Robust Influence Maximization

Wei Chen — Microsoft Research

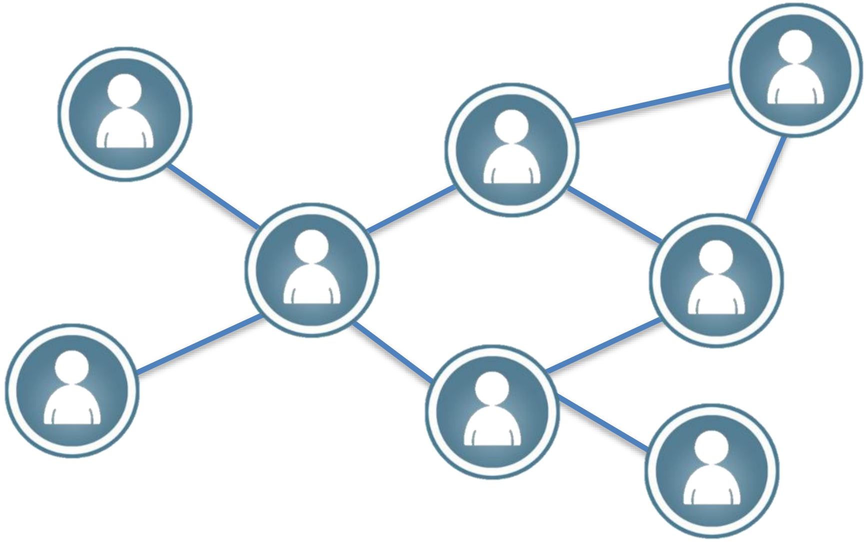
Joint work with
Tian Lin, Zihan Tan, Mingfei Zhao — Tsinghua University
Xuren Zhou — HKUST

Influence Maximization in Social Networks

- Influence maximization: selecting important nodes in a network to maximize the influence coverage [Kempe et al.'03]
- Applications of Influence Maximization
  - Viral marketing
  - Outbreak detection
  - Rumor monitoring and control
- In this work, we consider the robustness in Influence Maximization.

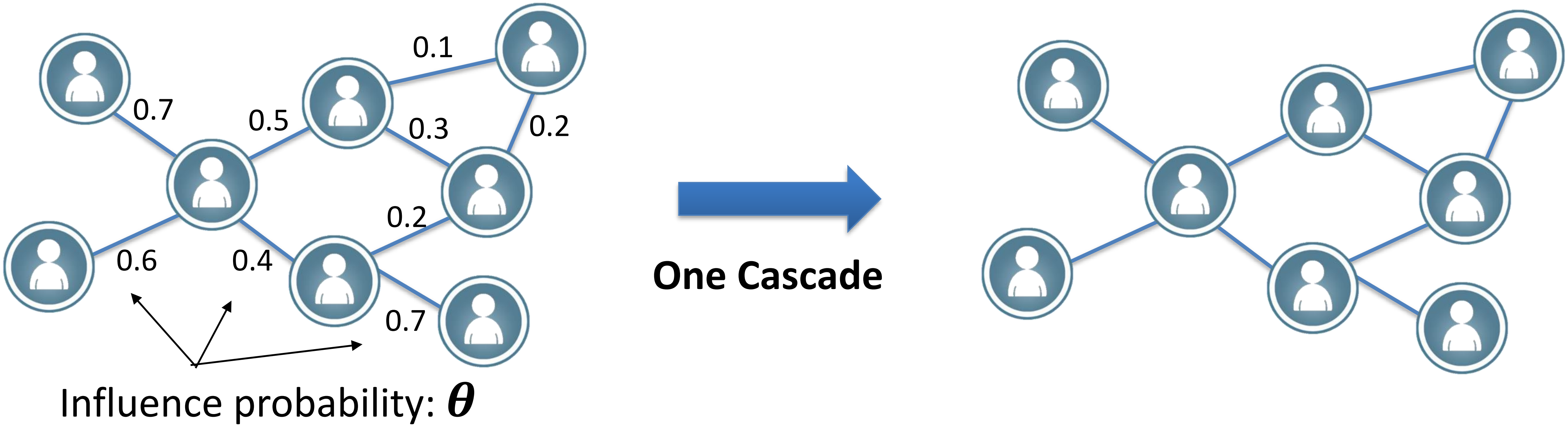
Motivating Example

- A company is to carry out the promotion campaign for their product, by sending free samples to initial users.

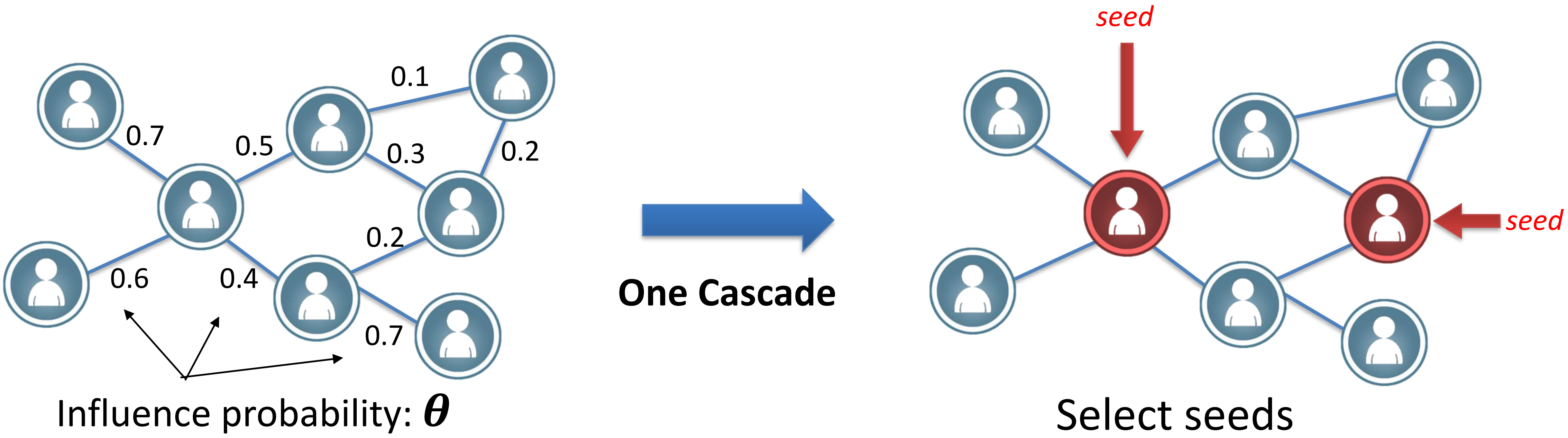


Independent Cascade (IC) Model

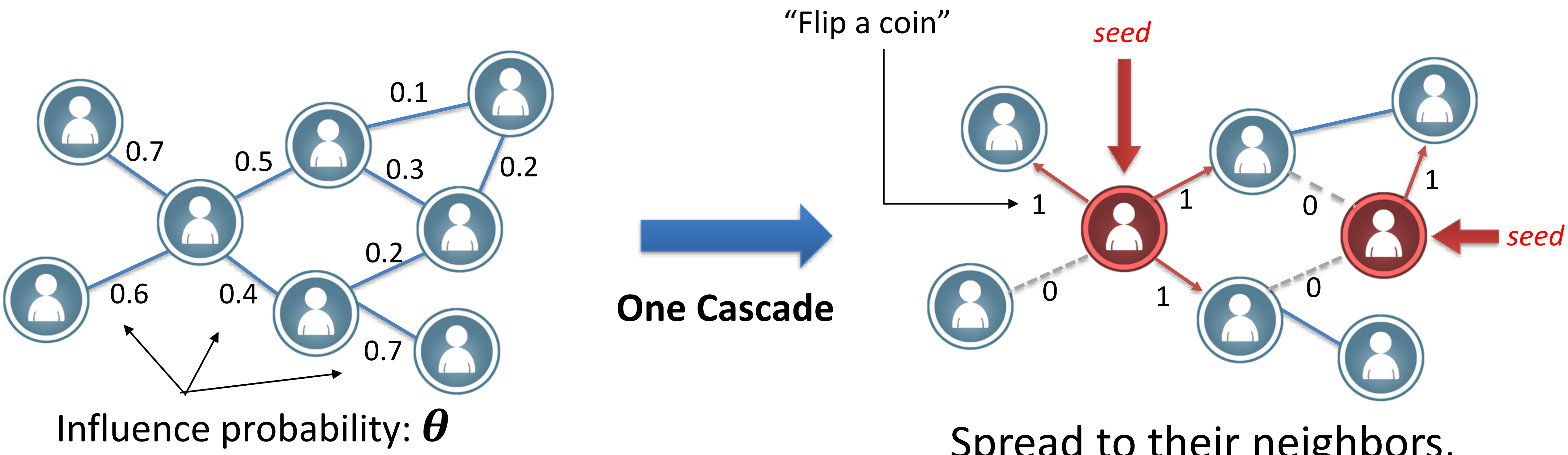
- Nodes are users, and edges are their relation.



Independent Cascade (IC) Model



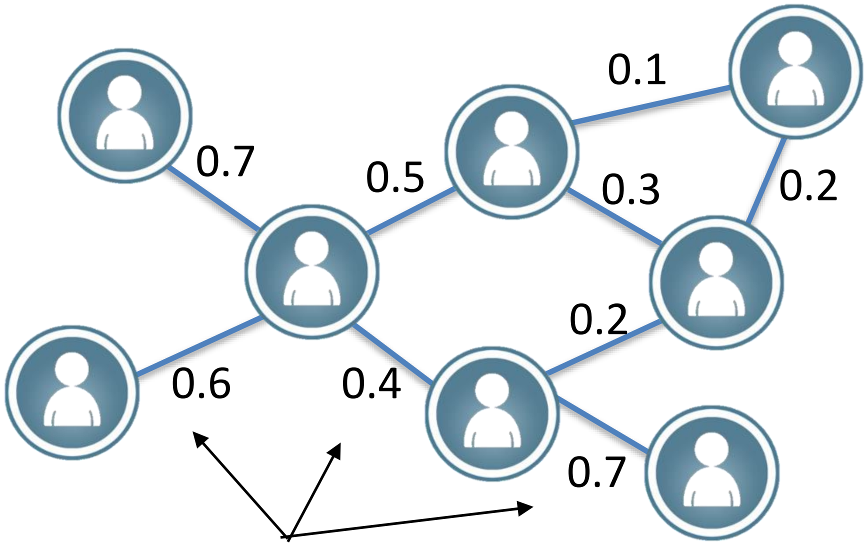
Independent Cascade (IC) Model



Spread to their neighbors, at time step 1

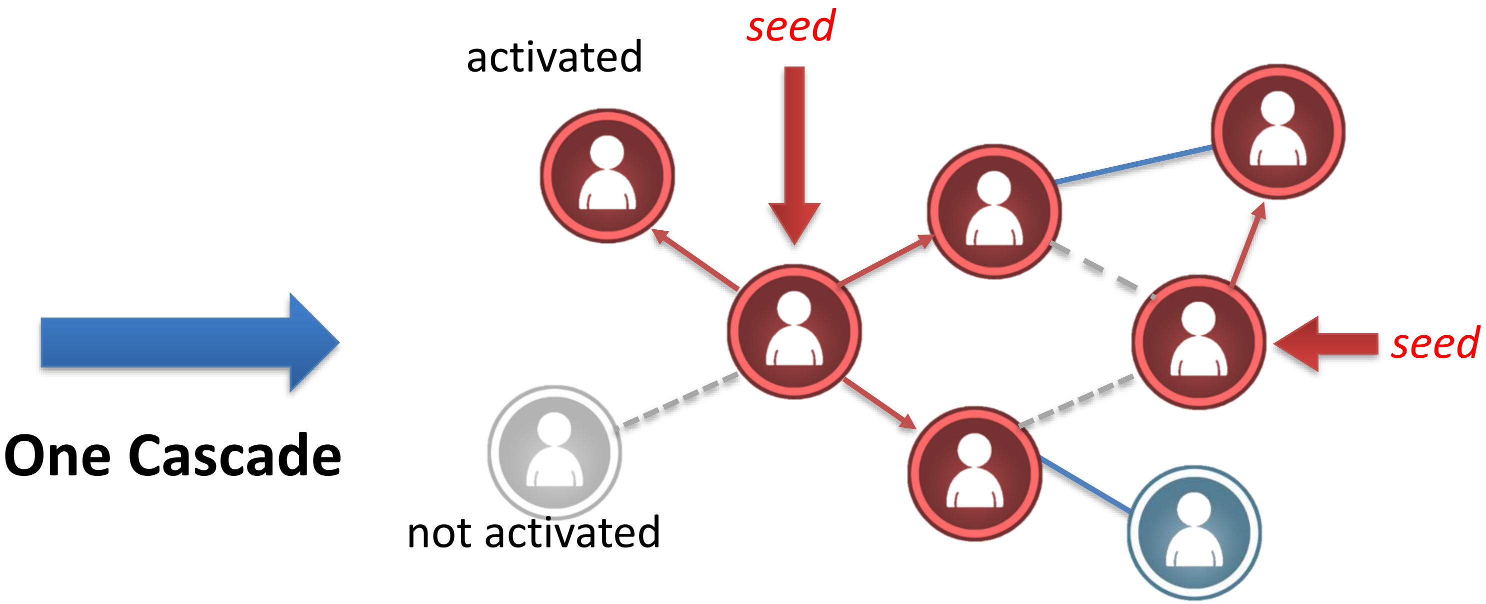
(Word-of-mouth effect)

Independent Cascade (IC) Model



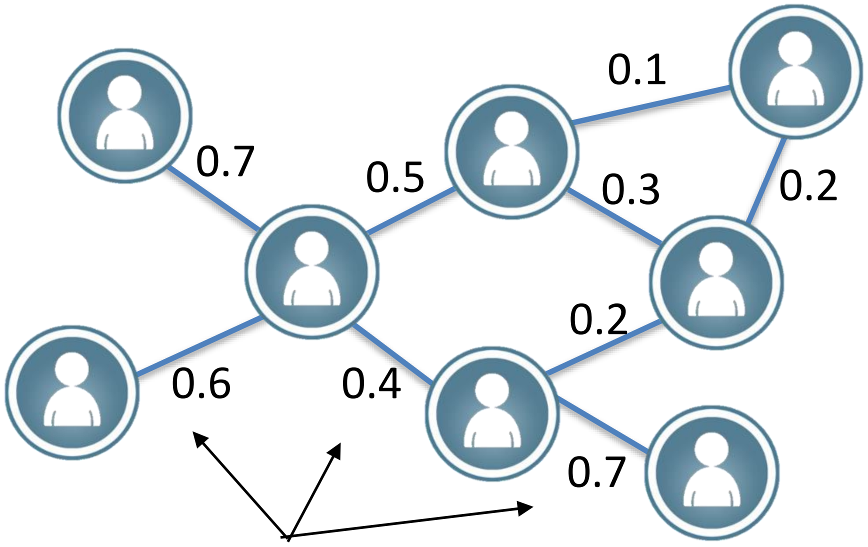
Influence probability: $\boldsymbol{\theta}$

One Cascade



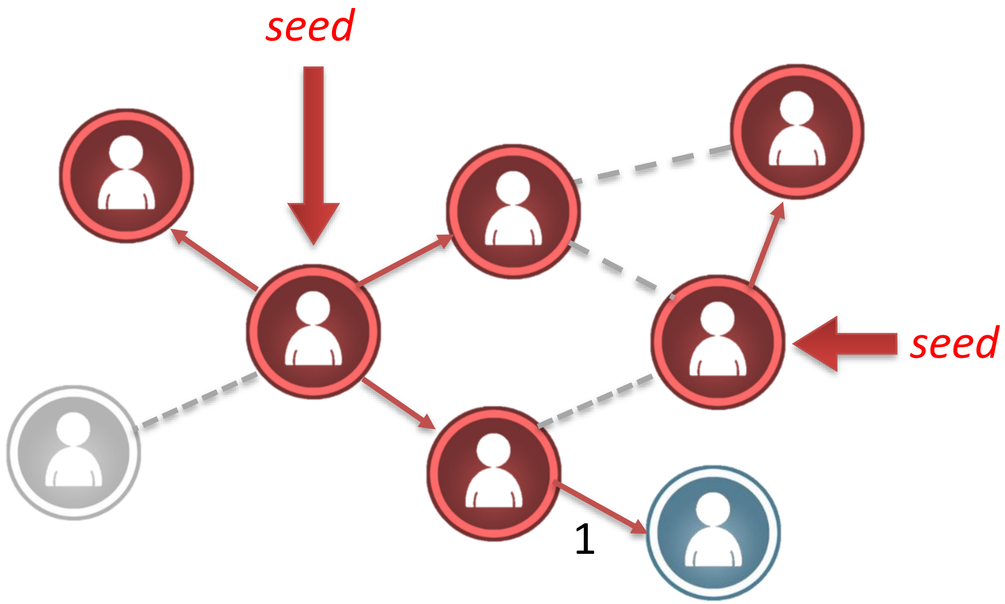
Spread to their neighbors,
at time step 1

Independent Cascade (IC) Model



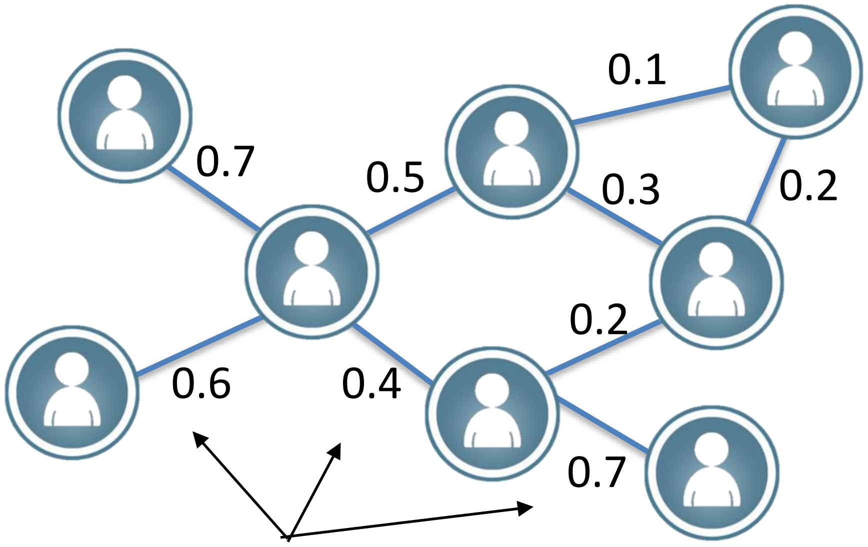
Influence probability: $\boldsymbol{\theta}$

One Cascade

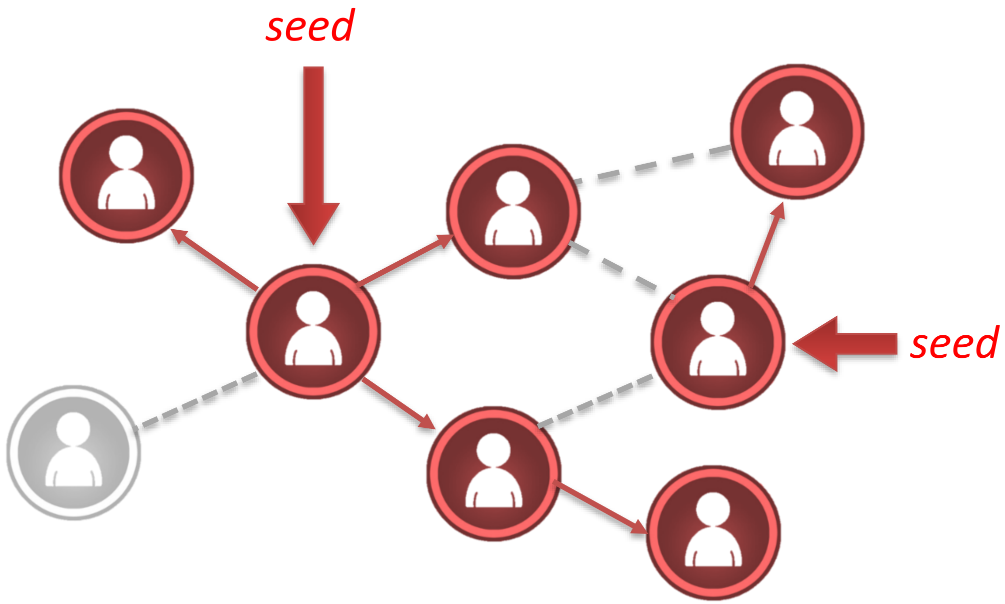


Continue to neighbors' neighbors, at time step 2

Independent Cascade (IC) Model

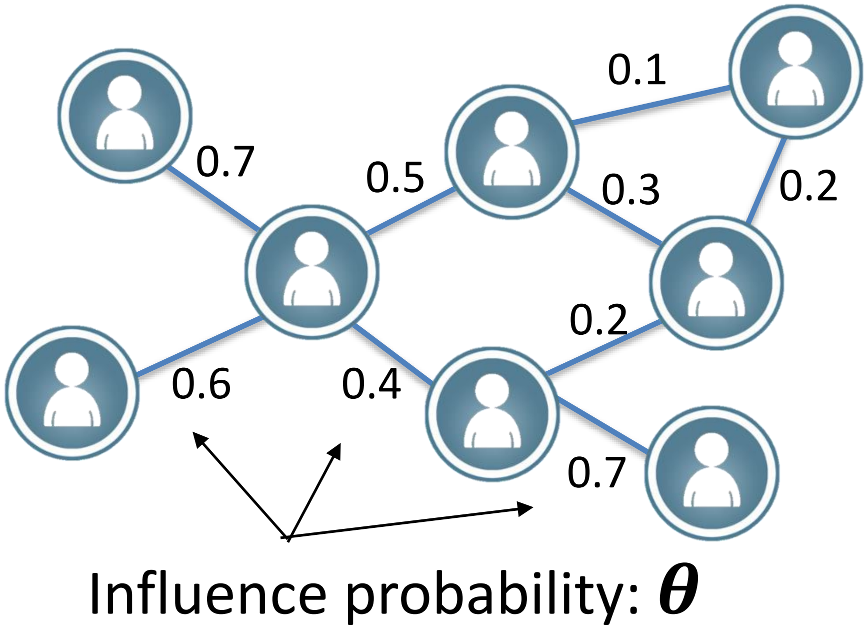


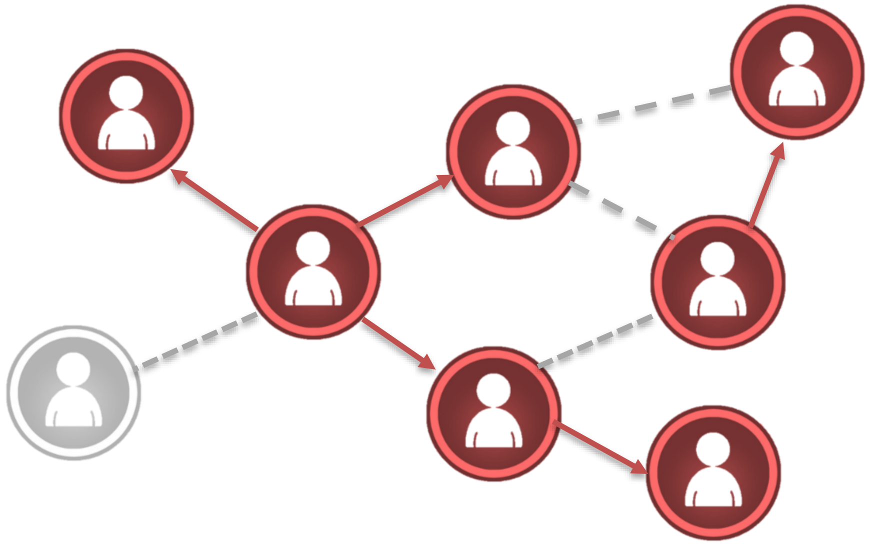
Influence probability: $\boldsymbol{\theta}$



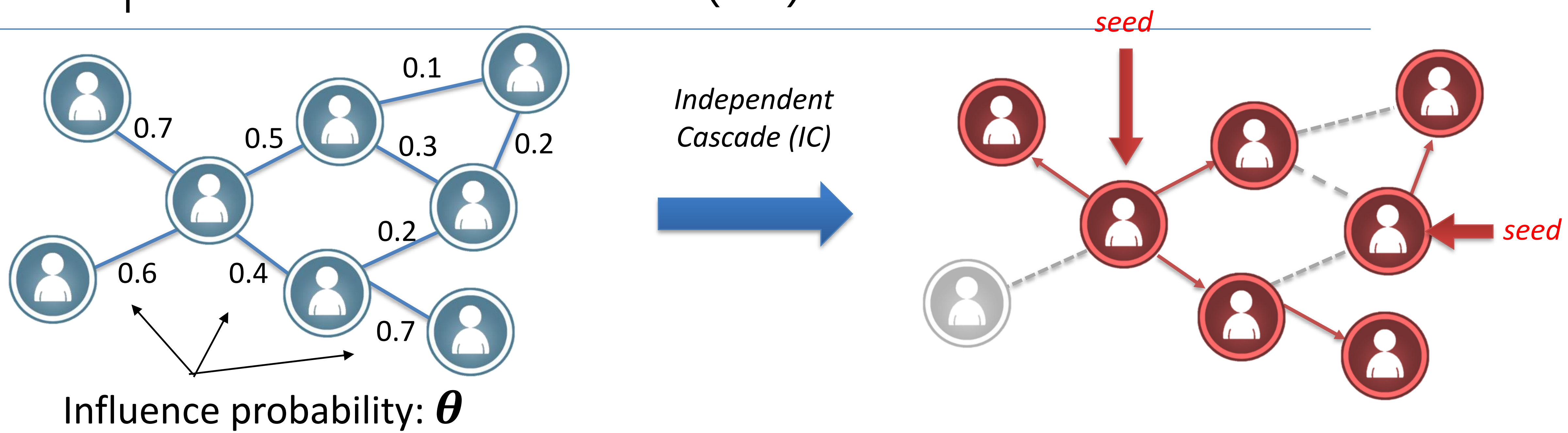
So on and so forth, until no more node is activated

Independent Cascade (IC) Model





Independent Cascade (IC) Model



Influence spread $\sigma_{\boldsymbol{\theta}}(S)$: expected number of activated nodes given seed set S and edge parameters θ

Influence Maximization Problem [Kempe et al.'03]: Given $G = (V, E, \boldsymbol{\theta})$, find k nodes $S \subseteq V$ as *seeds* to maximize $\sigma_{\boldsymbol{\theta}}(S)$.

Greedy Algorithm: achieves $1 - 1/\mathrm{e} - \epsilon$ approximation ratio

Robustness in Influence Maximization

Influence probability: $\boldsymbol{\theta}$

But where are these influence parameters from?

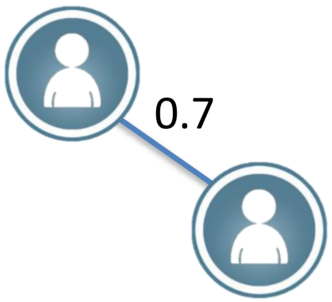
Learned from actual cascade data;
Ground-truth never known;
using confidence intervals is more realistic

Will influence maximization still work with these intervals?

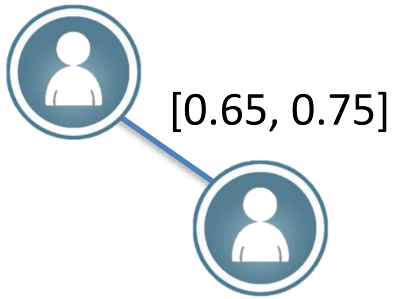
Model and Problem

- Probability of information diffusion is usually learned from data.
- Uncertainty caused by insufficient samples, noise, etc.
- What if the estimation error occurs?

(exact probability)



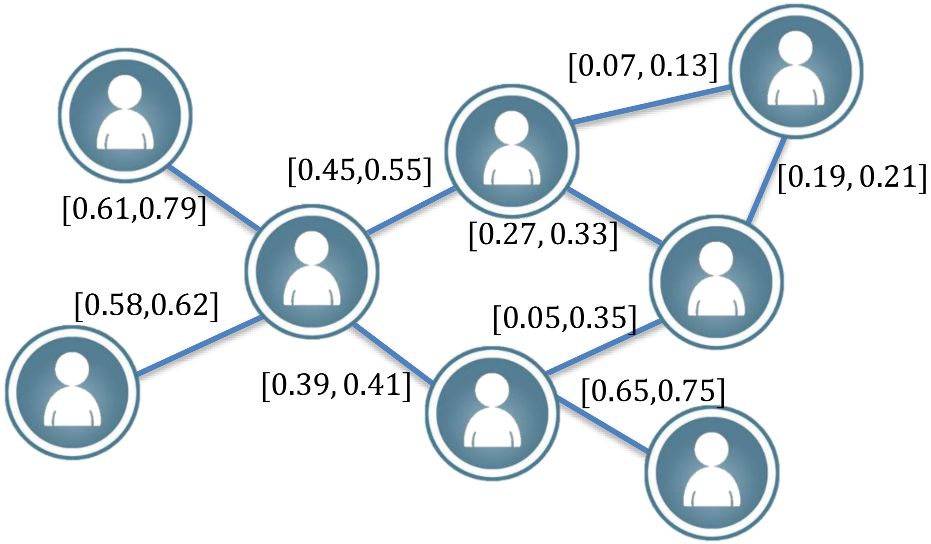
(with estimation error)



The true probability is somewhere in [0.65, 0.75] and is unknown.

No distribution assumption made for ground truth within the interval

Model and Problem



Assume that the probability on edges is given with their uncertainty.

Parameter space: $\mathbf{\Theta} := \times_{e \in E} [l_e, r_e]$

Given solution

Robust ratio: $g(\mathbf{\Theta}, S) := \min_{\boldsymbol{\theta} \in \mathbf{\Theta}} \frac{\sigma_{\boldsymbol{\theta}}(S)}{\sigma_{\boldsymbol{\theta}}(S^*_{\boldsymbol{\theta}})}$

The worst-case

Optimal solution

$\sigma_{\boldsymbol{\theta}}(S^*_{\boldsymbol{\theta}}) = \max_{|S|=k} \sigma_{\boldsymbol{\theta}}(S)$

Robust Influence Maximization (RIM)

- Given $G = (V, E, \boldsymbol{\Theta})$, find k nodes $S \subseteq V$ as *seeds* to maximize the robust ratio $g(\boldsymbol{\Theta}, S)$

$$S^*_{\boldsymbol{\Theta}} \coloneqq \underset{S \subseteq V, |S|=k}{\arg\max}\, g(\boldsymbol{\Theta}, S) = \underset{S \subseteq V, |S|=k}{\arg\max} \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\sigma_{\boldsymbol{\theta}}(S)}{\sigma_{\boldsymbol{\theta}}(S^*_{\boldsymbol{\theta}})}$$

Maximize the worst-case value

- **Follow the robust optimization approach in operation research**
- **Theorem 1.** RIM is **NP-hard**, and it is NP-hard to achieve RIM with robust ratio $1 - 1/\mathrm{e} + \epsilon$ for any $\epsilon > 0$.

$(\approx 63\%)$

LUGreedy for RIM with Solution-Dependent Bound

$$\boldsymbol{\theta}^- = (l_e)_{e\in E}, \boldsymbol{\theta}^+ = (r_e)_{e\in E}$$

Algorithm LUGreedy(G, k, Θ)

Input: Graph $G = (V, E)$, budget k, parameter space $\Theta = \times_{e\in E}[l_e, r_e]$

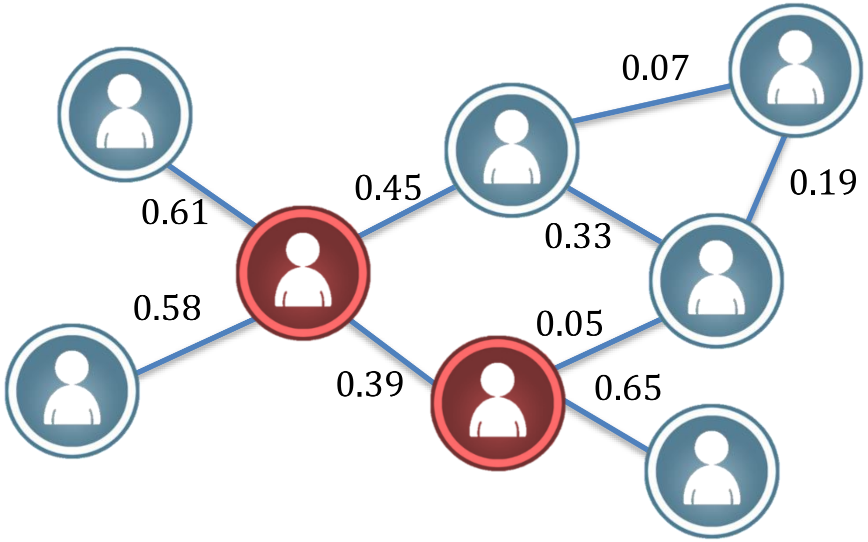
1: $S^g_{\theta^-} \leftarrow$ Greedy(G, k, θ^-)

2: $S^g_{\theta^+} \leftarrow$ Greedy(G, k, θ^+)

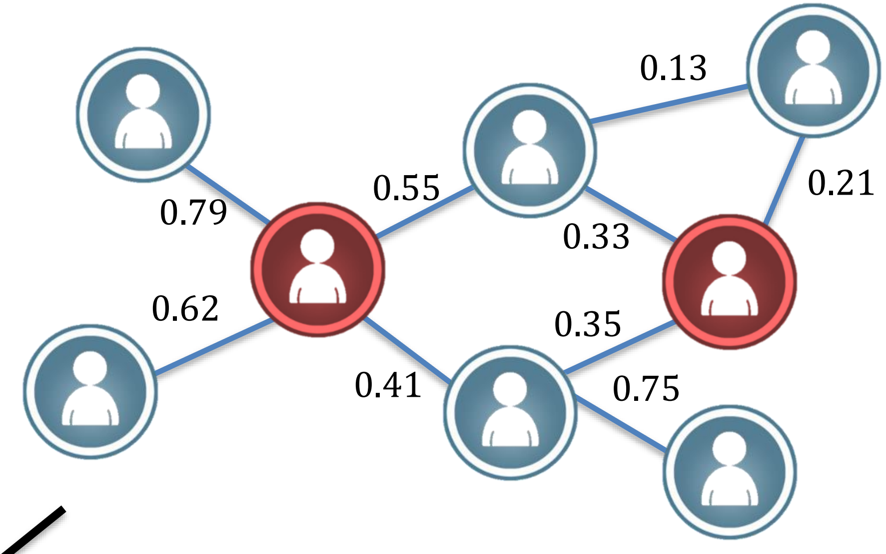
3: **return** $\arg\max_{S\in\{S^g_{\theta^-}, S^g_{\theta^+}\}} \{\sigma_{\theta^-}(S)\}$

Demonstration of LUGreedy

Lower Greedy $S^g_{\boldsymbol{\theta}^-}$



Upper Greedy $S^g_{\boldsymbol{\theta}^+}$



Select the maximum based on the lower probability

$$\sigma_{\boldsymbol{\theta}^-}(S^{LU}_{\boldsymbol{\Theta}}) = \max_{S\in\{S^g_{\boldsymbol{\theta}^-}, S^g_{\boldsymbol{\theta}^+}\}} \{\sigma_{\boldsymbol{\theta}^-}(S)\}$$

Solution-Dependent Bound

- Define *gap ratio*:

LUGreedy solution

$$\alpha(\boldsymbol{\Theta}) := \frac{\sigma_{\boldsymbol{\theta}^-}(S^{LU}_{\boldsymbol{\Theta}})}{\sigma_{\boldsymbol{\theta}^+}(S^{g}_{\boldsymbol{\theta}^+})} = \frac{\max\{\sigma_{\boldsymbol{\theta}^-}(S^{g}_{\boldsymbol{\theta}^-}), \sigma_{\boldsymbol{\theta}^-}(S^{g}_{\boldsymbol{\theta}^+})\}}{\sigma_{\boldsymbol{\theta}^+}(S^{g}_{\boldsymbol{\theta}^+})}.$$

- **Theorem 2.** LUGreedy outputs a seed set $\boldsymbol{S^{LU}_{\Theta}}$ such that:

$$g\left(\boldsymbol{\Theta}, S^{LU}_{\boldsymbol{\Theta}}\right) \geq \alpha(\boldsymbol{\Theta}) \cdot \left(1 - \frac{1}{\mathrm{e}}\right).$$

Example: When $\boldsymbol{\alpha(\Theta)}$ is large (e.g., $\geq \mathbf{0.7}$), then the result is reasonably good!

Worst-case Bound on Robust Ratio

- Unfortunately, a good input $\boldsymbol{\Theta} = \times_{e \in E} [l_e, r_e]$ is required (when the graph is bad)
  - Argument related to sharp threshold for the emergence of giant components in Erdös-Rényi Graphs

Theorem 3. For RIM, denote $\delta = \max_{e \in E} |r_e - l_e|$ as the maximum *interval width*.

- No constraint on δ. There exists a graph, such that $\max_{S \subseteq V, |S|=k} g(\boldsymbol{\Theta}, S) = O\left(\frac{k}{n}\right)$;
- Restrict $\delta = O\left(\frac{1}{n}\right)$. There exists a graph, such that $\max_{S \subseteq V, |S|=k} g(\boldsymbol{\Theta}, S) = O\left(\frac{\log n}{n}\right)$;
- Restrict $\delta = O\left(\frac{1}{\sqrt{n}}\right)$ and allow random seeds $\tilde{S}$. There exists a graph, such that

$$\max_{\Omega} \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\tilde{S} \sim \Omega} \left[\frac{\sigma_{\boldsymbol{\theta}}(\tilde{S})}{\sigma_{\boldsymbol{\theta}}(S^*_{\boldsymbol{\theta}})} \right] = O\left(\frac{\log n}{\sqrt{n}}\right).$$

- How to improve this?
  - Sampling to improve $\boldsymbol{\Theta}$
  - Study on the impact of graph structures?

Sampling for Improving RIM

- Intuition: sampling edges to shrink the confidence intervals in $\boldsymbol{\Theta}$
  - Law of large numbers

$$\text{Empirical mean } \hat{p}_t = \frac{1}{t}\sum_{i=1}^{t} X_i, \text{ true mean } \lim_{t\to\infty} \hat{p}_t = p.$$

  - "Tail probability diminishes fast"
- Sampling method
  - Uniform sampling: every edge has the same number of samples
  - Non-uniform / adaptive sampling

Theoretical Result on Uniform Sampling (US)

- Based on the additive and multiplicative relationship between influence spread error bound and sampling complexity:

Theorem 6. For any $\epsilon, \gamma > 0$, denote empirical vector $\boldsymbol{\theta} = (p_e)_{e \in E}$, $|V| = n$, and $|E| = m$. Then,

(1) Set $t = \frac{2m^2n^2 \ln(2m/\gamma)}{k^2\epsilon^2}$, and $\delta_e = \frac{k\epsilon}{nm}$;

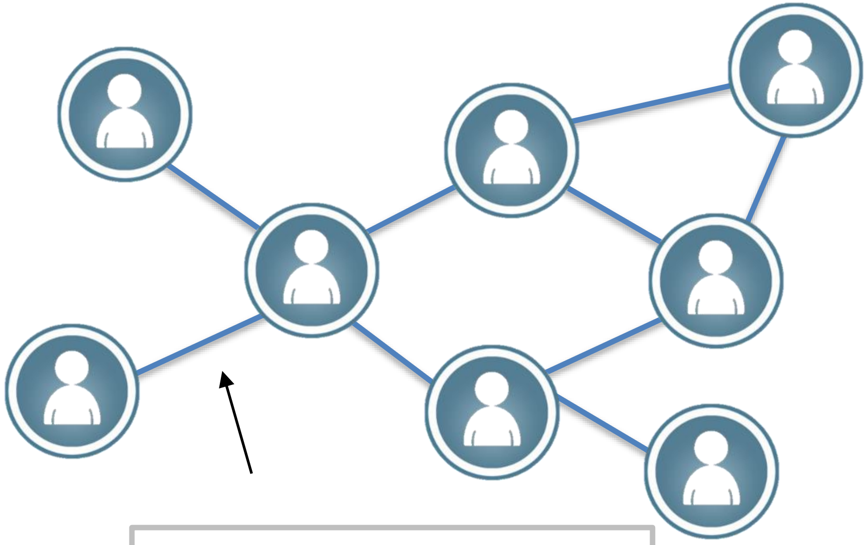
(2) Or, assume that the lower bound p': $0 < p' < \min_e p_e$. Set $t = \frac{3\ln(2m/\gamma)}{p'}\left(\frac{2n}{\ln(1/_{1-\epsilon})} + 1\right)$, and $\delta_e = \frac{1}{n}\hat{p}_e \ln(1/\gamma)$.

We have

$$g(\boldsymbol{\Theta}_{out}, S_{out}) \geq (1 - 1/e)(1 - \epsilon)$$

and

$$\Pr[\theta \in \boldsymbol{\Theta}_{out}] \geq 1 - \gamma.$$



$\boldsymbol{\Theta}_{out} = \times_{e \in E} \ [l_e, r_e]$

$l_e = \max\{0, \hat{p}_e - \delta_e\}$

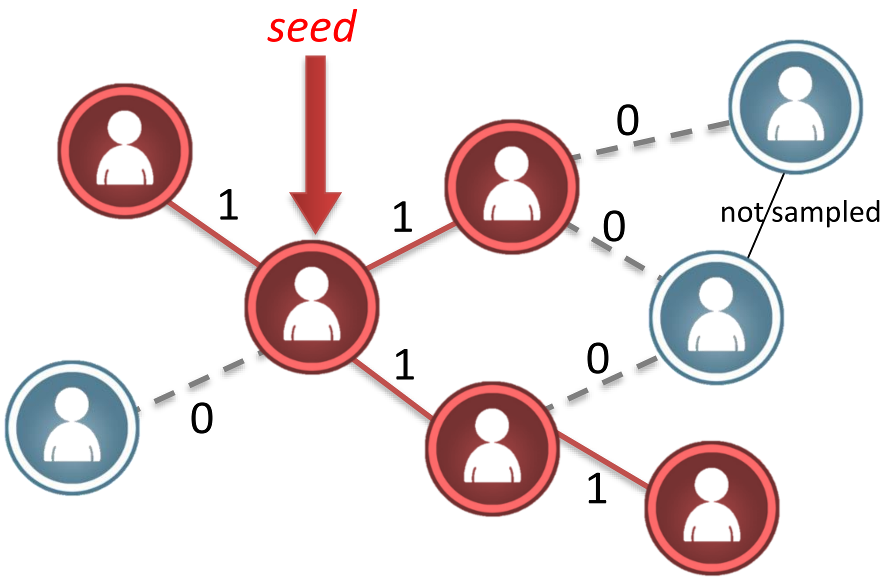
$r_e = \min\{1, \hat{p}_e + \delta_e\}$

Adaptive Sampling: Information Cascade Sampling (ICS)

- Idea: important edges should be sampled more; edges appear in cascades may be more important

Given threshold $\epsilon > 0$.
repeat:
- Call **LUGreedy** to get seeds S_i^g.
- Starting from seeds S_i^g, do *information cascade* and sample touched edges.
- $\mathbf{\Theta}_{i+1} \leftarrow \text{Update}(\mathbf{\Theta}_i)$; $i \leftarrow i + 1$;

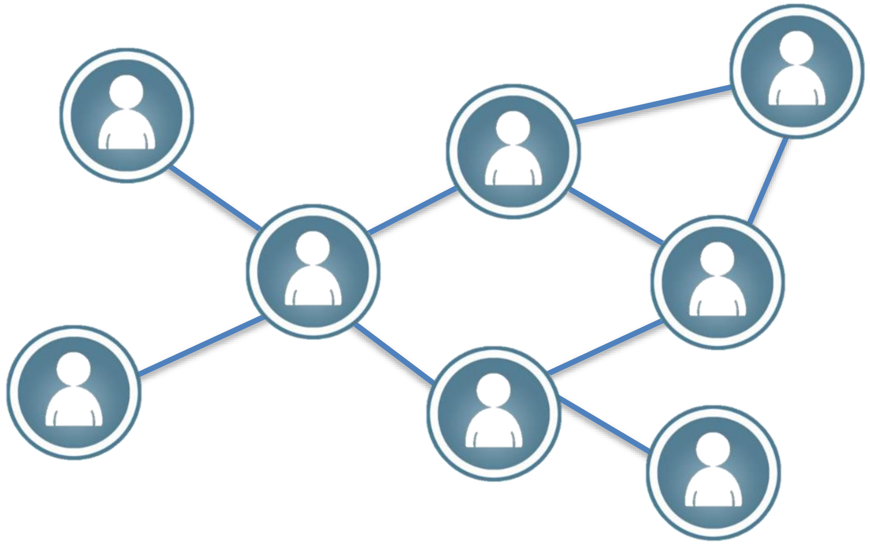
until $(\alpha(\mathbf{\Theta}_i) > 1 - \epsilon)$
return $\mathbf{\Theta}_{out} \leftarrow \mathbf{\Theta}_{i+1}, S_{out} \leftarrow S_i^g$



In practice, we samples τ times of information cascade, then change the seed set.

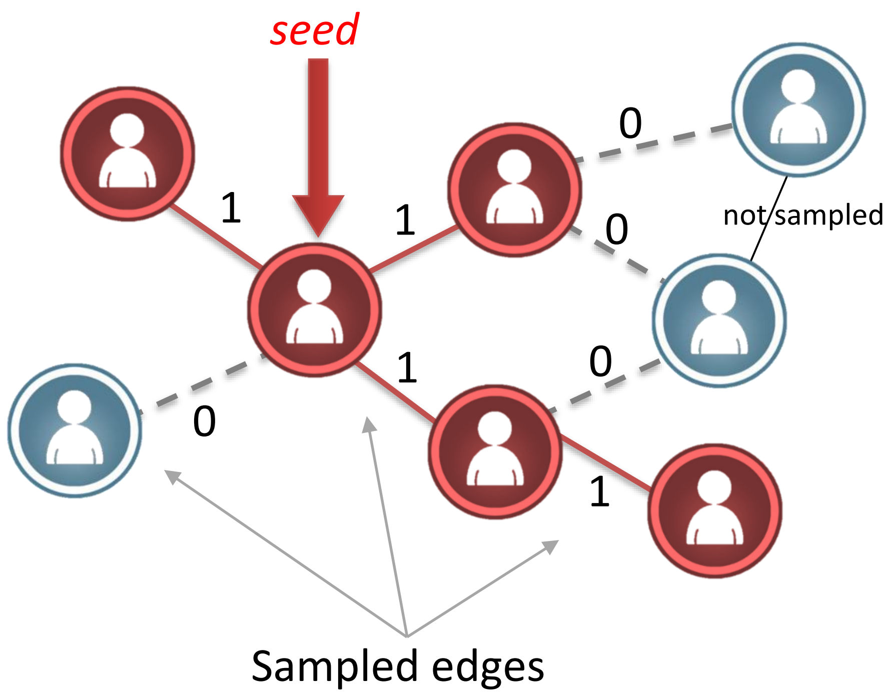
Adaptive Sampling: Influence Cascade Sampling (ICS)

- **The information cascade naturally samples edges along its trace.**



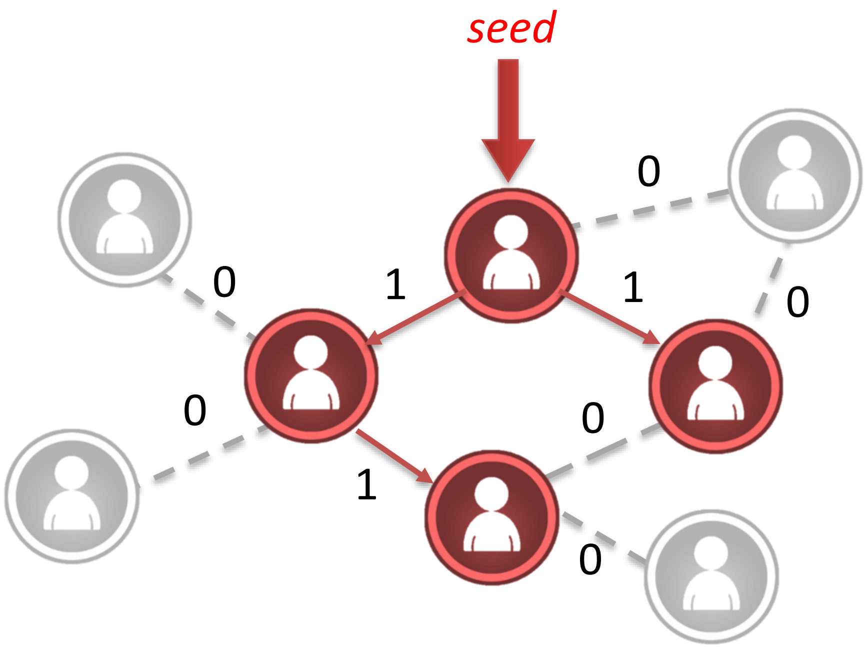
Adaptive Sampling: Influence Cascade Sampling (ICS)

Use **LUGreedy** to select seeds, and sample the trace of the information cascade.



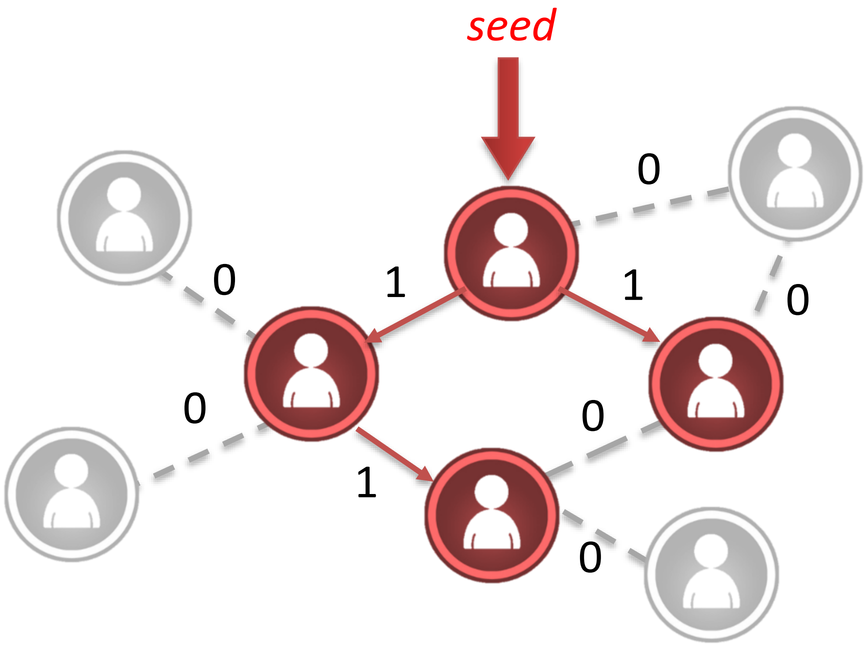
Adaptive Sampling: Influence Cascade Sampling (ICS)

Next time, we may sample different seeds and information cascade.



Adaptive Sampling: Influence Cascade Sampling (ICS)

From time to time, we can refine parameter space $\mathbf{\Theta}_{out}$.



Update($\mathbf{\Theta}$): $\times_{e \in E} \; [l_e, r_e]$

$$l_e = \max \left\{ 0, \hat{p}_e + \frac{c_e^2}{2} - c_e \sqrt{\frac{c_e^2}{4} + \hat{p}_e} \right\}$$

$$r_e = \min \left\{ 1, \hat{p}_e + \frac{c_e^2}{2} + c_e \sqrt{\frac{c_e^2}{4} + \hat{p}_e} \right\}$$

$\hat{p}_e$: empirical mean

t_e: estimated number

$$c_e = \sqrt{\frac{3}{t_e} \ln(2m/\gamma)}$$

Given any $\epsilon, \gamma > 0$, when the algorithm stops, it outputs S_{out} with robust ratio

$$g(\mathbf{\Theta}_{out}, S_{out}) \geq \left(1 - \frac{1}{\mathrm{e}}\right)(1 - \epsilon),$$

with probability $\Pr[\theta \in \mathbf{\Theta}_{out}] \geq 1 - \gamma$.

Empirical Evaluation

- Datasets

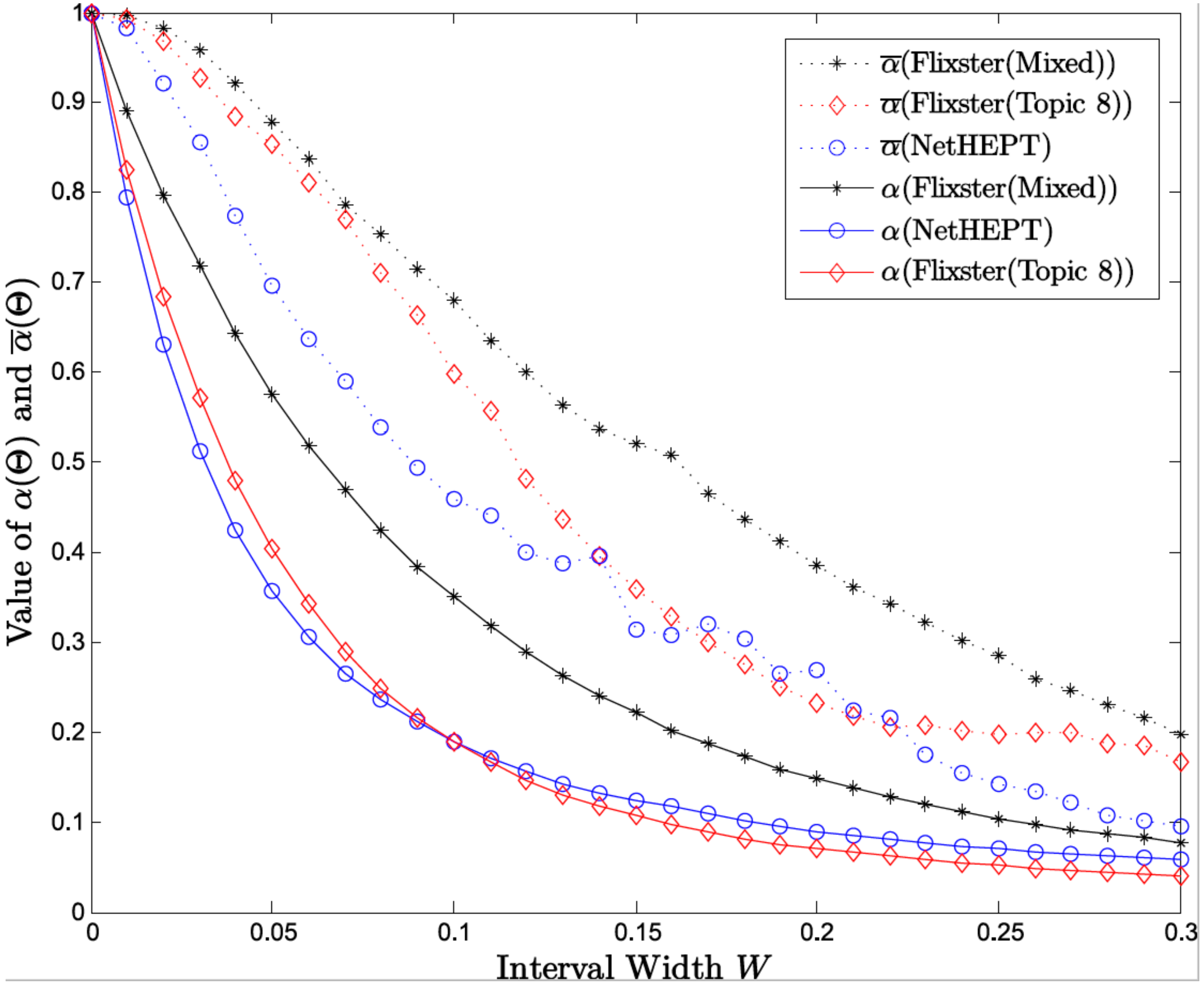
Name	Description	# of nodes	# of edges	Edge probability
NetHEPT	Academic collaboration network	15233	62774	Weighted cascade (synthetic)
Flixster (topic 8)	Movie rating induced network	14473	64934	Learned from trace
Flixster (Mixed, topics 1 & 4)	Movie rating induced network	7118	23252	Learned from trace, then evenly mixed between topics 1 & 4

Trend for Gap Ratio vs. Interval Width

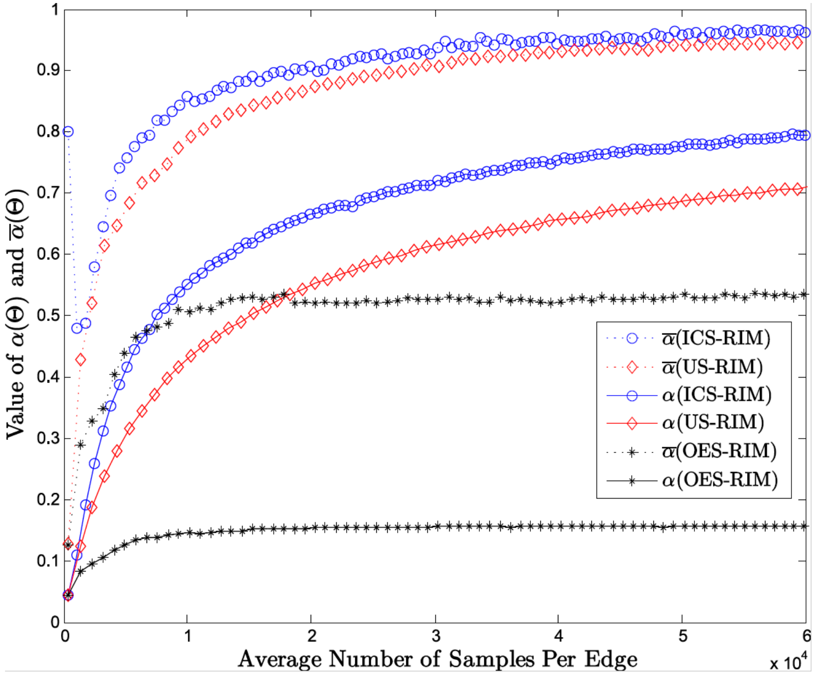
α: gap ratio (lower bound)

$\bar{\alpha}$: upper bound (estimated)

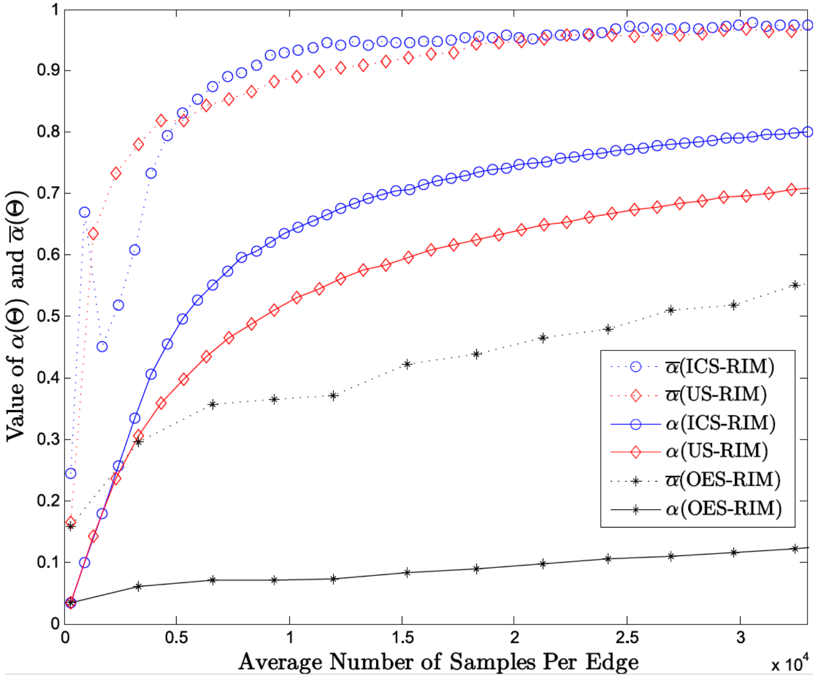
$k = 50$



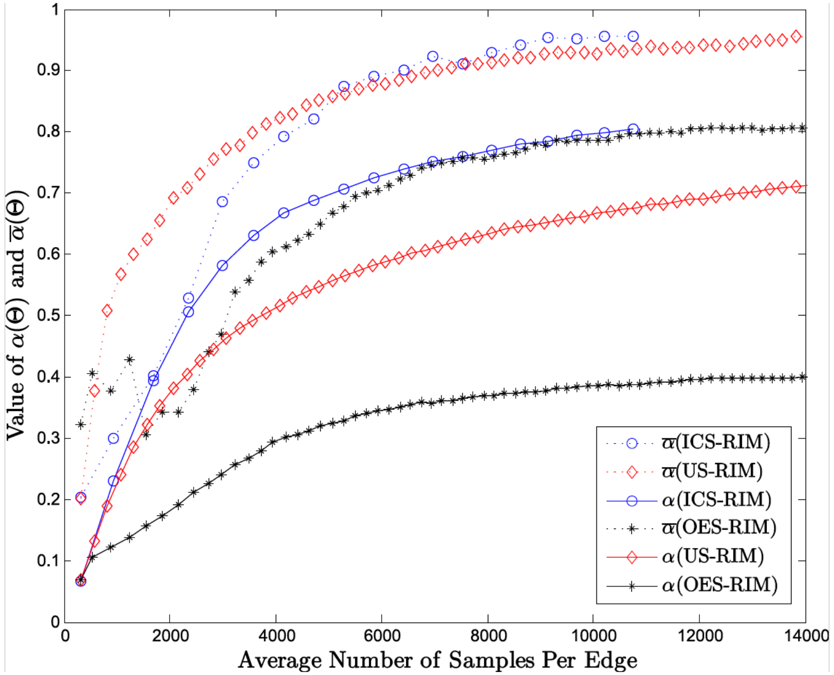
Comparison of Different Sampling Algorithms



NetHEPT



Flixster (Topic 8)



Flixster (Mixed)

Sampling Algorithm:

US: Uniform sampling ICS: Information cascade sampling OES: Outgoing edge only sampling

Related Works

- [Saito et al.'08] [Tang et al.'09] [Rodriguez et al.'11] etc., on methods to learn the probability on edges.
- [Chen et al. '09, '10] [Borgs et al. 14] [Tang et al. '14 '14] etc., on scalable influence maximization
- [He & Kempe'15] attempt to address the uncertainty by using a different model.
- [Krause et al.'08]: the hardness of general robust submodular optimization on a finite set of submodular functions; and bi-criteria solution
- [He & Kempe'16] (next talk): same objective function, but
  - Using the bi-criteria approach of [Krause et al'08]
  - For finite number of possible choices of diffusion models

Conclusion and Future Work

- We propose
  - the RIM problem to handle data uncertainty
  - the LUGreedy algorithm with a provable bound
  - the information cascade based sampling method to reduce the uncertainty and increase the robustness.
- Future work
  - The upper bound of the best robust ratio given a graph?
  - How to provide confidence intervals for a learning algorithm (e.g. MLE)?
  - The big data challenge for social influence analysis
    - Data is not big enough!
    - How to do better sampling, better model learning, and better optimization under the data constraint?

Thank You!