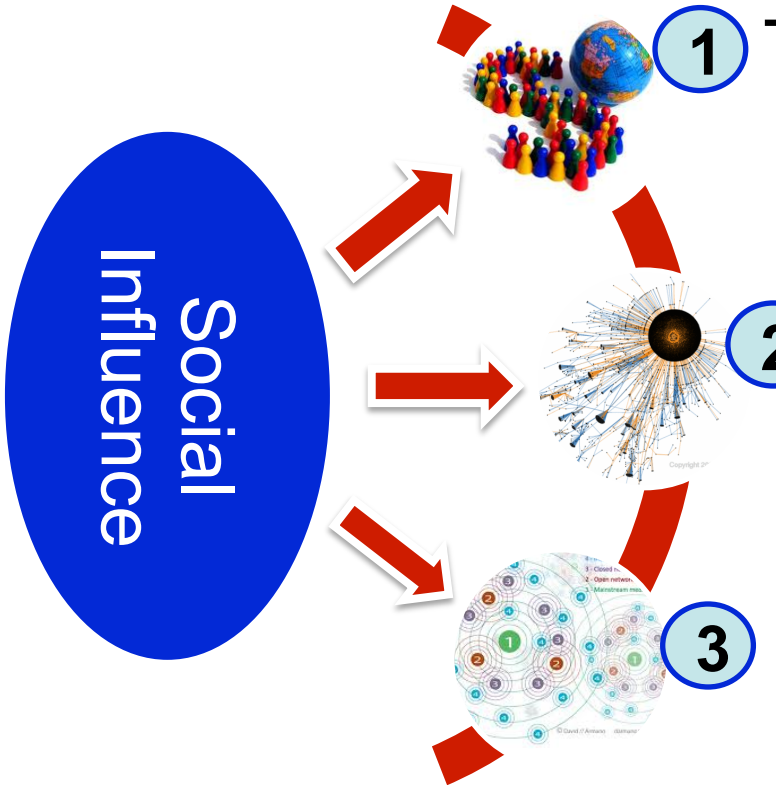
Models and Algorithms for Social Influence Analysis

Jie Tang and Jimeng Sun

Tsinghua University, China

IBM TJ Watson, USA

Agenda



1 Test

- Randomization test
- Shuffle test
- Reverse test

2 Measure

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

3 Models

- Linear Threshold Model
- Cascade Model
- Algorithms

Social Networks

- >**1000 million** users
- The **3rd** largest "Country" in the world
- More visitors than Google

- >**780 million** users

- 2013, **400 million** users, 40% 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

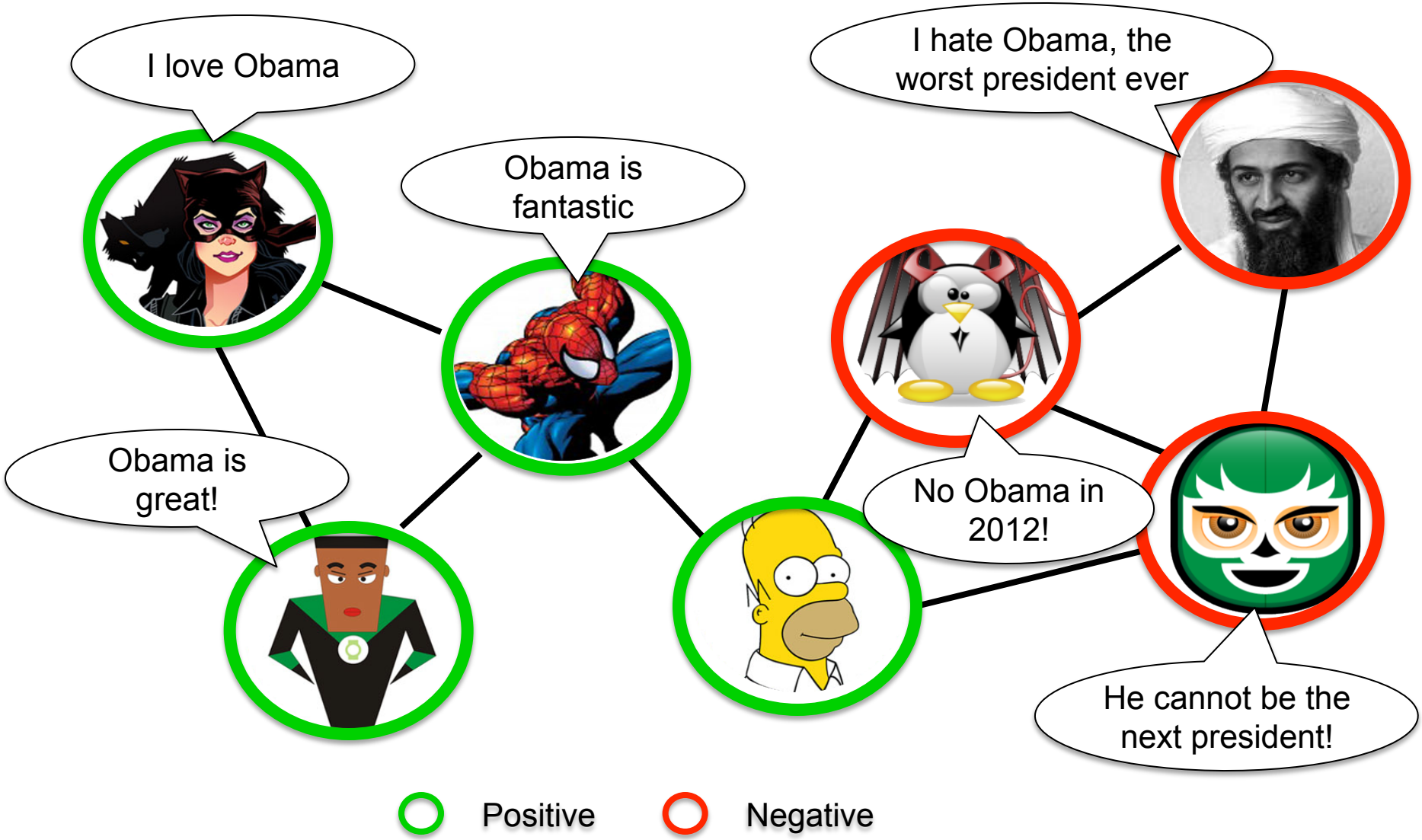
A Trillion Dollar Opportunity

Social networks already become a bridge to connect our daily **physical** life and the **virtual** web space

On2Off [1]

[1] Online to Offline is trillion dollar business
http://techcrunch.com/2010/08/07/why-online2offline-commerce-is-a-trillion-dollar-opportunity/

“Love Obama”


I love Obama
Obama is fantastic
I hate Obama, the worst president ever
Obama is great!
No Obama in 2012!
He cannot be the next president!
Positive
Negative

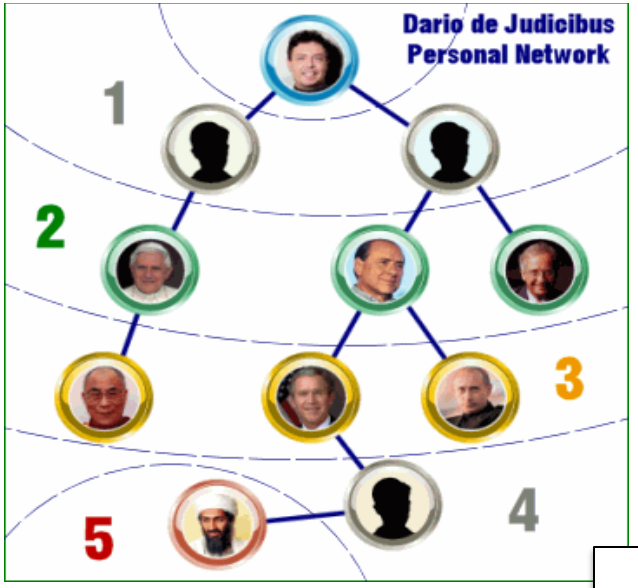
What is Social Influence?

- Social influence occurs when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally.[1]
  - **Informational social influence**: to accept information from another;
  - **Normative social influence**: to conform to the positive expectations of others.

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

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.

Does Social Influence really matter?

- Case 1: 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.
- Control group: did not receive any msg.

[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.

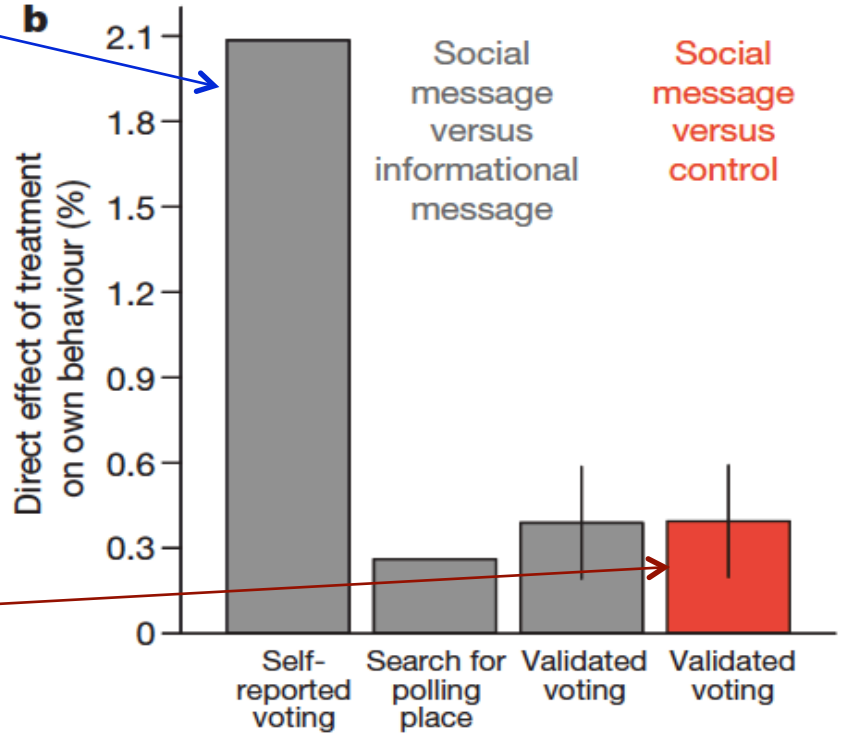
Case 1: Social Influence and Political Mobilization

Social msg group v.s.
Info msg group

Result: The former were 2.08% (*t*-test, $P<0.01$) more likely to click on the "I Voted" button

Social msg group v.s.
Control group

Result: The former were 0.39% (*t*-test, $P=0.02$) more likely to **actually vote** (via examination of public voting records)



[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.

Case 2: Klout[1]—Social Media Marketing

- Toward measuring real-world influence
  - Twitter, Facebook, G+, LinkedIn, etc.
  - Klout generates a score on a scale of 1-100 for a social user to represent her/his ability to engage other people and inspire social actions.
  - Has built 100 million profiles.
- Though controversial[2], in May 2012, Cathay Pacific opens SFO lounge to Klout users
  - A high Klout score gets you into Cathay Pacific's SFO lounge

[1] http://klout.com

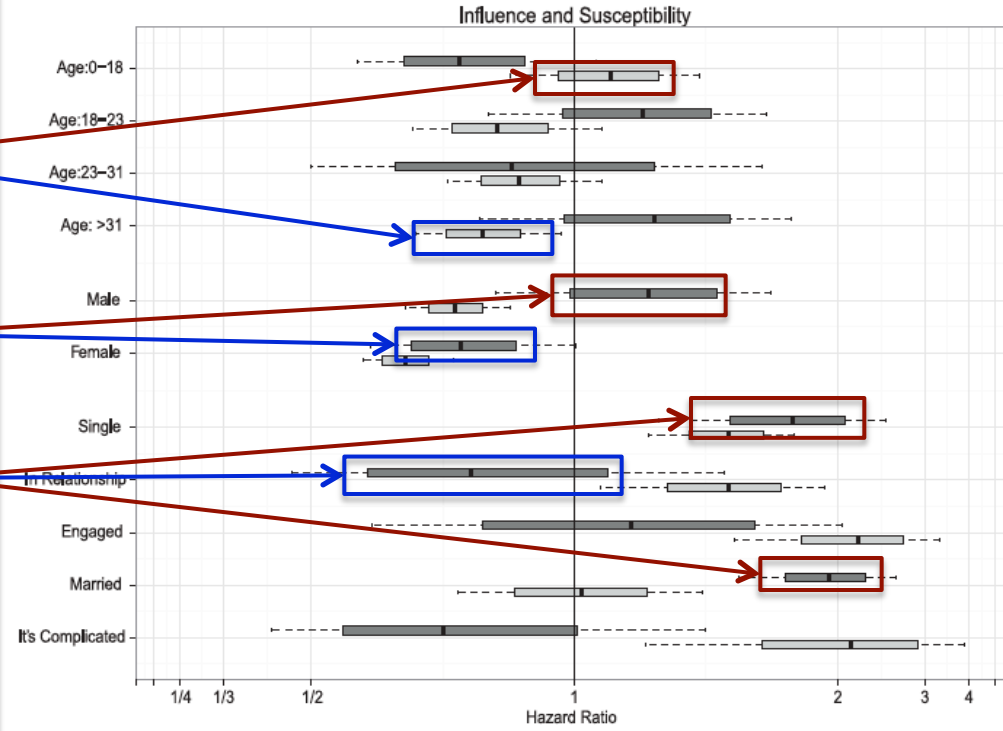
[2] Why I Deleted My Klout Profile, by Pam Moore, at Social Media Today, originally published November 19, 2011; retrieved November 26 2011

Case 3: Influential verse Susceptible[1]

- Study of product adoption for 1.3M FB users

Results:

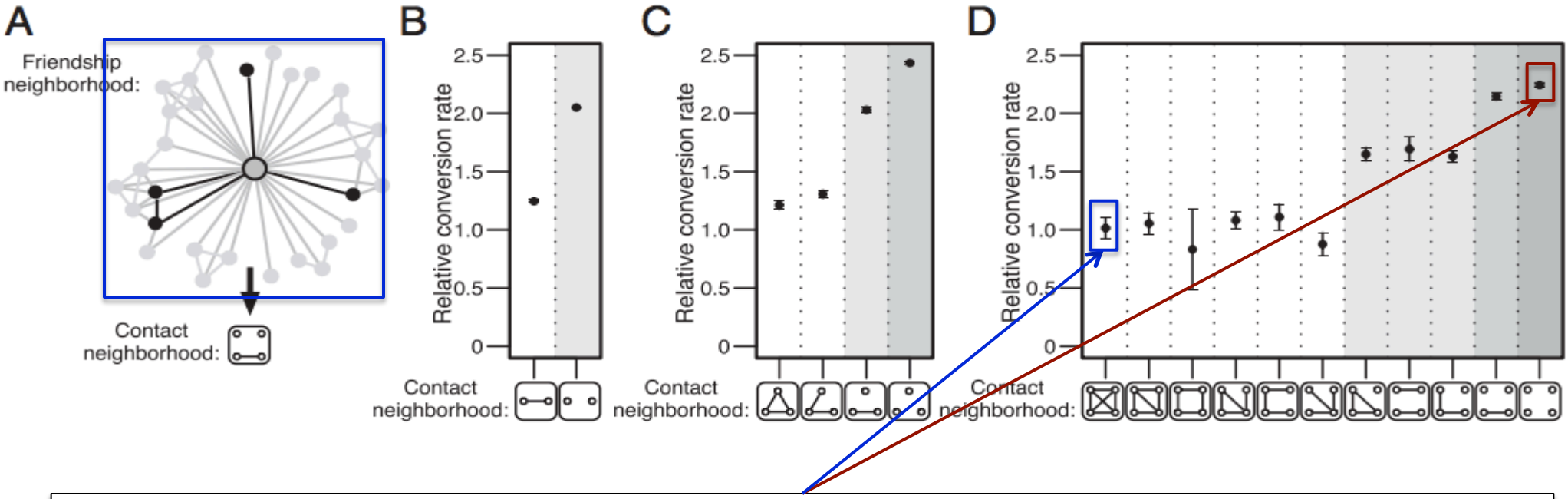
- Younger users are more (18%, $P<0.05$) susceptible to influence than older users
- Men are more (49%, $P<0.05$) influential than women
- Single and Married individuals are significantly more (>100%, $P<0.05$) influential than those who are in a relationship
- Married individuals are the least susceptible to influence



[1] S. Aral and D Walker. Identifying Influential and Susceptible Members of Social Networks. Science, 337:337-341, 2012.

Case 4: Who influenced you and How?

- Magic: the structural diversity of the ego network[1]

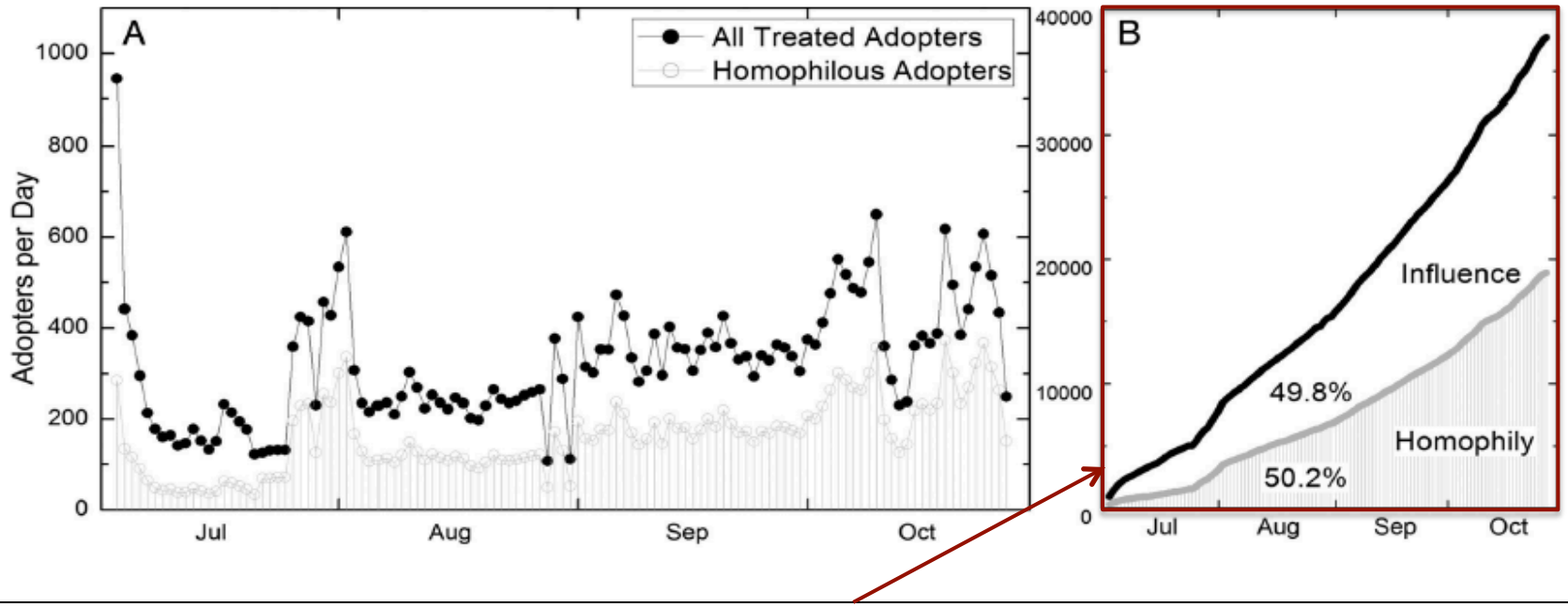


Results: Your behavior is influenced by the "structural diversity" (the number of connected components in your ego network) instead of the number of your friends.

[1] J. Ugandera, L. Backstromb, C. Marlowb, and J. Kleinberg. Structural diversity in social contagion. PNAS, 109 (20): 7591-7592, 2012.

Case 5: Influence and Correlation

- "Break" the myth of social influence



Results:

- Homophily explains >50% of the perceived behavioral contagion
- Previous methods overestimate peer influence by 300-700%

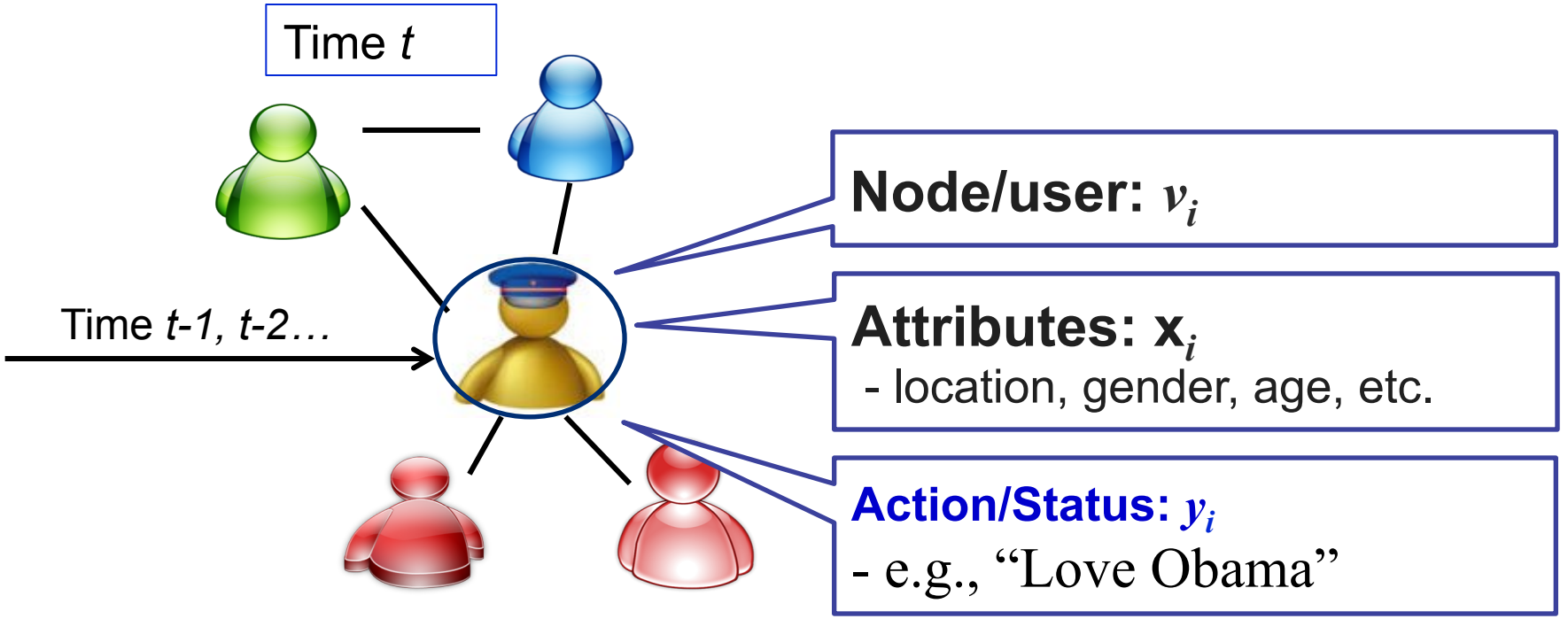
[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS, 106 (51):21544-21549, 2009.

Challenges: WH^3

1. Whether social influence exist?
2. How to measure influence?
3. How to model influence?
4. How influence can help real applications?

Preliminaries

Notations



$$G = (V, E, X, Y)$$

G^t — the superscript t represents the time stamp

$e_{ij}^t \in E^t$ — represents a link/relationship from v_i to v_j at time t

Homophily

- Homophily
  - A user in the social network tends to be similar to their connected neighbors.
- Originated from different mechanisms
  - Social influence
    - Indicates people tend to follow the behaviors of their friends
  - Selection
    - Indicates people tend to create relationships with other people who are already similar to them
  - Confounding variables
    - Other unknown variables exist, which may cause friends to behave similarly with one another.

Influence and Selection[1]

$$Selection = \frac{p(e_{ij}^{t} = 1 \mid e_{ij}^{t-1} = 0, \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle > \varepsilon)}{p(e_{ij}^{t} = 1 \mid e_{ij}^{t-1} = 0)}$$

Similarity between user *i* and *j* at time *t*-1 is larger than a threshold

There is a link between user *i* and *j* at time *t*

- Denominator: the conditional probability that an unlinked pair will become linked
- Numerator: the same probability for unlinked pairs whose similarity exceeds the threshold

$$Influence = \frac{p(\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{t} \rangle > \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle \mid e_{ij}^{t} = 1, e_{ij}^{t-1} = 0)}{p(\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{t} \rangle > \langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \rangle \mid e_{ij}^{t-1} = 0)}$$

- Denominator: the probability that the similarity increase from time *t*-1 to time *t* between two nodes that were not linked at time *t*-1
- Numerator: the same probability that became linked at time *t*

- A Model is learned through matrix factorization/factor graph

[1] J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In KDD'09, pages 747–756, 2009.

Other Related Concepts

- Cosine similarity
- Correlation factors
- Hazard ratio
- *t*-test

Cosine Similarity

- A measure of similarity
- Use a vector to represent a sample (e.g., user)

$$\mathbf{x} = (x_1, \ldots, x_n)$$

- To measure the similarity of two vectors **x** and **y**, employ cosine similarity:

$$sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Correlation Factors

- Several correlation coefficients could be used to measure correlation between two random variables *x* and *y*.
- Pearsons' correlation

$$\rho_{x,y} = corr(x,y) = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

mean

Standard deviation

- It could be estimated by

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- Note that **correlation** does NOT imply **causation**

Hazard Ratio

- **Hazard Ratio**
  - Chance of an event occurring in the treatment group divided by its chance in the control group
  - Example:

    $$\frac{\text{Chance of users to buy iPhone with >=1 iPhone user friend(s)}}{\text{Chance of users to buy iPhone without any iPhone user friend}}$$

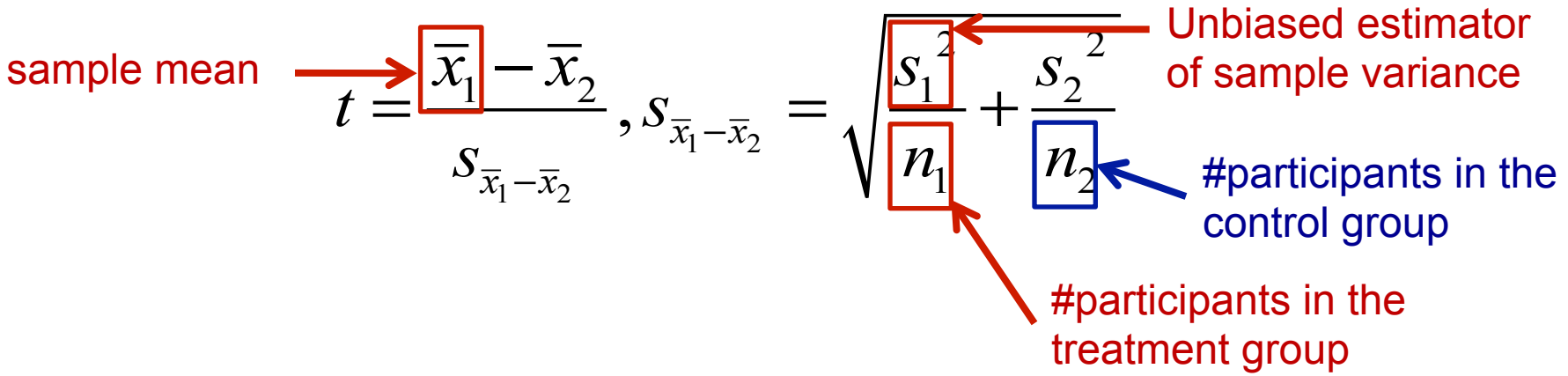
  - Measuring instantaneous chance by *hazard rate h(t)*

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{observed events in interval}[t, t+\Delta t]/N(t)}{\Delta t}$$

  - The hazard ratio is the relationship between the instantaneous hazards in two groups
  - Proportional hazards models (e.g. Cox-model) could be used to report hazard ratio.

t-test

- A *t*-test usually used when the test statistic follows a Student's *t* distribution if the null hypothesis is supported.
- To test if the difference between two variables are significant
- Welch's *t*-test
  - Calculate *t*-value



  - Find the *p*-value using a table of values from Student's t-distribution
  - If the *p*-value is below chosen threshold (e.g. 0.01) then the two variables are viewed as significant different.

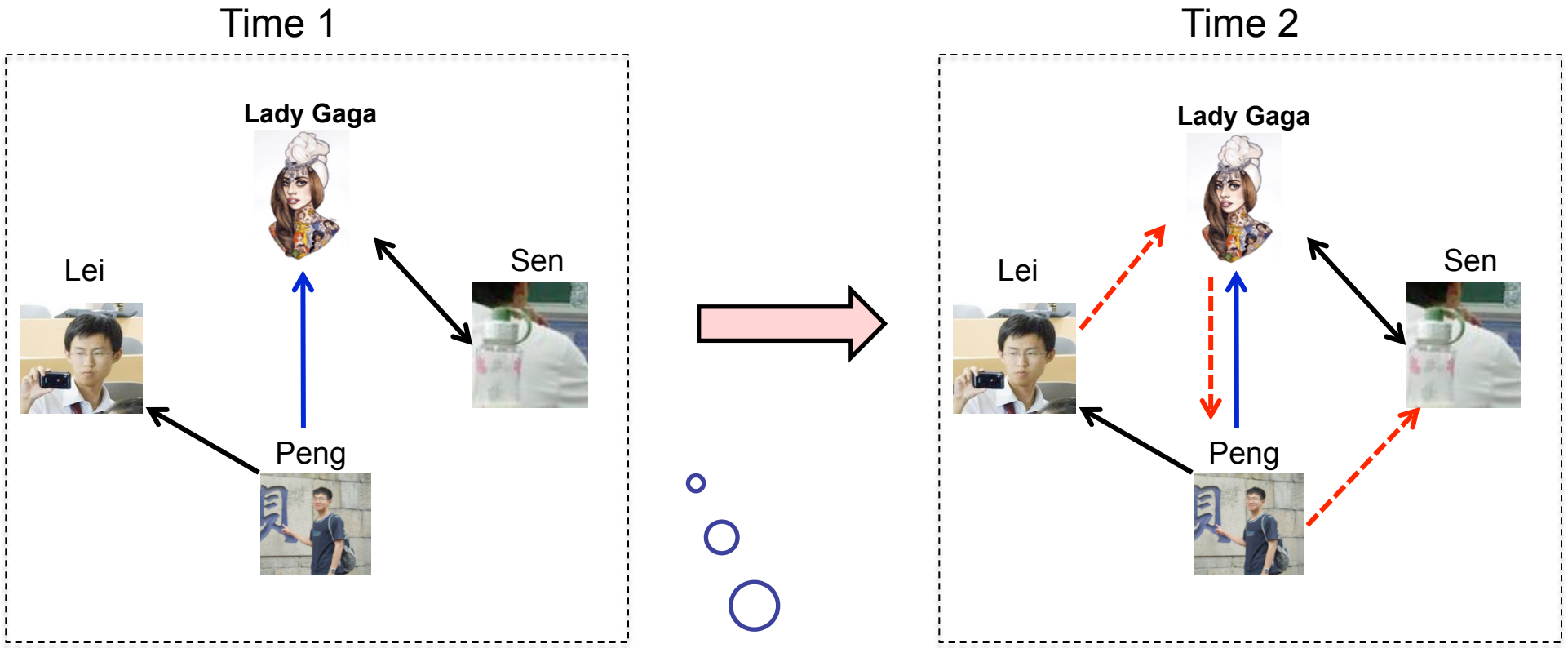
Data Sets

Ten Cases

Network	#Nodes	#Edges	Behavior
Twitter-net	111,000	450,000	Follow
Weibo-Retweet	1,700,000	400,000,000	Retweet
Slashdot	93,133	964,562	Friend/Foe
Mobile (THU)	229	29,136	Happy/Unhappy
Gowalla	196,591	950,327	Check-in
ArnetMiner	1,300,000	23,003,231	Publish on a topic
Flickr	1,991,509	208,118,719	Join a group
PatentMiner	4,000,000	32,000,000	Patent on a topic
Citation	1,572,277	2,084,019	Cite a paper
Twitter-content	7,521	304,275	Tweet “Haiti Earthquake”

Most of the data sets will be publicly available for research.

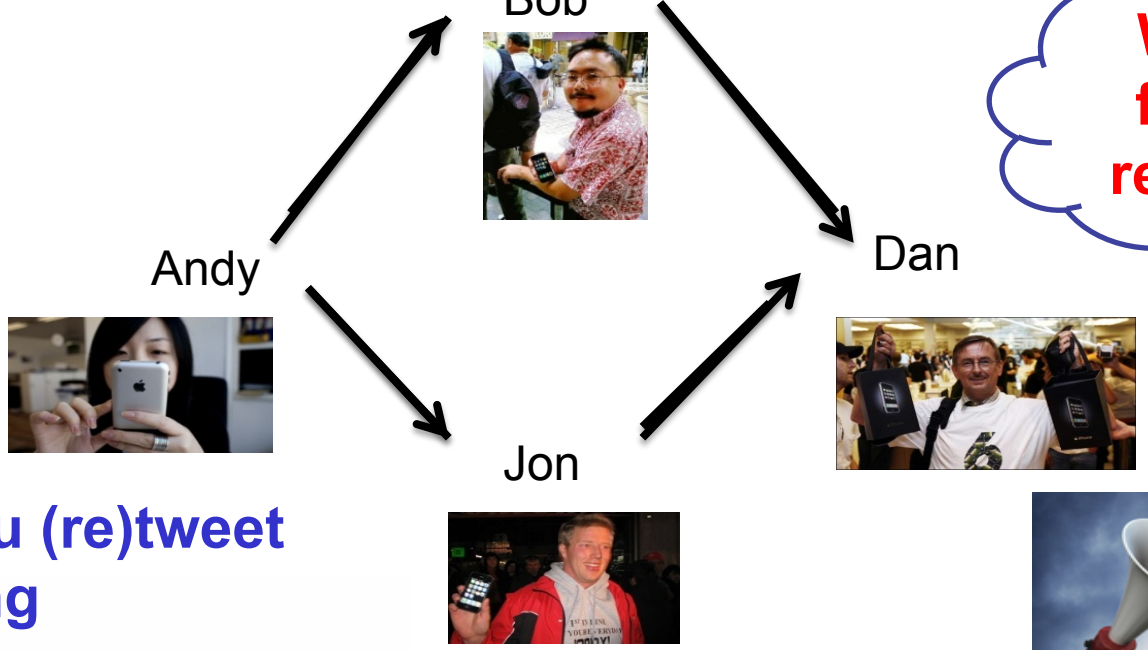
Case 1: Following Influence on Twitter



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

Case 2: Retweeting Influence

新浪微博 weibo.com



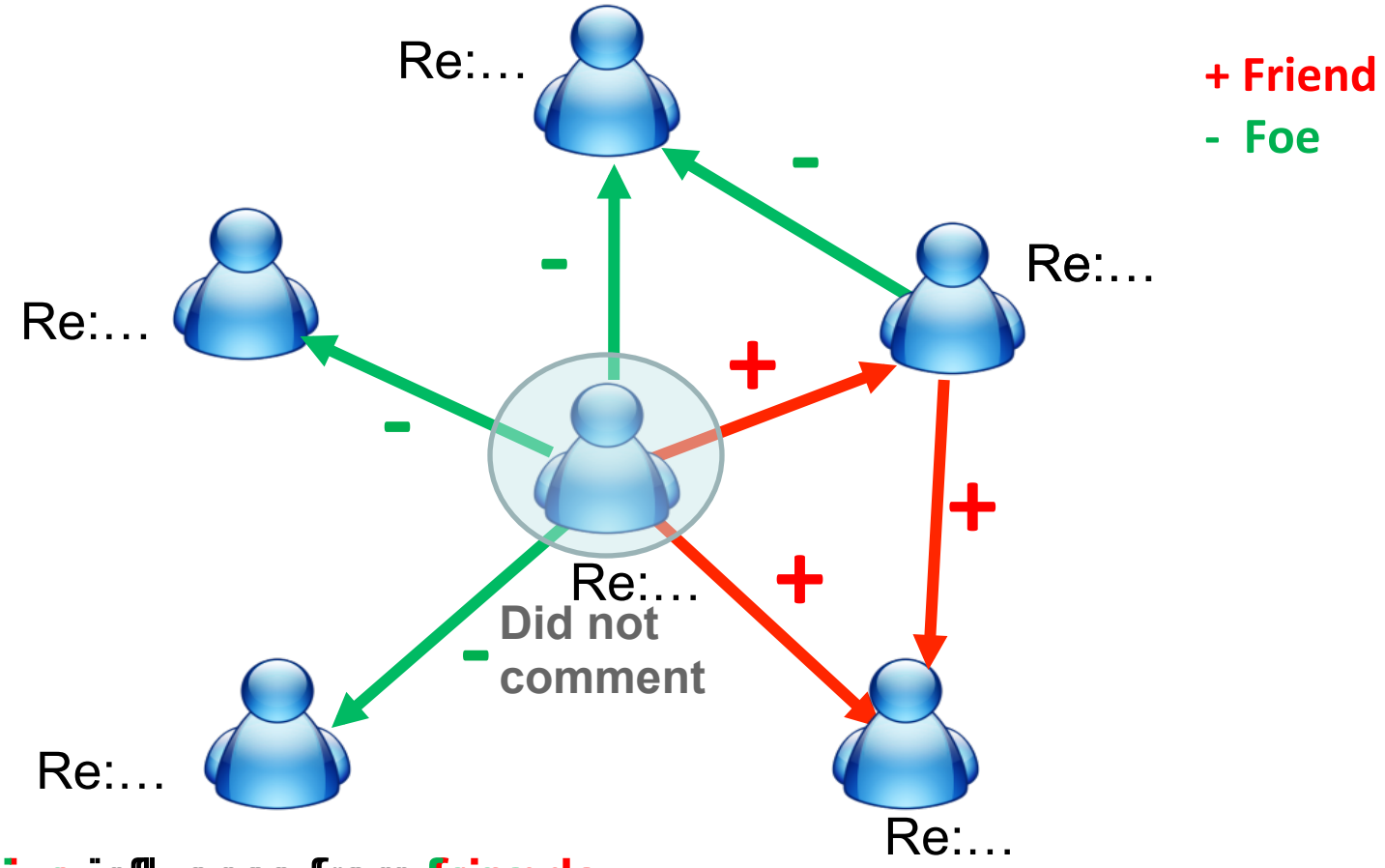
Who will follow to retweet it?

When you (re)tweet something

Case 3: Commenting Influence

News: Alan Cox Exists Intel.

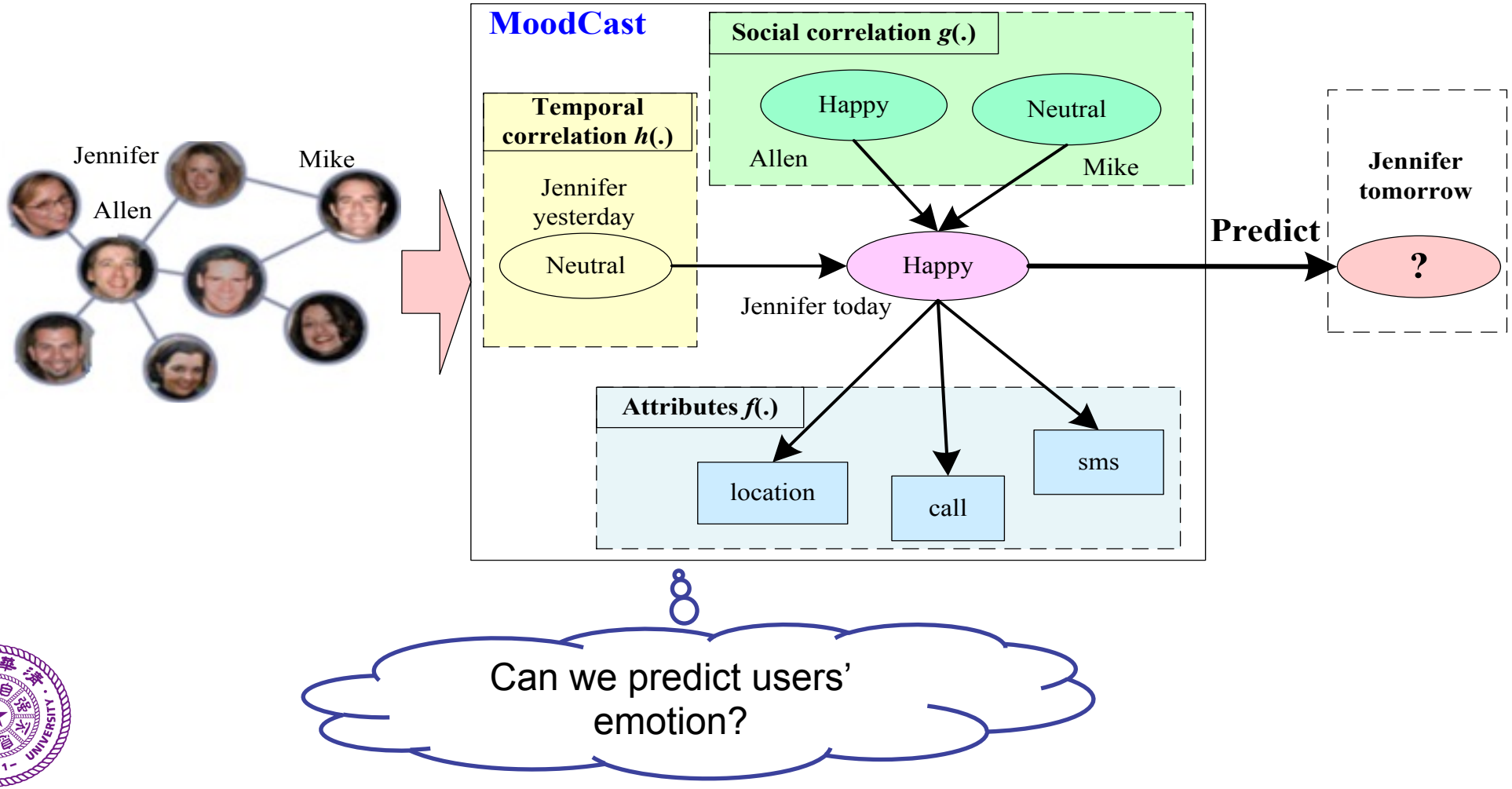
Governments Want Private Data



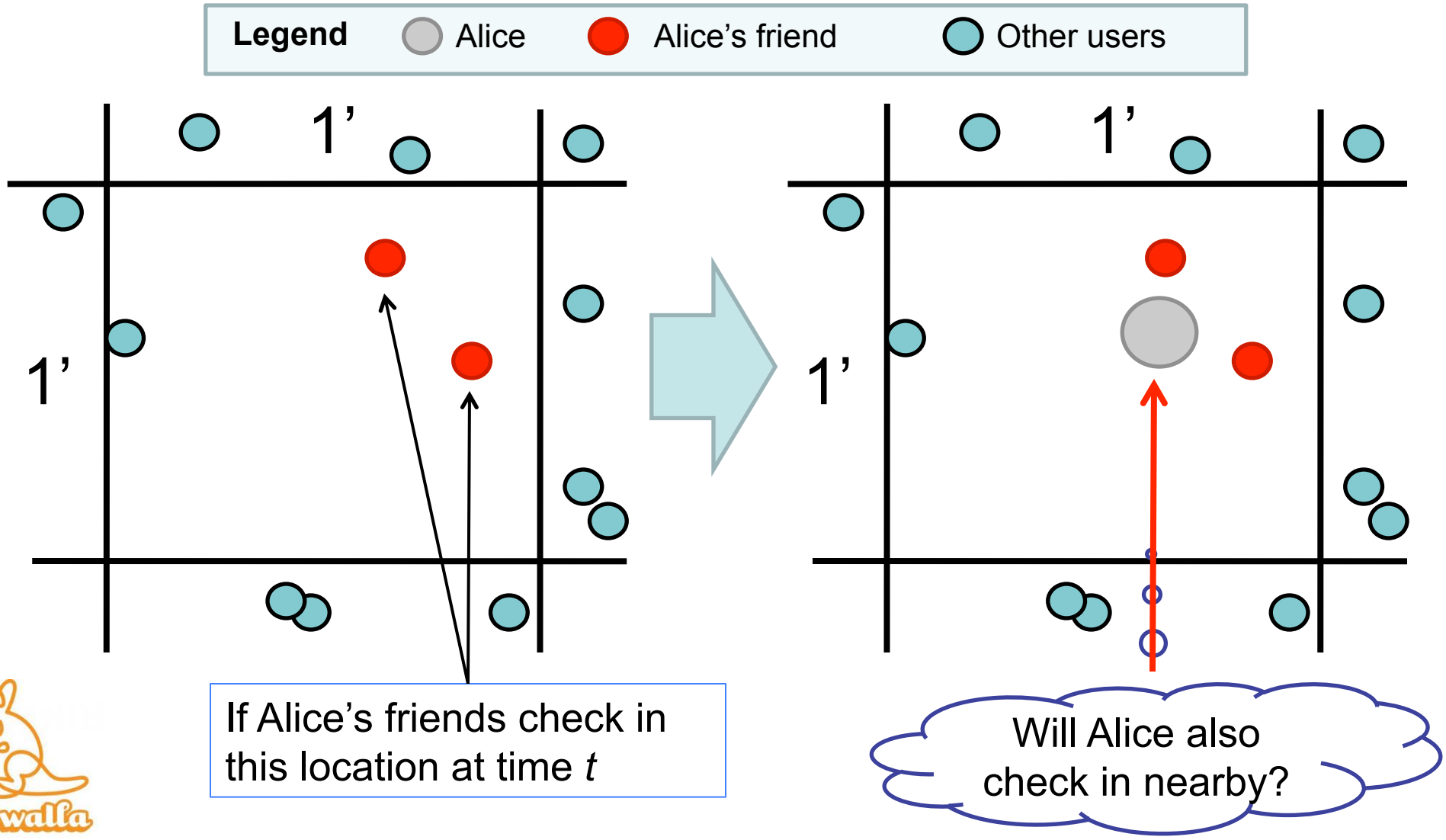
negative influence from friends

Case 4: Emotion Influence

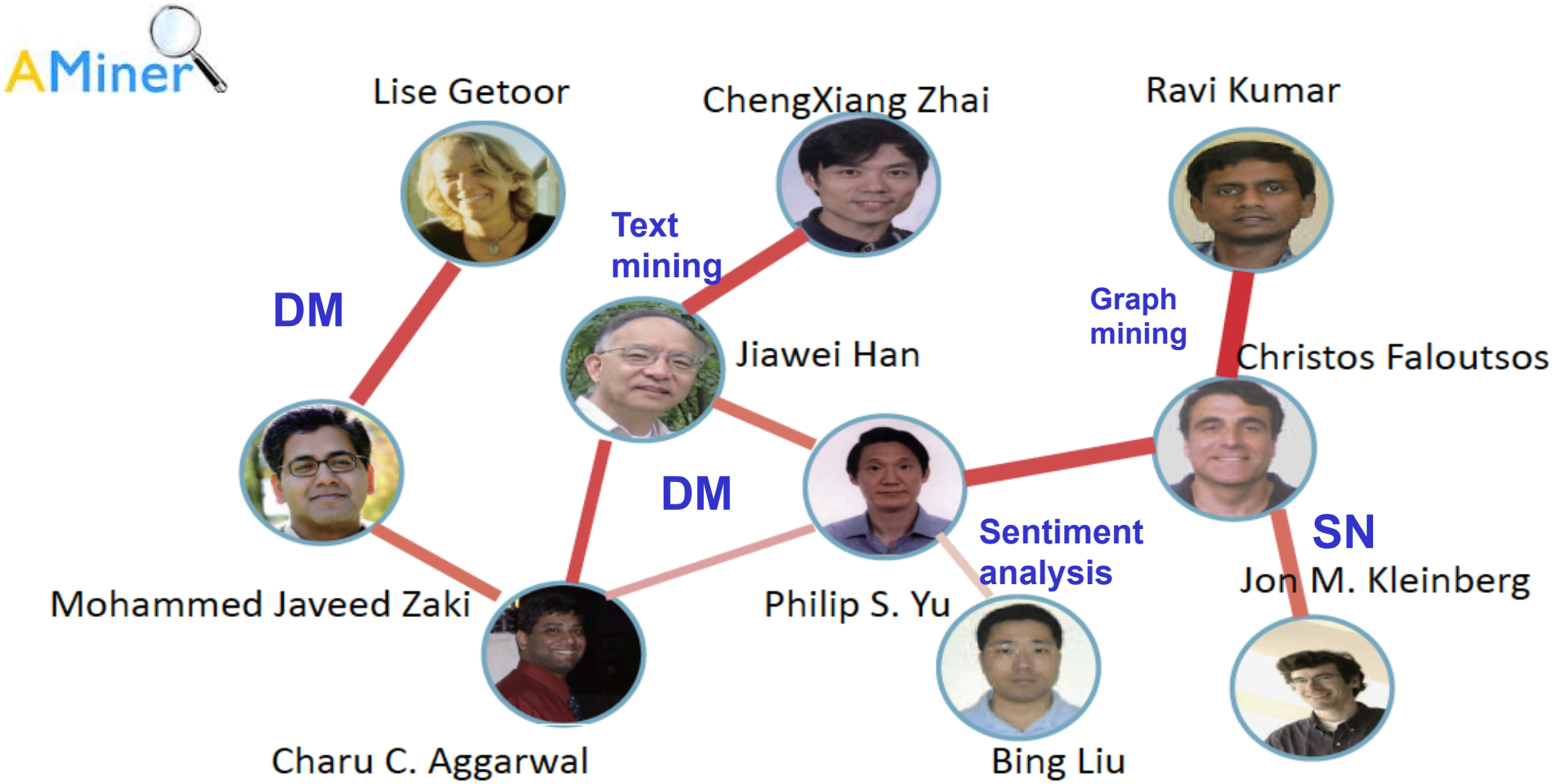
Case 4: Emotion Influence (cont.)



Case 5: Check-in Influence in Gowalla

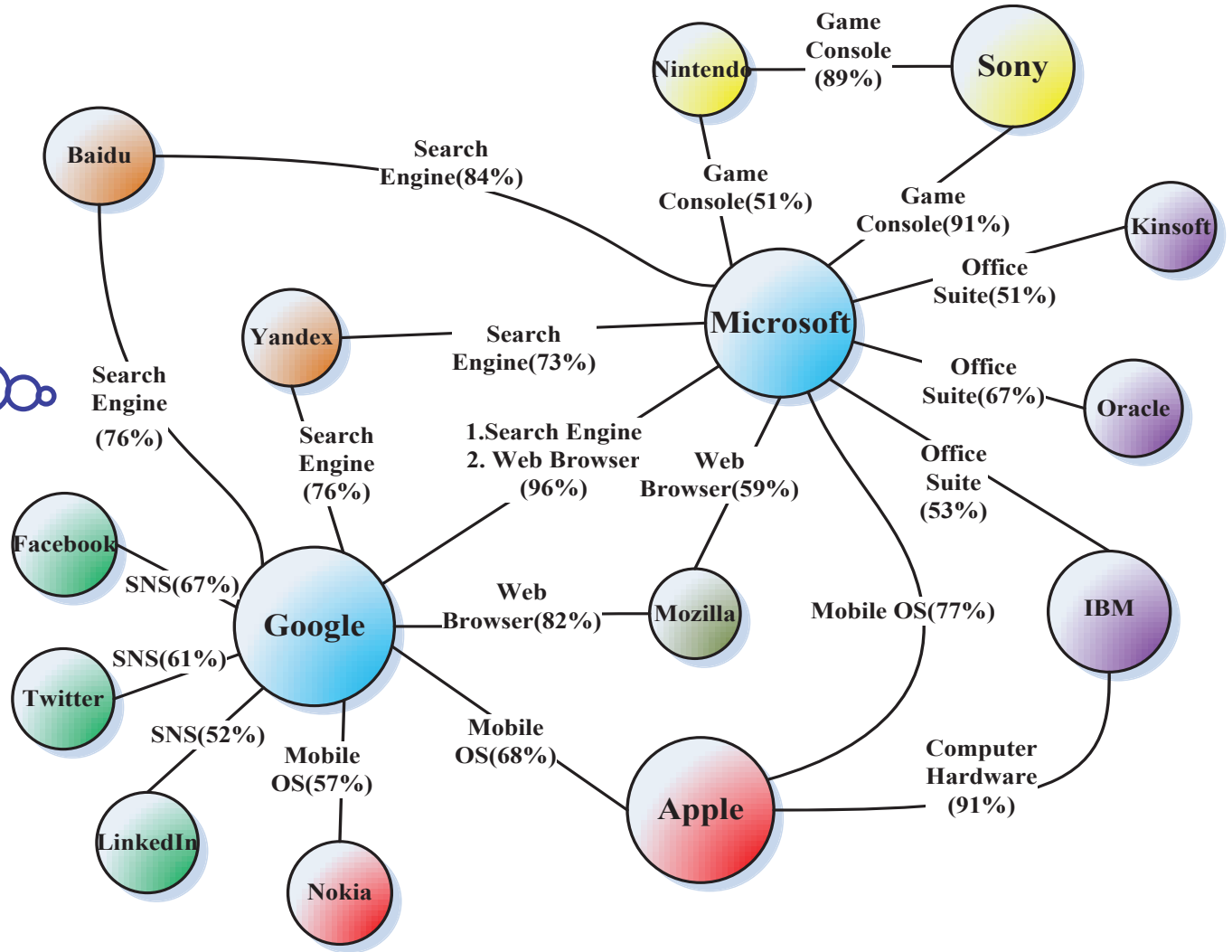


Case 6: Correlation & Influence in Academia



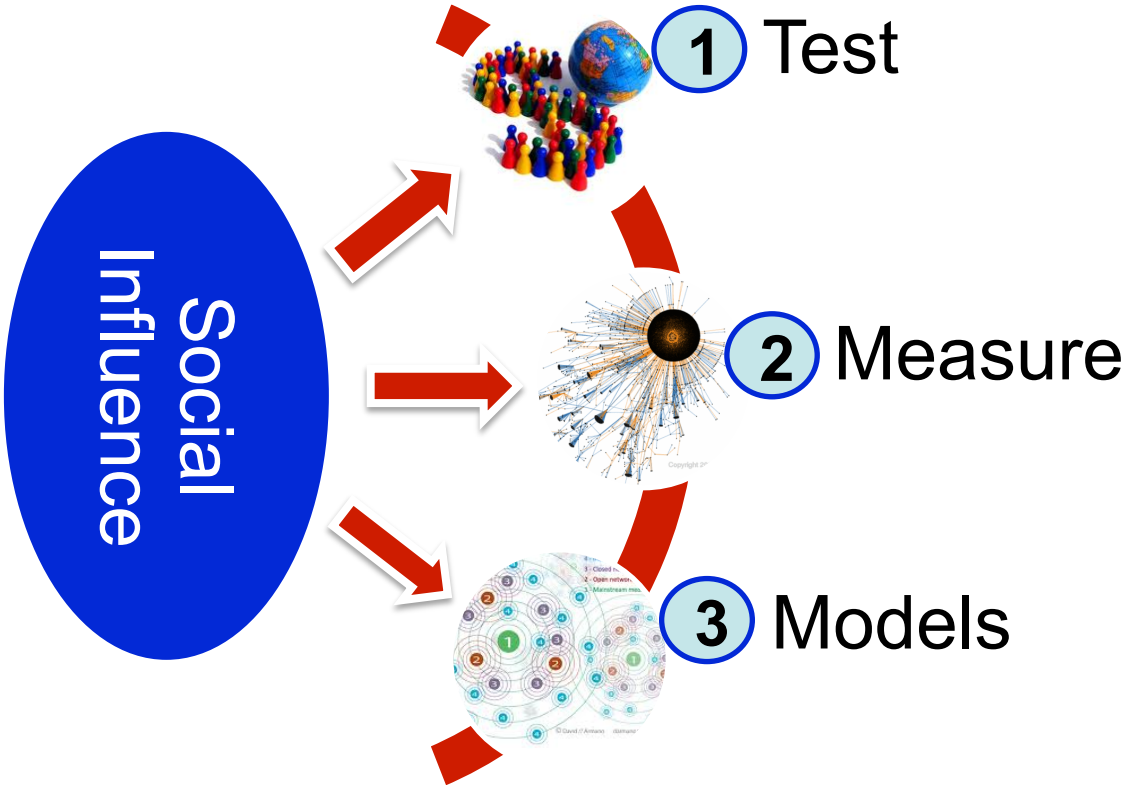
Case 7: Patenting Influence

How competitors' patenting behaviors influence each other

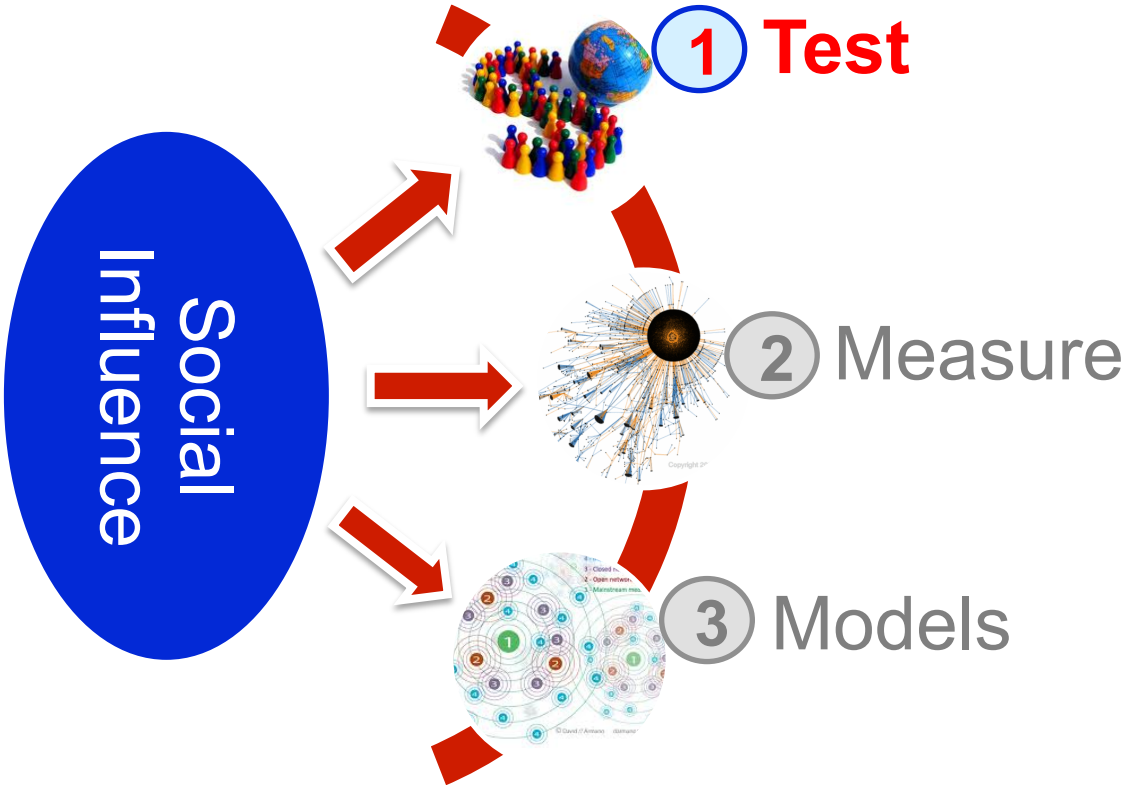


PatentMiner

Social Influence



Social Influence



Randomization

- Theoretical fundamentals[1, 2]
  - In science, randomized experiments are the experiments that allow the greatest reliability and validity of statistical estimates of treatment effects.
- Randomized Control Trials (RCT)
  - People are randomly assigned to a “treatment” group or a “controlled” group;
  - People in the treatment group receive some kind of “treatment”, while people in the controlled group do not receive the treatment;
  - Compare the result of the two groups, e.g., survival rate with a disease.

[1] Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 5, 688–701.

[2] http://en.wikipedia.org/wiki/Randomized_experiment

RCT in Social Network

- We use RCT to test the influence and its significance in SN.

- Two challenges:
  - How to define the treatment group and the controlled group?
  - How to find a real random assignment?

Example: Political mobilization

- There are two kinds of treatments.

A controlled trial

Treatment Group 1

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

Treatment for Group 2

- Informational msg group: was shown with msg that indicates how many other.
- Control group: did not receive any msg.

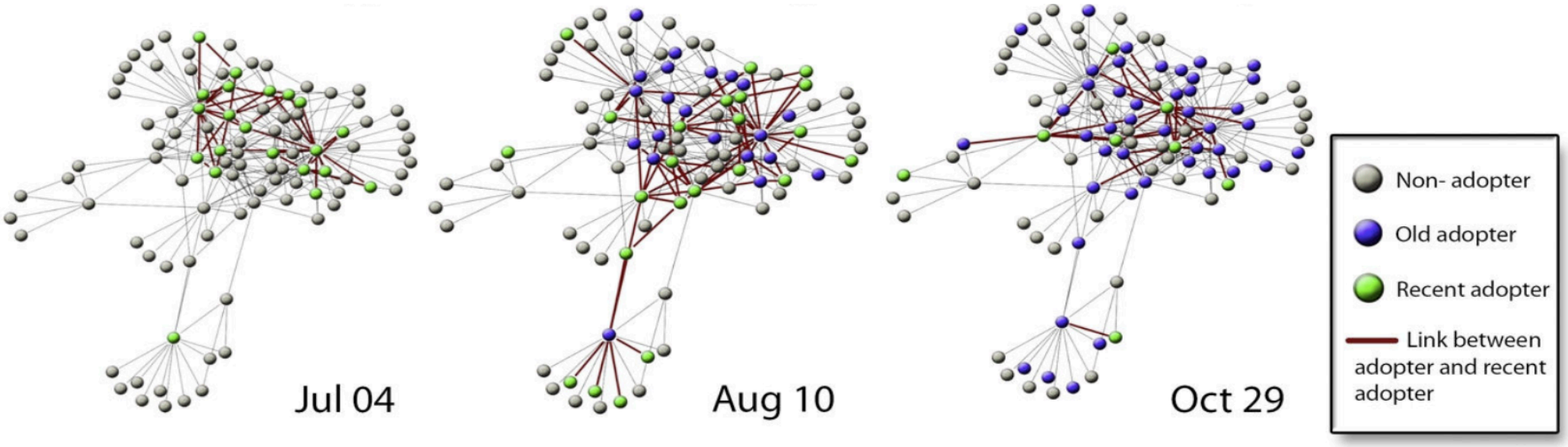
Treatment for Group 1

Treatment for Group 1&2

[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.

Adoption Diffusion of Y! Go

Yahoo! Go is a product of Yahoo to access its services of search, mailing, photo sharing, etc.



RCT:

- **Treatment group:** people who did not adopt Y! Go but have friend(s) adopted Y! Go at time *t*;
- **Controlled group:** people who did not adopt Y! Go and also have no friends adopted Y! Go at time *t*.

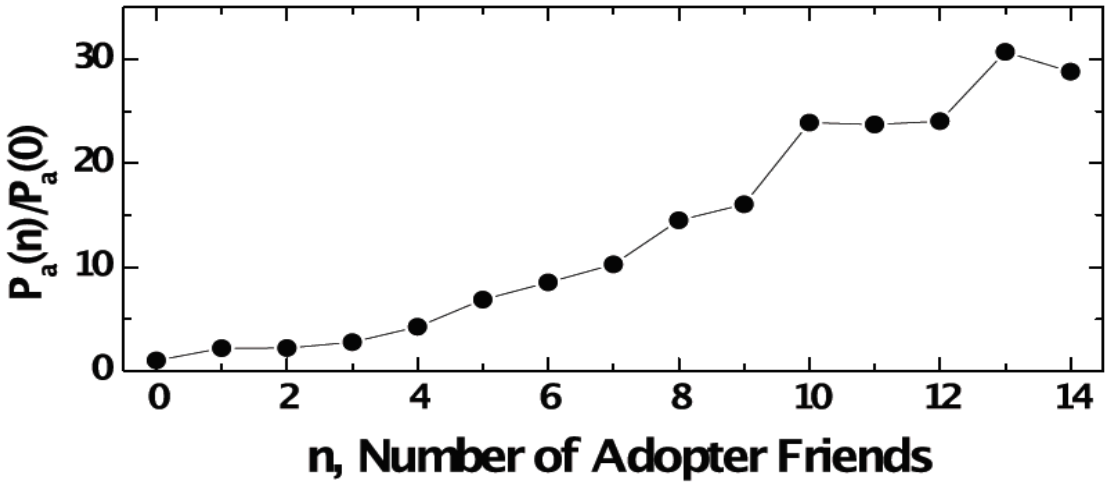
[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS, 106 (51):21544-21549, 2009.

For an example

- Yahoo! Go
  - 27.4 M users, 14 B page views, 3.9 B messages
- The RCT
  - Control seeds: random sample of 2% of the entire network (3.2M nodes)
  - Experimental seeds: all adopters of Yahoo! Go from 6/1/2007 to 10/31/2007 (0.5M nodes)

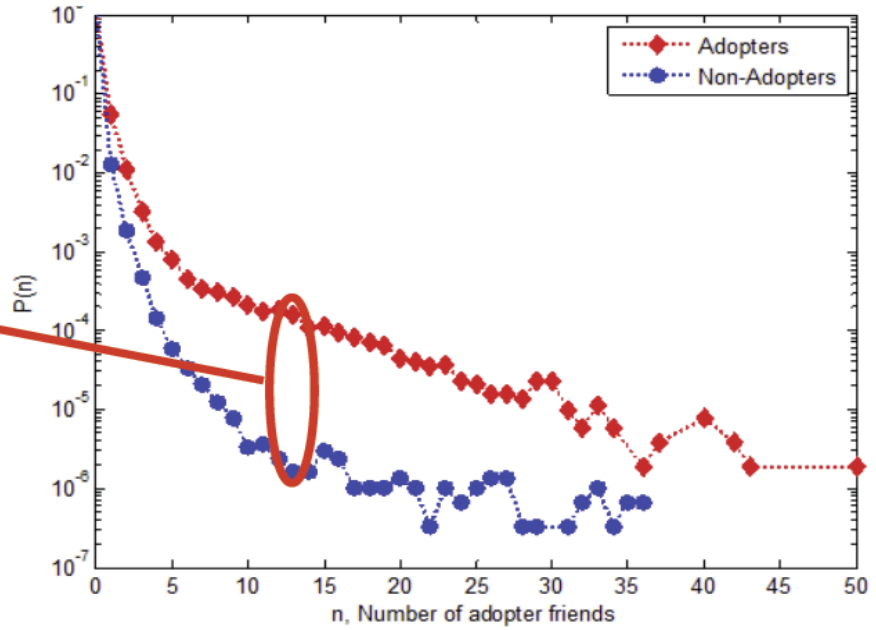
[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS, 106 (51):21544-21549, 2009.

Evidence of Influence?



Adopters are 100 times more likely to have 12 adopter friends than non-adopters

Is the setting fair?



Matched Sampling Estimation

- Bias of existing randomized methods
  - Adopters are more likely to have adopter friends than non-adopters
- Matched sampling estimation
  - Match the treated observations with untreated who are as likely to have been treated, conditional on a vector of observable characteristics, but who were not treated

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

All attributes associated with user i at time t

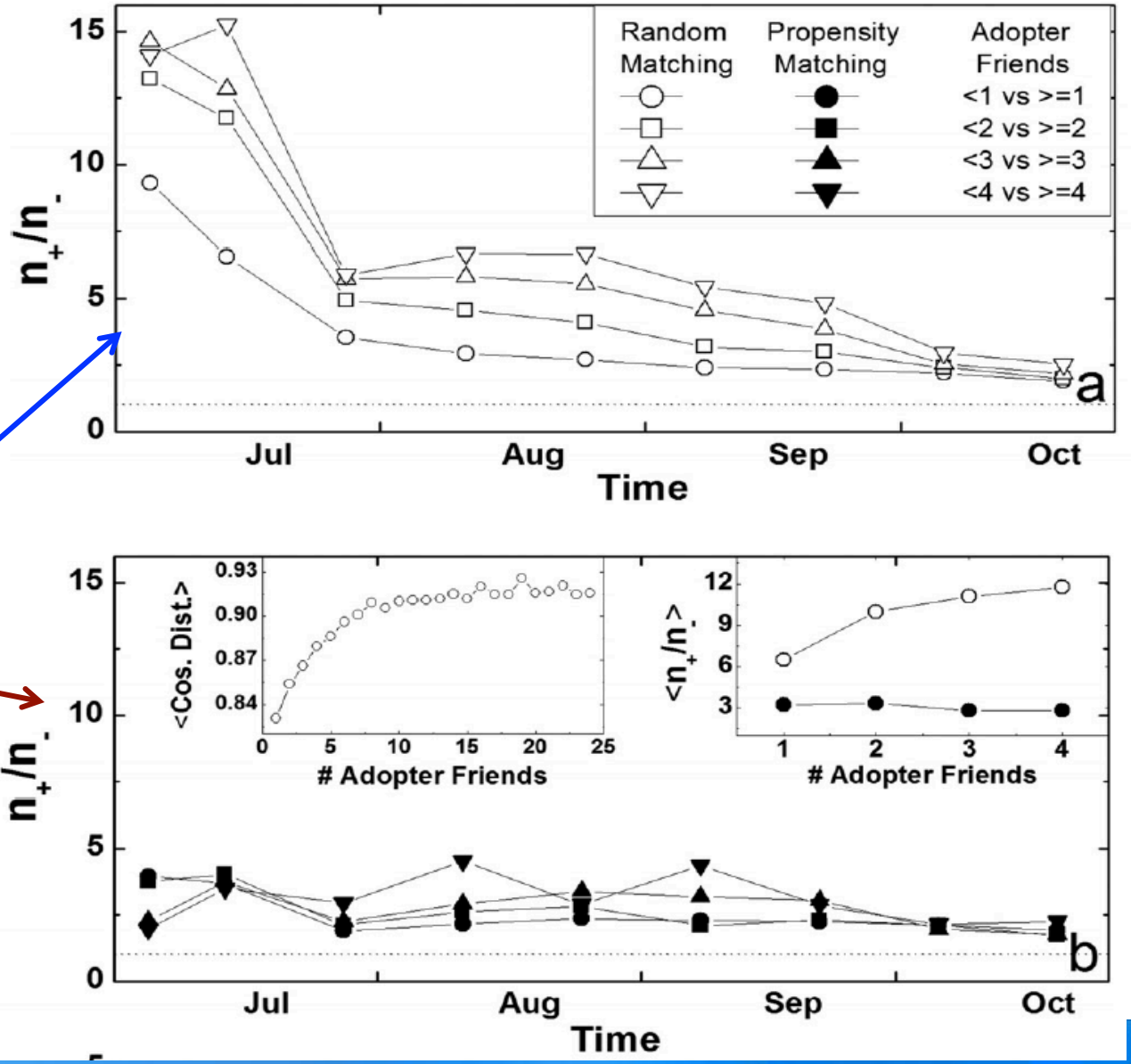
A binary variable indicating whether user i will be treated at time t

The new RCT:

- **Treatment group:** a user i who have k friends have adopted the Y! Go at time t;
- **Controlled group:** a matched user j who do not have k friends adopt Y! Go at time t, but is very likely to have k friends to adopt Y!Go at time t, i.e., $|p_{it} - p_{jt}| < \sigma$

Results—Random sampling and Matched sampling

The fraction of observed treated to untreated adopters (n_+/n_-) under:
(a) Random sampling;
(b) Matched sampling.

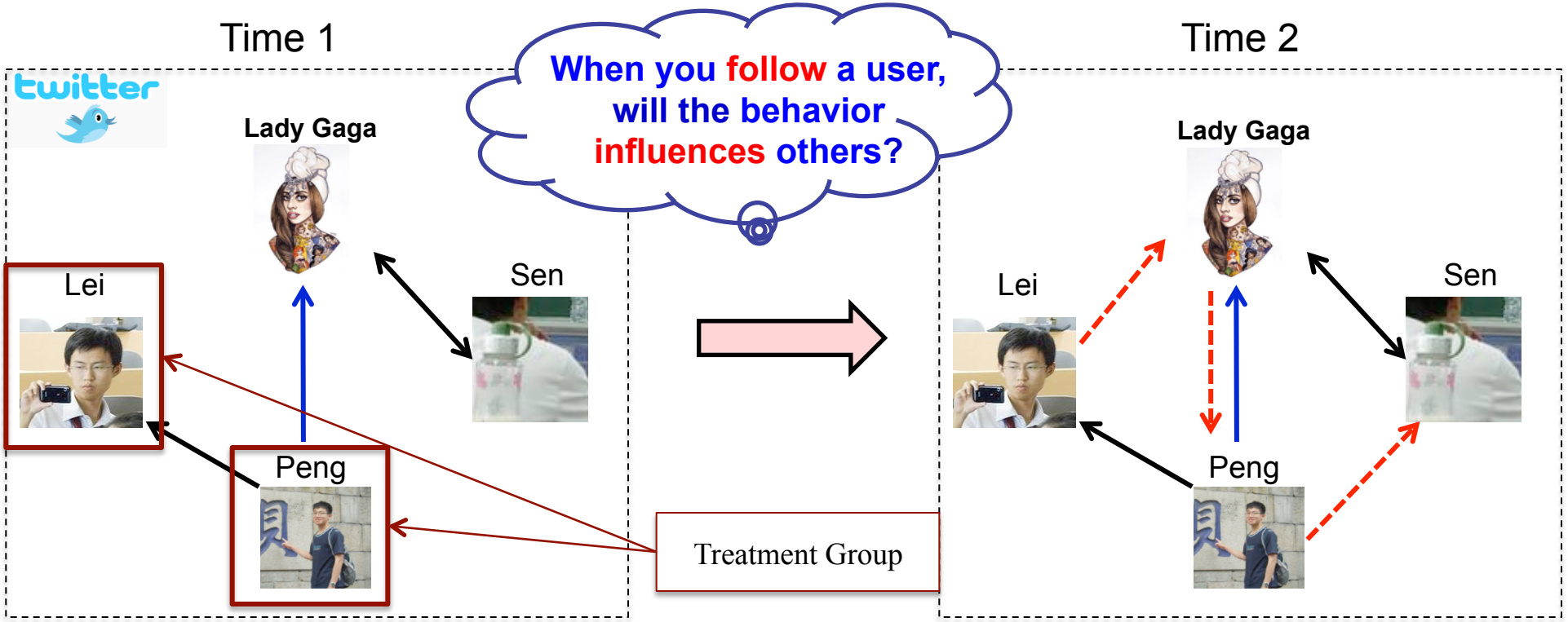


Two More Methods

- Shuffle test: shuffle the activation time of users.
  - If social influence does not play a role, then the timing of activation should be independent of the timing of activation of others.

- Reverse test: reserve the direction of all edges.
  - Social influence spreads in the direction specified by the edges of the graph, and hence reversing the edges should intuitively change the estimate of the correlation.

Example: Following Influence Test



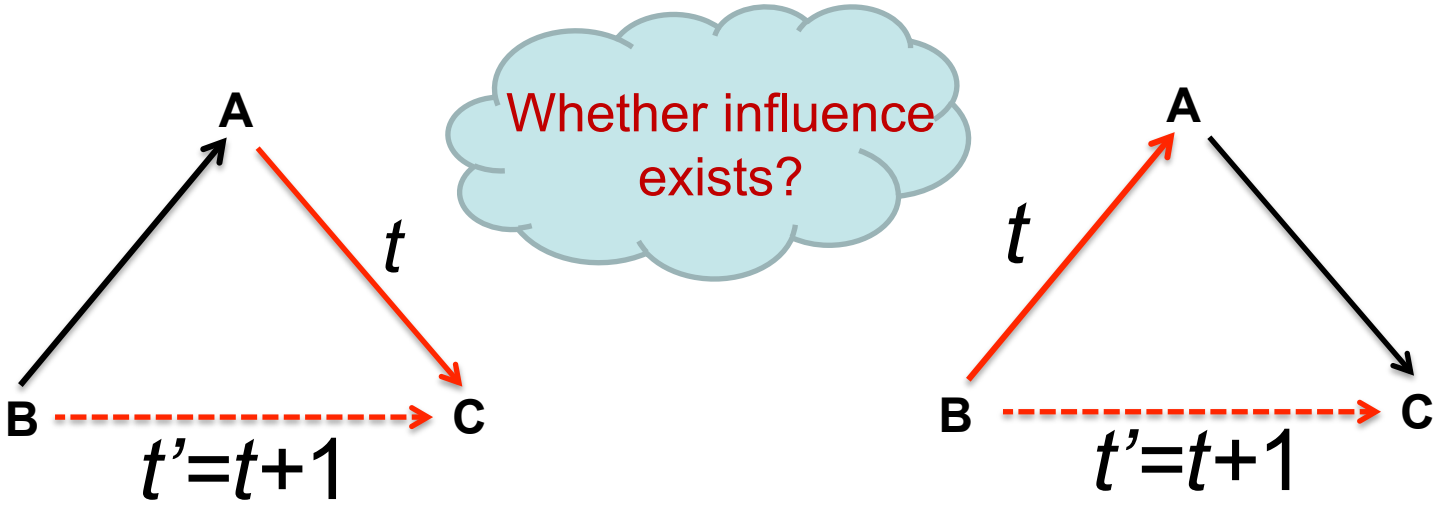
RCT:

- **Treatment group:** people who followed some other people or who have friends following others at time *t*;
- **Controlled group:** people who did not follow anyone and do not have any friends following others at time *t*.

[1] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. ACM TKDD, (accepted).

Influence Test via Triad Formation

Two Categories of Following Influences



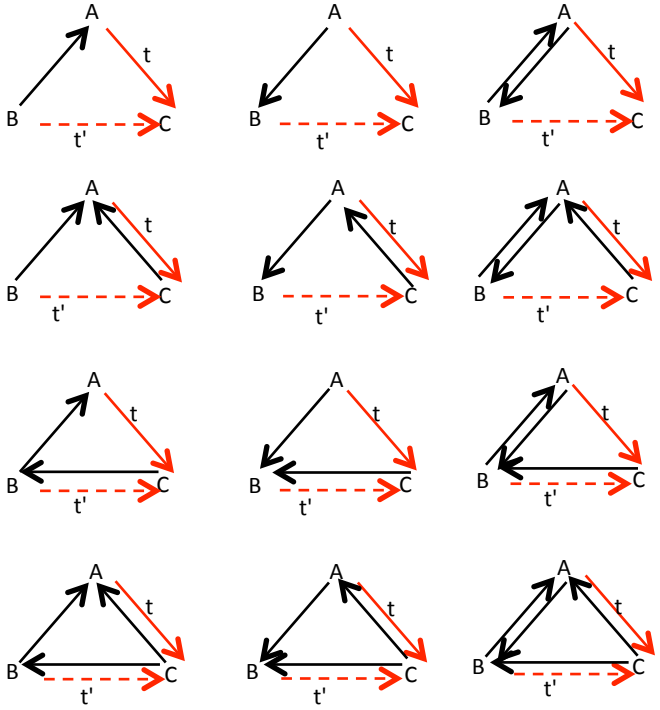
Follower diffusion

Followee diffusion

–>: pre-existed relationships
–>: a new relationship added at *t*
-->: a possible relationship added at *t*+1

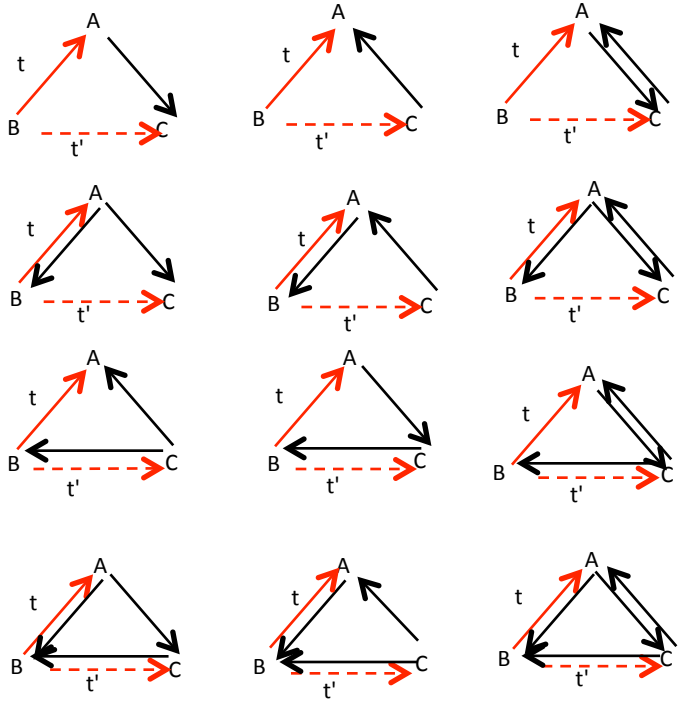
24 Triads in Following Influence

Follower diffusion



12 triads

Followee diffusion



12 triads

Twitter Data

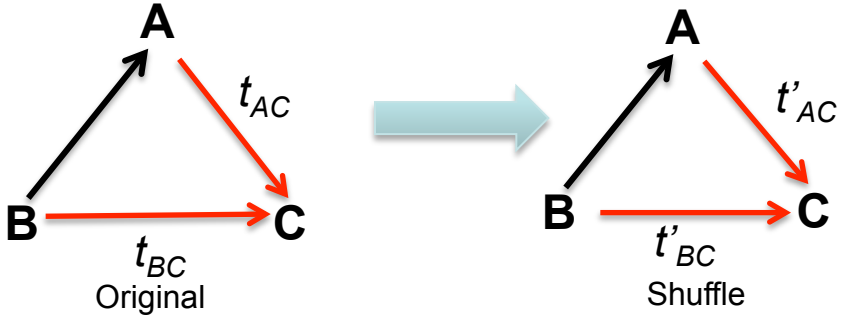
- Twitter data
  - “Lady Gaga” -> 10K followers -> millions of followers;
  - 13,442,659 users and 56,893,234 following links.
  - 35,746,366 tweets.

- A complete dynamic network
  - We have all followers and all followees for every user
  - 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.

Shuffle test

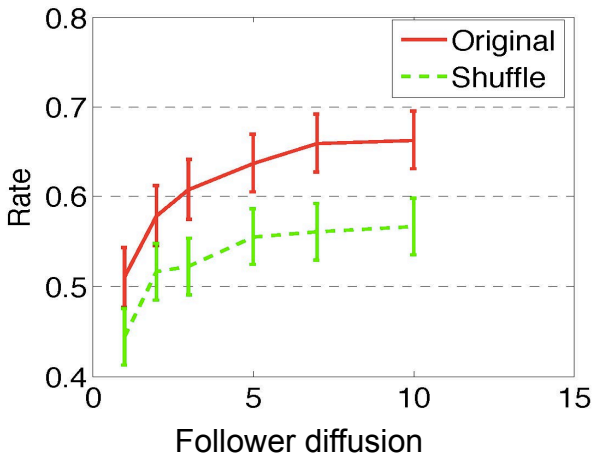


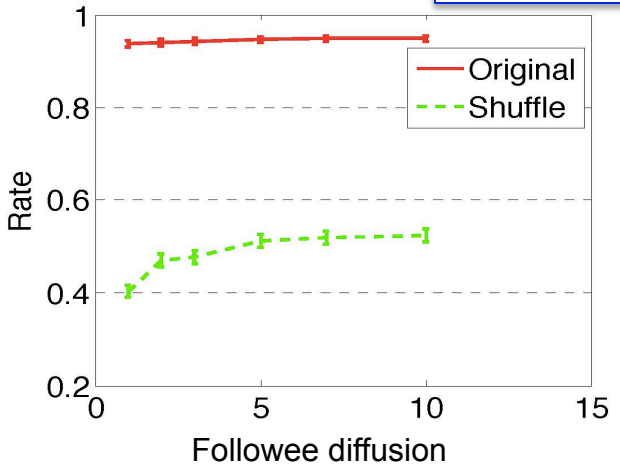
- Compare the rate under the original and shuffled dataset.

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

- Result

t-test, $P<0.01$

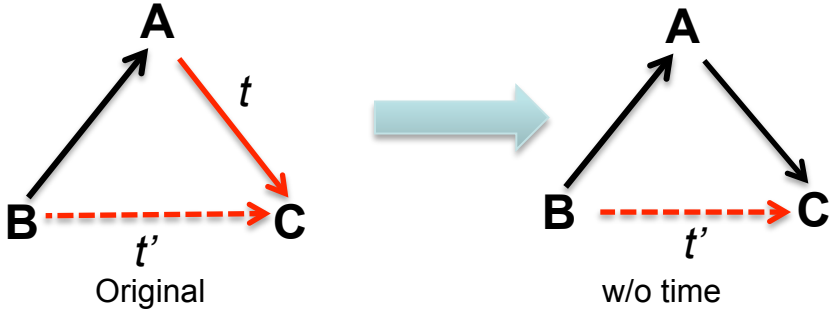




[1] A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks. In KDD, pages 7-15, 2008.

Test 2: Influence Decay Test

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



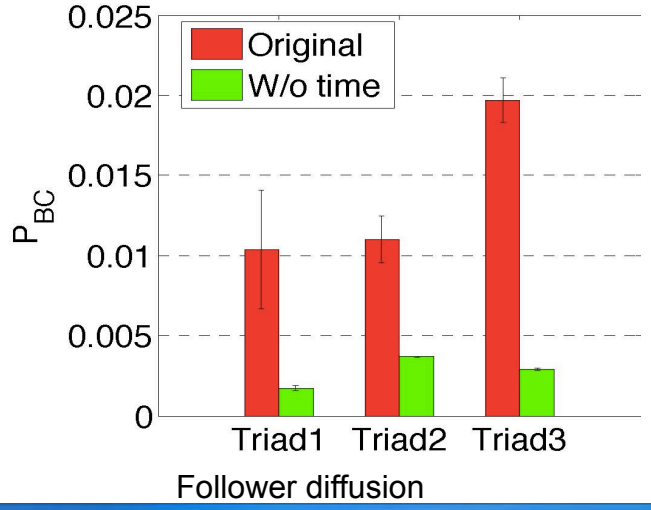
Shuffle test

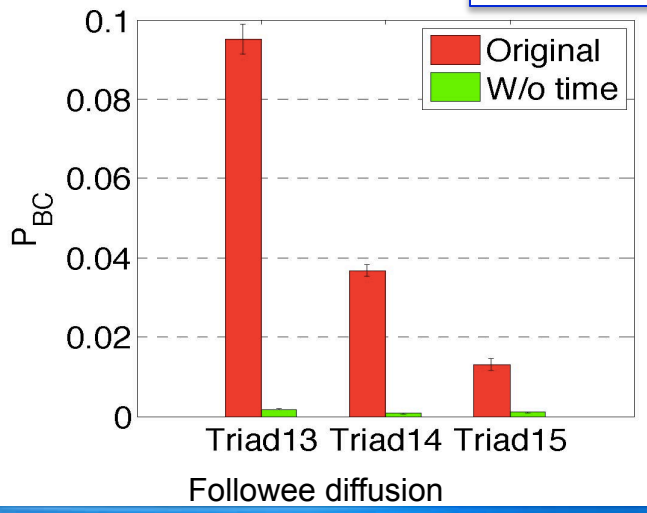
  - Compare the probability of B following C under the original and w/o time dataset.

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

t-test, $P<0.01$

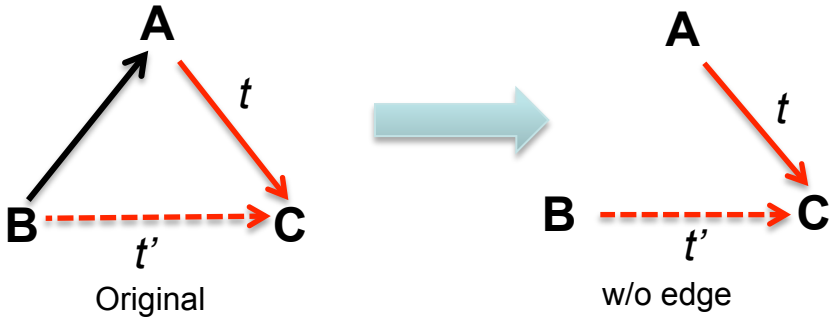
- Result





Test 3: Influence Propagation Test

- Method: Remove the relationship between A and B.



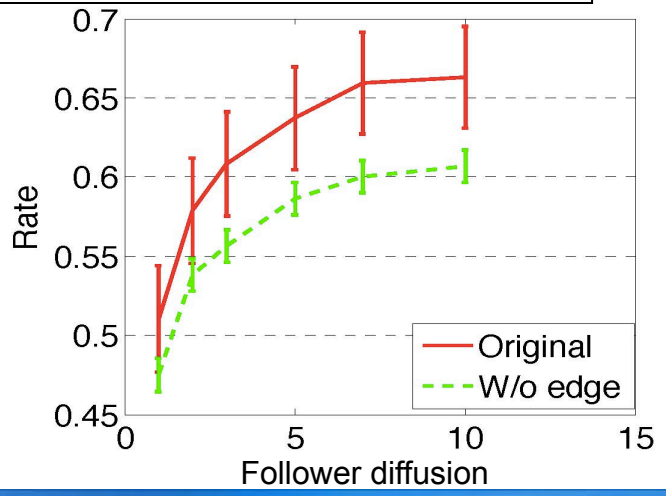
Reverse test

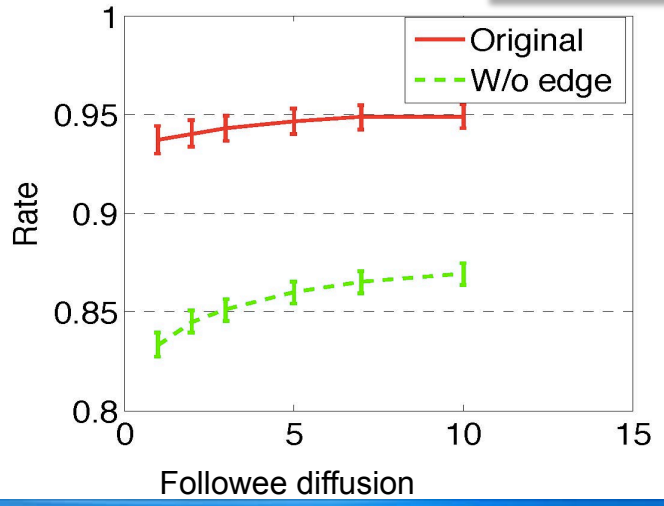
- Compare the rate under the original and w/o edge dataset.

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

t-test, $P<0.01$

- Result

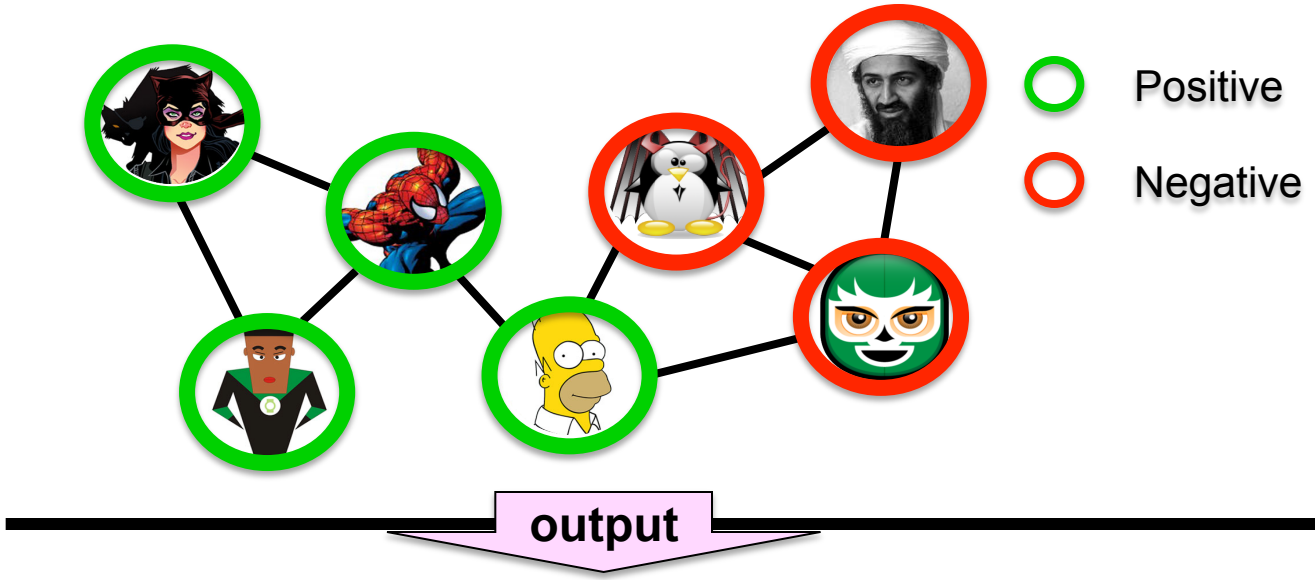




Summary

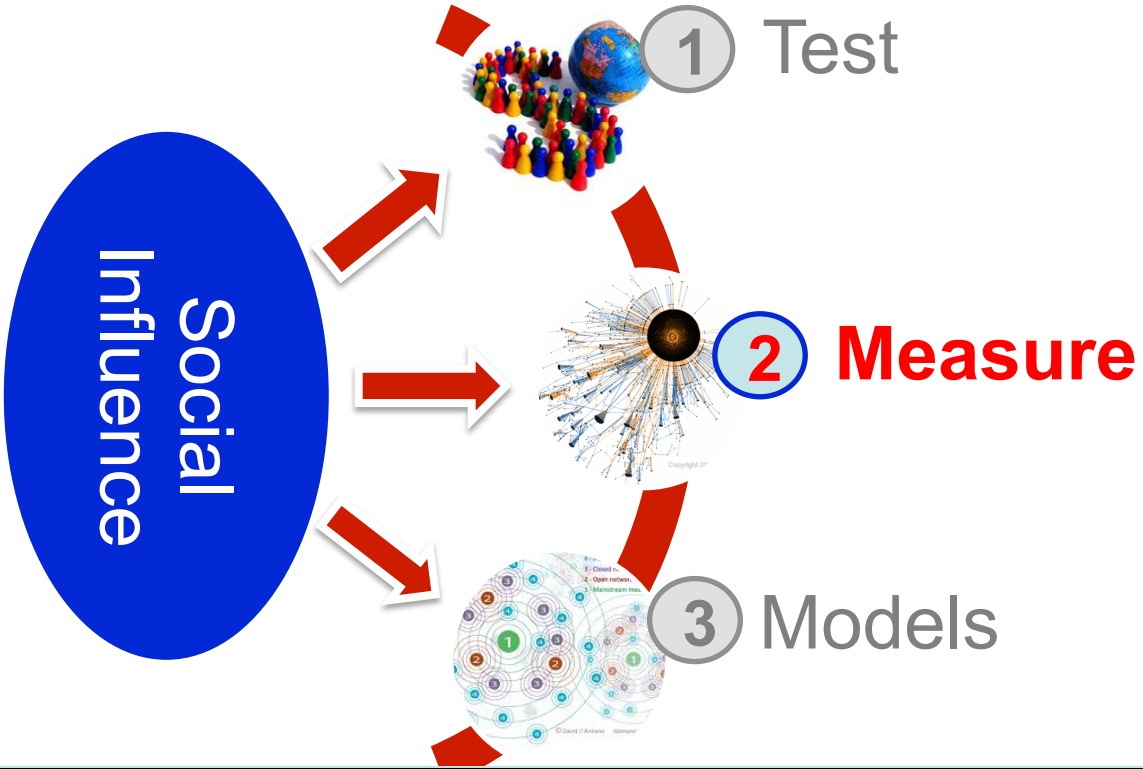
- Randomization test
  - Define “treatment” group
  - Define “controlled” group
  - Random assignment
- Shuffle test
- Reverse test

Output of Influence Test



There indeed exists influence!

Social Influence



"The idea of measuring influence is kind of crazy. Influence has always been something that we each see through our own lens."

—by CEO and co-founder of Klout, Joe Fernandez

Methodologies

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

Reachability-based Method

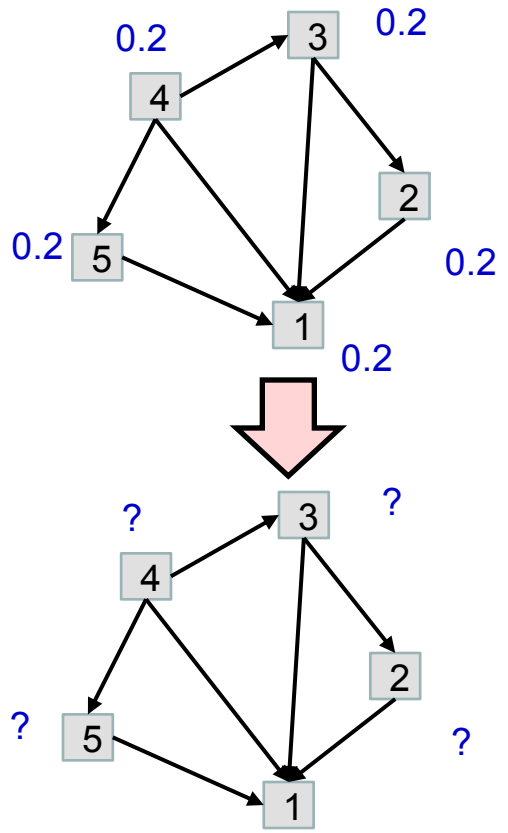
- Let us begin with PageRank[1]

$$\mathbf{r} = (1-\alpha)\mathbf{M}\cdot\mathbf{r} + \alpha\mathbf{U}$$

$$M_{ij} = \frac{1}{\text{outdeg}(v_i)}$$

$$U_i = \frac{1}{N}$$

$$\alpha = 0.15$$



(0.2+0.2*0.5+0.2*1/3+0.2)0.85+0.15*0.2

[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

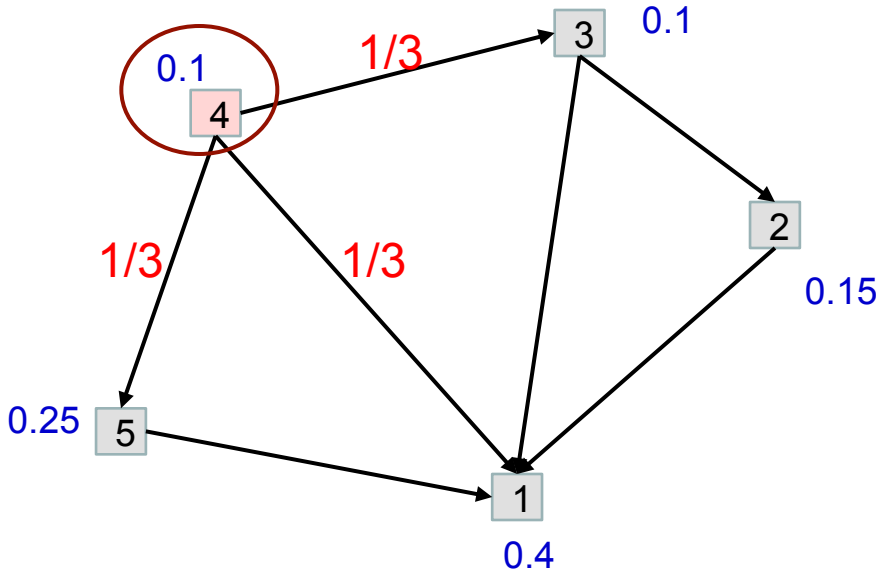
Random Walk Interpretation

- Probability distribution

$$P(t) = r$$

- Stationary distribution

$$P(t+1) = M\,P(t)$$

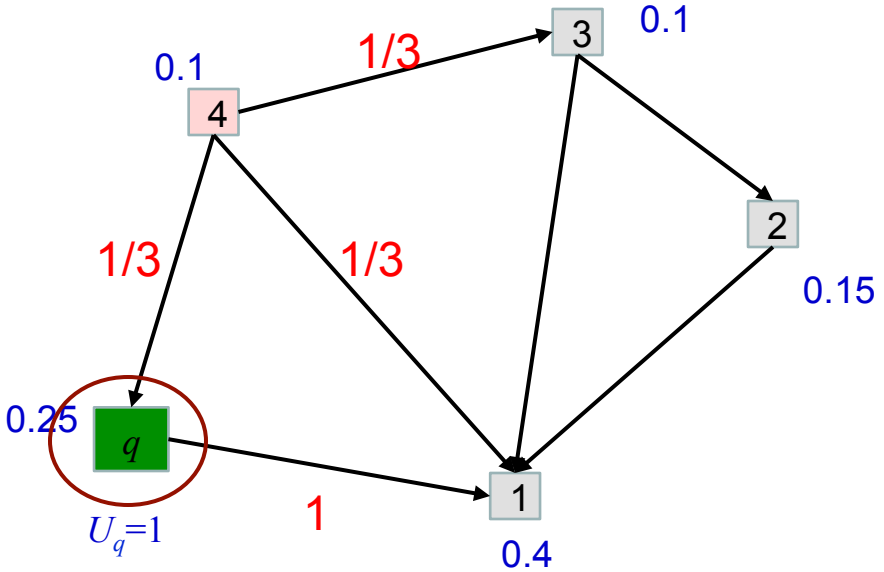


Random Walk with Restart[1]

$$\mathbf{r}_q = (1-\alpha)\mathbf{M}\cdot\mathbf{r}_q + \alpha\mathbf{U}$$

$$M_{ij} = \frac{1}{\text{outdeg}(v_i)}$$

$$U_i = \begin{cases} 1, & i = q \\ 0, & i \neq q \end{cases}$$

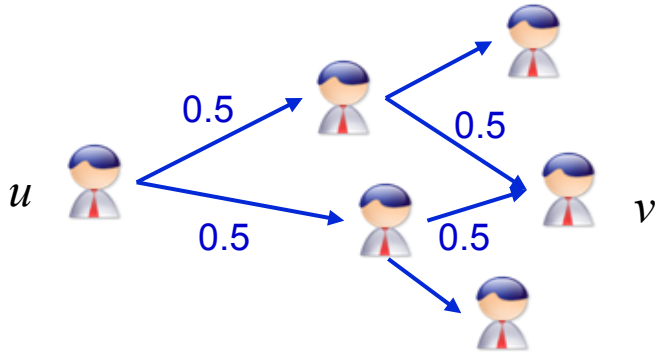


[1] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In ICDM'05, pages 418–425, 2005.

Measure Influence via Reachability[1]

- Influence of a path

$$\inf(p) = \prod_{v_i \in p} \frac{1}{\text{outdeg}(v_i)}$$



Influence(u, v)
=0.5*0.5+0.5*0.5

- Influence of user u on v

$$\text{influence}(u,v) = \lim_{t \to \infty} \sum_{p \in path_t(u,v)} \inf(p)$$

All paths from u to v within path length t

Note: The method only considers the network information and does not consider the content information

[1] G. Jeh and J. Widom. Scaling personalized web search. In WWW '03, pages 271-279, 2003.

Methodologies

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

SimRank

- SimRank is a general similarity measure, based on a simple and intuitive graph-theoretic model (Jeh and Widom, KDD'02).

C is a constant between 0 and 1, e.g., C=0.8

$$sim(u,v) = \frac{C}{|I(u)||I(v)|} \sum_{i=1}^{|I(u)|} \sum_{j=1}^{|I(v)|} sim(I_i(u), I_j(v))$$

$$\text{Initialization}: sim(u,u) = 1, \text{ if } u = v;$$

$$sim(u,v) = 0, \text{if } u \neq v.$$

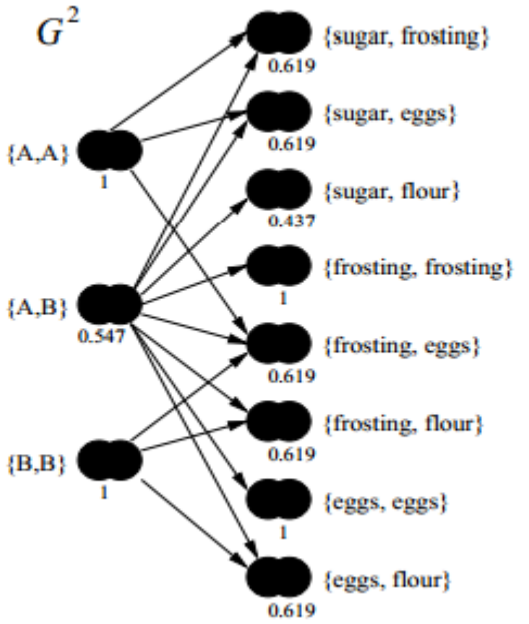
The set of pages which have inks pointing to *u*

[1] G. Jeh and J. Widom, SimRank: a measure of structural-context similarity. In KDD, pages 538-543, 2002.

Bipartite SimRank

Extend the basic SimRank equation to bipartite domains consisting of two types of objects
{A, B} and {a, b}.

E.g.,
People *A* and *B* are similar if they purchase similar items.
Items *a* and *b* are similar if they are purchased by similar people.



$$sim(A,B) = \frac{C_1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} sim(O_i(A), O_j(B))$$

$$sim(a,b) = \frac{C_2}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} sim(I_i(a), I_j(b))$$

MiniMax Variation

In some cases, e.g., course similarity, we are more care about the maximal similarity of two neighbors.

$$sim_A(A,B) = \frac{C_1}{|O(A)|}\sum_{i=1}^{|O(A)|}\max_{j=1}^{|O(B)|} sim(O_i(A),O_j(B))$$

$$sim_B(A,B) = \frac{C_1}{|O(B)|}\sum_{j=1}^{|O(B)|}\max_{i=1}^{|O(A)|} sim(O_i(A),O_j(B))$$

$$sim(A,B) = \min(sim_A(A,B), sim_B(A,B))$$

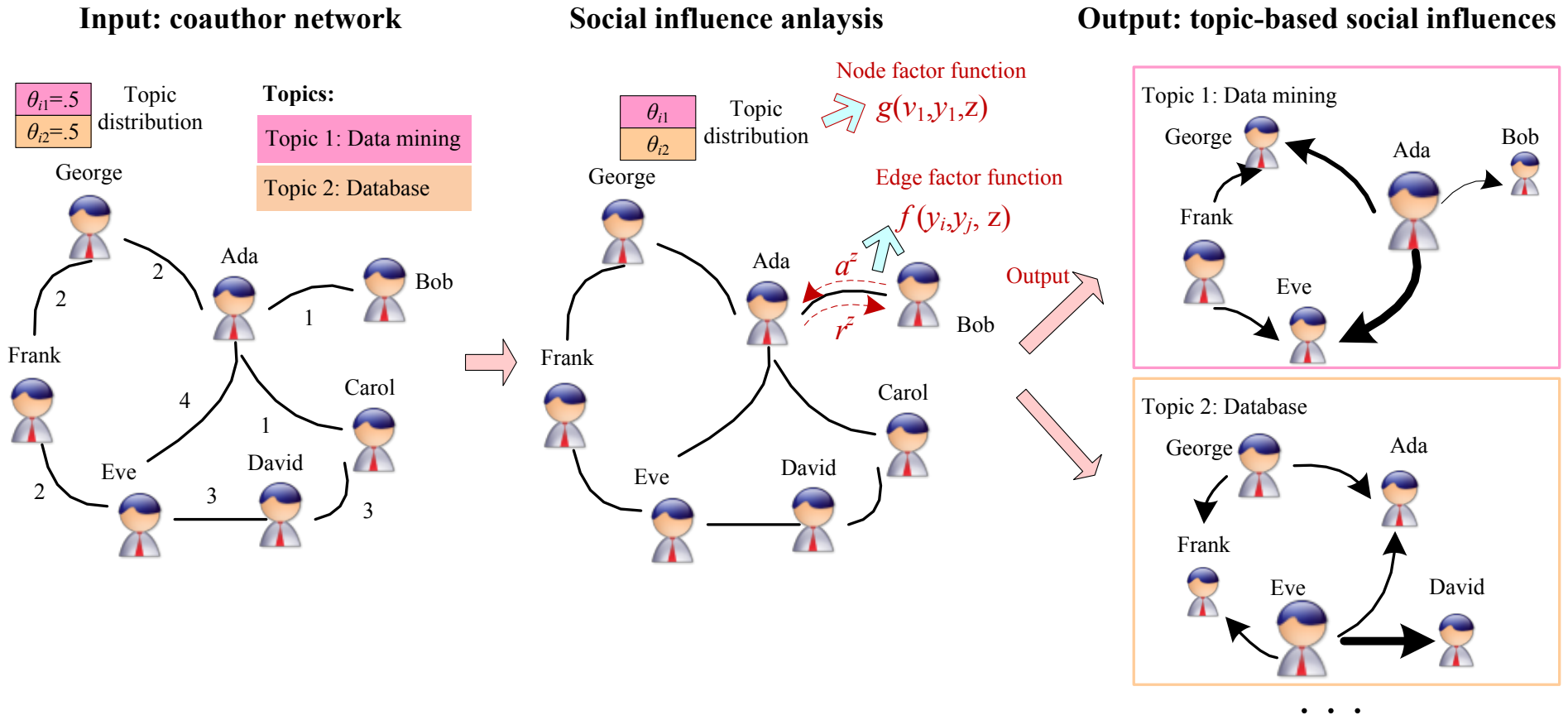
Note: Again, the method only considers the network information.

Methodologies

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

Topic-based Social Influence Analysis

- Social network -> Topical influence network



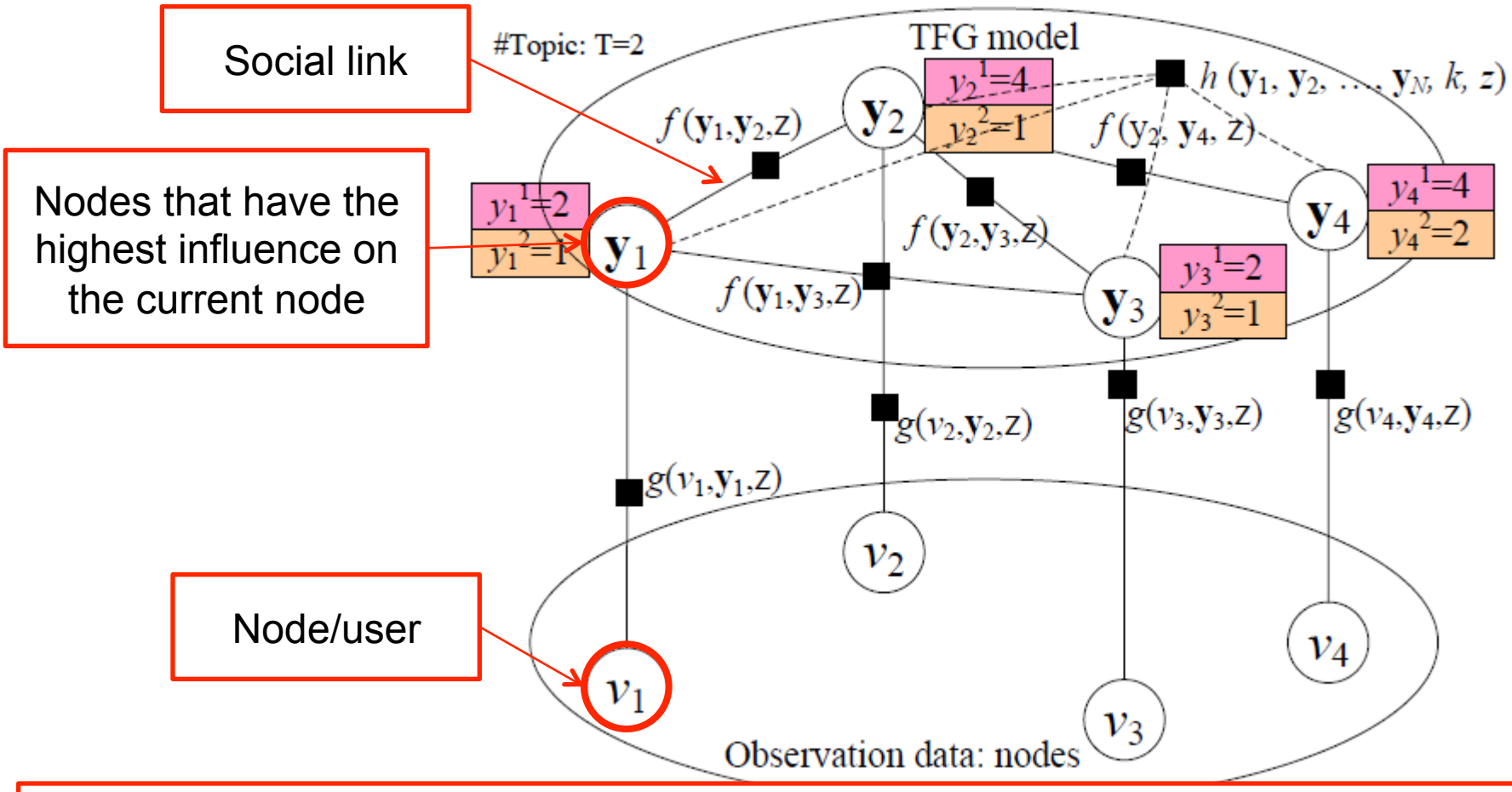
[1] J. Tang, J. Sun, C. Wang, and Z. Yang. Social Influence Analysis in Large-scale Networks. In KDD'09, pages 807-816, 2009.

The Solution: Topical Affinity Propagation

- Topical Affinity Propagation
  - Topical Factor Graph model
  - Efficient learning algorithm
  - Distributed implementation

[1] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD, pages 807-816, 2009.

Topical Factor Graph (TFG) Model



The problem is cast as identifying which node has the **highest probability** to **influence** another node on a **specific topic** along with the edge.

Topical Factor Graph (TFG)

Objective function:

$$P(\mathbf{v}, \mathbf{Y}) = \frac{1}{Z} \prod_{k=1}^{N} \prod_{z=1}^{T} h(\mathbf{y}_1, \ldots, \mathbf{y}_N, k, z) \prod_{i=1}^{N} \prod_{z=1}^{T} g(v_i, \mathbf{y}_i, z) \prod_{e_{kl} \in E} \prod_{z=1}^{T} f(\mathbf{y}_k, \mathbf{y}_l, z)$$

1. How to define?

2. How to optimize?

- The learning task is to find a configuration for all $\{y_i\}$ to maximize the joint probability.

How to define (topical) feature functions?

– Node feature function

similarity

$$g(v_i, \mathbf{y}_i, z) = \begin{cases} \dfrac{w^z_{i y^z_i}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i \neq i \\ \dfrac{\sum_{j \in NB(i)} w^z_{ji}}{\sum_{j \in NB(i)} (w^z_{ij} + w^z_{ji})} & y^z_i = i \end{cases}$$

– Edge feature function

$$f(y_i, y_j) = \begin{cases} w[v_i \sim v_j] & y_i = y_j \\ 1 - w[v_i \sim v_j] & y_i \neq y_j \end{cases}$$

or simply binary

– Global feature function

$$h(\mathbf{y}_1, \ldots, \mathbf{y}_N, k, z) = \begin{cases} 0 & \text{if } y^z_k = k \text{ and } y^z_i \neq k \text{ for all } i \neq k \\ 1 & \text{otherwise.} \end{cases}$$

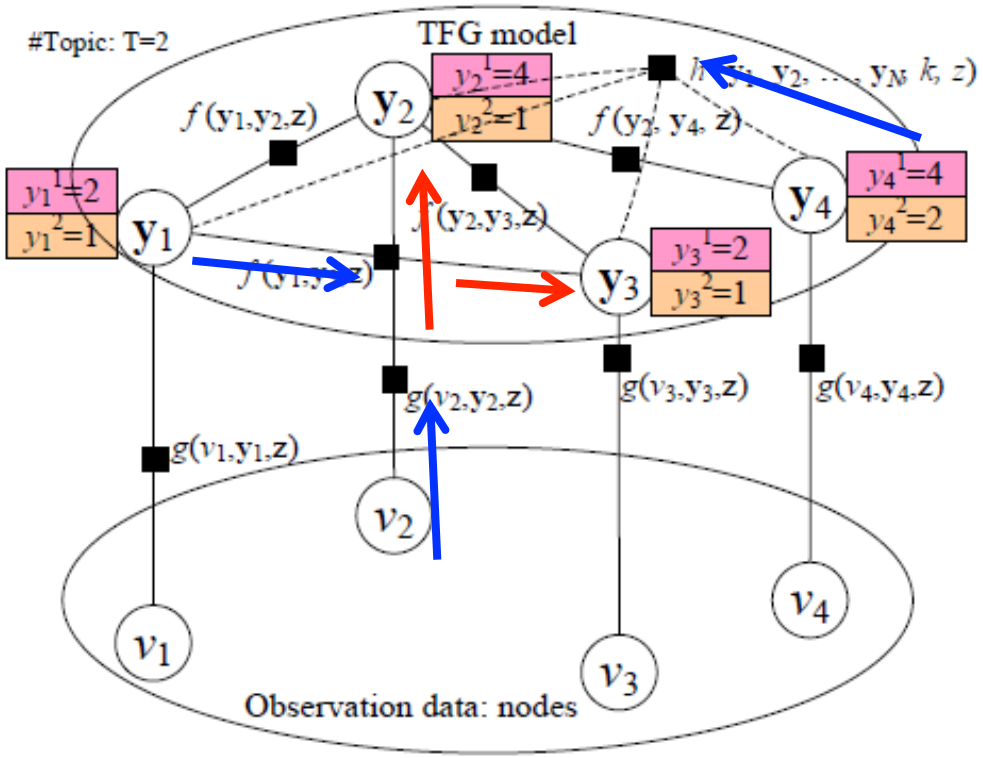
Model Learning Algorithm

Sum-product:

$$m_{y\to f}(y,z) = \prod_{f'\sim y\setminus f} m_{f'\to y}(y,z) \prod_{z'\neq z} \prod_{f'\sim y\setminus f} m_{f'\to y}(y,z')^{(\tau_{z'z})}$$

$$m_{f\to y}(y,z) = \sum_{\sim\{y\}} \left(f(Y,z) \prod_{y'\sim f\setminus y} m_{y'\to f}(y',z) \right)$$

$$+ \sum_{z'\neq z} \tau_{z'z} \sum_{\sim\{y\}} \left(f(Y,z') \prod_{y'\sim f\setminus y} m_{y'\to f}(y',z') \right) \quad (4)$$

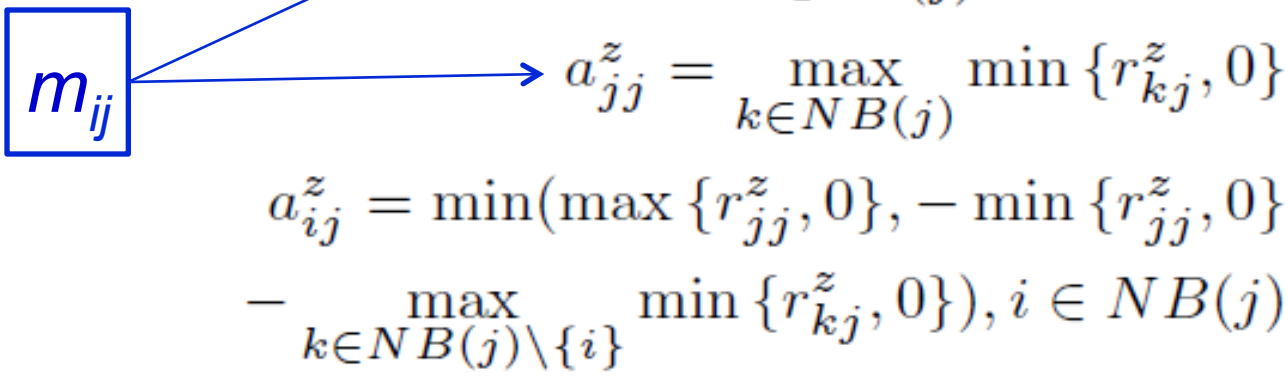


- Low efficiency!
- Not easy for distributed learning!

New TAP Learning Algorithm

1. Introduce two new variables *r* and *a, to* replace the original message *m*.

2. Design new update rules:



[1] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD, pages 807-816, 2009.

The TAP Learning Algorithm

Input: $G = (V, E)$ and topic distributions $\{\theta_v\}_{v \in V}$
Output: topic-level social influence graphs $\{G_z = (V_z, E_z)\}^T$

1.1 Calculate the node feature function $g(v_i, \mathbf{y}_i, z)$;
1.2 Calculate b_{ij}^z according to Eq. 8;
1.3 Initialize all $\{r_{ij}^z\} \leftarrow 0$;
1.4 **repeat**
1.5 **foreach** *edge-topic pair* (e_{ij}, z) **do**
1.6 Update r_{ij}^z according to Eq. 5;
1.7 **end**
1.8 **foreach** *node-topic pair* (v_j, z) **do**
1.9 Update a_{jj}^z according to Eq. 6;
1.10 **end**
1.11 **foreach** *edge-topic pair* (e_{ij}, z) **do**
1.12 Update a_{ij}^z according to Eq. 7;
1.13 **end**
1.14 **until** *convergence*;
1.15 **foreach** *node* v_t **do**
1.16 **foreach** *neighboring node* $s \in NB(t) \cup \{t\}$ **do**
1.17 Compute μ_{st}^z according to Eq. 9;
1.18 **end**
1.19 **end**
1.20 Generate $G_z = (V_z, E_z)$ for every topic z according to $\{\mu_{st}^z\}$;

$$b_{ij}^z = \log \frac{g(v_i, \mathbf{y}_i, z)|_{y_i^z = j}}{\sum_{k \in NB(i) \cup \{i\}} g(v_i, \mathbf{y}_i, z)|_{y_i^z = k}}$$

$$r_{ij}^z = b_{ij}^z - \max_{k \in NB(j)} \{b_{ik}^z + a_{ik}^z\}$$

$$a_{jj}^z = \max_{k \in NB(j)} \min\{r_{kj}^z, 0\}$$

$$a_{ij}^z = \min(\max\{r_{jj}^z, 0\}, -\min\{r_{jj}^z, 0\} - \max_{k \in NB(j) \setminus \{i\}} \min\{r_{kj}^z, 0\}), i \in NB(j)$$

$$\mu_{st}^z = \frac{1}{1 + e^{-(r_{ts}^z + a_{ts}^z)}}$$

Distributed TAP Learning

- Map-Reduce
  - Map: (key, value) pairs
    - $e_{ij}/a_{ij} \rightarrow e_{i*}/a_{ij}$; $e_{ij}/b_{ij} \rightarrow e_{i*}/b_{ij}$; $e_{ij}/r_{ij} \rightarrow e_{*j}/r_{ij}$.
  - Reduce: (key, value) pairs
    - e_{ij} / * → new r_{ij}; e_{ij}/* → new a_{ij}
- For the global feature function

THEOREM 1. *If the global feature function h can be factorized into $h = \prod_{k=1}^{N} h_k$, for every $i \in \{1, \ldots, N\}, y_i \neq k, y_i' \neq k, h_k(y_1, \ldots, y_i, \ldots, y_N) = h_k(y_1, \ldots, y_i', \ldots, y_N)$, then the message passing update rules can be simplified to influence update rules.* ■

Experiments

- Data set: (http://arnetminer.org/lab-datasets/soinf/)

Data set	#Nodes	#Edges
Coauthor	640,134	1,554,643
Citation	2,329,760	12,710,347
Film (Wikipedia)	18,518 films 7,211 directors 10,128 actors 9,784 writers	142,426

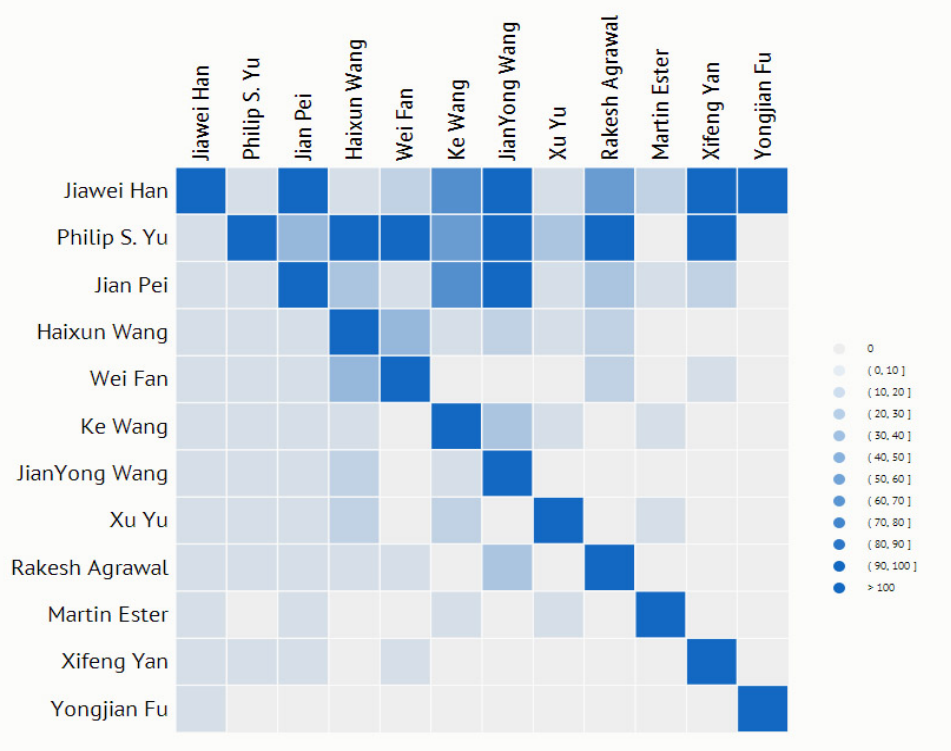
- Evaluation measures
  - CPU time
  - Case study
  - Application

Social Influence Sub-graph on “Data mining”

Table 4: Dynamic influence analysis for Dr. Jian Pei during 2000-2009. Due to space limitation, we only list coauthors who most influence on/by Dr. Pei in each time window.

Year	Pairwise	Influence
2000 - 2001	Influence on Dr. Pei	Jiawei Han (0.4961)
	Influenced by Dr. Pei	Jiawei Han (0.0082)
2002 - 2003	Influence on Dr. Pei	Jiawei Han (0.4045), Ke Wang (0.0418), Jianyong Wang (0.019), Xifeng Yan (0.007), Shiwei Tang (0.0052)
	Influenced by Dr. Pei	Shiwei Tang (0.436), Hasan M.Jamil (0.4289), Xifeng Yan (0.2192), Jianyong Wang (0.1667), Ke Wang (0.0687)
2004 - 2005	Influence on Dr. Pei	Jiawei Han (0.2364), Ke Wang (0.0328), Wei Wang (0.0294), Jianyong Wang (0.0248), Philip S. Yu (0.0156)
	Influenced by Dr. Pei	Chun Tang (0.5929), Shiwei Tang (0.5426), Hasan M.Jamil (0.3318), Jianyong Wang (0.1609), Xifeng Yan (0.1458), Yan Huang (0.1054)
2006 - 2007	Influence on Dr. Pei	Jiawei Han (0.1201), Ke Wang (0.0351), Wei Wang (0.0226), Jianyong Wang (0.018), Ada Wai-Chee Fu (0.0125)
	Influenced by Jian Pei	Chun Tang (0.6095), Shiwei Tang (0.6067), Byung-Won On (0.4599), Hasan M.Jamil (0.3433), Jaewoo Kang (0.3386)
2008 - 2009	Influence on Dr. Pei	Jiawei Han (0.2202), Ke Wang (0.0234), Ada Wai-Chee Fu (0.0208), Wei Wang (0.011), Jianyong Wang (0.0095)
	Influenced by Dr. Pei	ZhaoHui Tang (0.654), Chun Tang (0.6494), Shiwei Tang (0.5923), Zhengzheng Xing (0.5549), Hasan M.Jamil (0.3333), Jaewoo Kang (0.3057)

On “Data Mining” in 2009



Results on Coauthor and Citation

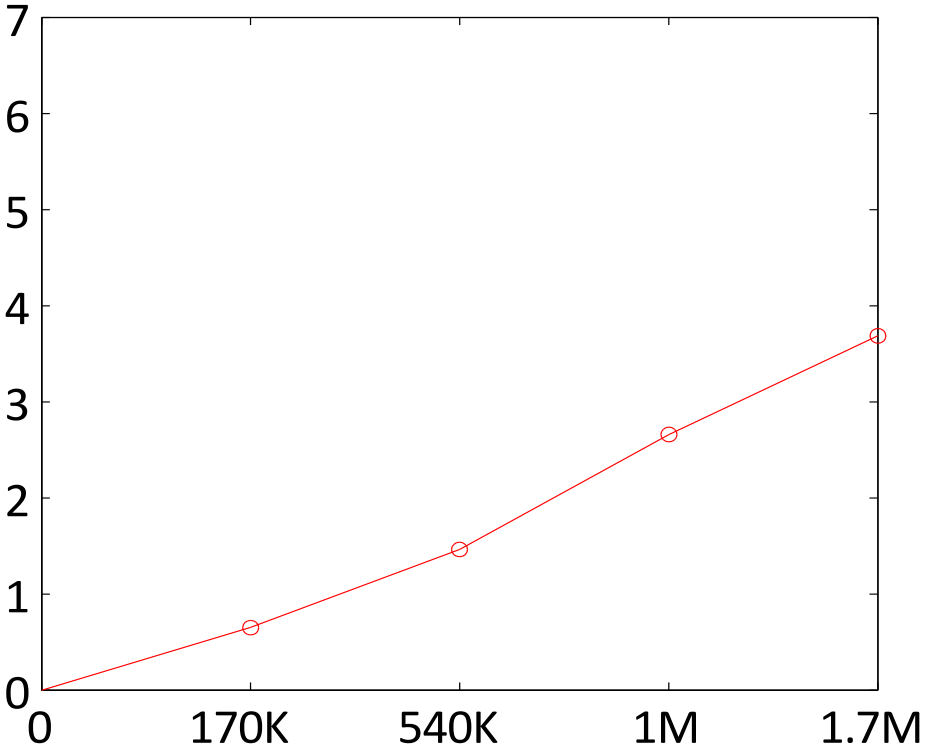
Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell, C. Lee Giles, Terrence J. Sejnowski, Samy Bengio, Daphne Koller, Luc De Raedt, Vasant Honavar, Floriana Esposito, Bernhard Scholkopf
	Database System	Gerhard Weikum, John Mylopoulos, Michael Stonebraker, Barbara Pernici, Philip S. Yu, Sharad Mehrotra, Wei Sun, V. S. Subrahmanian, Alejandro P. Buchmann, Kian-Lee Tan, Jiawei Han
	Information Retrieval	Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Frieder, Alan F. Smeaton, Rong Jin
	Web Services	Yan Wang, Liang-jie Zhang, Schahram Dustdar, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah
	Semantic Web	Wolfgang Nejdl, Daniel Schwabe, Steffen Staab, Mark A. Musen, Andrew Tomkins, Juliana Freire, Carole A. Goble, James A. Hendler, Rudi Studer, Enrico Motta
	Bayesian Network	Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets
Citation	Data Mining	Fast Algorithms for Mining Association Rules in Large Databases, Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Discovery of Multiple-Level Association Rules from Large Databases, Interleaving a Join Sequence with Semijoins in Distributed Query Processing
	Machine Learning	Object Recognition with Gradient-Based Learning, Correctness of Local Probability Propagation in Graphical Models with Loops, A Learning Theorem for Networks at Detailed Stochastic Equilibrium, The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, A Unifying Review of Linear Gaussian Models
	Database System	Mediators in the Architecture of Future Information Systems, Database Techniques for the World-Wide Web: A Survey, The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles, Fast Algorithms for Mining Association Rules in Large Databases
	Web Services	The Web Service Modeling Framework WSMF, Interval Timed Coloured Petri Nets and their Analysis, The design and implementation of real-time schedulers in RED-linux, The Self-Serv Environment for Web Services Composition
	Web Mining	Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Fast Algorithms for Mining Association Rules in Large Databases, The OO-Binary Relationship Model: A Truly Object Oriented Conceptual Model, Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations, Improving Fault Tolerance and Supporting Partial Writes in Structured Coterie Protocols for Replicated Objects
	Semantic Web	FaCT and iFaCT, The GRAIL concept modelling language for medical terminology, Semantic Integration of Semistructured and Structured Data Sources, Description of the RACER System and its Applications, DL-Lite: Practical Reasoning for Rich Dls

Scalability Performance

Table 2: Scalability performance of different methods on real data sets. >10hr means that the algorithm did not terminate when the algorithm runs more than 10 hours.

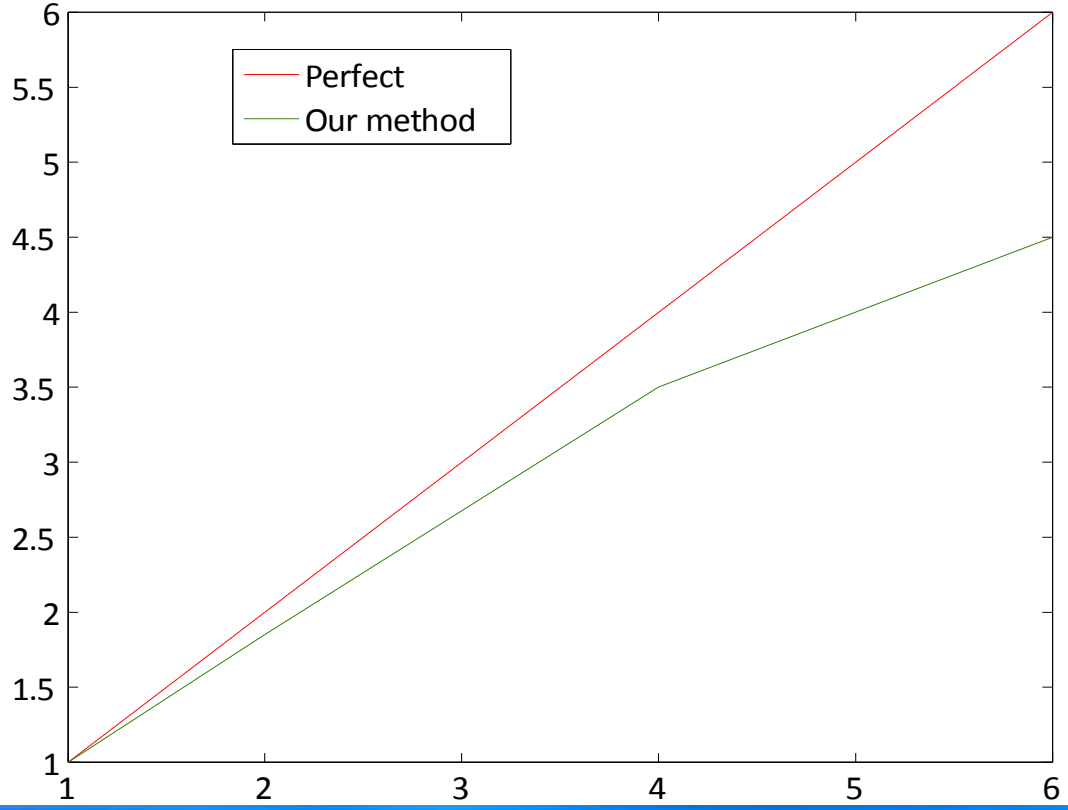
Methods	Citation	Coauthor	Film
Sum-Product	N/A	>10hr	1.8 hr
Basic TAP Learning	>10hr	369s	**57s**
Distributed TAP Learning	**39.33m**	**104s**	148s

Speedup results

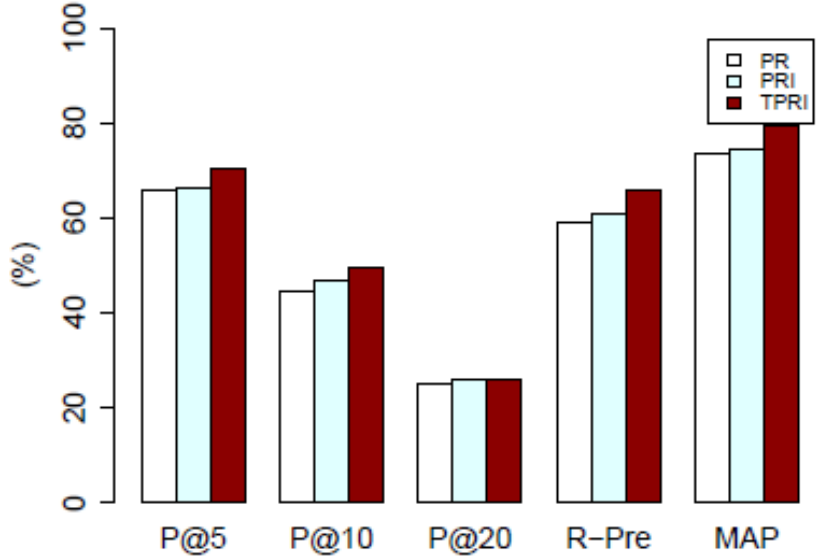


Speedup vs. Dataset size

Speedup vs. #Computer nodes



Application—Expert Finding[1]



Note: Well though this method can combine network and content information, it does not consider users' action.

Table 7: Performance of expert finding with different approaches.

Expert finding data from
http://arnetminer.org/lab-datasets/expertfinding/

[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In KDD'08, pages 990-998, 2008.

Methodologies

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

Influence and Action

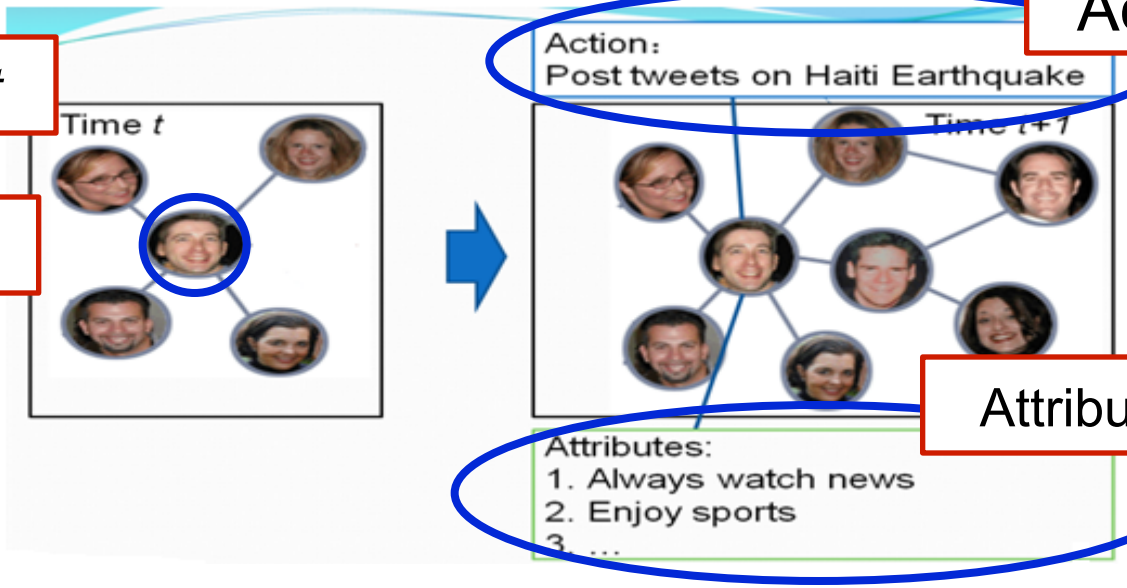
$G^t = (V^t, E^t, X^t, Y^t)$

Nodes at time *t*

Edges at time *t*

Actions at time *t*

Attribute matrix at time *t*



Input:

$G^t = (V^t, E^t, X^t, Y^t)$

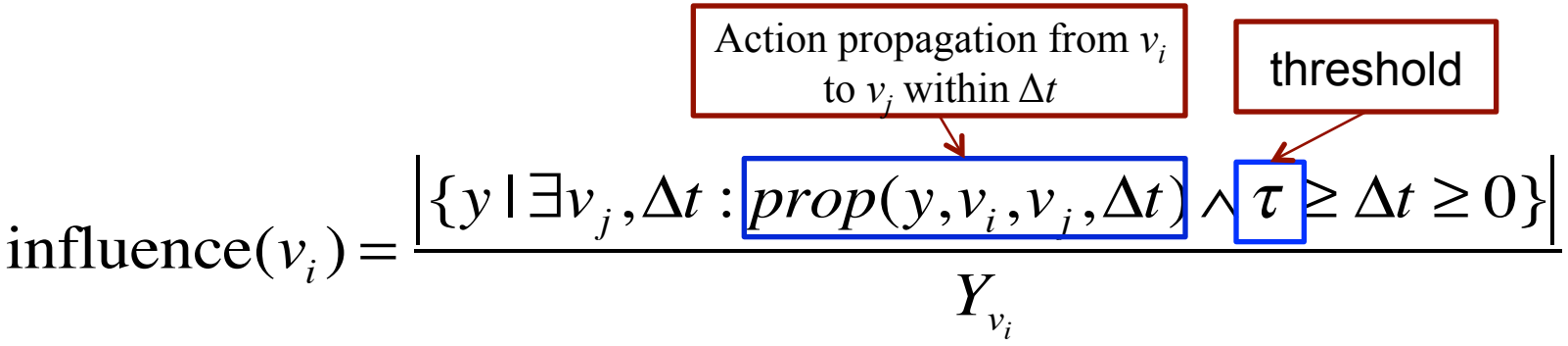
$t = 1,2,\ldots T$

Output:

$F: f(G^t) \rightarrow Y^{(t+1)}$

(a) Learning Influence Probabilities [1]

- **Goal:** Learn user influence and action influence from historical actions
- **Assumption**
  - If user v_i performs an action y at time t and later his friend v_j also perform the action, then there is an influence from v_i to v_j
- **User Influenceability: quantifies how influenceable a user is.**



where $\Delta t = t_j - t_i$ is the difference between the time when v_j performing the action and the time when user v_i performing the action, given e_{ij}=1.

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In WSDM'10, pages 207–217, 2010.

(a) Learning Influence Probabilities [1]

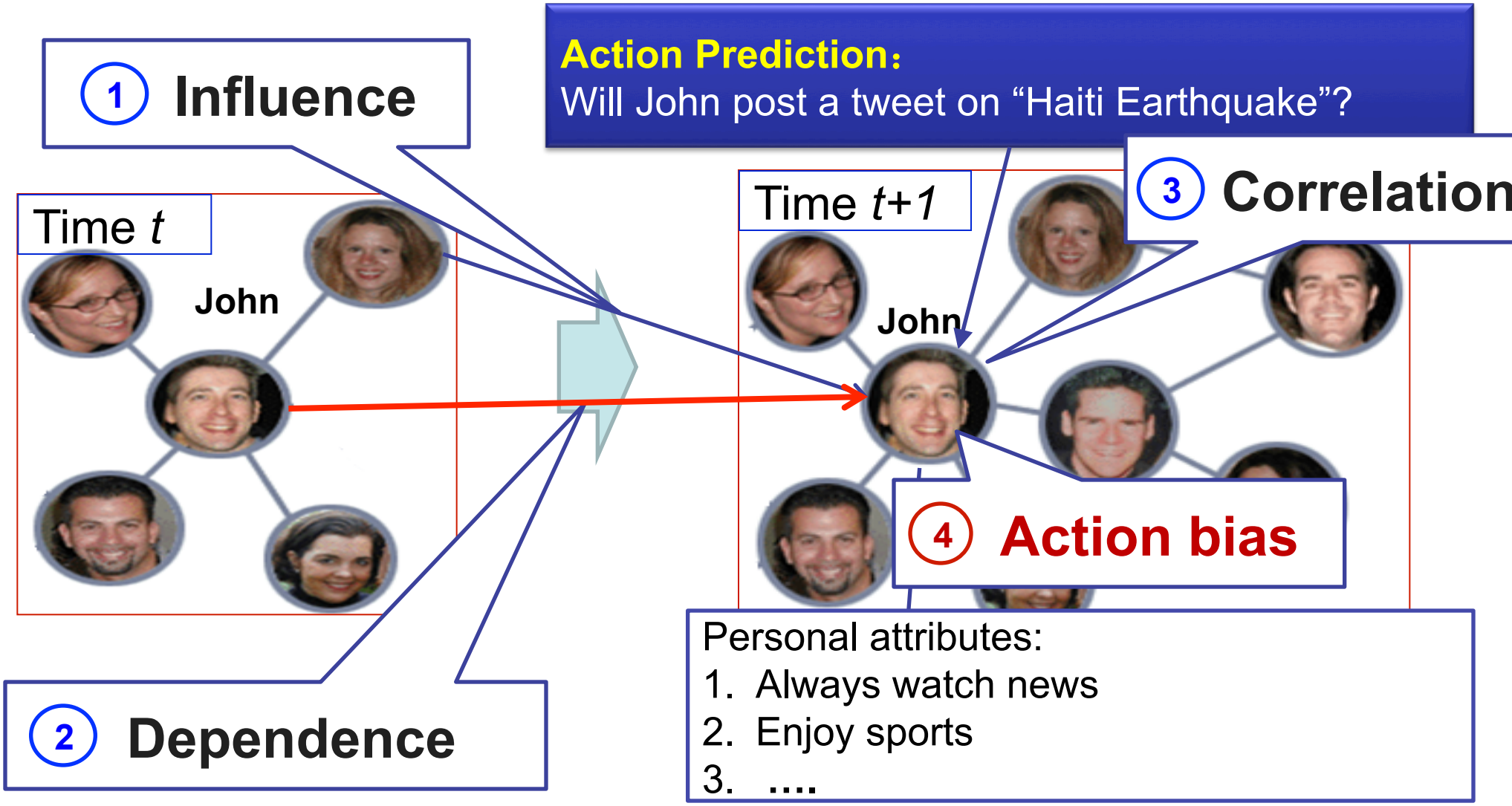
- **Action Influenceability: quantify how influenceable an action is.**

$$\text{influence}(y) = \frac{\left|\{v_j \mid \exists v_i, \Delta t : prop(y, v_i, v_j, \Delta t) \wedge \tau \geq \Delta t \geq 0\}\right|}{\text{number of users performing } y}$$

where $\Delta t = t_j - t_i$ is the difference between the time when v_j performing the action and the time when user v_i performing the action, given e_{ij}=1; $prop(y, v_i, v_j, \Delta t)$ represents the action propagation score

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In WSDM'10, pages 207–217, 2010.

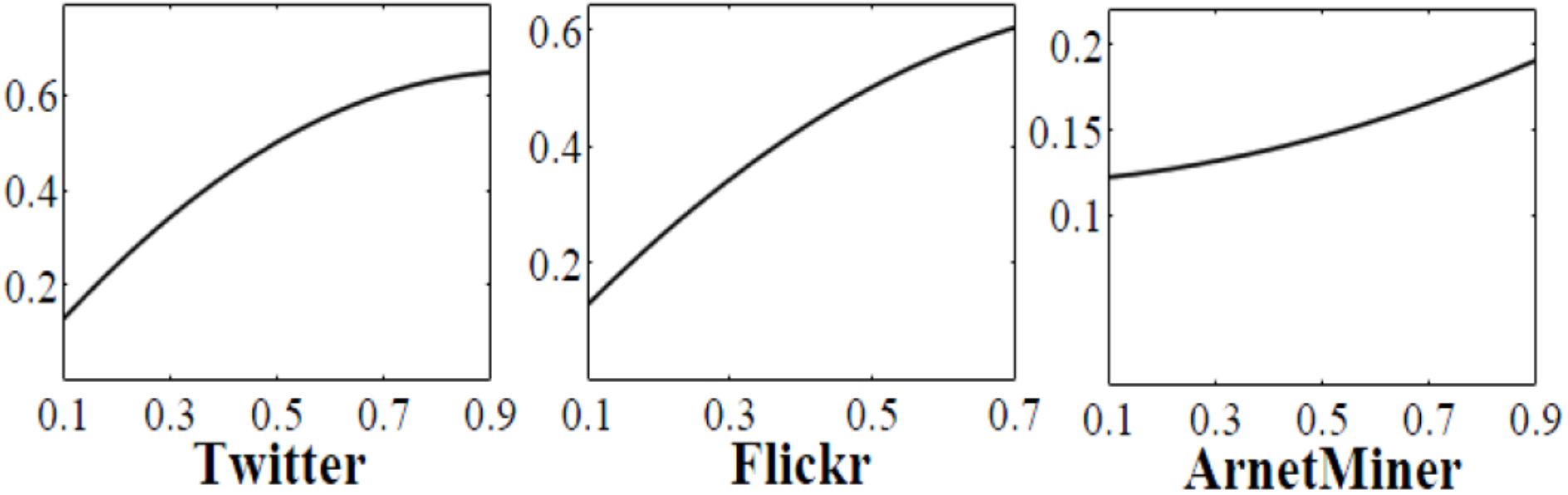
(b) Social Influence & Action Modeling[1]



[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In KDD’10, pages 807–816, 2010.

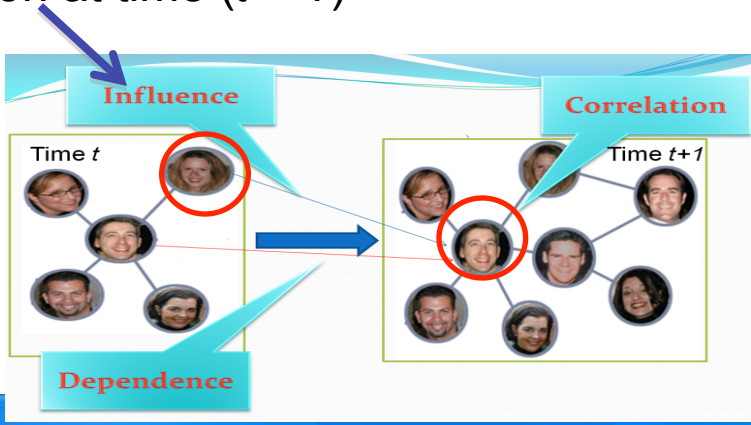
Statistical Study: Influence

Y-axis: the likelihood that the user also performs the action at time *t*



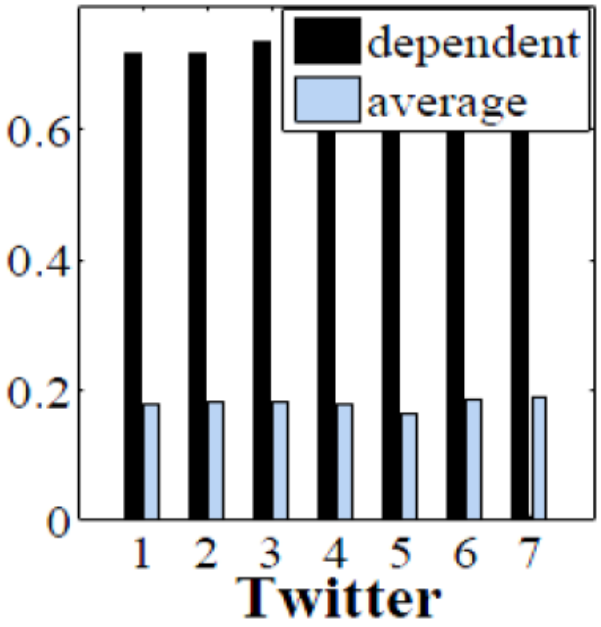
X-axis: the percentage of one's friends who perform an action at time (*t* − *1*)

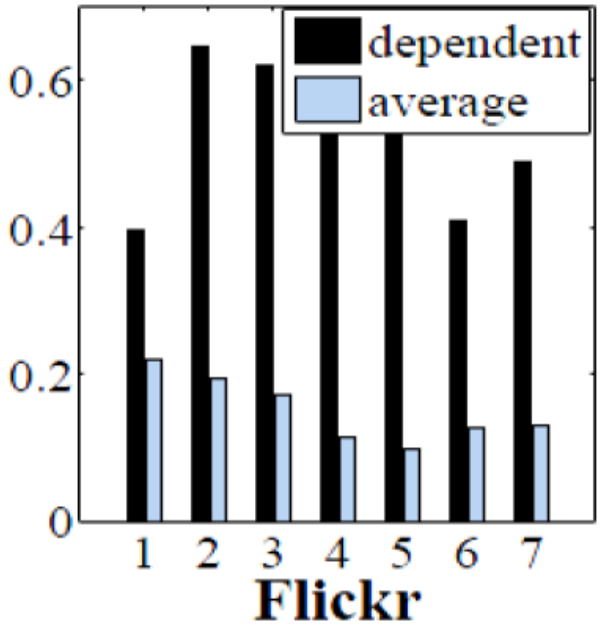
Twitter Action: Tweet on "Haiti Earthquake"
Flickr Action: Add a picture into favorite list
ArnetMiner Action: Publish on a conference

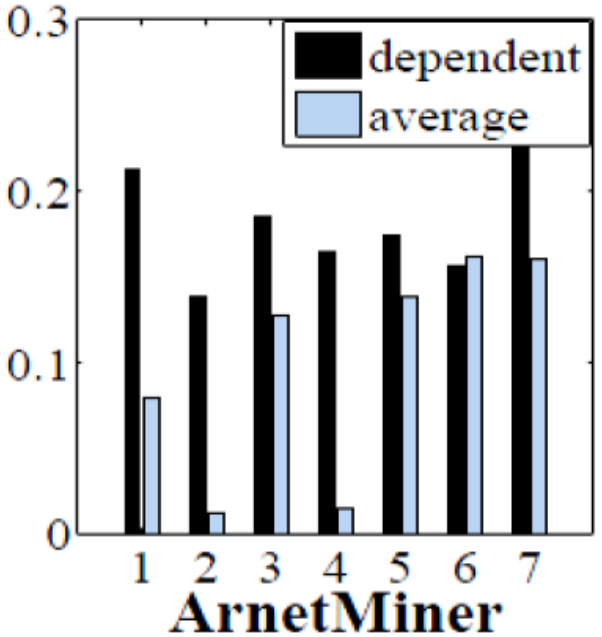


Statistical Study: Dependence

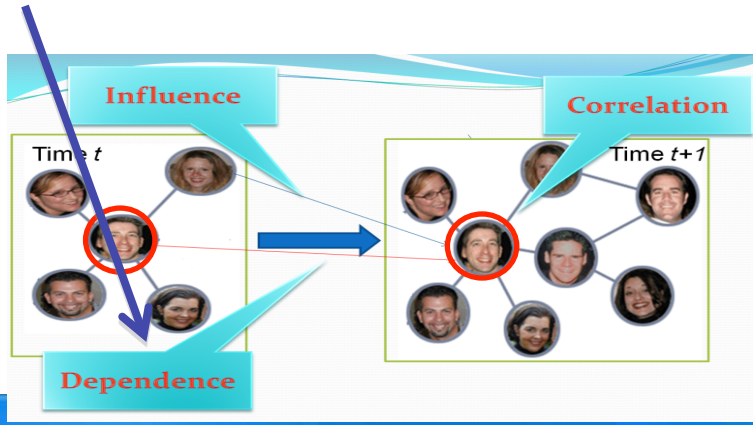
Y-axis: the likelihood that a user performs an action





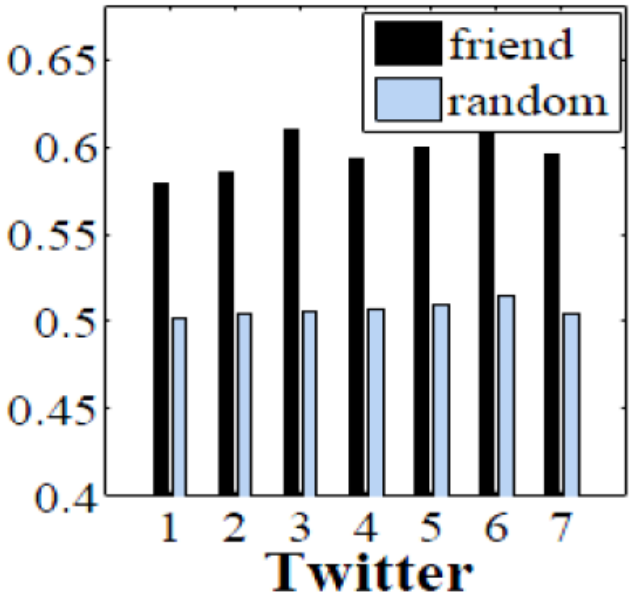


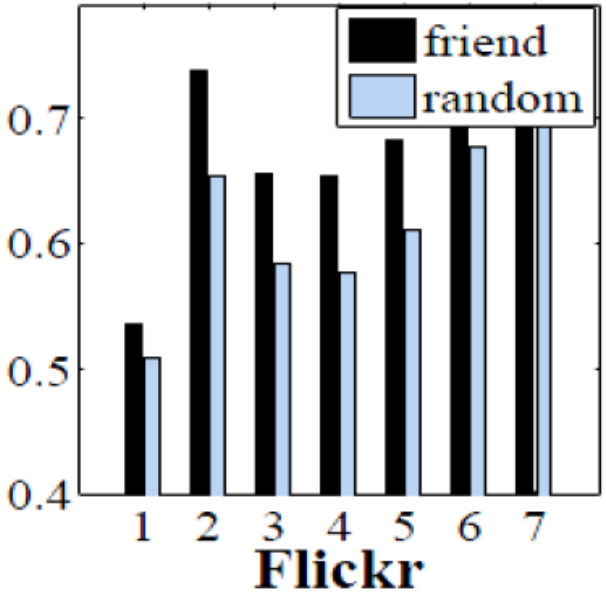
X-axis: different time windows (1-7)

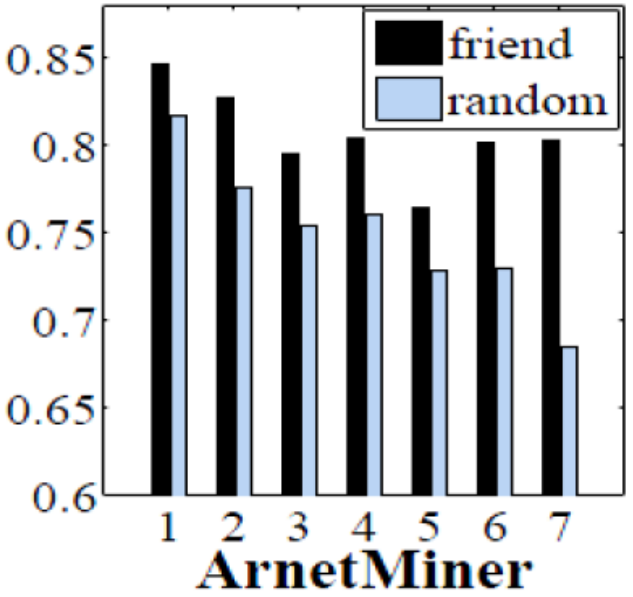


Statistical Study: Correlation

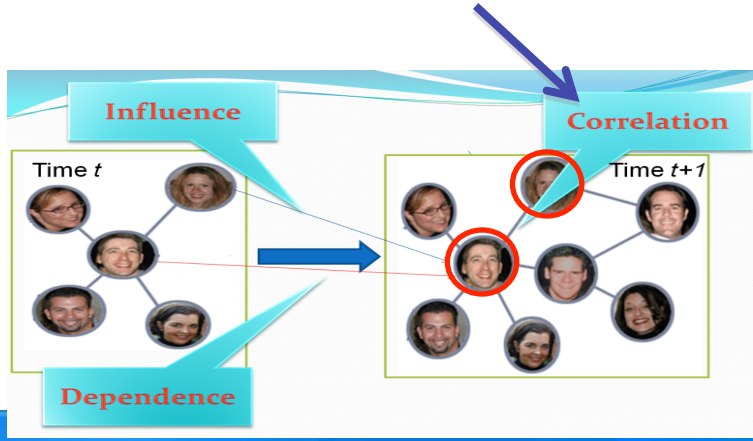
Y-axis: the likelihood that two friends(random) perform an action together



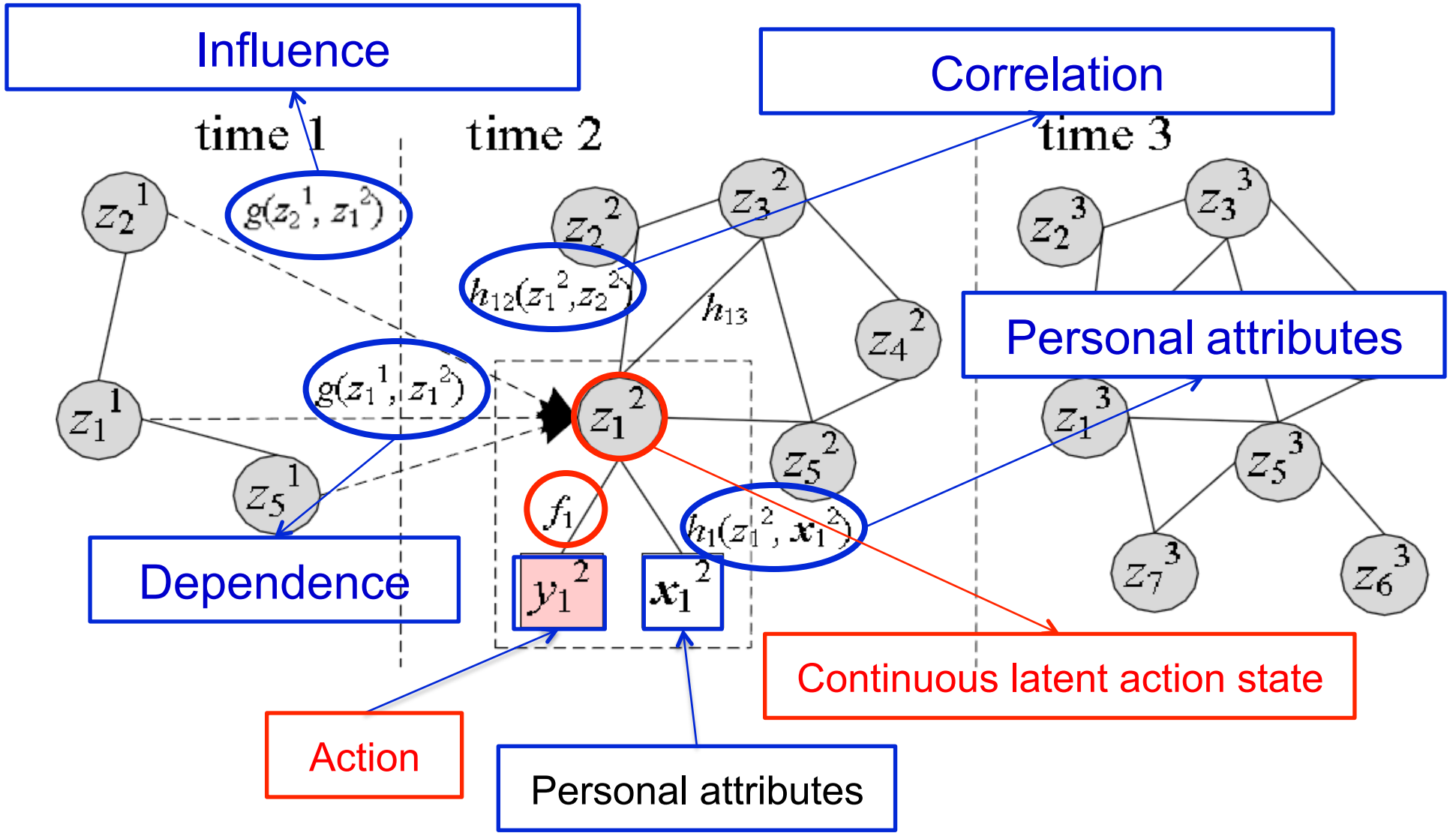




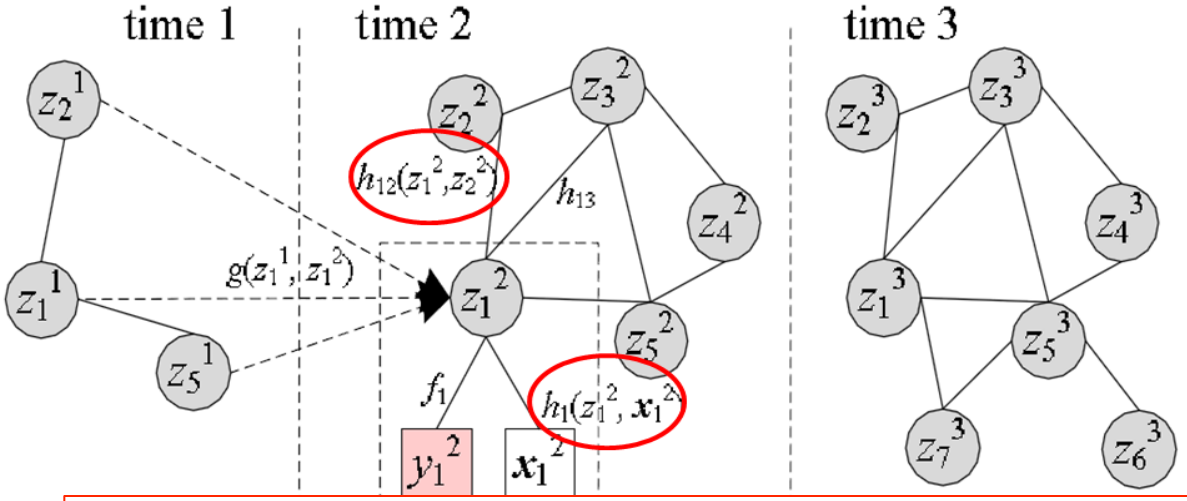
X-axis: different time windows (1-7)



A Discriminative Model: NTT-FGM



Model Instantiation



$$
\begin{aligned}
g_{ji}(z_i^t, z_j^{t-1}) &= -(z_i^t - z_j^{t-1})^2 \\
h_{ij}(z_i^t, z_j^t) &= -(z_i^t - z_j^t)^2 \\
h_k(z_i^t, x_{ik}^t) &= -(z_i^t - x_{ik}^t)^2
\end{aligned}
$$

How to estimate the parameters?

$$
p(\mathbf{Y}|\mathbf{G}) = \frac{1}{Z}\exp\{\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{(y_i^t - z_i^t)^2}{2\sigma^2} + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij} m_{ji}^{t-1} g(z_i^t, z_j^{t-1})
$$

$$
+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_{ij} m_{ij}^t h_{ij}(z_i^t, z_j^t) + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{d}\alpha_k h_k(z_i^t, x_{ik}^t)\}
$$

Model Learning—Two-step learning

Input: number of iterations I and learning rate η;
Output: learned parameters $\theta = (\{z_i\}, \{\alpha_k\}, \{\beta_{ij}\}, \{\lambda_{ij}\})$;

Initialize $\mathbf{z} = \mathbf{y}$;
Initialize α, β, λ;
repeat
  E Step: % fix $\mathbf{z}$, learn α, β, λ;
  for $i = 1$ *to* I **do**
    Compute gradient $\nabla_{\log \alpha_k}, \nabla_{\log \beta_{ij}}, \nabla_{\log \lambda_{ij}}$;
    Update $\log \alpha_k = \log \alpha_k + \eta \times \nabla_{\log \alpha_k}$;
    Update $\log \beta_{ij} = \log \beta_{ij} + \eta \times \nabla_{\log \beta_{ij}}$;
    Update $\log \lambda_{ij} = \log \lambda_{ij} + \eta \times \nabla_{\log \lambda_{ij}}$;
  end
  M Step: % fix α, β, λ learn $\mathbf{z}$;
    Solve the following linear equation:

$$(A + \mathbf{I})\mathbf{z} = \mathbf{y} + X\alpha$$

until *convergence*;

[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In KDD'10, pages 807–816, 2010.

Experiment

- Data Set (http://arnetminer.org/stnt)

	Action	Nodes	#Edges	Action Stats
Twitter	Post tweets on "Haiti Earthquake"	7,521	304,275	730,568
Flickr	Add photos into favorite list	8,721	485,253	485,253
Arnetminer	Issue publications on KDD	2,062	34,986	2,960

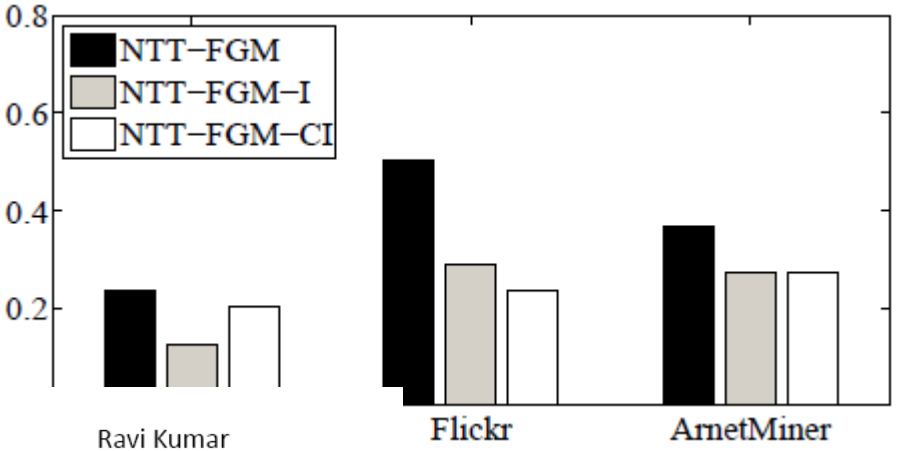
- Baseline
  - SVM
  - wvRN (Macskassy, 2003)
- Evaluation Measure:

  Precision, Recall, F1-Measure

Results

Table 1: Performance of action prediction with different approaches (%).

Data set	Method	Recall	Precision	F1-Measure
Twitter	SVM	10.41	16.71	13.85
	wvRN	0.45	7.89	0.86
	NTT-FGM	26.40	21.14	23.47
Flickr	SVM	34.48	45.05	39.06
	wvRN	60.02	48.81	53.84
	NTT-FGM			
ArnetMiner	SVM			
	wvRN			
	NTT-FGM			



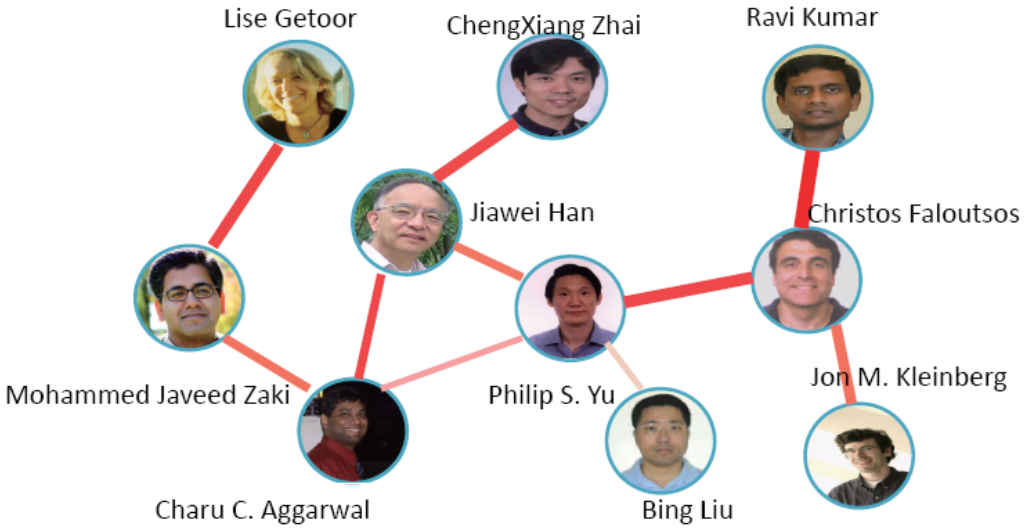
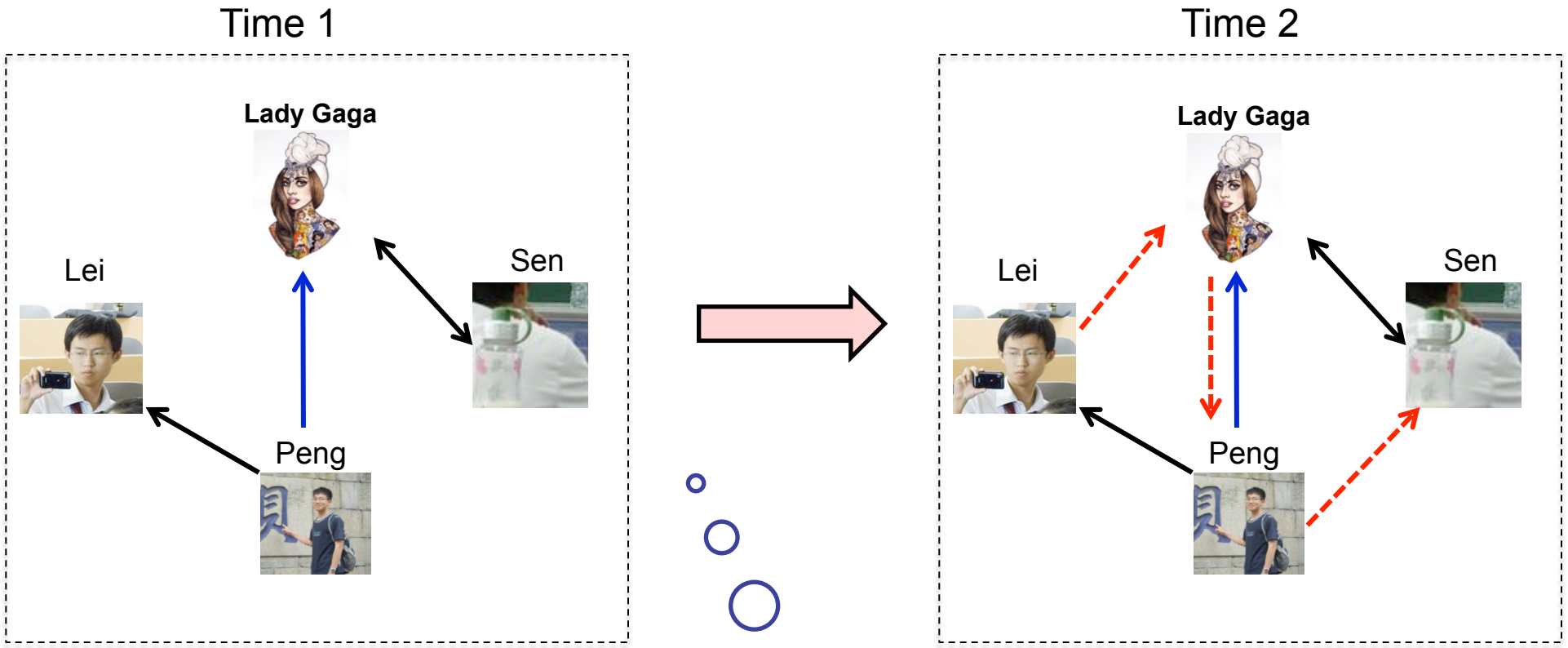


Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.

Measuring Following Influence -A Generative Model

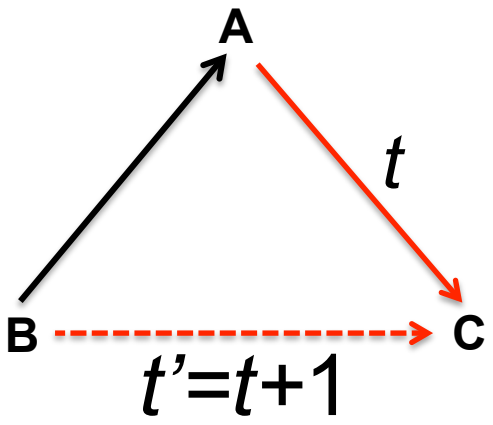
Measuring Following Influence



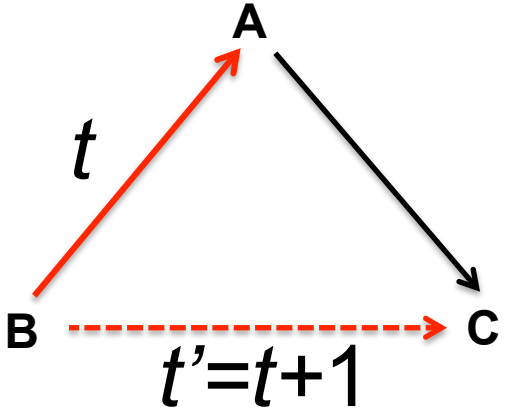
When you follow a user in a social network, will the behavior influences your friends to also follow her?

Recall we defined two kinds of influence..

Two Categories of Following Influences



Follower diffusion

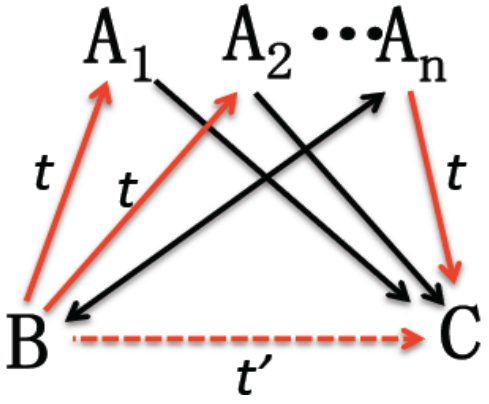


Followee diffusion

–>: pre-existed relationships
–>: a new relationship added at *t*
-->: a possible relationship added at *t*+1

A Generative Model: FCM

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 cascaded model)

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

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.

$$\theta = \{p_{ee'}\} \quad \Rightarrow \quad \theta = \{p_{\triangle}\}$$

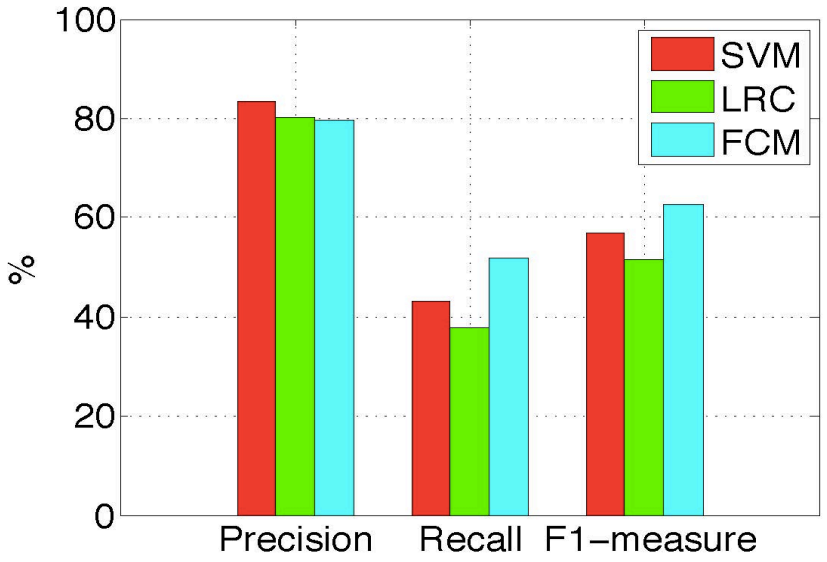
Input: network $G = (V, E, t)$
Output: $\theta = \{p_{\triangle}\}$
while *not converged* **do**
  E-step : calculate $p_{e'}$ using $p_{e'} = 1 - \prod_{e \in N^-(e')} (1 - \hat{p}_{\triangle})$
  M-step: calculate $p_{\triangle}$ using $p_{\triangle} = \frac{1}{|\triangle^A| + |\triangle^U|} \sum_{e'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

Experiments

- Improving link prediction
  - 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.
- Comparison methods
  - FCM (our approach)
  - CF
  - Katz
  - SimRank

Link Prediction Performance



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

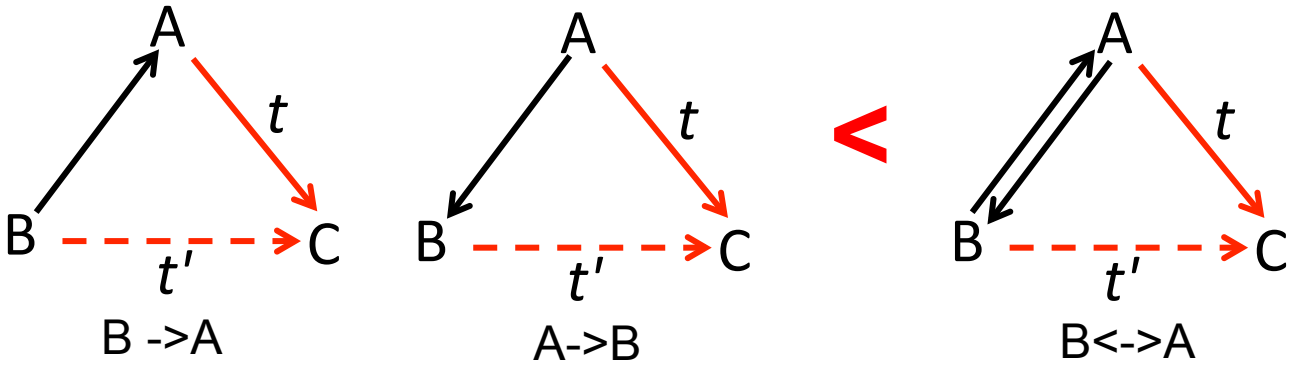
Link predction as classification

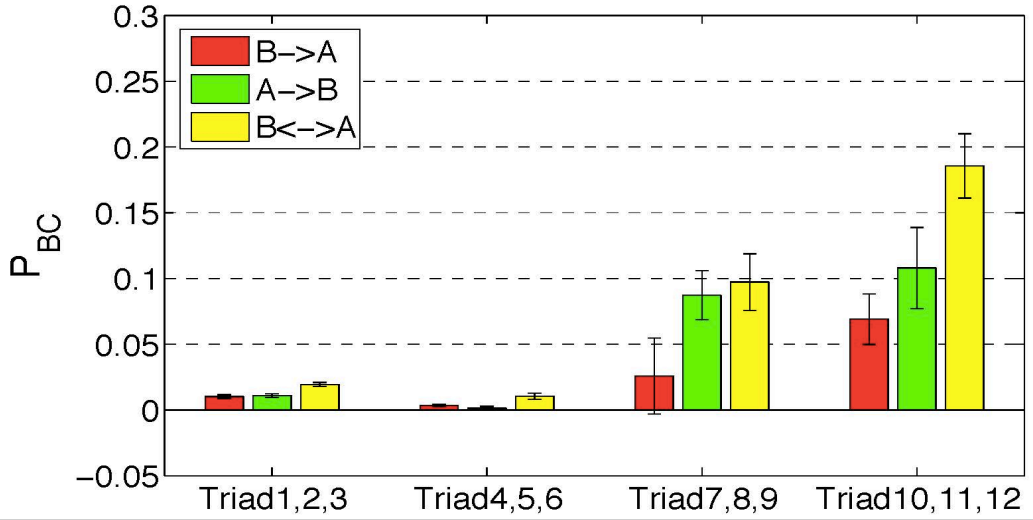
Model	P@1	P@2	P@5	P@10	MAP
CF	39.96	37.55	30.88	26.41	55.08
SimRank	26.35	26.06	26.22	24.39	44.15
Katz	46.24	41.84	32.77	26.61	59.40
FCM	72.88	55.69	37.15	27.88	77.91

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').

Follower Diffusion: Power of Reciprocity

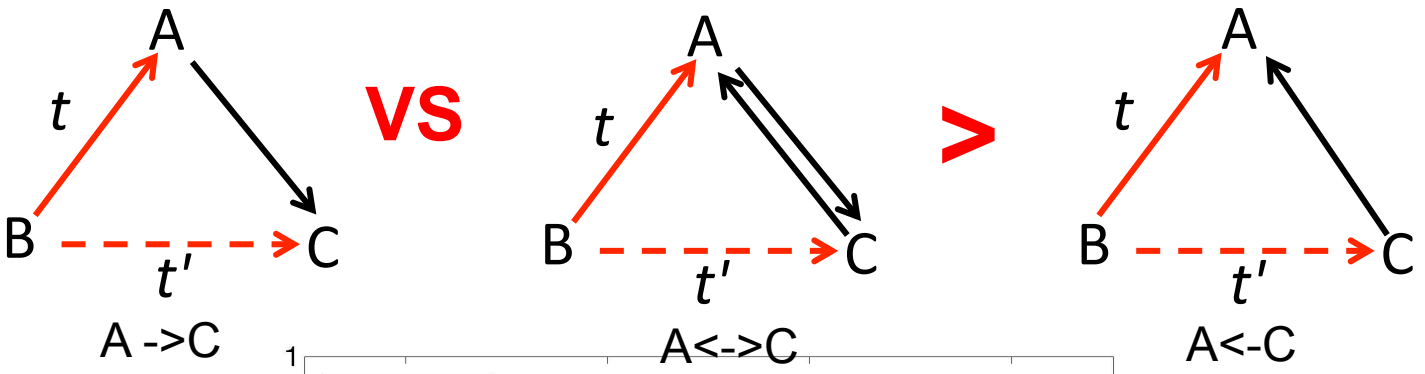


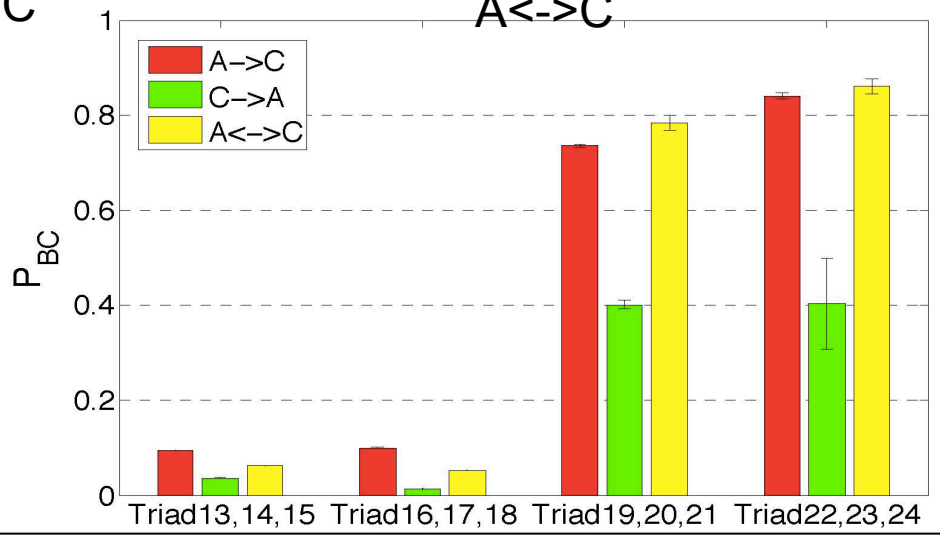


Observation: Following influence is more significant when there is a **reciprocal** relationship between B and A.

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

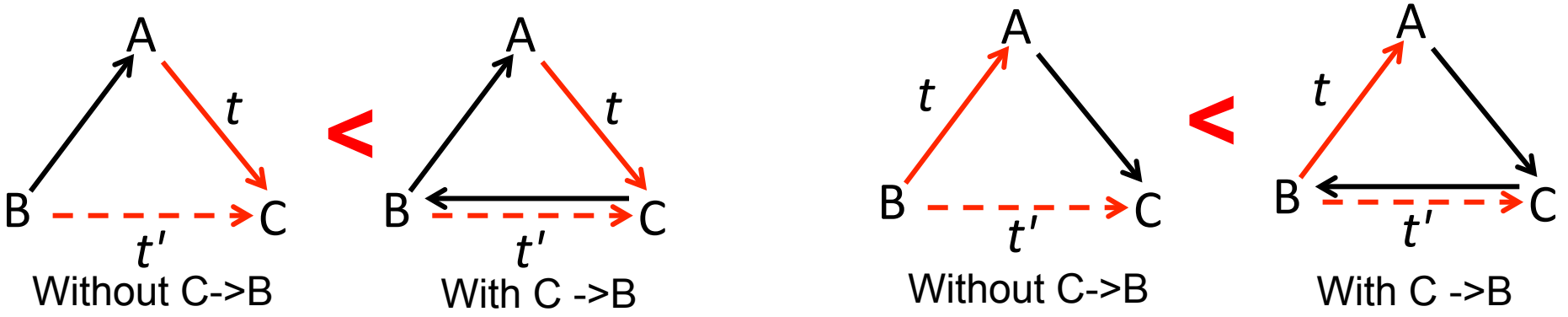


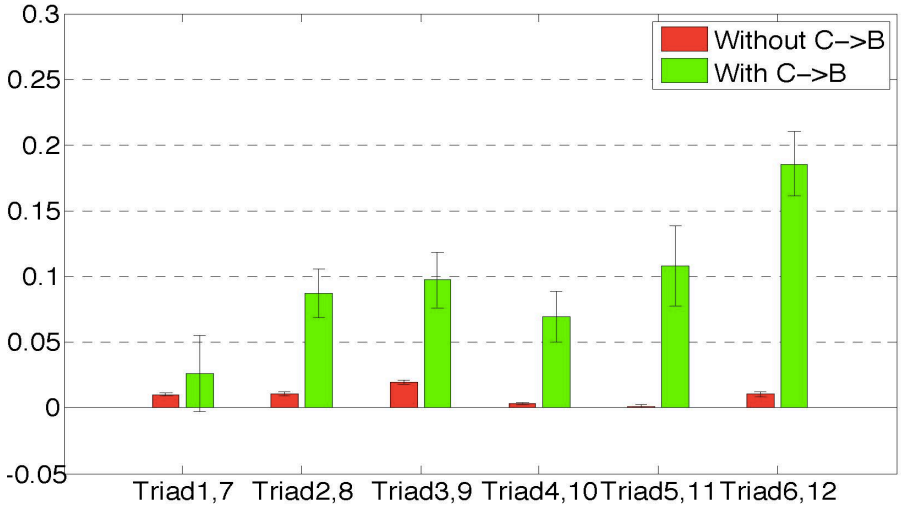


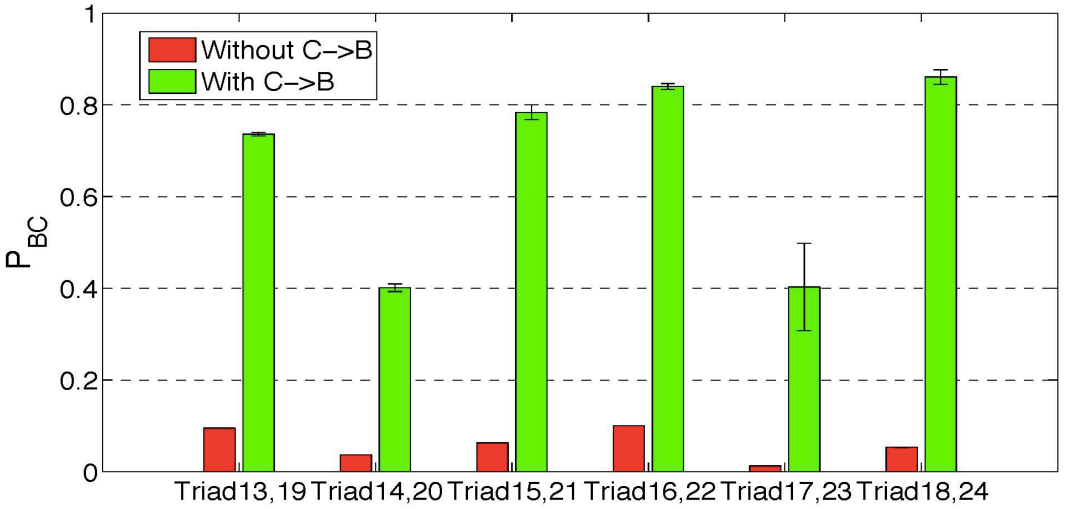
Observation: Following influence is more significant when there is a one-way relationship from A to C.

Explanation: Users usually prefer to check their followee's followees, from whom they select those they may be interested to follow.

Reversed Relationship



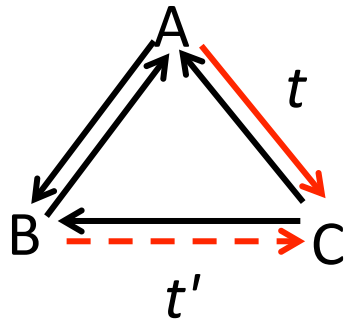




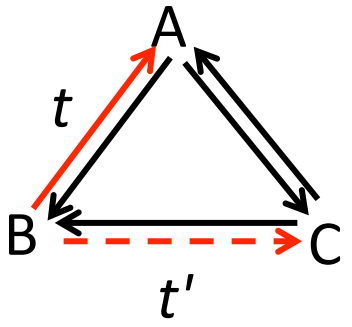
Observation: Following influence is more significant when there is a reversed relationship from C to B.

Explanation: Users are highly encouraged to follow their followers.

Social Theories: Structural Balance[1]



Follower diffusion



Followee diffusion

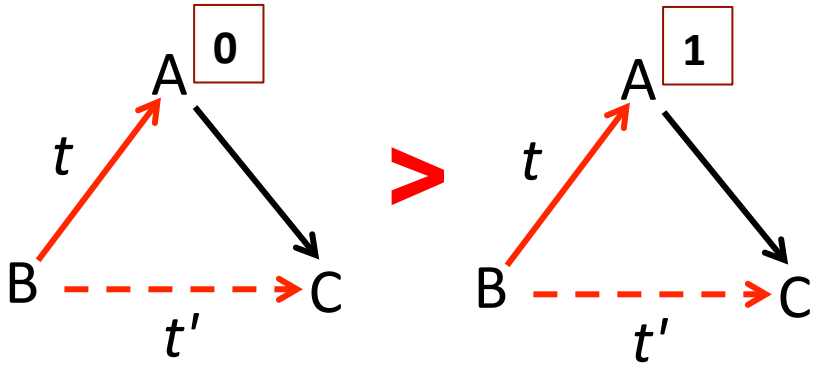
Social Balance: my friend's friend is also my friend
The probabilities of B following C in the two triads are higher than others in their respective categories.

Explanation: Users have tendency to form a balanced triad

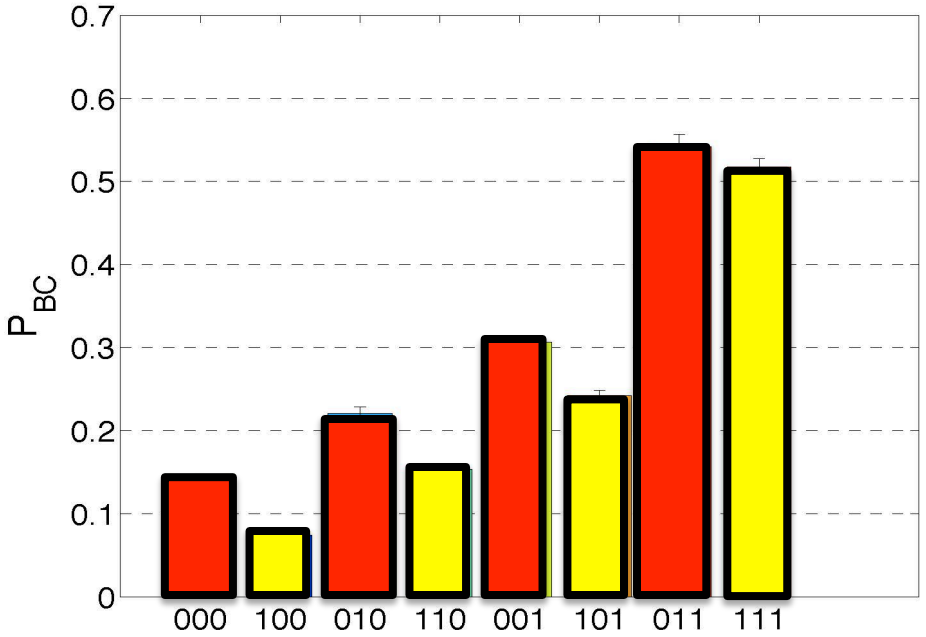
Fritz Heider (1958). The Psychology of Interpersonal Relations. John Wiley & Sons.

Social Theories: Social Status

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



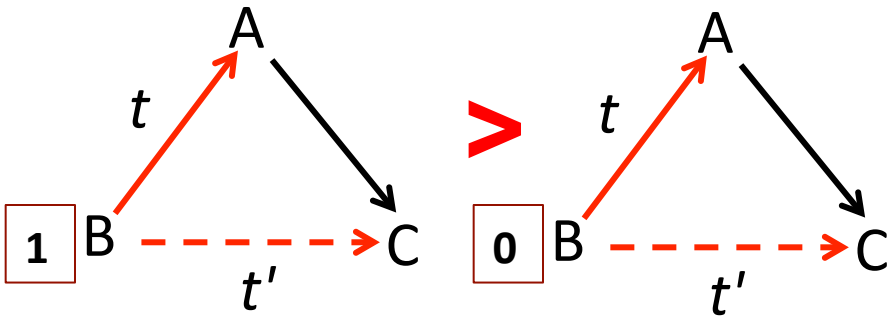
1: Elite user
0: Low-status user



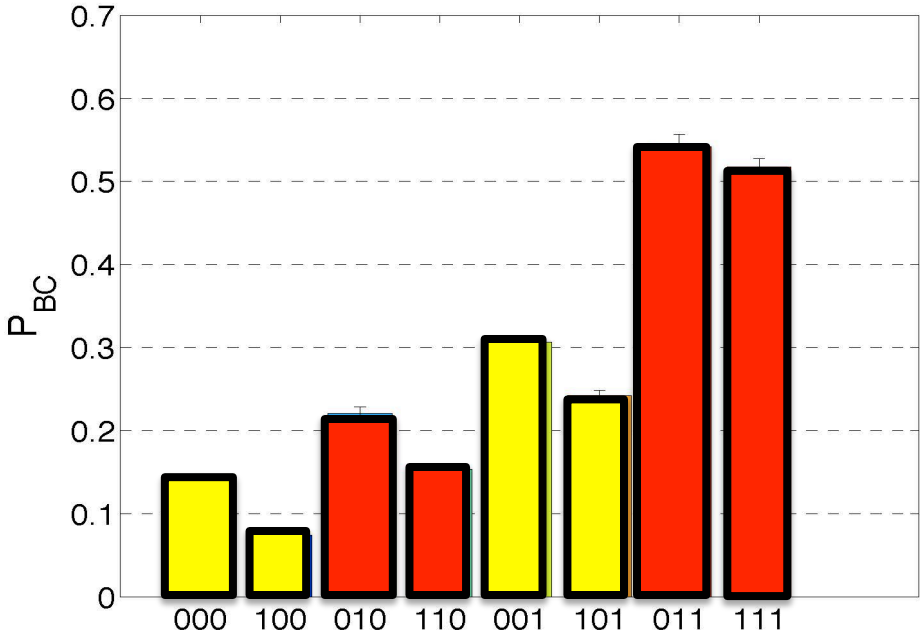
- Low-status users act as a bridge to connect users so as to form a closure triad.
- The likelihood of 0XX is 1.4 times of 1XX.

Social Theories: Social Status

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



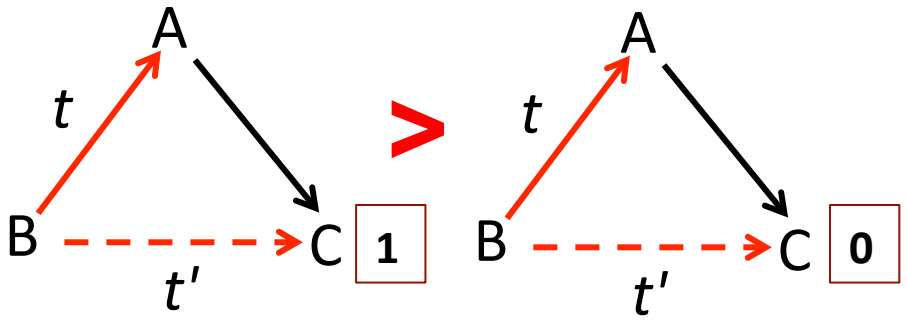
1: Elite user
0: Low-status user



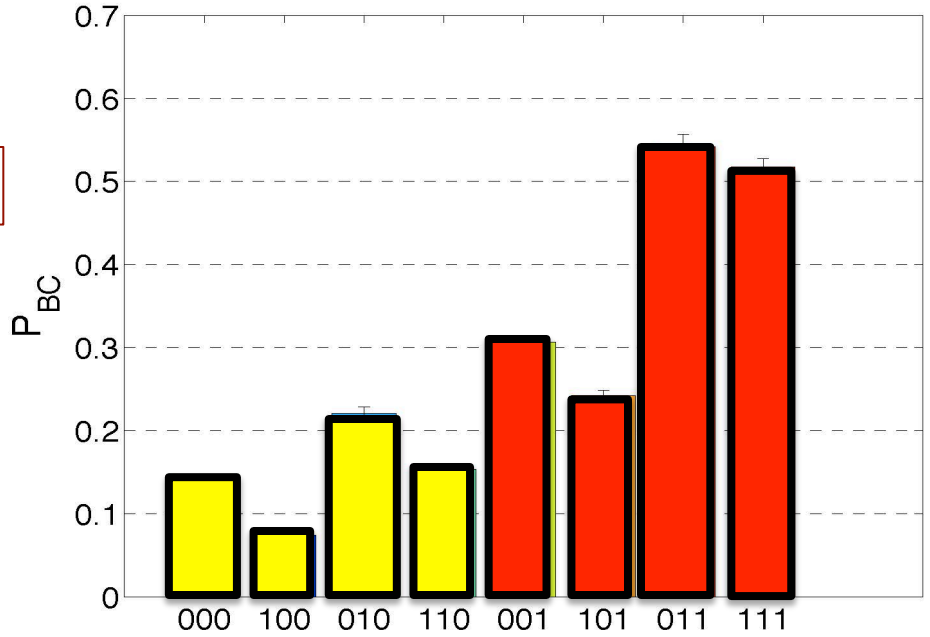
- Elite users play a more important role to form the triadic closure.
- The likelihood of X1X is almost double the probability of X0X.

Social Theories: Social Status

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



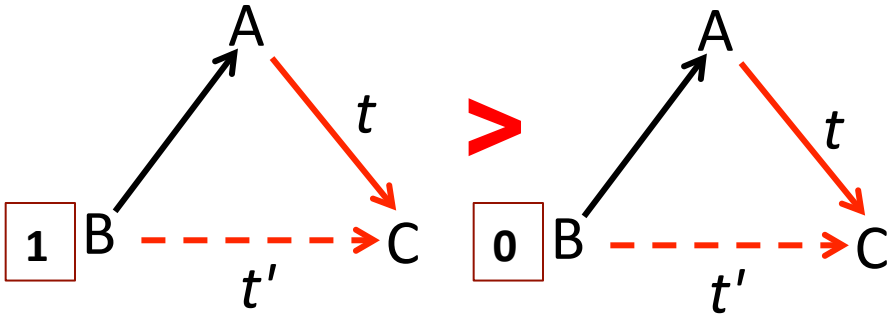
1: Elite user
0: Low-status user



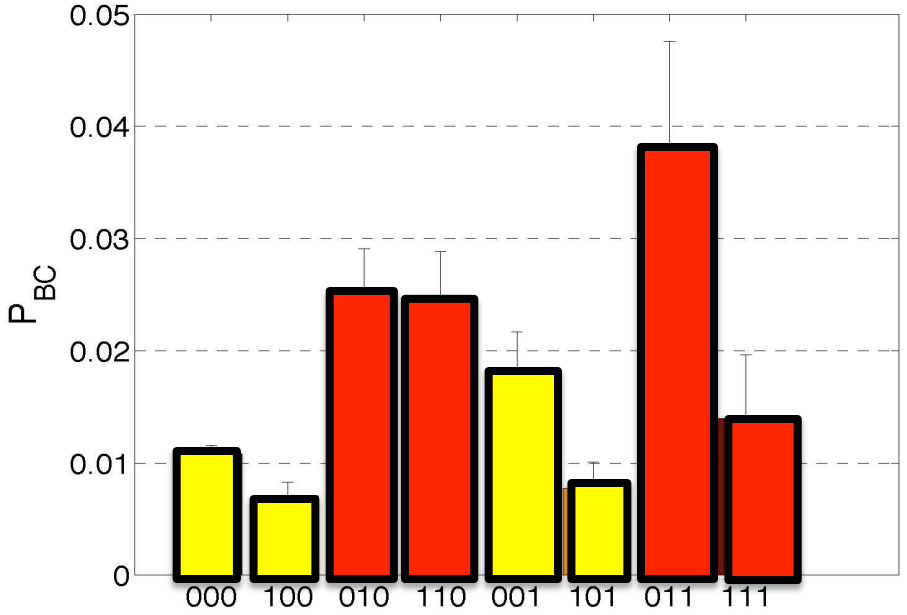
- The rich gets richer.
- The likelihood of XX1 is nearly 2 times higher than that of XX0.
- This phenomenon validates the mechanism of preferential attachment.

Social Theories: Social Status

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



1: Elite user
0: Low-status user

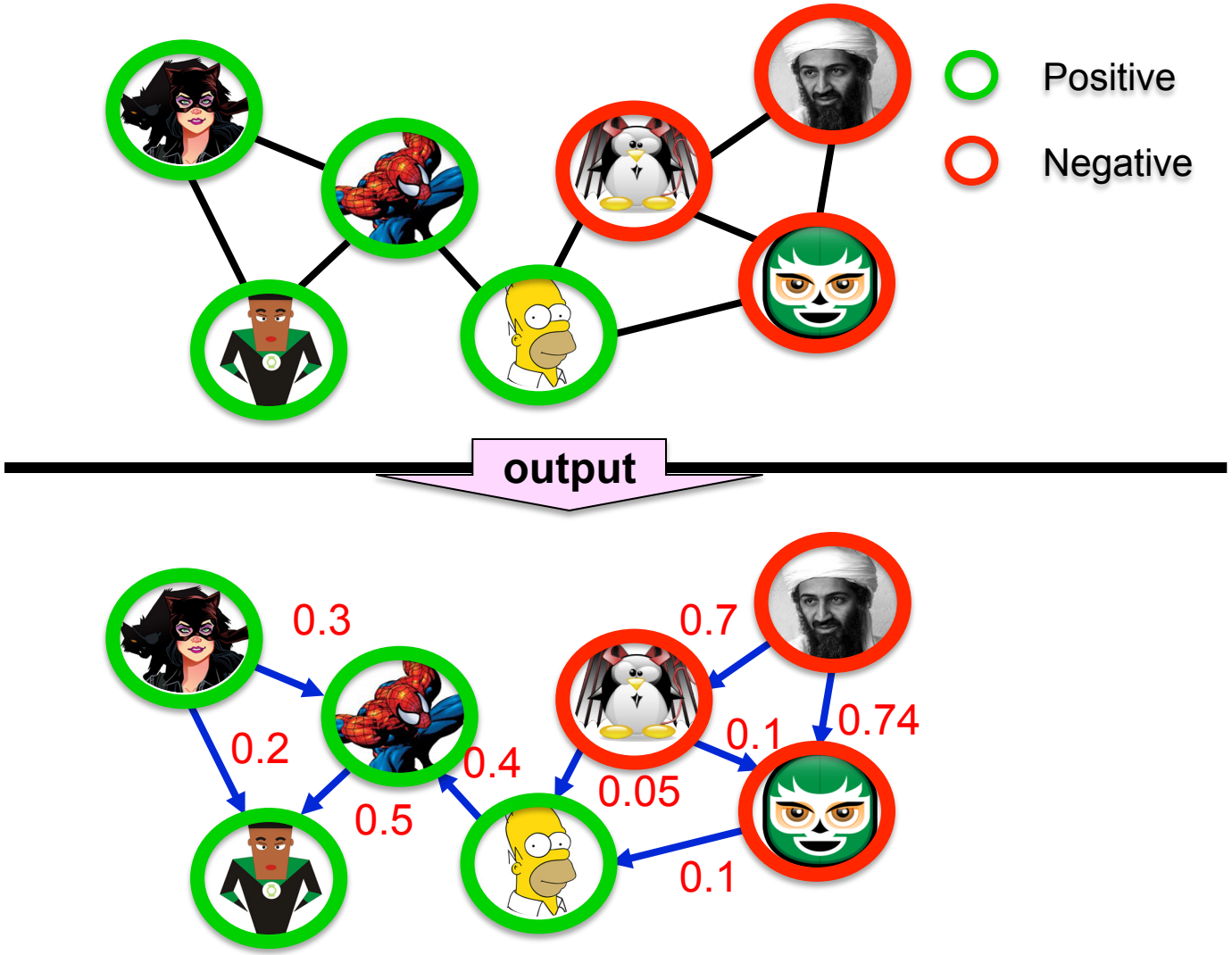


- Elite users play a more important role to form the triadic closure.
- The likelihood of X1X is almost double the probability of X0X.

Summaries

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
  - Topical Affinity Propagation (TAP)
- Action-based methods
  - A discriminative model: NTT-FGM
  - A generative model: FCM

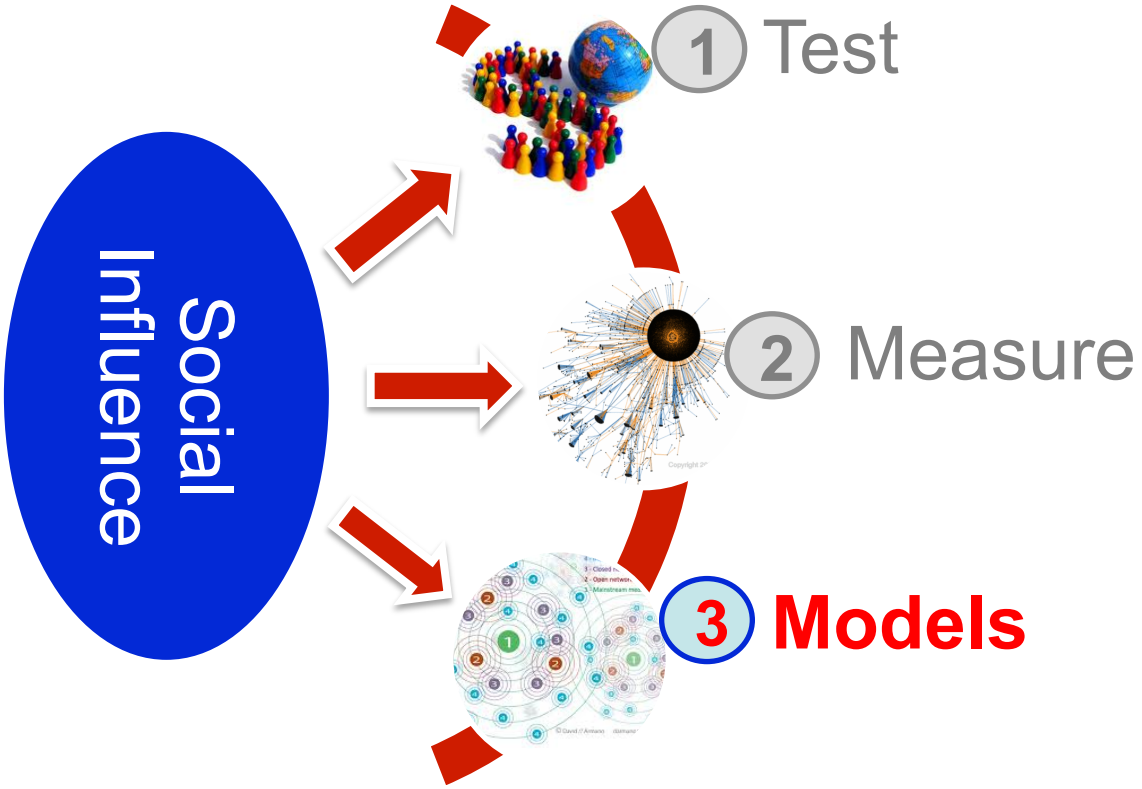
Output of Measuring Influence



Understanding the Emotional Impact in Social Networks

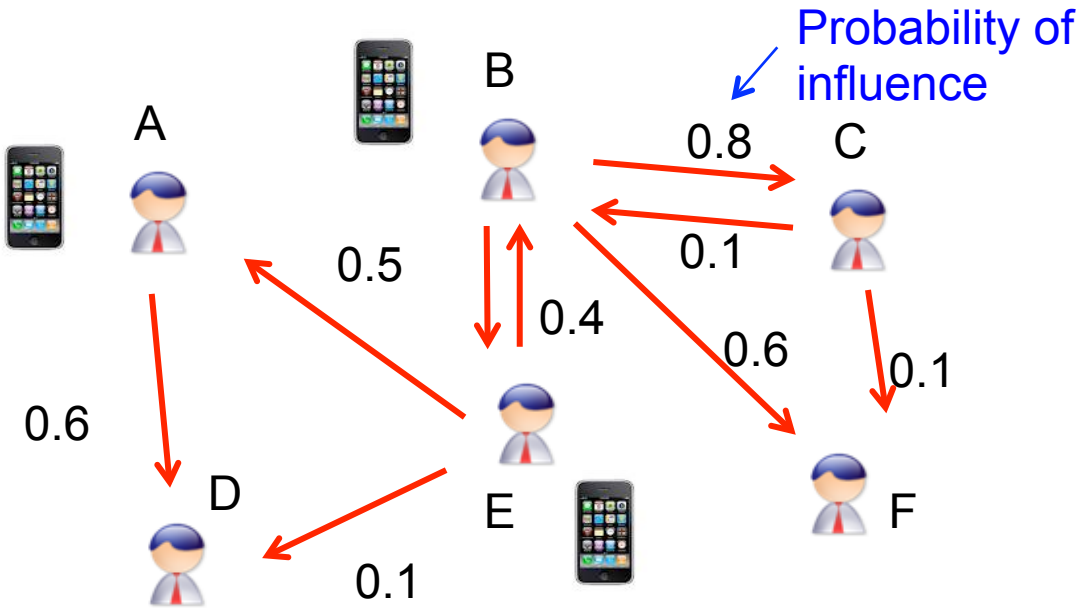
[1] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can We Understand van Gogh's Mood? Learning to Infer Affects from Images in Social Networks. In ACM Multimedia, pages 857-860, 2012.

Social Influence



Influence Maximization

- Influence maximization
  - Minimize marketing cost and more generally to maximize profit.
  - E.g., to get a small number of influential users to adopt a new product, and subsequently trigger a large cascade of further adoptions.



[1] P. Domingos and M. Richardson. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01), pages 57–66, 2001.

Problem Abstraction

- We associate each user with a status:
  - Active or Inactive
  - The status of the chosen set of users (seed nodes) to market is viewed as active
  - Other users are viewed as inactive
- Influence maximization
  - Initially all users are considered inactive
  - Then the chosen users are activated, who may further influence their friends to be active as well

Diffusion Influence Model

- Linear Threshold Model
- Cascade Model

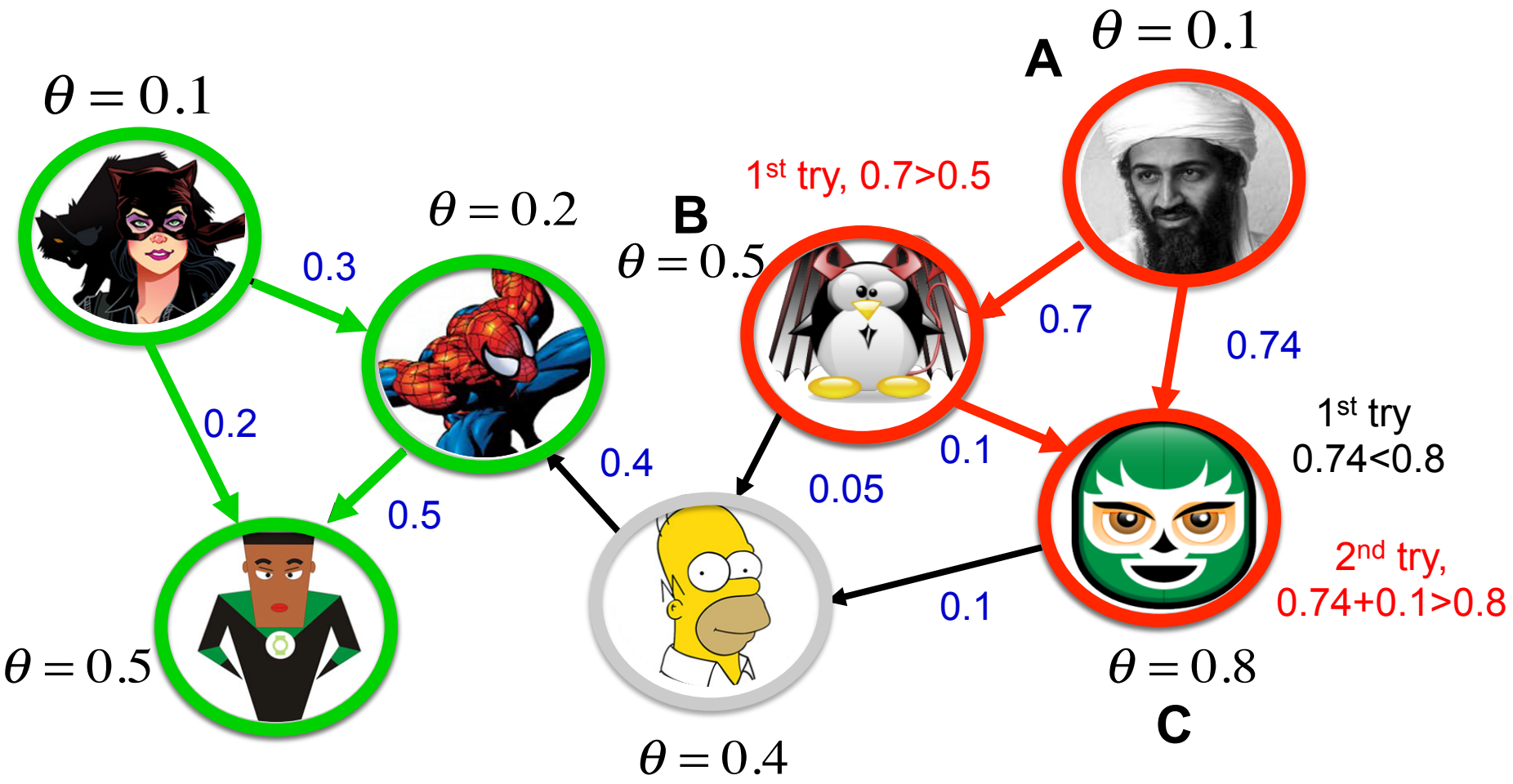
Linear Threshold Model

- General idea
  - Whether a given node will be active can be based on an arbitrary monotone function of its neighbors that are already active.
- Formalization
  - f_v : map subsets of v's neighbors' influence to real numbers in [0,1]
  - θ_v : a threshold for each node
  - S: the set of neighbors of v that are active in step t-1
  - Node v will turn active in step t if $f_v(S) > \theta_v$
- Specifically, in [Kempe, 2003], f_v is defined as $\sum_{u \in S} b_{v.u}$, where $b_{v,u}$ can be seen as a fixed weight, satisfying

$$\sum_{v \in N(u)} b_{u,v} \leq 1$$

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), pages 137–146, 2003.

Linear Threshold Model: An example



Cascade Model

- Cascade model
  - $p_v(u,S)$: the success probability of user *u* activating user v
  - User *u* tries to activate v and finally succeeds, where *S* is the set of *v*'s neighbors that have already attempted but failed to make *v* active
- Independent cascade model
  - $p_v(u,S)$ is a constant, meaning that whether *v* is to be active does not depend on the order *v*'s neighbors try to activate it.
  - Key idea: Flip coins *c* in advance -> live edges
  - $F_c(A)$: People influenced under outcome *c* (set cover)
  - $F(A) = \text{Sum}_c P(c) F_c(A)$ is submodular as well

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), pages 137–146, 2003.

Theoretical Analysis

- NP-hard [1]
  - Linear threshold model
  - General cascade model
- Kempe Prove that approximation algorithms can guarantee that the influence spread is within(1-1/e) of the optimal influence spread.
  - Verify that the two models can outperform the traditional heuristics
- Recent research focuses on the efficiency improvement
  - [2] accelerate the influence procedure by up to 700 times
- It is still challenging to extend these methods to large data sets

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'03), pages 137–146, 2003.
[2] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07), pages 420–429, 2007.

Objective Function

- **Objective function:**

  - $f(S)$ = Expected #people influenced when targeting a set of users S

- Define $f(S)$ as a monotonic submodular function

$$f(S \cup \{v\}) - f(S) \quad \geq \quad f(T \cup \{v\}) - f(T)$$

$$f(S \cup \{v\}) \geq f(S)$$

where $S \subseteq T$.

[1] P. Domingos and M. Richardson. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01), pages 57–66, 2001.
[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'03), pages 137–146, 2003.

Maximizing the Spread of Influence

- Solution
  - Use a submodular function to approximate the influence function
  - Then the problem can be transformed into finding a *k*-element set *S* for which *f* (*S*) is maximized.

THEOREM 7.3 *[19, 50] For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let $S^\star$ be a set that maximizes the value of f over all k-element sets. Then $f(S) \geq (1 - 1/e) \cdot f(S^\star)$; in other words, S provides a $(1 - 1/e)$-approximation.*

approximation ratio

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03), pages 137–146, 2003.

Algorithms

- General Greedy
- Low-distance Heuristic
- High-degree heuristic
- Degree Discount Heuristic

General Greedy

- General idea: In each round, the algorithm adds one vertex into the selected set *S* such that this vertex together with current set *S* maximizes the influence spread.

Algorithm 1 GeneralGreedy(G, k)

```
1:  initialize S = ∅ and R = 20000
2:  for i = 1 to k do
3:      for each vertex v ∈ V \ S do
4:          s_v = 0.
5:          for i = 1 to R do
6:              s_v += |RanCas(S ∪ {v})|
7:          end for
8:          s_v = s_v / R
9:      end for
10:     S = S ∪ {arg max_{v∈V\S} {s_v}}
11: end for
12: output S.
```

Any random diffusion process

Low-distance Heuristic

- Consider the nodes with the shortest paths to other nodes as seed nodes
- Intuition
  - Individuals are more likely to be influenced by those who are closely related to them.

High-degree heuristic

- Choose the seed nodes according to their degree.
- Intuition
  - The nodes with more neighbors would arguably tend to impose more influence upon its direct neighbors.
  - Know as “degree centrality”

Degree Discount Heuristic[1]

- General idea: If u has been selected as a seed, then when considering selecting v as a new seed based on its degree, we should not count the edge v->u
- Specifically, for a node v with d_v neighbors of which t_v are selected as seeds, we should discount v's degree by

$$2t_v + (d_v - t_v)\, t_v p$$

where p=0.1.

Algorithm 4 DegreeDiscountIC(G, k)

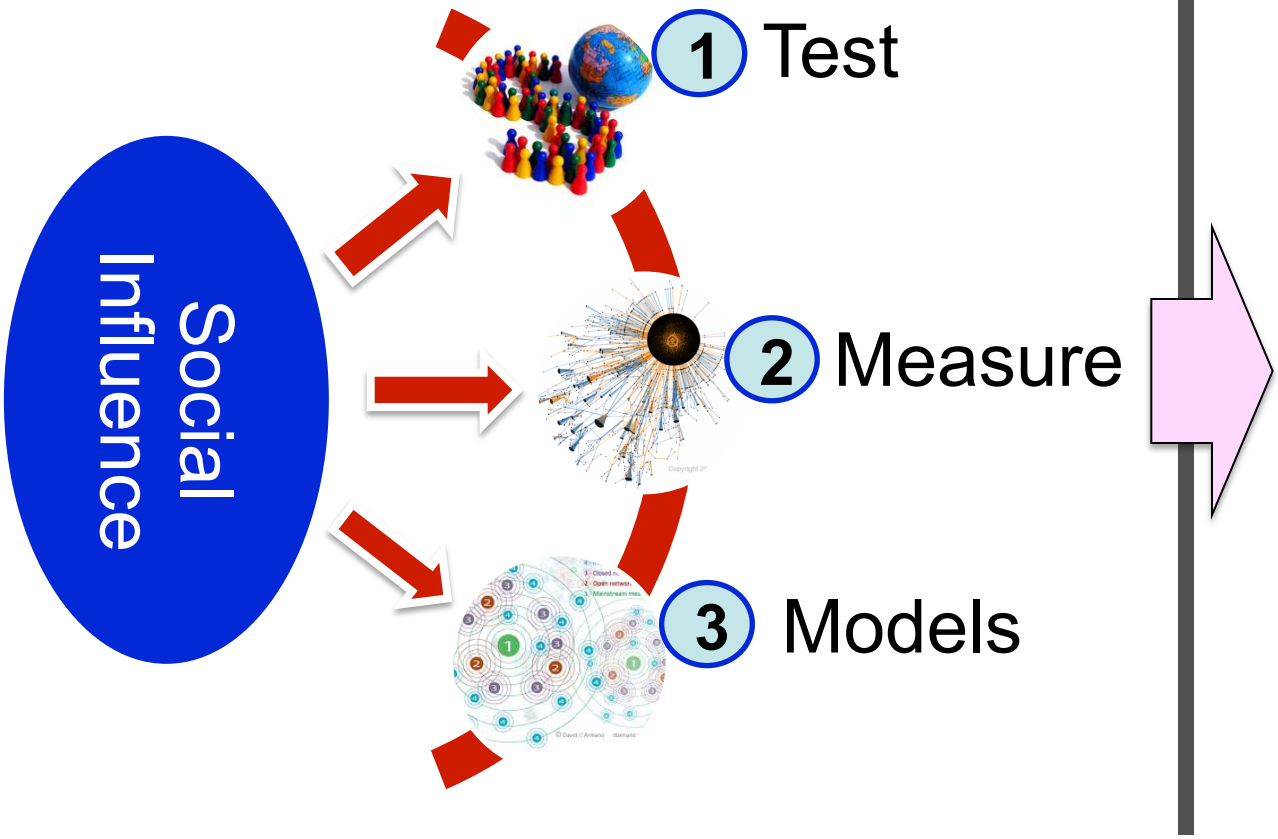
```
initialize S = ∅
for each vertex v do
    compute its degree d_v
    dd_v = d_v
    initialize t_v to 0
end for
for i = 1 to k do
    select u = arg max_v{dd_v | v ∈ V \ S}
    S = S ∪ {u}
    for each neighbor v of u and v ∈ V \ S do
        t_v = t_v + 1
        dd_v = d_v − 2t_v − (d_v − t_v)t_v p
    end for
end for
output S
```

[1] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In KDD'09, pages 199-207, 2009.

Summaries

- Influence Maximization Models
  - Linear Threshold Model
  - Cascade Model
- Algorithms
  - General Greedy
  - Low-distance Heuristic
  - High-degree heuristic
  - Degree Discount Heuristic

Social Influence



Application: Social Advertising[1]

- Conducted two very large field experiments that identify the effect of social cues on consumer responses to ads on Facebook
- Exp. 1: measure how responses increase as a function of the number of cues.
- Exp. 2: examines the effect of augmenting traditional ad units with a minimal social cue
- Result: Social influence causes significant increases in ad performance

[1] E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In EC'12, pages 146-161, 2012.

Application: Opinion Leader[1]

- Propose viral marketing through frequent pattern mining.
- Assumption
  - Users can see their friends actions.
- Basic formation of the problem
  - Actions take place in different time steps, and the actions which come up later could be influenced by the earlier taken actions.
- Approach
  - Define leaders as people who can influence a sufficient number of people in the network with their actions for a long enough period of time.
  - Finding leaders in a social network makes use of action logs.

[1] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In CIKM'08, pages 499–508, 2008.

Application: Influential Blog Discovery[1]

- Influential Blog Discovery
  - In the web 2.0 era, people spend a significant amount of time on user-generated content web sites, like blog sites.
  - Opinion leaders bring in new information, ideas, and opinions, and disseminate them down to the masses.
- Four properties for each bloggers
  - **Recognition**: A lot of inlinks to the article.
  - **Activity generation**: A large number of comments indicates that the blog is influential.
  - **Novelty**: with less outgoing links.
  - **Eloquence**: Longer articles tend to be more eloquent, and can thus be more influential.

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In WSDM'08, pages 207–217, 2008.

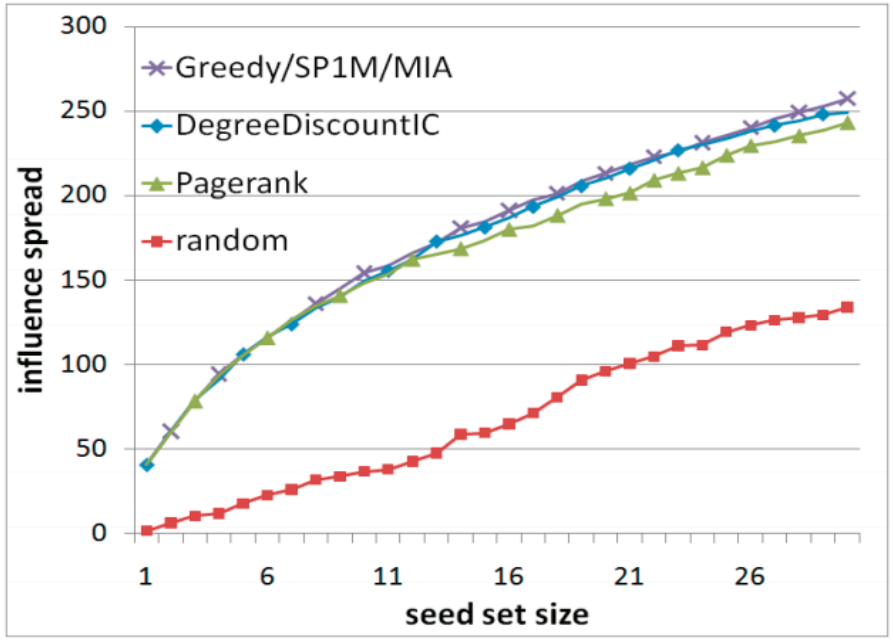
Example 1: Influence maximization with the learned influence probabilities

Maximizing Influence Spread

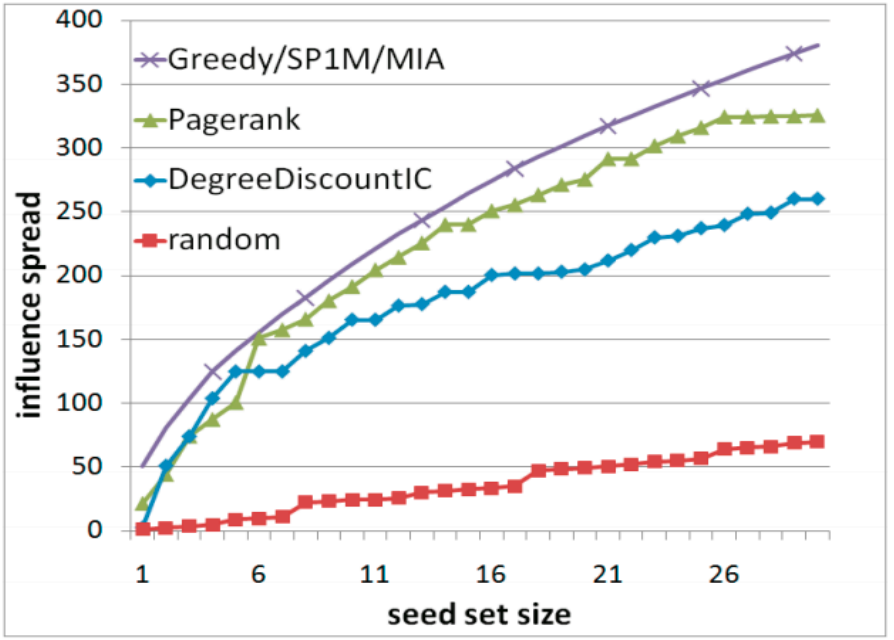
- Goal
  - Verify whether the learned influence probability can help maximize influence spread.
- Data sets
  - Citation and Coauthor are from Arnetminer.org;
  - Film is from Wikipedia, consisting of relationships between directors, actors, and movies.

Data Set	#Node	#Edge	Density
Citation	127K	374K	10^{-5}
Coauthor	61K	152K	10^{-3}
Film	34K	142K	10^{-2}

Influence Maximization



(a) With uniform influence



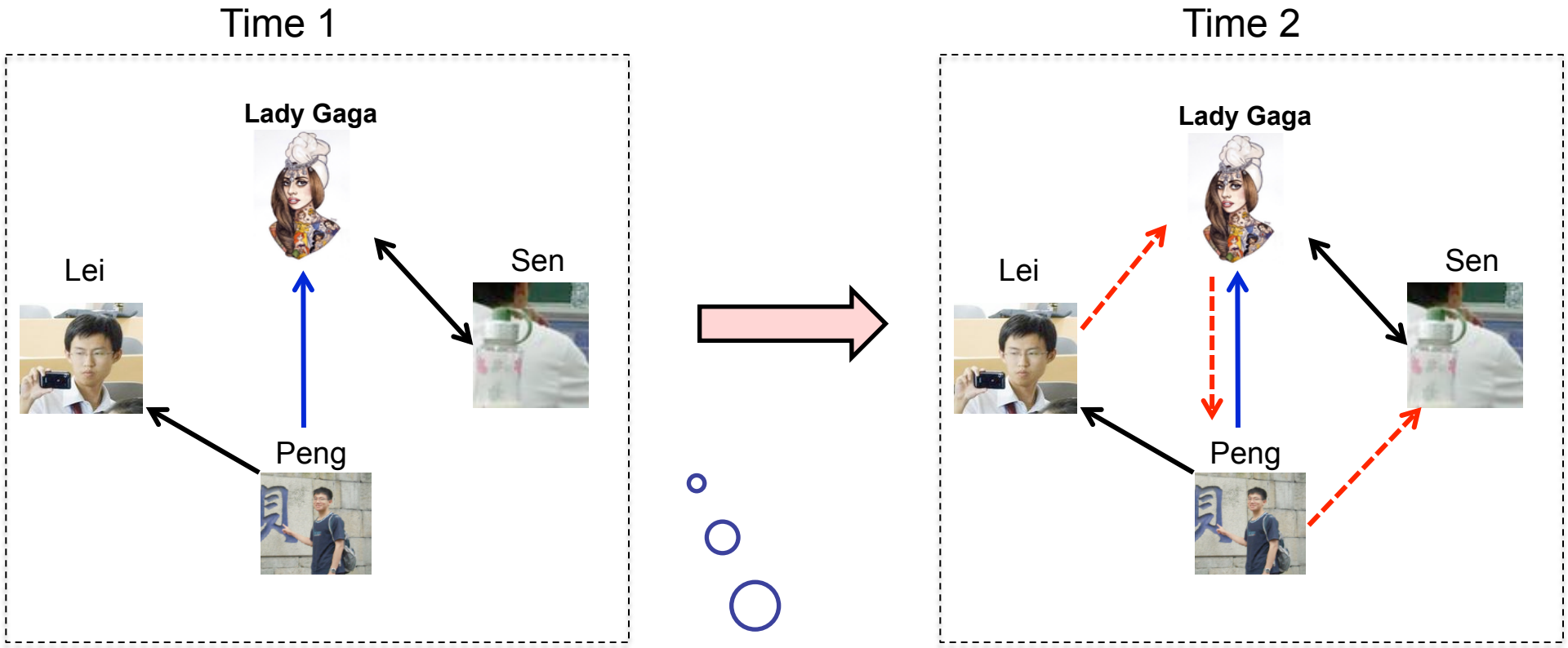
(b) With the learned influence

a) The influence probability from v_i to v_j is simply defined as as $\frac{1}{d_j}$, where d_j is the in-degree of v_j.
a) Influence probability learned from the model we introduced before.

[1] C. Wang, J. Tang, J. Sun, and J. Han. Dynamic Social Influence Analysis through Time-dependent Factor Graphs. In ASONAM'11, pages 239-246, 2011.

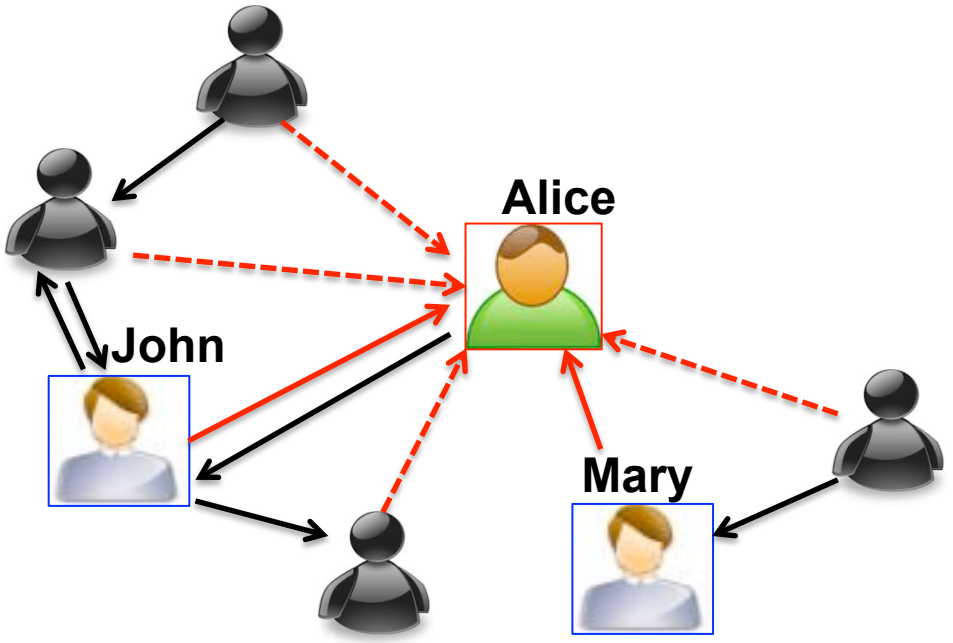
Example 2: Following Influence Applications

Following Influence Applications



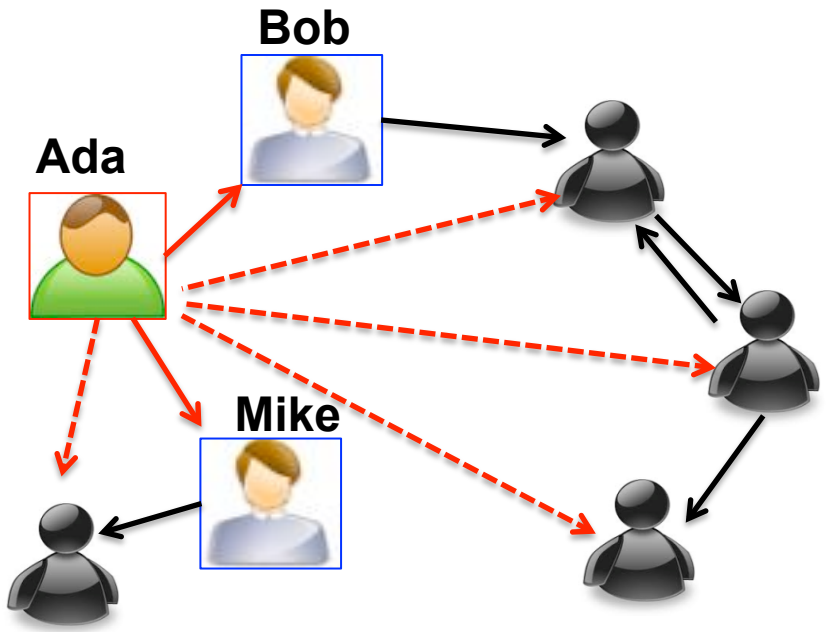
When you follow a user in a social network, will the behavior influences your friends to also follow her?

Applications: Influence Maximization



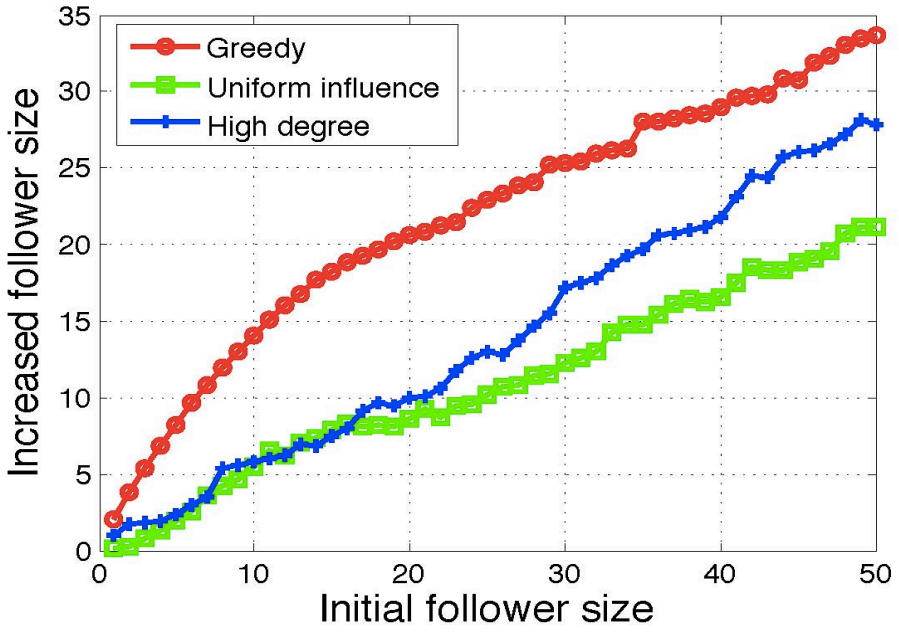
Find a set *S* of *k* initial followers to follow user *v* such that the number of newly activated users to follow *v* is maximized.

Applications: Friend Recommendation

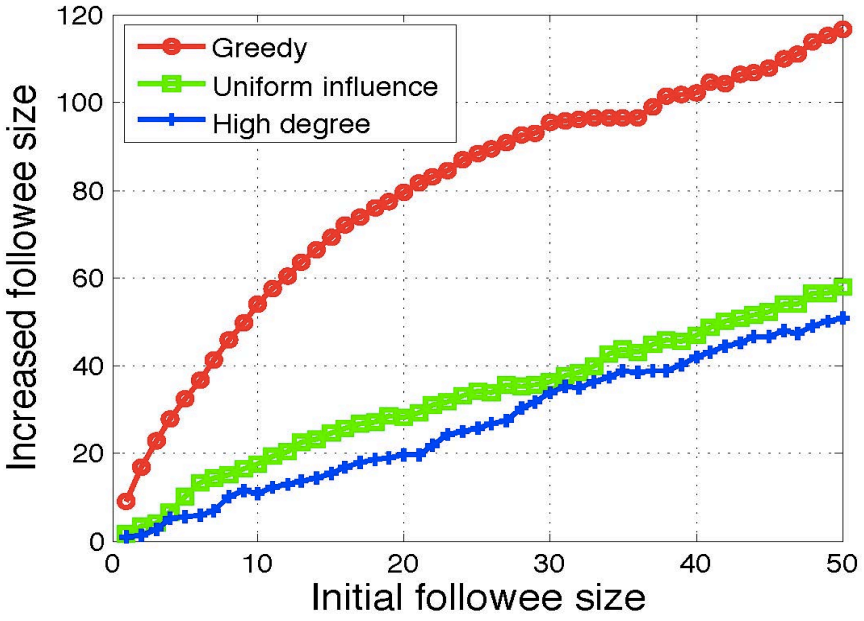


Find a set *S* of *k* initial followees for user *v* such that the total number of new followees accepted by *v* is maximized

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)

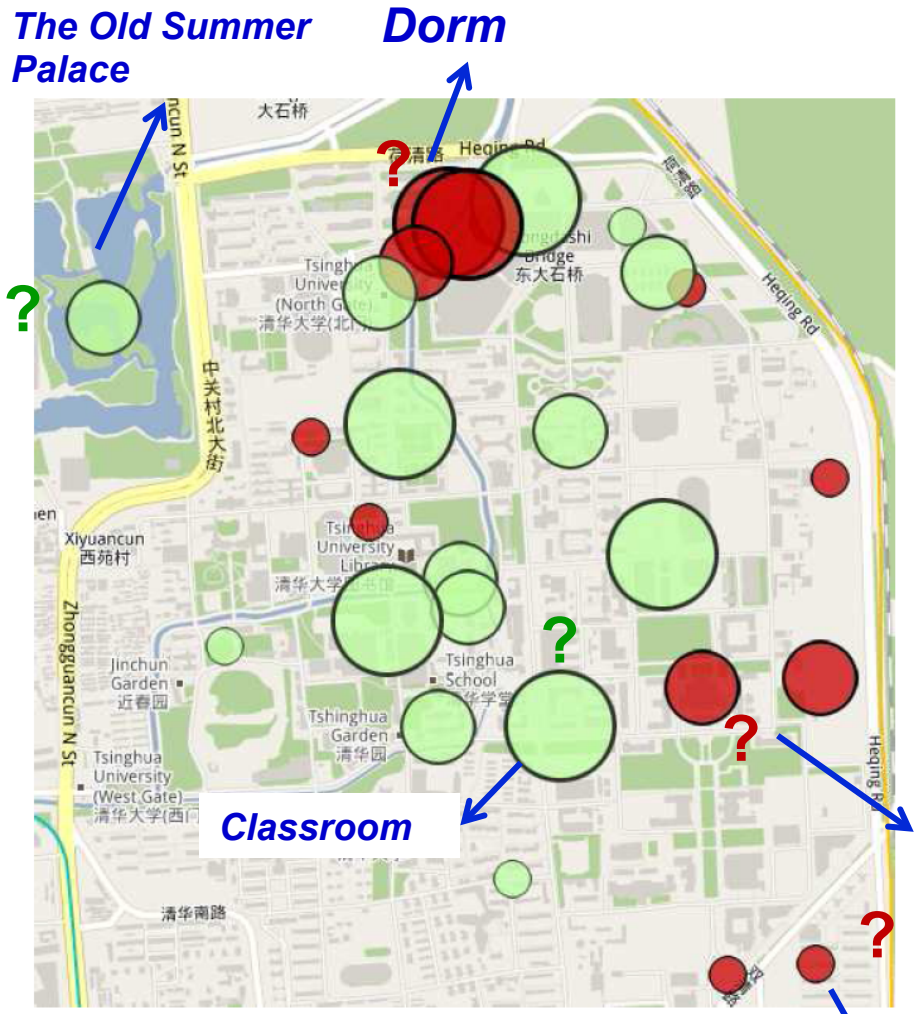
Example 3: Emotion Influence

[1] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and ACM Fong. Quantitative Study of Individual Emotional States in Social Networks. IEEE TAC, 2012, Volume 3, Issue 2, Pages 132-144.

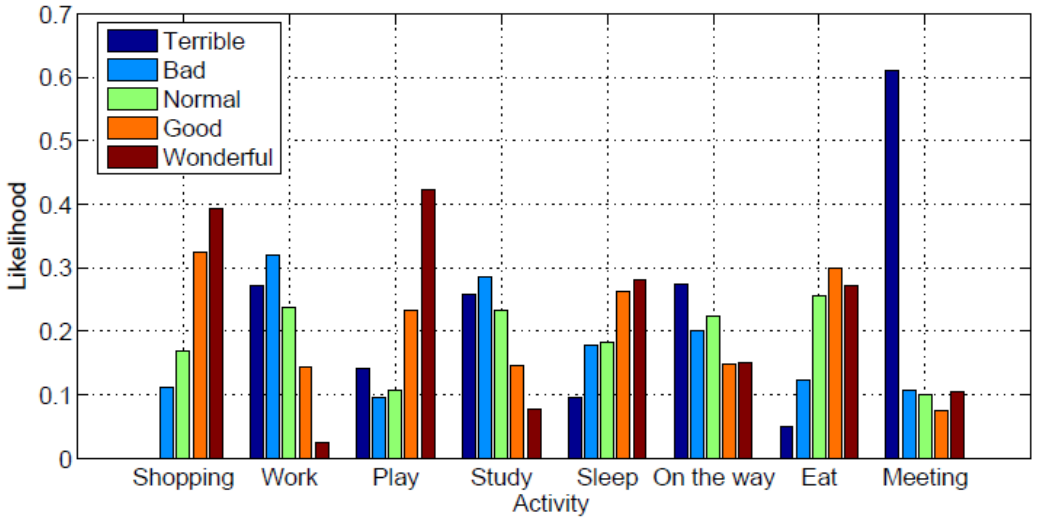
Happy System

Can we predict users' emotion?

Observations (cont.)

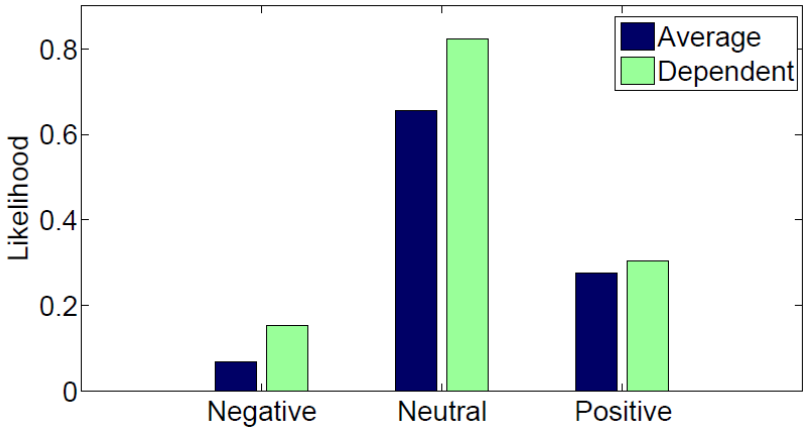


Location correlation
(Red-happy)



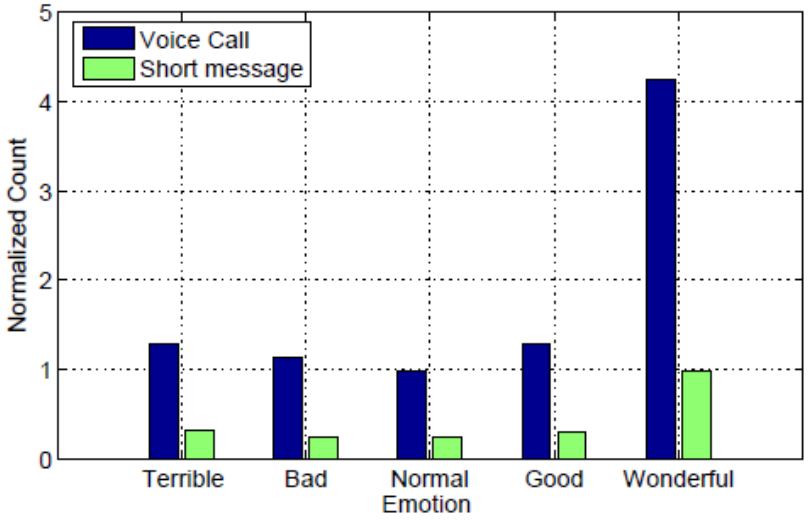
Activity correlation

Observations



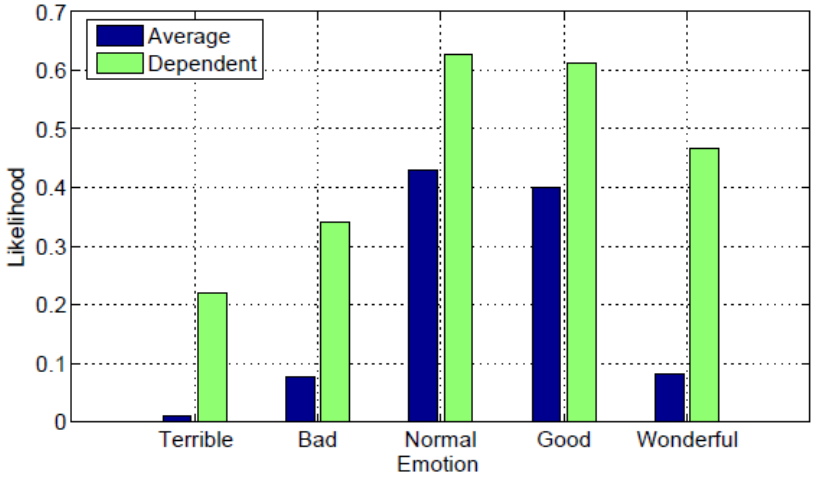
(a) Social correlation

(a) Implicit groups by emotions

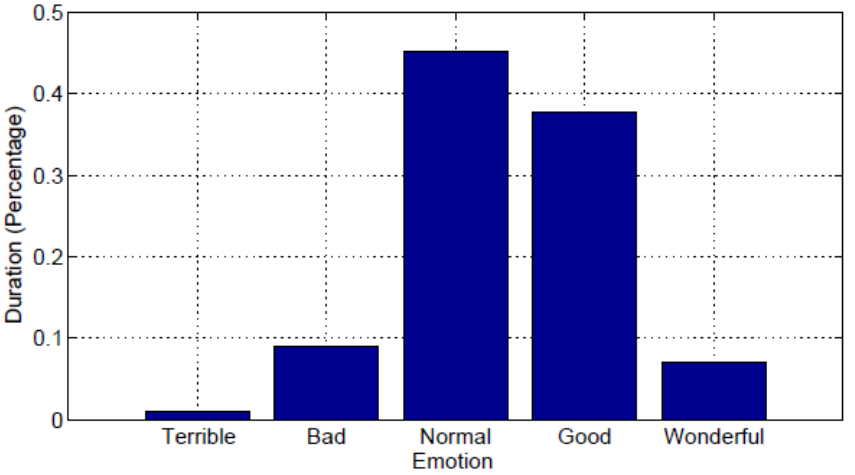


(c) Calling (SMS) correlation

Observations (cont.)

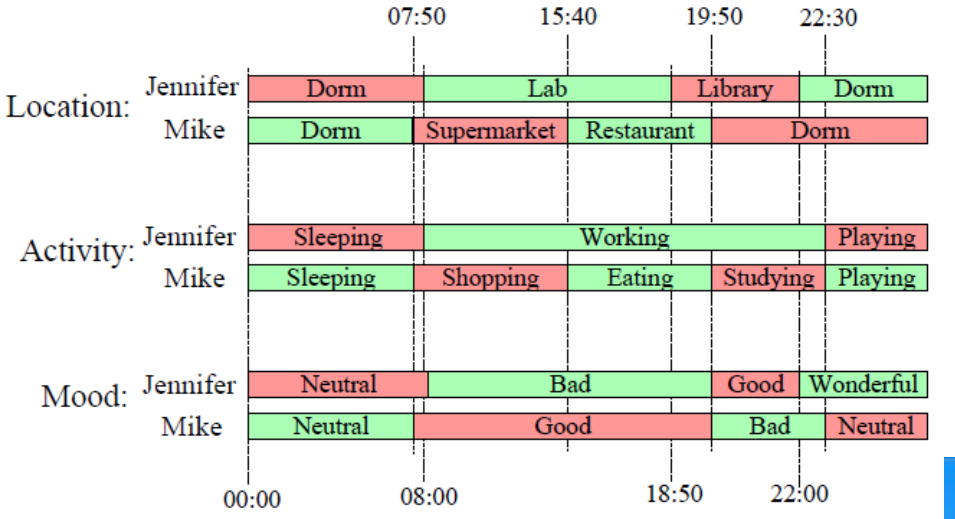


(a) Temporal correlation

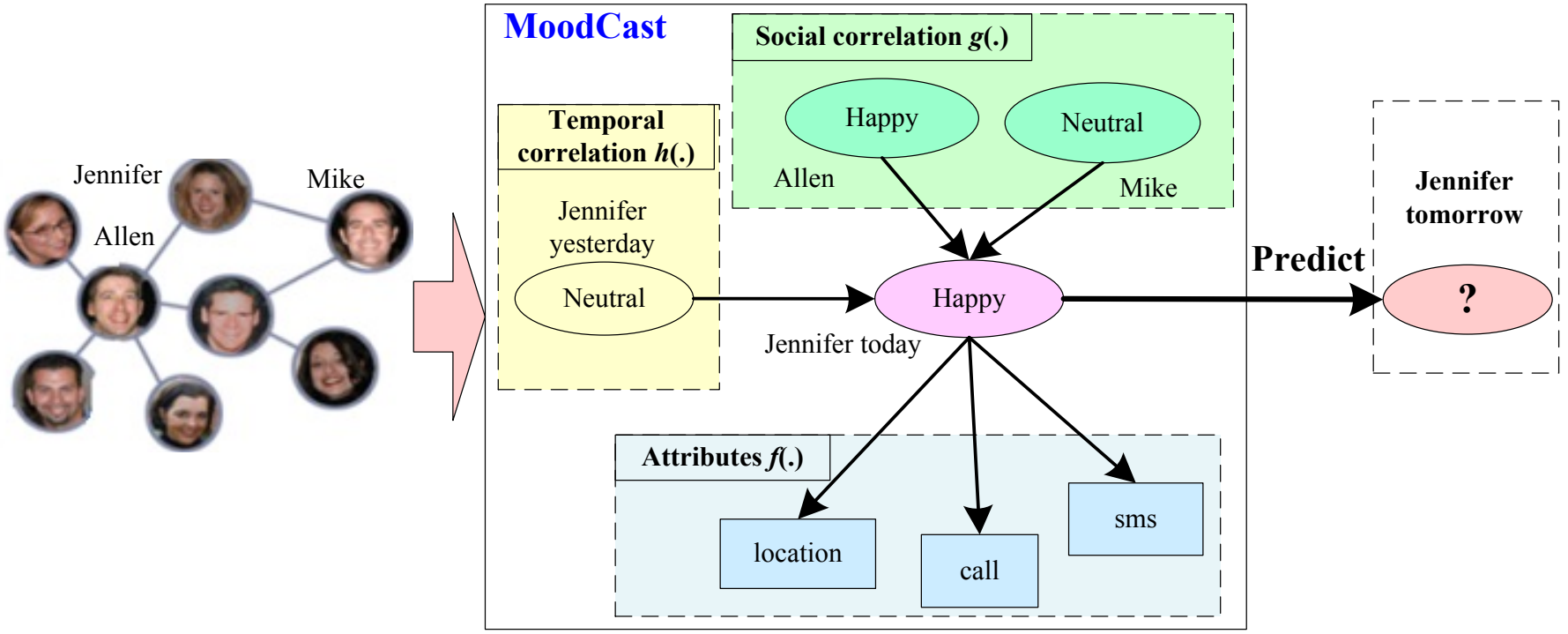


(b) Time duration

Temporal correlation



MoodCast: Dynamic Continuous Factor Graph Model

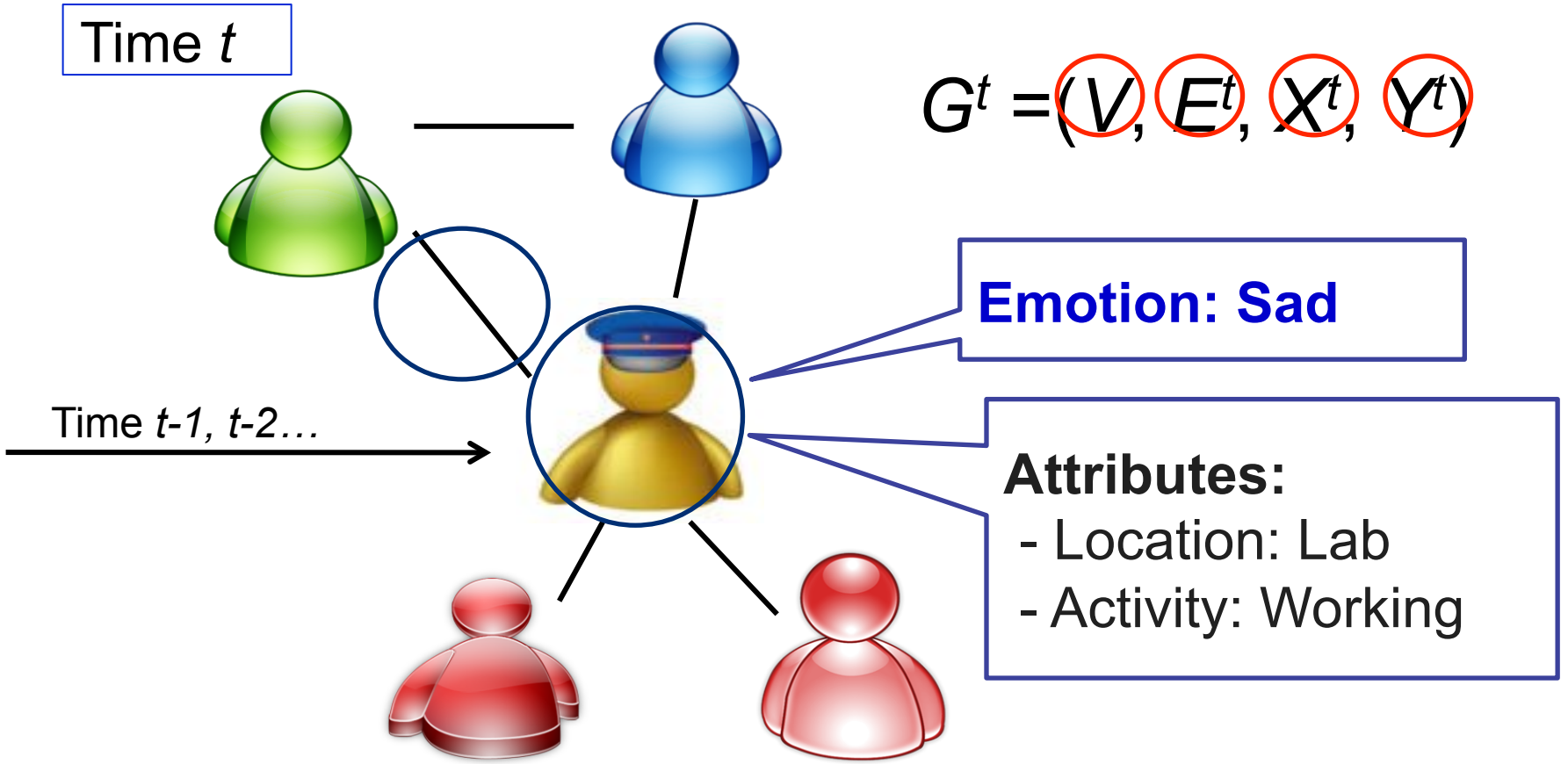


Our solution

1. We directly define continuous feature function;

2. Use Metropolis-Hasting algorithm to learn the factor graph model.

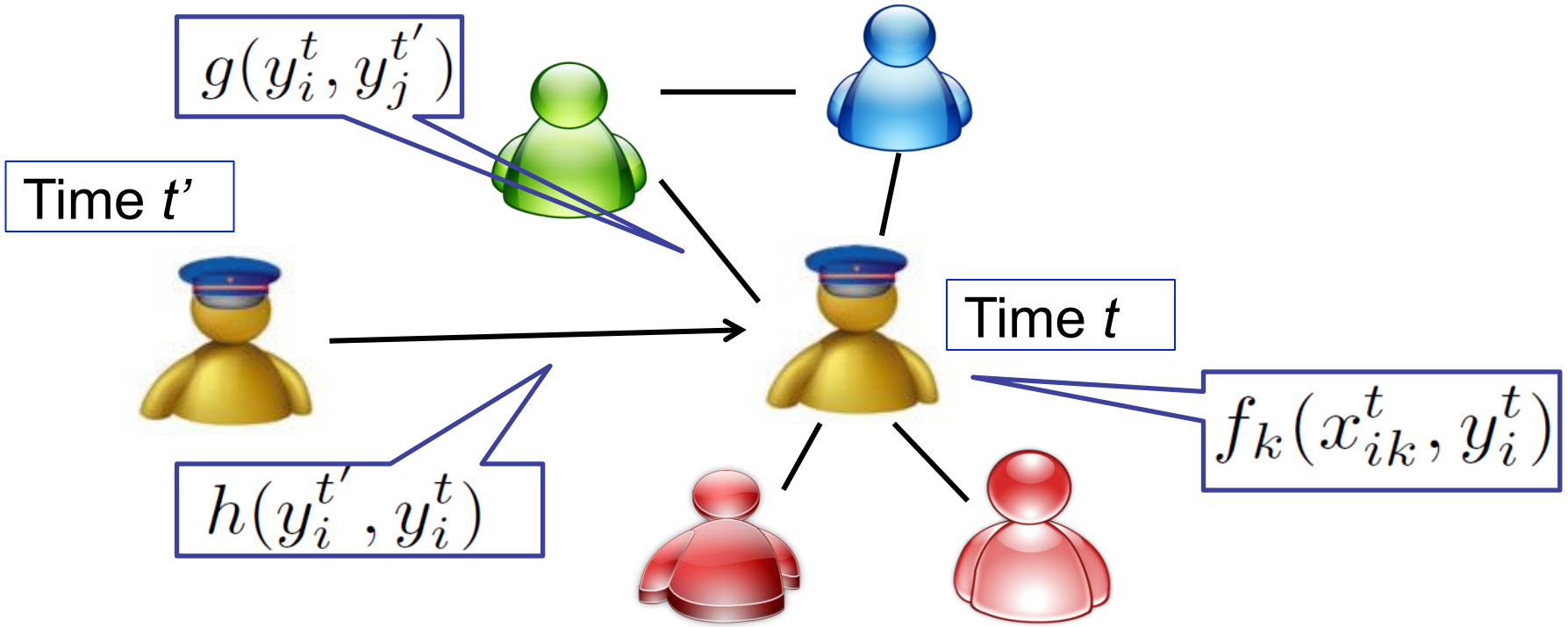
Problem Formulation



Learning Task:

$$f(V, E^{(t+1)}, X^{(t+1)} | G^t) \rightarrow Y^{(t+1)}$$

Dynamic Continuous Factor Graph Model

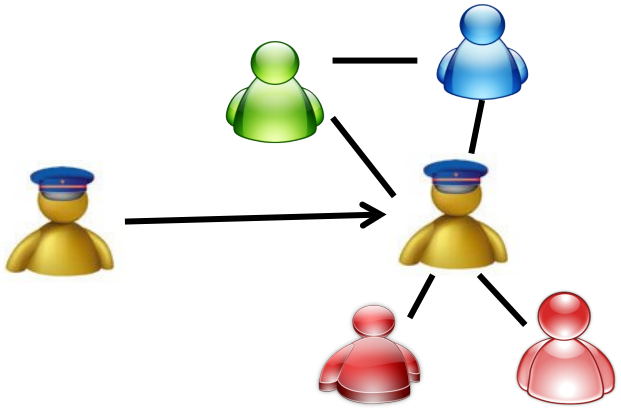


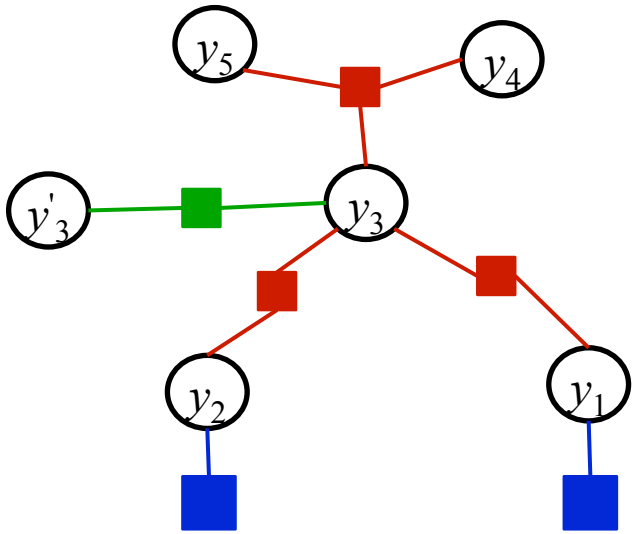
$f_k(x_{ik}^t, y_i^t)$: Binary function

$$g(y_i^t, y_j^{t'}) = \exp\{-\beta_{ji}(t - t')(y_i^t - y_j^{t'})^2\}$$

$$h(y_i^{t'}, y_i^t) = \exp\{-\lambda_i(t - t')(y_i^t - y_i^{t'})^2\}$$

Learning with Factor Graphs





$$p(Y|G^t) = \frac{1}{Z}\exp\{\sum_{v_i \in V}\sum_{x_{ik}^t \in X} \alpha_k f_k(x_{ik}^t, y_i^t)$$ **Attribute**

$$+ \sum_{v_j \in NB(v_i)}\sum_{(y_i^t, y_j^{t'}) \in Y^t} -\beta_{ji}(t-t')(y_i^t - y_j^{t'})^2$$ **Social**

$$+ \sum_{v_i \in V}\sum_{(y_i^t, y_i^{t'}) \in Y^t} -\lambda_i(t-t')(y_i^t - y_i^{t'})^2\}$$ **Temporal**

$$\theta^\star = \arg\max_\theta \ \log\ p(Y = y|x, \theta)$$

MH-based Learning algorithm

Input: number of iterations and learning rate η;
Output: learned parameters $\theta = (\{\alpha_k\}, \{\beta_{ji}\}, \{\lambda_i\})$;

Initialize $\theta = \{\alpha, \beta, \lambda\}$;
repeat
  % sample a new Y' according to $q(Y'|Y)$;
  $Y' \leftarrow q(Y'|Y)$;
  $\tau \sim \min(\frac{p(Y'|G^t,\theta)}{p(Y|G^t,\theta)}, 1)$;
  toss a coin s according to a $Bernoulli(\tau, (1-\tau))$;
  if $(s = 1)$ **then**
    % accept the new configuration Y';
    $Y \leftarrow Y'$;
    if $(Err(Y') < Err(Y)$ & $\Delta\theta F < 0)$ **then**
      $\theta^{new} \leftarrow \theta^{old} + \eta(\Delta\theta F)$;
    end
    else if $(Err(Y') > Err(Y)$ & $\Delta\theta F \geq 0)$ **then**
      $\theta^{new} \leftarrow \theta^{old} - \eta(\Delta\theta F)$;
    end
  end
until *convergence*;

Random Sampling

Update

[1] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and ACM Fong. Quantitative Study of Individual Emotional States in Social Networks. IEEE TAC, 2012, Volume 3, Issue 2, Pages 132-144.

Experiment

- Data Set

	#Users	Avg. Links	#Labels	Other
MSN	30	3.2	9,869	>36,000hr
LiveJournal	469,707	49.6	2,665,166	

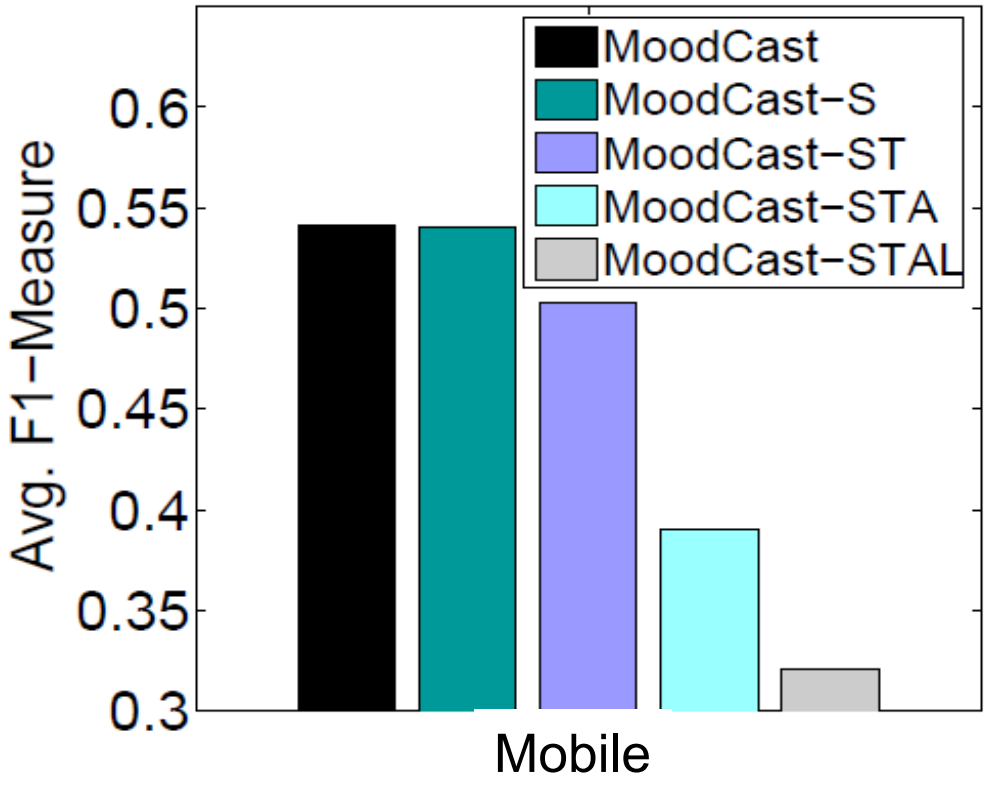
- Baseline
  - SVM
  - SVM with network features
  - Naïve Bayes
  - Naïve Bayes with network features
- Evaluation Measure:

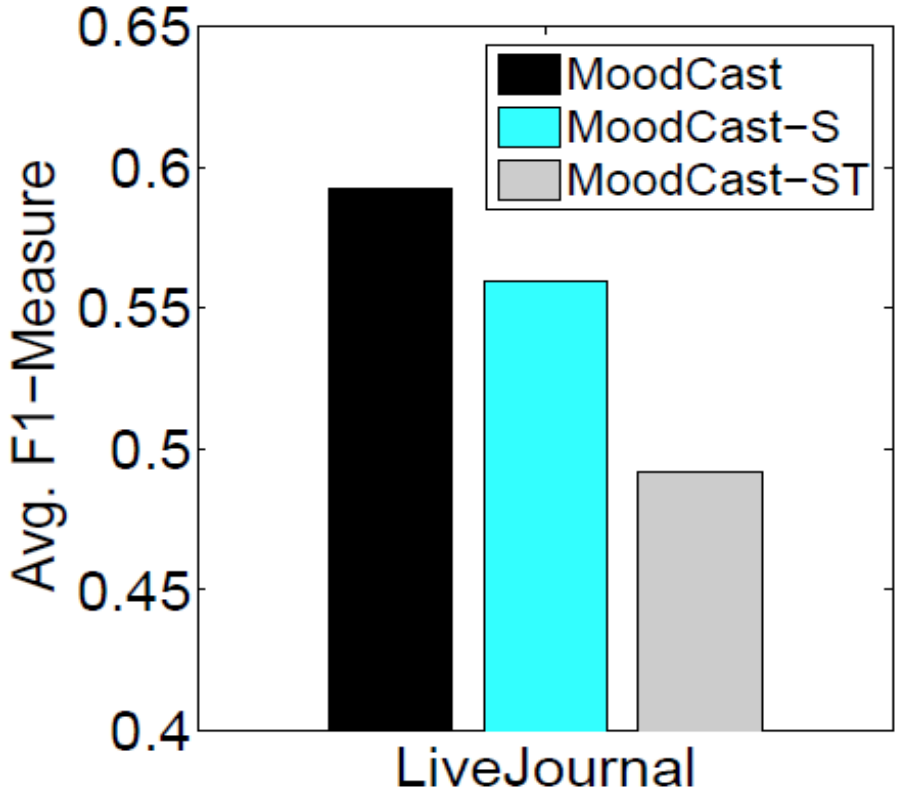
  Precision, Recall, F1-Measure

Performance Result

Classifier	Method	MSN Dataset			LiveJournal Dataset		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Positive	MoodCast	68.42	69.23	**68.82**	52.50	73.68	**61.32**
	SVM-Simple	60.88	71.08	65.58	49.56	48.57	49.06
	SVM-Net	59.12	72.70	65.21	50.72	60.29	55.09
	NB-Simple	67.30	56.21	61.25	57.08	43.34	49.27
	NB-Net	71.89	56.59	63.33	59.1	47.38	52.59
Neutral	MoodCast	67.78	76.57	**71.90**	59.61	84.92	**75.44**
	SVM-Simple	67.39	59.73	63.33	67.58	78.69	72.71
	SVM-Net	68.42	55.11	61.05	71.21	78.13	74.51
	NB-Simple	54.14	68.04	60.30	65.95	54.14	59.46
	NB-Net	51.06	71.62	59.62	61.70	61.53	61.61
Negative	MoodCast	30.77	13.95	19.20	45.45	54.98	**49.77**
	SVM-Simple	5.63	4.54	5.03	71.67	37.39	49.14
	SVM-Net	8.18	16.90	11.02	68.78	37.68	48.68
	NB	14.70	28.16	19.32	54.77	36.61	43.89
	NB-Net	17.88	32.08	**22.96**	51.70	41.18	45.84
Average	MoodCast	55.66	53.25	**53.31**	52.52	71.19	**62.17**
	SVM-Simple	44.63	45.12	44.65	62.94	54.83	56.97
	SVM-Net	45.24	48.23	45.76	63.57	58.70	59.42
	NB-Simple	45.38	50.80	46.95	59.26	44.69	50.87
	NB-Net	46.94	53.43	48.63	57.5	50.03	53.35

Factor Contributions



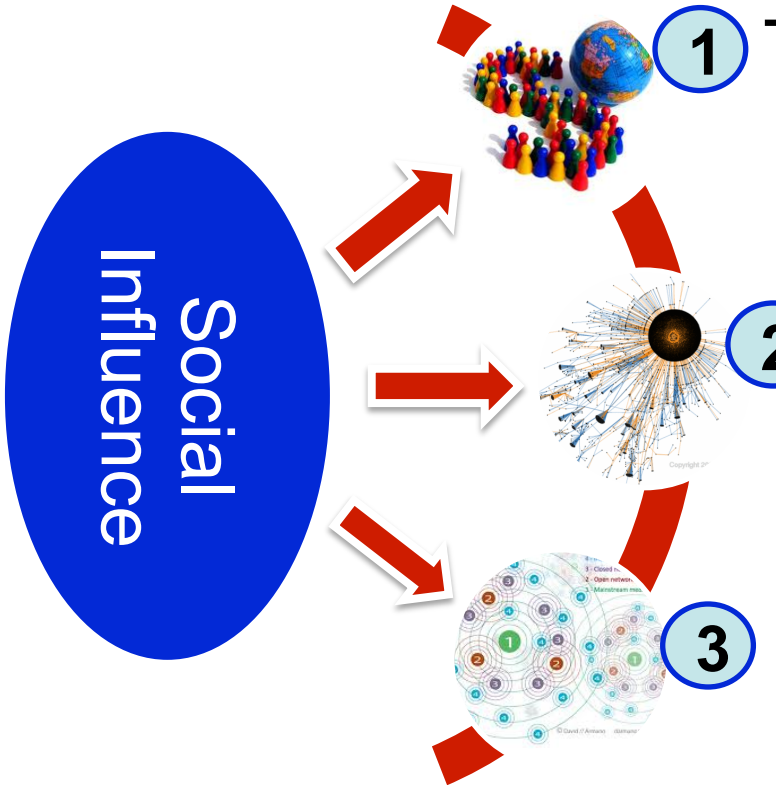


- All factors are important for predicting user emotions

Summaries

- Applications
  - Social advertising
  - Opinion leader finding
  - Social recommendation
  - Emotion analysis
  - etc.

Social Influence Summaries



- Randomization test
- Shuffle test
- Reverse test

- Reachability-based methods
- Structure Similarity
- Structure + Content Similarity
- Action-based methods

- Linear Threshold Model
- Cascade Model
- Algorithms

Related Publications

- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In **KDD'09**, pages 807-816, 2009.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In **KDD'08**, pages 990-998, 2008.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social action tracking via noise tolerant time-varying factor graphs. In **KDD'10**, pages 807–816, 2010.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In **KDD'11**, pages 1397–1405, 2011.
- Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. Can We Understand van Gogh's Mood? Learning to Infer Affects from Images in Social Networks. In **ACM MM**, pages 857-860, 2012.
- Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining Topic-Level Influence in Heterogeneous Networks. In **CIKM'10**, pages 199-208, 2010.
- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. In **TKDD**.
- Jimeng Sun and Jie Tang. Models and Algorithms for Social Influence Analysis. In **WSDM'13**. (Tutorial)
- Lu Liu, Jie Tang, Jiawei Han, and Shiqiang Yang. Learning Influence from Heterogeneous Social Networks. In **DMKD**, 2012, Volume 25, Issue 3, pages 511-544.
- Jimeng Sun and Jie Tang. A Survey of Models and Algorithms for Social Influence Analysis. Social Network Data Analytics, Aggarwal, C. C. (Ed.), Kluwer Academic Publishers, pages 177–214, 2011.
- Chi Wang, Jie Tang, Jimeng Sun, and Jiawei Han. Dynamic Social Influence Analysis through Time-dependent Factor Graphs. In **ASONAM'11**, pages 239-246, 2011.
- Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Social Influence on User Following Behaviors in Social Networks. (submitted)

References

- S. Milgram. The Small World Problem. **Psychology Today**, 1967, Vol. 2, 60–67
- 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
- R. Dunbar. Neocortex size as a constraint on group size in primates. **Human Evolution**, 1992, 20: 469–493.
- 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.
- http://klout.com
- Why I Deleted My Klout Profile, by Pam Moore, at **Social Media Today**, originally published November 19, 2011; retrieved November 26 2011
- S. Aral and D Walker. Identifying Influential and Susceptible Members of Social Networks. **Science**, 337:337-341, 2012.
- J. Ugandera, L. Backstromb, C. Marlowb, and J. Kleinberg. Structural diversity in social contagion. **PNAS**, 109 (20):7591-7592, 2012.
- S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. **PNAS**, 106 (51):21544-21549, 2009.
- J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In **KDD'09**, pages 747–756, 2009.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology** 66, 5, 688–701.
- http://en.wikipedia.org/wiki/Randomized_experiment

References(cont.)

- A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks. In **KDD'08**, pages 7-15, 2008.
- L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.
- G. Jeh and J. Widom. Scaling personalized web search. In **WWW '03**, pages 271-279, 2003.
- G. Jeh and J. Widom, SimRank: a measure of structural-context similarity. In **KDD'02**, pages 538-543, 2002.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In **WSDM'10**, pages 207–217, 2010.
- P. Domingos and M. Richardson. Mining the network value of customers. In **KDD'01**, pages 57–66, 2001.
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In **KDD'03**, pages 137–146, 2003.
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In **KDD'07**, pages 420–429, 2007.
- W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In **KDD'09**, pages 199-207, 2009.
- E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In **EC'12**, pages 146-161, 2012.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In **CIKM'08**, pages 499–508, 2008.
- N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In **WSDM'08**, pages 207–217, 2008.

References(cont.)

- E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In **EC '09**, pages 325–334, New York, NY, USA, 2009. ACM.
- P. Bonacich. Power and centrality: a family of measures. **American Journal of Sociology**, 92:1170–1182, 1987.
- R. B. Cialdini and N. J. Goldstein. Social influence: compliance and conformity. **Annu Rev Psychol**, 55:591–621, 2004.
- D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In **KDD'08**, pages 160–168, 2008.
- P. W. Eastwick and W. L. Gardner. Is it a game? evidence for social influence in the virtual world. **Social Influence**, 4(1):18–32, 2009.
- S. M. Elias and A. R. Pratkanis. Teaching social influence: Demonstrations and exercises from the discipline of social psychology. **Social Influence**, 1(2):147–162, 2006.
- T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In **WWW'10**, 2010.
- M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In **KDD'10**, pages 1019–1028, 2010.
- M. E. J. Newman. A measure of betweenness centrality based on random walks. **Social Networks**, 2005.
- D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. **Nature**, pages 440–442, Jun 1998.
- J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In **ICDM'05**, pages 418–425, 2005.

Thank you !

Collaborators: John Hopcroft, Lillian Lee, Chenhao Tan (**Cornell**)

Jiawei Han and Chi Wang (**UIUC**)

Tiancheng Lou (**Google**)

Wei Chen, Ming Zhou, Long Jiang (**Microsoft**)

Jing Zhang, Zhanpeng Fang, Zi Yang, Sen Wu, Jia Jia (**THU**)

Jie Tang, KEG, Tsinghua U, http://keg.cs.tsinghua.edu.cn/jietang

Jimeng Sun, IBM TJ Watson, http://www.dasfa.net/jimeng

Download all data & Codes, http://arnetminer.org/download