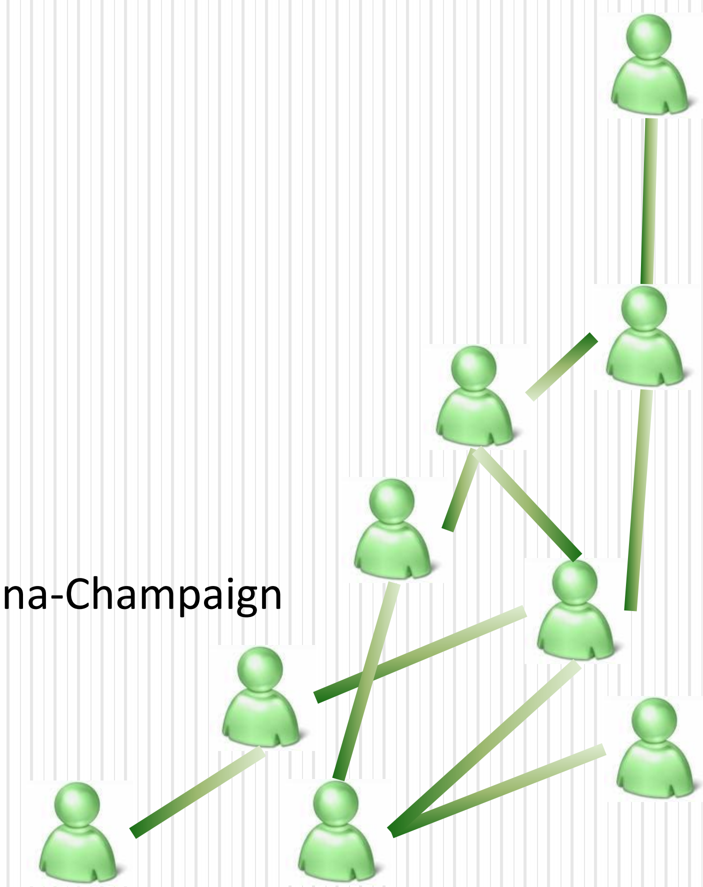
Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks

Wei Chen
Microsoft Research Asia

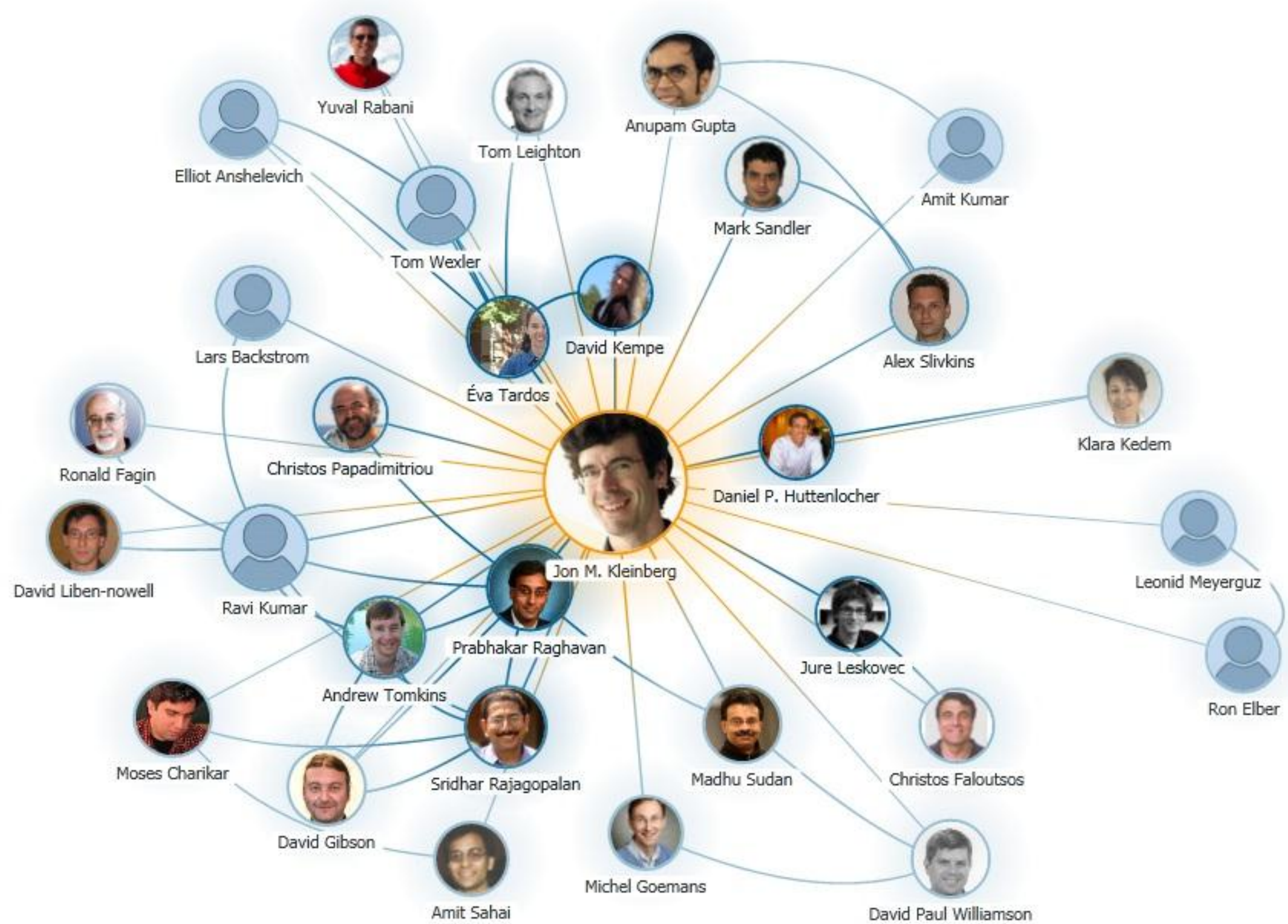
In collaboration with
Chi Wang University of Illinois at Urbana-Champaign
Yajun Wang Microsoft Research Asia



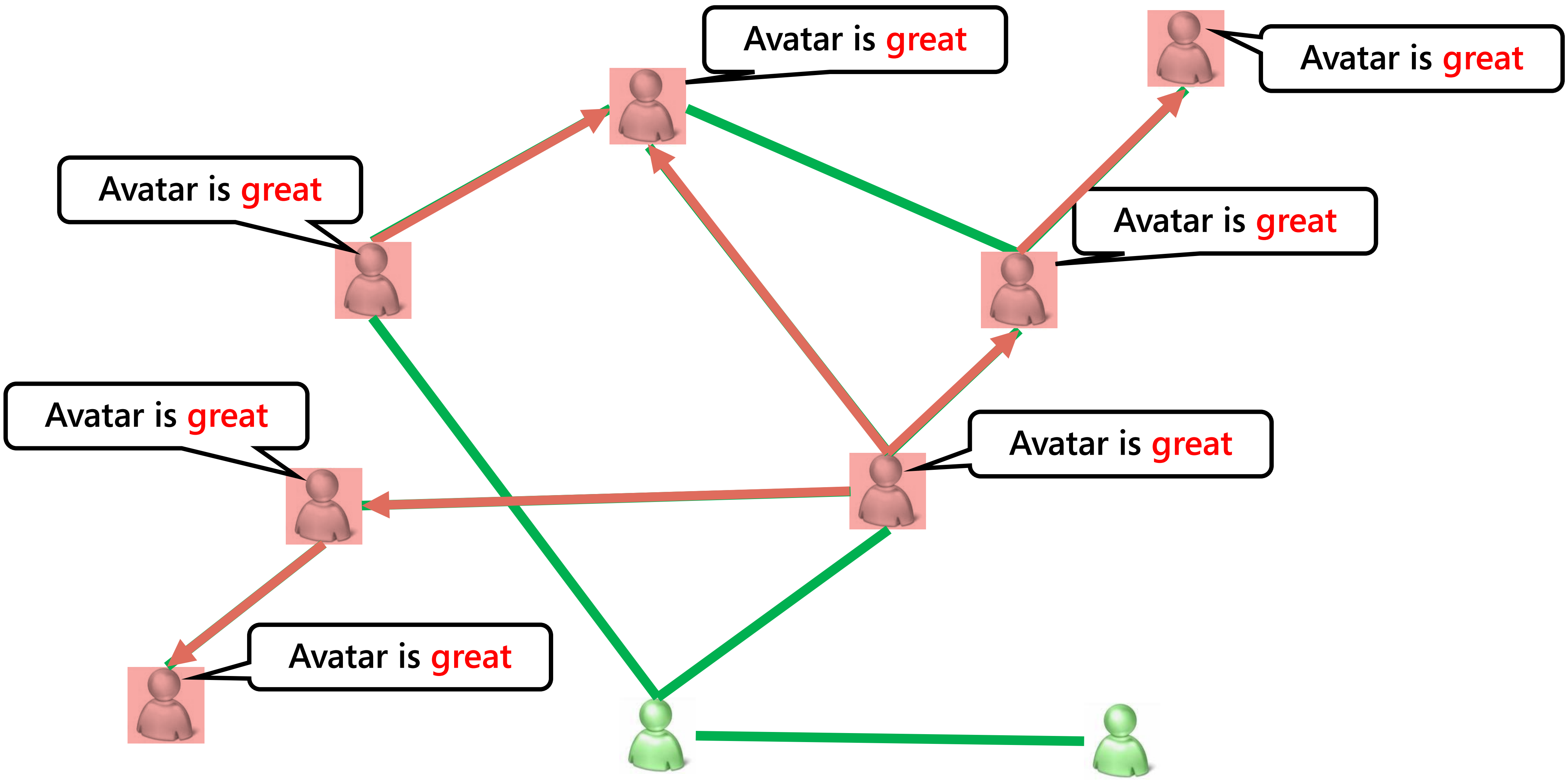
Outline

- Background and problem definition
- Maximum Influence Arborescence (MIA) heuristic
- Experimental evaluations
- Related work and future directions

Ubiquitous Social Networks

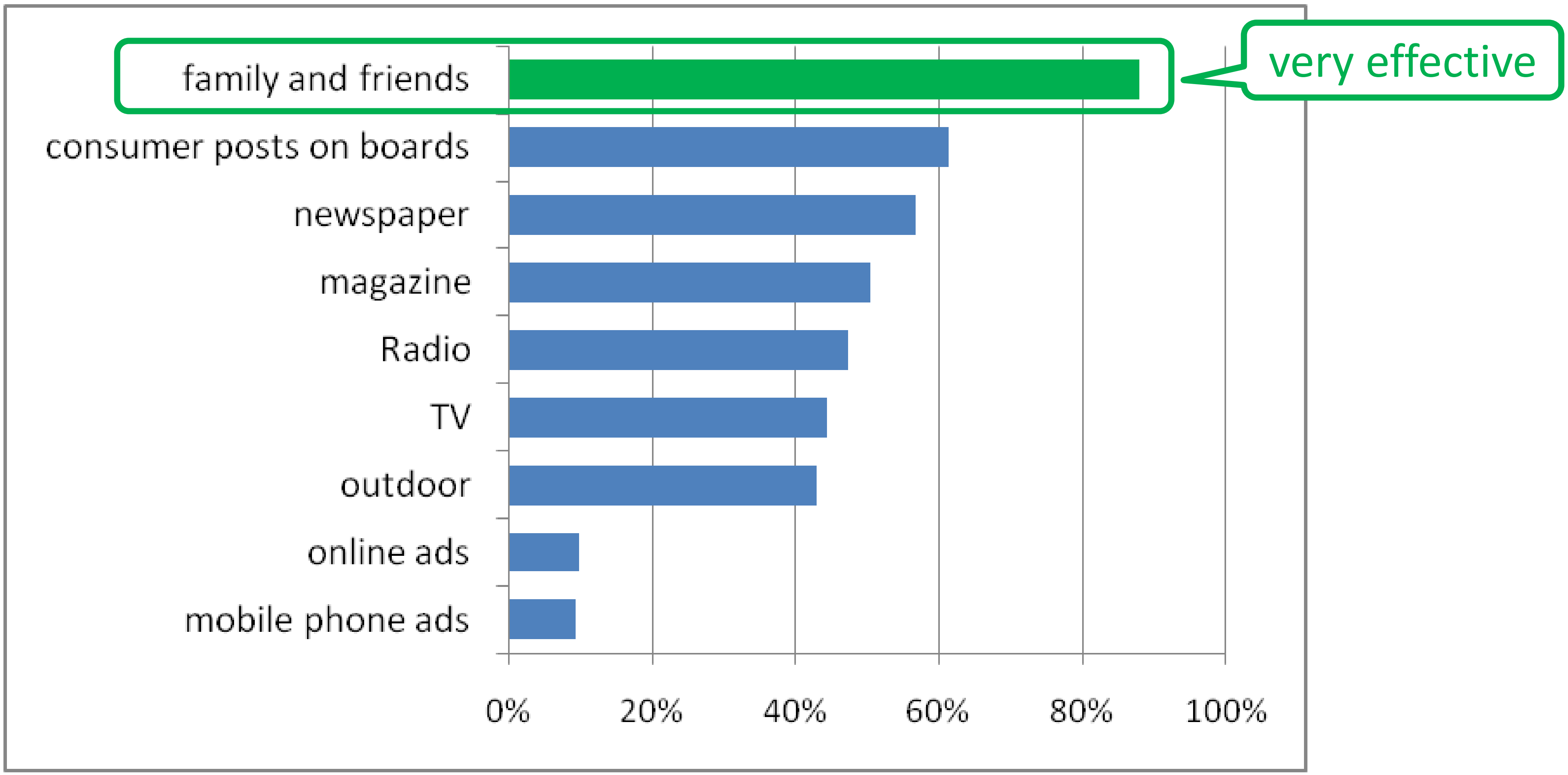


A Hypothetical Example of Viral Marketing



Effectiveness of Viral Marketing

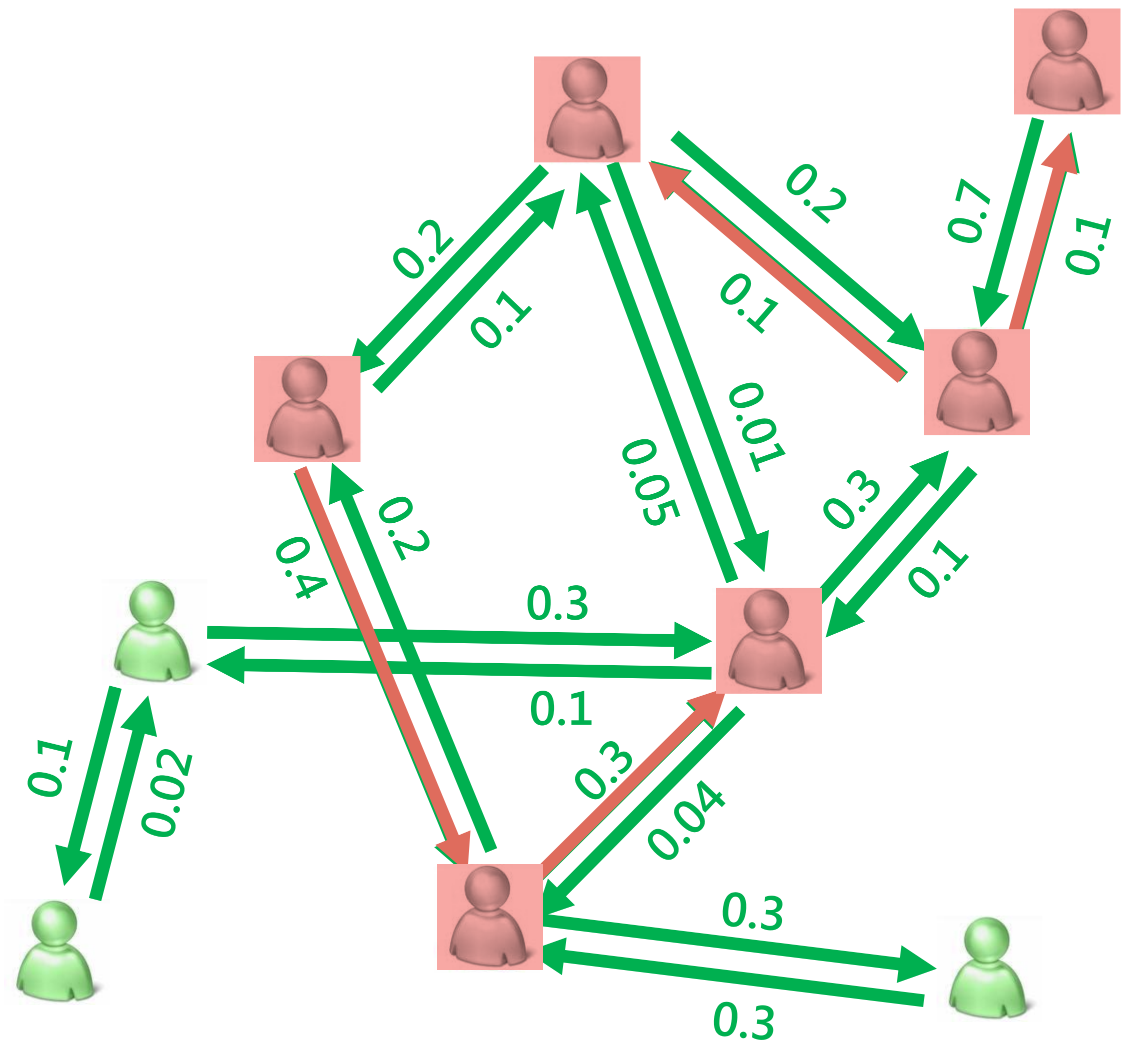
level of trust on different types of ads *



*source from Forrester Research and Intelliseek

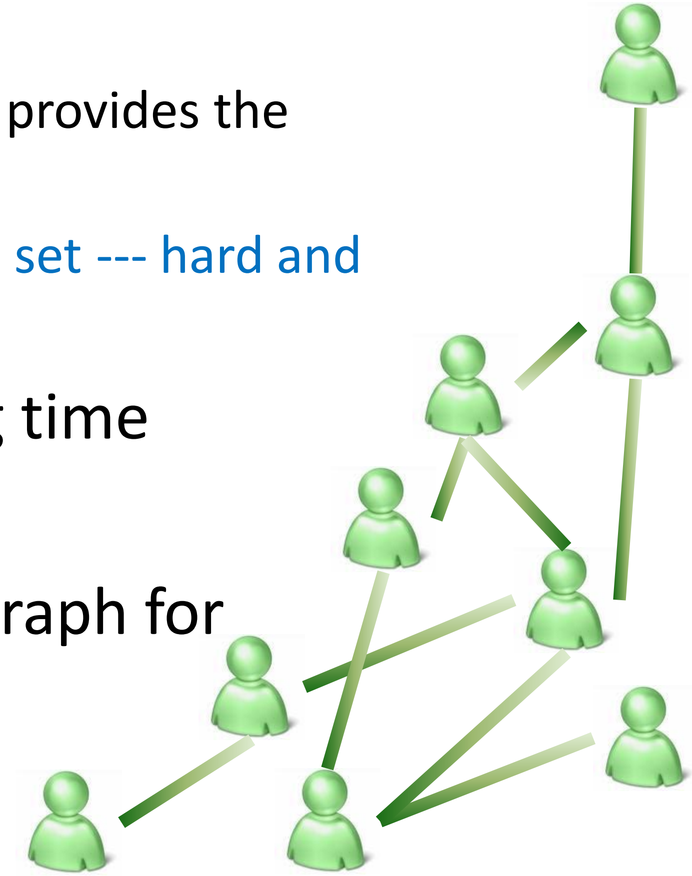
The Problem of Influence Maximization

- 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 spread*: expected number of nodes activated
- Influence maximization:
  - find *k* seeds that generate the largest influence spread



Research Background

- Influence maximization as a discrete optimization problem proposed by Kempe, Kleinberg, and Tardos, in KDD'2003
  - Finding optimal solution is provably hard (NP-hard)
  - Greedy approximation algorithm, 63% approximation of the optimal solution
    - Repeat k rounds: in the i-th round, select a node v that provides the largest marginal increase in influence spread
    - require the evaluation of influence spread given a seed set --- hard and slow
- Several subsequent studies improved the running time
- Serious drawback:
  - very slow, not scalable: > 3 hrs on a 30k node graph for 50 seeds

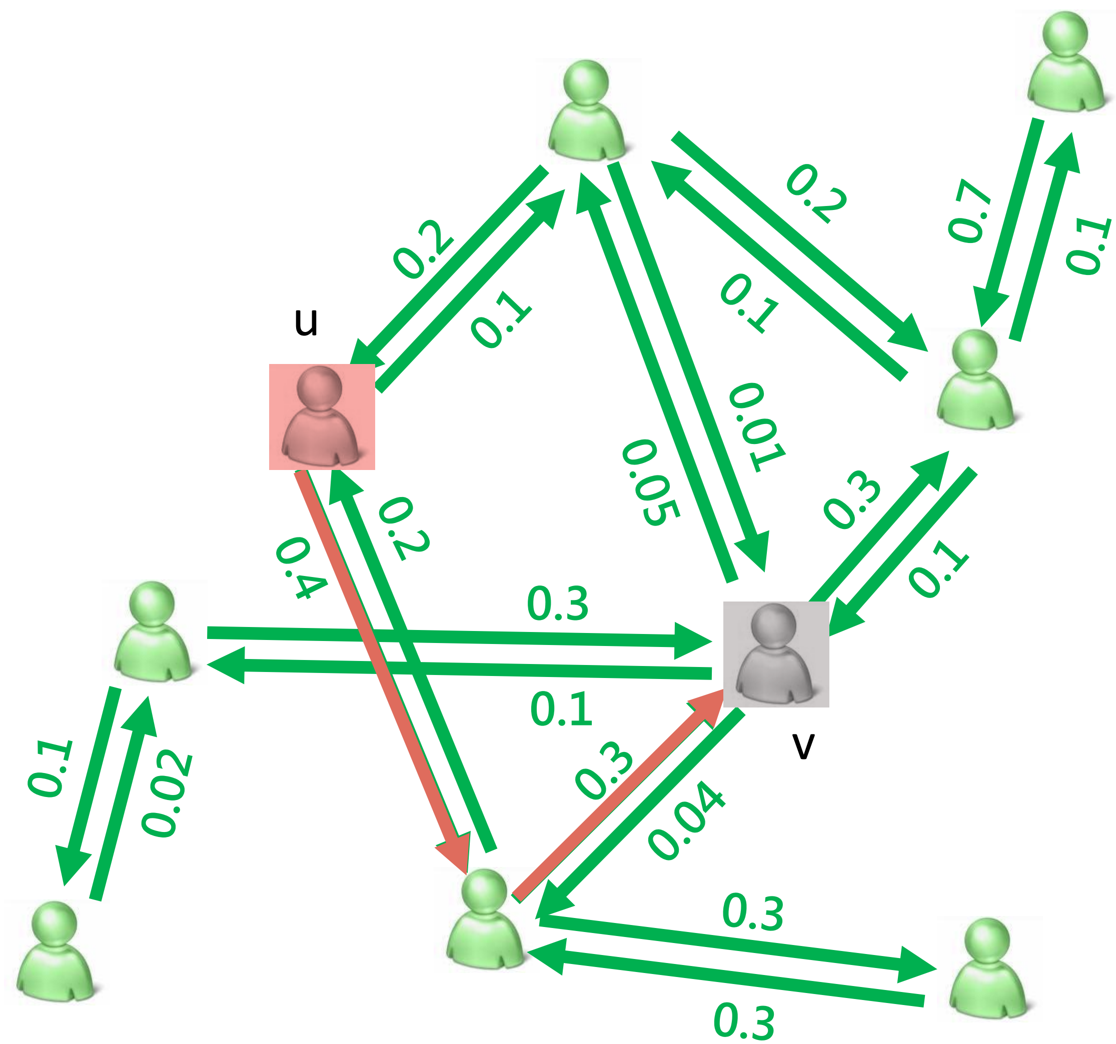


Our Work

- Design new heuristics
  - MIA (maximum influence arborescence) heuristic
    - for general independent cascade model
    - 10^3 speedup --- from hours to seconds (or days to minutes)
    - influence spread close to that of the greedy algorithm of [KKT'03]
- We also show that computing exact influence spread given a seed set is #P-hard (counting hardness)
  - resolve an open problem in [KKT'03]
  - indicate the intrinsic difficulty of computing influence spread

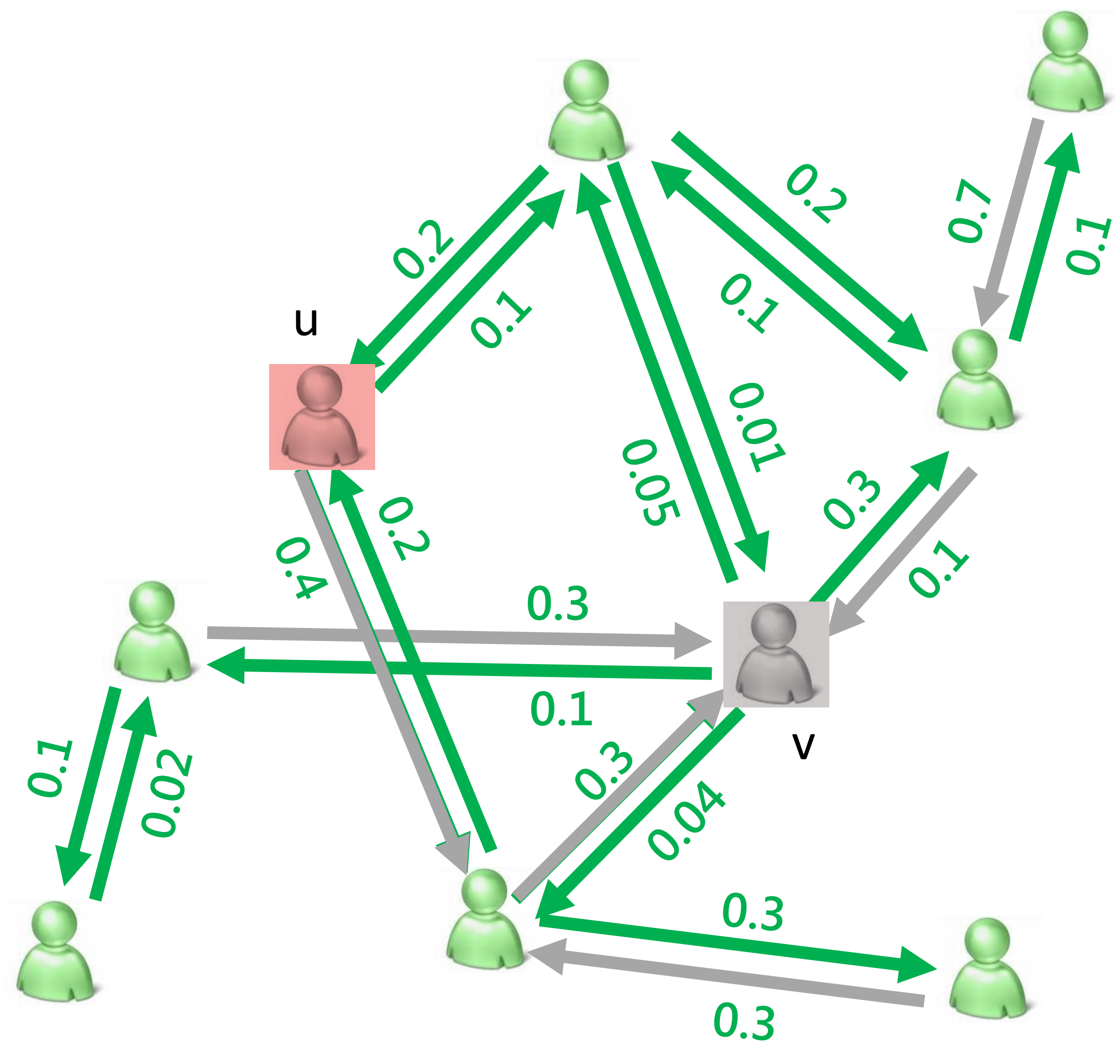
Maximum Influence Arborescence (MIA) Heuristic I: Maximum Influence Paths (MIPs)

- 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)



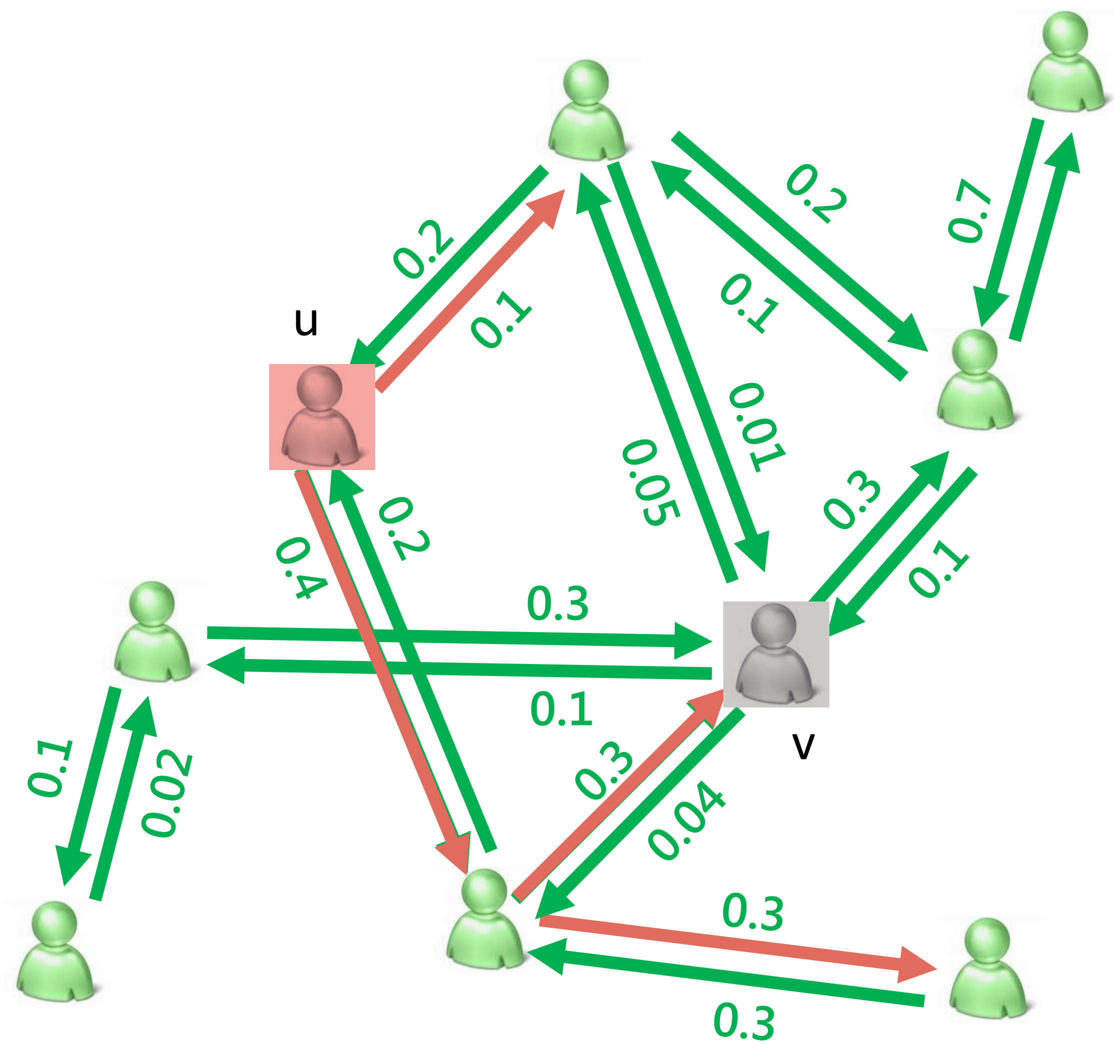
MIA Heuristic II: Maximum Influence in-(out-) Arborescences

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



MIA Heuristic II: Maximum Influence in-(out-) Arborescences

- 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)
  - These MIIAs and MIOAs can be computed efficiently using the Dijkstra shortest path algorithm



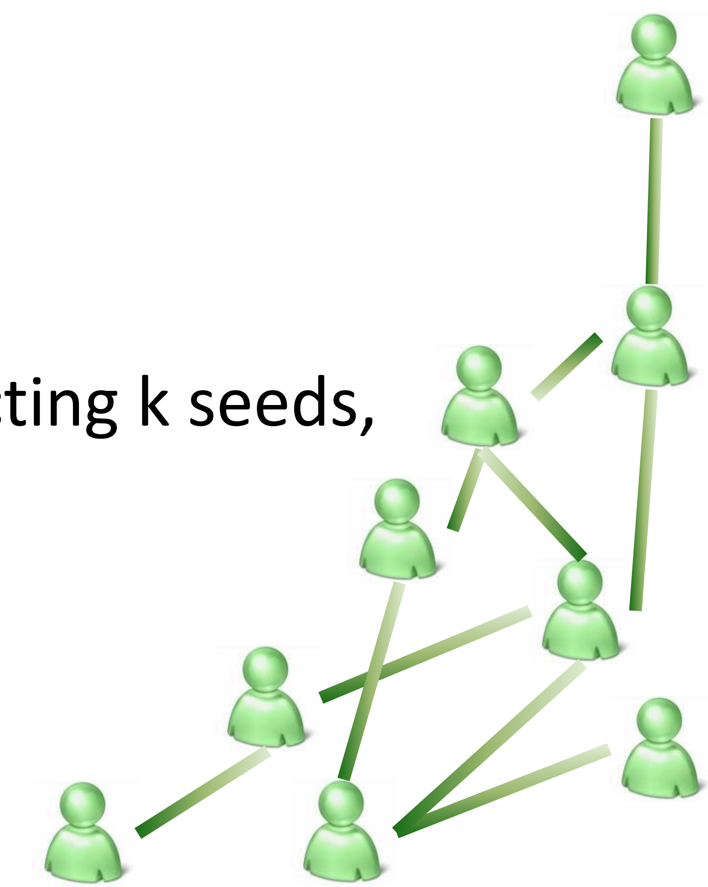
MIA Heuristic III: Computing Influence through the MIA structure

- 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

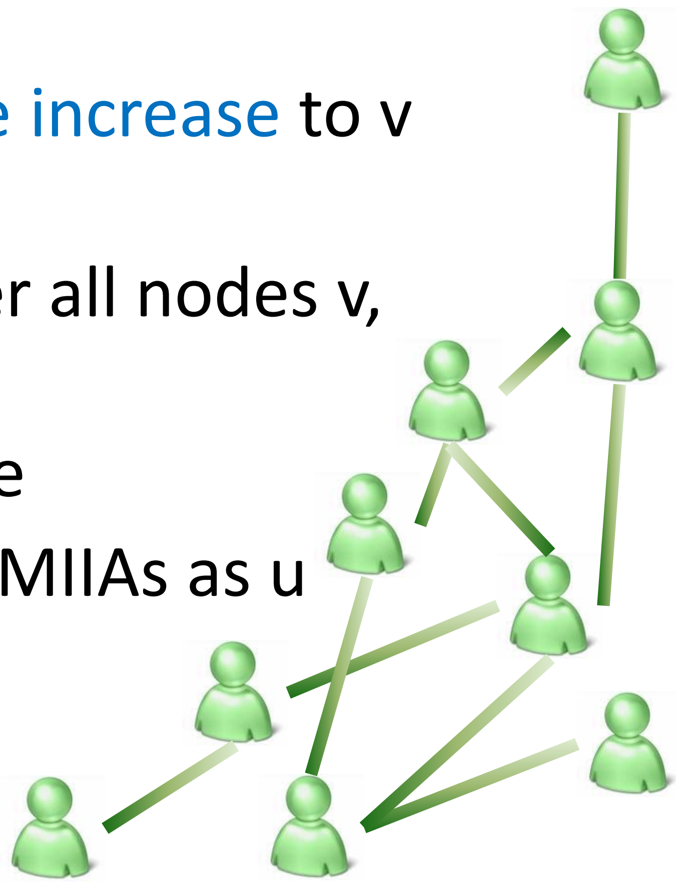


MIA Heuristic IV: Efficient Updates on Activation Probabilities

- If v is the root of a MIIA, and u is a node in the MIIA, then their activation probabilities have a linear relationship:

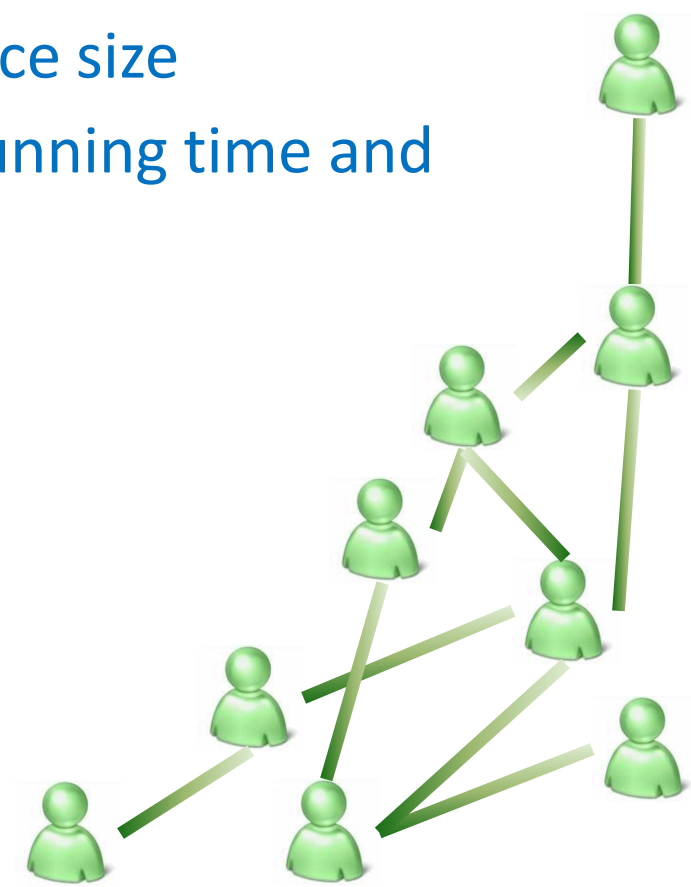
$$ap(v) = \alpha(v, u) \cdot ap(u) + \beta(v, u)$$

- All $\alpha(v, u)$'s in a MIIA can be recursively computed
  - time reduced from quadratic to linear time
- If u is selected as a seed, its marginal influence increase to v is $\alpha(v, u) \cdot (1 - ap(u))$
- Summing up the above marginal influence over all nodes v, we obtain the marginal influence of u
- Select the u with the largest marginal influence
- Update $\alpha(v, w)$ for all w's that are in the same MIIAs as u



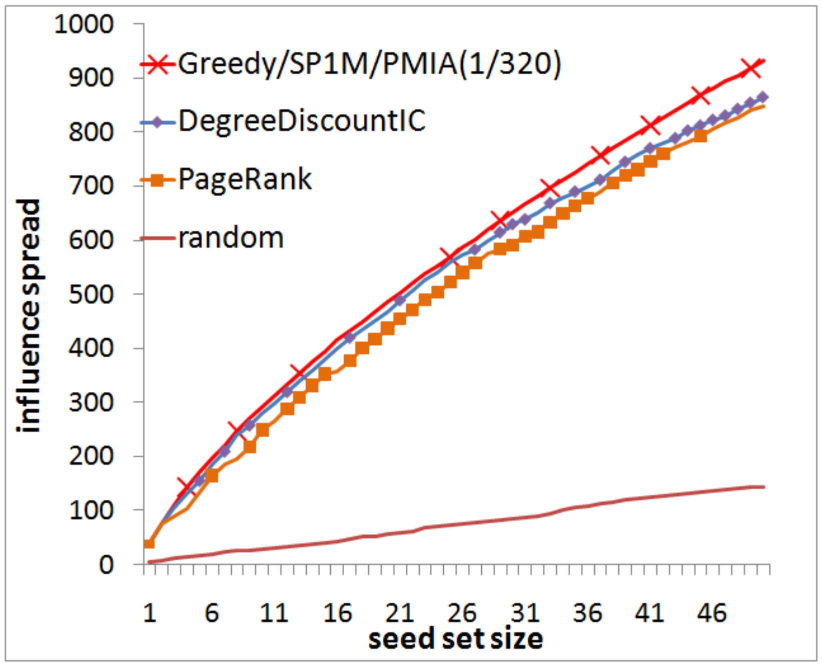
MIA Heuristic IV: Summary

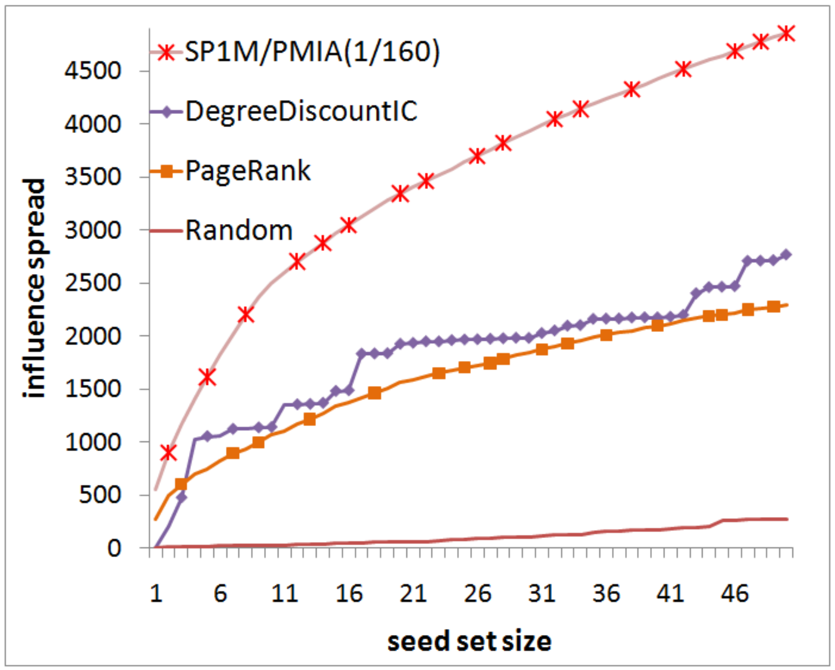
- Iterating the following two steps until finding k seeds
  - Selecting the node u giving the largest marginal influence
  - Update MIAs (linear coefficients) after selecting u as the seed
- Key features:
  - updates are local, and linear to the arborescence size
  - tunable with parameter θ: tradeoff between running time and influence spread



Experiment Results on MIA Heuristic

Influence spread vs. seed set size





NetHEPT dataset:

- collaboration network from physics archive
- 15K nodes, 31K edges

Epinions dataset:

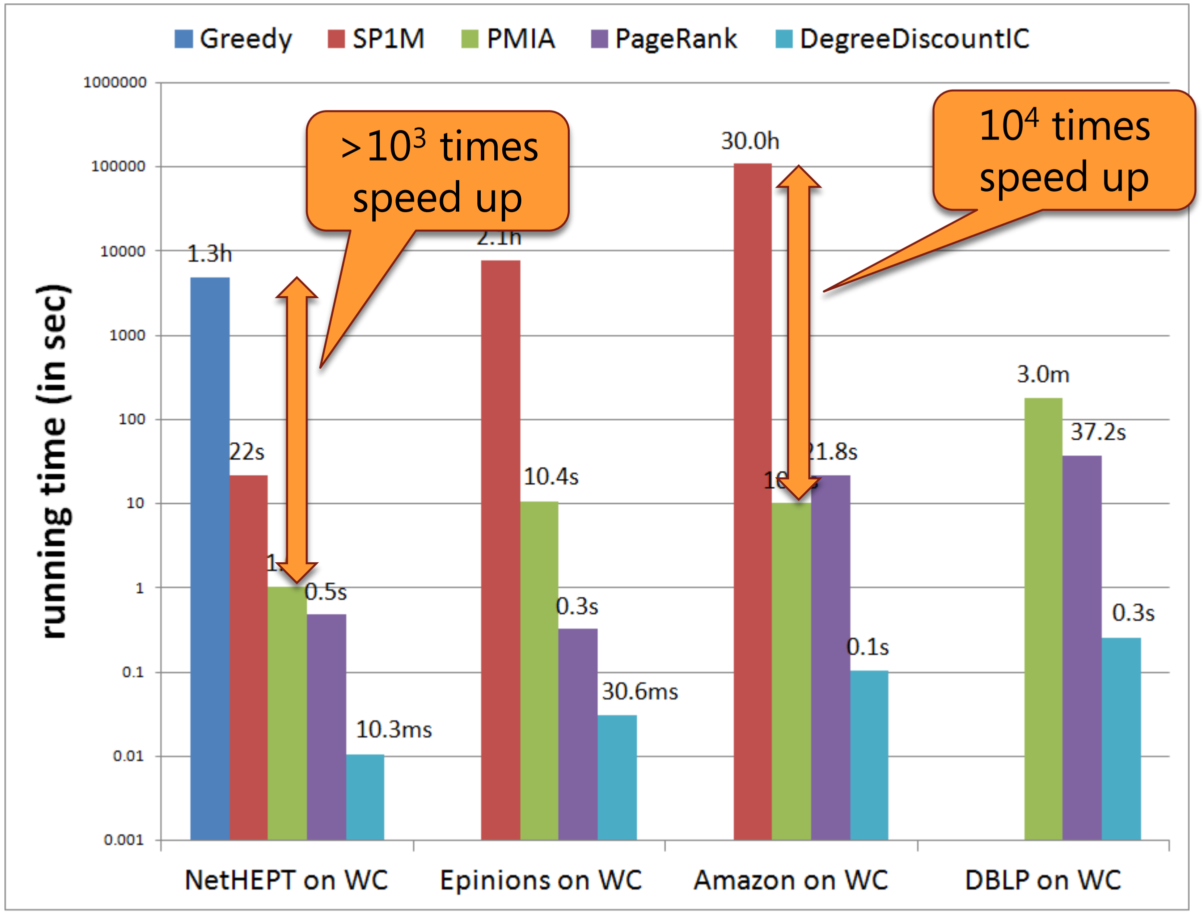
- who-trust-whom network of Epinions.com
- 76K nodes, 509K edges

weighted cascade model:

- influence probability to a node v = 1 / (# of in-neighbors of v)

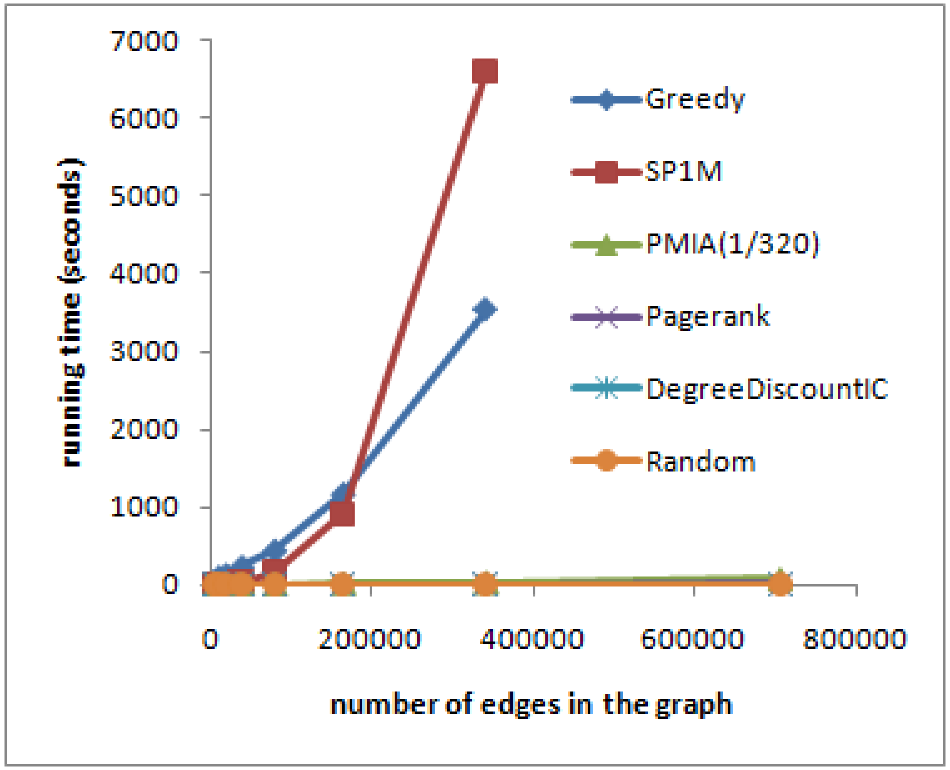
Experiment Results on MIA Heuristic

running time



Running time is for selecting 50 seeds

Scalability of MIA Heuristic



- synthesized graphs of different sizes generated from power-law graph model
- weighted cascade model
- running time is for selecting 50 seeds

Related Work

- Greedy approximation algorithms
  - Original greedy algorithm [Kempe, Kleinberg, and Tardos, 2003]
  - Lazy-forward optimization [Leskovec, Krause, Guestrin, Faloutsos, VanBriesen, and Glance, 2007]
  - Edge sampling and reachable sets [Kimura, Saito and Nakano, 2007; C., Wang, and Yang, 2009]
  - reduced seed selection from days to hours (with 30K nodes), but still not scalable
- Heuristic algorithms
  - SPM/SP1M based on shortest paths [Kimura and Saito, 2006], not scalable
  - SPIN based on Shapley values [Narayanam and Narahari, 2008], not scalable
  - Degree discounts [C., Wang, and Yang, 2009], designed for the uniform IC model
  - CGA based on community partitions [Wang, Cong, Song, and Xie 2010]
    - complementary
    - our local MIAs naturally adapt to the community structure, including overlapping communities

Future Directions

- Theoretical problem: efficient approximation algorithms:
  - How to efficiently approximate influence spread given a seed set?
- Practical problem: Influence analysis from online social media
  - How to mine the influence graph?

Thanks!

and

questions?