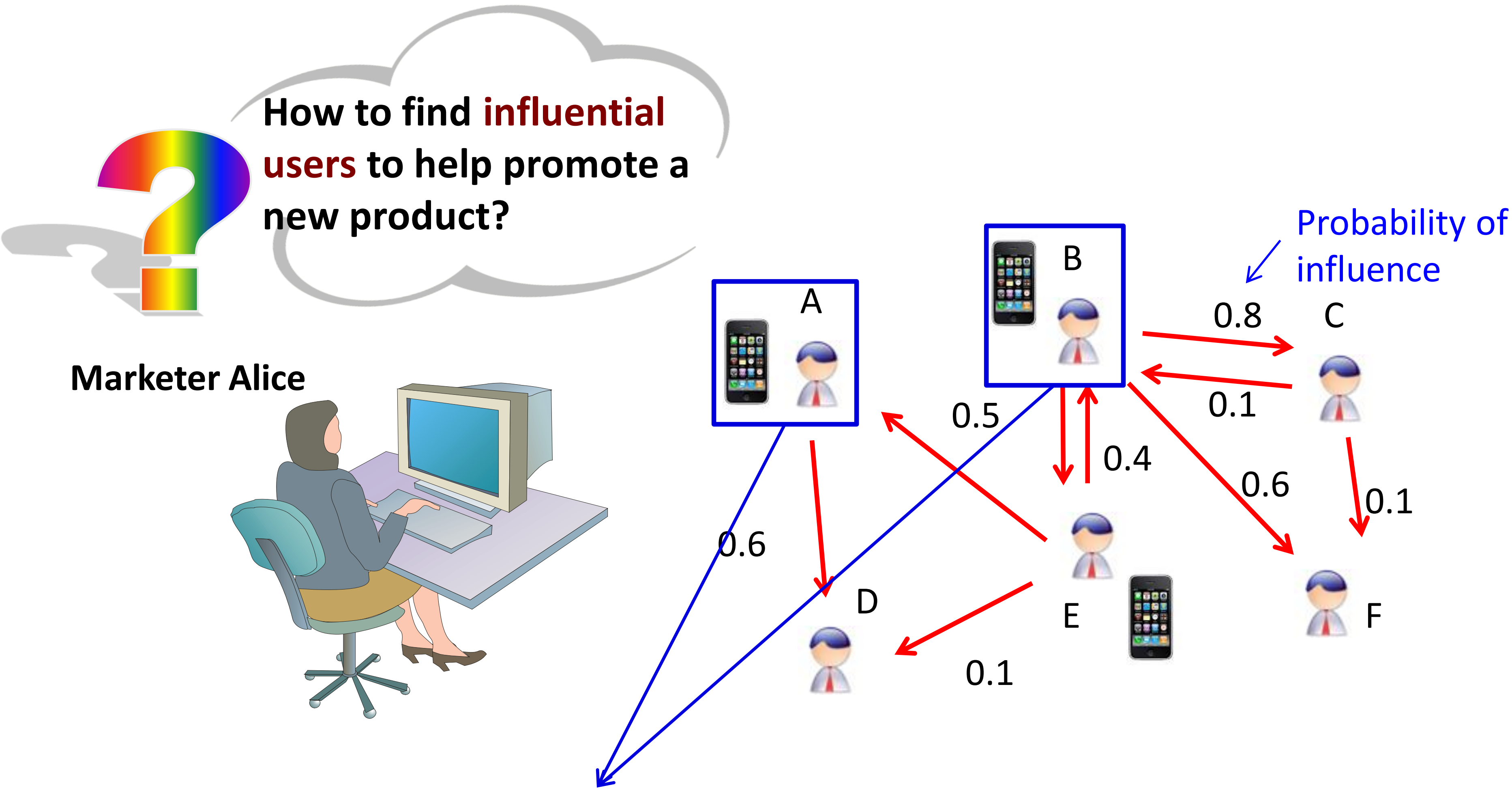
Influence Maximization in Dynamic Social Networks

Honglei Zhuang, Yihan Sun, Jie Tang,
Jialin Zhang, Xiaoming Sun

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Influence Maximization



Find K nodes (users) in a social network that could maximize the spread of influence (Domingos, 01; Richardson, 02; Kempe, 03)

Influence Maximization

- Influence model
  - Initially all users are considered inactive
  - Then the chosen users are activated, who may further influence their friends to be active as well
- Models
  - Linear Threshold model
  - Independent Cascading model

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD'03, pages 137–146, 2003.

Approximate Solution

- NP-hard [1]
  - Linear Threshold Model
  - Independent Cascading Model

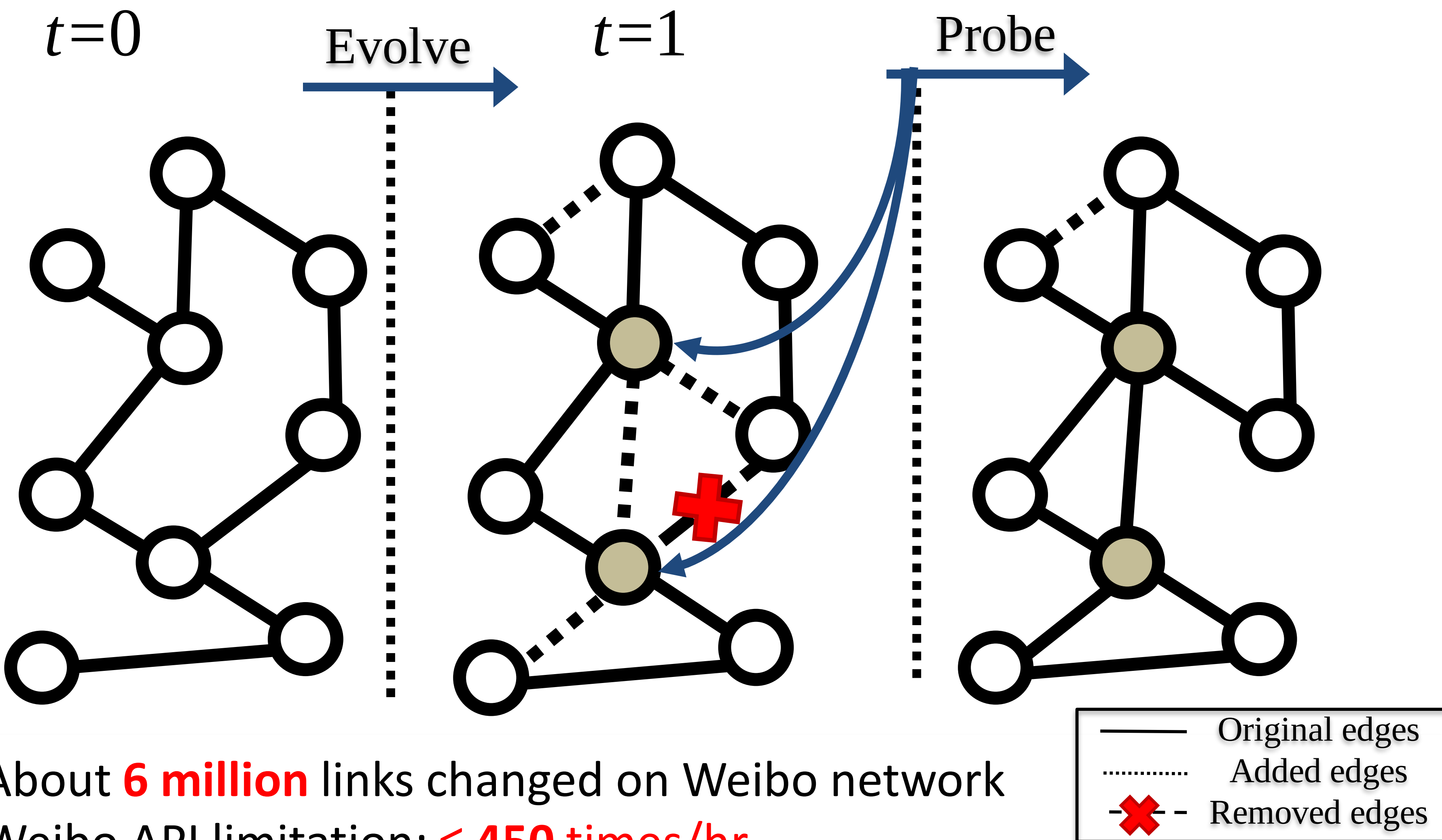
The problem is solved by optimizing a monotonic submodular function

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

- Kempe Prove that approximation algorithms can guarantee that the influence spread is within $\boxed{(1-1/e)}$ of the optimal influence spread.
  - Verify that the two models can outperform the traditional heuristics
- Recent research focuses on the efficiency improvement
  - [2] accelerate the influence procedure by up to 700 times
- It is still challenging to extend these methods to large data sets

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD'03, pages 137–146, 2003.
[2] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. KDD'07, pages 420–429, 2007.

Influence Maximization in Dynamic Networks



About **6 million** links changed on Weibo network

Weibo API limitation: ≤ **450** times/hr

Problem

- Input: For a dynamic social network $\{G^0, \ldots, G^t\}$, we have observed G^0, but for all $t>0$, G^t is unknown
- Problem: To probe b nodes, observe their neighbors to obtain an observed network $\hat{G}^t$ from $\hat{G}^{t-1} / G^0$, such that influence maximization on the real network G^t can be approximated by that on the observed network.
- Challenge: How to find the k influential users, if we only partially observe the update of the social network?

Basic Idea

- Estimate how likely the neighborhood of a node will change in a dynamic social network
  - Probe nodes that change a lot
- Estimate how much the influence spread can be improved by probing a node
  - Probe the one maximizes the improvement

Methodologies and Results

Preliminary Theoretical Analysis

- Formal definition of loss

Max seed set on fully observed network

$$\ell = E_{G|\hat{G}}\left[\left|Q\left(\boxed{S^*}\right) - Q\left(\boxed{T^*}\right)\right|\right]$$

Max seed set on partially observed network

- With an specified evolving graph model
  - At each time stamp an edge is chosen uniformly
  - and its head will point to a node randomly chosen with probability proportional to the in-degree

Preliminary Theoretical Analysis

- Error bound of Random probing strategy

$$\begin{aligned}\ell^t_{Rand} \quad &\leq \quad \sum_{x \in S^*} \frac{4np}{m} \left[\hat{d}^{t'_x}(x) + \frac{1}{4} p \left(\hat{d}^{t'_x}(x) \right)^2 \right] \\ &+ \quad \sum_{x \in T^*} \frac{4np}{m} \left[\hat{d}^{t'_x}(x) + \frac{1}{4} p \left(\hat{d}^{t'_x}(x) \right)^2 \right]\end{aligned}$$

- Error bound of Degree weighted probing strategy

$$\ell^t_{DegRR} \quad \leq \quad 16pk + 2p^2 \left[\sum_{x \in S^*} \hat{d}^{t'_x}(x) + \sum_{x \in T^*} \hat{d}^{t'_x}(x) \right]$$

- In most cases, degree weighted probing strategy performs better than random probing strategy

Maximum Gap Probing

- Basic Idea
  - Estimate how much the influence spread can be improved by probing a node
  - Probe the one which maximizes the improvement
- Formally,
  - For a given tolerance probability ε
  - The minimum value β that satisfies the following inequality is defined as performance gap $\beta(v)$

$$P\left[\hat{Q}_v\left(\boxed{S'_o(v)}\right)-\hat{Q}_v\left(\boxed{S_o}\right)\geq\beta\right]\leq\varepsilon$$

Best solution if v is probed

Best solution before probing

*To simplify problem, define the quality function as the sum of degree in the seed set.

Maximum Gap Probing

- Assume the degree of a node is a martingale. We can estimate the degree gap of each node by

$$P\left[d^{t}(v)-\boxed{d^{t-c_v}(v)}\geq\boxed{\sqrt{-2c_v\ln\varepsilon}}\right]\leq\varepsilon$$

Last time when v is probed

Defined as z_v

- Considering the node to probe is in/not in the current seed set.

$$\beta(v)=\begin{cases}\max\left\{0,\hat{d}(v)+z_v-\min_{w\in S_o}\hat{d}(w)\right\},v\notin S_O\\ \max\left\{0,\max_{u\notin S_o}\hat{d}(u)-\hat{d}(v)+z_v\right\},v\in S_O\end{cases}$$

- Each time, choose the one with maximum gap $\beta(v)$ to probe

MaxG Algorithm

Input: G^0, T, ϵ, b
Output: Seed set S^t at $t = 1, 2, \cdots, T$

$\hat{G} \leftarrow G^0$; $\forall v \in V, c_v \leftarrow 0$;
for $t = 1$ **to** T **do**
  $\forall v \in V, c_v \leftarrow c_v + 1$;
  for b *times* **do**
    $S_o \leftarrow k$ nodes with maximum $\hat{d}_{in}(v)$;
    $\hat{d}_{max} = \max_{u \notin S_o} \hat{d}_{in}(u)$;
    $\hat{d}_{min} = \min_{w \in S_o} \hat{d}_{in}(w)$;
    foreach $v \in V$ **do**
      $z_v \leftarrow \sqrt{-2 c_v \ln \epsilon}$;
      if $v \in S$ **then**
      $\beta_v \leftarrow \max \left\{ 0, \hat{d}_{max} - \hat{d}_{in}(v) + z_v \right\}$;
      else $\beta_v \leftarrow \max \left\{ 0, \hat{d}_{in}(v) + z_v - \hat{d}_{min} \right\}$;
    $v^* \leftarrow \arg\max_{v \in V} \beta_v$, $c_{v^*} \leftarrow 0$;
    Probe v^* in G^t and update $\hat{G}$;
  `// Degree discount heuristics`
  $S^t \leftarrow \emptyset$;
  for k *times* **do**
    $v^* \leftarrow \arg\max_{v \in V \setminus S^t} \hat{h}_{S^t}(v)$;
    $S^t \leftarrow S^t \cup \{v^*\}$;
    foreach *neighbor* u *of* v^* **do**
      Update $\hat{h}_{S^t}(u)$;
  Output S^t;

Finding nodes to probe by maximizing the degree gap

Perform the standard greedy algorithm (degree discount heuristics) for influence maximization

Experiment Setup

- Data sets

Data sets	#Users	#Relationships	#Time stamps
Synthetic	500	12,475	200
Twitter	18,089,810	21,097,569	10
Coauthor	1,629,217	2,623,832	27

- Evaluation
  - Take optimal seed set S' obtained from partially observed network
  - Calculate its influence spread on real network

Experiment Setup

- Comparing methods
  - *Rand, Enum:* Uniform probing
  - *Deg, DegRR:* Degree-weighted probing
  - *BEST:* Suppose network dynamics fully observed
- Configurations
  - Probing budget:
    - b=1,5 for Synthetic; b=100,500 for Twitter and Coauthor
  - Seed set size for influence maximization:
    - k=30 for Synthetic; k=100 for Twitter and Coauthor
  - Independent Cascade Model, with uniform p=0.01

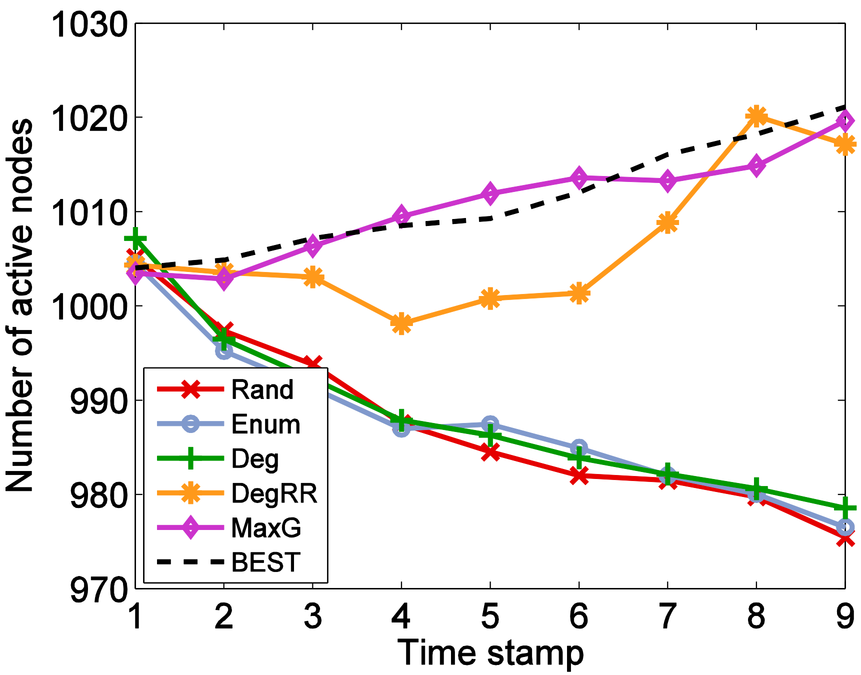
Experimental Results

- Average influence spread

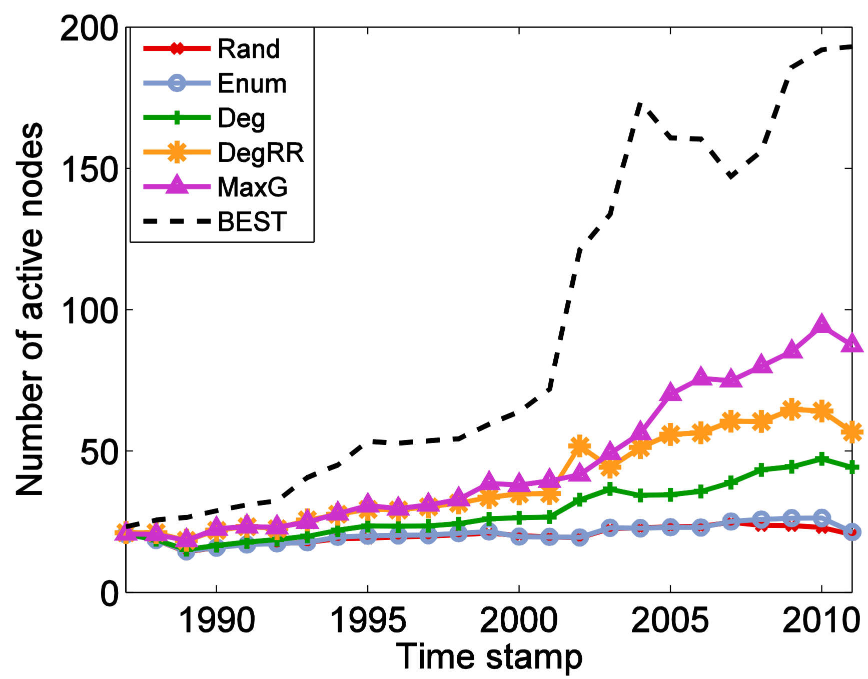
Data Set	b	Rand	Enum	Deg	DegRR	MaxG	BEST
Synthetic	1	13.83	13.55	13.78	14.30	**14.79**	15.95
	5	15.07	15.33	15.09	15.40	**15.60**	
Twitter	100	987.74	987.62	988.41	1001.47	**1005.12**	1011.15
	500	987.45	987.67	988.36	1006.38	**1010.61**	
Coauthor	100	20.34	20.82	28.67	38.94	**45.51**	91.51
	500	20.35	22.93	44.27	56.68	**61.74**	

The large, the best

Influence Maximization Results (b=100)

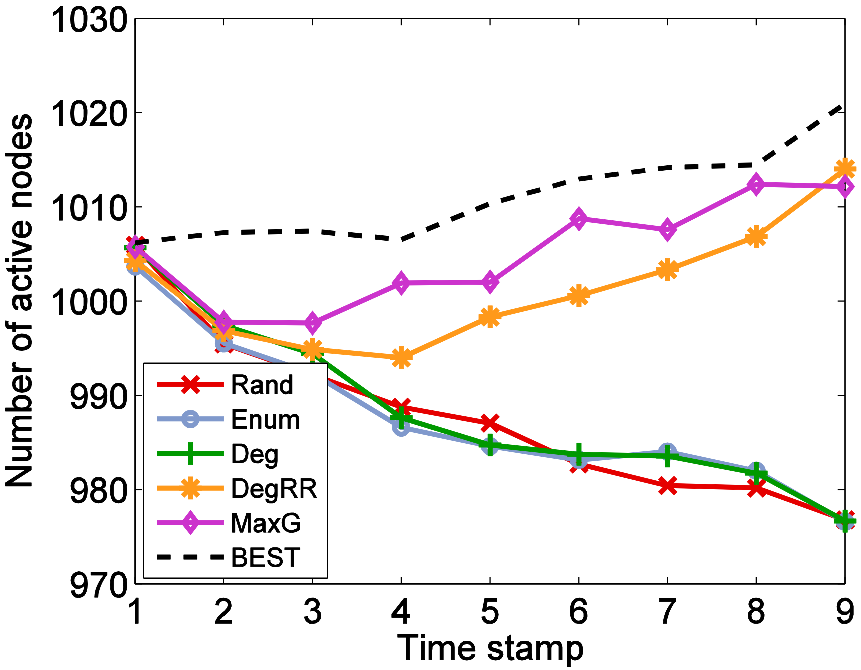


Twitter

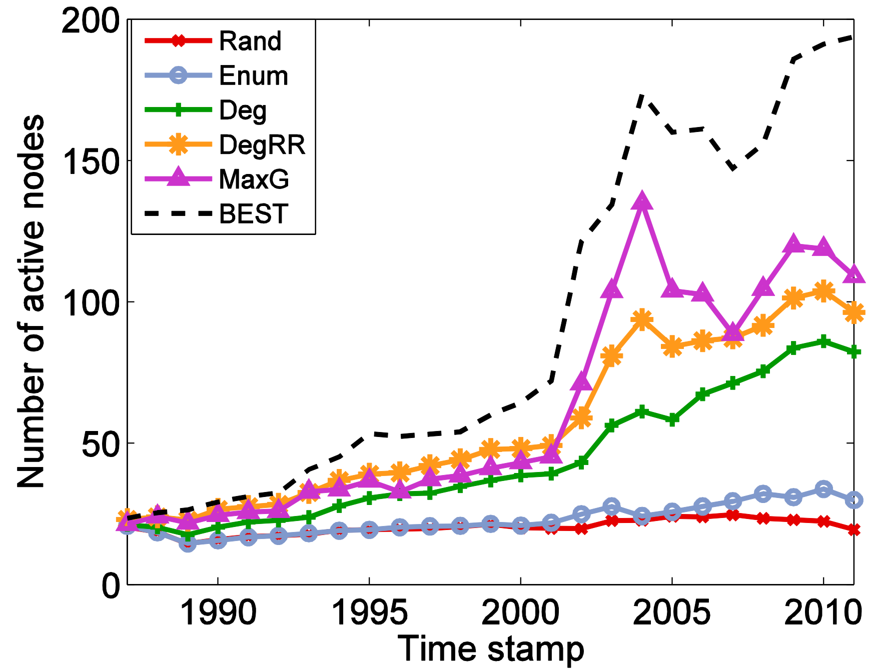


Coauthor

Influence Maximization Results (b=500)



Twitter



Coauthor

Conclusions

Conclusions

- Propose a probing algorithm to partially update a dynamic social network, so as to guarantee the performance of influence maximization in dynamic social networks

- Future work include:
  - Online updating seed set in dynamic social networks
  - Probing for other applications, e.g. PageRank[1]

[1] B. Bahmani, R. Kumar, M. Mahdian, and E. Upfal. PageRank on an evolving graph. In KDD, pages 24–32, 2012.

Thank you

- Dec 8, 2013
- *Go to…*
- Motivating Example
- Problem Formulation
- Approach
- Experiments