中国人民大學
RENMIN UNIVERSITY OF CHINA

Social Influence Analysis and Measurement

Jing Zhang
Information School
Renmin University

Collaborate with

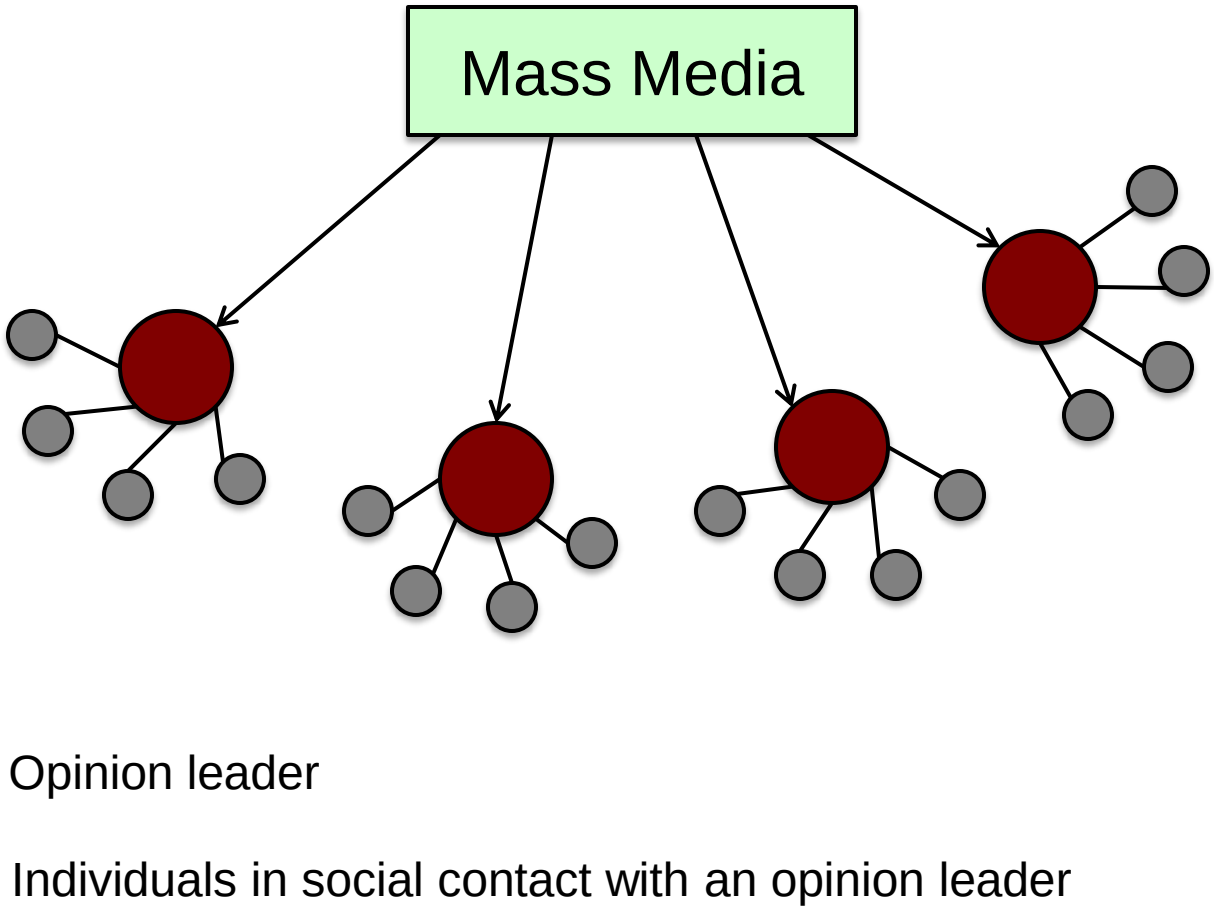
Wei Chen (*MSRA*), Cane Leung (*Huawei Noah's Ark*), Hanghang Tong(*ASU*), Jimeng Sun (*GIT*), Jie Tang (*THU*), Juanzi Li (*THU*)

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
  - Conformity
  - …

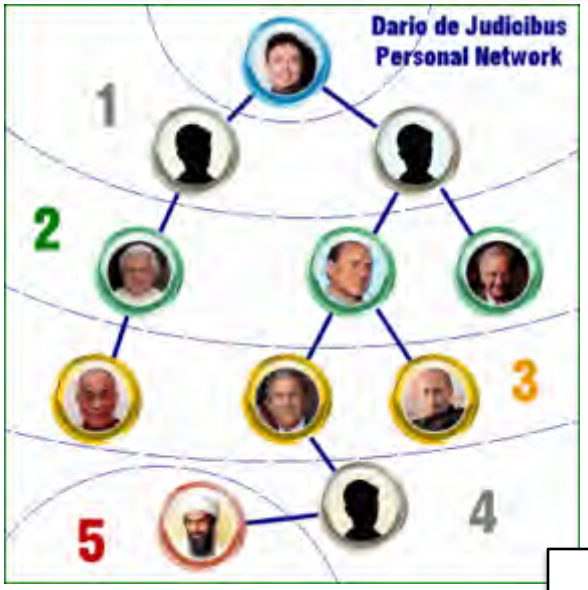
[1] http://en.wikipedia.org/wiki/Social_influence

Two-step Flow Theory



The theory of “Three Degree of Influence”

Six degree of separation[1]



Three degree of Influence[2]

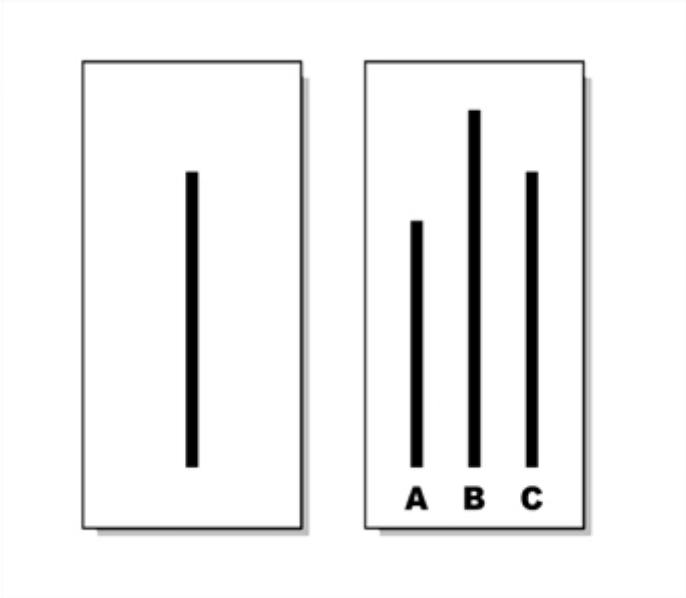
You are able to **influence** up to >1,000,000 persons in the world, according to the Dunbar’s number[3].

[1] S. Milgram. The Small World Problem. Psychology Today, 1967, Vol. 2, 60–67

[2] J.H. Fowler and N.A. Christakis. The Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. British Medical Journal 2008; 337: a2338

[3] R. Dunbar. Neocortex size as a constraint on group size in primates. Human Evolution, 1992, 20: 469–493.

Asch's Experiment



Which line matches the first line, A, B, or C?

74% of the participants followed the majority judgment on at least one trial, even when the majority was wrong.

Experiment on Voting

- Social influence and political mobilization[1]
  - Will online political mobilization really work?

A controlled trial (with 61M users on FB)

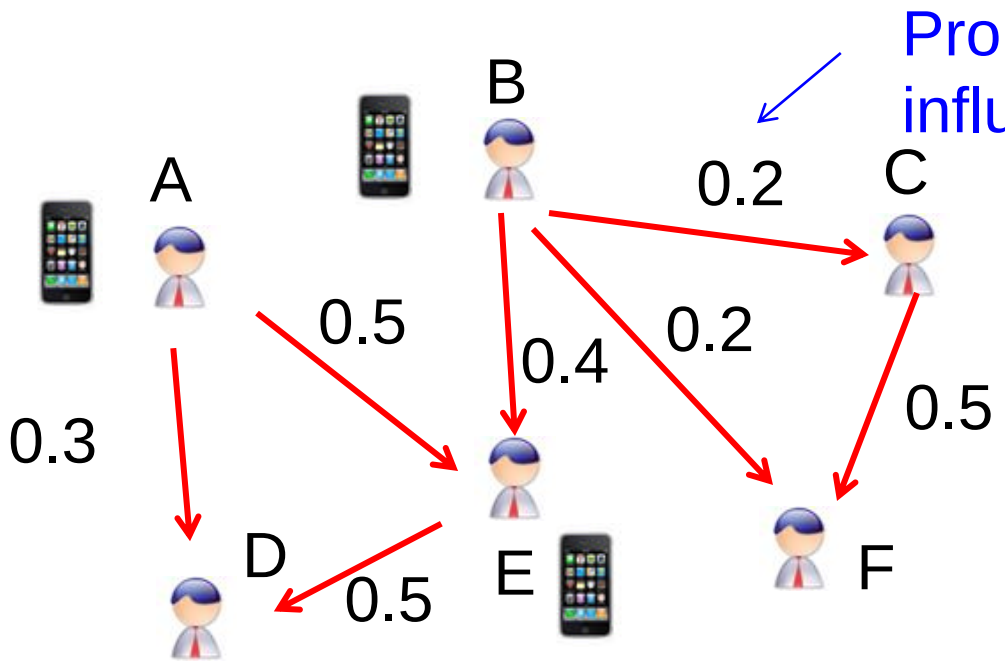
- Social msg group: was shown with msg that indicates one's friends who have made the votes.
- Informational msg group: was shown with msg that indicates how many other.

Social msg group were **2.08%** more likely to click on the "I Voted" button

[1] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. Nature, 489:295-298, 2012.

Virtual Marketing

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



Existing Research

- Influence Test
  - Statistical causal inference

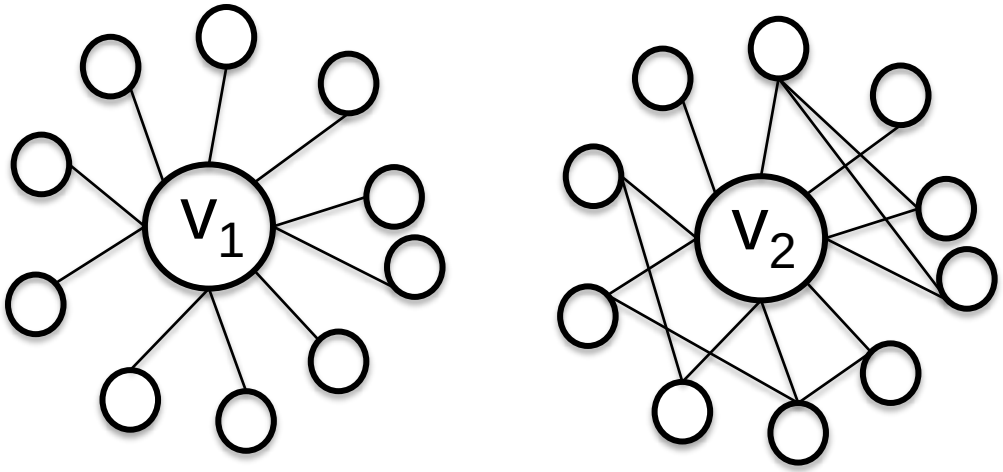
  [Arala et al. 2009] [La Fond and Neville 2010] [Anagnostopoulos et al. 2008]
  - Real controlled trials [Bakshy et al. 2012] [Bond et al. 2012]
- Influence Learning
  - Node influence [Weng et al. 2010]
  - Pairwise infuence [Saito et al. 2008]
  - Group influence [Tang et al. 2013]
- Influence Model
  - Independent cascade model [Kemp et al, 2003]
  - Linear threshold model [Kemp et al, 2003]

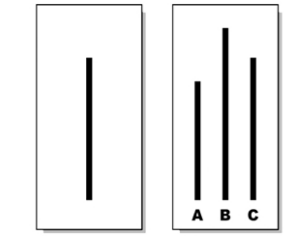
Outline

- Node influence
  - Conformity influence
- Pairwise influence
  - Link influence
- Group influence
  - Structural influence

- An important assumption
  - A is more likely to be influence by B if A's behaviors frequently follow B's.

Conformity Influence





Asch's Experiment

Who is more likely to conform to others, v_1 or v_2?

- Conformity is the inclination of a person to be influenced by others by yielding to perceived group pressure and copying the behavior and beliefs of others [Jenness 1932; Sherif 1935].

Formalize Conformity Influence

- Conformity theory [Bernheim 1994]
  - Everyone in a group expresses her own individuality.
  - Yet, even individualists pursue somewhat for status (esteem or popularity) and change their choices toward the social norm.

- Formalize conformity theory by a utility function:

$$f(y_i) = (1-\lambda_i)\, d(y_i, \hat{y}_i) + \lambda_i \sum_{j \in N(i)} d(y_i, y_j)$$

2 → Esteem acquired through conforming

1 → Individual's intrinsic utility

y_i=1: adopt an action
y_i=0: do not adopt an action

λ_i represents the conformity tendency of v_i

- There exists Nash equilibria if all users in a network make the decisions for a given action according to the utility function.

Measure Conformity Influence

To solve the **data sparsity** problem, we extend the utility function by incorporating role and topic.

- Conformity tendency is different for persons with different roles.
- Conformity tendency is different on actions with different topics.

Binary action y_i vs a set of actions W={w}

$$f(y_i) = (1-\lambda_i)\, d(y_i, \hat{y}_i) + \lambda_i \sum_{j \in N(i)} d(y_i, y_j)$$

$$\gamma_{i,r}^{w} = \left[(1-\lambda_r) \sum_{z=1}^{K} \theta_i^z \phi_z^w + \lambda_r \frac{1}{|N_i|} \sum_{j \in N_i} \sum_{z=1}^{K} \theta_j^z \phi_z^w \right]$$

Conformity tendency of role r

Topic tendency of user v_j on topic z

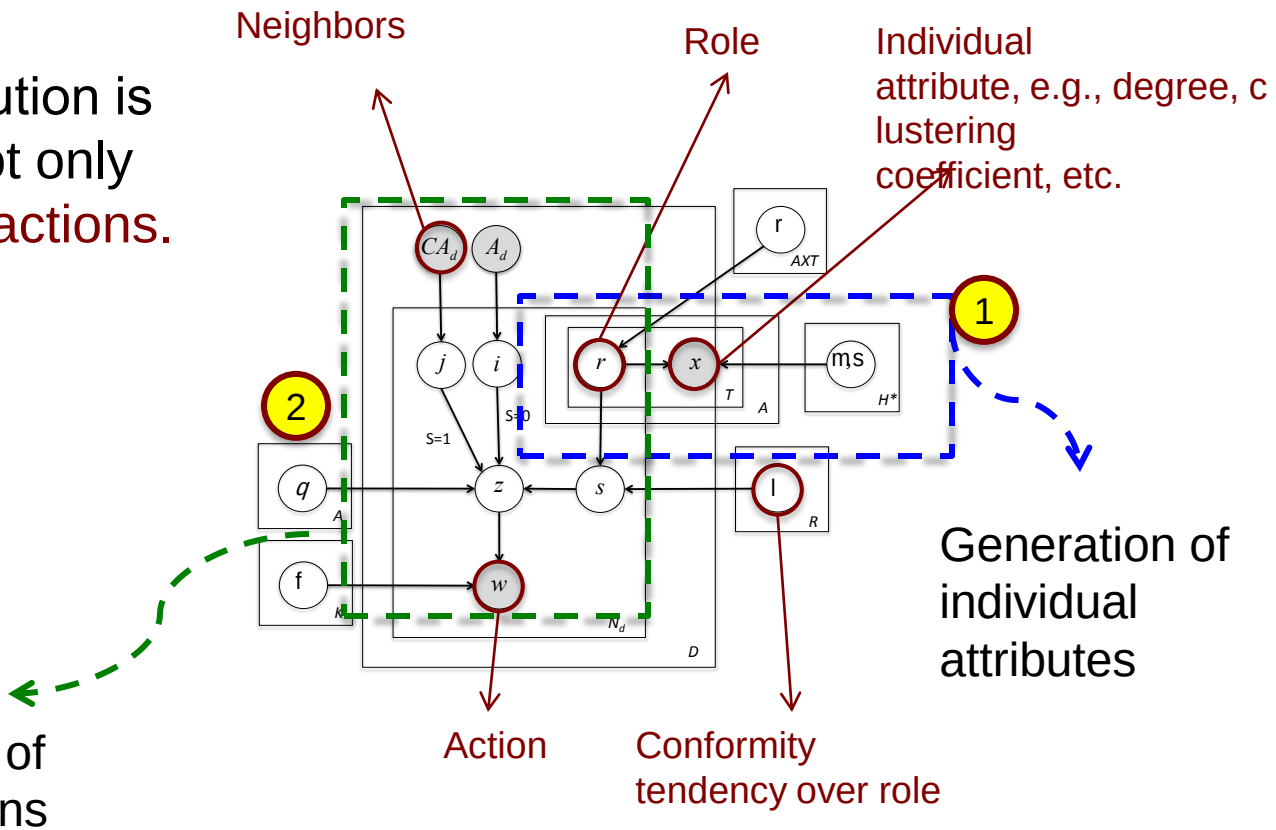
A score of taking action *w* under topic *z*

Model Details

Basic Idea：

Users' role distribution is determined by not only attributes but also actions.

Probabilistic explanation



Parameter Estimation

- The objective is to estimate λ_r , i.e., the conformity over role.
- The method is to maximize the likelihood of generating both the individual attributes and the actions.

$$\mathcal{L}_1 = \prod_{i=1}^{A}\prod_{t=1}^{T}\prod_{h=1}^{H}\sum_{r=1}^{R}\frac{\rho_{i,t}^{r}}{\sqrt{2\pi\sigma_{r,h}^{2}}}\exp\left[-\frac{(x_{i,t,h}-\mu_{r,h})^{2}}{2\sigma_{r,h}^{2}}\right]$$

$$\mathcal{L}_2 = \prod_{d,w}\sum_{i\in A_d}\frac{\sum_{r=1}^{R}\rho_{i,t}^{r}\gamma_{r,i}^{w}}{|A_d|}$$

- We iteratively optimize *L1* and *L2* by using EM algorithms and solve the parameters.

$$\mu_{r,h} = \frac{\sum_{i=1}^{A}\sum_{t=1}^{T} q_{i,t,h}^{r} x_{i,t,h}}{\sum_{i=1}^{A}\sum_{t=1}^{T} q_{i,t,h}^{r}}$$

$$\sigma_{r,h} = \sqrt{\frac{\sum_{i=1}^{A}\sum_{t=1}^{T} q_{i,t,h}^{r}(x_{i,t,h}-\mu_{r,h})^{2}}{\sum_{i=1}^{A}\sum_{t=1}^{T} q_{i,t,h}^{r}}}$$

$$\theta_i^z \propto \sum_{d=1}^{D}\sum_{w\in N_d}\sum_{r=1}^{R} a_{d,w,i}^{r} b_{d,w,i,r} c_{d,w,i,r}^{z} + \sum_{d=1}^{D}\sum_{w\in N_d}\sum_{j\in A_d}\sum_{r=1}^{R}\left[a_{d,w,i}^{r}(1-b_{d,w,i,r})\sum_{j\in CA_d} c_{d,w,j,r}^{z}\right]$$

$$\phi_z^w \propto \sum_{d=1}^{D}\sum_{i\in A_d}\sum_{r=1}^{R} a_{d,w,i}^{r} b_{d,w,i,r} c_{d,w,i,r}^{z} + \sum_{d=1}^{D}\sum_{i\in A_d}\sum_{r=1}^{R}\left[a_{d,w,i}^{r}(1-b_{d,w,i,r})\sum_{j\in CA_d} c_{d,w,j,r}^{z}\right]$$

$$\lambda_r = \sum_{d=1}^{D}\sum_{w\in N_d}\sum_{i\in A_d}\sum_{r=1}^{R} a_{d,w,i}^{r} b_{d,w,i,r}$$

$$\rho_{i,t}^{r} = \frac{\sum_{h=1}^{H} q_{i,t,h}^{r} + \sum_{d,w} a_{d,w,i}^{r}}{\sum_{r=1}^{R}(\sum_{h=1}^{H} q_{i,t,h}^{r} + \sum_{d,w} a_{d,w,i}^{r})}$$

Evaluate Conformity through "Wording" Behavior Prediction

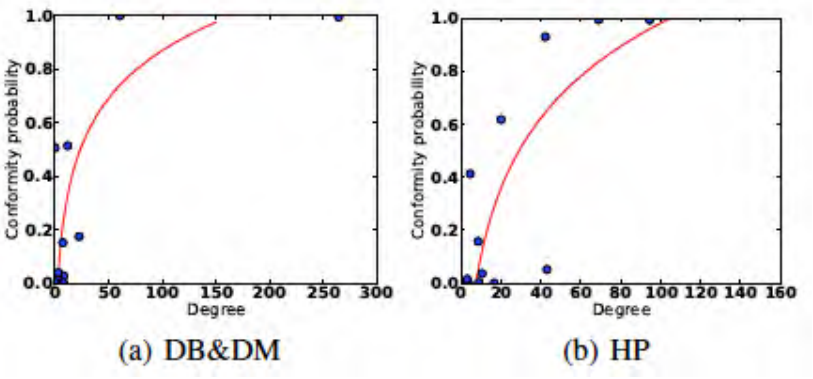
Table 2: Performance of word usage prediction (%).

Query	Method	P@5	P@10	MAP	AUC
DB&DM	PLSA	20.10	15.49	09.26	77.61
	CIM	22.26	17.98	11.59	85.50
	RCM	**30.40**	**24.94**	**14.16**	**86.90**
HCI	PLSA	14.25	11.65	5.71	67.37
	CIM	18.67	15.34	8.12	73.39
	RCM	**19.16**	**15.40**	**8.92**	**75.32**
HP	PLSA	17.33	14.39	8.47	79.96
	CIM	19.62	16.25	10.83	88.67
	RCM	**20.57**	**17.12**	**11.37**	**89.21**
SE	PLSA	4.20	2.60	2.60	81.15
	CIM	21.43	16.42	12.16	**85.55**
	RCM	**25.31**	**19.98**	**12.54**	85.27
CT	PLSA	17.52	13.37	9.88	81.09
	CIM	19.36	14.50	11.04	85.31
	RCM	**20.13**	**15.20**	**11.46**	**85.93**
AI&ML	PLSA	19.92	15.50	9.40	84.10
	CIM	21.24	16.41	10.85	90.70
	RCM	**23.60**	**18.02**	**11.41**	**90.92**
CN	PLSA	26.68	20.33	12.99	80.63
	CTM	29.36	21.62	14.75	86.92
	RCM	**31.20**	**23.35**	**15.22**	**88.41**
CV&MM	PLSA	19.88	14.78	09.64	78.85
	CTM	22.09	16.12	11.10	85.02
	RCM	**24.49**	**18.37**	**11.50**	**85.63**
Avg	PLSA	17.49	13.51	8.49	78.85
	CTM	21.75	16.83	11.31	85.13
	RCM	**24.36**	**19.05**	**12.07**	**85.95**

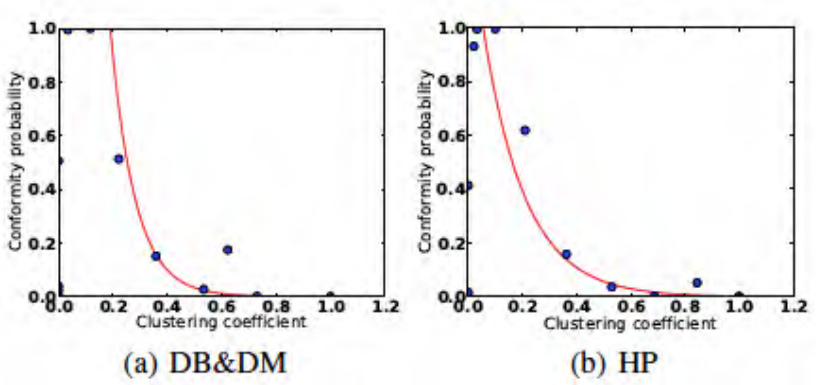
- PLSA only consider the intrinsic preference, and ignores the situation where a user's topic distribution may change and become closer to her neighbors' topic distribution over time.
- CTM (Citation influence model) directly learns the conformity tendency of each user, which becomes very difficult to be estimated accurately when very few historical actions of the user and/or her neighbors are available.
- Our model RCM clusters similar users into roles, and then learn the conformity tendency of each role.

The Correlation between Role and Conformity

People with higher degree and lower clustering coefficient are more likely to conform to others.



Mean degree of role

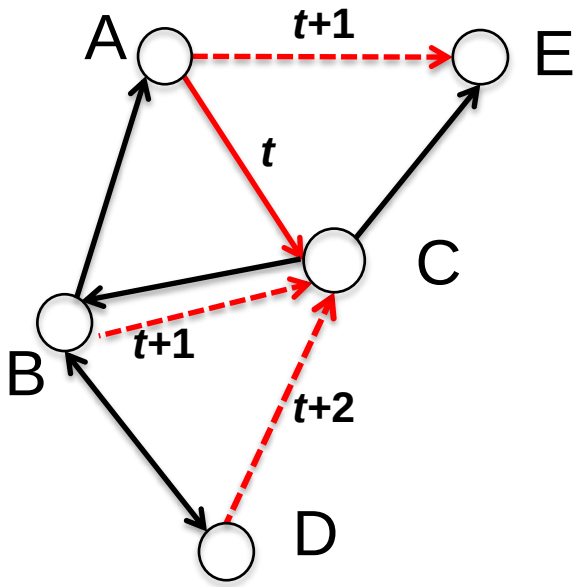


Mean clustering coefficient of role

When a person collaborates with more authors and the coauthors are more structurally diverse, she may become more open-minded and tend to accept new ideas from others.

When the social circle of the user is restricted to a few coauthors forming a dense collaboration network, the person will be more conservative and tend not to accept other ideas.

Pairwise Influence between Links

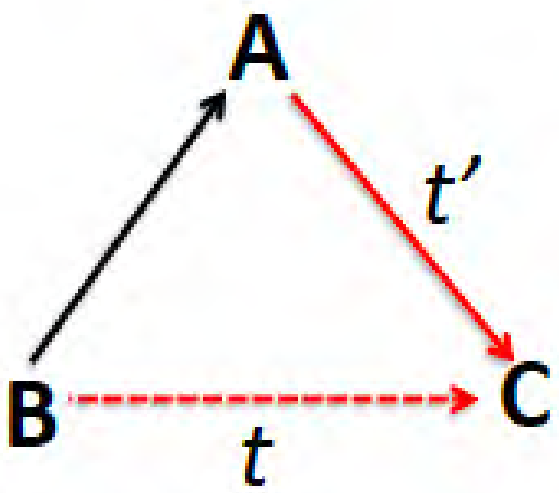


Will the formation of AC influences AE, BC, and DC to be formed?

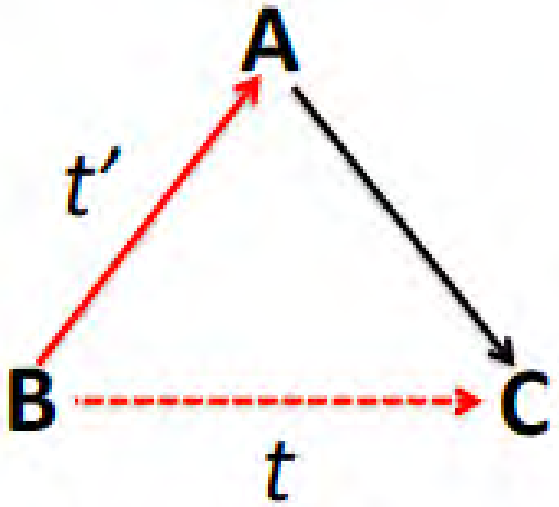
Active link

Link to be influenced

Two Categories of Link Influence



(a) Follower diffusion



(b) Followee diffusion

–>: pre-existing relationships
–>: a new link added at time *t'*
-->: a possible link added at time *t*

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

Randomization Test

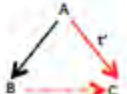
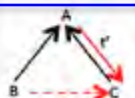
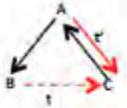
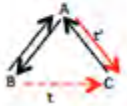
- Randomization test is a model-free, computationally intensive statistical technique for hypothesis testing, the main steps are
  1. Compute some test statistic using the set of original observations;
  2. Carry out the random shuffle according to the null hypothesis a large number of times, and compute 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 more or less than the observed value of the test statistic.
- Null hypothesis: the formation of neighboring links is temporally independent of one another.
- Test statistic:

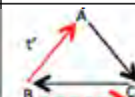
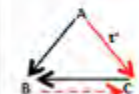
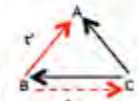
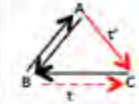
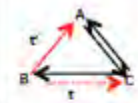
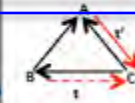
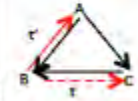
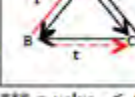
$$\text{rate} = \frac{|\{\text{triad(A,B,C)}|0 \leq t_{BC} - t_{AC} \leq \delta\}|}{|\{\text{triad(A,B,C)}\}|}$$

P-values on 24 Triads

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

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				
	△	$\|C_\triangle\|$	$\|C^+_\triangle\|$	$r_\triangle$		△	$\|C_\triangle\|$	$\|C^+_\triangle\|$	$r_\triangle$	
1		22870	233	0.0102	***	13		24162	2298	0.0951	***
2		22527	246	0.0109	**	14		62411	2293	0.0367	***
3		33122	642	0.0194	***	15		63092	3985	0.0632	***
4		29830	100	0.0034		16		23099	2314	0.1002	***
5		2370	3	0.0013		17		25049	324	0.0129	***
6		7283	76	0.0104	*	18		65219	3469	0.0532	***

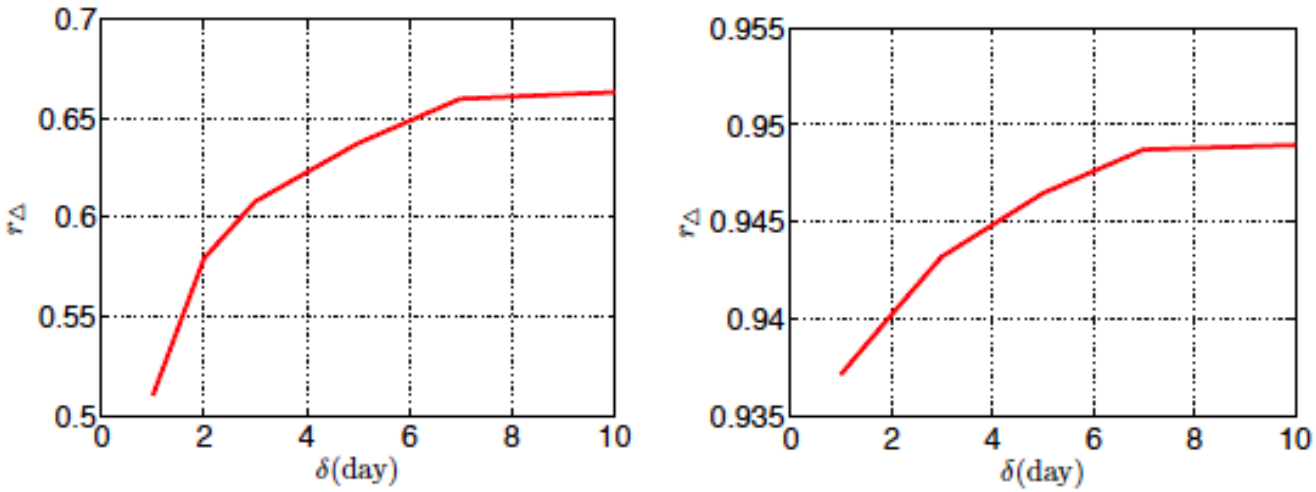
	Follower Diffusion					Followee Diffusion				
	△	$\|C_\triangle\|$	$\|C^+_\triangle\|$	$r_\triangle$		△	$\|C_\triangle\|$	$\|C^+_\triangle\|$	$r_\triangle$	
7		116	3	0.0259		19		428	315	0.7360	***
8		883	77	0.0872		20		5729	2300	0.4015	***
9		730	71	0.0973		21		4372	3427	0.7839	***
10		666	46	0.0691	**	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

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

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

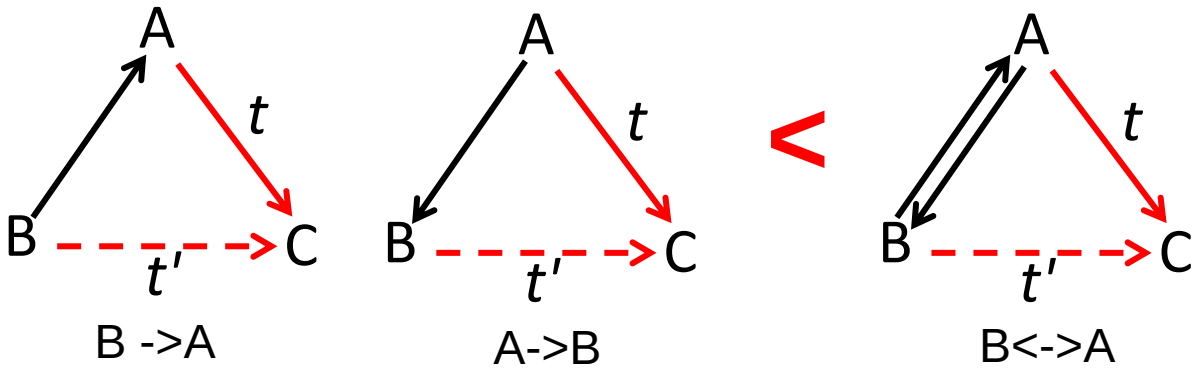
Diffusion Decay



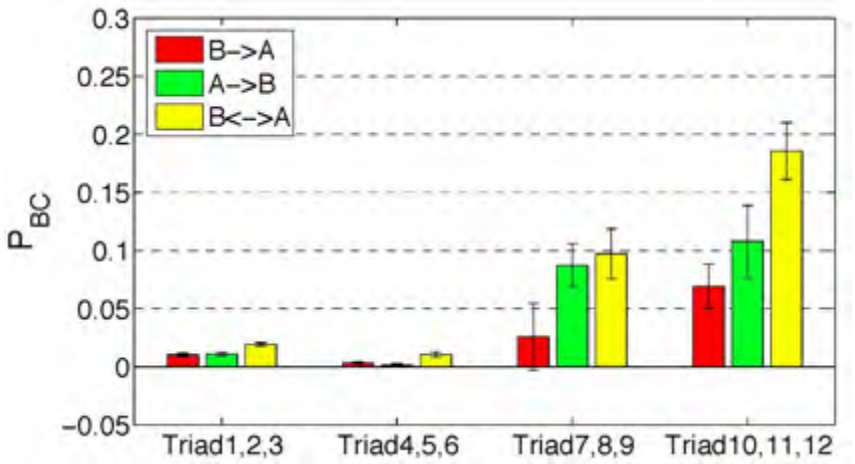
(a) Follower diffusion (b) Followee diffusion

- 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.

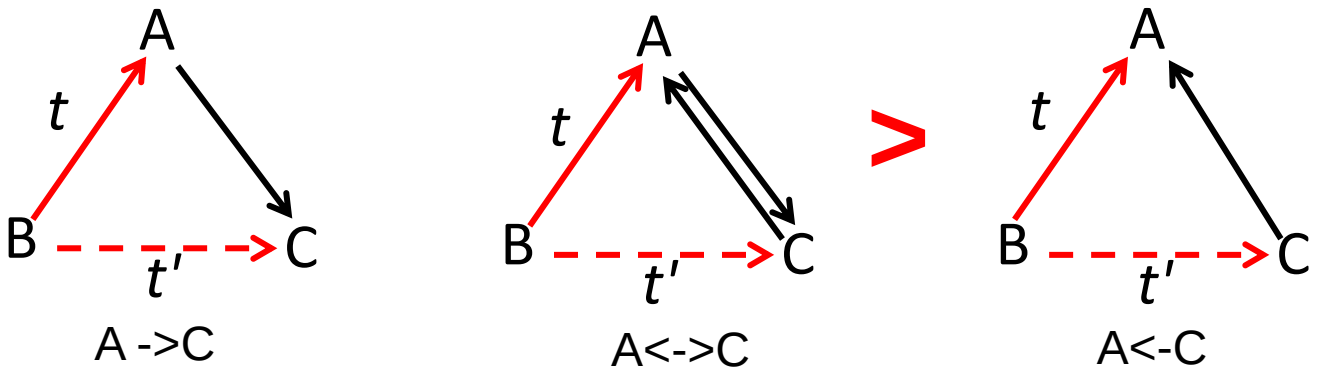
Follower Diffusion: Power of Reciprocity



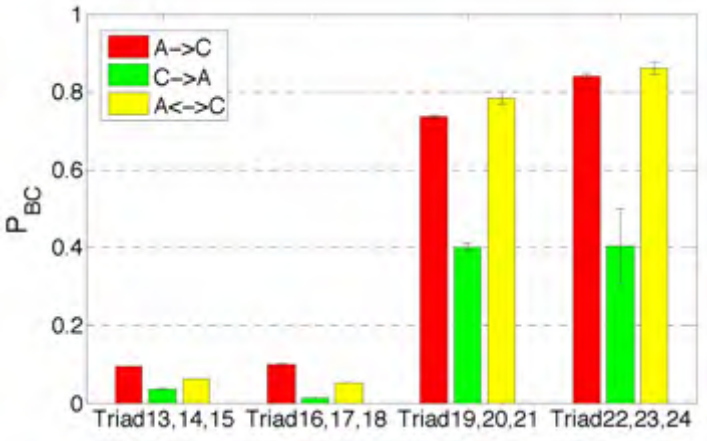
Observation: Reciprocal relationships are much more likely to be actual "social" relationships, rather than "celebrity following", and thus have stronger social influence.



Followee Diffusion: Easy Discovery

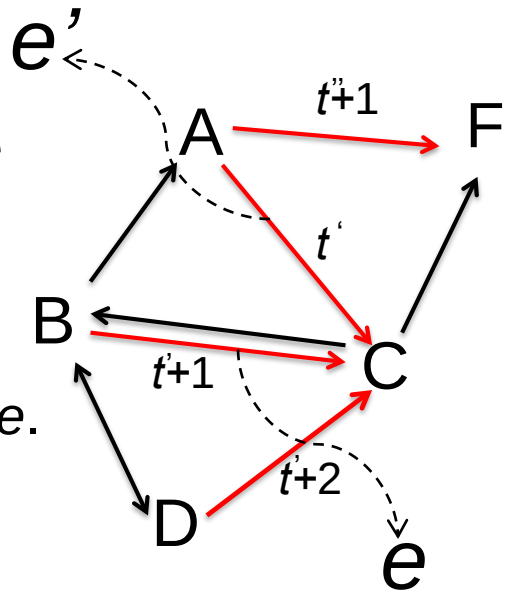


Observation: When a user B follows another user A, who already follows user C, B is likely to discover C through browsing A's retweets of C's messages or directly checking A's followee list, and A's interest in C may indicates that B would also be interested in C.



“Following” Link Cascade Model

- When a link *e’* is added at time *t’*, at each time slot from time *t’* to *t’*+*δ*:
  - 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*’+*λ* that *e*’ could activate *e*.



Influence Estimation

- The objective 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.

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

1

We formalize the formation of each newly added link.

2

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 $\boldsymbol{\alpha}_{Se} = \{\alpha_{e'}\}_{e' \in Se}$ 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 p(α_{Se}) is uniformly distributed.

Assume *e*' activates *e* independently

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

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

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

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

$$y_{e'e} = 1 - h_\triangle g_\triangle \sum_{t=t_{e'}}^{t_e} (1 - g_\triangle)^{t - t_{e'}}$$
$$= h_\triangle (1 - g_\triangle)^{t_e - t_{e'} + 1} + (1 - h_\triangle)$$

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

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

- Introduce a posterior distribution $q(e|\boldsymbol{\alpha}_{Se})$ of $p(e|\boldsymbol{\alpha}_{Se})$, 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**

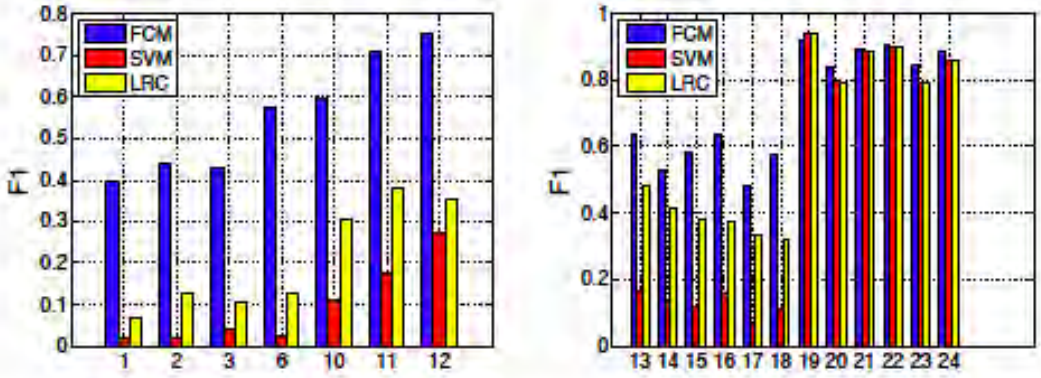
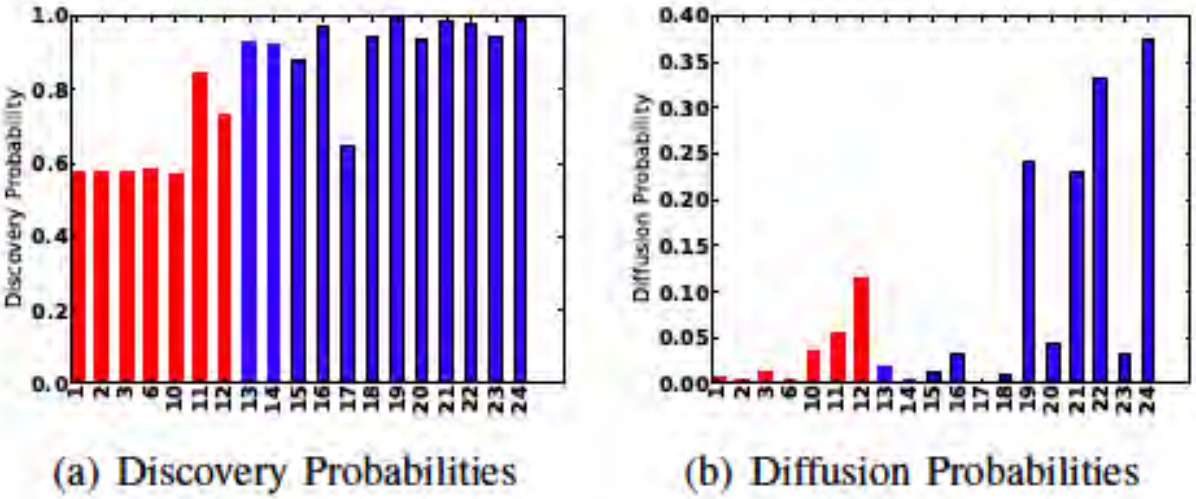


Fig. 8. Performance analysis in different triadic structures on Twitter. X-axis: triadic structure index. Y-axis: F1-measure

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

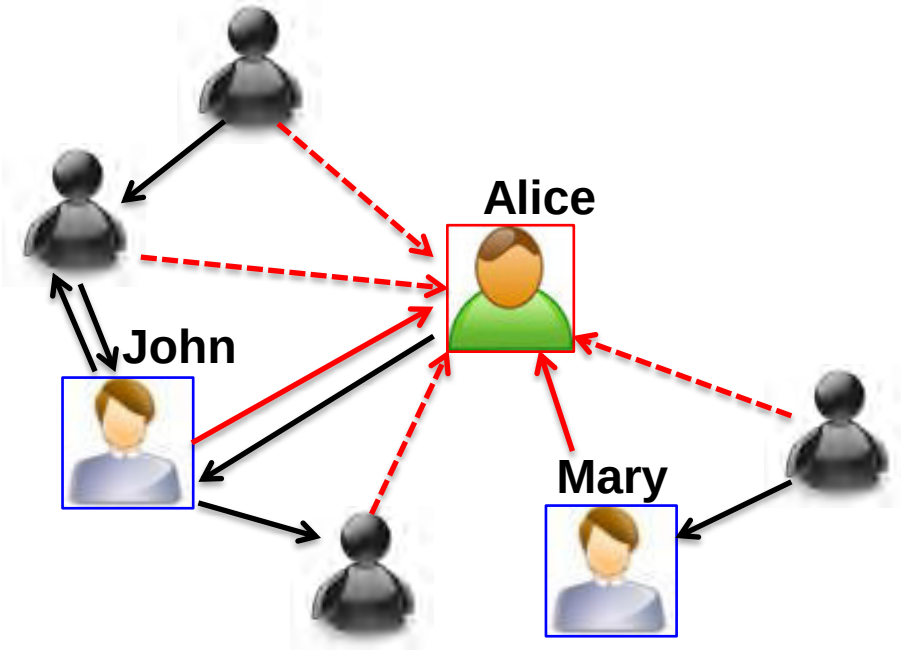


(a) Discovery Probabilities (b) Diffusion Probabilities

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

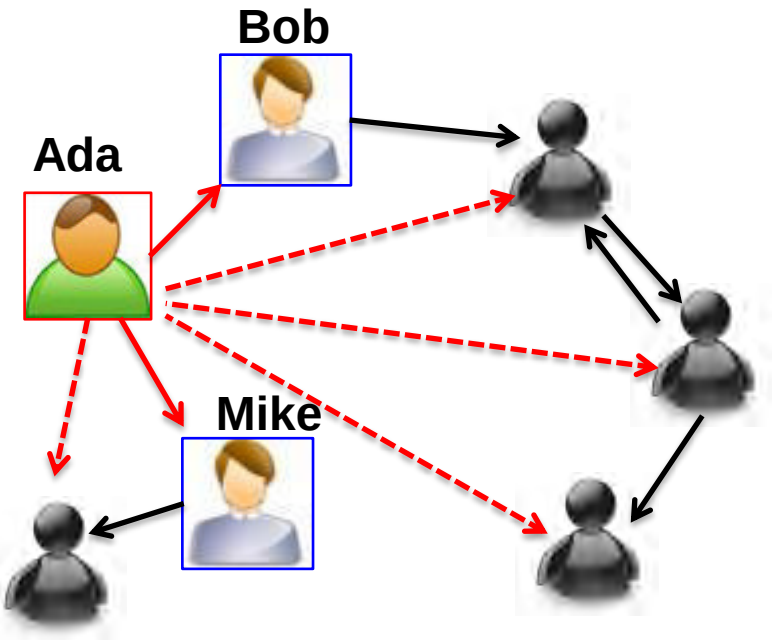
- The discoveries in followee diffusion are easier than those in follower diffusion.
- The diffusion effects in followee diffusion are stronger than those in follower diffusion.

Application: Follower Maximization



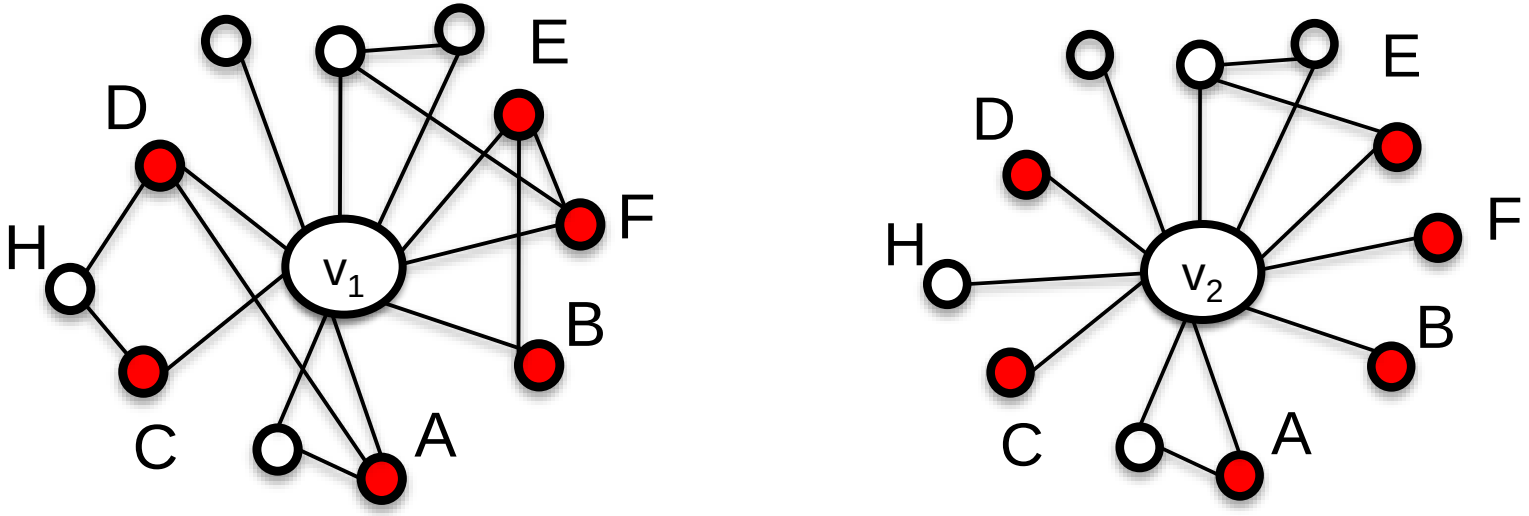
Find a set *S* of *k* initial followers to follow user *v* such that the number of subsequent new followers to follow *v* is maximized.

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.

Structural Influence



Whose ego network has more influence, v_1 or v_2?

Active neighbor Inactive neighbor v User to be influenced

Test Influence Locality

Goal: evaluate the correlation between active probability and the active neighbors.

Randomized experiment

Treatment group	vs	Control group
Users have > 1 active neighbors		Users have =1 active neighbor

新浪微博 weibo.com

Selection bias: users assigned in the treatment group are more likely to retweet than those in the control group even though they do not have >1 active neighbors, because of homophily.

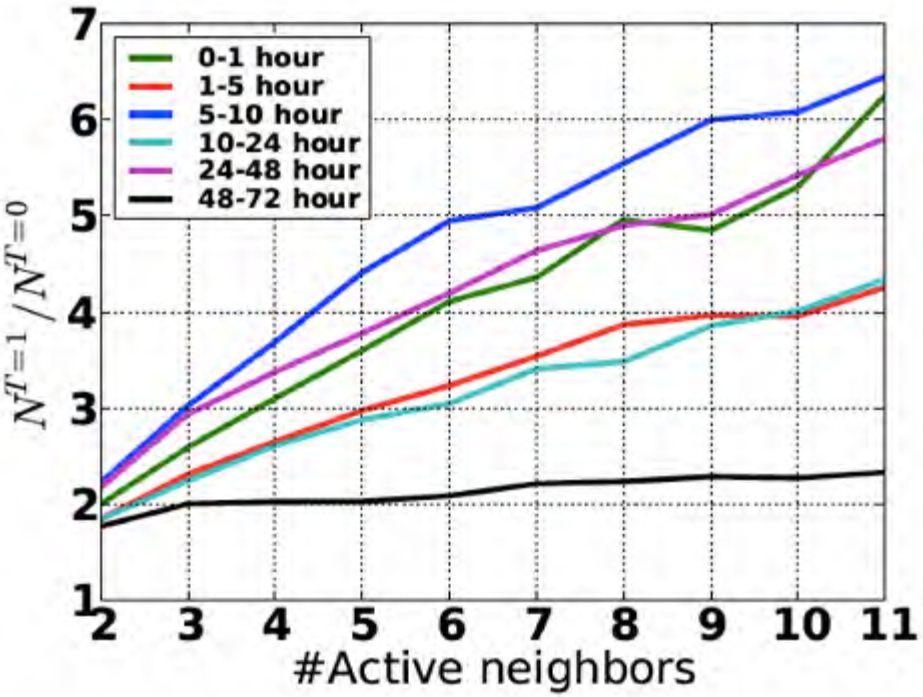
Matched sampling: Match the users in treatment group to those in control group with similar probability to be treated.

$$p_{it} = P(T_{it} = 1 \mid X_{it})$$

A binary variable indicating whether user *i* will be treated at time *t*

All attributes associated with user *i* at time *t*

Test Result



The fraction of active users with 2 active neighbors is about **2 times** the fraction of active users with only 1 active neighbor.

The ratio increases with the number of active neighbors.

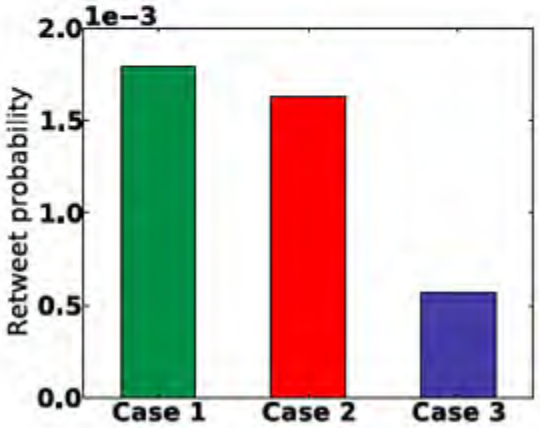
After 48 hours when the original tweet has been published, the increasing rate slows down.

$N^{T=1}$: the average number of active users in the treatment group.
$N^{T=0}$: the average number of active users in the control group.

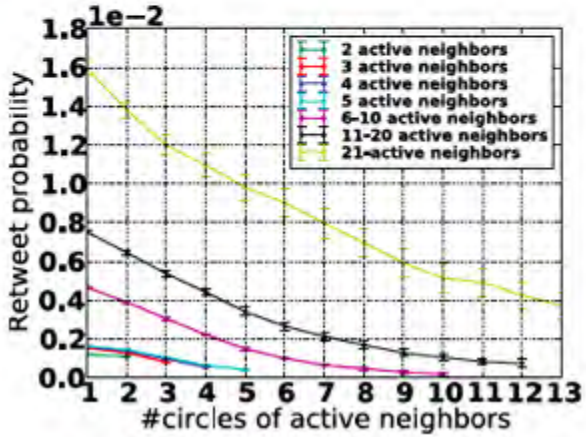
The ratio $N^{T=1}/N^{T=0}>1$ indicates the influence locality exerts positive effect on users' retweet behaviors.

Ego Network Structure and Influence

(a) Case 1: #circles=2 (b) Case 2: #circles=4 (c) Case 3: #circles=6



(d) Retweet probability in case 1,2 and 3



(e) Retweet probability in various cases

The probability of a user retweeting a microblog is **negatively** correlated with the structure diversity of the active neighbors.

Evaluate through Retweet Prediction

Ego network Influence Q = #active neighbors + #circles formed by active neighbors

Table 2: Performance of retweet behavior prediction. (%)

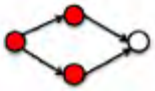
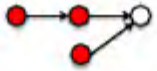
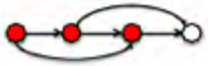
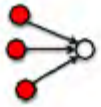
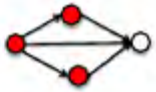
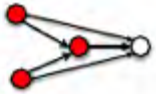
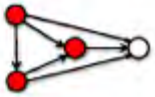
Model	Prec.	Rec.	F1	Acc.
LRC-B	68.11	74.26	71.05	69.74
LRC-Q	66.82	77.22	71.65	69.44
LRC-BQ	69.89	77.06	73.30	71.93

With only ego network influence factor, we can obtain a F1-score of 71.65%.

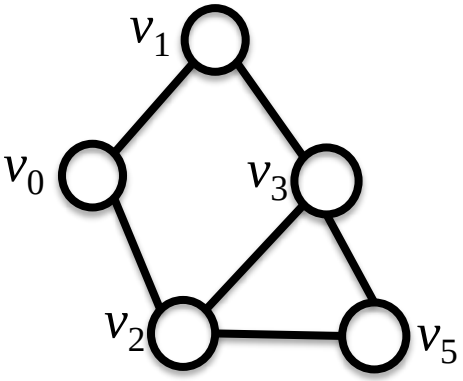
LRC-B: logistic regression classifier with only basic features
LRC-Q: logistic regression classifier with only the feature of ego network influence.
LRC-BQ: Combine basic features and influence locality function together.

Basic features: Gender, verification status, #followers, #two-way following relationships, #one-way following relationships, #historical microblogs, topic propensity, the elapsed time

Structural Influence

k	C_k	IP_k	$\widetilde{IP}_k$	U_{IP_k}	k	C_k	IP_k	$\widetilde{IP}_k$	U_{IP_k}
1		0.066	0.066	0.020	11		0.038	0.038	0.720
2		0.074	0.074	0.085	12		0.186	0.186	0.088
3		0.111	0.110	0.425	13		0.399	0.392	1.785
4		0.307	0.304	0.928	14		0.063	0.062	0.616
5		0.069	0.069	0.530	15		0.619	**0.615**	0.548
6		0.091	0.090	0.358	16		0.444	**0.439**	1.378
7		0.067	0.067	0.236	17		0.070	0.070	0.074
8		0.106	0.099	5.852	18		0.420	**0.416**	0.890
9		0.381	0.388	1.666	19		0.662	**0.645**	2.696
10		0.165	0.162	1.128	20		0.485	**0.479**	1.239

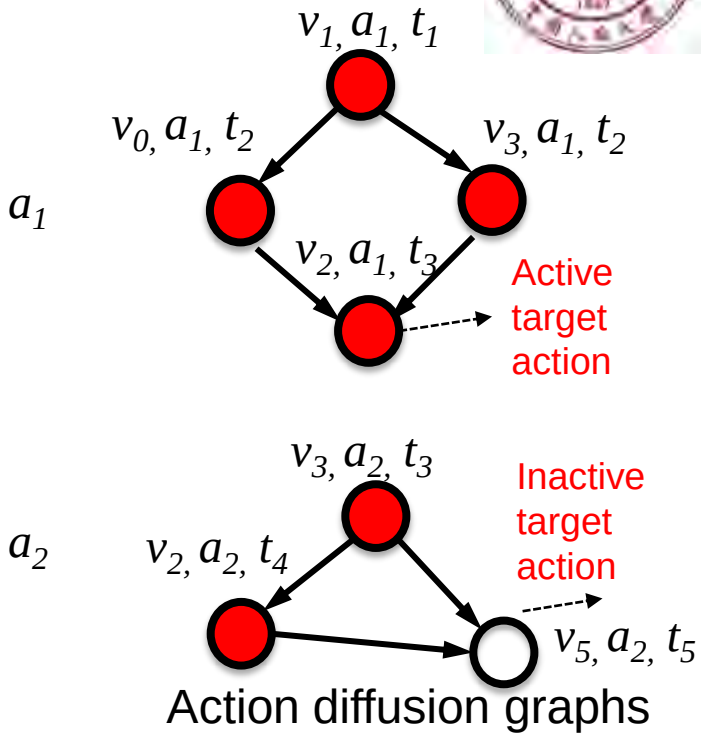
Problem of Structural Influence Measurement



Network

v_1, a_1, t_1
v_0, a_1, t_2
v_3, a_1, t_2
v_2, a_1, t_3
v_3, a_2, t_3
v_2, a_2, t_4
…

Action logs

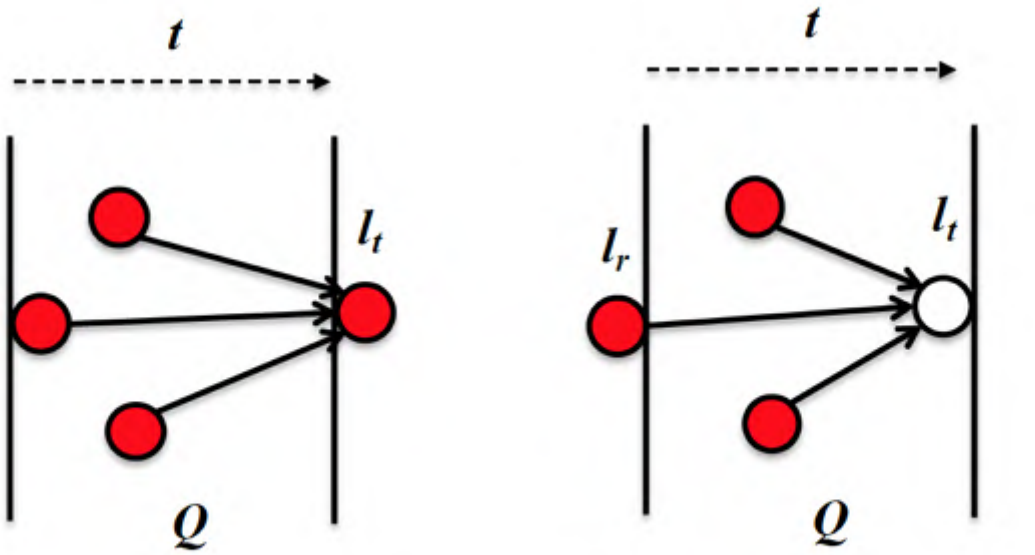


Action diffusion graphs

Influence Probability		$l_t \in L$	$l_t \notin L$
$\mathrm{IP}(C_k) = \frac{x_k}{x_k + y_k}$	C_k	x_k	y_k
	$\overline{C_k}$	z_k	w_k

Approach: StructInf-Basic

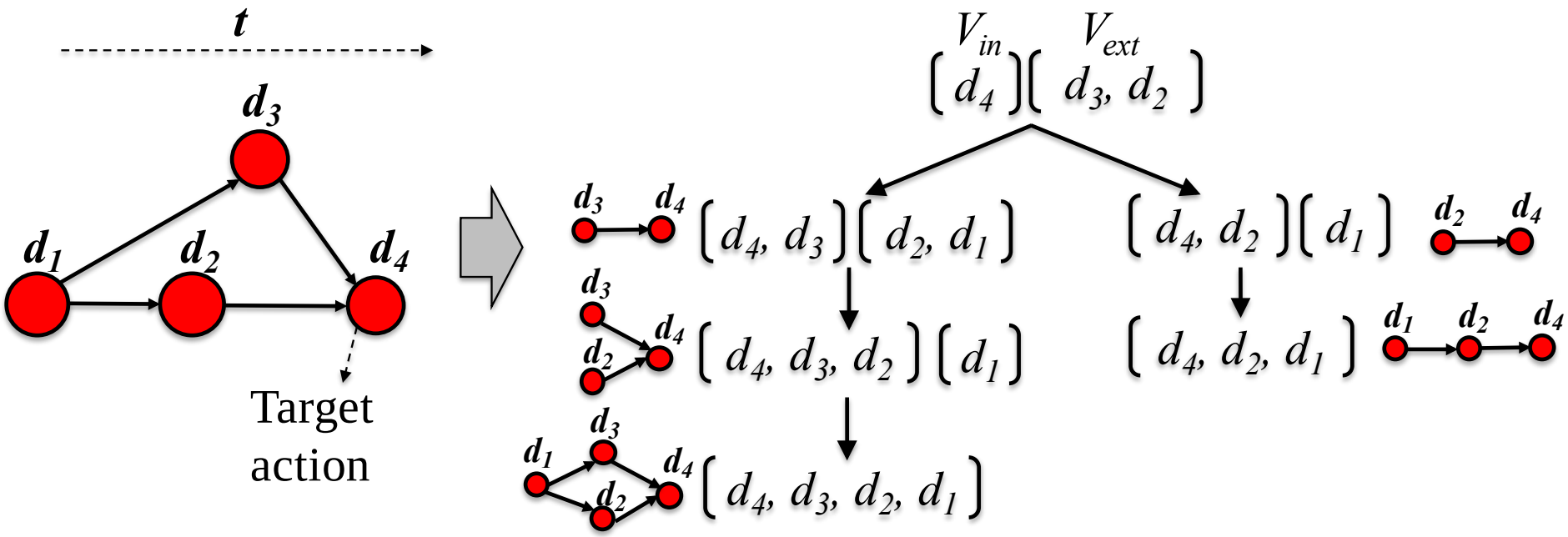
- Identify active and inactive target actions
  - Count active actions when an action newly arrives
  - Count inactive actions when an action is outdated



(a) Active target action (b) Inactive target action

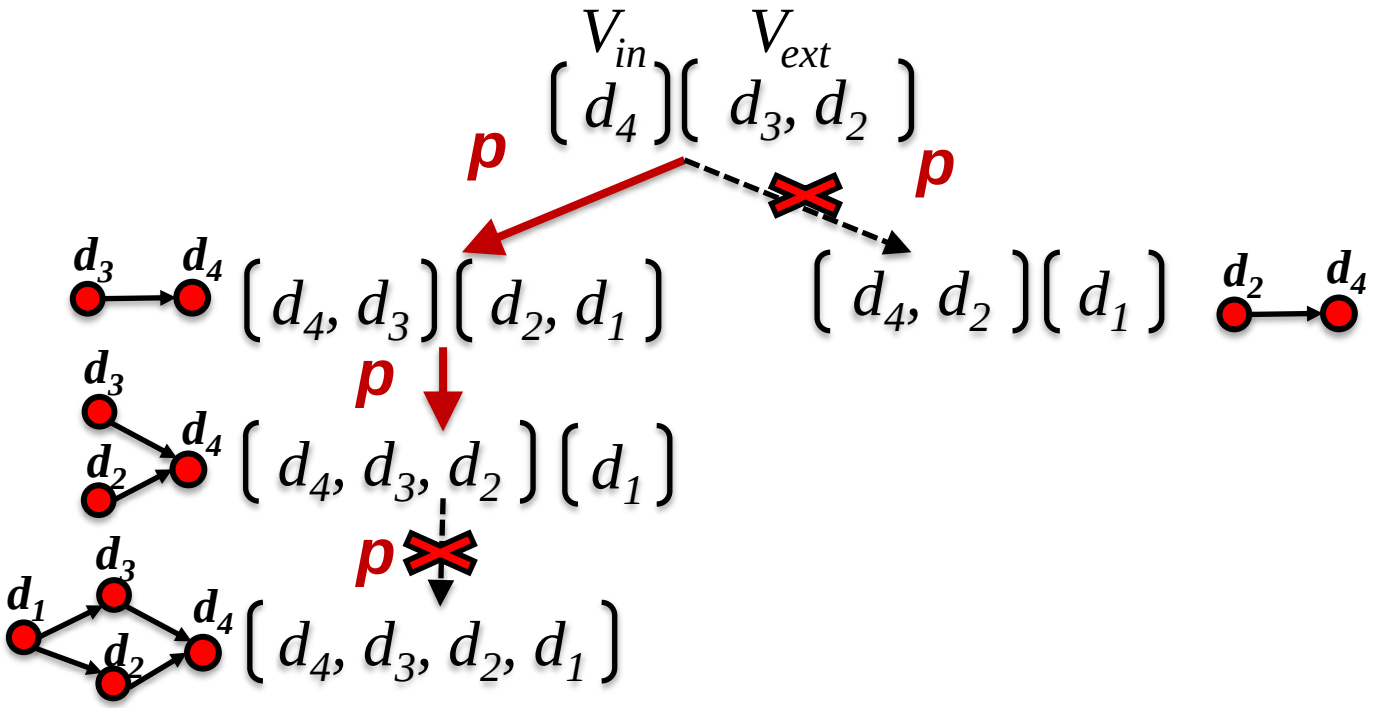
Approach: StructInf-Basic

- Enumerate influence patterns
  - Extend nodes instead of edges
  - Dynamic labeling to avoid duplication



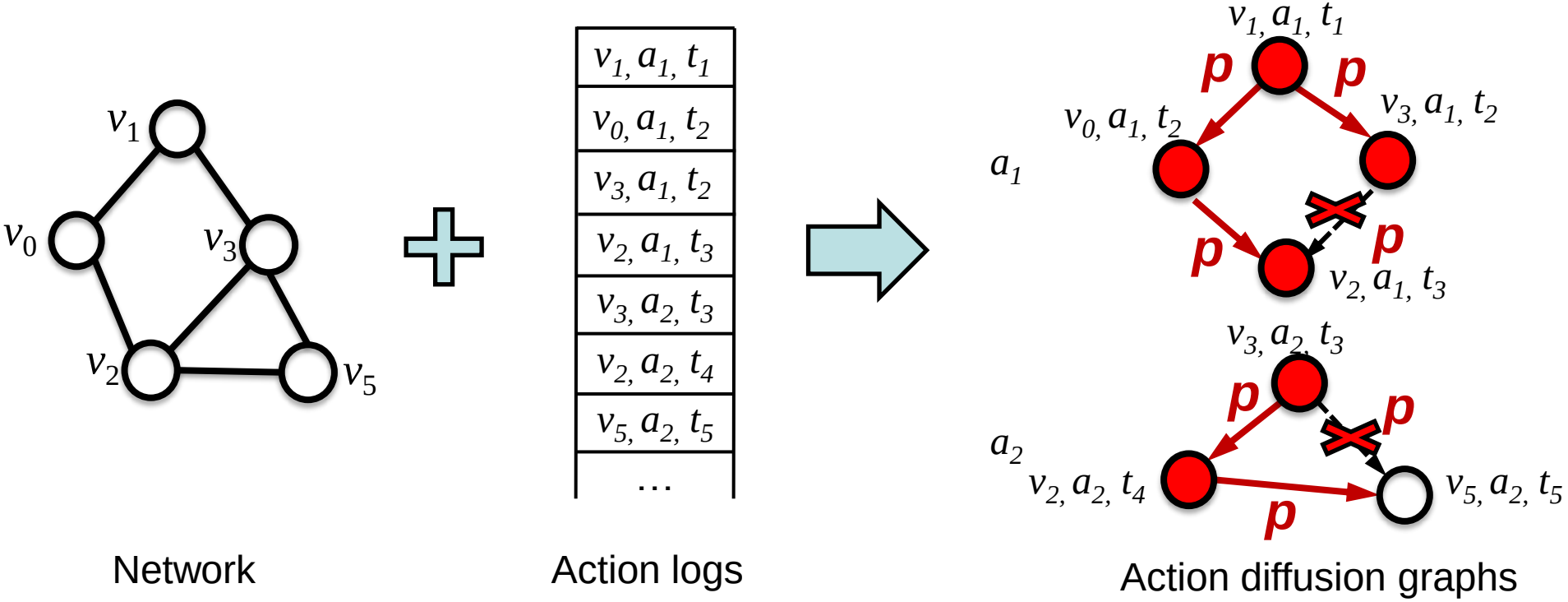
Fast Sampling: StructInf-S1

- Randomly sample nodes when enumerating influence patterns using **sampling probability** *p*.



Fast Sampling: StructInf-S2

- Randomly sample edges when constructing action diffusion graphs using **sampling probability** *q*.



Fast Sampling: StructInf-S3

- Combine StructInf-S1 and StructInf-S2
- Randomly sample edges when constructing action diffusion graphs using **sampling probability** *q* and sample nodes when enumerating influence patterns using **sampling probability** *p* together.

UnBiasness Property

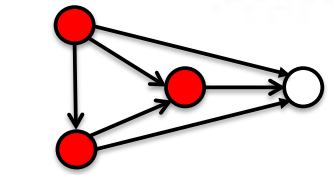
- StructInf-S1

$$\widetilde{x}_k = \hat{x}_k / p^{n_k}$$

- StructInf-S2
  - Complete subgraph $\hat{x}_k / q^{m_k}$
  - Incomplete subgraph

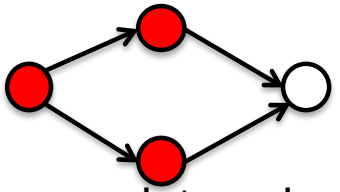
$$\widetilde{x}_k = \frac{\hat{x}_k + \sum_{C_i:C_k \subset C_i \& n_k = n_i} n_{ik} \hat{x}_i}{q^{m_k}} - \sum_{\substack{C_i:C_k \subset C_i \\ \& n_k = n_i}} n_{ik} \widetilde{x}_i$$

- StructInf-S3
  - Complete subgraph $\hat{x}_k / (p^{n_k} q^{m_k})$
  - Incomplete subgraph



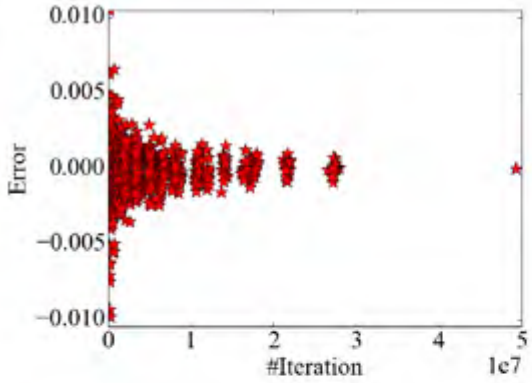
Complete subgraph

$$\widetilde{x}_k = \frac{\hat{x}_k + \sum_{C_i:C_k \subset C_i \& n_k = n_i} n_{ik} \hat{x}_i}{p^{n_k} q^{m_k}} - \sum_{\substack{C_i:C_k \subset C_i \\ \& n_k = n_i}} n_{ik} \widetilde{x}_i$$

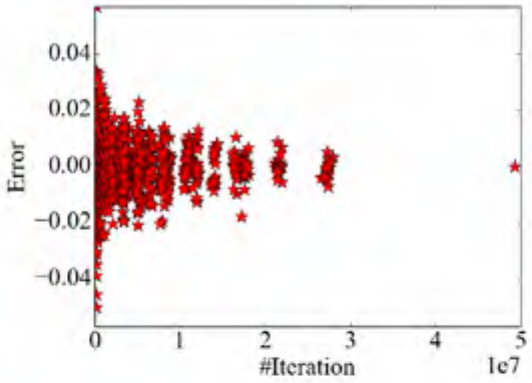


Incomplete subgraph

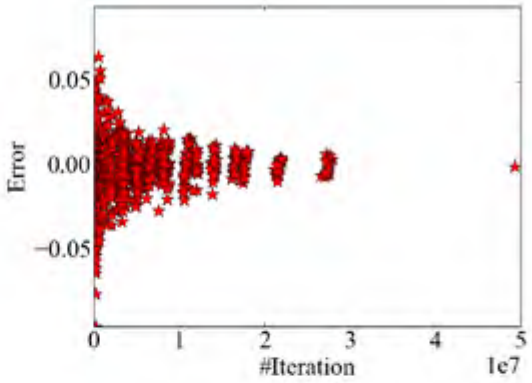
Results



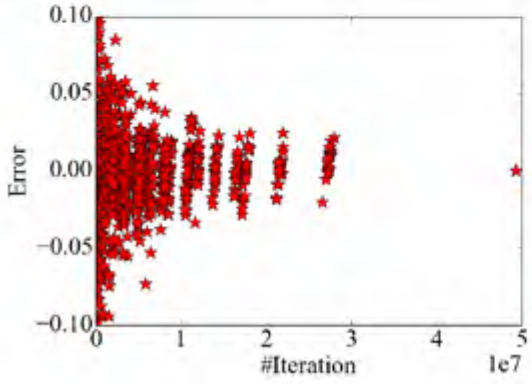
(a) C_1



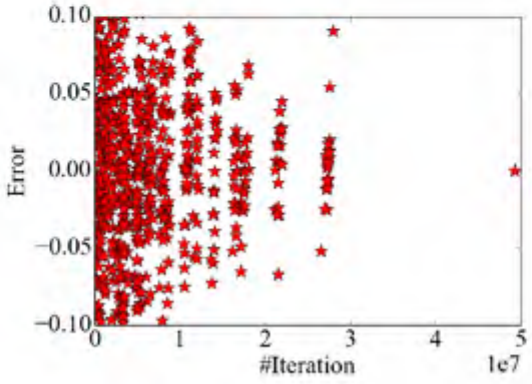
(b) C_2



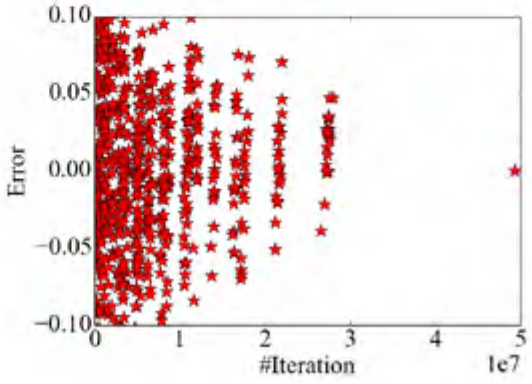
(c) C_3



(d) C_4



(e) C_{15}



(f) C_{20}

Sampling Variance and Time

- Variance:

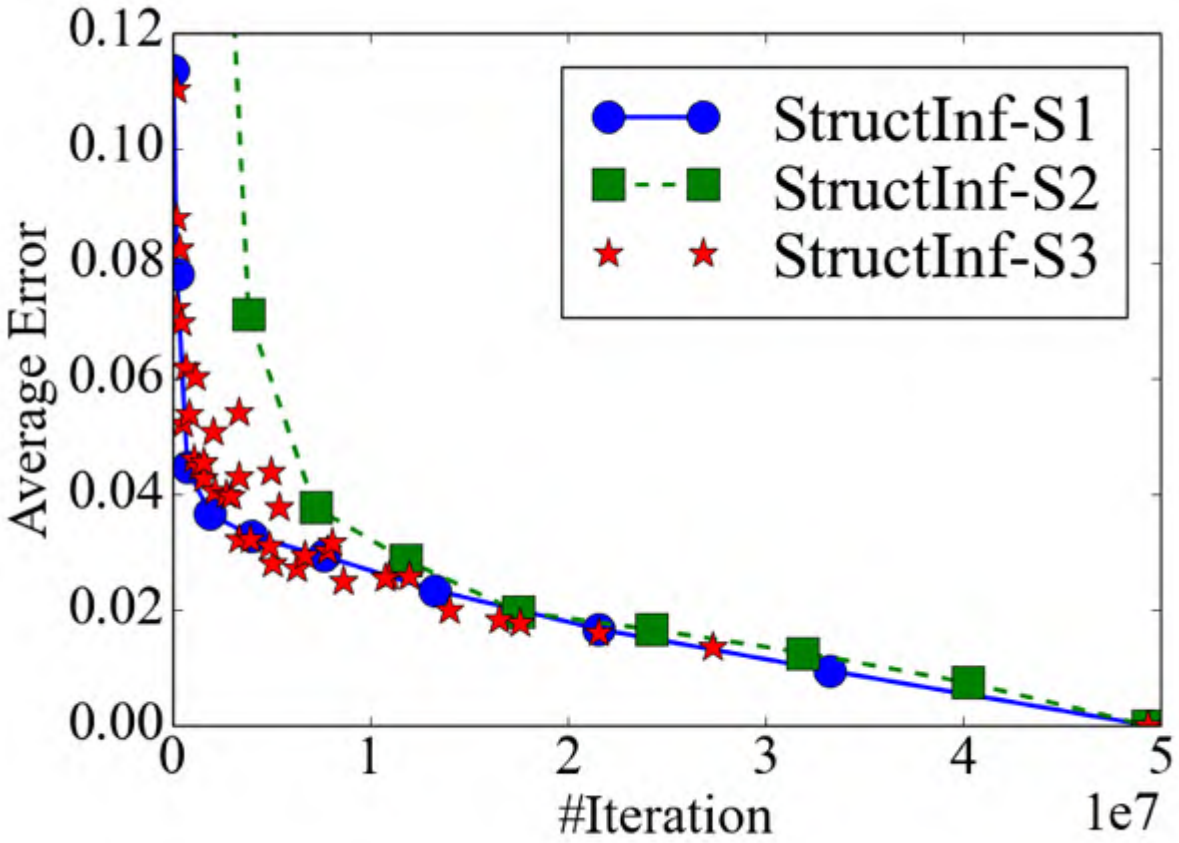
$$\widetilde{V}(\widetilde{x}_k) = \sum_{i=1}^{\hat{x}_k} \frac{1 - p(c_i^k)}{p^2(c_i^k)} + \sum_{i \neq j}^{\hat{x}_k} \frac{p(c_i^k c_j^k) - p(c_i^k)p(c_j^k)}{p(c_i^k c_j^k)p(c_i^k)p(c_j^k)}$$

- The higher the sampling probability $p(c_i^k)$, the smaller the variance will be, while the sampling speed will be slower.

- Trade off error and time by ***p*** and ***q***

Results

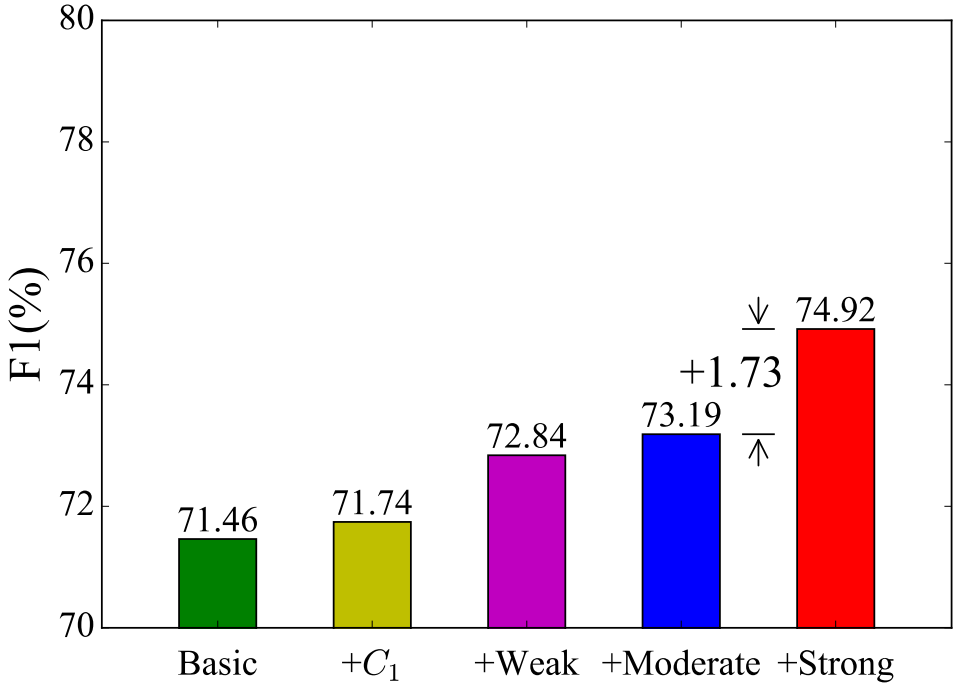
Varying the probabilities ***p*** and ***q***.



Weibo dataset
1,787,443 nodes
413,503,687 edges
20,134,307 actions

- StructInf-S1 performs better than StructInf-S2
- StructInf-S3 is less sensitive to the parameters

Application: Retweet Prediction



Basic: #friends, gender, status, etc.

C_1 : the number of active neighbors

Weak: $\widetilde{\mathrm{IP}}_k < 0.1$

Moderate: $0.1 \leq \widetilde{\mathrm{IP}}_k < 0.3$

Strong: $\widetilde{\mathrm{IP}}_k > 0.3$

Summary

- Node conformity influence
  - People with higher degree and lower clustering coefficient are more likely to conform to others.
- Pairwise link influence
  - A two-way relationship between two users can trigger more links than a one-way relationship.
- Group influence
  - Structural diversity
    - The probability of a user retweeting a tweet is negatively correlated with structural diversity of the active neighbors.
  - Structural influence
    - Sampling algorithms can achieve a 10 speedup compared to the exact influence pattern mining algorithm

Future Work

- How to design a diffusion model that considers different kinds of influence together?
- What's the difference between influence in different kinds of social medias?
- How to leverage different kinds of influence to do social recommendation?

Code & Dataset

- Conformity Influence on “wording” behavior
  - http://arnetminer.org/roleconformity
  - Jing Zhang, Jie Tang, Honglei Zhuang, Cane Wing-Ki Leung and Juanzi Li**.** Role-aware Conformity Influence Modeling and Analysis in Social Networks. In AAAI’14. pp. 1-7
- Link Influence
  - http://cs.aminer.org/followinf
  - Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Diffusion of “Following” Links in Microblogging Networks. IEEE Transaction on Knowledge and Data Engineering (TKDE)
- Structural influence
  - http://arnetminer.org/influencelocality
  - Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social Influence Locality for Modeling Retweeting Behaviors. In IJCAI'13. pp. 2761-2767.
  - https://cn.aminer.org/structinf
  - Jing Zhang, Jie Tang, Yuanyi Zhong, Yuchen Mo, Jimeng Sun, and Juanzi Li. StructInf: Mining Structural Influence from Social Streams. In AAAI’17.

中国人民大學

RENMIN UNIVERSITY OF CHINA

Thank You

http://info.ruc.edu.cn/academic_professor.php?teacher_id=163
https://scholar.google.com/citations?user=T7Wa3GQAAAAJ&hl=en