f_\eta(S U \{u\}) - f_\eta(S) = 0 \]
\[ f_\eta(T U \{u\}) - f_\eta(T) = 1 \]

edge probabilities are 0.5,
\[ P = 0.8. \]
\[ g_P(S U \{c\}) - g_P(S) = 0 \]
\[ g_P(T U \{c\}) - g_P(T) = 1 \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\text{Inf}_U(S) \geq \eta) \]
\[ S^* = \arg\min_{f_\eta(S) \geq P} |S| \]

\[ g_P(S) = \max_{\eta': \Pr(\text{Inf}_U(S) \geq \eta') \geq P} \eta' \]
\[ S^* = \arg\min_{g_P(S) \geq \eta} |S| \]

edge probabilities are 1, \( \eta = 5 \).
\[ f_\eta(S \cup \{u\}) - f_\eta(S) = 0 \]
\[ f_\eta(T \cup \{u\}) - f_\eta(T) = 1 \]

edge probabilities are 0.5, \( P = 0.8 \).
\[ g_P(S \cup \{c\}) - g_P(S) = 0 \]
\[ g_P(T \cup \{c\}) - g_P(T) = 1 \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\inf_U(S) \geq \eta) \]
\[ S^* = \arg\min_{f_\eta(S) \geq P} |S| \]
\[ g_P(S) = \max_{\eta': \Pr(\inf_U(S) \geq \eta') \geq P} \eta' \]
\[ S^* = \arg\min_{g_P(S) \geq \eta} |S| \]

Edge probabilities are 1, \( \eta = 5 \).
\[ f_\eta(S \cup \{u\}) - f_\eta(S) = 0 \]
\[ f_\eta(T \cup \{u\}) - f_\eta(T) = 1 \]

Edge probabilities are 0.5, \( P = 0.8 \).
\[ g_P(S \cup \{c\}) - g_P(S) = 0 \]
\[ g_P(T \cup \{c\}) - g_P(T) = 1 \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\inf_U(S) \geq \eta) \]
\[ S^* = \text{argmin}_{f_\eta(S) \geq P} |S| \]

\[ g_P(S) = \max_{\eta'}:\Pr(\inf_U(S) \geq \eta') \geq P \eta' \]
\[ S^* = \text{argmin}_{g_P(S) \geq \eta} |S| \]

edge probabilities are 1, \( \eta = 5 \).
\[ f_\eta(S \cup \{u\}) - f_\eta(S) = 0 \]
\[ f_\eta(T \cup \{u\}) - f_\eta(T) = 1 \]

edge probabilities are 0.5, \( P = 0.8 \).
\[ g_P(S \cup \{c\}) - g_P(S) = 0 \]
\[ g_P(T \cup \{c\}) - g_P(T) = 1 \]
Nonsubmodularity of SM-PCG

\[ f_\eta(S) = \Pr(\text{Inf}_U(S) \geq \eta) \]

\[ S^* = \arg\min_{f_\eta(S) \geq P} |S| \]

\[ g_P(S) = \max_{\eta': \Pr(\text{Inf}_U(S) \geq \eta') \geq P} \eta' \]

\[ S^* = \arg\min_{g_P(S) \geq \eta} |S| \]

edge probabilities are 1, \( \eta = 5 \).

\[ f_\eta(S \cup \{u\}) - f_\eta(S) = 0 \]

\[ f_\eta(T \cup \{u\}) - f_\eta(T) = 1 \]

edge probabilities are 0.5, \( P = 0.8 \).

\[ g_P(S \cup \{c\}) - g_P(S) = 0 \]

\[ g_P(T \cup \{c\}) - g_P(T) = 1 \]

Nonsubmodular!
Connect SM-PCG to *Seed Minimization with Expected Coverage Guarantee (SM-ECG)*.

- $E[\ln U(S)]$ is submodular $\rightarrow \ln n + O(1)$ multiplicative error;
- stopping criteria: $\Pr(\ln U(S) \geq \eta) \geq P$ $\rightarrow$ additive error.
Approximation Algorithm

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

**Input:** $G = (V, E), \{p_{u,v}\}_{(u,v) \in E}, U, \eta, P$

**Output:** seed set $S$

1: $S_0 \leftarrow \emptyset$
2: for $i = 1$ to $n$ do
3:    $u \leftarrow \arg\max_v \{E[\inf_U(S_{i-1} \cup \{v\})] - E[\inf_U(S_{i-1})]\}$
4:    $S_i \leftarrow S_{i-1} \cup \{u\}$
5:    if $\hat{\Pr}(\inf_U(S_i) \geq \eta) \geq P + \varepsilon$ then
6:        return $S_i$
7:    end if
8: end for
Analysis

Theorem
Let $S_a$ be the output of MinSeed-PCG[$\varepsilon$] and $a$ is the index. Let

$$c = \max\{\eta - E[\ln f(U(S^*))], 0\} \text{ and } c' = \max\{E[\ln f(S_{a-1})] - \eta, 0\}.$$ 

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

Where,

$$c \leq \sqrt{\frac{\text{Var}(\ln f(U(S^*)))}{P}} \quad \text{and} \quad c' \leq \sqrt{\frac{\text{Var}(\ln f(U(S_{a-1})))}{1 - P - 2\varepsilon}}.$$ 

Remark. Consider $m = \Theta(n)$ and $c + c' = O(\sqrt{m})$, then

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

Theorem
Let $S_a$ be the output of MinSeed-PCG[$\varepsilon$] and $a$ is the index. Let
\[ c = \max\{\eta - E[\inf_U(S^*)], 0\} \quad \text{and} \quad c' = \max\{E[\inf(S_{a-1})] - \eta, 0\}. \]

\[ |S_a| \leq \left\lfloor \ln \left( \frac{\eta n}{m - \eta} \right) \right\rfloor \cdot |S^*| + \frac{(c + c')n}{m - (\eta + c')} + 3. \]

Where,
\[ c \leq \sqrt{\frac{\text{Var}(\inf_U(S^*))}{P}} \quad \text{and} \quad c' \leq \sqrt{\frac{\text{Var}(\inf_U(S_{a-1}))}{1 - P - 2\varepsilon}}. \]

Remark. Consider $m = \Theta(n)$ and $c + c' = O(\sqrt{m})$, then
\[ |S_a| \leq (\ln n + O(1)) \cdot |S^*| + O(\sqrt{n}). \]
## Experiments (Datasets)

<table>
<thead>
<tr>
<th>Graph</th>
<th># of nodes</th>
<th># of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>wiki-Vote</td>
<td>7,115</td>
<td>103,689</td>
</tr>
<tr>
<td>NetHEPT</td>
<td>15,233</td>
<td>58,891</td>
</tr>
<tr>
<td>Flixster 1</td>
<td>28,317</td>
<td>206,012</td>
</tr>
<tr>
<td>Flixster 2</td>
<td>25,474</td>
<td>135,618</td>
</tr>
</tbody>
</table>
Experiments (Concentration)

Standard deviation of influence distributions \( c + c' = O(\sqrt{n}) \).

Flixster graph 1, 28317 nodes
standard deviation \( \leq 760 \).

Flixster graph 2, 25474 nodes
standard deviation \( \leq 270 \).
Experiments (Performance)

Fix $P = 0.1$, change $\eta$.

Flixster graph 1
MinSeed-PCG selects seeds
94.4% less than Random,
54.0% less than High-degree,
29.2% less than PageRank.

Flixster graph 2
MinSeed-PCG selects seeds
91.1% less than Random,
73.0% less than High-degree,
24.4% less than PageRank.
Conclusion

Our work

▶ We are the first to propose the problem Seed Minimization with Probabilistic Coverage Guarantee (SM-PCG).
▶ We show that neither of the two set functions corresponding to the objective is submodular.
▶ We approximate SM-PCG with theoretical analysis.
Conclusion

Our work

▶ We are the first to propose the problem Seed Minimization with Probabilistic Coverage Guarantee (SM-PCG).
▶ We show that neither of the two set functions corresponding to the objective is submodular.
▶ We approximate SM-PCG with theoretical analysis.

Future work

▶ Nonsubmodular influence maximization.
▶ Concentration property of graphs.
▶ Influence maximization problem where becoming a hot topic is the first step, which is followed by further diffusion steps.
Thank You