



Diffusion of “Following” Links in Microblogging Networks

Jing Zhang

Tsinghua University

Collaborate with

Wei Chen (*MSRA*)

Zhanpeng Fang and Jie Tang (*THU*)

What is Social Influence?

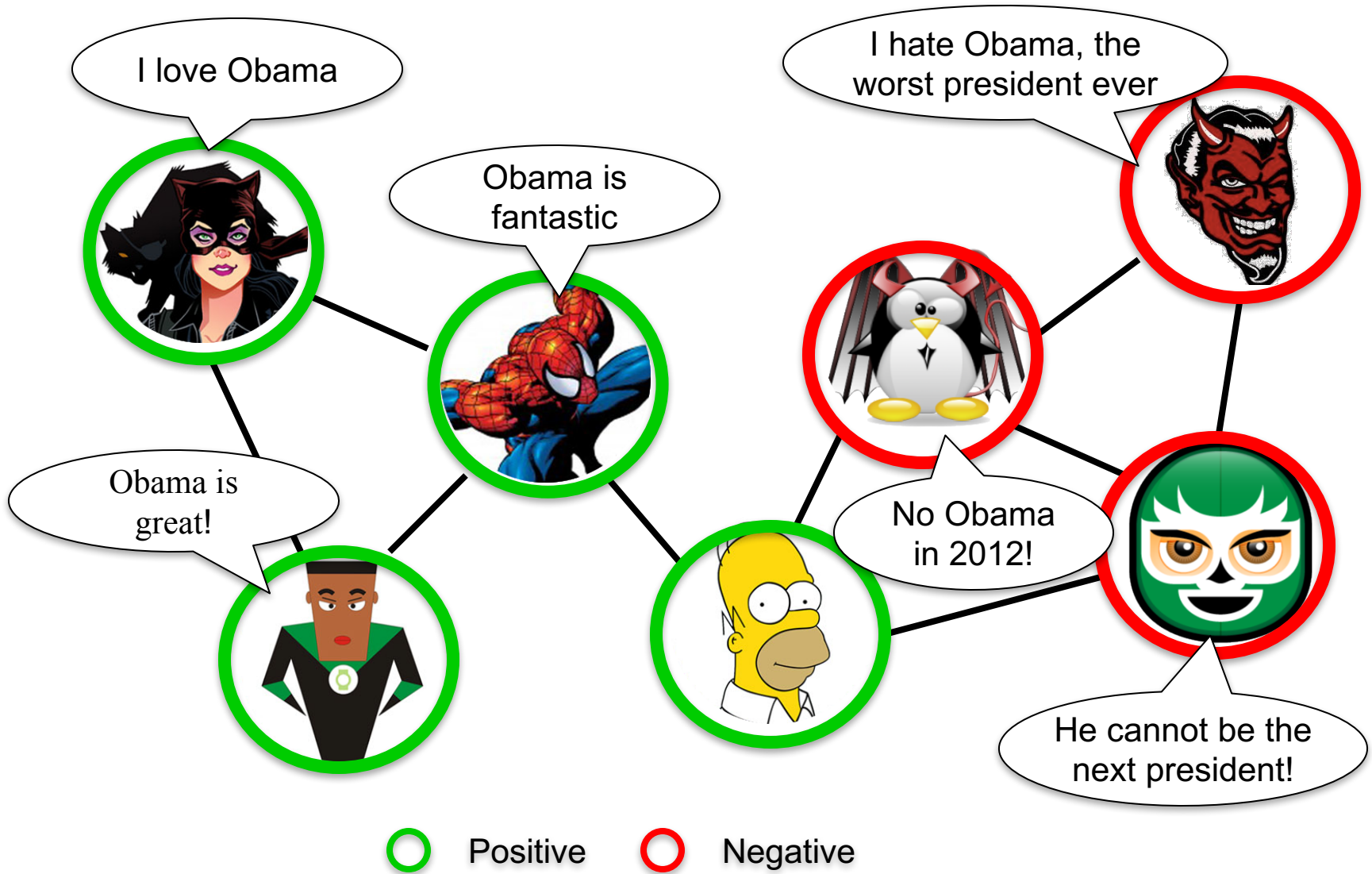
- Social influence occurs when one's **opinions**, **emotions**, or **behaviors** are affected by others, intentionally or unintentionally.^[1]

- Peer Pressure
- Opinion leadership
- Group Influence
- ...



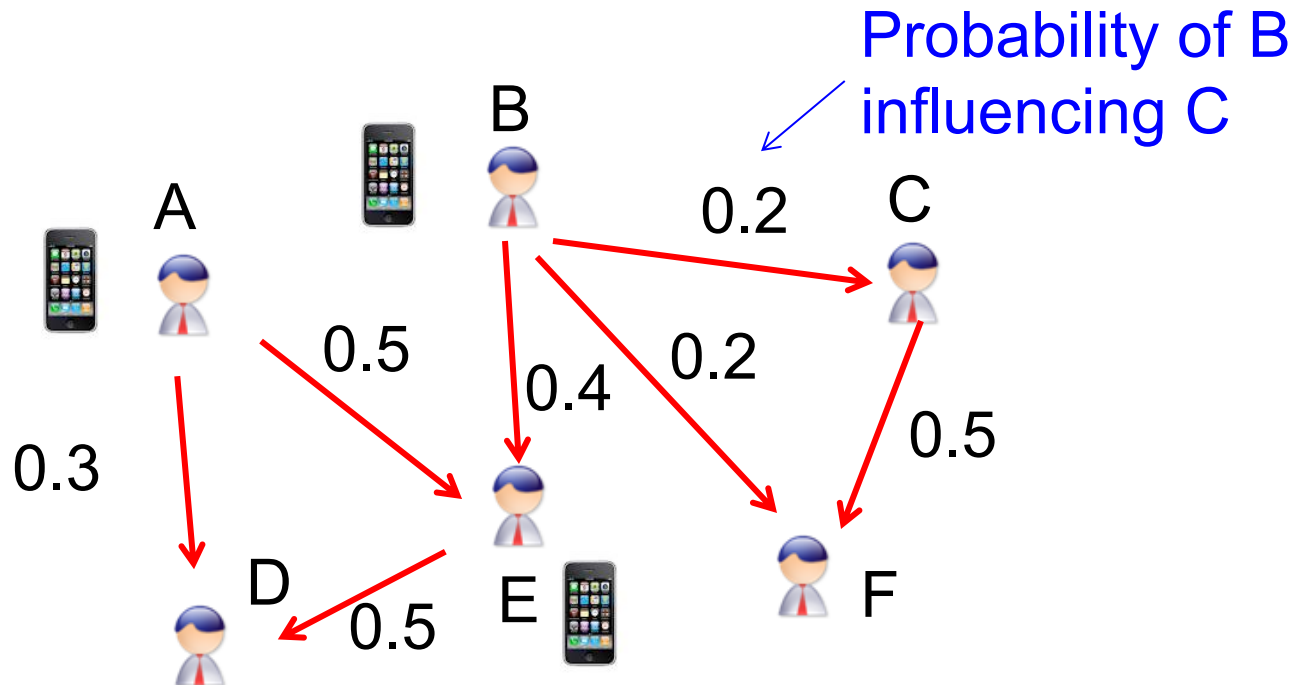
[1] http://en.wikipedia.org/wiki/Social_influence

“Love Obama”

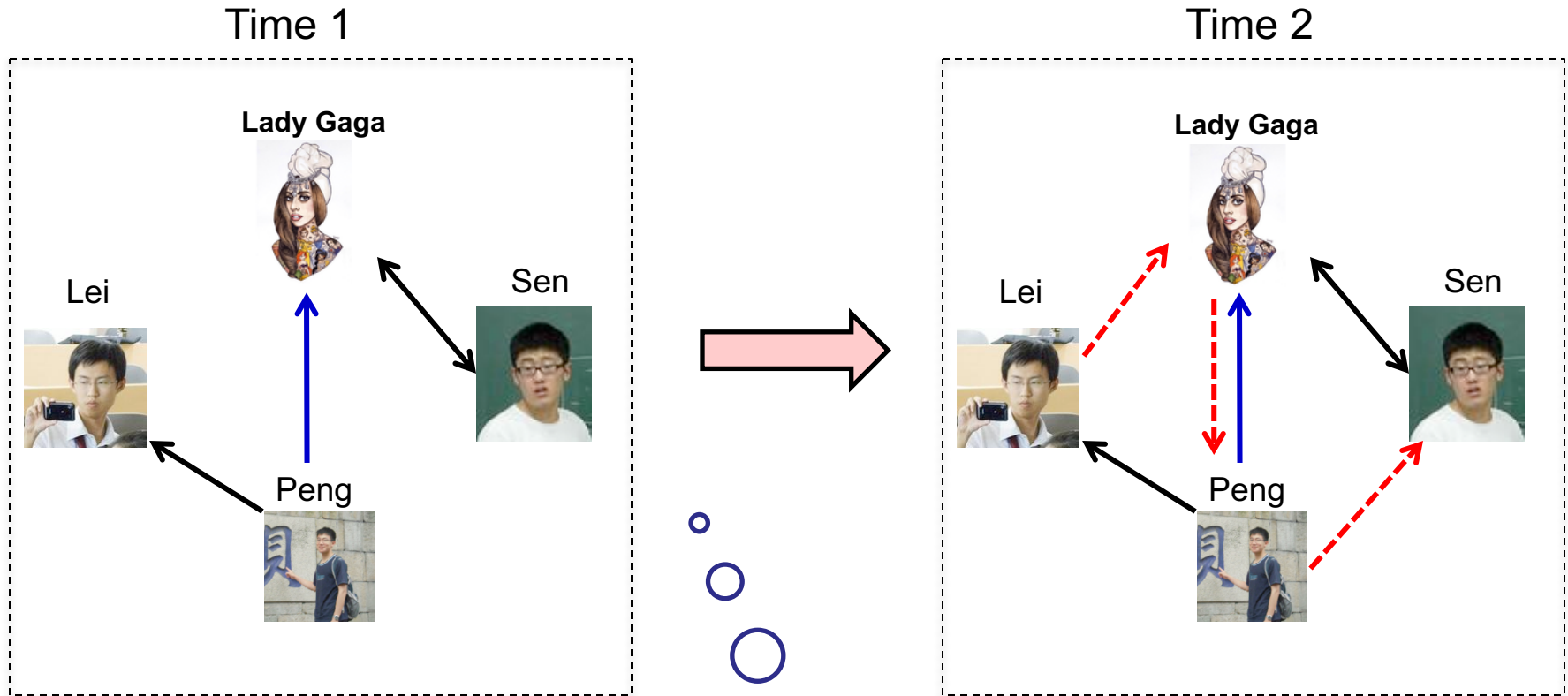


Influence Maximization

- Initially targeting a few “influential” seeds, to trigger a maximal number of individuals to adopt the opinions/products through friend recommendation.



Following Influence on Twitter

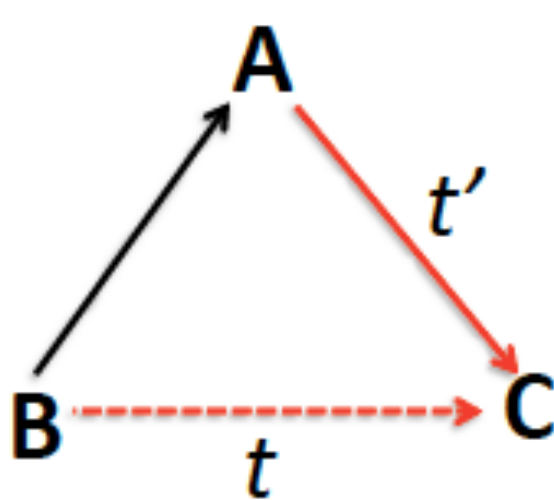


When you **follow** a user in a social network, will the behavior **influences** your friends to also follow her?

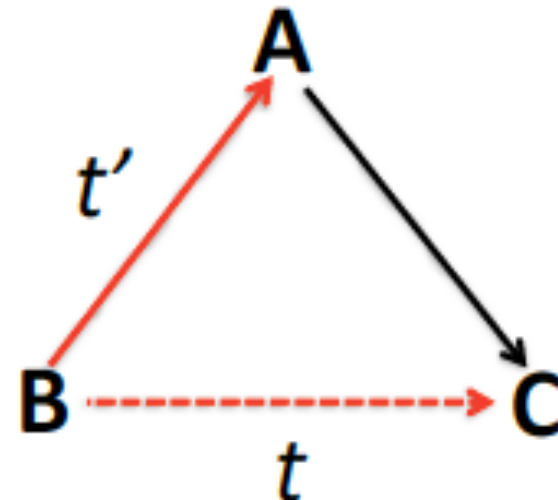
Link Influence



Two Categories of Link Influence



(a) Follower diffusion



(b) Followee diffusion

\rightarrow : pre-existing relationships

$\color{red}\rightarrow$: a new link added at time t'

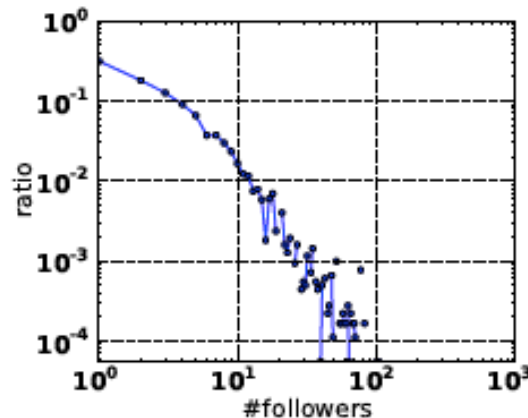
$\color{red}\dashrightarrow$: a possible link added at time t

$$0 \leq t - t' \leq \delta$$

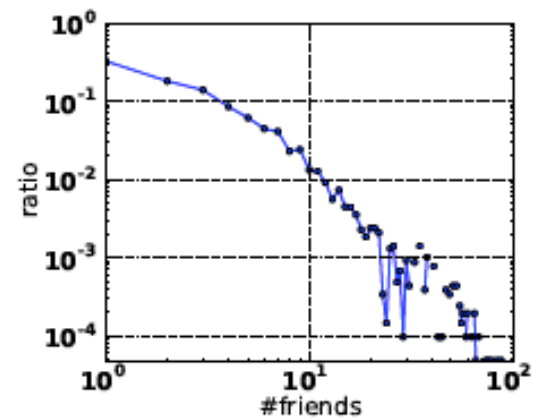
Twitter Data



- Twitter data
 - “Lady Gaga” -> 10K followers -> millions of followers;
 - 13,442,659 users and 56,893,234 following links.
- A **complete dynamic** network
 - 112,044 users and 468,238 follows
 - From 10/12/2010 to 12/23/2010, 13 timestamps by viewing every 4 days as a timestamp



(a) Follower distribution



(b) Followee distribution

Randomization Test

- Randomization test is a model-free, computationally intensive statistical technique for hypothesis testing, the main steps are
 1. Compute some test statistic using the set of original observations;
 2. Carry out the random shuffle according to the null hypothesis a large number of times, and compute the test statistic for each random data;
 3. By the law of large numbers, the permutation p-value is approximated by the proportion of randomly generated values that exceed or equal the observed value of the test statistic.
- **Null hypothesis:** the formation of neighboring links is temporally independent of one another.
- **Test statistic:**

$$\text{rate} = \frac{|\{\text{triad}(A,B,C) | 0 \leq t_{BC} - t_{AC} \leq \delta\}|}{|\{\text{triad}(A,B,C)\}|}$$

P-values on 24 Triads

The link e_{AC} is formed most probably due to the “following” behavior from ordinary user to celebrity user.

Follower Diffusion				Followee Diffusion					
	Δ	$ C_{\Delta} $	$ C_{\Delta}^+ $	r_{Δ}		Δ	$ C_{\Delta} $	$ C_{\Delta}^+ $	r_{Δ}
1		22870	233	0.0102 ***	3		24162	2298	0.0951 ***
2		22527	246	0.0109 **	4		62411	2293	0.0367 ***
3		33122	642	0.0194 ***	5		63092	3985	0.0632 ***
4		29830	100	0.0034	6		23099	2314	0.1002 ***
5		2370	3	0.0013	7		25049	324	0.0129 ***
6		7283	76	0.0104 *	8		65219	3469	0.0532 ***

The most probable reason why A follows C is “following” back, and thus C is more likely to be an ordinary user.

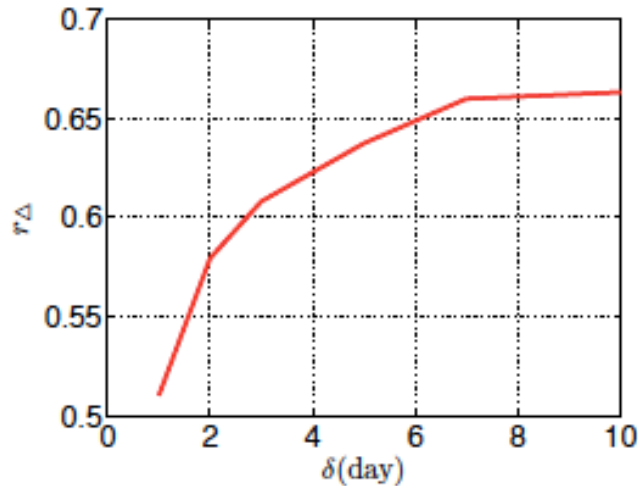
The most probable reason of B “following” C is C “following” B before and B “following” back, rather than the influence from A “following” C.

Follower Diffusion				Followee Diffusion					
	Δ	$ C_{\Delta} $	$ C_{\Delta}^+ $	r_{Δ}		Δ	$ C_{\Delta} $	$ C_{\Delta}^+ $	r_{Δ}
7		116	3	0.0259	19		428	315	0.7360 ***
8		883	77	0.0872	20		5729	2300	0.4015 ***
9		730	71	0.0973	21		4372	3427	0.7839 ***
10		666	46	0.0691 **	22		3889	3267	0.8401 ***
11		389	42	0.1080 ***	23		8145	3280	0.4027 ***
12		970	180	0.1856 **	24		27076	23310	0.8609 ***

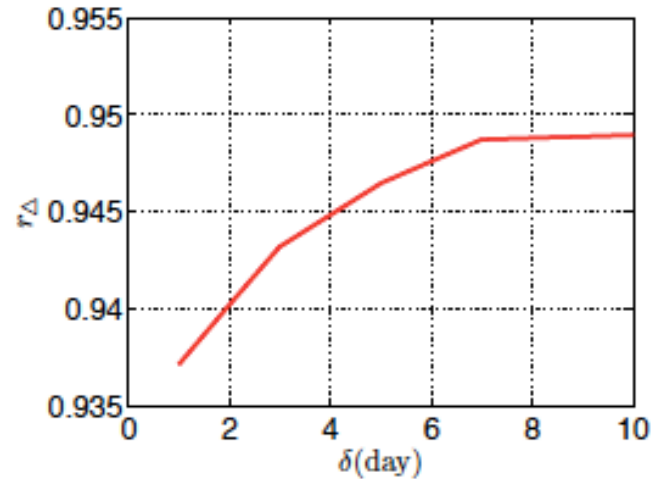
Notes: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001

There are more two-way links in a triadic closure, which can strengthen the diffusion effect from e_{AC} .

Diffusion Decay



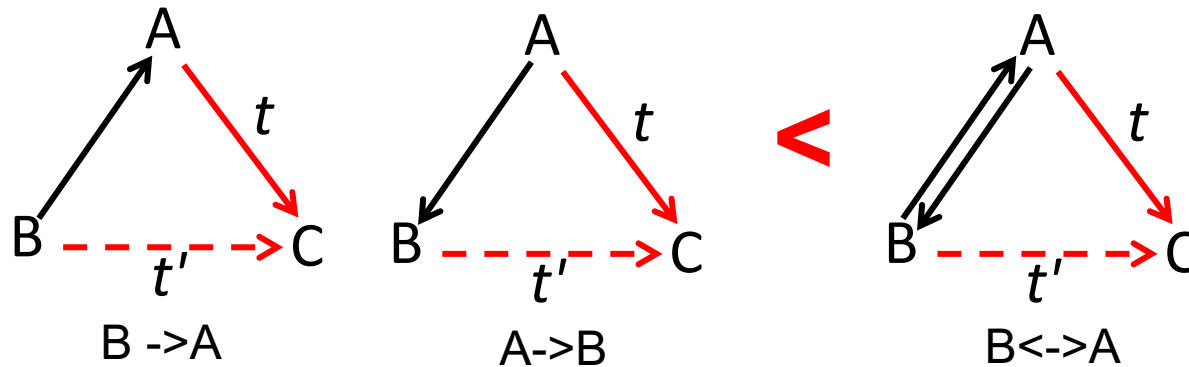
(a) Follower diffusion



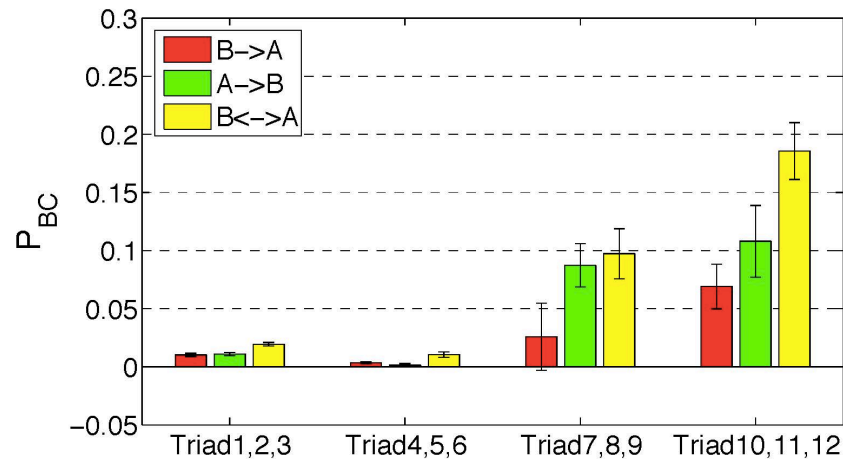
(b) Followee diffusion

- The increasing rate becomes slower over time.
- When δ is larger than 7 days, the rate almost stops increasing.
- The formation of B following C in followee diffusion is easier than that in follower diffusion.

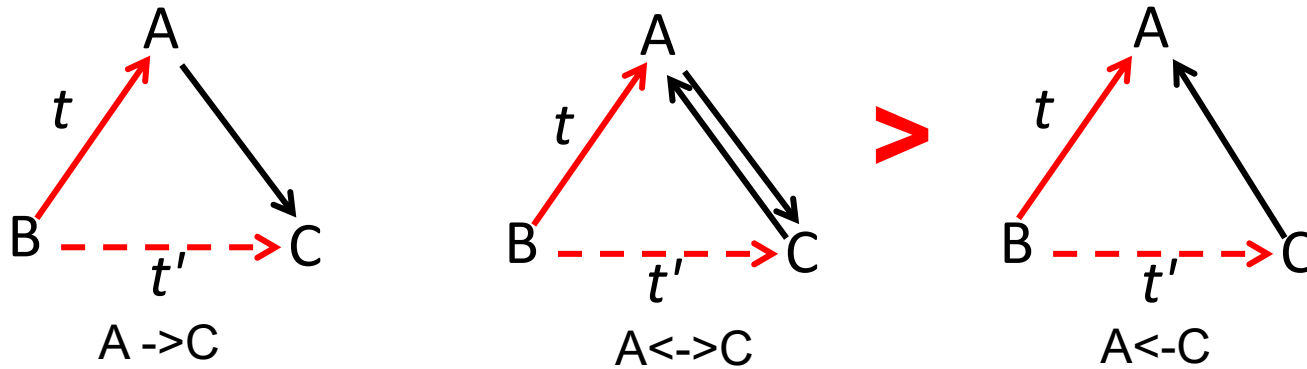
Follower Diffusion: Power of Reciprocity



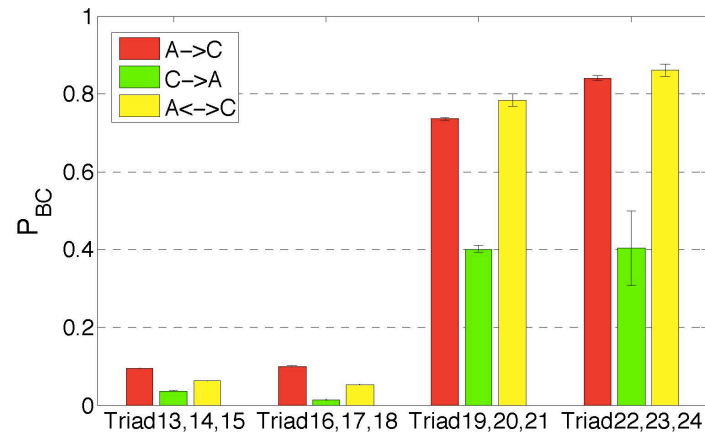
Observation: Reciprocal relationships are much more likely to be actual “social” relationships, rather than “celebrity following”, and thus have stronger social influence.



Followee Diffusion: Easy Discovery

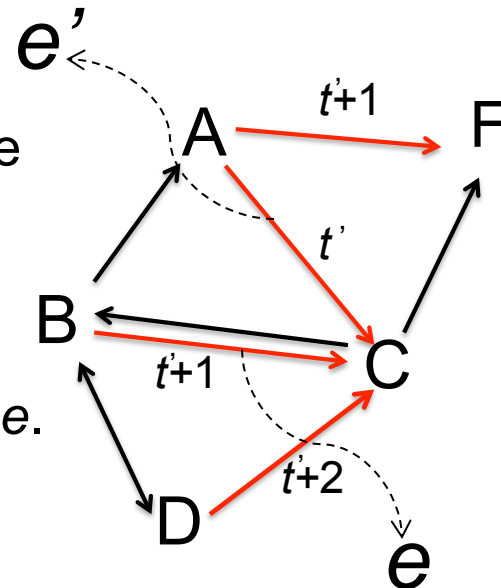


Observation: When a user B follows another user A, who already follows user C, B is likely to discover C through browsing A's retweets of C's messages or directly checking A's followee list, and A's interest in C may indicate that B would also be interested in C.




“Following” Link Cascade Model

- When a link e' is added at time t' , at each time slot from time t' to $t'+\delta$:
 - The follower end point B of link e may discover the link e' with **discovery probability** $g_{e'e}$.
 - Once discovered, e' may trigger e to be formed with **influence probability** $h_{e'e}$.
 - If failed, e' will have no chance to activate e again.
 - When multiple links activate e , e is activated at the time of the first successful attempt.
- The time delay λ for discovery follows a geometric distribution with parameter $g_{e'e}$ and after discovery there is one chance at time $t'+\lambda$ that e' could activate e .



Influence Estimation

- The object is to estimate $h_{e'e}$ and $g_{e'e}$.
- The method is to maximize the likelihood of generating all the links and solve the parameters in the likelihood function.

$$\mathcal{L} = \prod_{e \in \mathcal{E}} \left\{ p(e|S_e) \prod_{e' \in R_e} y_{ee'} \right\}.$$


We formalize the formation of each newly added link.

For each newly added link, we also formalize its effect on its unformed neighboring links.

Log-likelihood

- A link e is successfully added if at least one of its recently added neighboring links $e' \in S_e$ successfully activated it.
- Use a latent binary vector $\alpha_{S_e} = \{\alpha_{e'}\}_{e' \in S_e}$ to represent the statuses of S_e .
 - $\alpha_{e'}=1$: e' tried to activate e and succeeded.
 - $\alpha_{e'}=0$: e' failed to activate e within $[t_{e'}, t_e]$.

$$p(e|S_e) = \sum_{\vec{\alpha}_{S_e}} p(e|\vec{\alpha}_{S_e})p(\vec{\alpha}_{S_e})$$

Assume e' activates e independently

Assume $p(\alpha_{S_e})$ is uniformly distributed.

$$p(e|\vec{\alpha}_{S_e}) = \prod_{e' \in S_e} x_{e'e}^{\alpha_{e'}} y_{e'e}^{1-\alpha_{e'}}$$

The probability of e' activating e at time t_e successfully.

$$x_{e'e} = h_{\Delta} g_{\Delta} (1 - g_{\Delta})^{t_e - t_{e'}}$$

The probability of e' not activating e within $[t_{e'}, t_e]$

$$\begin{aligned} y_{e'e} &= 1 - h_{\Delta} g_{\Delta} \sum_{t=t_{e'}}^{t_e} (1 - g_{\Delta})^{t-t_{e'}} \\ &= h_{\Delta} (1 - g_{\Delta})^{t_e - t_{e'} + 1} + (1 - h_{\Delta}) \end{aligned}$$

The final log-likelihood:

$$\log \mathcal{L} = \sum_{e \in \mathcal{E}} \left\{ \log \sum_{\vec{\alpha}_{S_e}} \prod_{e' \in S_e} x_{e'e}^{\alpha_{e'}} y_{e'e}^{1-\alpha_{e'}} + \sum_{e' \in R_e} \log y_{ee'} \right\}$$

EM Algorithm

- Estimate the influence probabilities associated to 24 triads instead of link pairs.
 - Associate each link pair (e,e') to a triad structure.
 - Aggregate different pairs with the same structure together.

$$\theta = \{h_{e'e}, g_{e'e}\} \longrightarrow \theta = \{h_{\Delta}, g_{\Delta}\}$$

- Introduce a posterior distribution $q(e|\alpha_{Se})$ of $p(e|\alpha_{Se})$, and get a lower bound of the original log-likelihood function.
- Differentiate the lower bound with respect to each parameter and set the partial differential to zero.

Ranking-based Link Prediction

Model	P@1	P@2	P@5	P@10	MAP
CF	47.69	44.24	35.78	30.26	61.55
SimRank	27.44	30.11	28.90	27.53	46.11
Katz	50.46	45.38	36.22	30.16	62.54
RR	54.57	46.87	36.11	29.99	64.53
PAC	47.69	40.85	33.36	28.99	59.68
FCM	75.54	60.43	40.37	31.17	79.66

- CF, SimRank, and Katz
 - They only consider the static structure information and ignore the dynamic evolution of the network structure.
- RR and PAC
 - They fit the distributions of some macroscopic properties such as clustering coefficient and closure ratio.
 - They also do not consider the temporal dependence between two links.

Classification-based Link Prediction

Model	Precision	Recall	F1-measure	AUC
Basic	74.09	54.66	62.90	77.00
SVM	73.54	56.18	63.69	75.28
LRC	63.37	63.51	63.43	88.67
FCM	70.58	60.04	64.88	91.95

- SVM and LRC perform poorer than FCM on the triads presenting relatively weak diffusion effects, especially on triads 1, 2, 3, and 6.

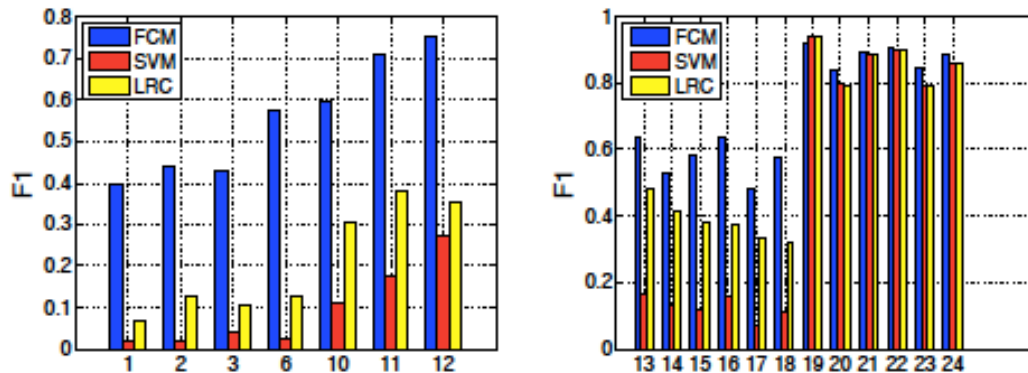
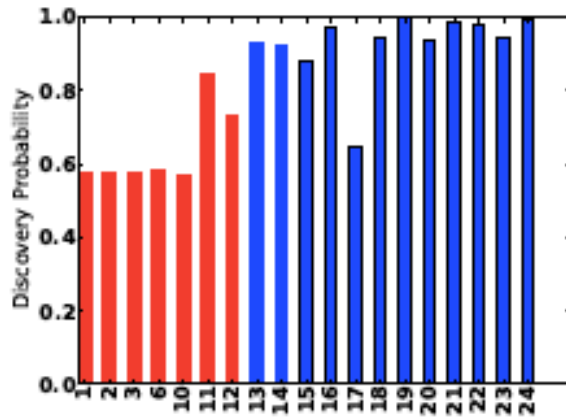


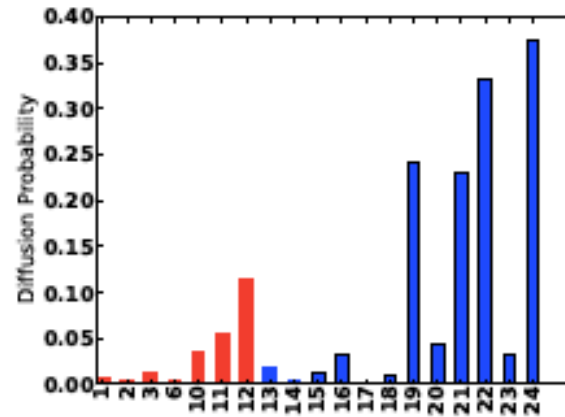
Fig. 8. Performance analysis in different triadic structures on Twitter. X-axis: triadic structure index. Y-axis: F1-measure

- The performance of SVM and LRC may be dominated by the effects from the statistically significant triads.
- FCM smooths the effects from different factors using a generative process.

Learned Model Parameters



(a) Discovery Probabilities

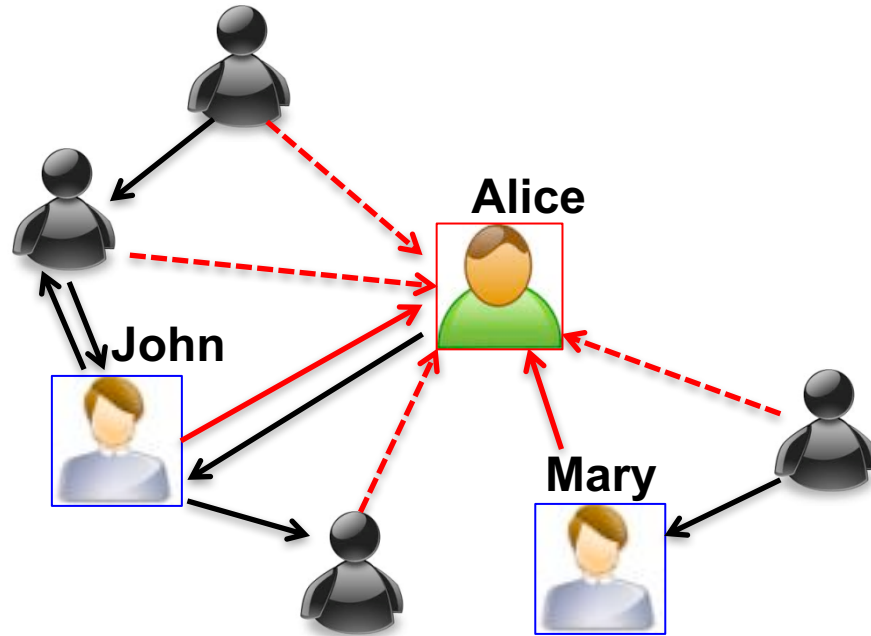


(b) Diffusion Probabilities

Fig. 9. Learned model parameters on Twitter. X-axis: triadic structure index. Y-axis: Discovery/Diffusion probability.

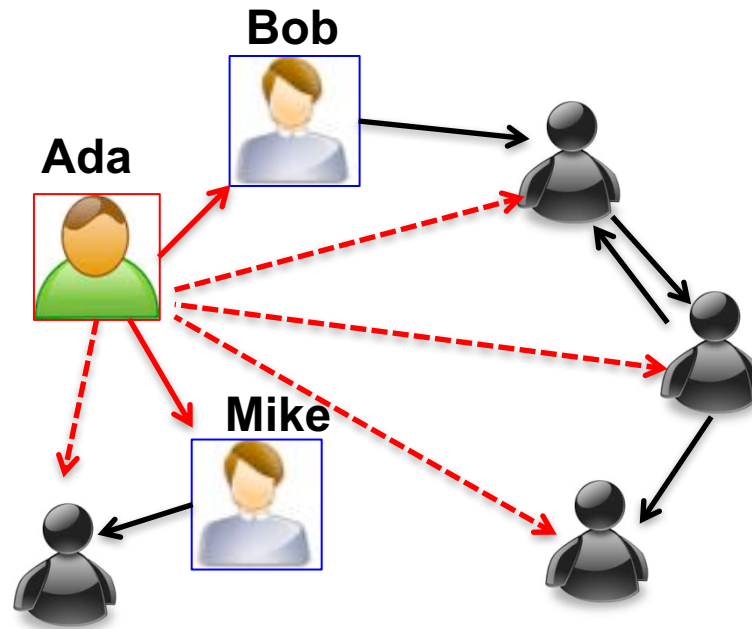
- The discovery probabilities learned for followee diffusion patterns are generally higher than follower diffusion patterns, which indicate that the discoveries in followee diffusion are easier than those in follower diffusion.
- The learned diffusion probabilities are consistent with the rates in Table 1, which suggests that the diffusion effects in followee diffusion are stronger than those in follower diffusion.

Application: Follower Maximization



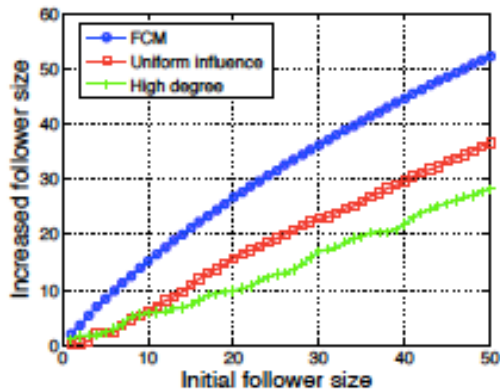
Find a set S of k initial followers to follow user v such that the number of subsequent new followers to follow v is maximized.

Application: Friend Recommendation

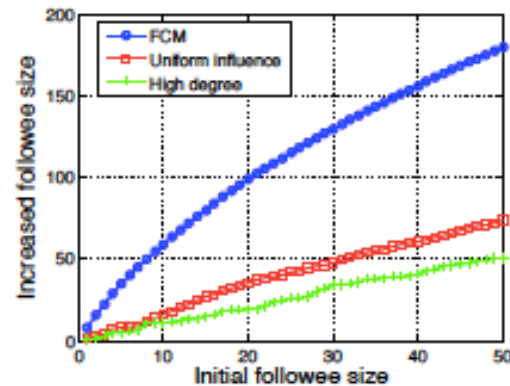


Find a set S of k initial followees for user v such that the total number of subsequent new followees accepted by v is maximized.

Application Performance



(a) Follower maximization



(b) Followee maximization

- High degree
 - May select the users that do not have large influence during link diffusion process.
- Greedy algorithm with uniform configured influence
 - Can not accurately describe the influence between links.
- Greedy algorithm with learned influence by FCM
 - Distinguish the influence in different triad structures.

Conclusion

- Observations
 - Conduct a randomization test to demonstrate the formation of two links in some triads is temporally dependent.
 - The diffusion effect between two links decays over time.
 - A two-way relationship between two users can trigger more links (+1%) than a one-way relationship.
 - A relationship directed from A to C improves the diffusion likelihood from A following C to B following C (+3-40%).
- Propose a “following” link cascade model to depict the link diffusion process by considering the time delay and different diffusion patterns.
- Learn the diffusion strength in different triadic structures by maximizing an objective function based on the proposed model.
- Apply the model into two specific influence maximization applications, follower maximization and followee maximization.



Thank You

Data&Codes: <http://cs.aminer.org/followinf>