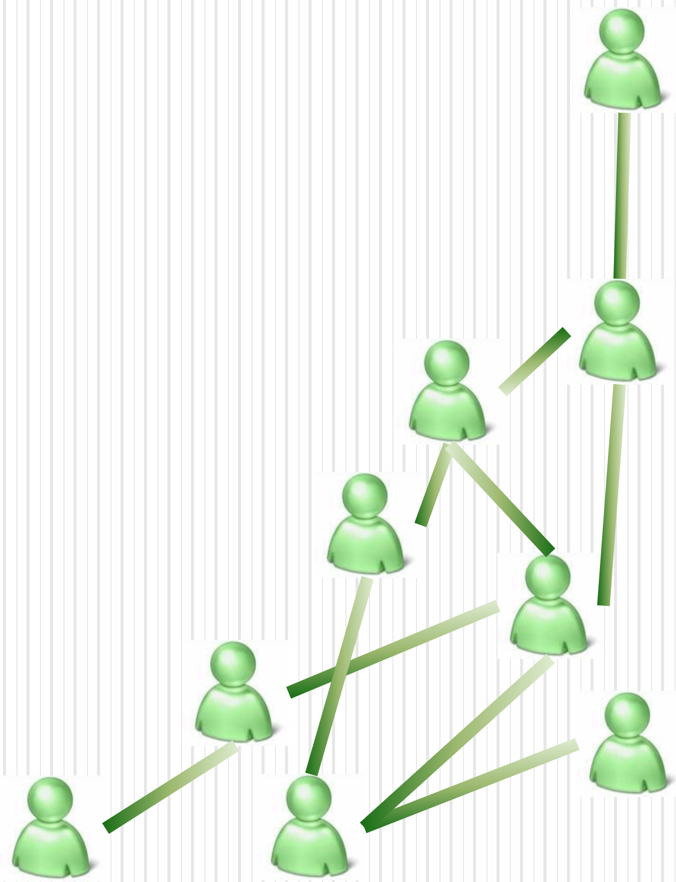
Influence diffusion dynamics and influence maximization in complex social networks

Wei Chen

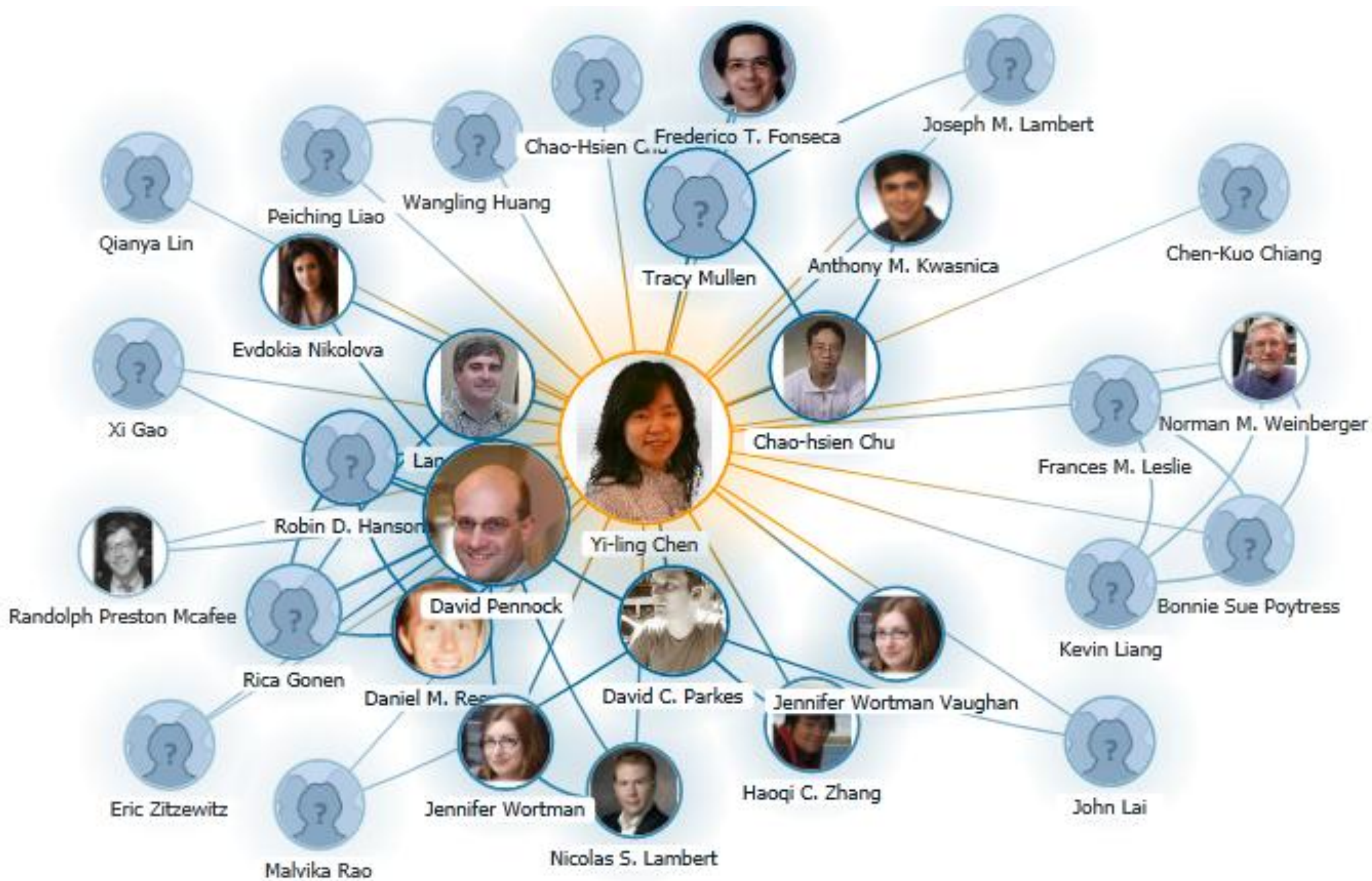
陈卫

Microsoft Research Asia



Harvard, Oct. 18, 2011

(Social) networks are natural phenomena



Booming of online social networks

facebook

开心网

myspace

人人网 renren.com

twitter

天涯社区 www.tianya.cn

Opportunities and challenges on the research of online social networks

- Opportunities
  - massive data set, real time, dynamic, open
  - help social scientists to understand social interactions in a large scale
  - help marketing people to target to the right audience
  - help economists to understand social economic networks
- Challenges
  - graph structure based large scale data analysis
  - scalable graph algorithm design
  - realistic modeling of network formation, evolution, and information/influence diffusion in networks

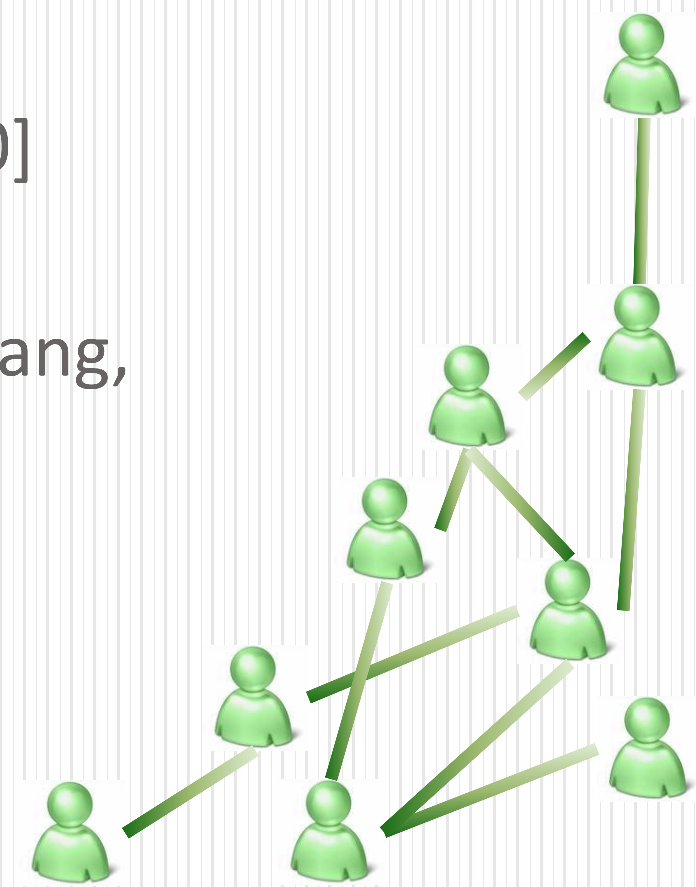
Our recent work on social network related research

- Social influence in social networks
  - scalable influence maximization
  - influence maximization with complex social interactions
- Game-theoretic based modeling of social interaction
  - bounded budget betweenness centrality game for network formation
  - Optimal pricing in social networks with networked effect
- Fundamental algorithms for large graphs
  - fast distance queries in power-law graphs
  - game-theoretic approach to community detection

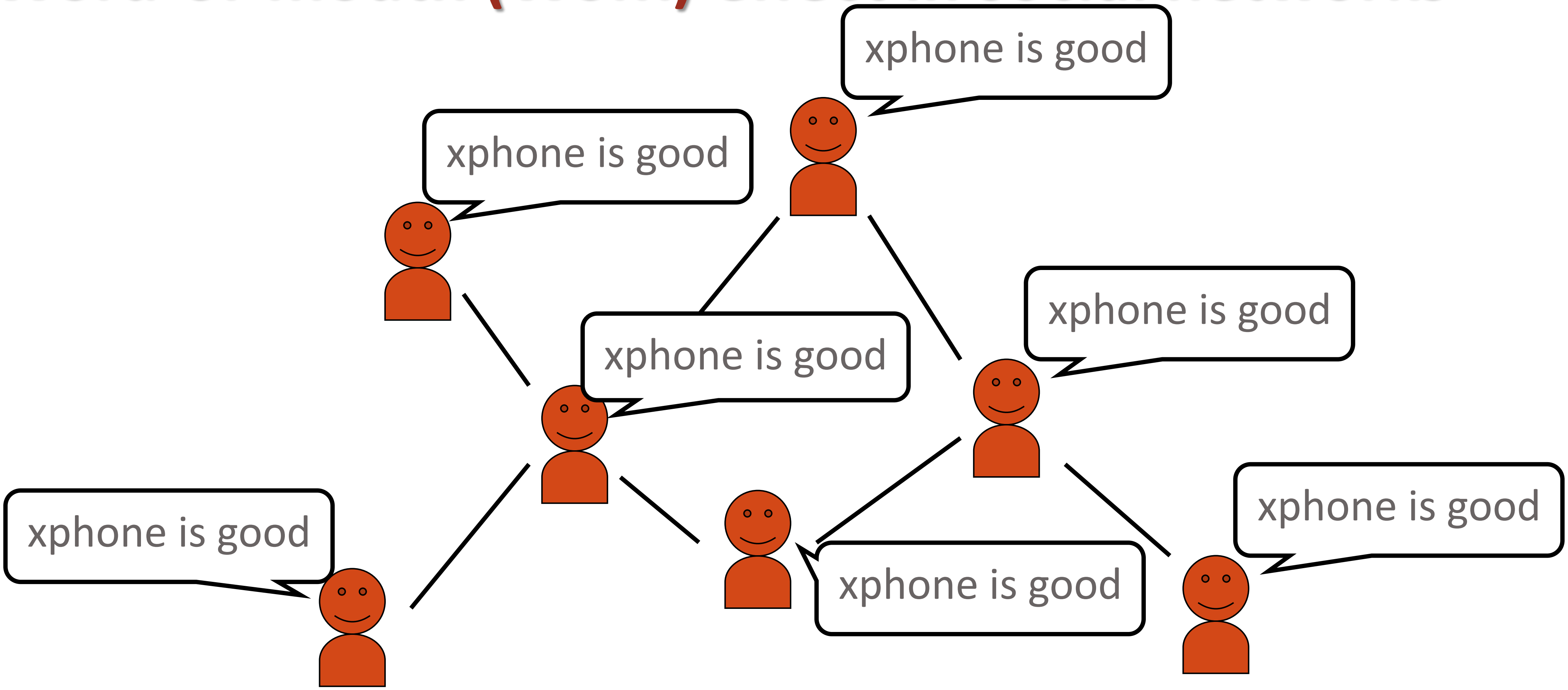
Scalable Influence Maximization in Social Networks

[KDD’09, KDD’10, ICDM’10]

Collaborators:
Chi Wang, Yajun Wang, Siyu Yang,
Yifei Yuan, Li Zhang



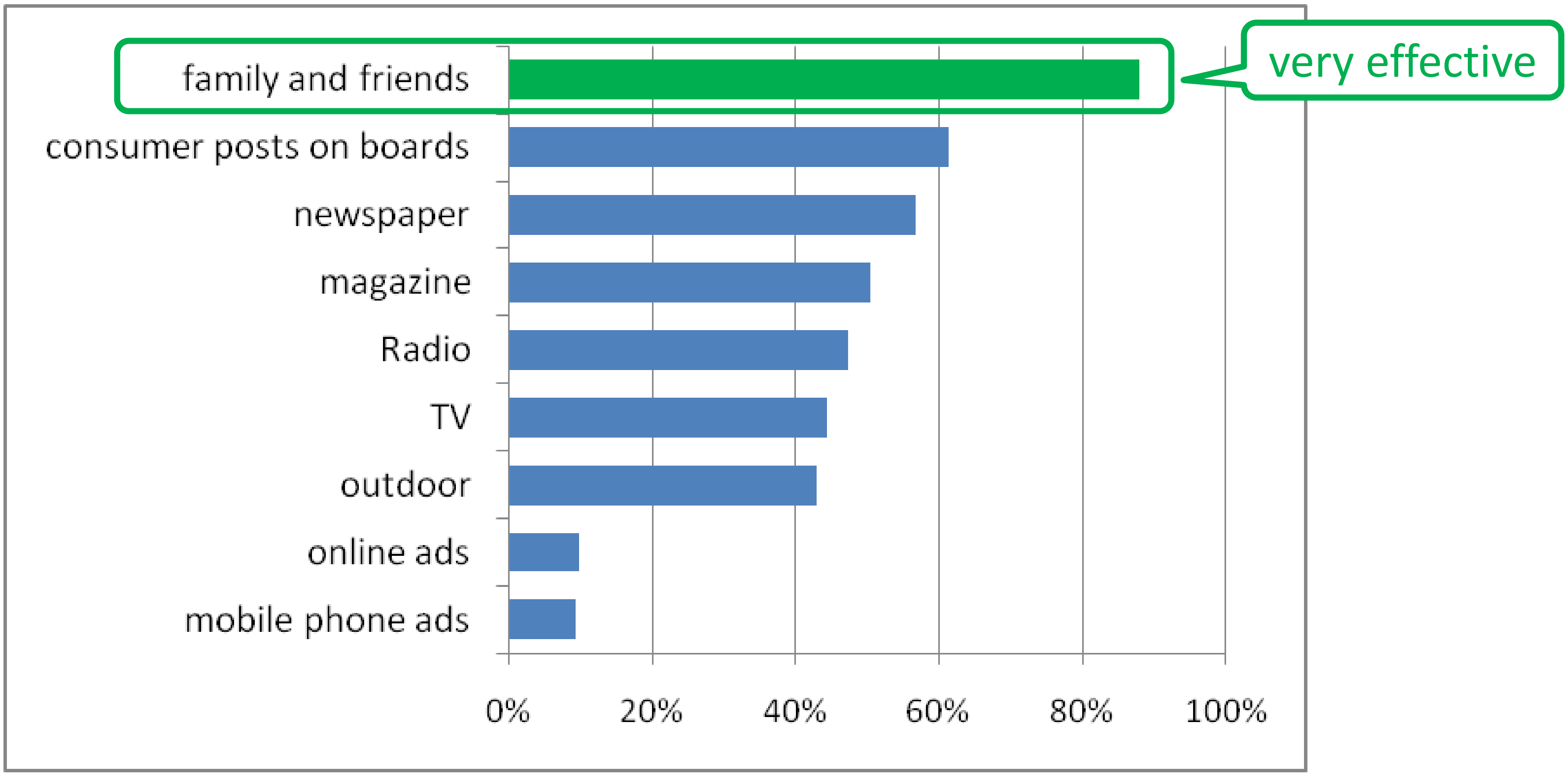
Word-of-mouth (WoM) effect in social networks



- Word-of-mouth effect is believed to be a promising advertising strategy.
- Increasing popularity of online social networks may enable large scale WoM marketing

WoM (or Viral) Marketing

level of trust on different types of ads *



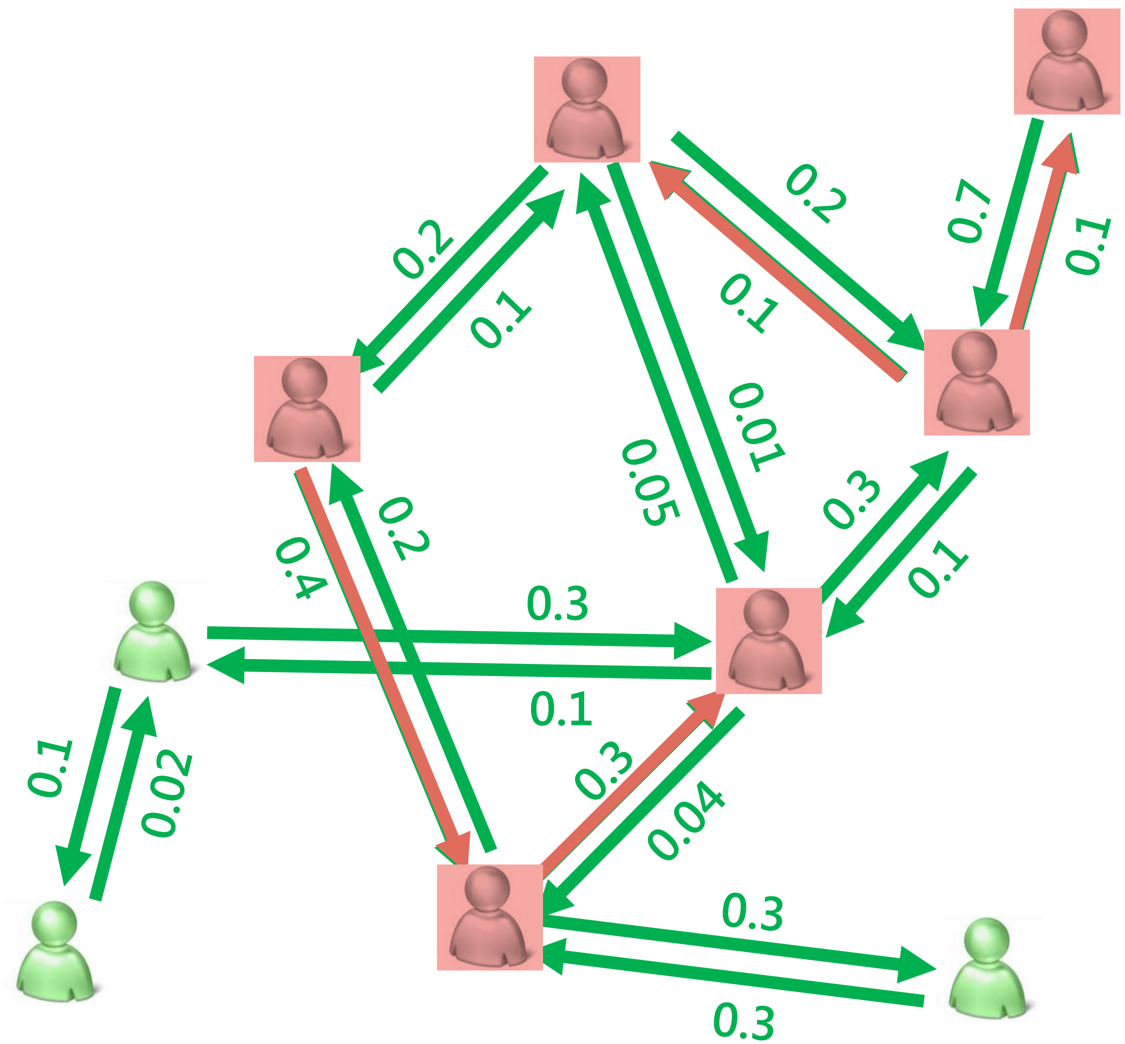
*source from Forrester Research and Intelliseek

Two key components for studying WoM marketing

- Modeling influence diffusion dynamics, prior work includes:
  - independent cascade (IC) model
  - linear threshold (LT) model
  - voter model
- Influence maximization, prior work includes:
  - greedy approximation algorithm
  - centrality based heuristics

The Problem of Influence Maximization

- Social influence graph
  - vertices are individuals
  - links are social relationships
  - number p(u,v) on a directed link from u to v is the probability that v is activated by u after u is activated
- Independent cascade model
  - initially some *seed* nodes are activated
  - At each step, each newly activated node u activates its neighbor v with probability p(u,v)
- Influence maximization:
  - find *k* seeds that generate the largest expected influence



Prior Work

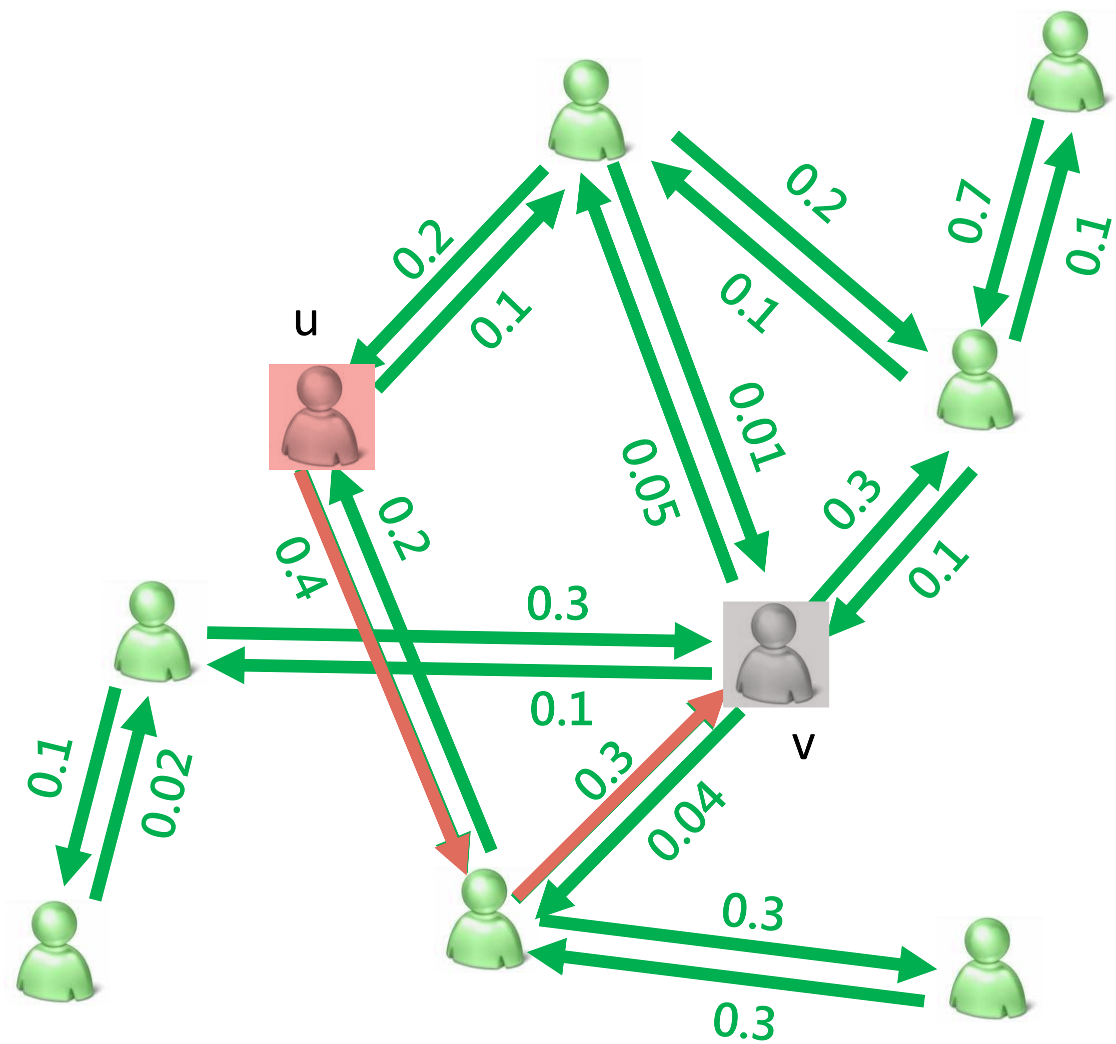
- Influence maximization as a discrete optimization problem proposed by Kempe, Kleinberg, and Tardos, 2003
  - Introduce Independent Cascade (IC) and Linear Threshold (LT) models
  - Finding optimal solution is provably hard (NP-hard)
  - Greedy approximation algorithm, 63% approximation of the optimal solution
    - select k seeds in k iterations
    - in each iteration, select one seed that provides the largest marginal increase in influence spread
- Several subsequent studies improved the running time
- Serious drawback:
  - very slow, not scalable: > 3 hrs on a 30k node graph for 50 seeds

Our Work

- Exact influence computation is #P hard, for both IC and LT models --- computation bottleneck
- Design new heuristics
  - MIA (maximum influence arborescence) heuristic [KDD'10]
    - for general independent cascade model (more realistic)
    - 10^3 speedup --- from hours to seconds
    - influence spread close to that of the greedy algorithm of [KKT'03]
  - Degree discount heuristic [KDD'09]
    - for uniform independent cascade model
    - 10^6 speedup --- from hours to milliseconds
  - LDAG (local directed acyclic graph) heuristic [ICDM'10]
    - for the linear threshold model
    - 10^3 speedup --- from hours to seconds

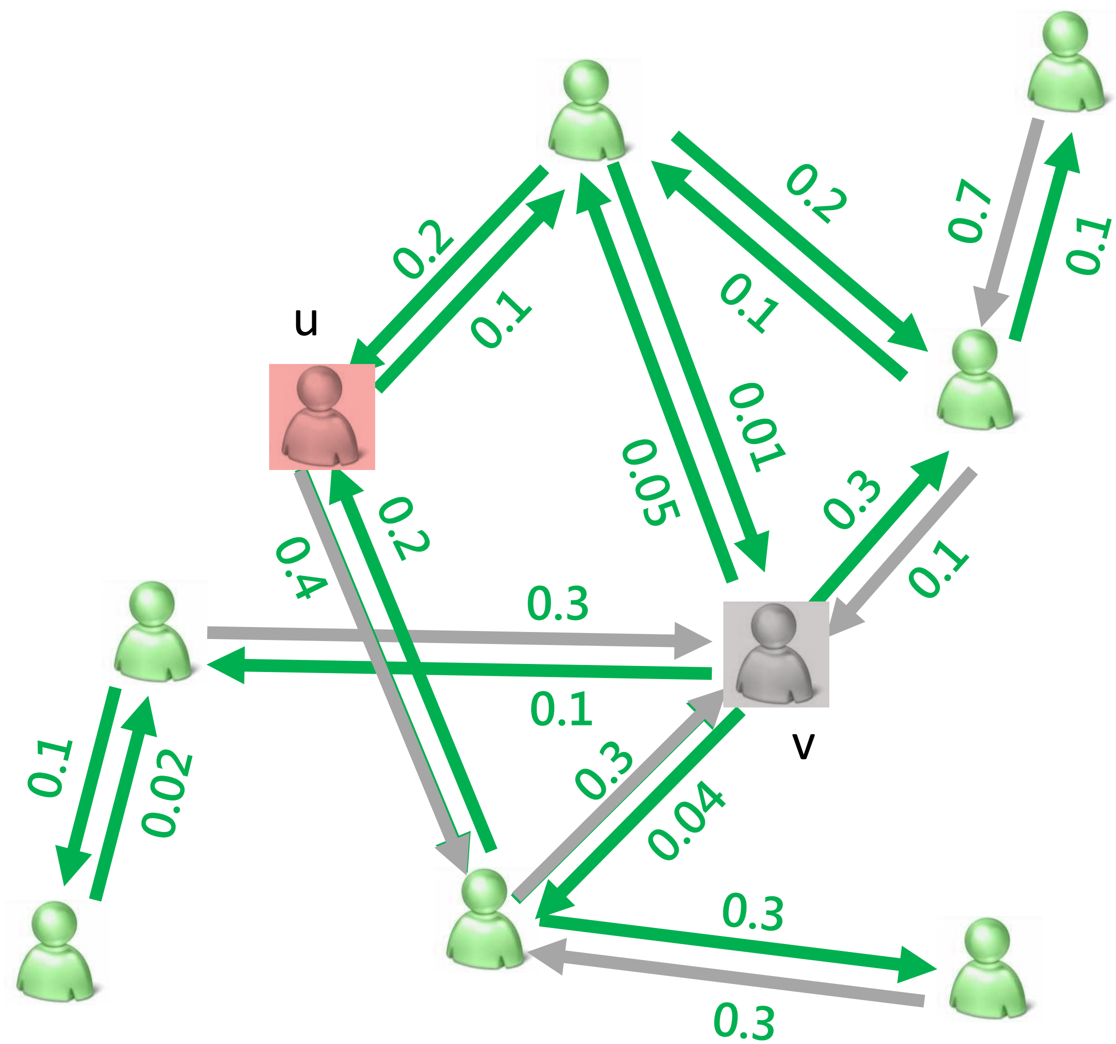
Maximum Influence Arborescence (MIA) Heuristic

- For any pair of nodes u and v, find the maximum influence path (MIP) from u to v
- ignore MIPs with too small probabilities (< parameter θ)



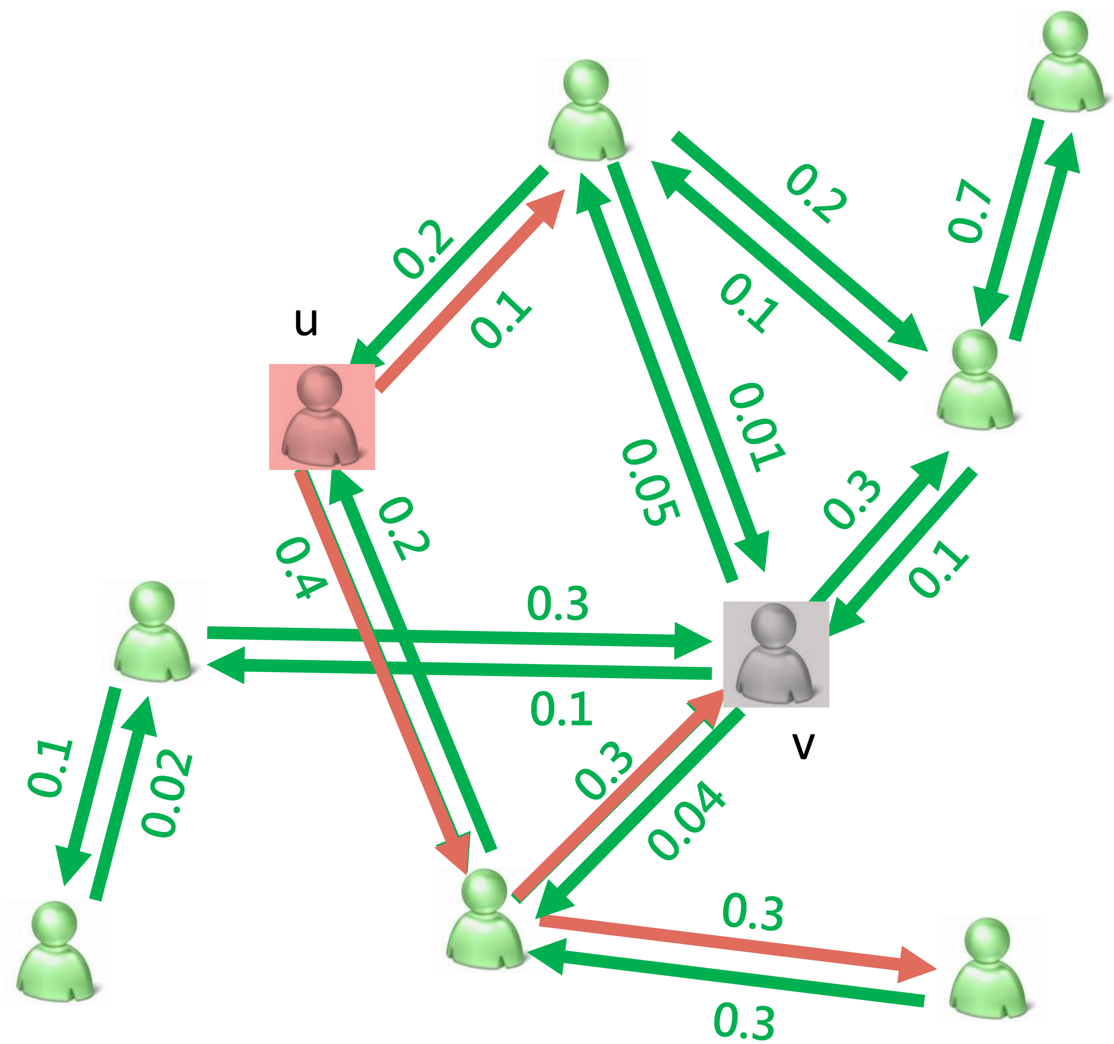
MIA Heuristic (cont'd)

- Local influence regions
  - for every node v, all MIPs to v form its maximum influence in-arborescence (MIIA)



MIA Heuristic (cont'd)

- Local influence regions
  - for every node v, all MIPs to v form its maximum influence in-arborescence (MIIA)
  - for every node u, all MIPs from u form its maximum influence out-arborescence (MIOA)
  - computing MIAs and the influence through MIAs is fast



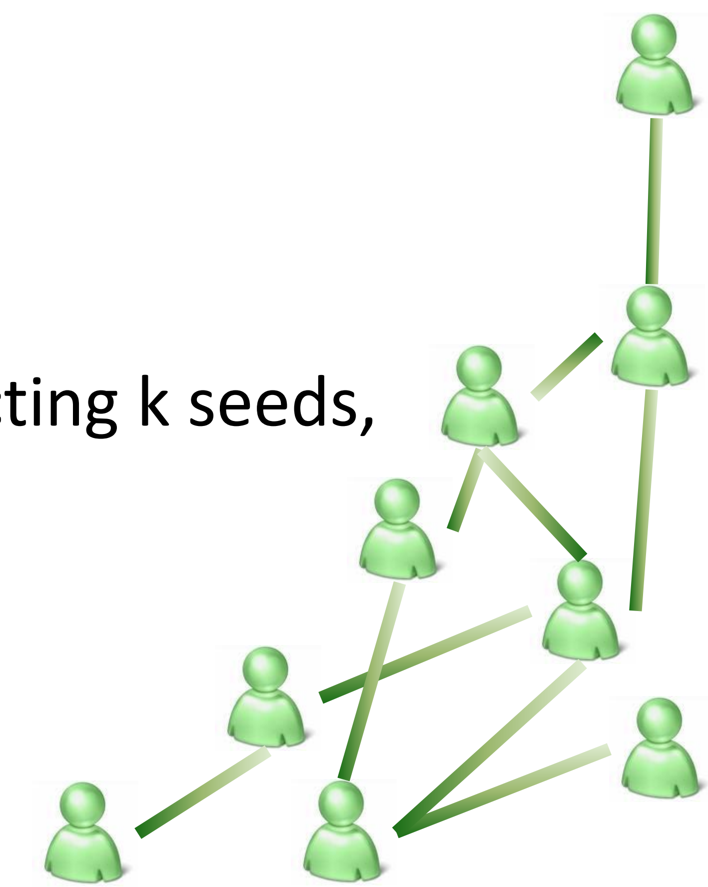
MIA Heuristic III: Computing Influence through the MIA structure

- Recursive computation of activation probability ap(u) of a node u in its in-arborescence, given a seed set S

Algorithm 2 $ap(u, S, MIIA(v, \theta))$

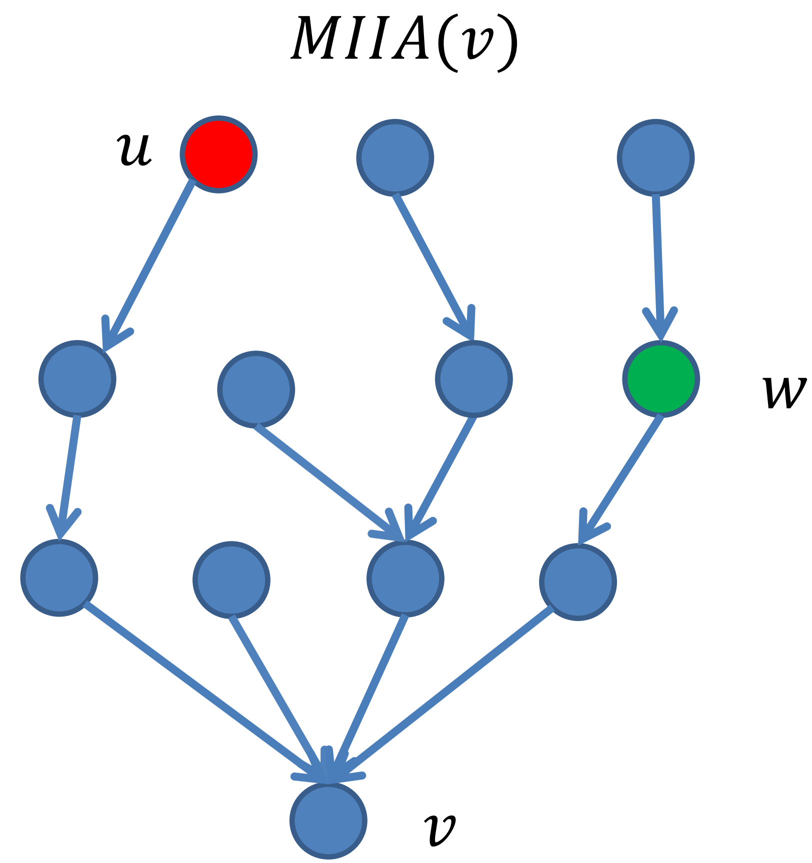
```
if u ∈ S then
    ap(u) = 1
else if Ch(u) = ∅ then
    ap(u) = 0
else
    ap(u) = 1 − Π_{w∈Ch(u)} (1 − ap(w) · pp(w, u))
end if
```

- Can be used in the greedy algorithm for selecting k seeds, but not efficient enough



MIA Heuristic IV: Efficient updates on incremental activation probabilities

- u is the new seed in $MIIA(v)$
- Naive update: for each candidate w, redo the computation in the previous page to compute w's incremental influence to v
  - $O(|MIIA(v)|^2)$
- Fast update: based on linear relationship of activation probabilities between any node w and root v, update incremental influence of all w's to v in two passes
  - $O(|MIIA(v)|)$

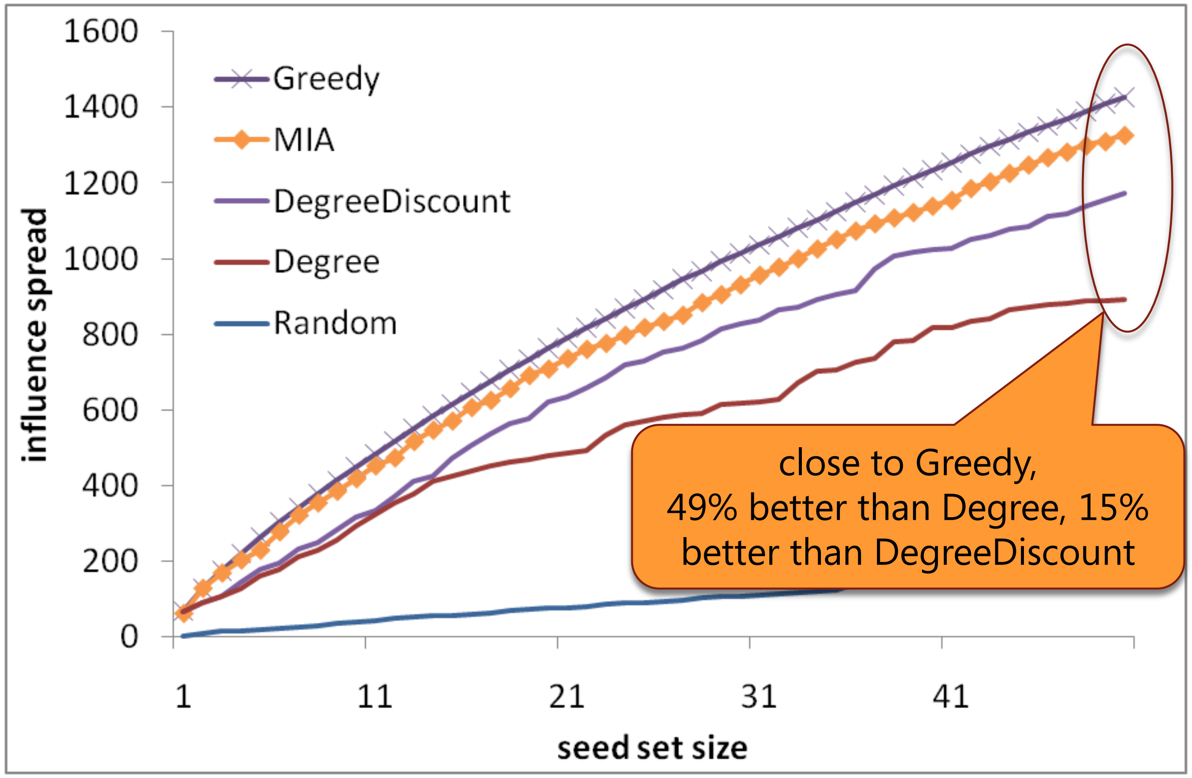


MIA Heuristic (cont'd)

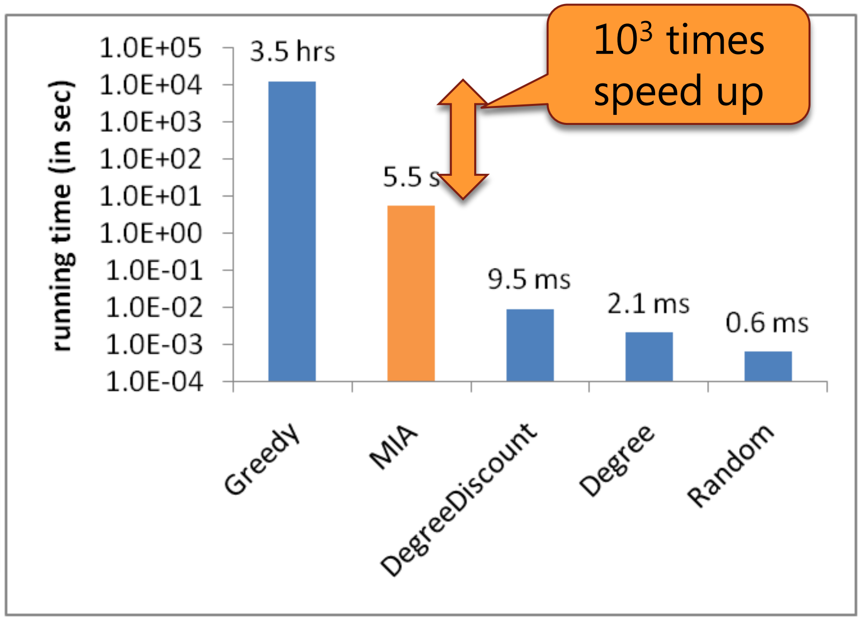
- Iteration between two steps
  - Selecting the node v giving the largest marginal influence
  - Update MIAs after selecting v as the seed
- Key features:
  - updates are local
  - local updates are linear to the local tree structure

Experiment Results on MIA heuristic

Influence spread vs. seed set size



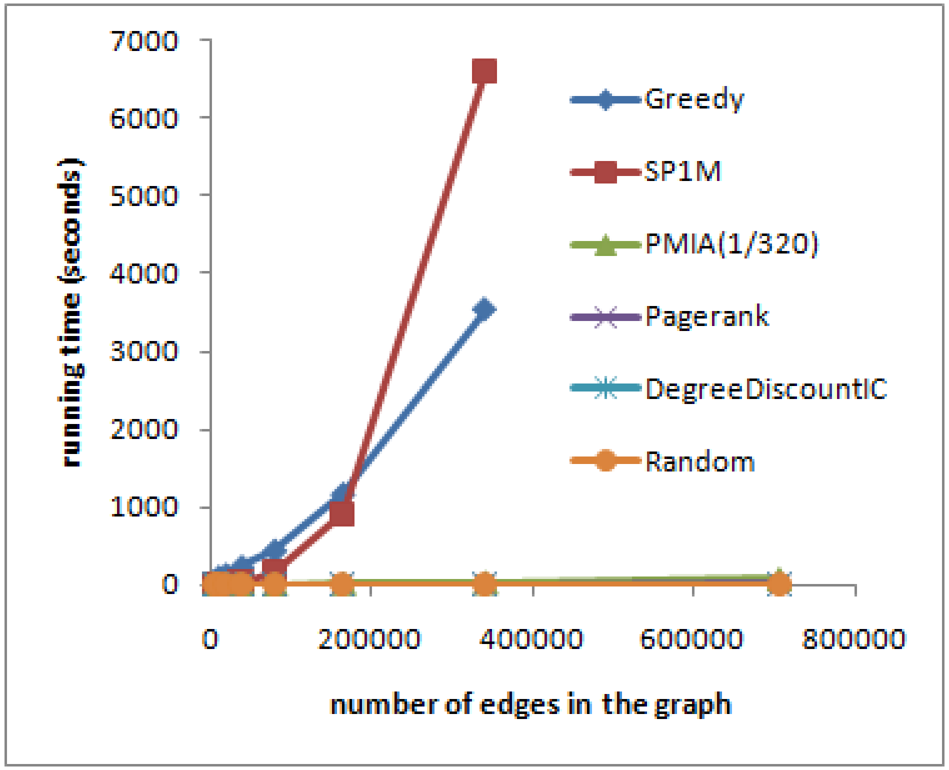
running time



Experiment setup:

- 35k nodes from coauthorship graph in physics archive
- influence probability to a node v = 1 / (# of neighbors of v)
- running time is for selecting 50 seeds

Scalability of MIA heuristic



Experiment setup:

- synthesized graphs of different sizes generated from power-law graph model
- influence probability to a node v = 1 / (# of neighbors of v)
- running time is for selecting 50 seeds

Summary

- Scalable influence maximization algorithms
  - MixedGreedy and DegreeDiscount [KDD'09]
  - PMIA for the IC model [KDD'10]
  - LDAG for the LT model [ICDM'10]
- PMIA/LDAG have become state-of-the-art benchmark algorithms for Inf. Max.
- Collective citation count above 110 in less than 2 years

Handling Complex Social Interactions

[SDM'11, others under submissions]

Alex Collins, Rachel Cummings, Te Ke, Zhenming Liu, David Rincon, Xiaorui Sun, Yajun Wang, Wei Wei, Yifei Yuan, Xinran He, Guojie Song, Yanhua Li, Katie Everett, Zhi-Li Zhang

Handling complex social interactions

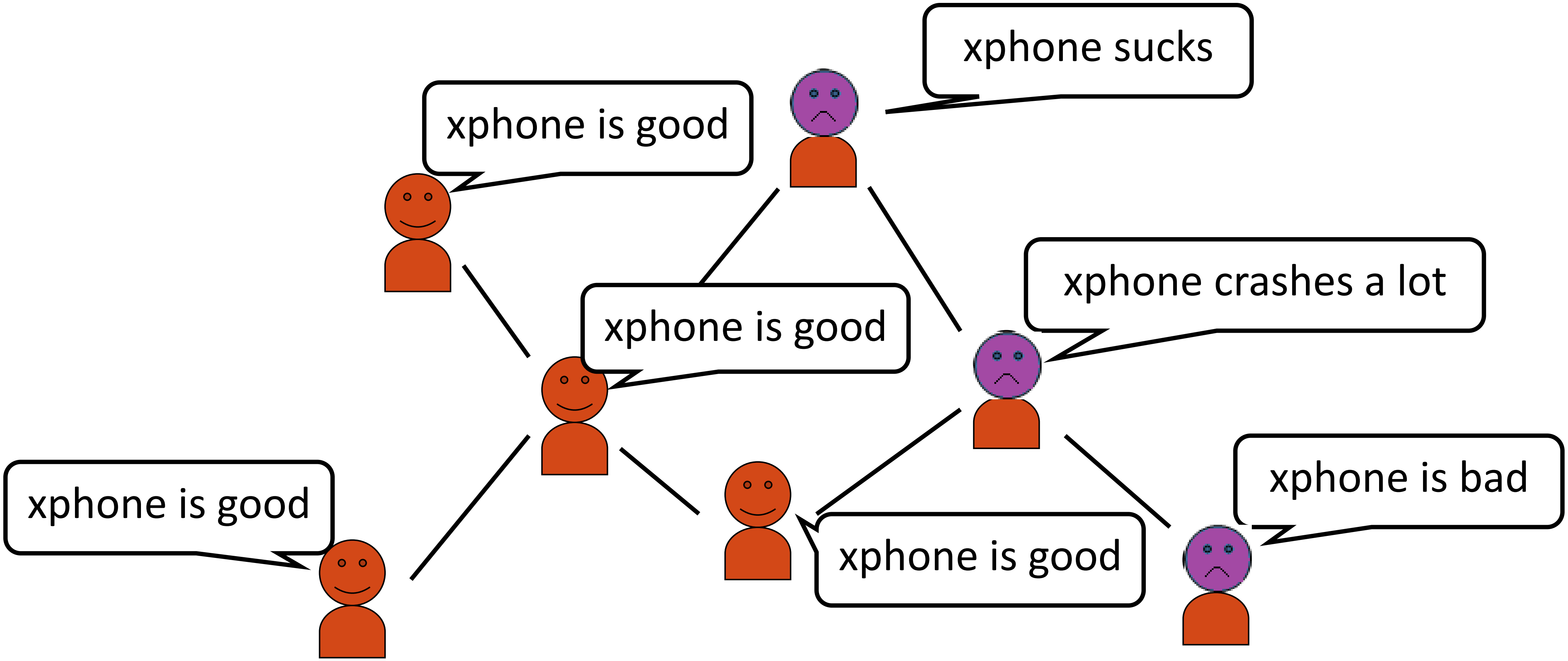
- people may dislike a product after usage and spread bad words about it
- a competing product may compete for social influence in the social network
- social relationships may be friends or foes

Our solutions

- people may dislike a product after usage and spread bad words about it
  - IC-N model and MIA-N algorithm
- a competing product may compete for social influence in the social network
  - CLT model and CLDAG algorithm for influence blocking maximization
- social relationships may be friends or foes
  - voter model in signed networks with exact inf. max. algorithm

IC-N model and MIA-N algorithm for the emergence and propagation of negative opinions

Negative opinions may emerge and propagate

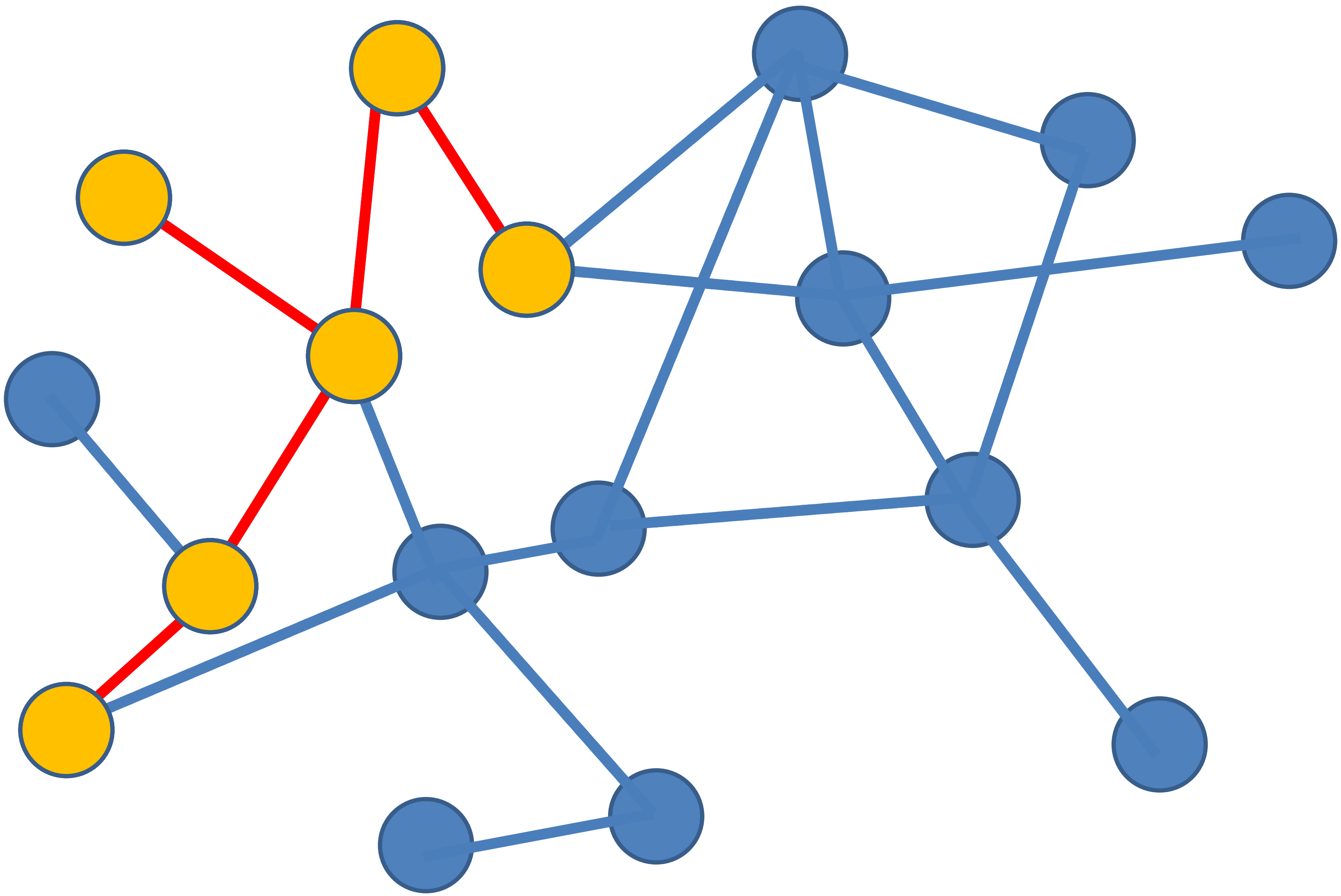


- Negative opinions originates from poor product/service quality
- Negative opinions may be more contagious --- *negativity bias*

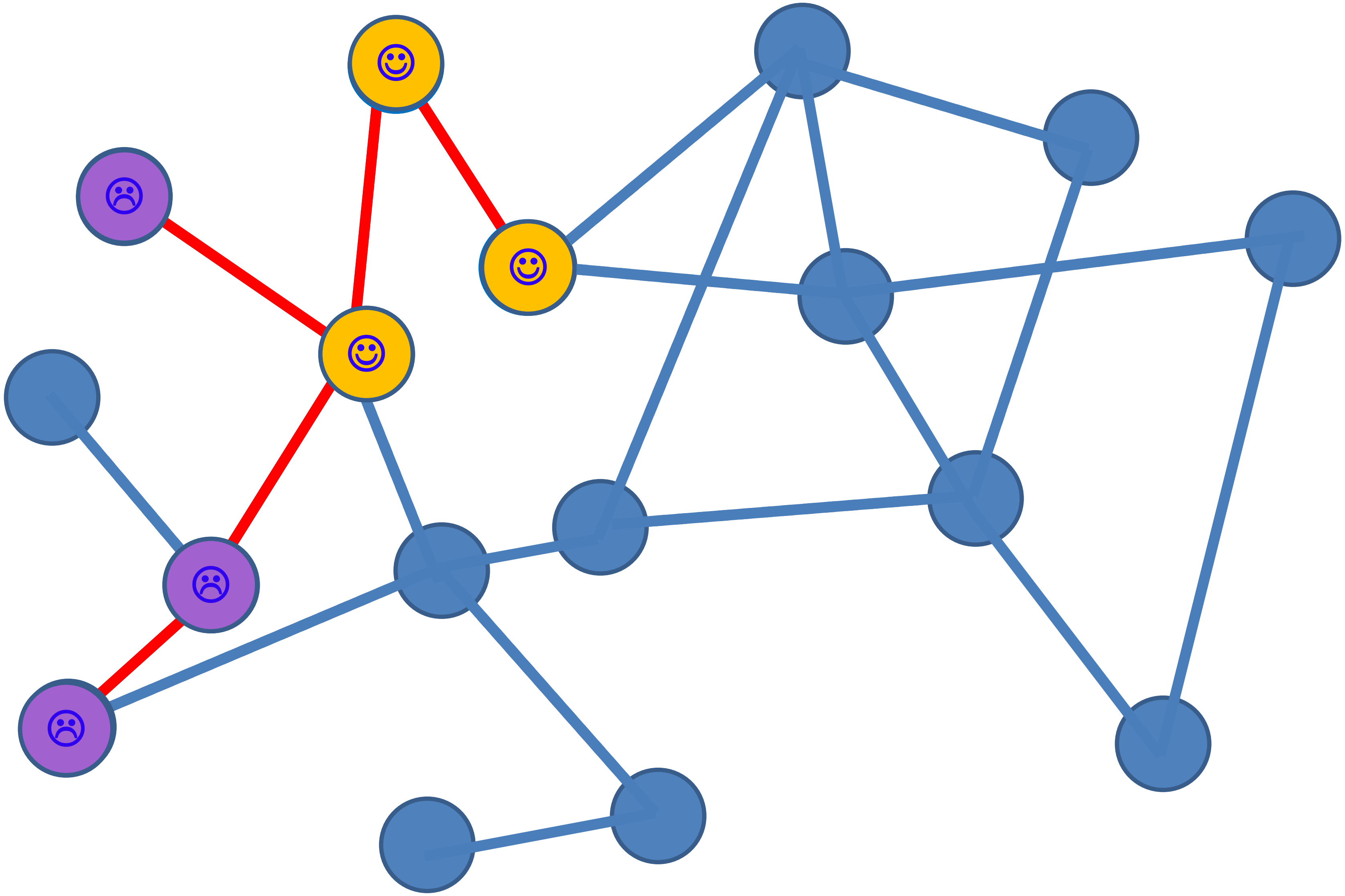
Negative opinion model

- Extention of the independent cascade model
- The quality of the product to be advertised is characterized by the quality factor (QF) $q \in [0,1]$.
- Each node could be in 3 states
  - Inactive, positive, and negative.
- When node v becomes active,
  - If the influencer is negative, the activated influencee is also negative (negative node generates negative opinions).
  - If the influencer is positive, the activated influencee
    - is positive with prob. q.
    - is negative with prob. $1 - q$.
  - If multiple activations of a node occur at the same step, randomly pick one
  - Asymmetric --- negativity bias

Independent Cascading Process (**without** considering QF)



Independent Cascading Process (**when** considering QF)



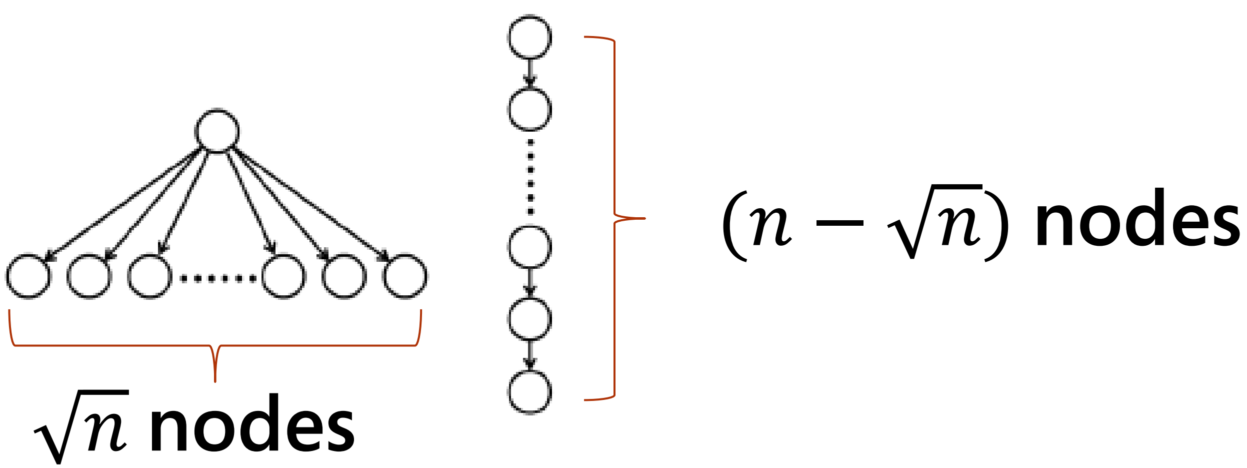
Our results (1)

- Complexity and approximation algorithm results

Scenario	Objective function	Algorithm result	Negative result
General directed graphs	Maximize expected positive nodes	$(1-\frac{1}{e}-\varepsilon)$-approx alg, due to submodularity	Exact sol. is NP hard.
General directed graphs	Maximize expected (positive – negative) nodes.	Exists an $(1-\frac{1}{e}-\varepsilon)$-approx alg. Only when q is sufficiently large	Same as above
Directed graphs with different q for different people	Maximize expected positive nodes	NA	Objective is non-submodular

Our results (2)

- Q: is the knowledge of quality factor important?
  - guess a "universally good" value q so that regardless of the actual quality factor, the seeds are good?
  - No: ∃ social networks s.t. a wrong guess of q could lead to a much worse result than the optimal one. ($\Theta(\sqrt{n/k})$)
  - Intuition: which one seed to select in the following graph?

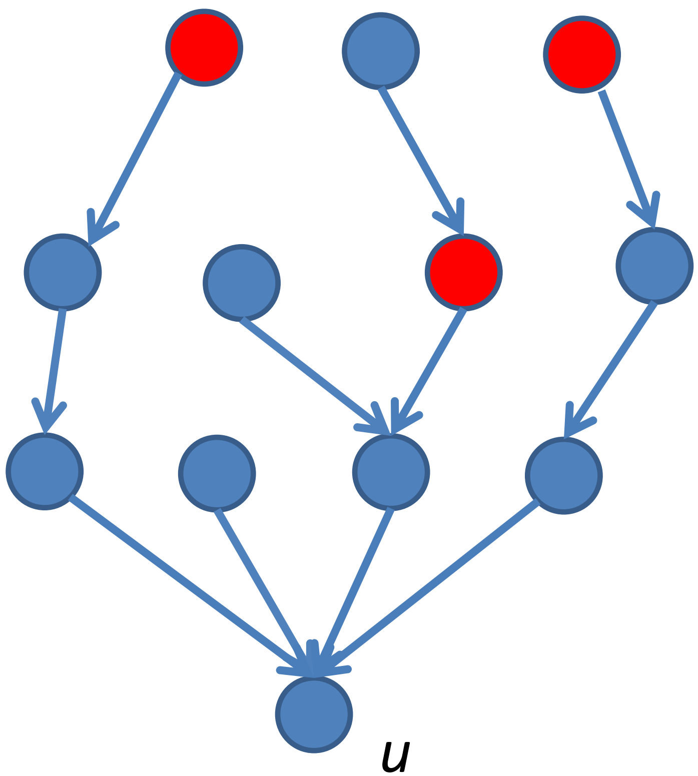


Our results (3)

- Q: what is the bottleneck of the approx. alg.
  - Given a specific seed set *S*, can we evaluate the expected number of positive nodes?
    - In general, #P-hard; can use Monte Carlo to approximate.
    - But exists efficient exact algorithm for arborescence (trees).
  - Developed scalable heuristic MIA-N based on influence calculation alg. for arborescences.

Computation in directed trees (in-arborescences)

- Without negative opinions, a simple recursion computes the activation probability of u:
  - $ap(u) = 1 - \prod_{w \in N^{in}(u)} (1 - ap(w)p(w,u))$
- Difficulty with negative opinions: needs to know whether the neighbors of u is positive or negative --- because of negativity bias



Solutions for in-arborescences

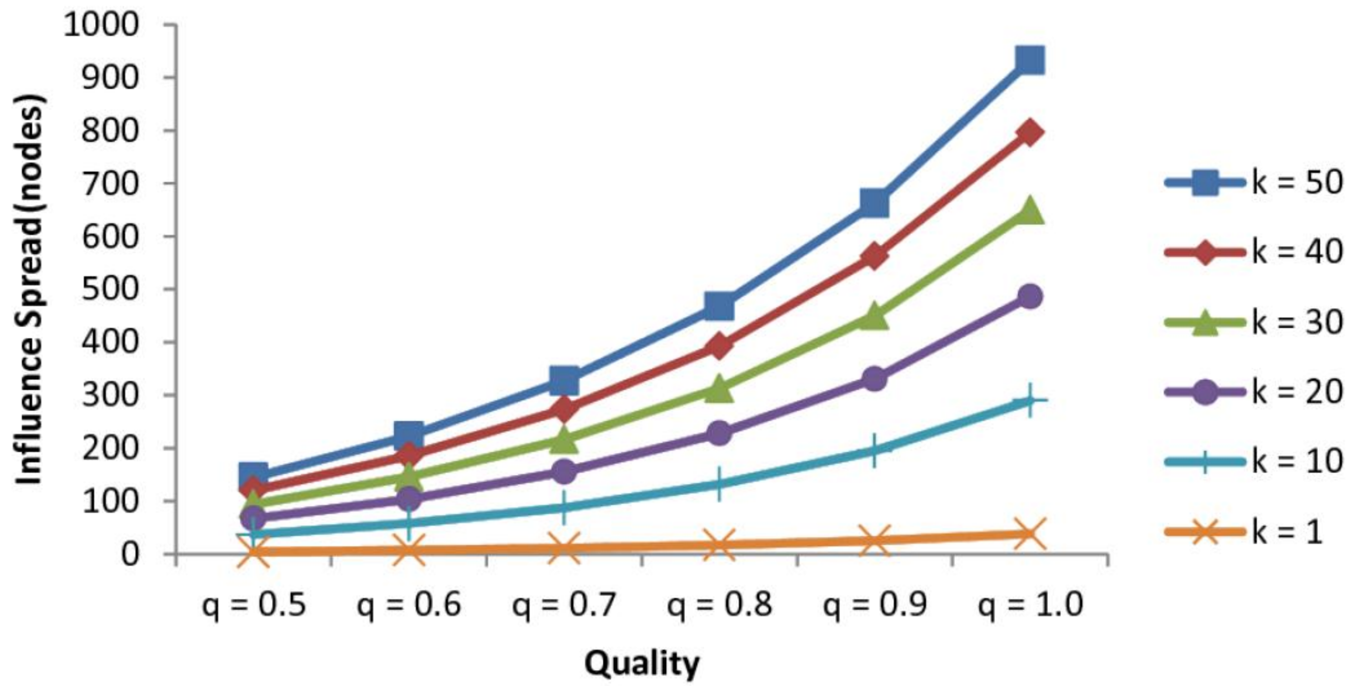
- Step 1: compute activation probability of u at step t (via dynamic programming):

$$ap(u,t) = \begin{cases} 1 & t = 0 \wedge u \in S, \\ 0 & t = 0 \wedge u \notin S, \\ 0 & t > 0 \wedge u \in S, \\ \prod_{w \in N^{in}(u)}[1 - \sum_{i=0}^{t-2} ap(w,i)p(w,u)] \\ \quad - \prod_{w \in N^{in}(u)}[1 - \sum_{i=0}^{t-1} ap(w,i)p(w,u)] & t > 0 \wedge u \notin S. \end{cases}$$

- Step 2: compute positive activation probability of u at step t:

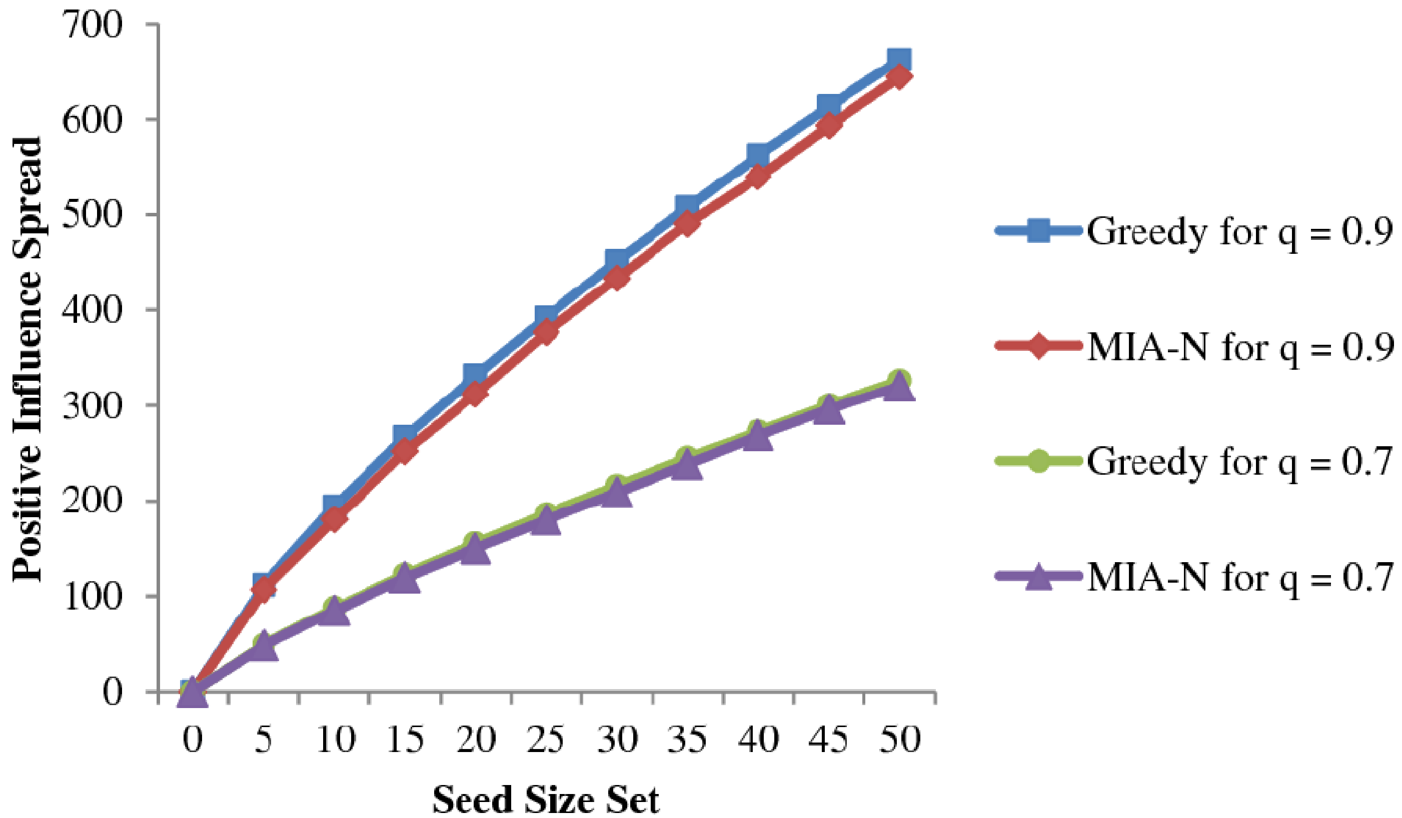
$$pap(u,t) = ap(u,t) \cdot q^{t+1}.$$

Influence spread and QF



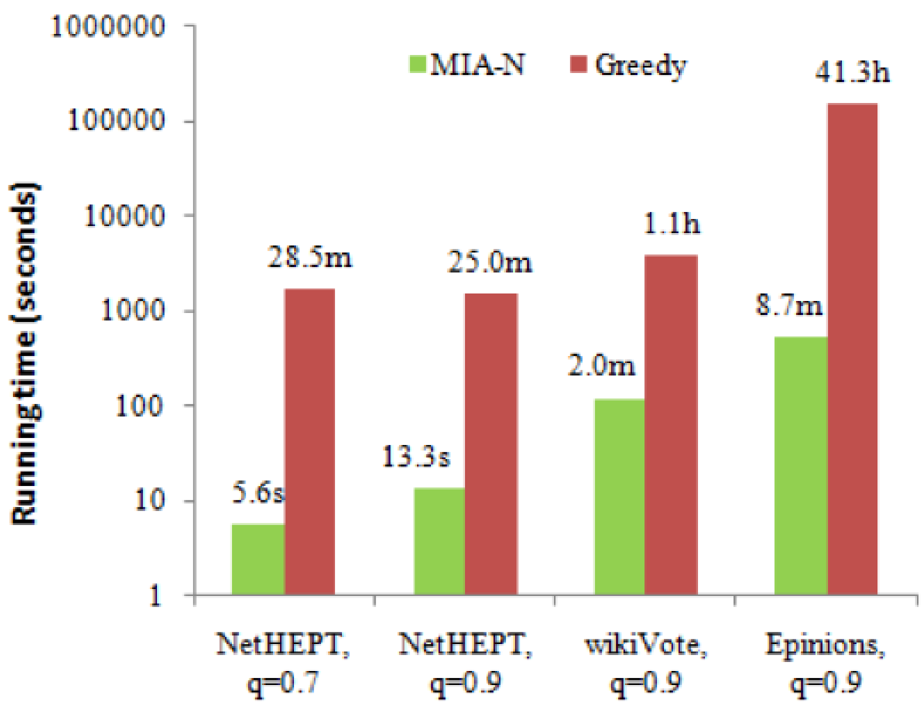
- Results on a collaboration network with 15K nodes.
- Convex function because of negativity bias

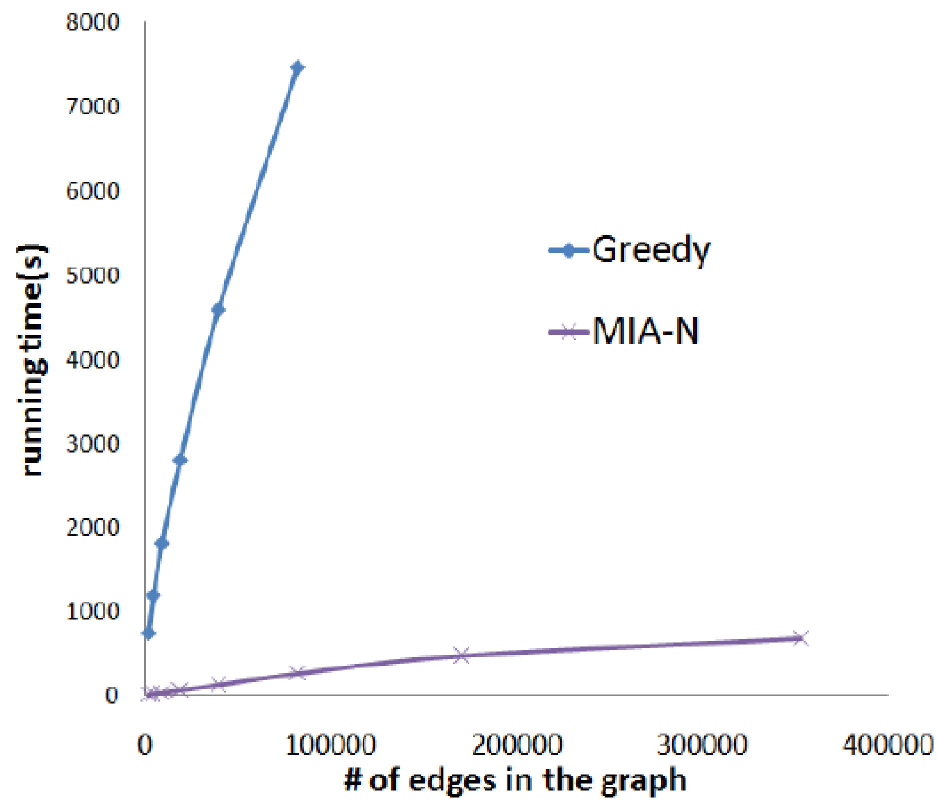
Performance of the heuristic



- **MIA-N heuristic performs nearly as good as the original greedy algorithm.**

Scalability





- **MIA-N heuristic is 3 orders of magnitude faster than Greedy**

CLT model for competitive influence diffusion and CLDAG algorithm for the influence blocking maximization problem

The problem

- Consider two competing influence diffusion process, one positive and one negative
- Inf. Blocking Max.: selecting positive seeds to block the negative influence diffusion as much as possible
  - e.g. stop rumors on a company, on a political candidate, on public safety events, etc.

Our solution

- Competitive linear threshold model
  - positive influence and negative influence diffuse concurrently in the network
  - negative influence dominates in direct competition
- Prove that the objective function is submodular
- Design scalable algorithm CLDAG to achieve fast blocking effect

Influence diffusion on networks with friends and foes

The problem

- You would positively influence your friends, but influence your foes in the reverse direction
- How to model such influence?
- How to design influence maximization algorithm?

Our solution

- Voter model in signed networks
  - suitable for opinion changes from positive to negative or reverse
  - individual takes the opposite opinion from his foe
- Provide complete characterization of short term dynamics and long-term steady state behavior
- Provide exact solutions to the influence maximization problem

On going and future directions

- Model validation and influence analysis from real data
- Even faster heuristic algorithms
- Fast approximate algorithms
- Online and adaptive algorithms

Questions?