Influence Maximization: The New Frontier --- Non-Submodular Optimizations

Weich Chen
Motivating Example: Viral Marketing in Social Networks

- Increasing popularity of online social networks may enable large scale viral marketing
Influence Maximization Problem

• Given a social network and an influence diffusion model
  – Find the seed set of certain size
  – Provide the largest influence spread

• Application
  – Viral marketing [Kempe et al. 2003, etc.]
  – Cascade detection [Leskovec et al., 2007]
  – Rumor control [Budak et al. 2011, He et al. 2012]
  – Text summarization [Wang et al. 2013]
  – Gang violence reduction [Shakarian et al. 2014]
Summary of My Past Work

• Scalable influence maximization
  – Fast heuristics algorithms with thousand times speedup
    • DegreeDiscount: No.2 most cited paper in KDD’09 (462 times)
    • PMIA: No.1 most cited paper in KDD’10 (340 times)
    • LDAG: No.2 most cited paper in ICDM’10 (169 times)

• Competitive diffusion modeling and optimization
  [SDM’11 ’12, WSDM’13]

• Alternative objectives: time-critical influence maximization [AAAI’12]; optimal influence route selection [KDD’13], etc.

• Monograph on influence diffusion, 2013
Common Theme

• Based on submodularity property
  – Diminishing marginal return
  – $f : 2^V \rightarrow R$; for all $S \subseteq T \subseteq V$, all $v \in V \setminus T$,
    $f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$

• Submodularity allow greedy solution
  – expected influence coverage is submodular
  – Select node with largest marginal influence one by one
  – Guarantee
    • $\left(1 - \frac{1}{e}\right)$ approximation for maximizing influence
    • $\ln n$ approximation for minimizing seed set size
Issue: Conformity (Group Psychology, Herd Mentality) in Influence Diffusion
Issue: Not All Diffusion Is Submodular

- Threshold behavior
  - tipping point: when diffusion reaches a critical mass, a drastic increase in further diffusion

New Frontier: Non-Submodular Influence Maximization
Seed Minimization with Probabilistic Coverage Guarantee

KDD’13, joint work with Peng Zhang, Purdue U.
Xiaoming Sun, Jialin Zhang, ICT of CAS
Yajun Wang, Microsoft
Motivation

• Our first attempt at non-submodular influence maximization

• Consider influencing mass media (e.g. sina.com)
  – Mass media pay attention only when a topic is discussed by a large portion of people (e.g. hot topic list on weibo.com)
    • Threshold behavior
  – Need probabilistic guarantee (e.g. 70%)
    • expected influence coverage is not informative enough
Independent Cascade Model

- Each edge \((u, v)\) has an influence probability \(p(u, v)\)
- Initially seed nodes in \(S_0\) are activated
- At each step \(t\), each node \(u\) activated at step \(t - 1\) activates its neighbor \(v\) independently with probability \(p(u, v)\)
Problem Definition

• Seed Minimization with Probabilistic Coverage Guarantee (SM-PCG)

• Input: directed graph $G = (V, E)$, influence probabilities $p_e$’s on edges under IC model, the target set $U$, coverage threshold $\eta < |U|$, probability threshold $P \in (0, 1)$.

• Output: $S^* = \text{argmin}_S \text{Pr}(\text{Inf}(S) \geq \eta \geq P |S|)$.

  - $\text{Inf}(S)$: random variable, number of nodes activated by seed set $S$
Non-Submodularity of Objective Functions

- Fix $\eta$, $f_\eta(S) = \Pr(\text{Inf}(S) \geq \eta)$,
  - $S^* = \text{argmin}_{S : f_\eta(S) \geq P} |S|$
  - not submodular

Edge probabilities are 1.
Fix $\eta = 5$,
$f_\eta(S \cup \{c\}) - f_\eta(S) = 0$,
$f_\eta(T \cup \{c\}) - f_\eta(T) = 1$.

- Fix $P$, $g_P(S) = \max_{\eta'}:Pr(\text{Inf}(S) \geq \eta') \geq P \eta'$,
  - $S^* = \text{argmin}_{S : g_P(S) \geq \eta} |S|$
  - not submodular

Edge probabilities are 0.5.
Fix $P = 0.8$,
$g_P(S \cup \{c\}) - g_P(S) = 0$,
$g_P(T \cup \{c\}) - g_P(T) = 1$. 
Influence Coverage Computation

- $P = f_\eta(S)$: #P-hard, but approximable by Monte Carlo simulation
  - Simulate diffusion from $S$ for $R$ times, use
    - $\hat{P} = \text{fraction of cascades with coverage at least } \eta$
  - To achieve $|\hat{P} - P| \leq \varepsilon$ with probability $1 - \frac{1}{n^\delta}$, set $R \geq \frac{\ln(2n^\delta)}{2\varepsilon^2}$.

- $\eta = g_P(S)$: #P-hard to approximate within any nontrivial multiplicative ratio
Idea for Solving SM-PCG

• Connect SM-PCG problem with another problem, Seed Minimization with Expected Coverage Guarantee (SM-ECG), which has submodular objective function
  – Output: \( S^* = \arg\min_S: E[\text{Inf}(S)] \geq \eta |S| \).
  – \( E[\text{Inf}(S)] \) is submodular \( \Rightarrow \ln n \) greedy approximation algorithm

• Need additional seeds for probabilistic guarantee, resulting in an additive term in approximation guarantee
  – related to the concentration of the influence coverage distribution
  – Our contribution: build such connection and detailed analysis
Approximation Algorithm

- Main idea: connect SM-PCG with SM-ECG

\[ \text{MinSeed-PCG}(\varepsilon): \quad \varepsilon \in \left[ 0, \frac{1-P}{2} \right] \text{ is a control parameter} \]

\[ S_0 = \emptyset \]

\textbf{For} \( i = 1 \) to \( n \) \textbf{do}

\[ u = \arg\max_{v \in V \setminus S_{i-1}} E[\Inf(S_{i-1} \cup \{v\})] - E[\Inf(S_{i-1})] \]

\[ S_i = S_{i-1} \cup \{u\} \]

\textit{prob} = Monte Carlo estimate of \( \Pr(\Inf(S_i) \geq \eta) \)

\textbf{if} \textit{prob} \( \geq P + \varepsilon \)

\textbf{return} \( S_i \)

\textbf{end if}

\textbf{End for}
Approximation Algorithm

• Let $n = |V|, m = |U|

• Theorem: Let $S_a$ be the output of $\text{MinSeed-PCG}(\varepsilon)$, $c = \max\{\eta - E[\text{Inf}(S^*)], 0\}, \ c' = \max\{E[\text{Inf}(S_{a-1})] - \eta, 0\}$. Then,

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

• Theorem: When using Monte Carlo estimate of $\Pr(\text{Inf}(S_{i}) \geq \eta)$ with at least $\ln \left( \frac{2n^2}{(2\varepsilon)^2} \right)$ iterations, with probability at least $1 - 1/n$, $\Pr(\text{Inf}(S_a) \geq \eta) \geq P$, and

$$c \leq \sqrt{\frac{\text{Var}(\text{Inf}(S^*))}{p}}, \ c' \leq \sqrt{\frac{\text{Var}(\text{Inf}(S_{a-1}))}{1-P-2\varepsilon}}.$$ 

• Assume $m = \Theta(n), \ c + c' = O(\sqrt{m})$, then

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

- Result on submodular function approximation:
Let $f$ be a real-valued nonnegative, monotone, submodular set function on $V$, $0 < \eta < f(V)$. Let $S^* = \arg\min_{S: f(S) \geq \eta} |S|$, $S$ be the greedy solution satisfying $f(S) \geq \eta$. Then,

$$|S| \leq \alpha |S^*| + 1, \quad \alpha = \max \left\{ \left\lceil \ln \frac{\eta |V|}{f(V) - \eta} \right\rceil, 0 \right\}.$$
Analysis II

- $\sigma(S) = E[\text{Inf}(S)]$
- Greedy seed sets: $S_1, S_2, \ldots S_i, \ldots, S_j, \ldots, S_n$

\[
\min i \text{ s.t. } \sigma(S_i) \geq \eta - c,
\text{Let } S_i^* = \arg\min_S \sigma(S) \geq \eta - c.
\Rightarrow |S_i| \leq \left[ \ln \left( \frac{m}{m-(\eta-c)} \right) \right] |S_i^*| + 1 \leq \left[ \ln \frac{m}{m-\eta} \right] |S^*| + 1.
\]

\[
\min j \text{ s.t. } \sigma(S_j) \geq \eta + c', \text{ thus } |S_a| \leq |S_j| + 1.
\]

By submodularity and greedy seed selection:

- $\forall i < t \leq k, \sigma(S_t) - \sigma(S_{t-1}) \geq \sigma(S_k) - \sigma(S_{k-1}),$
- $\Rightarrow \forall i < t < j, \sigma(S_t) - \sigma(S_{t-1}) \geq \frac{m-\sigma(S_{t-1})}{n} > \frac{m-(\eta+c')}{n},$
- $\Rightarrow |S_{j-1} \setminus S_i| \leq \frac{\sigma(S_{j-1}) - \sigma(S_i)}{\min_{i < t < j} \{ \sigma(S_t) - \sigma(S_{t-1}) \}} \leq \frac{(c+c')n}{m-(\eta+c')}.$

pdf of $\text{Inf}(S^*)$

pdf of $\text{Inf}(S_{a-1})$
Analysis III

\[ c \leq \sqrt{\frac{\text{Var}(\text{Inf}(S^*))}{P}} \]

\[ P \leq \Pr(\text{Inf}(S^*) \geq \eta) \]
\[ = \Pr(\text{Inf}(S^*) - E[\text{Inf}(S^*)] \geq \eta - E[\text{Inf}(S^*)]) \]
\[ \leq \Pr(|\text{Inf}(S^*) - E[\text{Inf}(S^*)]| \geq \eta - E[\text{Inf}(S^*)]) \]
\[ \leq \frac{\text{Var}(\text{Inf}(S^*))}{(\eta - E[\text{Inf}(S^*)])^2} \{ \text{Chebyshev's inequality} \} \]
\[ = \frac{\text{Var}(\text{Inf}(S^*))}{c^2}. \]

\[ c' \leq \sqrt{\frac{\text{Var}(\text{Inf}(S_{a-1}))}{1 - P - 2\varepsilon}} \text{ with high prob.} \]
Results on Bipartite Graphs

• $G = (V_1, V_2, E)$ is a one-way bipartite graph.

• Observation: activation of nodes in $U$ is mutually independent.
Results on Bipartite Graphs

• Pr(Inf(S) ≥ \eta) can be computed exactly by dynamic programming.

• \(A(S, i, j)\): probability that \(S\) activates \(j\) nodes of the first \(i\) nodes.

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

\[
A(S, i, j) = \begin{cases} 
A(S, i - 1, 0) \cdot (1 - p(S, v_i)), & j = 0 \\
A(S, i - 1, j - 1) \cdot p(S, v_i) + A(S, i - 1, j) \cdot (1 - p(S, v_i)), & 1 \leq j < i \\
A(S, i - 1, j - 1) \cdot p(S, v_i), & j = i 
\end{cases}
\]
Results on Bipartite Graphs

• Theorem:

\[ c \leq \sqrt{\frac{m}{2} \ln \frac{1}{P}}, \quad c' \leq \sqrt{\frac{m}{2} \ln \frac{2}{1 - P}}. \]

• Corollary:

\[ |S| \leq \left( \ln n + O(1) \right) |S^*| + O \left( \frac{n}{\sqrt{m}} \right). \]
## Experiment Datasets

<table>
<thead>
<tr>
<th>graph</th>
<th># of nodes</th>
<th># of edges</th>
<th>edge probabilities</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki-Vote</td>
<td>7,115</td>
<td>103,689</td>
<td>synthetic, weighted cascade</td>
<td>voting network in Wikipedia</td>
</tr>
<tr>
<td>NetHEPT</td>
<td>15,233</td>
<td>58,891</td>
<td>synthetic, weighted cascade</td>
<td>collaboration network in arxiv.org</td>
</tr>
<tr>
<td>Flixster 1</td>
<td>28,327</td>
<td>206,012</td>
<td>learned from action trace</td>
<td>rating network in movie rating site Flixster for topic 1</td>
</tr>
<tr>
<td>Flixster 2</td>
<td>25,474</td>
<td>135,618</td>
<td>learned from action trace</td>
<td>rating network in movie rating site Flixster for topic 2</td>
</tr>
</tbody>
</table>
Experiment (Concentration)

- Standard deviation of influence distribution \((c + c' = O(\sqrt{m}))\)

![Graphs showing standard deviation of influence coverage against number of seeds for Wiki-vote and NetHEPT networks.]

Wiki-vote, 7115 nodes, Standard deviation \(\leq 130\).

NetHEPT, 15233 nodes, Standard deviation \(\leq 105\).
Experiment (Concentration)

- Standard deviation of influence distribution \( (c + c' = O(\sqrt{m})) \)

Flixster with topic 1, 28317 nodes,
Standard deviation \( \leq 760. \)

Flixster with topic 2, 25474 nodes,
Standard deviation \( \leq 270. \)
Experiment (Performance)

- **MinSeed-PCG(ε):** generate seed set sequence by PMIA ([Chen et al, KDD 2010]), set $\epsilon = 0.01$.
- **Random:** generate seed set sequence randomly.
- **High-degree:** generate seed set sequence according to the decreasing order of out-degree of nodes.
- **PageRank:** generate seed set sequence according to the importance measured by PageRank.
Experiment (Performance)

- Performance of our algorithm ($P = 0.1$)

Wiki-vote,
88.2% less than Random,
20.2% less than High-degree,
30.9% less than PageRank.

NetHEPT,
56.7% less than Random,
46.0% less than High-degree,
24.4% less than PageRank.
Experiment (Performance)

- Performance of our algorithm ($P = 0.1$)

Flixster with topic 1,
94.4% less than Random,
54.0% less than High-degree,
29.2% less than PageRank.

Flixster with topic 2,
91.2% less than Random,
73.0% less than High-degree,
24.4% less than PageRank.
Experiment (Performance)

- Performance of our algorithm ($P = 0.5$)
Experiment (Performance)

- Performance of our algorithm (fixed $\eta$)

<table>
<thead>
<tr>
<th>Wiki-vote</th>
<th>$\eta = 3000$</th>
<th>$\eta = 4500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>86.4%</td>
<td>76.3%</td>
</tr>
<tr>
<td>High-degree</td>
<td>27.7%</td>
<td>30.8%</td>
</tr>
<tr>
<td>PageRank</td>
<td>34.1%</td>
<td>38.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NetHEPT</th>
<th>$\eta = 6000$</th>
<th>$\eta = 10500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>59.2%</td>
<td>49.6%</td>
</tr>
<tr>
<td>High-degree</td>
<td>51.8%</td>
<td>52.9%</td>
</tr>
<tr>
<td>PageRank</td>
<td>22.8%</td>
<td>36.1%</td>
</tr>
</tbody>
</table>
Experiment (Performance)

- Performance of our algorithm (fixed $\eta$)

<table>
<thead>
<tr>
<th>Flixster 1</th>
<th>$\eta = 2000$</th>
<th>$\eta = 4000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>98.3%</td>
<td>93.9%</td>
</tr>
<tr>
<td>High-degree</td>
<td>78.9%</td>
<td>70.0%</td>
</tr>
<tr>
<td>PageRank</td>
<td>44.1%</td>
<td>53.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flixster 2</th>
<th>$\eta = 2000$</th>
<th>$\eta = 4000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>95.8%</td>
<td>89.0%</td>
</tr>
<tr>
<td>High-degree</td>
<td>78.6%</td>
<td>76.2%</td>
</tr>
<tr>
<td>PageRank</td>
<td>59.0%</td>
<td>54.9%</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

- First to propose the problem emphasizing **probabilistic coverage guarantee**
  - Objective functions are not submodular
- Approximate SM-PCG with theoretical analysis
- Future work
  - Other nonsubmodular influence maximization tasks
    - Generating a hot topic as the first step, with further diffusion steps
  - Study concentration properties of influence coverage on graphs
Thank you!