Finding Influential Nodes in a Social Network from Information Diffusion Data

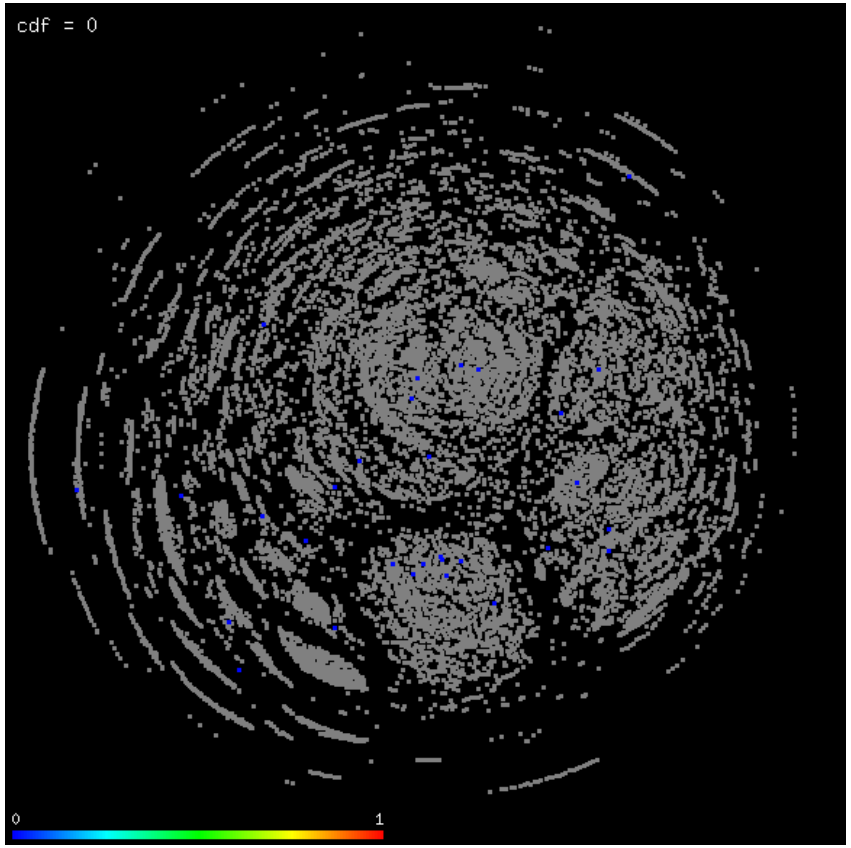
Masahiro Kimura (Ryukoku University)
Kazumi Saito (University of Shizuoka)
Ryohei Nakano (Chubu University)
Hiroshi Motoda (Osaka University)

Social Networks for Information Diffusion

- Innovation, hot topics and even malicious rumors can propagate through social networks (*word-of-mouth*).

- The rise of the Internet and the WWW accelerates the creation of various large-scale social networks.

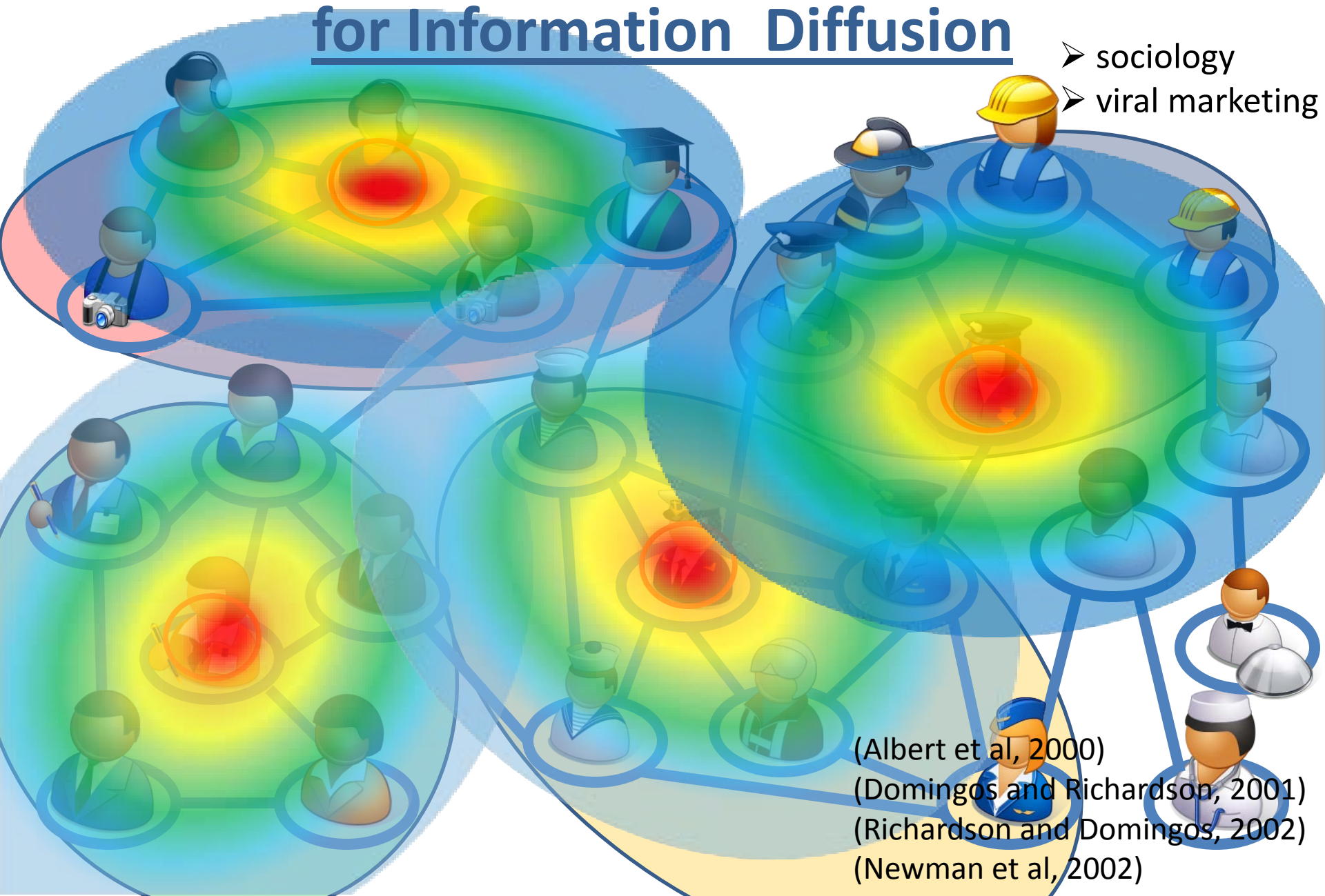
- Considerable attention has recently been devoted to social networks as an important medium for the spread of information.



e.g.,
(Gruhl et al, 2004)
(Adar and Adamic, 2005)
(Leskovec et al, 2006)

Finding Influential Nodes for Information Diffusion

- sociology
- viral marketing



(Albert et al, 2000)
(Domingos and Richardson, 2001)
(Richardson and Domingos, 2002)
(Newman et al, 2002)

Previous Work

A widely-used fundamental probabilistic model of information diffusion through a social network is the independent cascade (IC) model.

Note: The IC model can also be identified with the SIR model.

Using the IC model,

- the problem of finding a limited number of nodes that are effective for the spread of information has been extensively investigated (Kempe et al, 2003; Kimura et al ,2007),

- and further, yet another problem of minimizing the spread of undesirable information by blocking links has recently been addressed (Kimura et al, 2008).

Research Aim

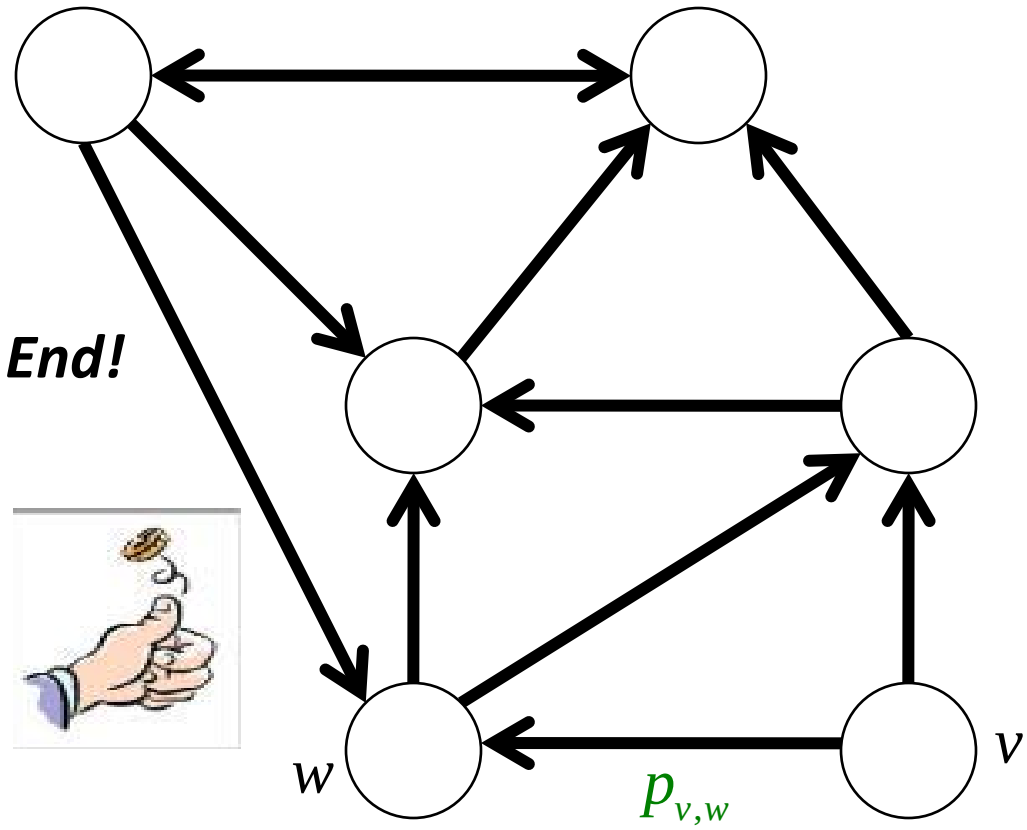
- Finding influential nodes in a social network is one of the most central problems in SNA.

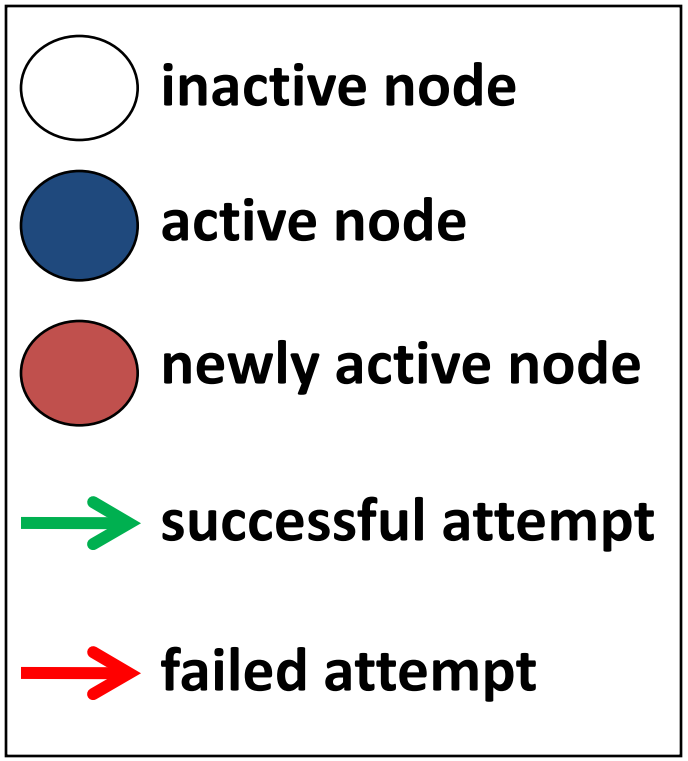
- There exist several methods for ranking nodes on the basis of the network structure.

We also address this problem, but from a different angle.

Propose a method for extracting influential nodes by ranking nodes in terms of *influence degrees* for the IC model based on the observed information diffusion data in the network.

IC Model





$\Theta = (p_{v,w})$: diffusion probabilities (parameters of the IC model)

For initial active node *v*, define the influence degree of node *v*, $\sigma(v;\Theta)$, as the expected number of active nodes at the end.

Proposed Method

$G = (V, E)$: a network (graph)

Assume that the IC model generates them.

Given:

$$\left\{ D_m = \left\langle D_m(0), \cdots, D_m(T_m) \right\rangle ; m = 1, \cdots, M \right\},$$

an observed data set of *M* independent information diffusion results (i.e., time sequences), where

$D_m(t)$: the set of nodes activated at time *t* in the *m*th information diffusion result

Estimate:

$\hat{\Theta} = \left(\hat{p}_{v,w} \right)$: the diffusion probabilities of the IC model

Extract influential nodes:

using the node-ranking based on influence degree $\sigma\left(v; \hat{\Theta}\right)$

Estimation Method (1/4)

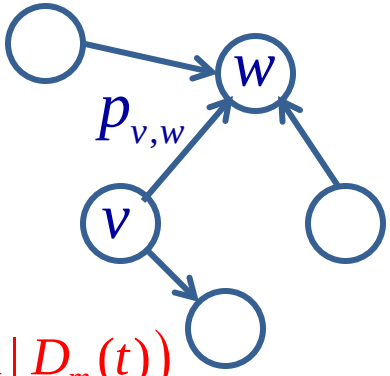
$D_m = \langle D_m(0), \cdots, D_m(T_m) \rangle$: the mth info. diffusion result

$\Rightarrow D_m(T_m + 1) = \varnothing$

$C_m(t) = D_m(0) \cup \cdots \cup D_m(t)$: the set of active nodes at time t

$F(v) = \{ w \in V; (v, w) \in E \}$: the forward set of v

$B(w) = \{ v \in V; (v, w) \in E \}$: the backward set of w



The likelihood L for D_m w.r.t. $\Theta = (p_{v,w})$:

$\mathrm{Prob}(D_m(t+1)$: the set of nodes activated at $t+1 | D_m(t))$

$$\mathrm{L}(\Theta; D_m) = \prod_{t=0}^{T_m - 1} \left(\prod_{w \in D_m(t+1)} \left(1 - \prod_{v \in B(w) \cap D_m(t)} (1 - p_{v,w}) \right) \right) \times \prod_{t=0}^{T_m} \left(\prod_{v \in D_m(t)} \prod_{w \in F(v) \setminus C_m(t+1)} (1 - p_{v,w}) \right)$$

succeeded

failed

$\mathrm{Prob}(V \setminus C_m(t+1)$: the inactive set at $t+1 | D_m(t))$

Estimation Method (2/4)

Objective function: $J(\Theta) = \sum_{m=1}^{M} \log L(\Theta; D_m)$

$$\log L(\Theta; D_m) = \sum_{w} \log\left(1 - \prod_{v \in B(w) \cap D_m(t)} (1 - p_{v,w})\right) + \sum_{v,w} \log(1 - p_{v,w})$$

$$\|$$

$$\log\left(\sum_{\mathbf{a}_w \neq \mathbf{0}} \prod_{v \in B(w) \cap D_m(t)} p_{v,w}^{a_{v,w}} (1 - p_{v,w})^{(1 - a_{v,w})}\right) \stackrel{\text{def}}{=} \log P_{m,t+1}(w; \Theta)$$

where

$\mathbf{a}_w = (a_{v,w});\ a_{v,w} = 1$ (succeeded), $a_{v,w} = 0$ (failed) <hidden variables>

To derive the EM algorithm, consider the posterior probability:

$$q_{m,t+1}(\mathbf{a}_w \mid w; \Theta) = \frac{\prod_{v \in B(w) \cap D_m(t)} p_{v,w}^{a_{v,w}} (1 - p_{v,w})^{(1 - a_{v,w})}}{P_{m,t+1}(w; \Theta)}$$

Estimation Method (3/4)

To derive the EM algorithm, construct *Q*-function:

$\Theta = (p_{v,w})$: new values, $\Theta' = (p'_{v,w})$: old values

$$\log P_{m,t+1}(w;\Theta) - \log P_{m,t+1}(w;\Theta')$$

$$\geq \sum_{\mathbf{a}_w \neq \mathbf{0}} q_{m,t+1}(\mathbf{a}_w \mid w;\Theta') \log \left(\prod_{v \in B(w) \cap D_m(t)} p_{v,w}^{\,a_{v,w}} (1-p_{v,w})^{(1-a_{v,w})} \right) + f(m,t,w,\Theta')$$

$$= \sum_{v \in B(w) \cap D_m(t)} \left(\frac{p'_{v,w}}{P_{m,t+1}(w;\Theta')} \log p_{v,w} + \left(1 - \frac{p'_{v,w}}{P_{m,t+1}(w;\Theta')} \right) \log(1-p_{v,w}) \right)$$

$$\stackrel{\text{def}}{=} Q_{m,t+1,w}(\Theta \mid \Theta')$$

Q-function:

$$Q(\Theta \mid \Theta') = \sum_{m,t,w} Q_{m,t+1,w}(\Theta \mid \Theta') + \sum_{m} \sum_{t} \sum_{v \in D_m(t)} \sum_{w \in F(v) \setminus C_m(t+1)} \log(1-p_{v,w})$$

Estimation Method (4/4)

Update formula:

$$p_{v,w} = \frac{1}{|\mathrm{M}^{+}_{v,w}| + |\mathrm{M}^{-}_{v,w}|} \sum_{m \in \mathrm{M}^{+}_{v,w}} \frac{p'_{v,w}}{P_{m,t(m,v,w)+1}(w;\Theta')}$$

where

$$\mathrm{M}^{+}_{v,w} = \left\{ m \in \{1,\cdots,M\};\ \exists t \text{ s.t. } w \in D_m(t+1), v \in B(w) \cap D_m(t) \right\}$$

(the set of info. diffusion results s.t. the activation attempts through (*v*, *w*) might succeed)

$$\mathrm{M}^{-}_{v,w} = \left\{ m \in \{1,\cdots,M\};\ \exists t \text{ s.t. } v \in D_m(t), w \in F(v) \setminus C_m(t+1) \right\}$$

(the set of info. diffusion results s.t. the activation attempts through (*v*, *w*) definitely failed)

Note:

$$P_{m,t+1}(w;\Theta') = 1 - \prod_{u \in B(w) \cap D_m(t)} (1 - p'_{u,w})$$

Experimental Settings (1/2)

Network data:

Employed two sets of large real (bidirectional) networks, which exhibit many of the key features of (bidirectional) social networks (Kimura et al, 2008):

- Blog network (12,047 nodes, 79,920 links)
- Wiki network (9,481 nodes, 245,044 links)

Diffusion probabilities:

Assumed the simplest case where the diffusion probability is uniform, $p_{v,w} = p, \forall (v,w) \in E$, and set the value of p as

p = 0.1 (Blog network) and p = 0.01 (Wiki network).

<ground truth>

Experimental Settings (2/2)

Observed data:

In the learning stage, a training sample was an information diffusion result,

$$D_m = \langle D_m(0), \cdots, D_m(T_m) \rangle, \quad (m \in \{1, \cdots, M\}),$$

which is a sequence of the activated nodes starting from a randomly selected initial active node $D_m(0) = \{v_m\}$.

Used M training samples, where M is a parameter.

Influence degrees:

Evaluated the influence degrees $\{\sigma(v; p); v \in V\}$ using the bond percolation method (Kimura et al, 2007).

Experimental Evaluation

$G = (V, E)$: a network

p_0 : the true value of diffusion probability

$L_0(r)$: the true set of top r nodes w.r.t. the influence degrees, $\{\sigma(v; p_0); v \in V\}$

$\hat{p}$: the value of diffusion probability estimated by the proposed method

$\hat{L}(r)$: the set of top r nodes w.r.t. the influence degrees, $\{\sigma(v; \hat{p}); v \in V\}$

Compare $\hat{L}(r)$ with $L_0(r)$ for high ranks r.

Learning Performance for Diffusion Probability

Evaluated the learning performance by the error rate $Er = \left| p_0 - \hat{p} \right| / p_0$ as a function of the number of training samples M:

M	E
20	0.036 (0.024)
40	0.018 (0.014)
60	0.016 (0.007)
80	0.009 (0.006)
100	0.006 (0.004)

Blog network

M	E
20	0.138 (0.081)
40	0.109 (0.066)
60	0.080 (0.041)
80	0.047 (0.018)
100	0.021 (0.013)

Wiki network

Show the average values (the standard deviations) for five experimental results.

Comparison Methods

Compared our method with four heuristics from SNA w.r.t. the predictive capability of high ranked influential nodes.

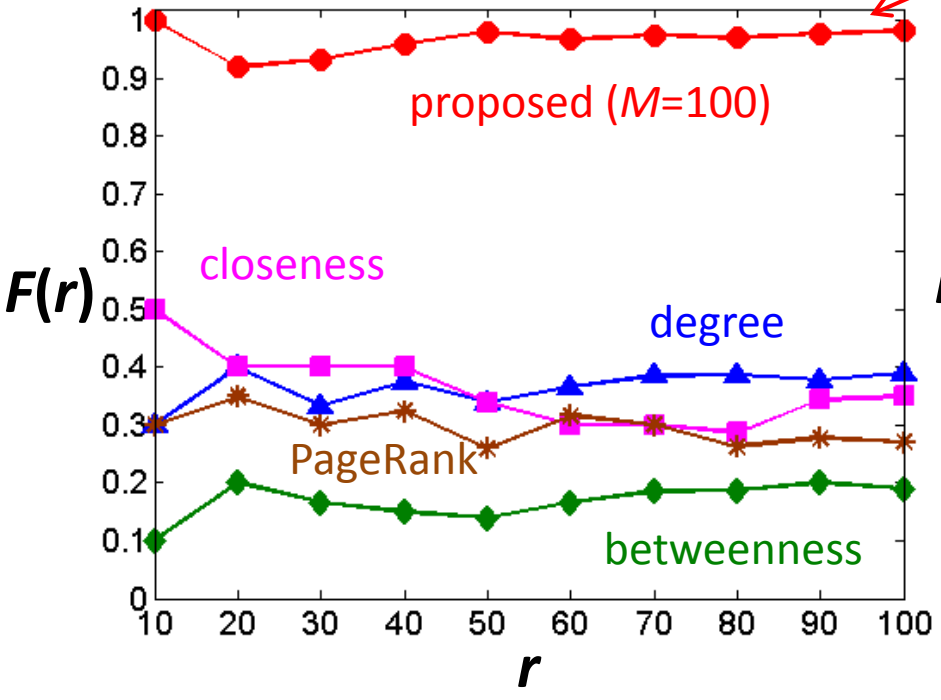
- Degree centrality;
  degree of *v*: the number of links attached to *v*
- Closeness centrality;
  closeness of *v*: the reciprocal of the average distance between *v* and other nodes
- Betweenness centrality;
  betweenness of *v*: the total number of shortest paths between pairs of nodes that pass through *v*
- PageRank (authoritativeness)

$L(r)$: the set of top *r* nodes for a given ranking method

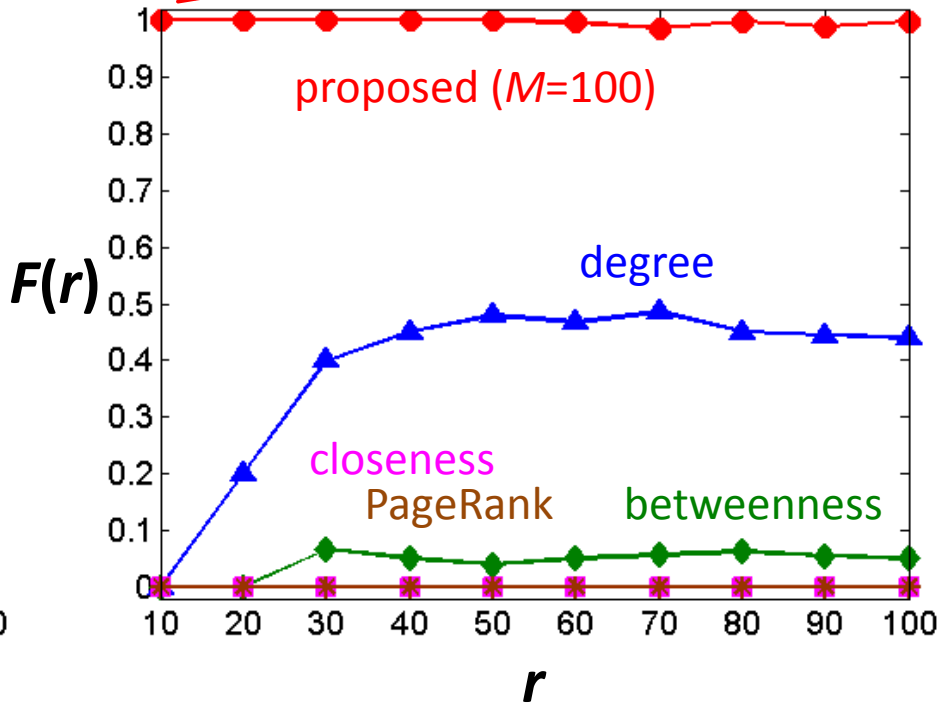
Experimental Results

Evaluated the performance of the ranking method by the ranking similarity $F(r) = |L_0(r) \cap L(r)| / r$ at high rank *r*.

Plotted the average values for five experimental results.



Blog network



Wiki network

Discussion (1/3)

- Experimental results show that nodes identified as higher ranked by our method are substantially different from those by each of the conventional methods.

- This means that our method enables a new type of SNA if past information diffusion data are available.

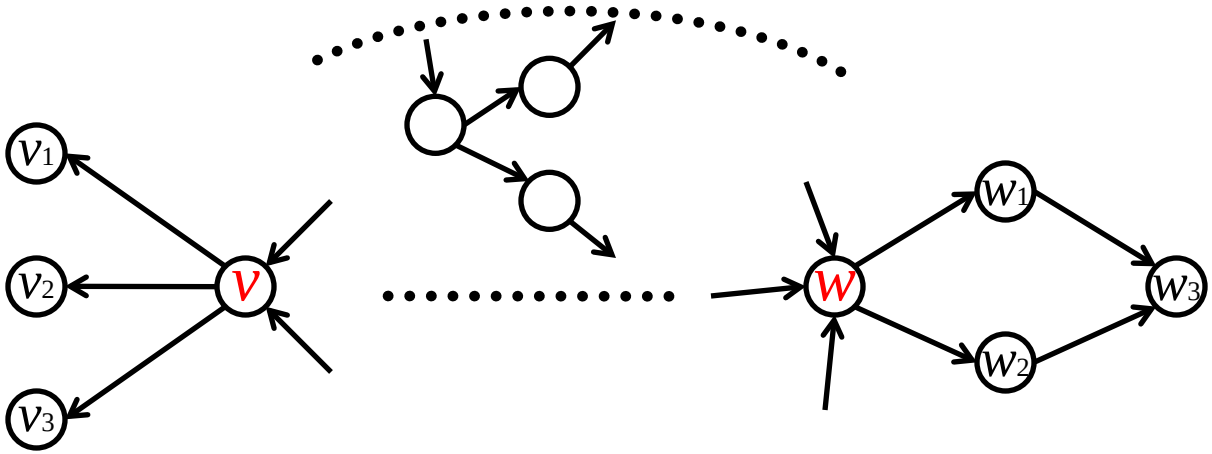
- Of course, it is beyond controversy that each conventional method has its own merit and usage, and our method is an addition to them which has a different merit in terms of information diffusion.

Discussion (2/3)

It is important to estimate the diffusion probability as accurately as possible in finding the influential nodes, since the probability affects the ranking:

Example:

For this graph, we have:



$$\sigma(v;p) = 3p$$

$$\sigma(w;p) = 2p + \left(1-(1-p^2)^2\right)$$

$$\Rightarrow \quad \sigma(v;p) - \sigma(w;p) = p(1-p)(1-p-p^2)$$

Hence,

$$\sigma(v;p) > \sigma(w;p) \quad \text{if} \quad p < (\sqrt{5}-1)/2$$

$$\sigma(v;p) \le \sigma(w;p) \quad \text{if} \quad p \ge (\sqrt{5}-1)/2$$

Discussion (3/3)

- The analysis we showed here is the simplest case where p takes a single value for all the links in E.

- In a more realistic setting, we can divide E into subsets E_1, …, E_n and assign a different value p_n for all the links in each E_n.

- If there is some background knowledge about the node grouping, our method can make the best use of it, one of the characteristic of the artificial intelligence approach.

Conclusion

- Proposed a method of ranking influential nodes in social networks by estimating diffusion probabilities from observed information diffusion data using the popular IC model.

- Applied this to two real networks in the simplest setting where the diffusion probability is uniform for all the links.

- Showed that the proposed method can estimate the diffusion probability accurately.

- Further showed that the proposed method can predict the high ranked influential nodes much more accurately than the well studied conventional four heuristic methods.