Social Influence Analysis in Large Social Networks

Jie Tang

Tsinghua University

Collaborate with

Jimeng Sun (*IBM*), Wei Chen (*MSRA*)

John Hopcroft (*Cornell*), Jiawei Han and Chi Wang (*UIUC*)

Tiancheng Lou, Jing Zhang, Zhanpeng Fang, Lu Liu (*THU*)

说几个数字给您听…

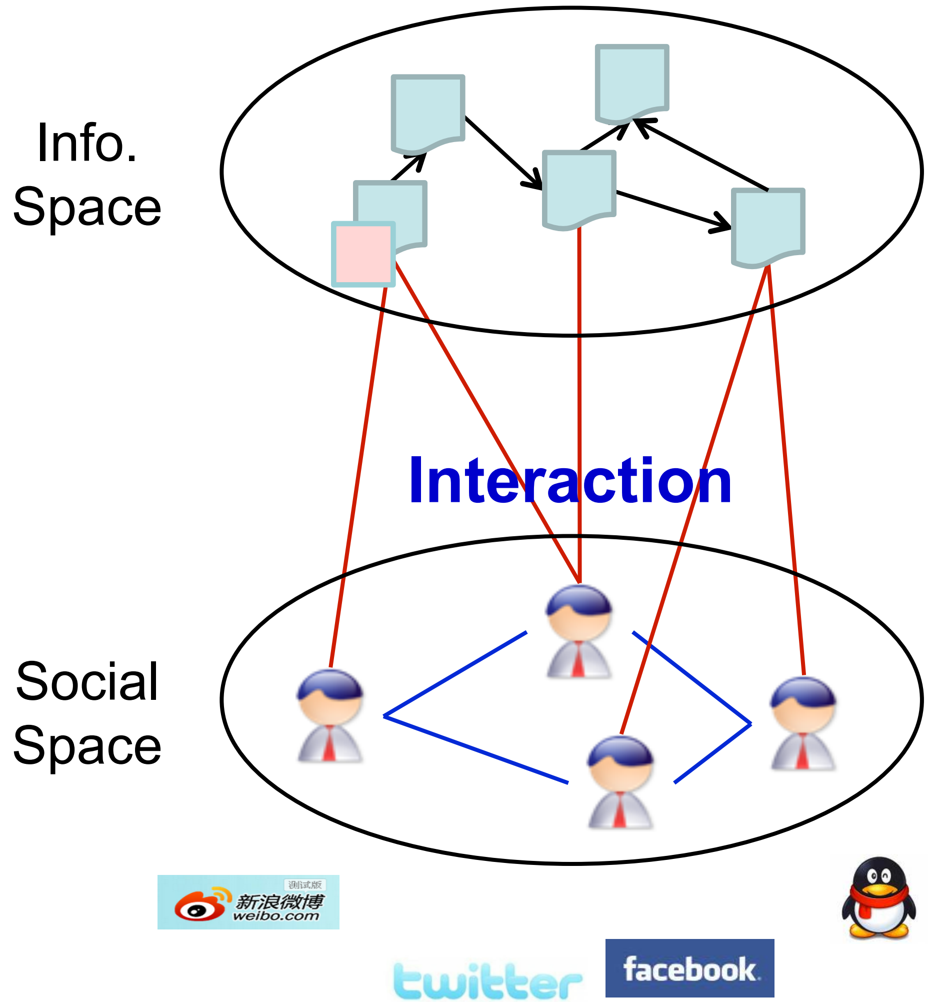
- >**1000 million** users
- The **3**rd largest "Country" in the world
- More visitors than Google
- >**721 million** users
- 2012, **400 million** users, 300% yearly increase

- 2009, **2 billion** tweets per quarter
- 2010, **4 billion** tweets per quarter
- 2011, **25 billion** tweets per quarter
- More than **6 billion** images
- Pinterest, with a traffic higher than Twitter and Google

再说几个数字给您听...

- Per Second: Email
  - 2.9 million emails per second
- Per Minute: Election Night 2012
  - a peak of 327,452 Tweets per minute
- Per Month: Facebook
  - "Waste" 700 billion minutes per month
- →Big Data
  - **2.5×10^{18}** Byte (2.5 EB) data per day

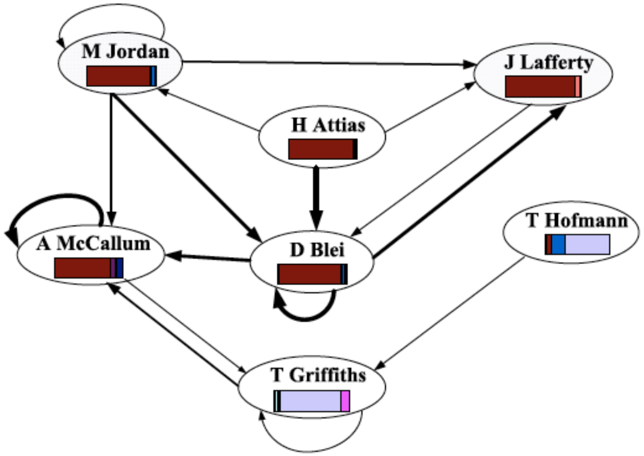
Social Web and Social Influence



1.Social Influence

2.Collective Intelligence

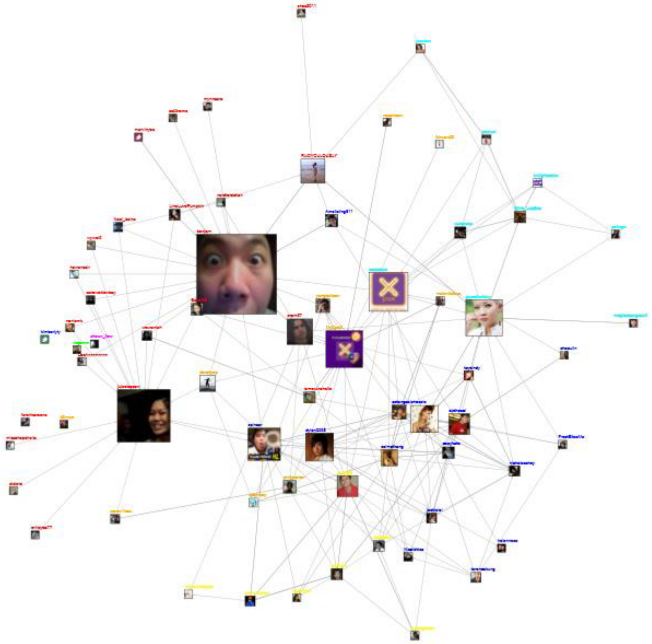
Social Influence



Author citation network

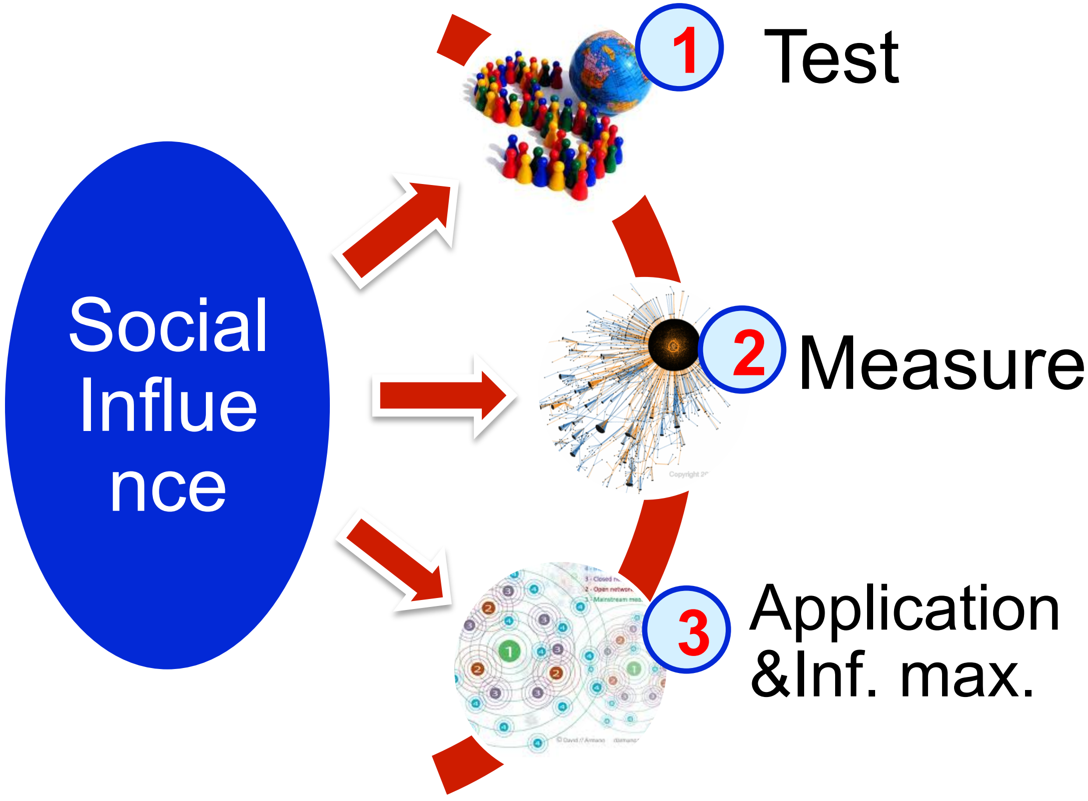
How researchers influence each other?

How people influences friends' following behaviors?



Twitter's following network

Social Influence

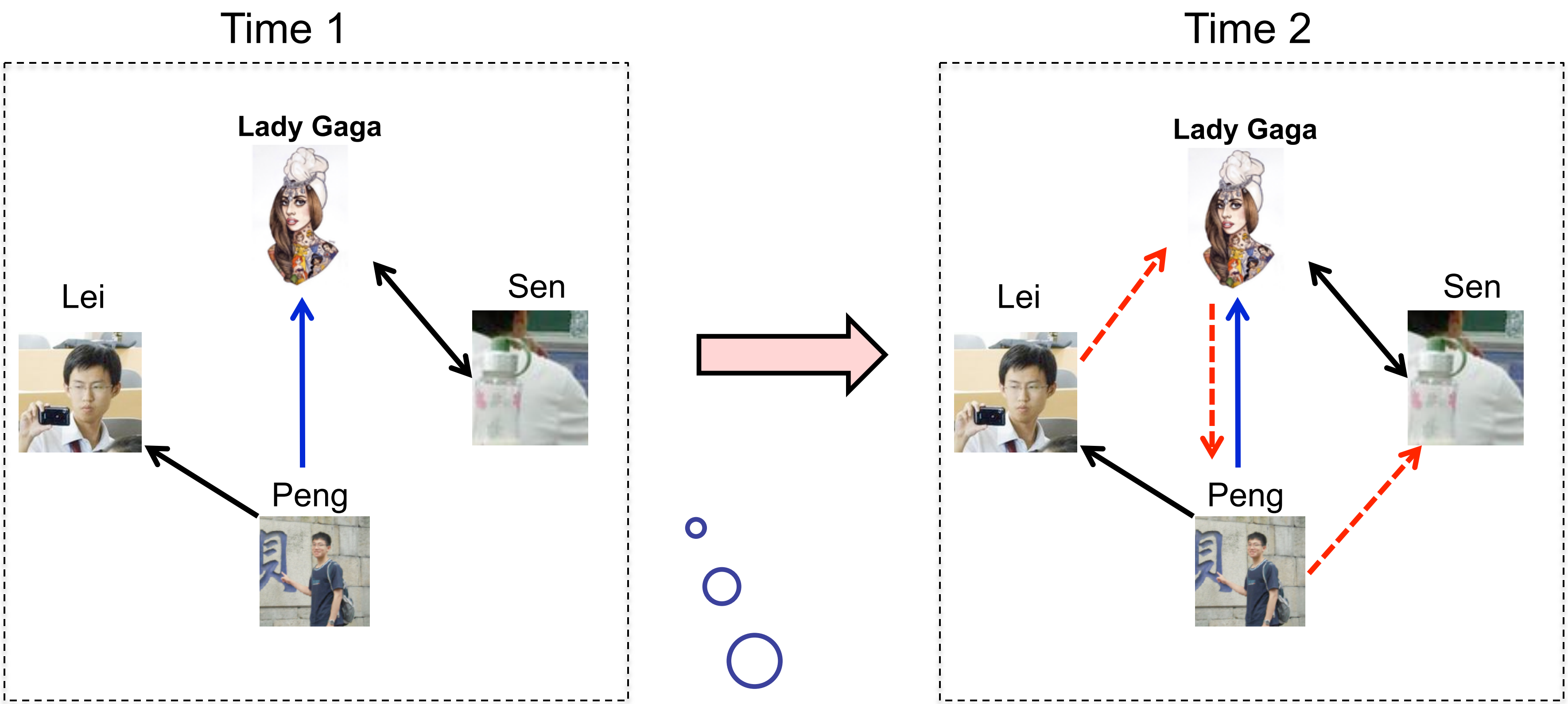


Following Influence Analysis

—A Case Study

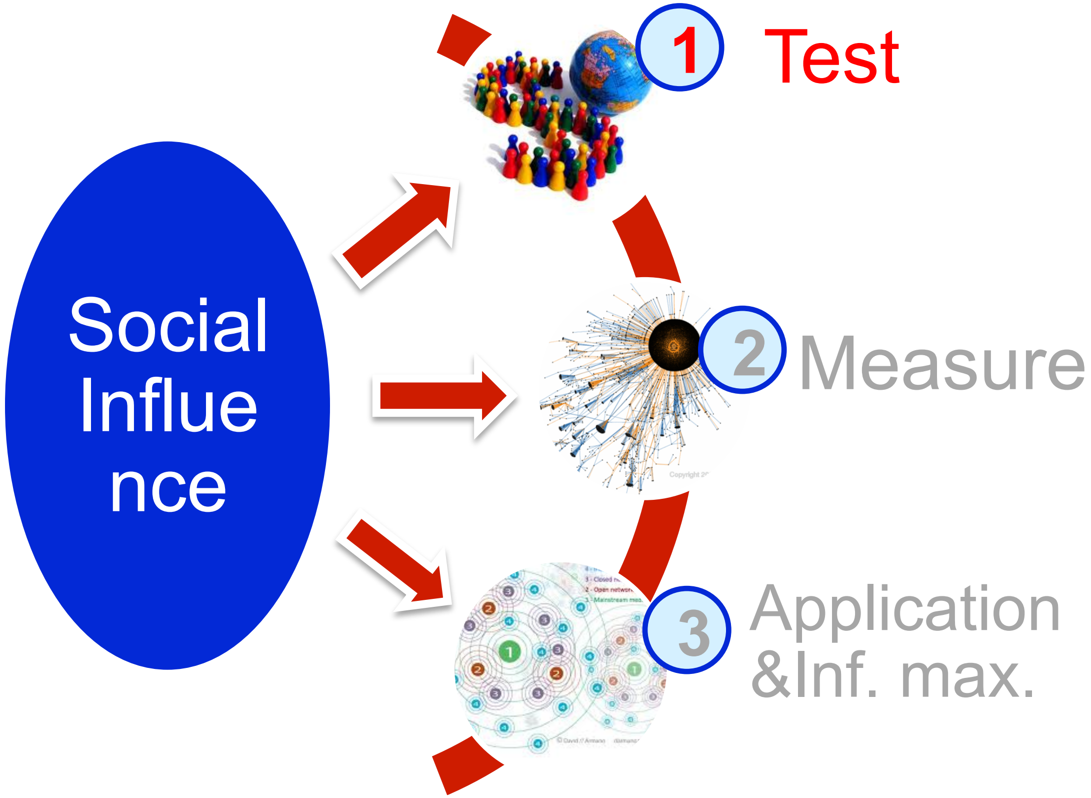
Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. ACM Transactions on Knowledge Discovery from Data (TKDD).

Following Influence on Twitter



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

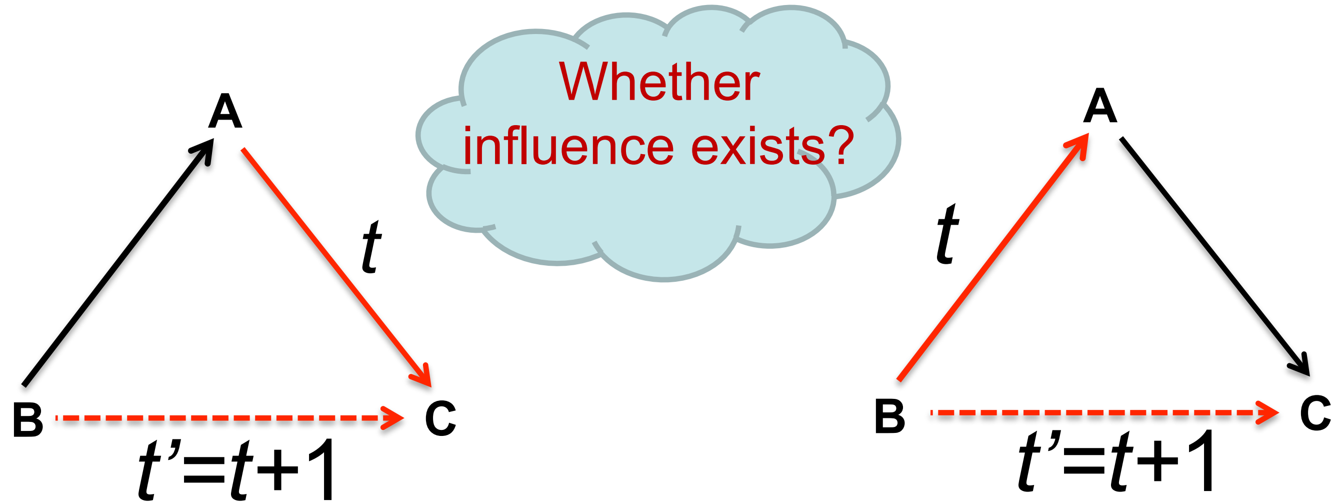
Social Influence



Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Social Influence on User Following Behaviors in Social Networks. (submitted)

Influence Test via Triad Formation

Two Categories of Following Influences



Follower diffusion

Followee diffusion

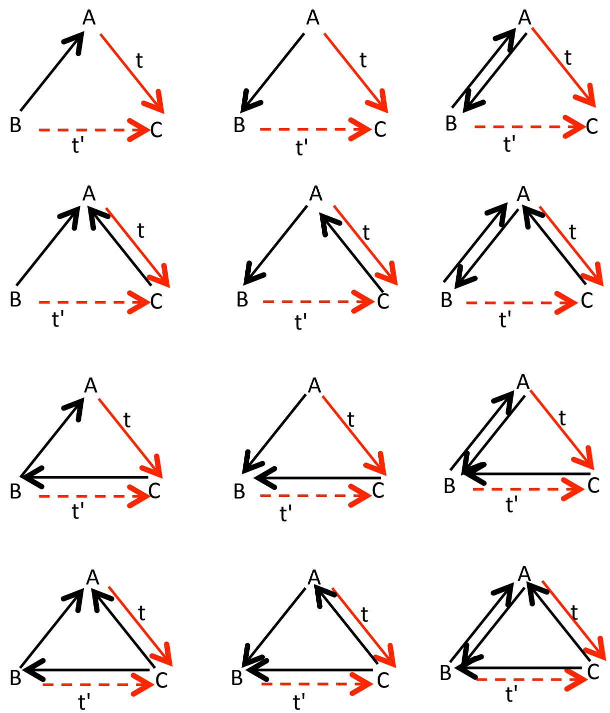
–>: pre-existed relationships

–>: a new relationship added at *t*

-->: a possible relationship added at *t*+1

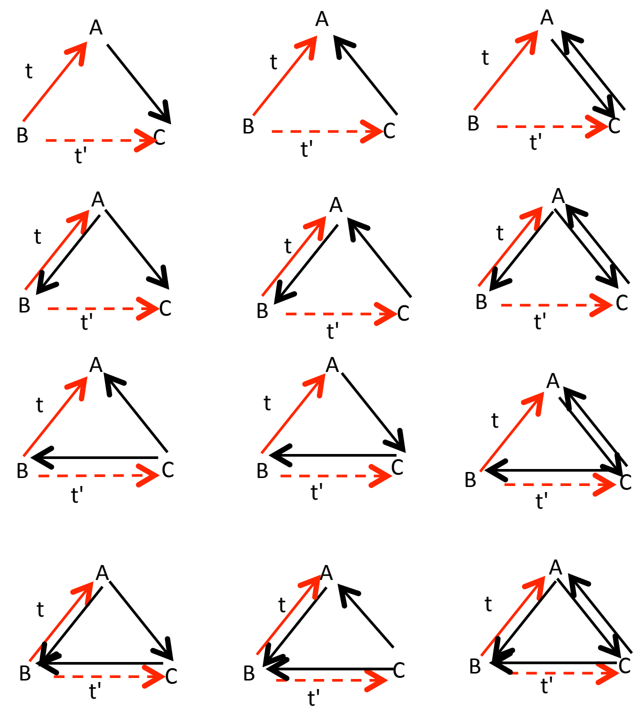
24 Triads in Following Influence

Follower diffusion



12 triads

Followee diffusion



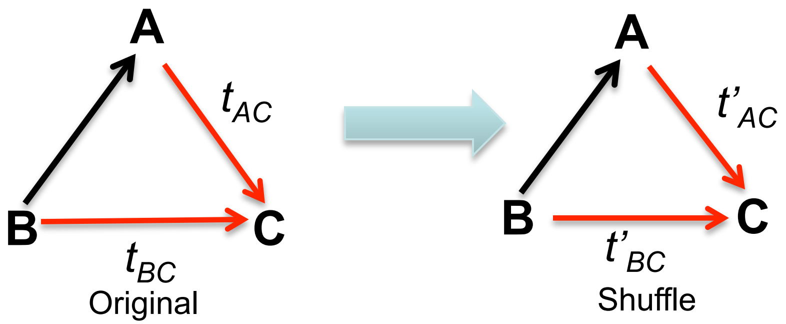
12 triads

Twitter Data

- Twitter data
  - "Lady Gaga" -> 10K followers -> millions of followers;
  - 13,442,659 users and 56,893,234 following links.
  - 35,746,366 tweets.
- 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

Test 1: Timing Shuffle Test

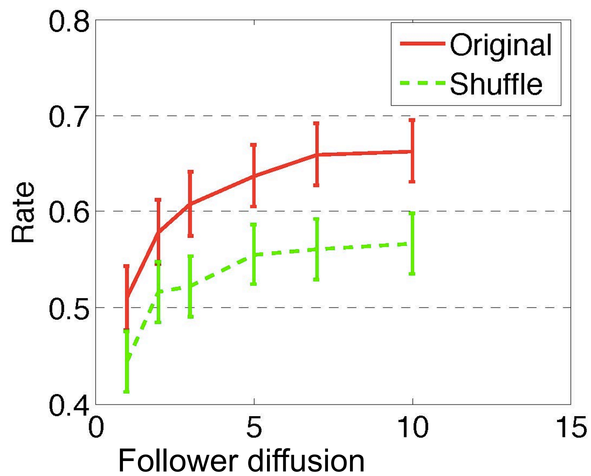
- Method: Shuffle the timing of all the following relationships.

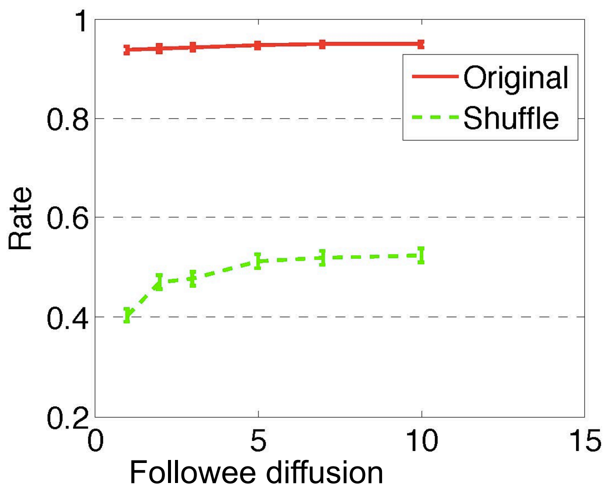


  - Compare the rate under the original and shuffled dataset.

$$Rate = \frac{\#Triad \mid 0 < t_{BC} - t_{AC} < \delta}{\#Triad \mid t_{BC} \text{ and } t_{AC} \text{ exist}}$$

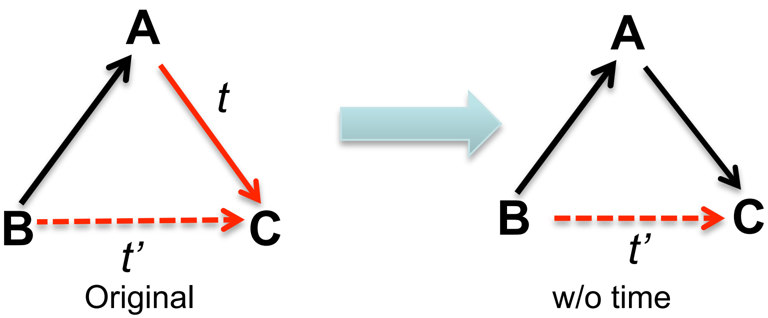
- Result





Test 2: Influence Decay Test

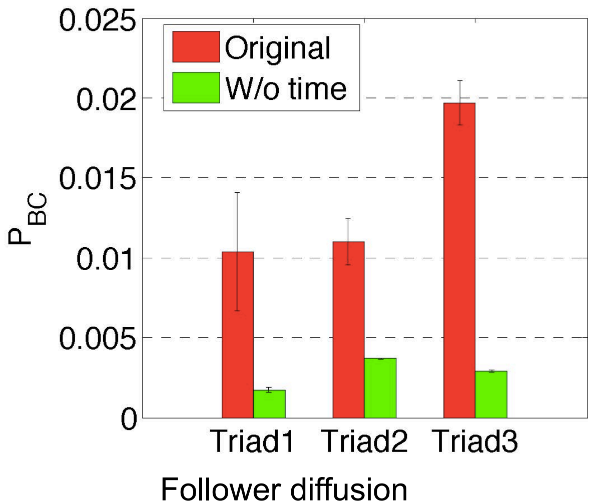
- Method: Remove the time information *t* of AC

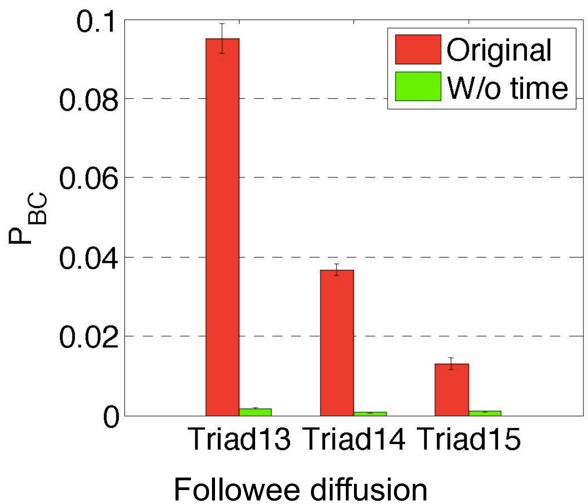


  - Compare the probability of B following C under the original and w/o time dataset.

$$P_{BC} = \frac{\#Triad \mid B \quad follows \quad C}{\#Triad}$$

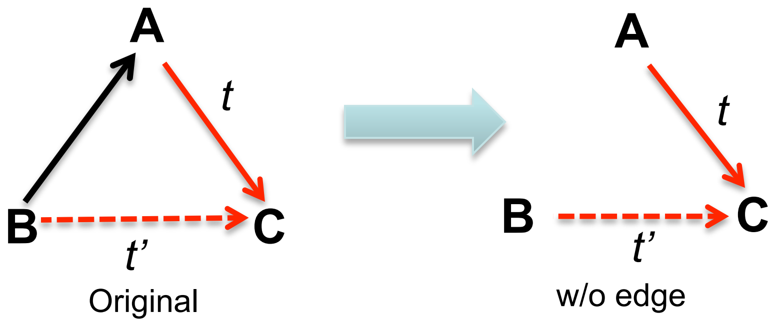
- Result





Test 3: Influence Propagation Test

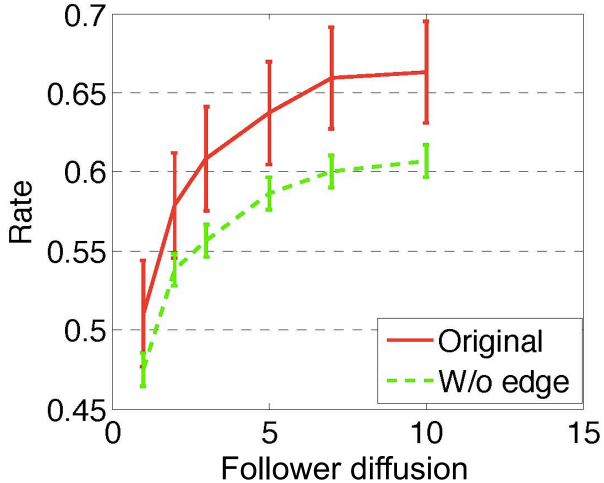
- Method: Remove the relationship between A and B.

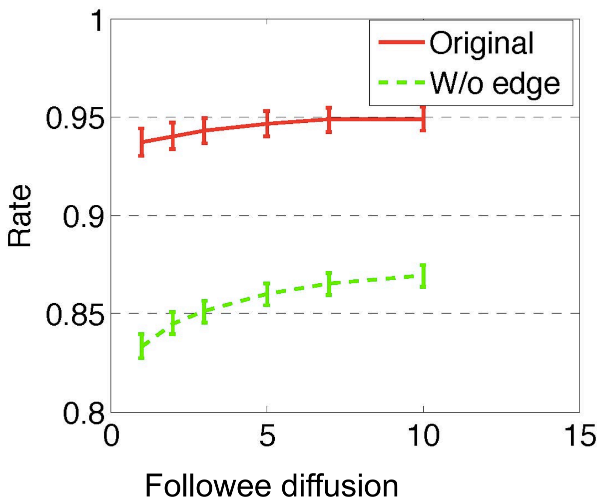


- Compare the rate under the original and w/o edge dataset.

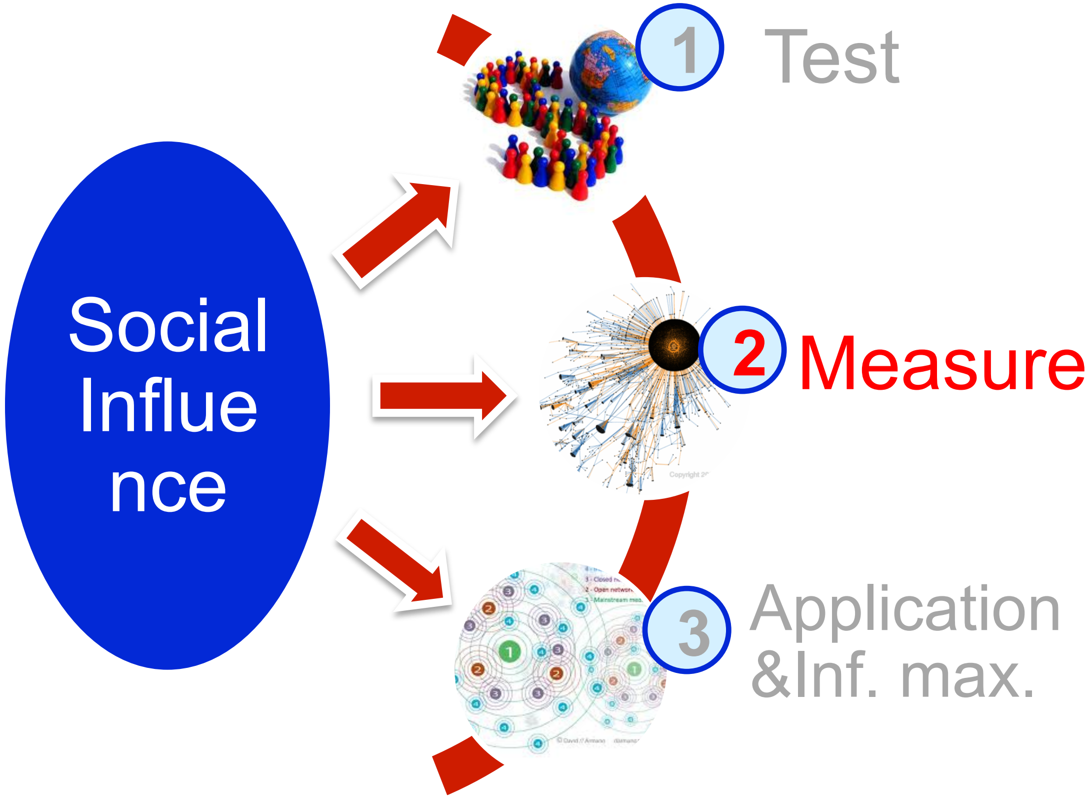
$$Rate = \frac{\#Triad \quad | \quad 0 < t_{BC} - t_{AC} < \delta}{\#Triad \quad | \quad t_{BC} \quad and \quad t_{AC} \quad exist}$$

- Result

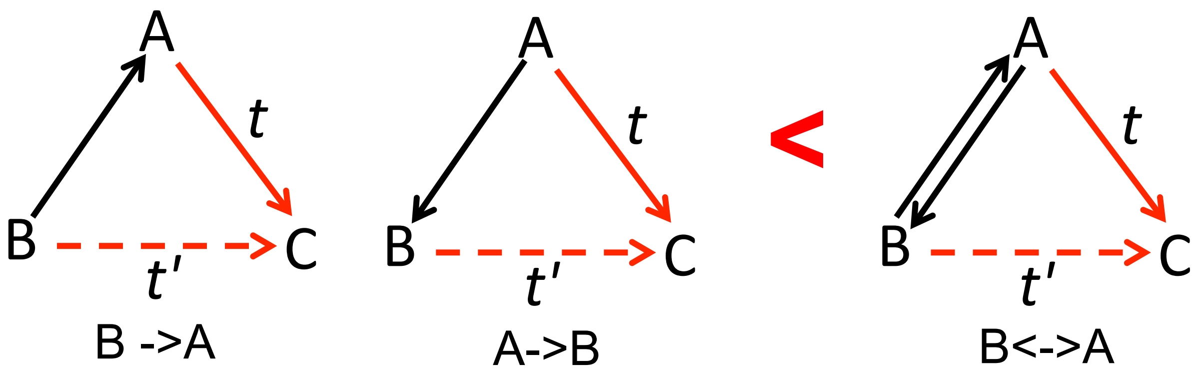


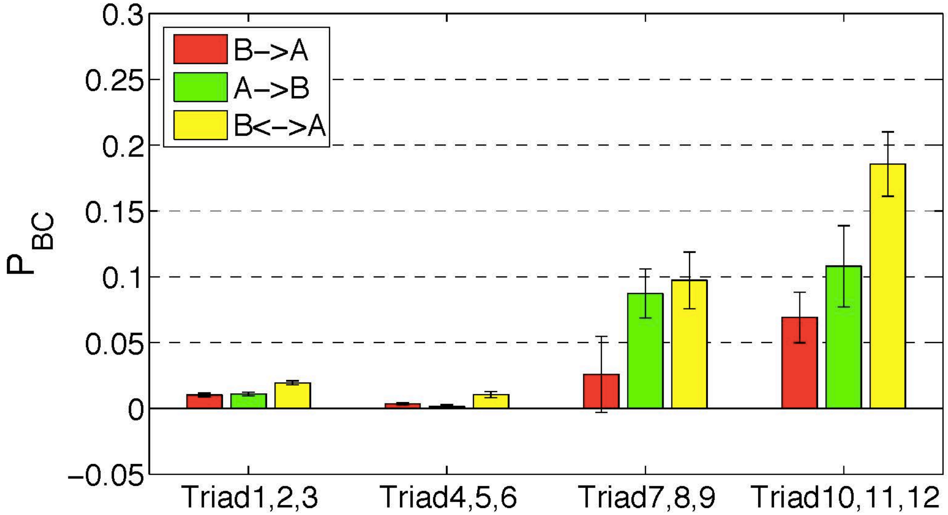


Social Influence



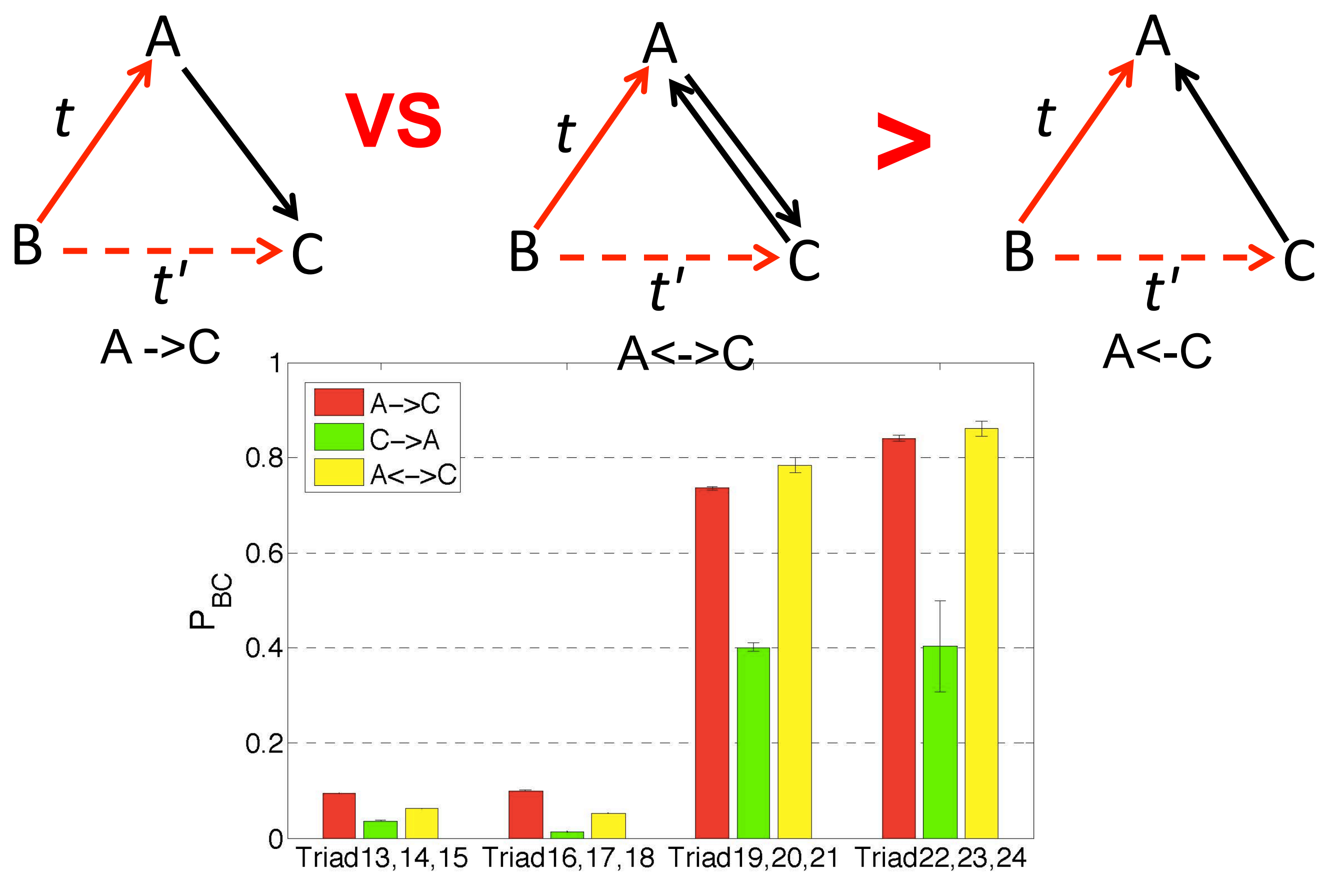
Follower Diffusion: Power of Reciprocity





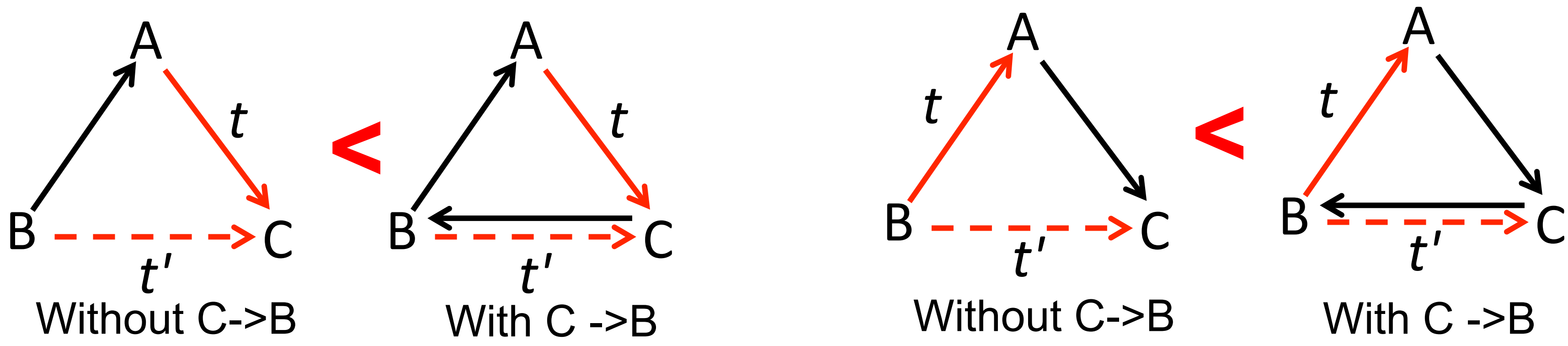
Observation: Following influence is more significant when there is a **reciprocal** relationship between B and A.
Explanation: "intimacy" is one of the three key factors that can increase people's likelihood to respond to social influence(social impact theory)

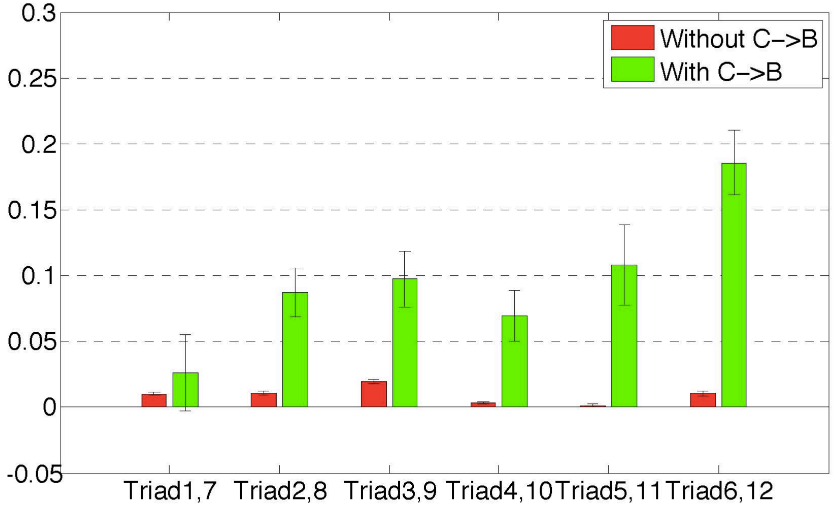
Followee Diffusion: One-way Relationship

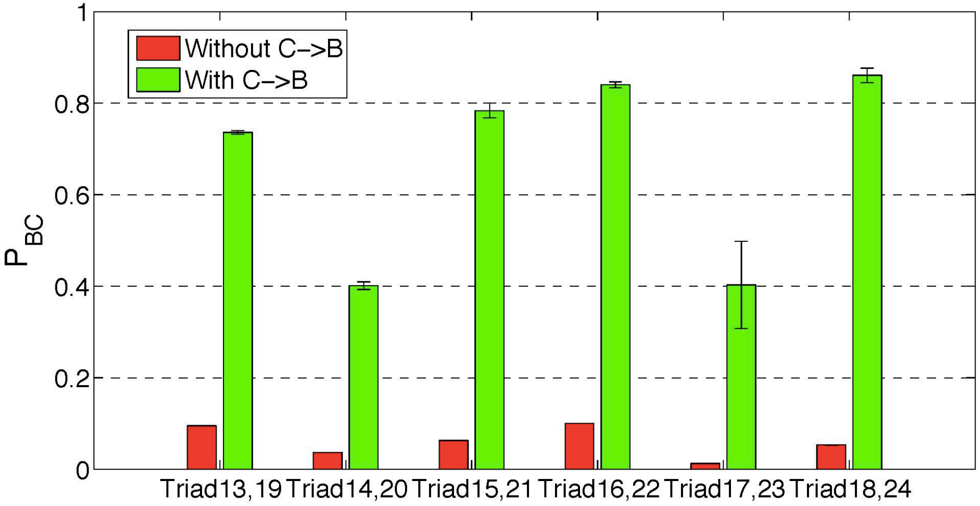


Observation: Following influence is more significant when there is a one-way relationship from A to C.
Explanation: Users usually prefer to check their followee's followees, from whom they select those they may be interested to follow.

Reversed Relationship



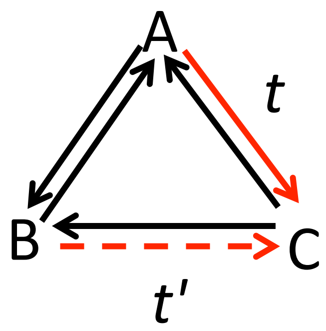




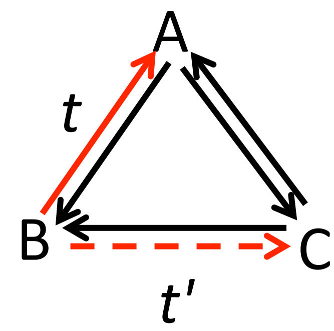
Observation: Following influence is more significant when there is a reversed relationship from C to B.

Explanation: Users are highly encouraged to follow their followers.

Social Theories: Structural Balance[1]



Follower diffusion



Followee diffusion

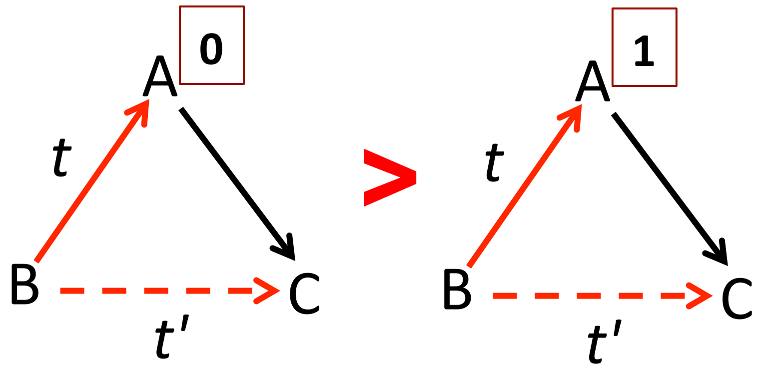
Social Balance: my friend's friend is also my friend
The probabilities of B following C in the two triads are higher than others in their respective categories.

Explanation: Users have tendency to form a balanced triad

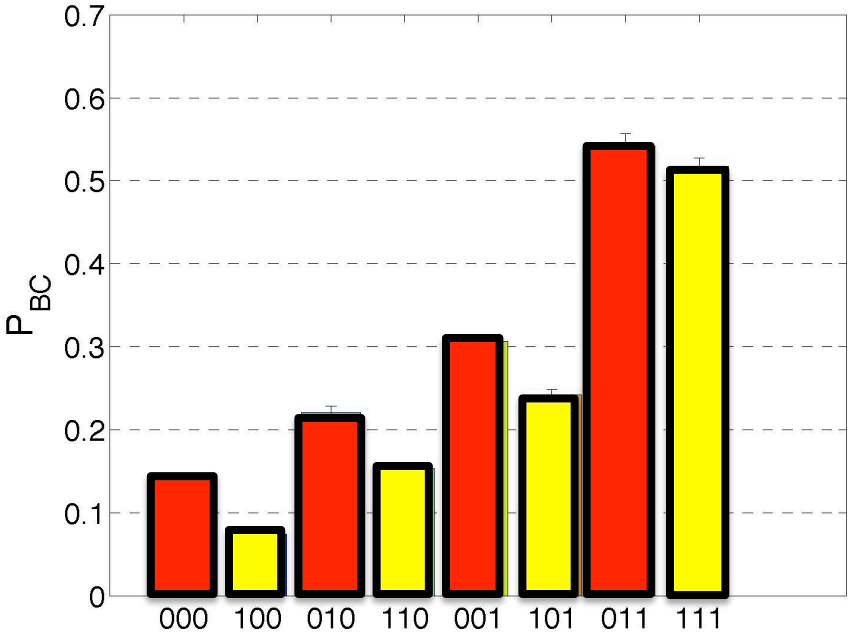
Fritz Heider (1958). The Psychology of Interpersonal Relations. John Wiley & Sons.

Social Theories: Social Status

Followee diffusion: P(0XX) > P(1XX)



1: Elite user
0: Low-status user

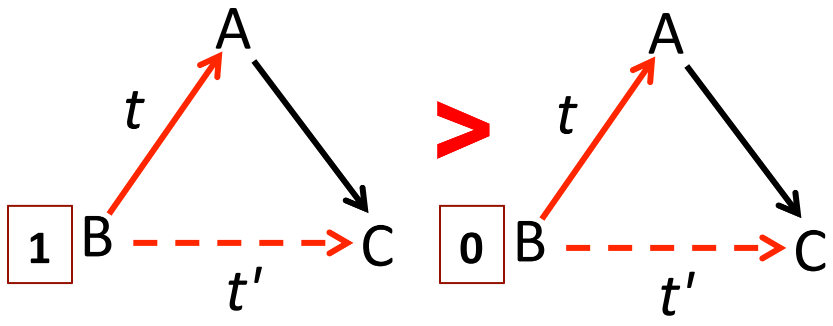


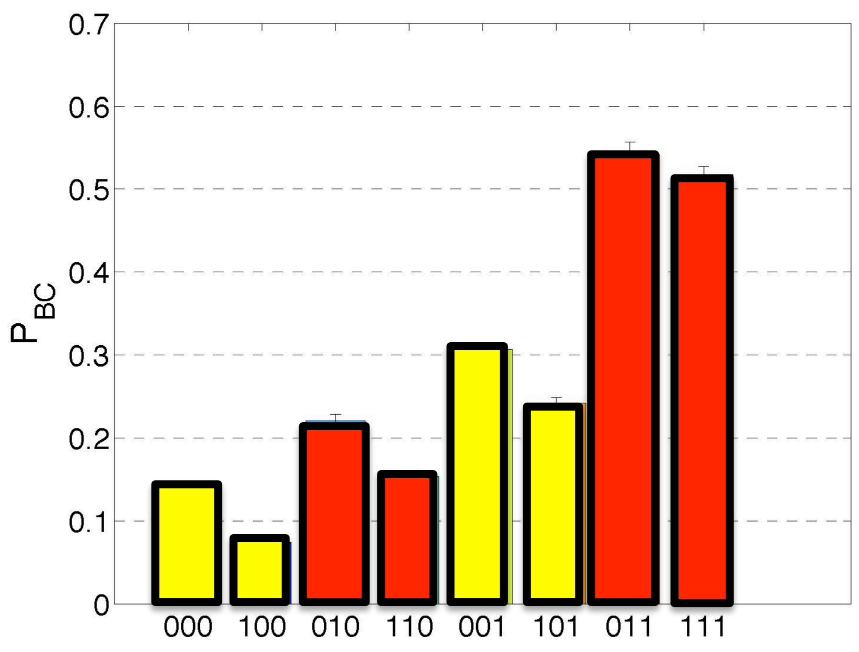
- Low-status users act as a bridge to connect users so as to form a closure triad.
- The likelihood of 0XX is 1.4 times of 1XX.

Social Theories: Social Status

Followee diffusion: P(X1X) > P(X0X)

1: Elite user
0: Low-status user



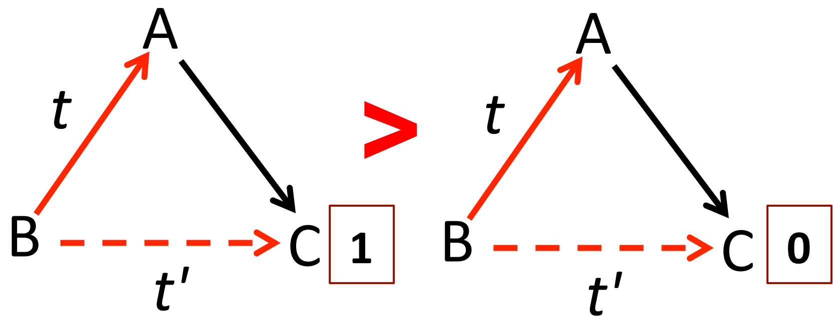


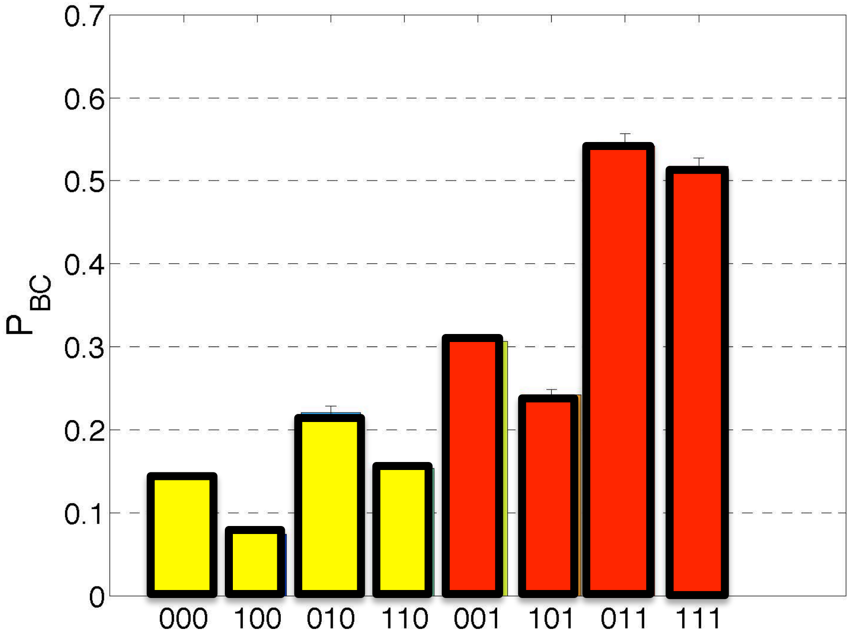
- Elite users play a more important role to form the triadic closure.
- The likelihood of X1X is almost double the probability of X0X.

Social Theories: Social Status

Followee diffusion: P(XX1) > P(XX0)

1: Elite user
0: Low-status user



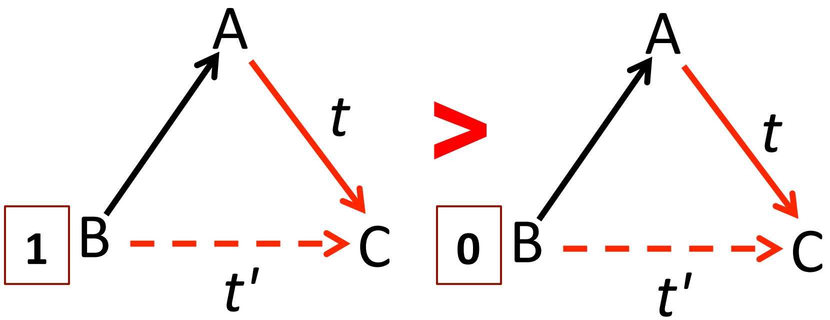


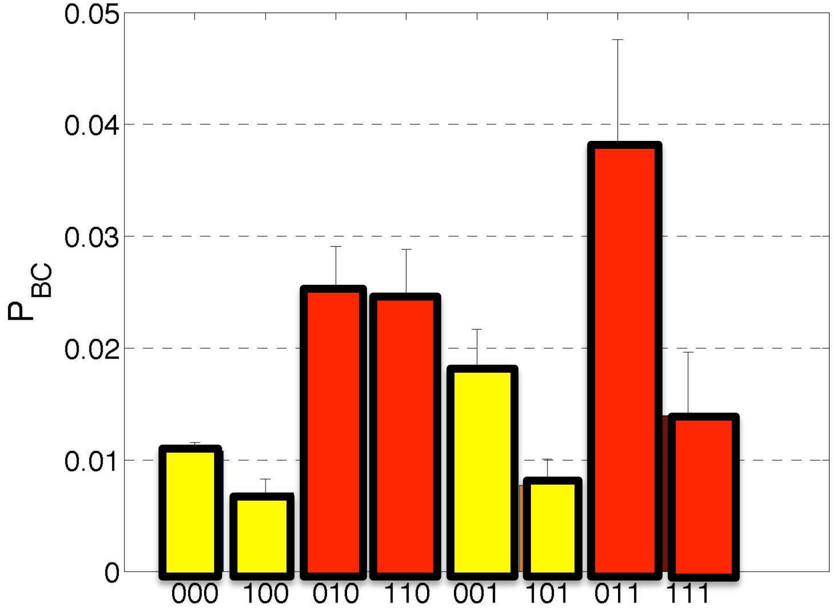
- The rich gets richer.
- The likelihood of XX1 is nearly 2 times higher than that of XX0.
- This phenomenon validates the mechanism of preferential attachment.

Social Theories: Social Status

Follower diffusion: P(X1X) > P(X0X)

1: Elite user
0: Low-status user

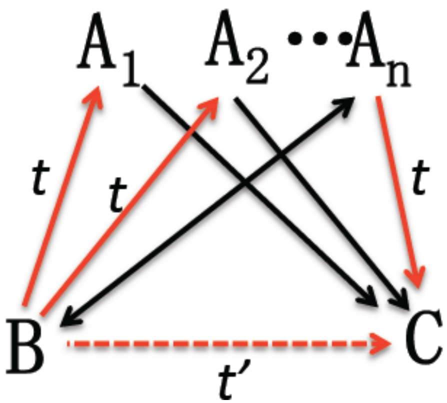




- Elite users play a more important role to form the triadic closure.
- The likelihood of X1X is almost double the probability of X0X.

Influence Learning Model

The formation of one following edge at time t' actually may be influenced by the formation of multiple neighbor edges e_{BA1} , e_{BA2} and e_{AnC} at time t.



We assume the neighbor edges activated at time t independently trigger a new edge.

The generative model FCM (Following cascade model)

$$L = \prod_{e' \in E} \boxed{\left(1 - \prod_{e \in N^-(e')} (1 - p_{ee'})\right)} \prod_{e \in E} \prod_{e' \in \neg N(e)} \boxed{(1 - p_{ee'})}$$

The formed edges

The unformed edges

$N(e)$	the neighbor edges of e activated before $t_e + \delta$
$N^-(e)$	the neighbor edges of e activated within $[t_e - \delta, t_e]$
$\neg N(e)$	the neighbor edges of e not activated within $[t_e, t_e + \delta]$
p_e	the probability of the formation of edge e
$p_{ee'}$	the influence probability of edge e on edge e'

Parameter Estimation

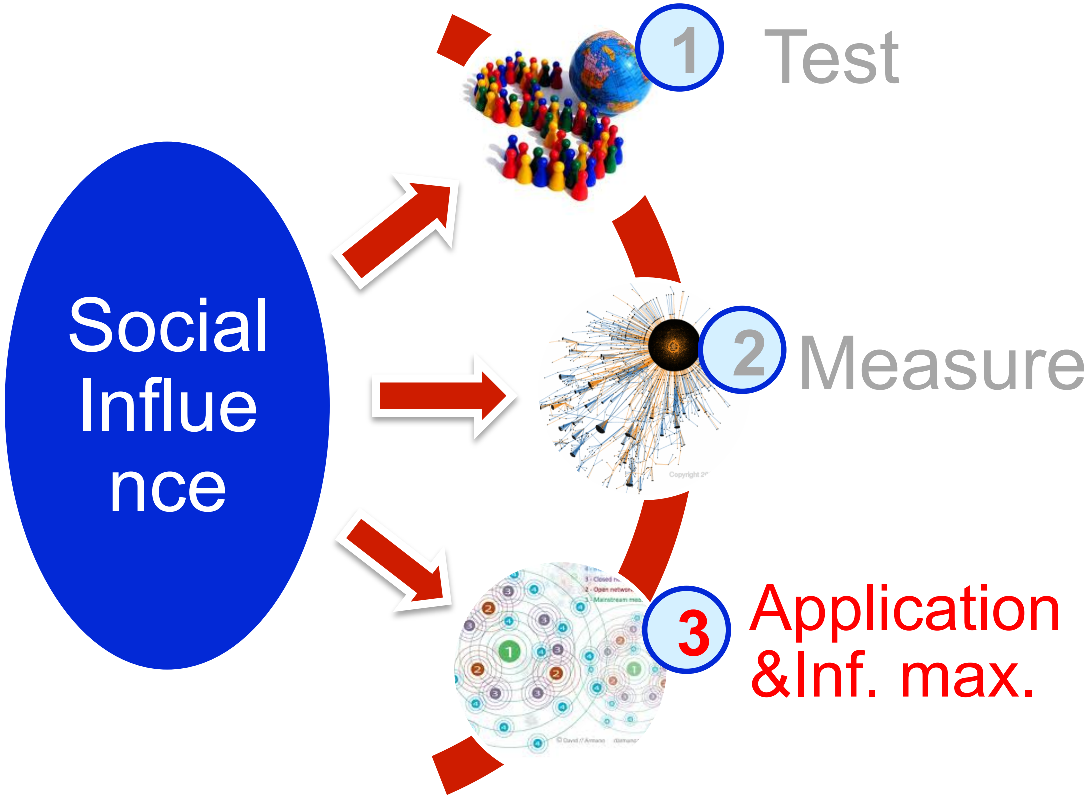
- We exact 24*8 features from the neighbor edges of each edge pair (e,e')
  - 24 triad structures and 8 triad statuses
- We aggregate different pairs with same features together and estimate the probabilities associated to 24*8 triads.

$$\boxed{\theta = \{p_{ee'}\}} \longrightarrow \boxed{\theta = \{p_{\triangle}\}}$$

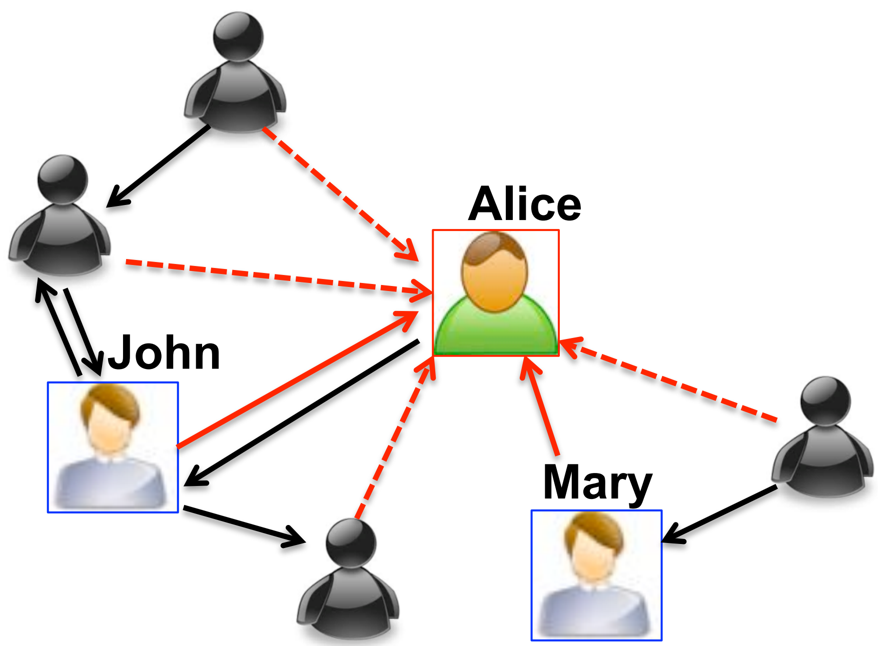
Input: network $G = (V, E, t)$
Output: $\theta = \{p_{\triangle}\}$
while *not converged* **do**
  E-step : calculate $p_{e'}$ using $p_{e'} = 1 - \prod_{e \in N^{-}(e')} (1 - \hat{p}_{\triangle})$
  M-step: calculate $p_{\triangle}$ using $p_{\triangle} = \frac{1}{|\triangle^A| + |\triangle^U|} \sum_{e' \in E_{\triangle}} \frac{\hat{p}_{\triangle}}{\hat{p}_{e'}}$
end

$\triangle$	the triad type associated with two edges
$p_{\triangle}$	the influence probability of the triad type $\triangle$
$\triangle^A$	the times of the triad type activated one edge
$\triangle^U$	the times of the triad type failed in activating one edge
$E_{\triangle}$	the edges activated by a triad type $\triangle$

Social Influence

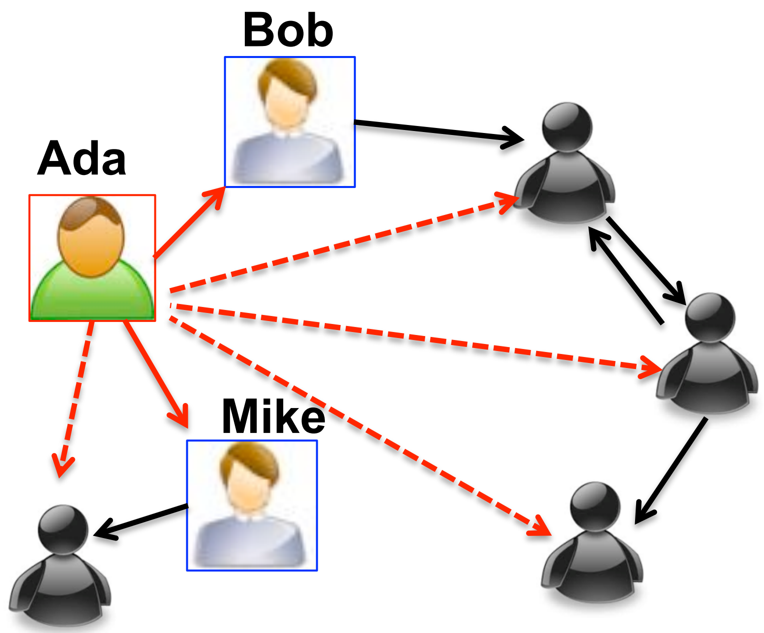


Applications: Influence Maximization



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

Applications: Friend Recommendation

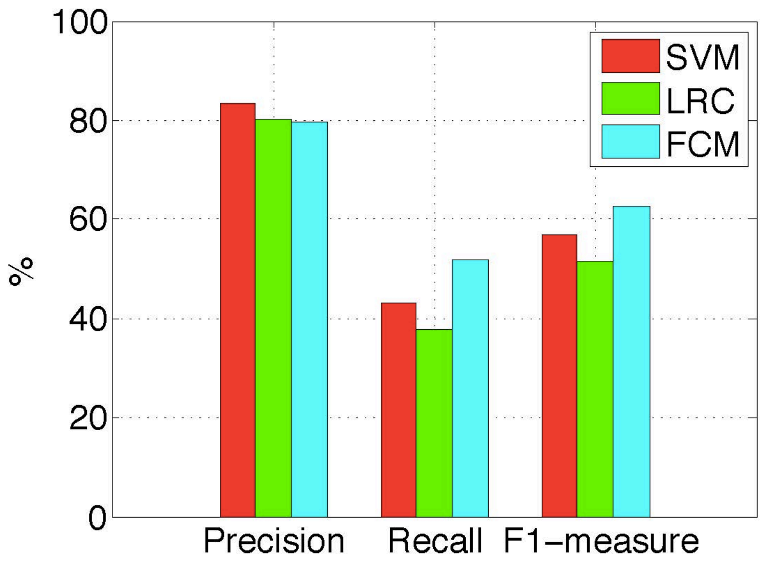


Find a set S of k initial followees for user v such that the total number of new followees accepted by v is maximized

Experiments

- Link Formation Accuracy
  - Link formation is used to verify the the influence probabilities learned by FCM.
  - A model has a good performance If it can best recover the process of link formation over time.
  - Link formation is modeled as both classification and ranking problem.
- Application improvement
  - Influence probabilities are applied to influence maximization and recommendation.

Link Formation Performance



Link formation as classification

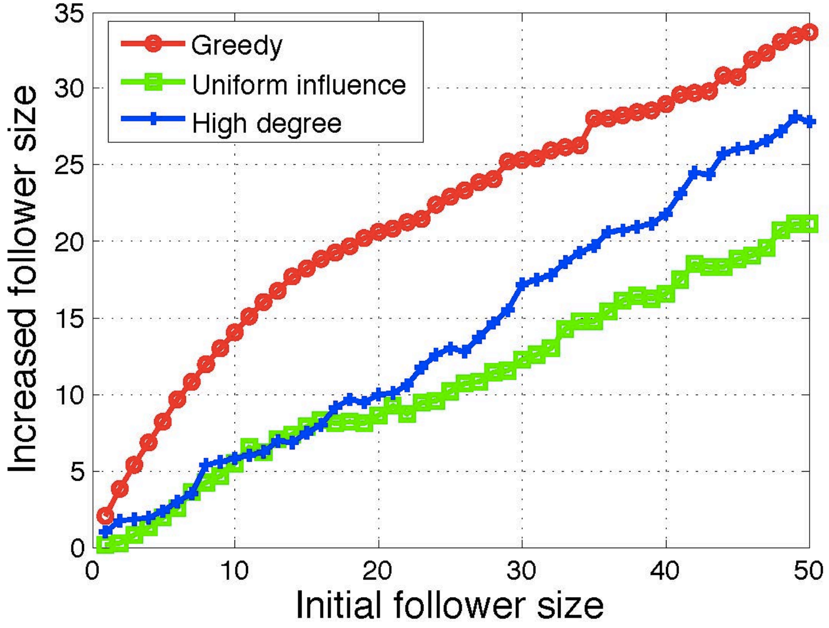
SVN, LRC, and FCM all use the same features except that FCM considers the diffusion process of following influence.

Model	P@1	P@2	P@5	P@10	MAP
CF	39.96	37.55	30.88	26.41	55.08
SimRank	26.35	26.06	26.22	24.39	44.15
Katz	46.24	41.84	32.77	26.61	59.40
FCM	72.88	55.69	37.15	27.88	77.91

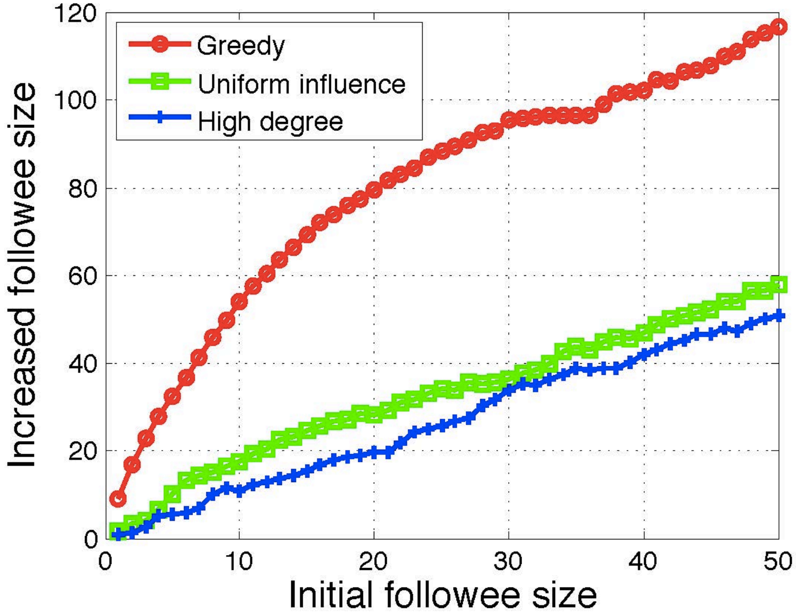
Link formation as ranking

CF, SimRank and Katz ignore the dynamic evolution of the network structure (e.g., an edge newly formed at t may trigger the neighbor edges at t').

Application Performance



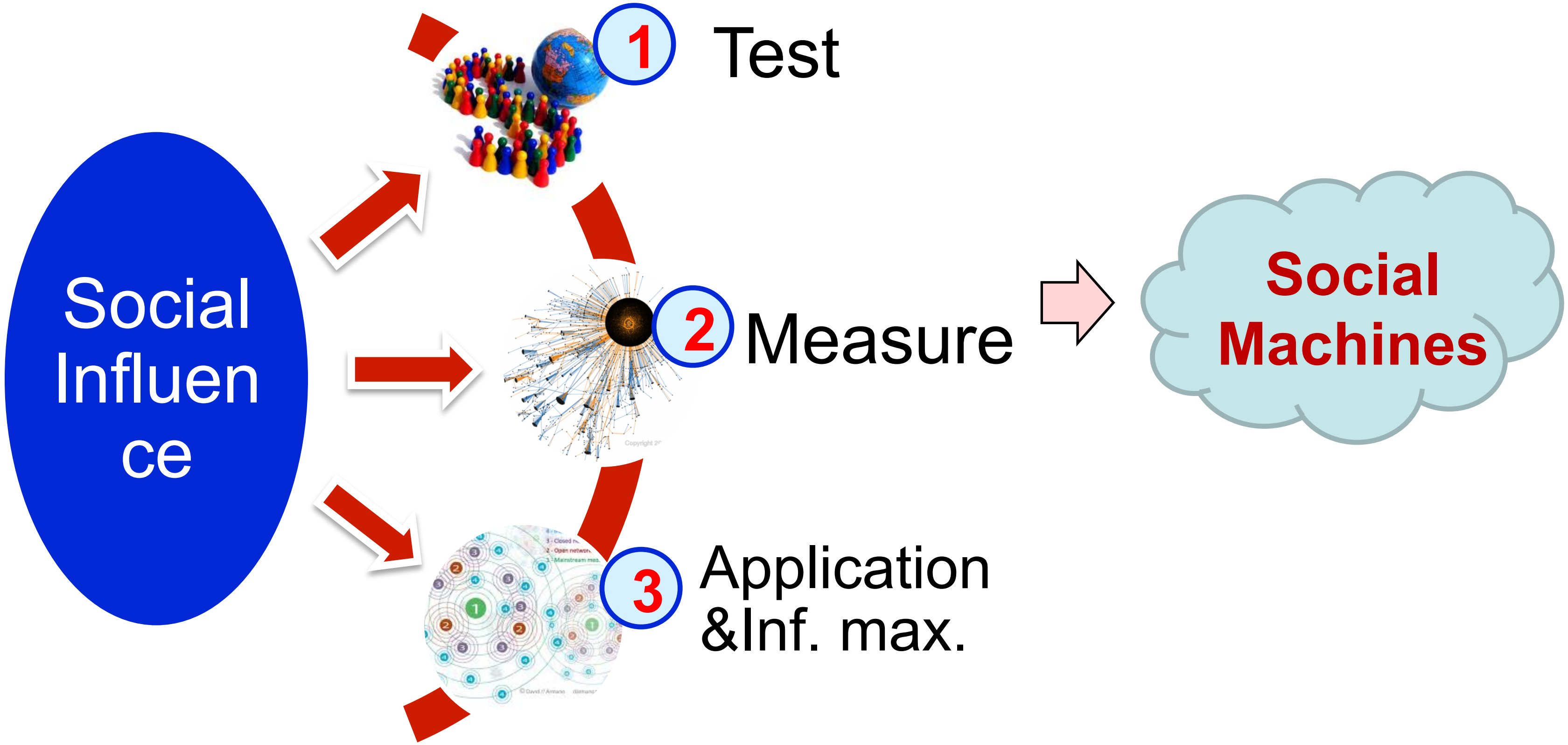
Influence Maximization



Recommendation

- High degree
    - May select the users that do not have large influence on following behaviors.
- Uniform configured influence
    - Can not accurately reflect the correlations between following behaviors.
- Greedy algorithm based on the influence probabilities learned by FCM
    - Captures the entire features of three users in a triad (i.e., triad structures and triad statuses)

Summaries


Social
Influen
ce
1
Test
2
Measure
3
Application
&Inf. max.
Social
Machines

“Social Machines”

- **Deploy** a “machine” on Weibo.com, the largest “Twitter” in China;
- **Act** as a person by auto follow/retweet/reply;
- **Attracted** thousands of fans.

Related Publications

- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**SIGKDD'2009**). pp. 807-816.
- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. ACM Transactions on Knowledge Discovery from Data (**TKDD**).
- Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Social Influence on User Following Behaviors in Social Networks. (submitted)
- Jimeng Sun and Jie Tang. A Survey of Models and Algorithms for Social Influence Analysis. In the book of Social Network Data Analysis. Charu C. Aggarwal (Ed.), Kluwer Academic Publishers, pages 177-214, 2011.
- Jimeng Sun and Jie Tang. Models and Algorithms for Social Influence Analysis. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (**WSDM** 2013). (Tutorial)
- Lu Liu, Jie Tang, Jiawei Han, and Shiqiang Yang. Learning Influence from Heterogeneous Social Networks. In Data Mining and Knowledge Discovery (**DMKD**), 2012, Volume 25, Issue 3, pages 511-544.

Thank you！

Data: http://arnetminer.org/download/

http://keg.cs.tsinghua.edu.cn/jietang/

Related Works

- Social Influence Testing
  - Randomized controlled trial [Bakshy,2012][Bond,2012]
  - Distinguish influence and homophily [Sinan, 2009]
  - Shuffle Test [Anagnostopoulos,2008]
- Social Influence Quantification
  - Directly count action number [Goyal, 2010]
  - Define likelihood function based on IC model [Myers, 2010][Gruhl,2004][Saito,2011]
- Influence Maximization
  - Algorithmic problem [Domingos, 2001]
  - Discrete optimization problem [Kempe, 2003]
  - Efficiency improvement [Leskovec, 2010][Chen, 2010]