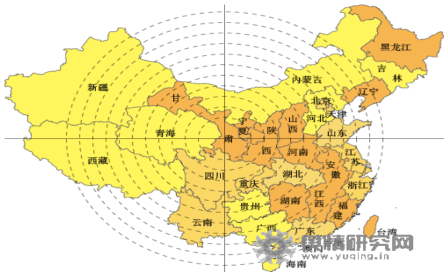
Computational Models for Social Networks

Jie Tang

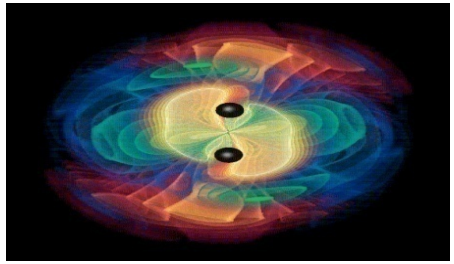
Tsinghua University, China

Social Networks

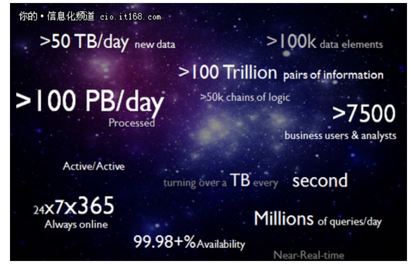
SN bridges our daily life and the **virtual** web space!



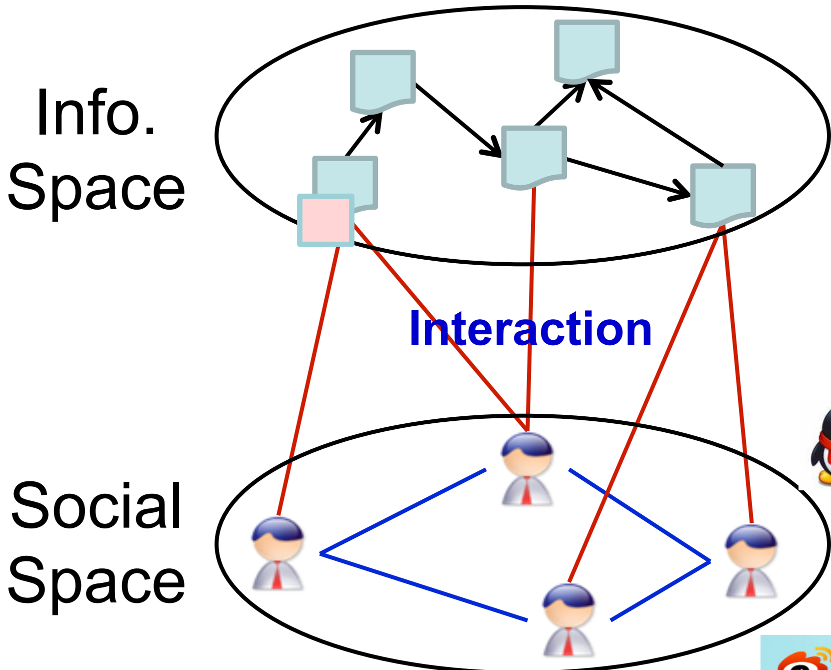
Opinion Mining



Innovation diffusion



Business Intelligence

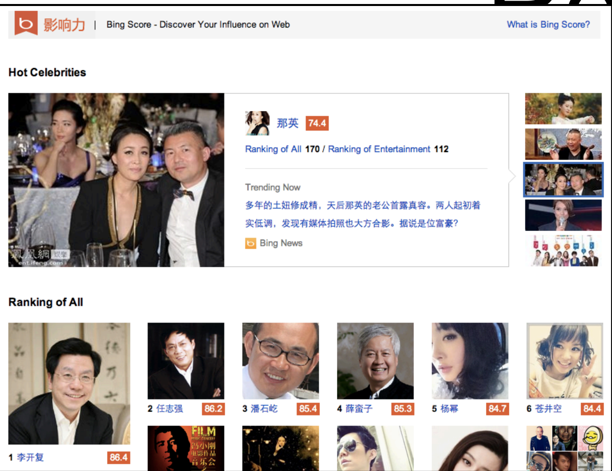


Revolutionary changes…

Req: Info.→ user

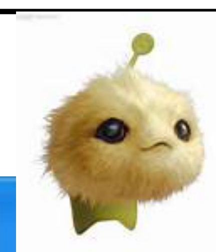
Interaction mechanism

Revolutionary Changes



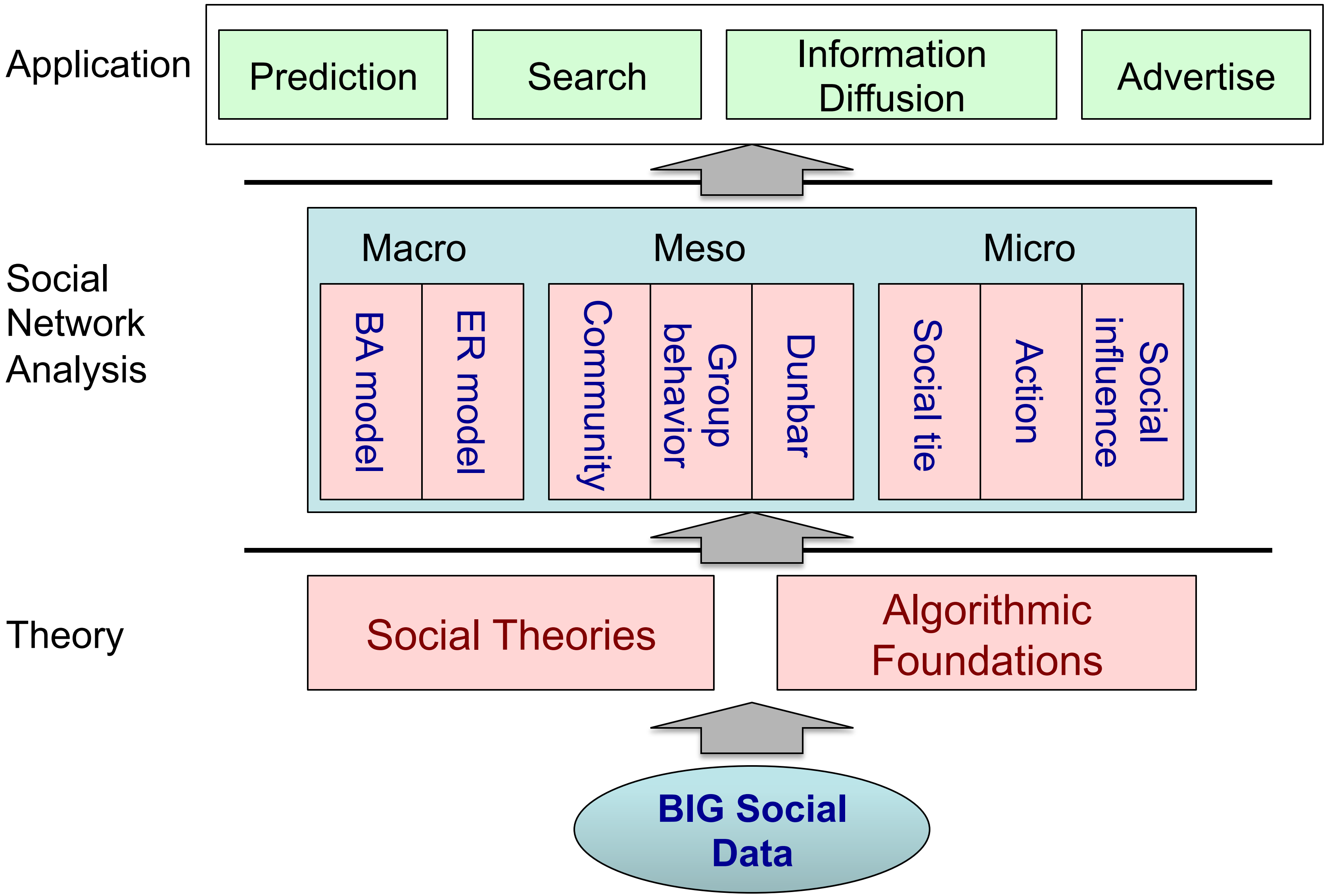
Social Networks

Search	Education	O2O	...
Embedding social in search: • Google plus • FB graph search • Bing's influence	Human Computation: • CAPTCHA + OCR • MOOC • Duolingo (Machine Translation)	The Web knows you than yourself: • Contextual computing • Big data marketing	More …

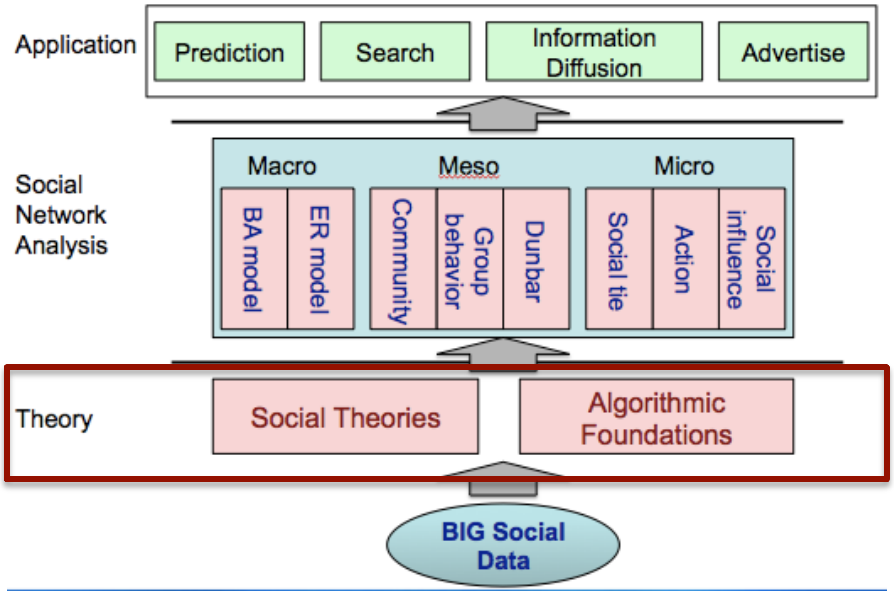


Part A: Overview of Core Research in Social Networks

Core Research in Social Network



Computational Foundations for Social Networks

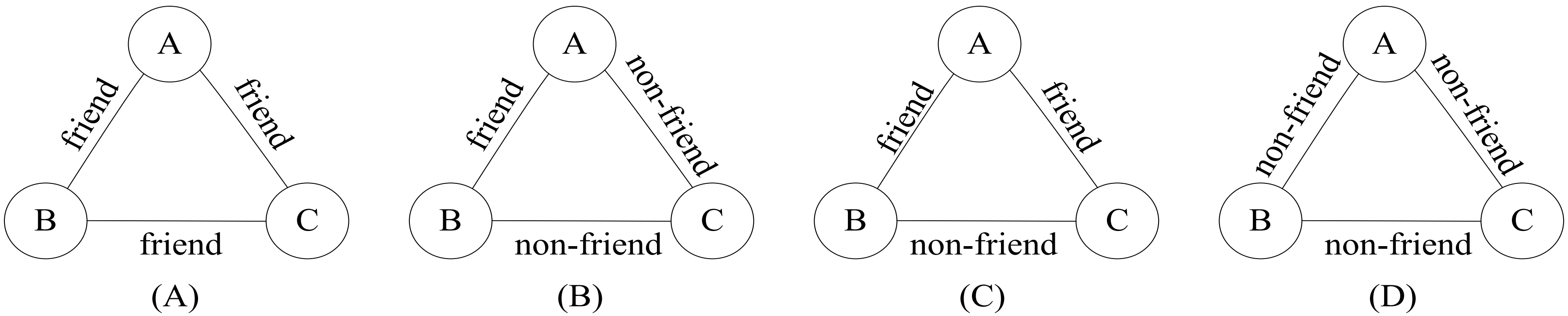


Computational Foundations

- Social Theories
  - Social balance
  - Social status
  - Structural holes
  - Two-step flow
- Algorithmic Foundations
  - Network flow
  - K-densest subgraph
  - Set cover

Social Theories—Social Balance

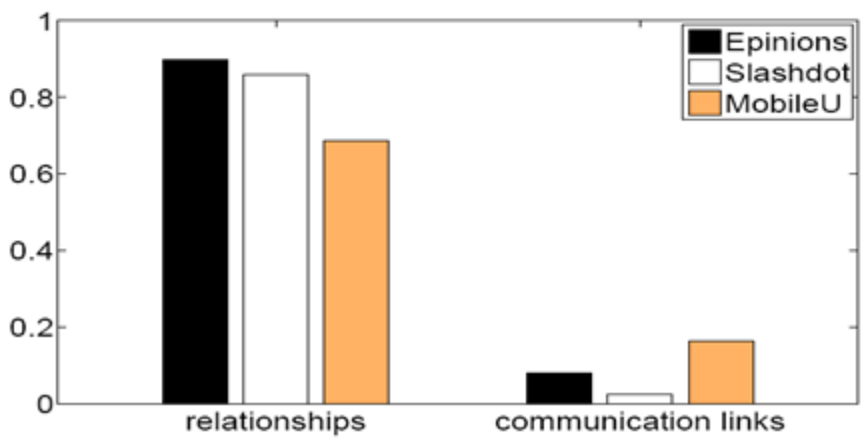
Your friend's friend is your friend, and your enemy's enemy is also your friend.



Examples on Epinions, Slashdot, and MobileU

(1) The underlying networks are unbalanced;

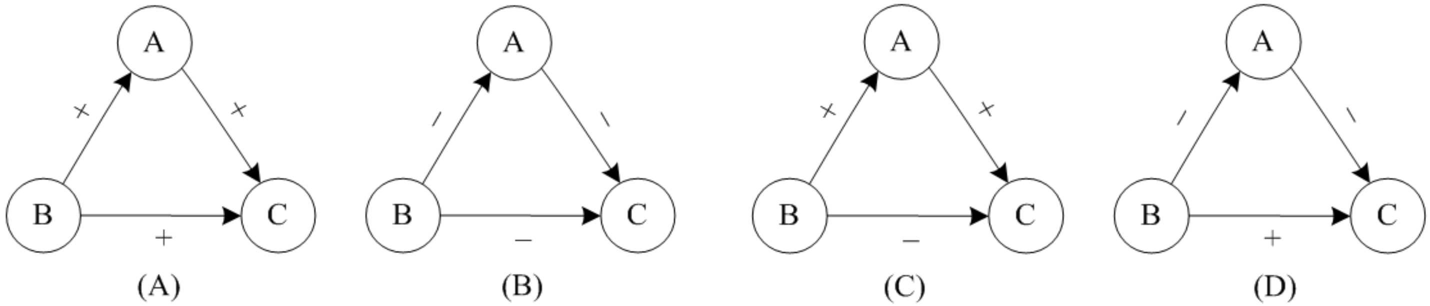
(2) While the friendship networks are balanced.



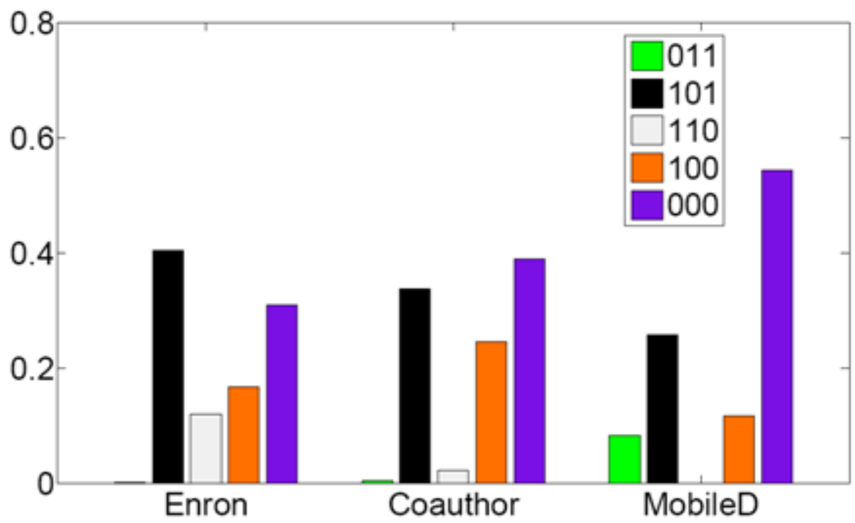
Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. In WSDM'2012. pp. 743-752.

Social Theories—Social status

Your boss's boss is also your boss…



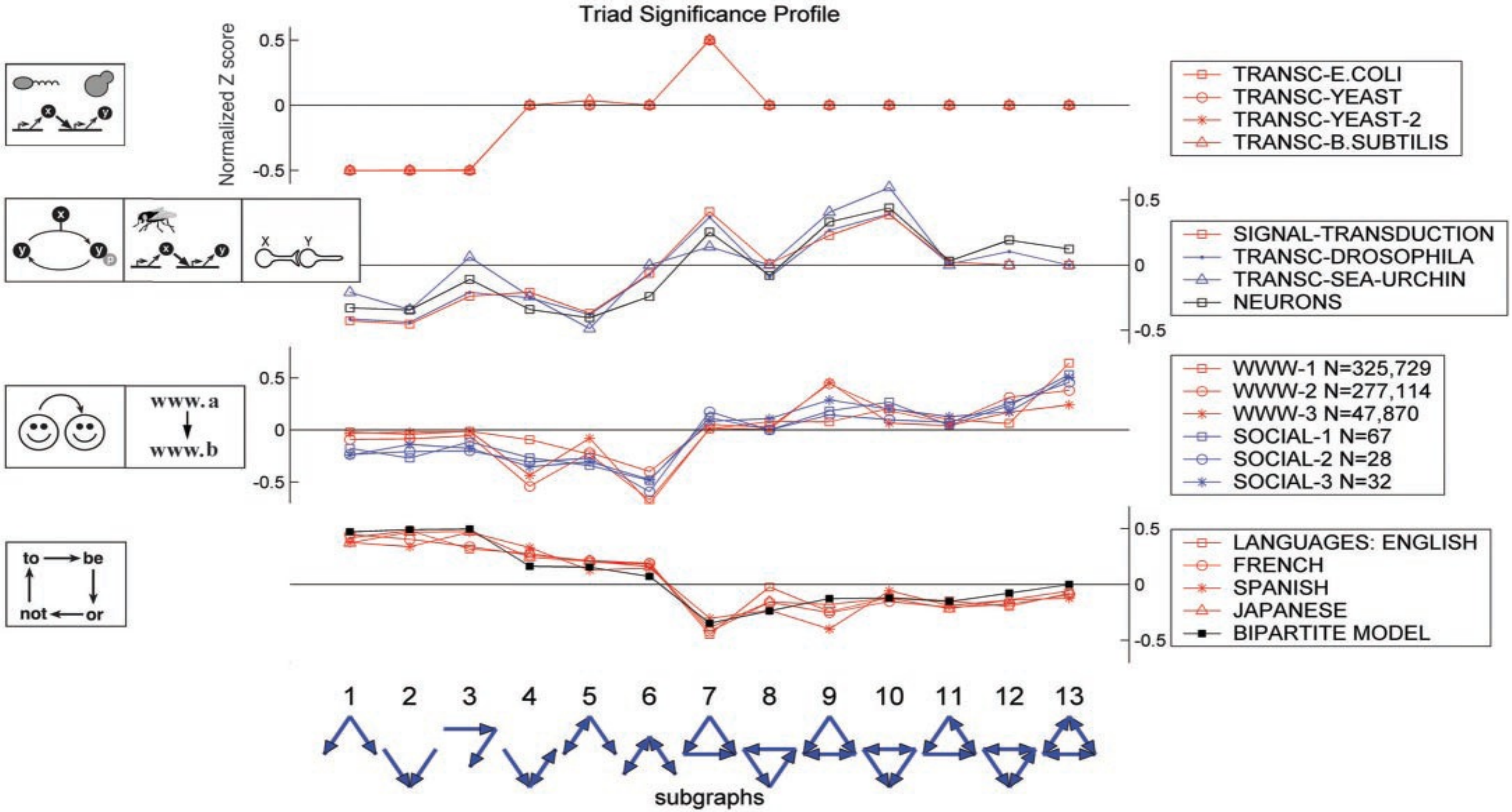
Observations: 99% of triads in the networks satisfy the social status theory
Examples: Enron, Coauthor, MobileD



Note: Given a triad (A,B,C), let us use 1 to denote the advisor-advisee relationship and 0 colleague relationship. Thus the number 011 to denote A and B are colleagues, B is C's advisor and A is C's advisor.

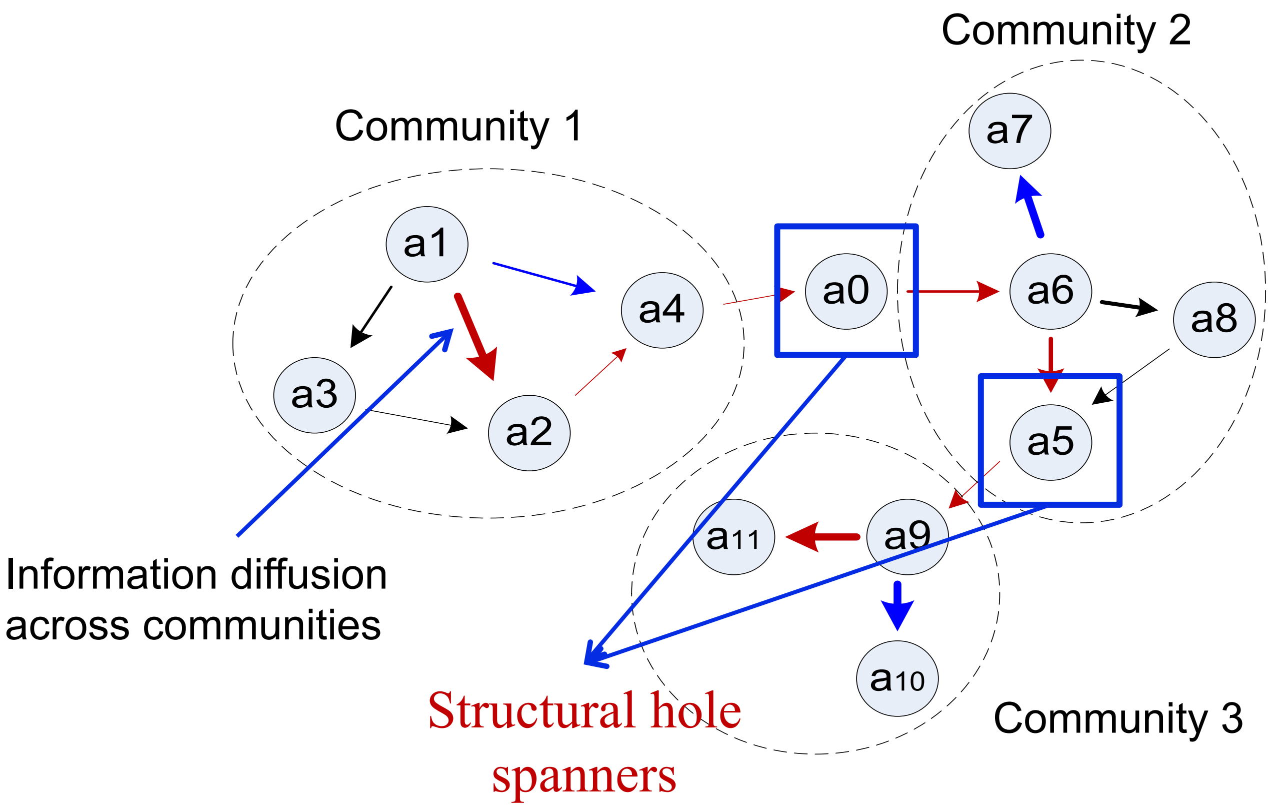
Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. In WSDM'2012. pp. 743-752.

Triadic Closure



R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. Science, 2004

Social Theories—Structural holes



twitter

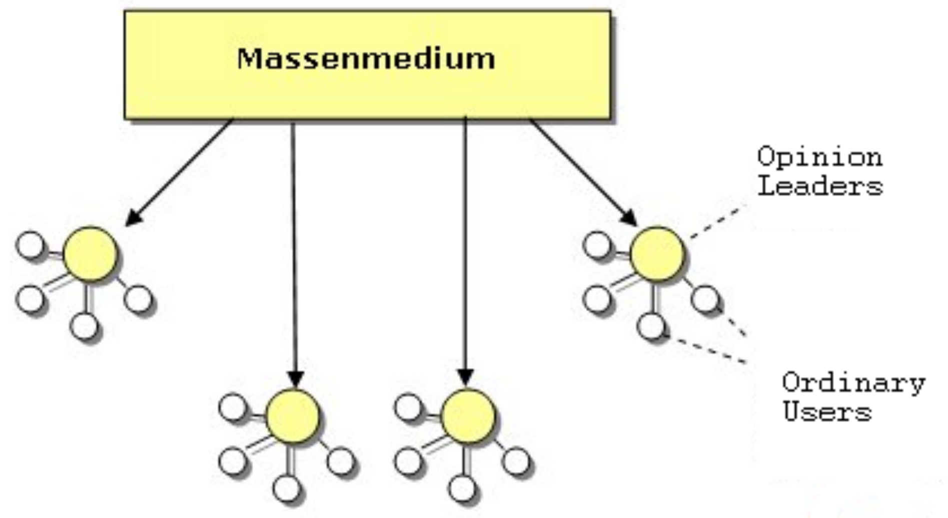
1% twitter users control **25%** retweeting behaviors between communities.

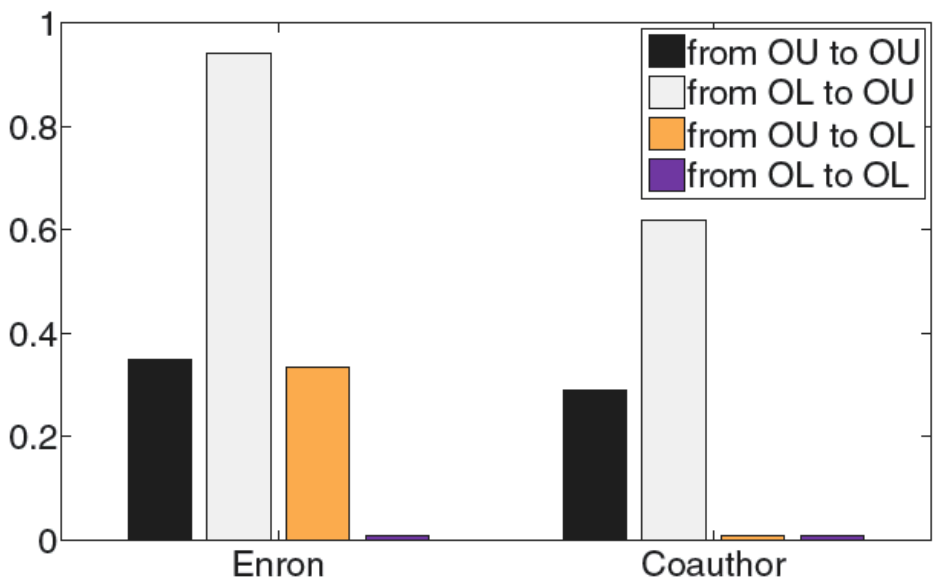
Structural hole users control the information flow between different communities (Burt, 92; Podolny, 97; Ahuja, 00; Kleinberg, 08; Lou & Tang, 13)

T. Lou and J. Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In **WWW'13**. pp. 837-848.

Social Theories—Two-step-flow

Lazarsfeld *et al* suggested that:
"ideas often flow from radio and print to the opinion leaders and from them to the less active sections of the population."





Estimate OL and OU by PageRank
OL : Opinion leader;
OU : Ordinary user.

Observations: Opinion leaders are more likely (+71%-84% higher than chance) to spread information to ordinary users.

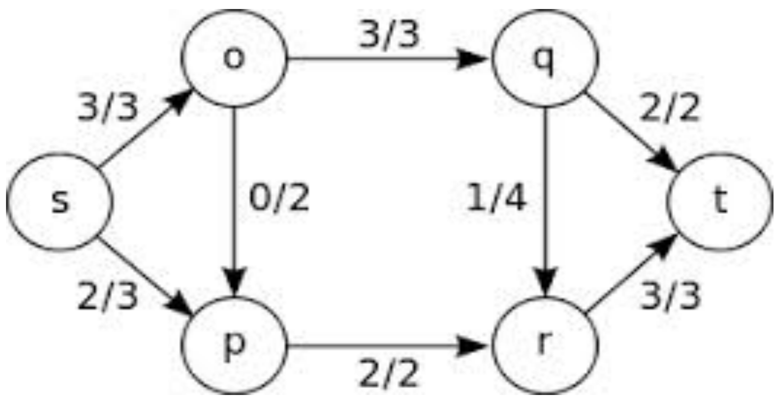
Lazarsfeld et al. (1944). The people's choice: How the voter makes up his mind in a presidential campaign.

Computational Foundations

- Social Theories
  - Social balance
  - Social status
  - Structural holes
  - Two-step flow
- Algorithmic Foundations
  - Network flow
  - K-densest subgraph
  - Set cover

Algorithm — Network Flow

- Classical problems:
  - Maximum flow / minimum cut
    - Ford-Fulkerson algorithm
    - Dinic algorithm
  - Minimum cut between multiple sets of vertices
    - NP hard when there are more than 2 sets
  - Minimum cost flow;
  - Circulation problem;
  - …



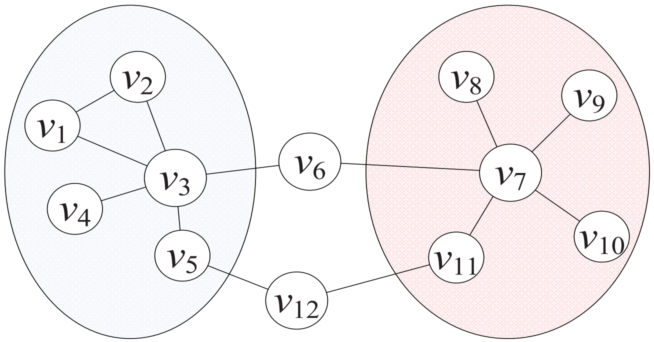
Algorithm — Network Flow (cont.)

- Ford-Fulkerson
  - As long as there is an augmenting path, send the minimum of the residual capacities on the path.
  - A maximum flow is obtained when the no augmenting paths left.
  - Time complexity: O(VE^2)

```
FORD-FULKERSON(G, s, t)
for each edge (u, v) ∈ E[G]
      do f[u, v] ← 0
         f[v, u] ← 0
while there exists a path p from s
to t in the residual network Gf
      do cf(p) ← min {cf(u, v) : (u, v) is in p}
         for each edge (u, v) in p
               do f[u, v] ← f[u, v] + cf(p)
                  f[v, u] ← -f[u, v]
```

Algorithm — K-densest subgraph

- NP Problem
  - Find the maximum density subgraph on exactly k vertices.
  - Reduced from the clique problem
- Application
  - Reduce the structural hole spanner detection problem to proof its NP hardness.
  - To find a subset of nodes, such that without them, the connection between communities would be minimized.



Algorithm — K-densest subgraph (cont.)

- An linear programming based solution
  - Approximation ratio: $O(n^{1/4+\varepsilon})$

Find the subgraph with the largest average degree in subgraph S_{t-1}

Find j which satisfy:

$$\mathrm{LP}_{\{y_{ij}/y_j | i \in V\}}(S \cap \Gamma(j)) \geq \frac{d \cdot \mathrm{LP}_{\{y_i\}}(S)}{2k}, \textit{ and}$$

$$\mathrm{LP}_{\{y_{ij}/y_j | i \in V\}}(S \cap \Gamma(j))/|S \cap \Gamma(j)| \geq \frac{d \cdot \mathrm{LP}_{\{y_i\}}(S)}{2\rho \cdot \max\{k, |S|\}}.$$

Update S by j's neighbors.

: of Procedure DkS-Local(S_{t-1}, k).

contains an integer, perform a **hair step**:

emma 4.4 (or for $t = 1$, choose any j_1 such

* Let $S_t = S_{t-1} \cap \Gamma(j_t)$.
* Replace the LP solution $\{y_i\}$ with $\{y_{ij_t}/y_{j_t} \mid i \in V\}$.

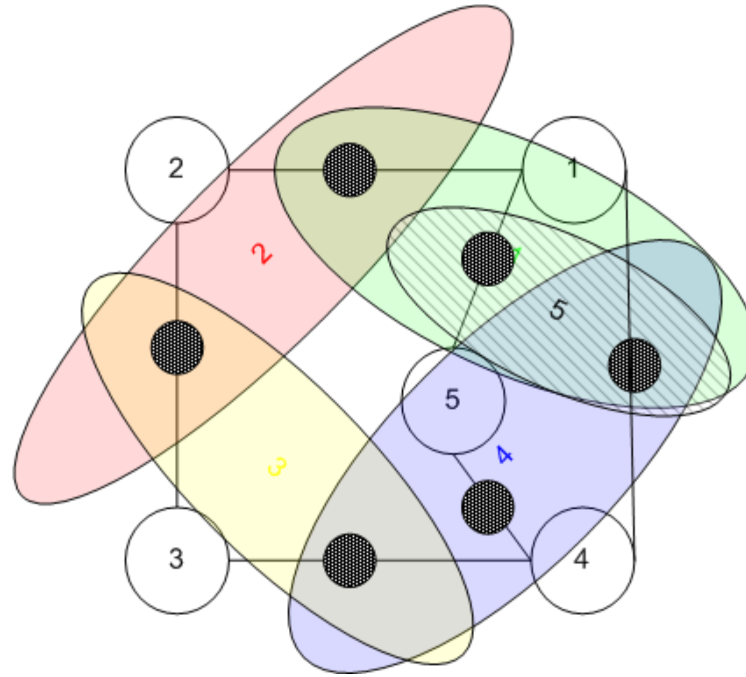
– Otherwise, perform a **backbone step**: Let $S_t = \Gamma(S_{t-1})$.

Replace S_t by neighbors of S_{t-1}

- Output the subgraph H_t with the highest average degree.

Algorithm — Set Cover

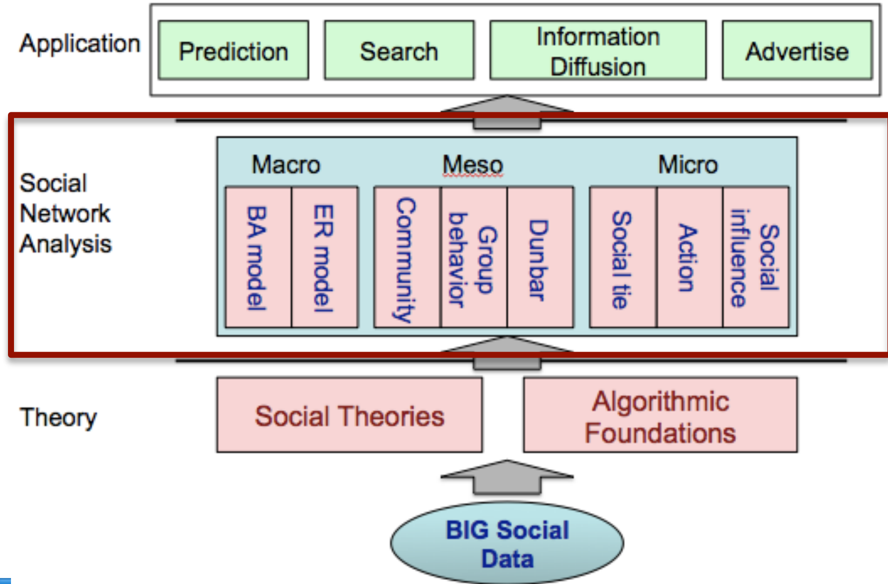
- Another NP problem
  - Given a set of elements (universe) and a set *S* of *n* sets whose union equals the universe;
  - Find the smallest subset of *S* to contains all elements in the universe;
  - The decision version is NP-complete.
- Greedy
  - Choose the set containing the most uncovered elements;
  - Approximation ratio: *H*(size(*S*)), where *H*(*n*) is the n-th harmonic number.



$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{k}.$$

Social Network Analysis

- Macro Level
- Meso Level
- Micro Level



Erdős–Rényi Model

In the *G*(*n*, *p*) model, each edge is included in the graph with probability p independent from every other edge.

Each random graph has the probability

$$p^M(1-p)^{\binom{n}{2}-M}.$$

- Properties

(1) Degree distribution-Poisson

$$p(k) = \frac{<k>^k}{k!} e^{-<k>}$$

(2) Clustering coefficient $\longrightarrow$ Small

$$p$$

(3) Average shortest path

$$L \sim \frac{\ln N}{\ln <k>}$$

Problem: In real social network, neighbors tend to be connected with each other, thus the clustering coefficient should not be too small.

Erdős, P.; Rényi, A. (1959), “On Random Graphs.”.

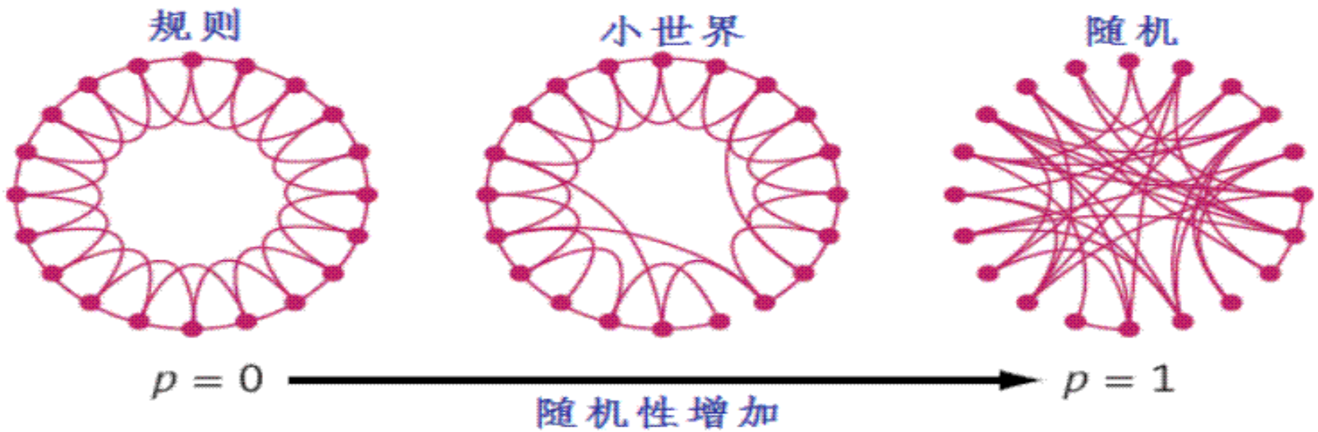
Small-World Model

Mechanism

1. Start from a regular wired ring, where each node is connected with its K-nearest neighbors

2. With probability p rewire each edge.

Problem: In real social network, degree distribution is power law.



- Properties

(1) Degree distribution

$$p(k) = \begin{cases} 0, k < K \\ \dfrac{<d>}{(k-K)!} e^{-<d>}, k \geq K \end{cases} \quad <d> = Kp$$

→ Not power law

(2) Clustering coefficient

$$C = \frac{3(K-2)}{4(K-1) + 4Kp(p+2)}$$

(3) Average shortest path

$$L = \frac{\ln NKp}{K^2 p}$$

Watts, D. J.; Strogatz, S. H. (1998). "Collective dynamics of 'small-world' networks". Nature 393 (6684): 440–442.

Barabási-Albert Model

Idea

- Growth
- Preferential attachment (rich-get-richer, the Matthew Effect)

Mechanism

1. Start from a small connected graph with m_0 nodes
2. At each time step, add one new node with m (m ≤ m_0) new edges; the probability that the new node is connected to node i is $p_i = \frac{k_i}{\sum_j k_j}$

- Degree distribution

$$p(k) = 2m^2 k^{-3}$$

Scale-free

- Clustering coefficient

$$C \sim \frac{(\ln t)^2}{t}$$

- Average longest shortest path

$$L \sim \frac{\ln N}{\ln \ln N}$$

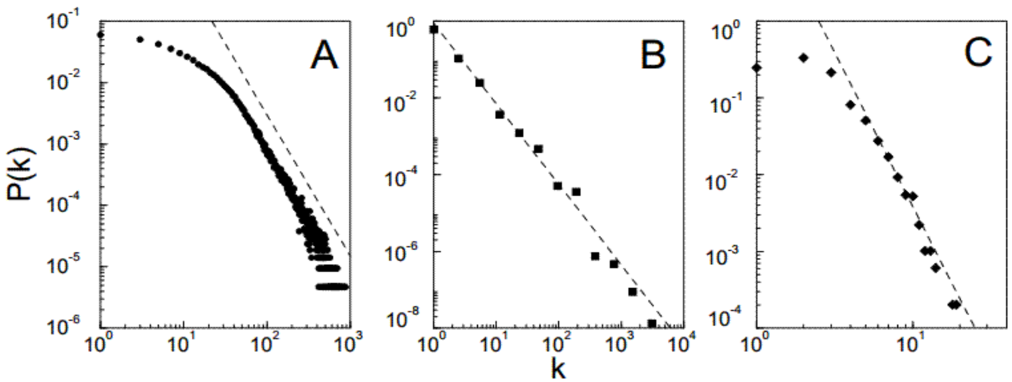
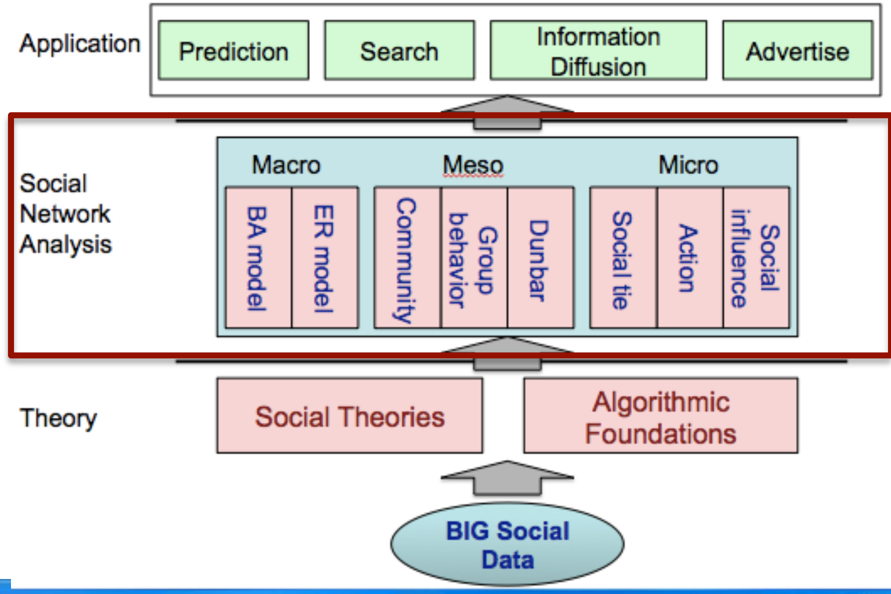


FIG. 1. The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with $N = 212,250$ vertices and average connectivity $\langle k \rangle = 28.78$; (**B**) World wide web, $N = 325,729$, $\langle k \rangle = 5.46$ (6); (**C**) Powergrid data, $N = 4,941$, $\langle k \rangle = 2.67$. The dashed lines have slopes (**A**) $\gamma_{actor} = 2.3$, (**B**) $\gamma_{www} = 2.1$ and (**C**) $\gamma_{power} = 4$.

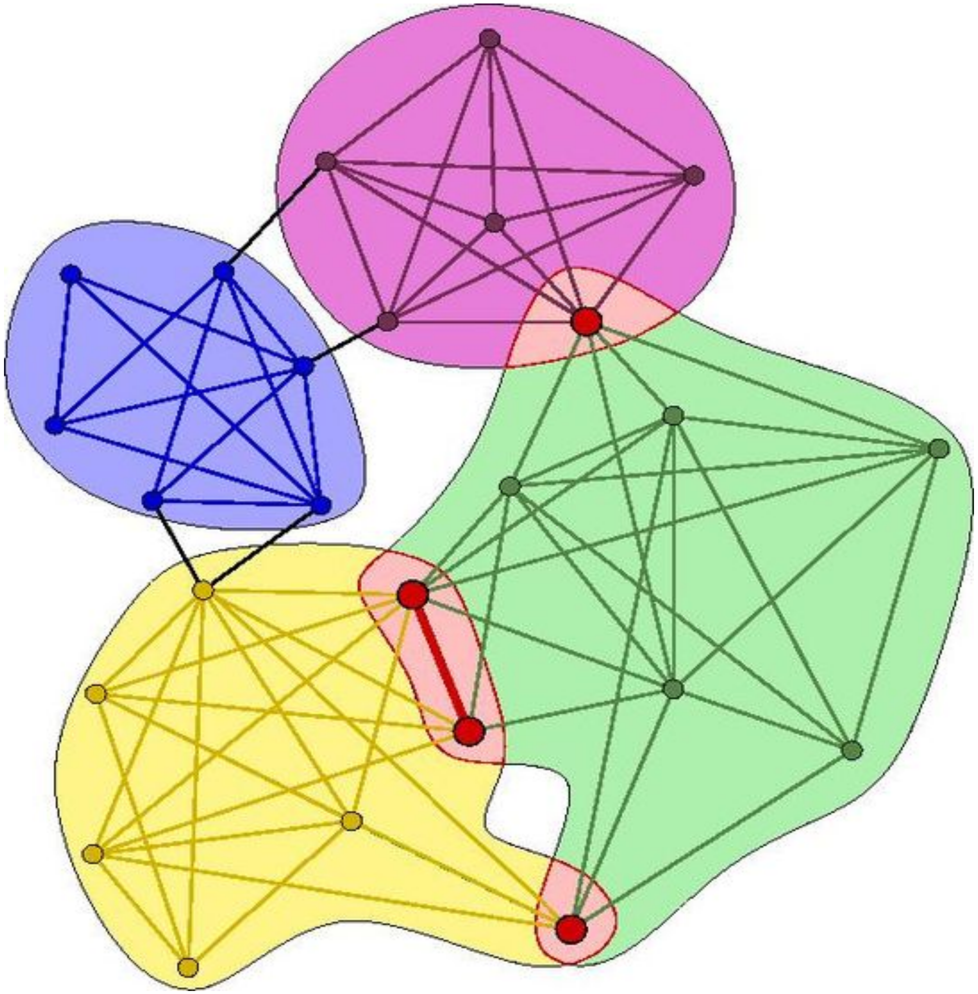
Barabasi and Albert(1999). Emergence of scaling n complex networks.

Social Network Analysis

- Macro Level
- Meso Level
- Micro Level



Community Detection



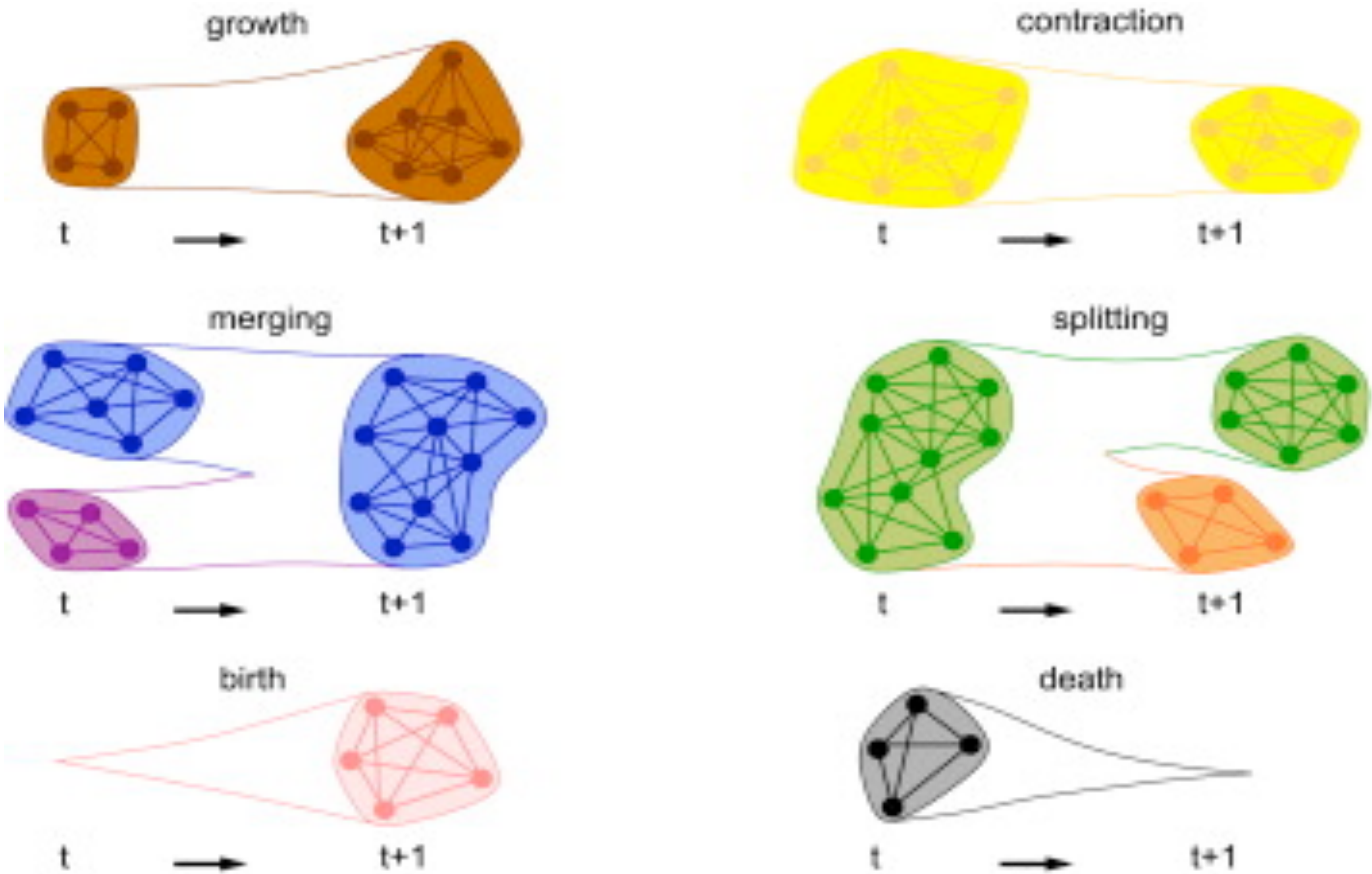
Node-Centric Community
- Each node in a group satisfies certain properties

Group-Centric Community
- Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level

Network-Centric Community
- Partition the whole network into several disjoint sets

Hierarchy-Centric Community
- Construct a hierarchical structure of communities

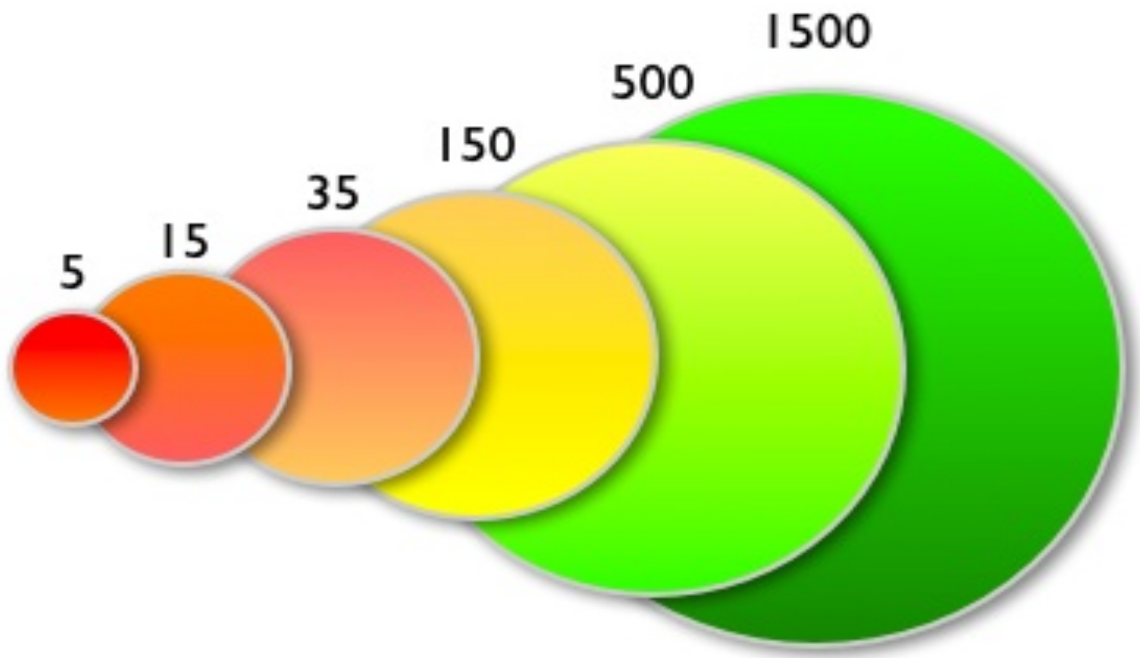
Community Evolution



Dunbar Number

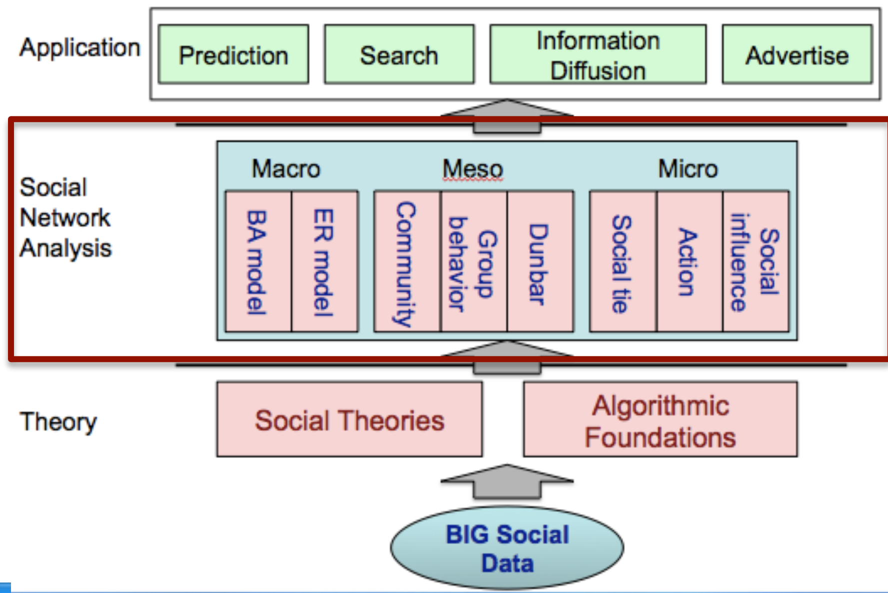
- Dunbar number:150. Dunbar's number is a suggested cognitive limit to the number of people with whom one can maintain stable social relationships

—Robin Dunbar, 2000



Social Network Analysis

- Macro Level
- Meso Level
- Micro Level



Social Action

- …the object is to interpret the meaning of social action and thereby give a causal explanation of the way in which the action proceeds and the effects which it produces...

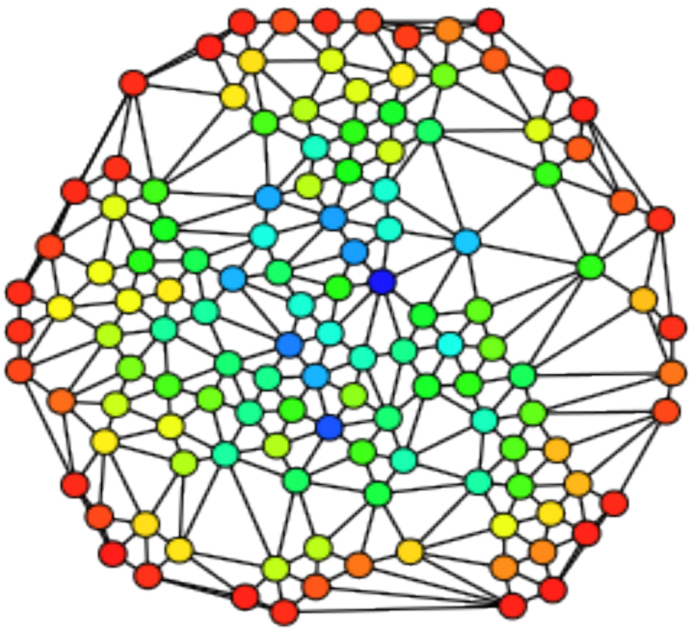
— Social Action Theory, by Max Weber, 1922

Social Action — User Characterization

- Betweenness
  - A centrality measure of a vertex within a graph
  - $C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$

#shortest paths pass through v

#shortest paths from s to t



Hue (from red=0 to blue=max) shows the node betweenness.

Social Action — User Characterization (cont.)

- Clustering Coefficient
  - A measure of degree to which nodes in a graph tend to cluster together.
  - Global clustering coefficient
    - $$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}.$$
    - A triangle consists of three closed triplets, and a closed triplet consists of three nodes connected to each other.
  - Local clustering coefficient

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

Social Action — User Characterization (cont.)

- Degree: the number of one vertex's neighbors.
- Closeness: the shortest path between one vertex and another vertex.

Social Action — Game Theory

- Example: a game theory model on Weibo.
  - Strategy: whether to follow a user or not;
  - Payoff:

The value of a user

The density of v's ego network

$$P(u) = \alpha_u \sum_{v \in B(u)} G(v) - \sum_{v \in L(u)} C + \sum_{v \in B(u)} \log_2\Big(\sum_{w \in L(v) \cap F(u)} C_2 \Big)$$

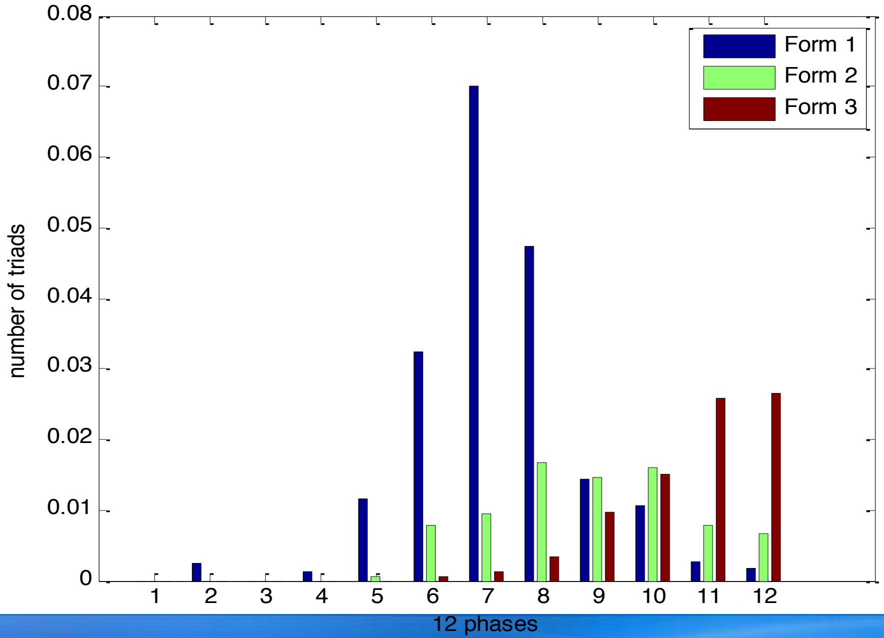
The frequency of a user to follow someone

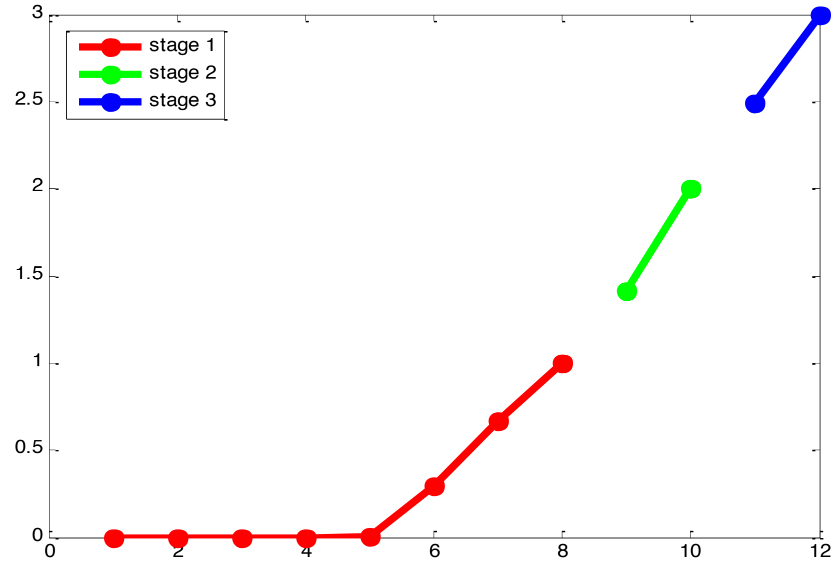
The cost of following a user

  - The model has a pure strategy Nash Equilibrium

Social Action — Game Theory (cont.)

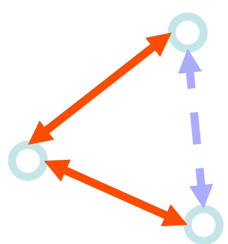
- Results: three stage life cycle
  - Stage 1: getting into a community
  - Stage 2: becoming an elite
  - Stage 3: bridging different communities (structural hole spanners)



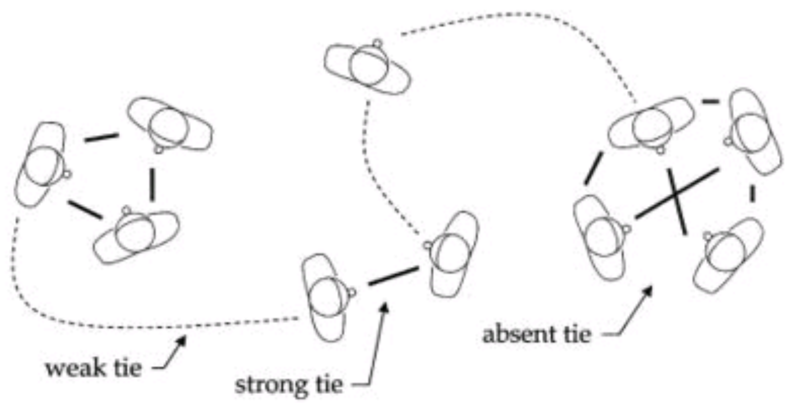


Strong/Weak Ties

- Strong ties
  - Frequent communication, but ties are redundant due to high clustering
- Weak ties
  - Reach far across network, but communication is infrequent…



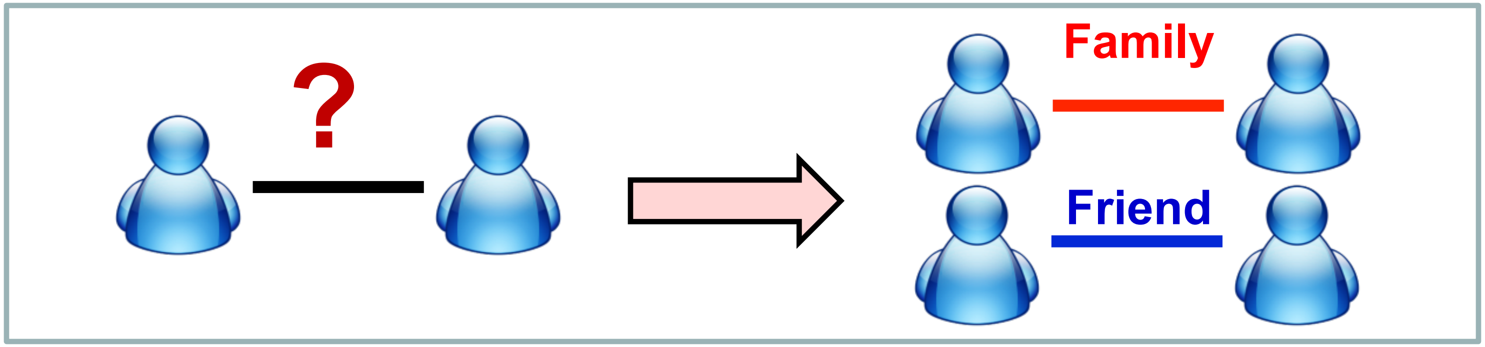
“forbidden triad”：
strong ties are likely to “close”



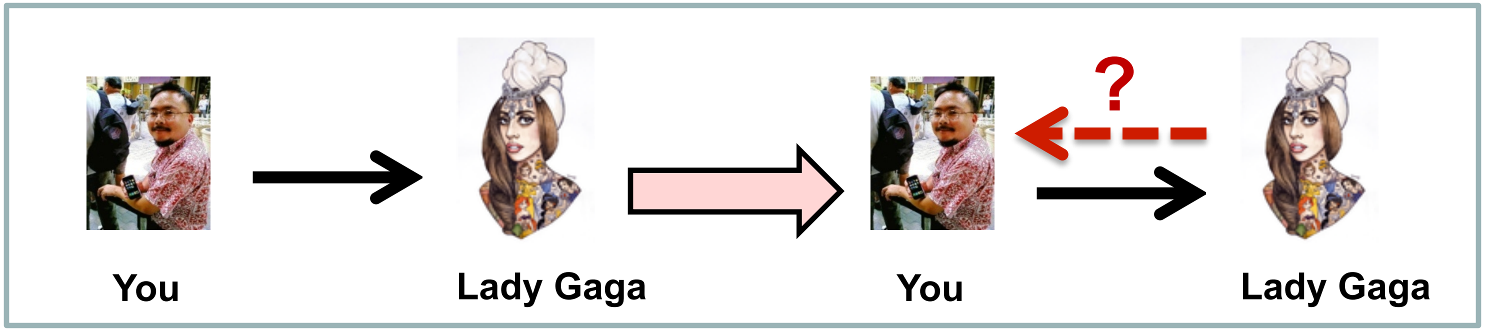
Weak ties act as local bridge

Social Ties

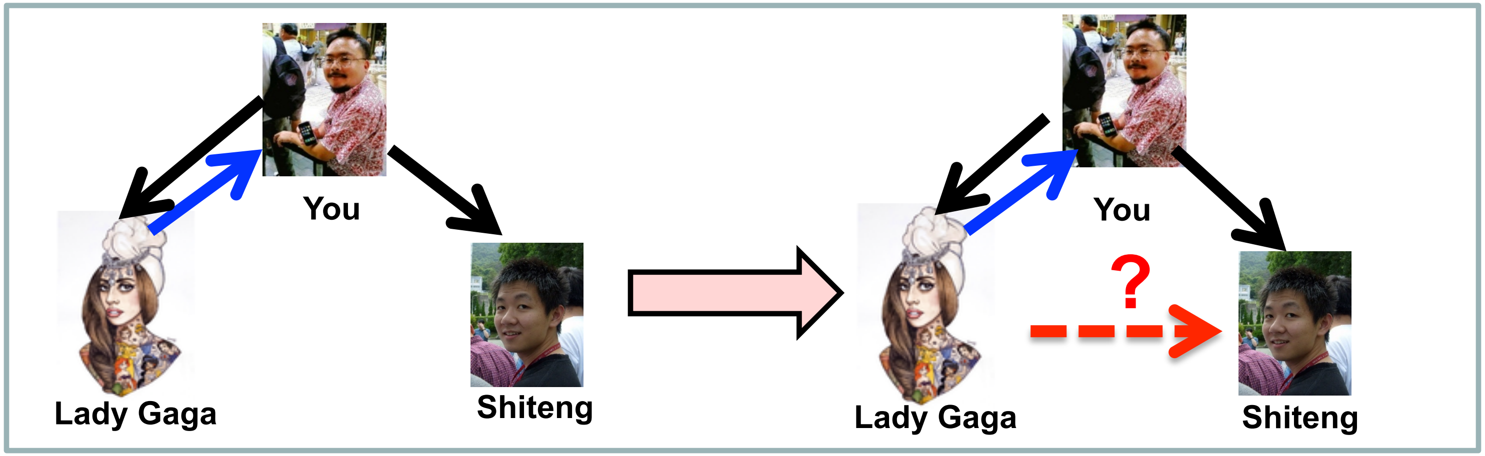
Inferring social ties



Reciprocity



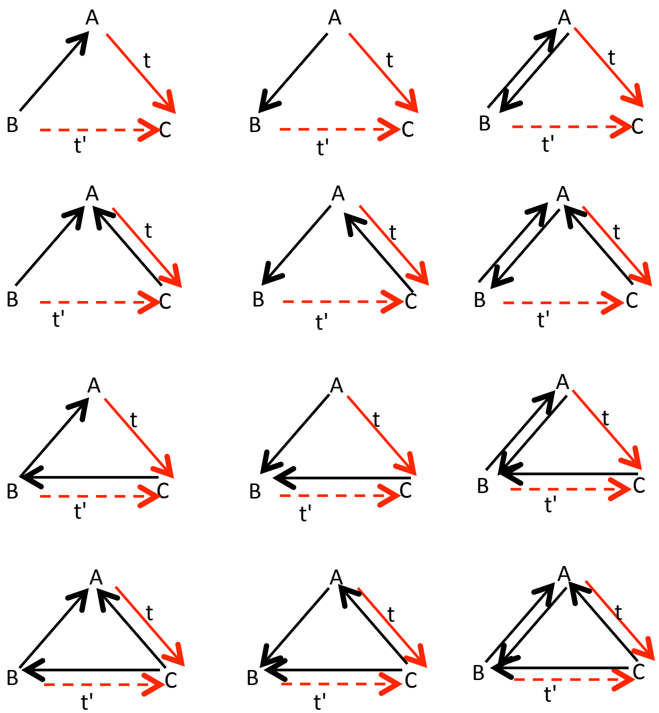
Triadic Closure



KDD 2010, PKDD 2011 (*Best Paper Runnerup*), WSDM 2012, ACM TKDD

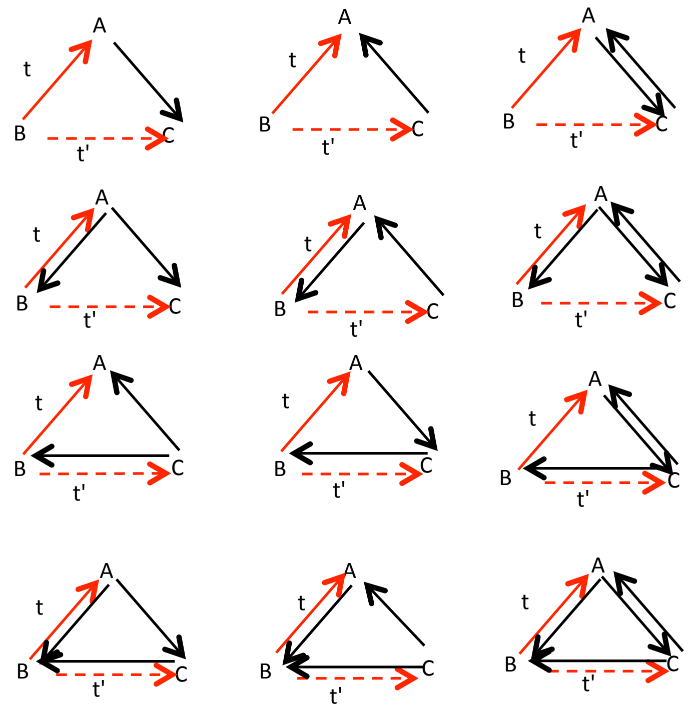
Triadic Closure

Follower diffusion



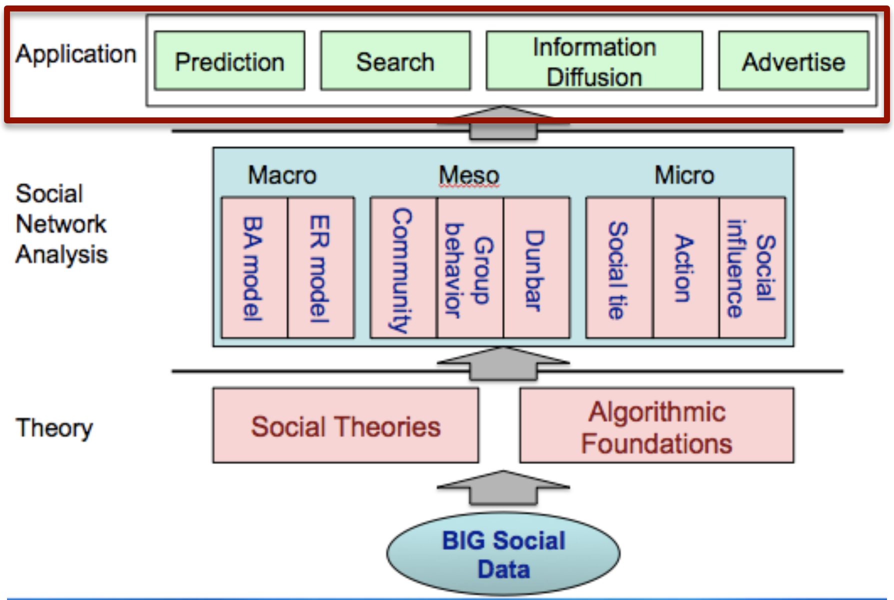
12 triads

Followee diffusion



12 triads

Information Diffusion



Disease-Propagation Models

- Classical disease-propagation models in epidemiology are based upon the cycle of disease in a host.
  - Susceptible
  - Infected
  - Recovered
  - …
- The transition rates from one cycle to another are expressed as derivatives.
- Classical models:
  - SIR
  - SIS
  - SIRS
  - …

SIR Model

- Created by Kermack and McKendrick in 1927.
- Considers three cycles of disease in a host:

- Transition rates:

$$\frac{dS}{dt} = -\beta S(t)I(t)$$
$$\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t)$$
$$\frac{dR}{dt} = \gamma I(t)$$

S(t) : #susceptible people at time t;

I(t) : #infected people at time t;

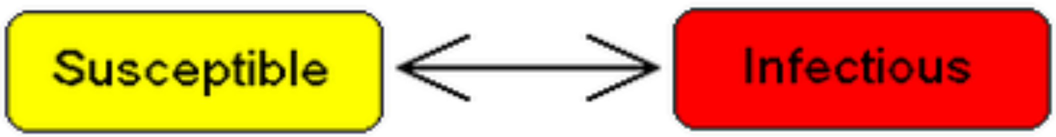
R(t) : #recovered people at time t;

β *: a parameter for infectivity;*

γ *: a parameter for recovery.*

SIS Model

- Designed for infections confer no long lasting immunity (e.g., common cold)
- Individuals are considered become susceptible again after infection:



- Model:

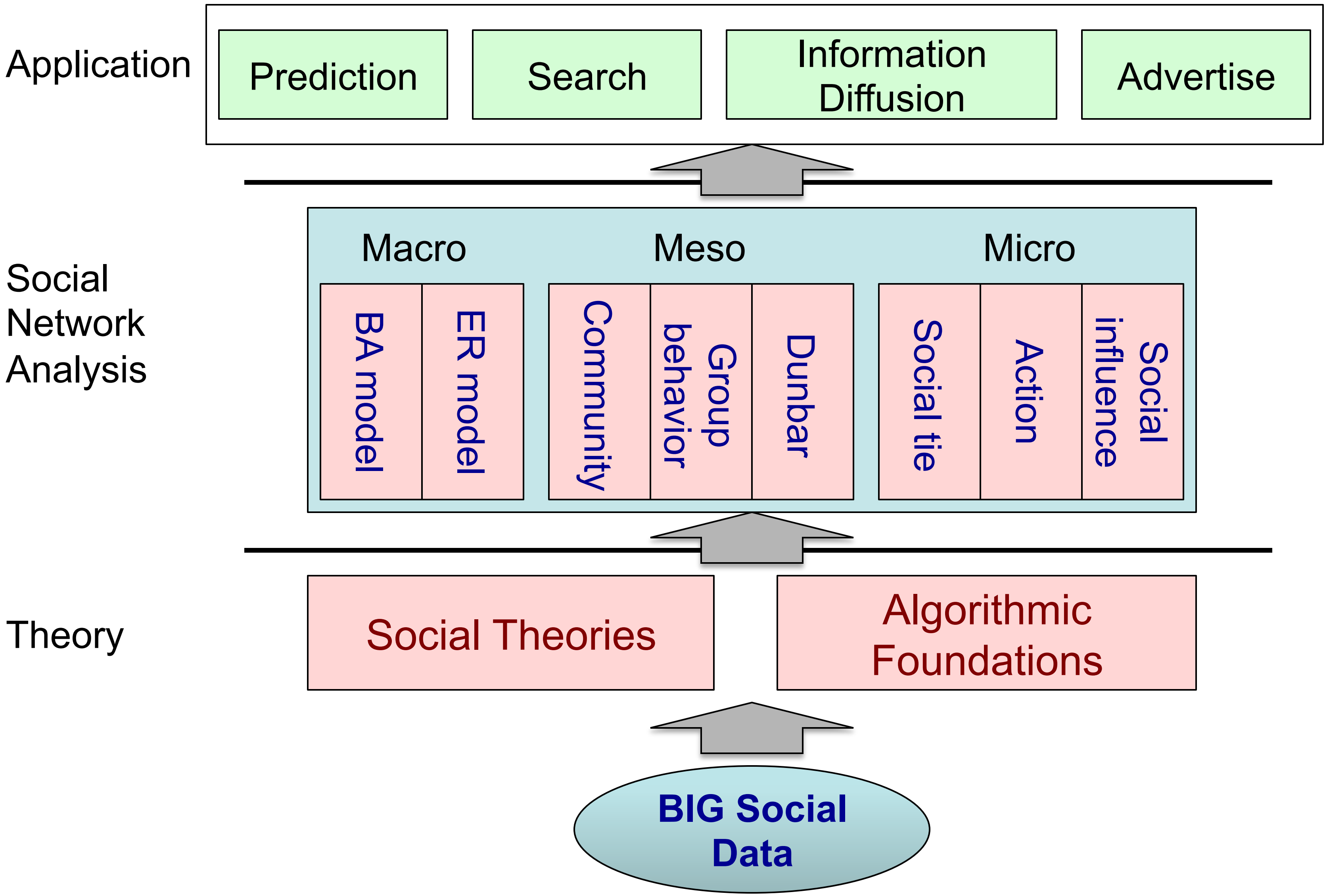
$$\frac{dS}{dt} = -\beta SI + \gamma I$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$

Notice for both SIR and SIS, it holds:

$$\frac{dS}{dt} + \frac{dI}{dt} = 0 \Rightarrow S(t) + I(t) = N$$

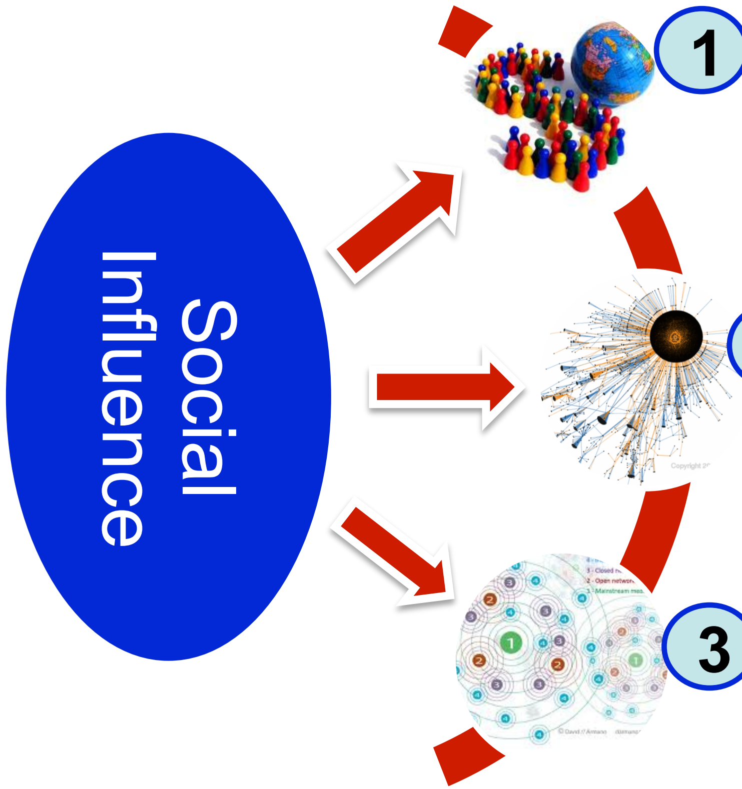
*where N is the **fixed** total population.*

Core Research in Social Network



Part B: Social Influence Analysis

Agenda



1 Test

- Randomization test
- Shuffle test
- Reverse test

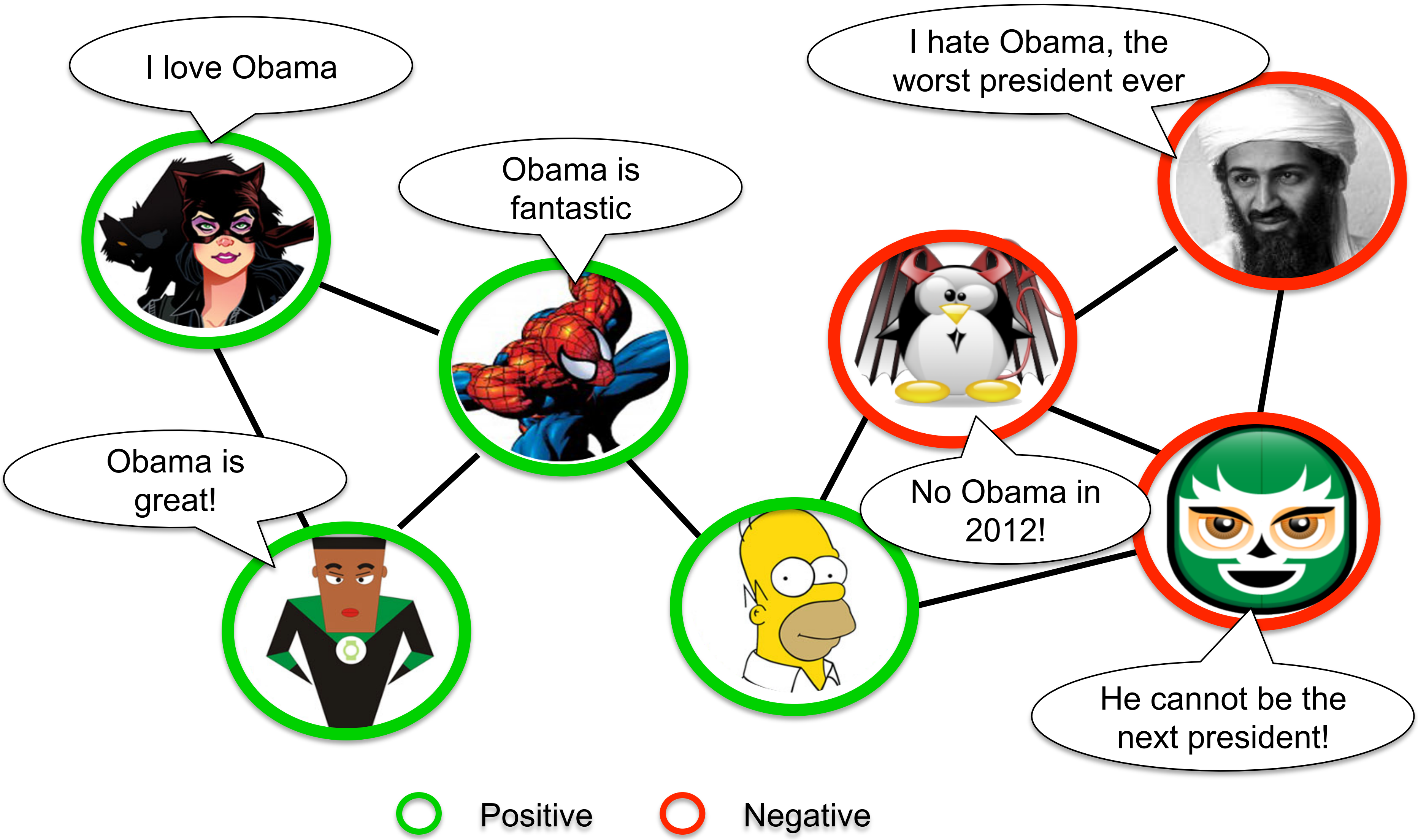
2 Measure

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

3 Models

- Linear Threshold Model
- Cascade Model
- Algorithms

“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

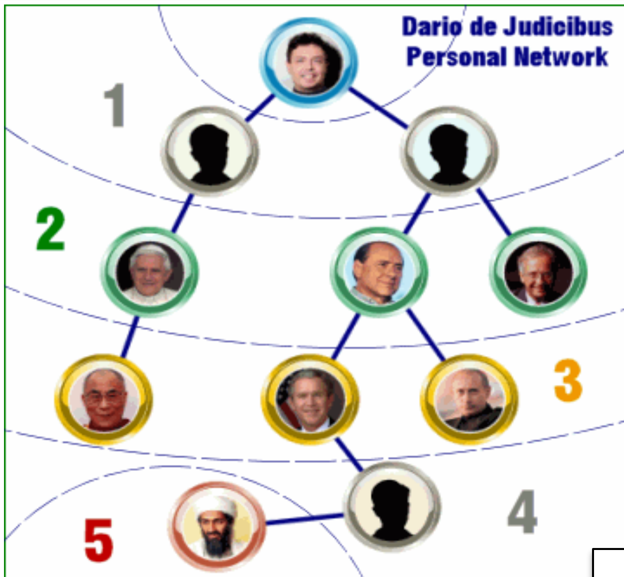
What is Social Influence?

- 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.

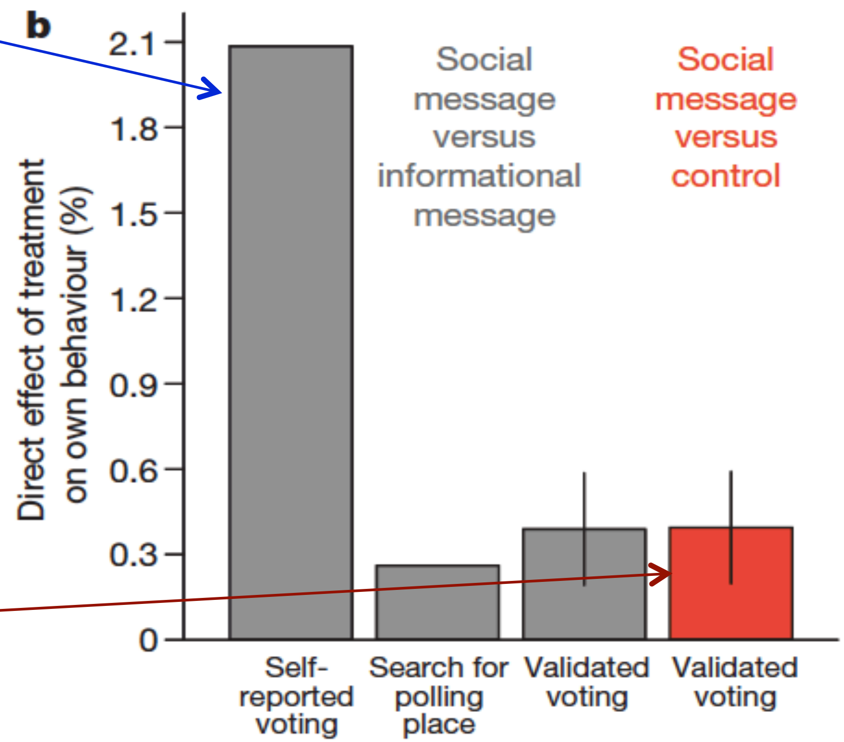
Case 1: Social Influence and Political Mobilization

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

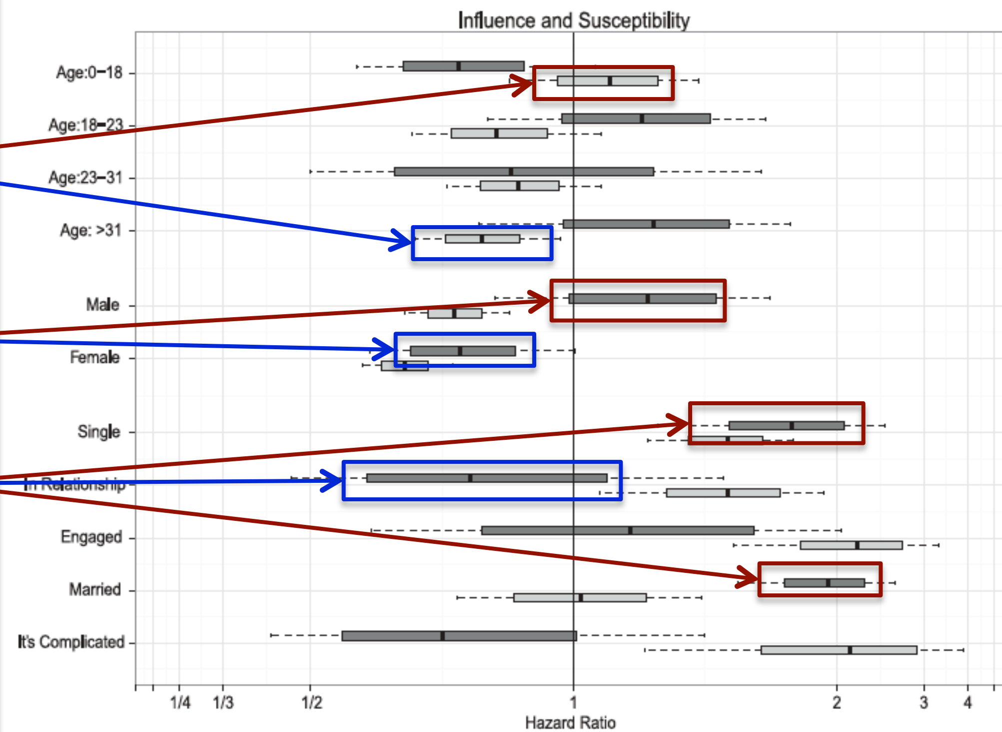
[2] Why I Deleted My Klout Profile, by Pam Moore, at Social Media Today, originally published November 19, 2011; retrieved November 26 2011

Case 3: Influential verse Susceptible[1]

- Study of product adoption for 1.3M FB users

Results:

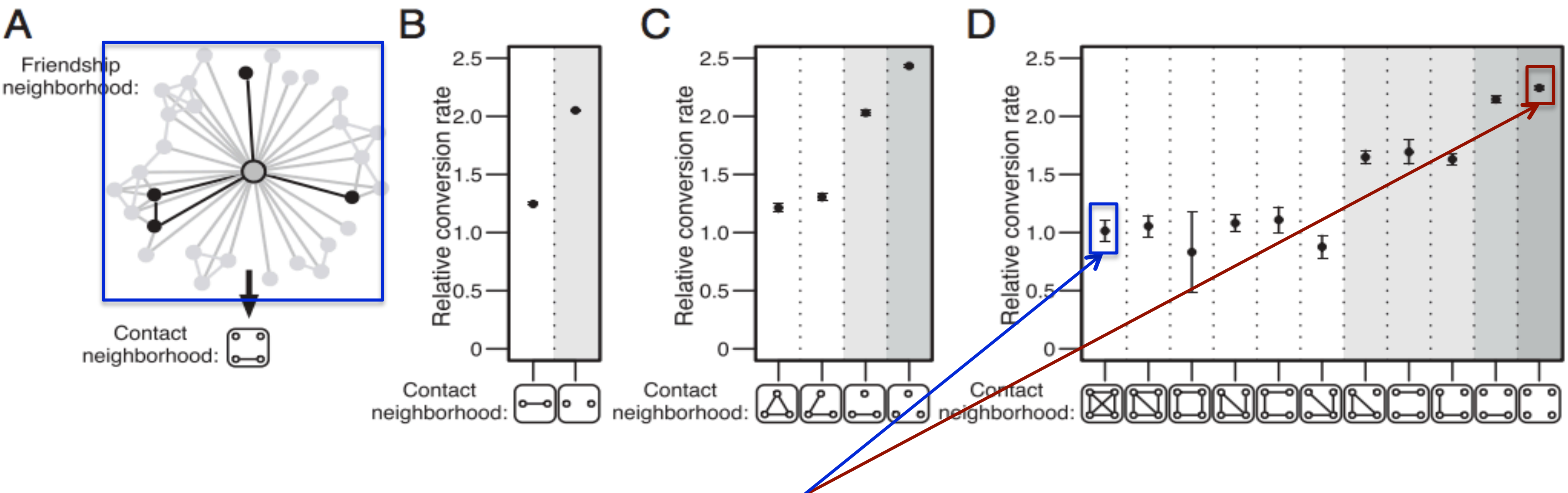
- Younger users are more (18%, *P*<0.05) susceptible to influence than older users
- Men are more (49%, *P*<0.05) influential than women
- Single and Married individuals are significantly more (>100%, *P*<0.05) influential than those who are in a relationship
- Married individuals are the least susceptible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

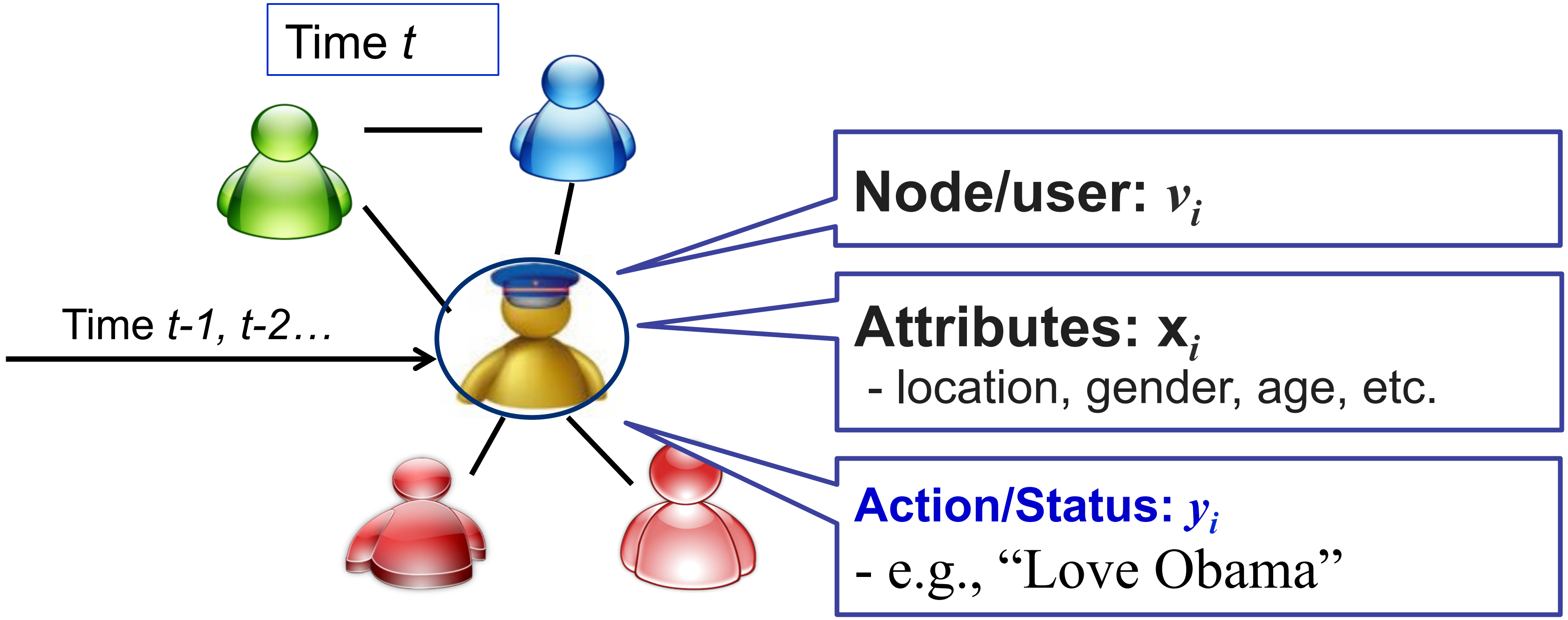
[1] J. Ugandera, L. Backstromb, C. Marlowb, and J. Kleinberg. Structural diversity in social contagion. PNAS, 109 (20): 7591-7592, 2012.

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]

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

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

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

- 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

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

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

sample mean

$$t = \frac{\boxed{\overline{x}_1} - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}}, \quad s_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\boxed{s_1}^2}{\boxed{n_1}} + \frac{s_2^{\ 2}}{\boxed{n_2}}}$$

Unbiased estimator of sample variance

#participants in the control group

#participants in the treatment group

  - 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.

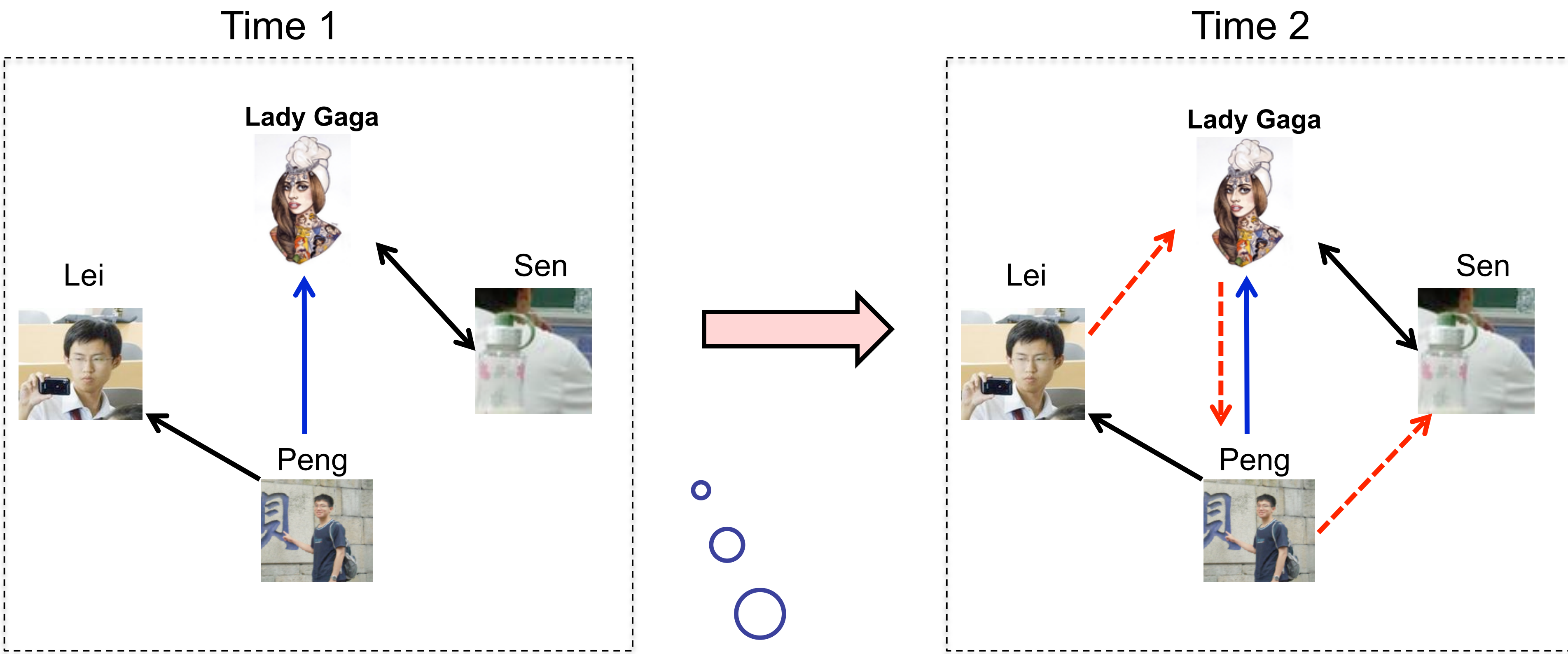
Data Sets

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.

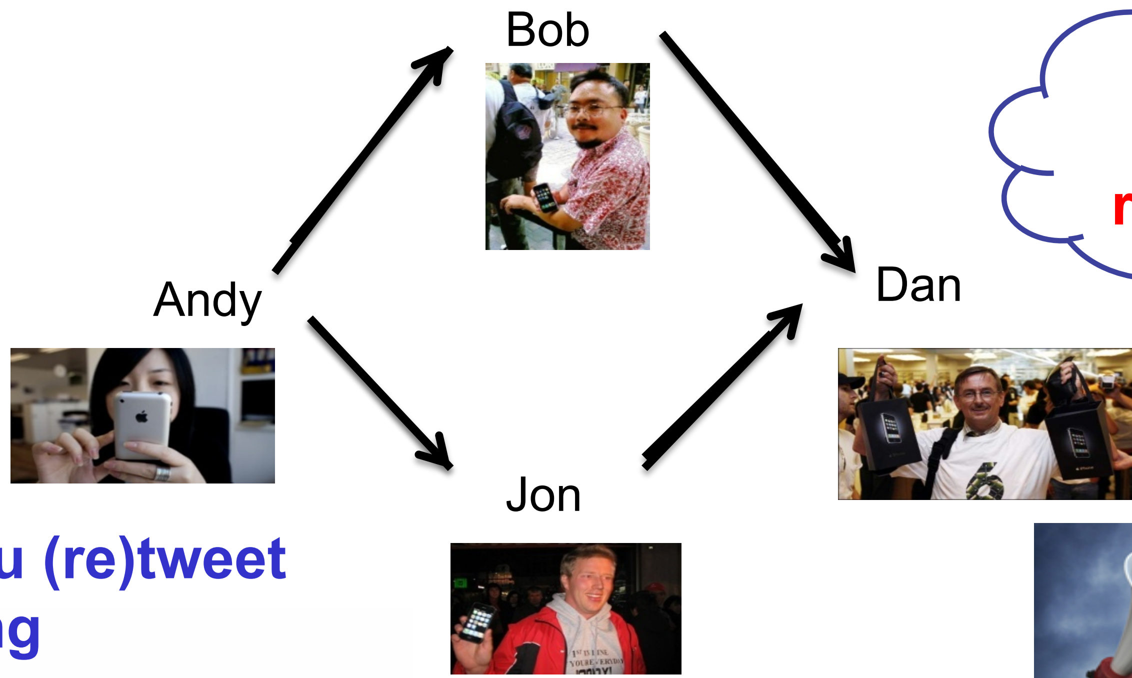
Case 1: Following Influence on Twitter



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

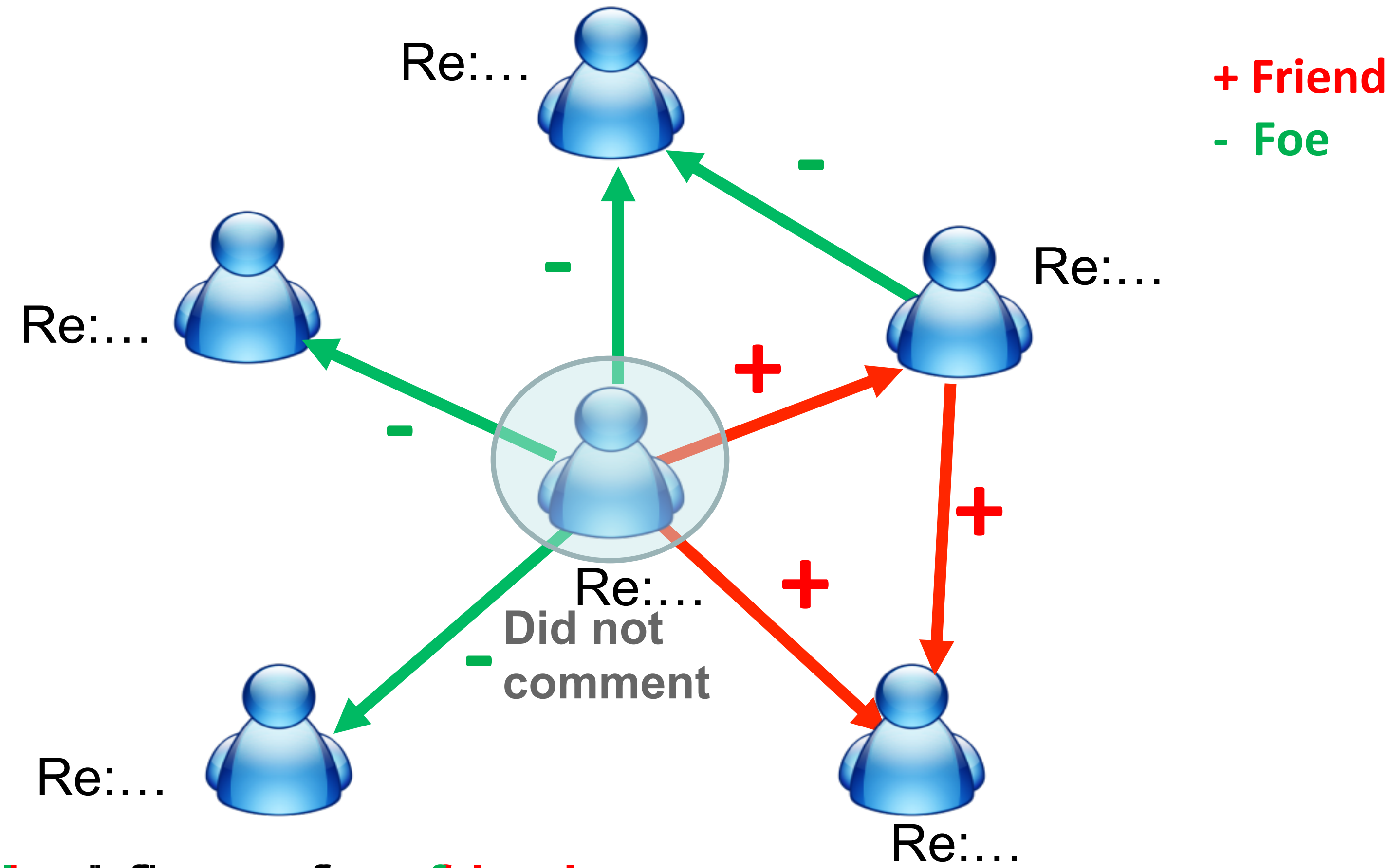


Who will follow to retweet it?

When you (re)tweet something

Case 3: Commenting Influence

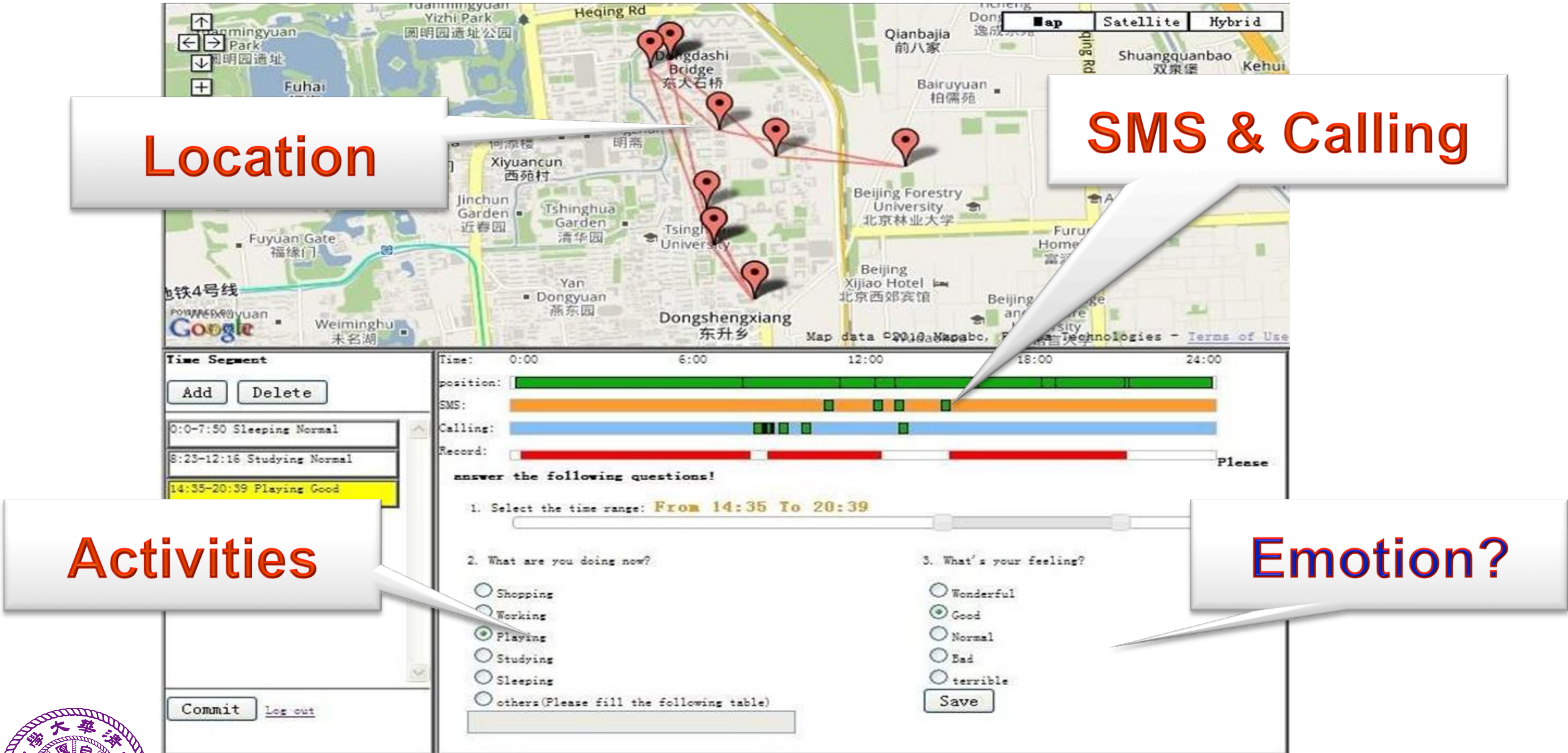
News: Government Wants Private Data / Alan Cox Exits Intel.



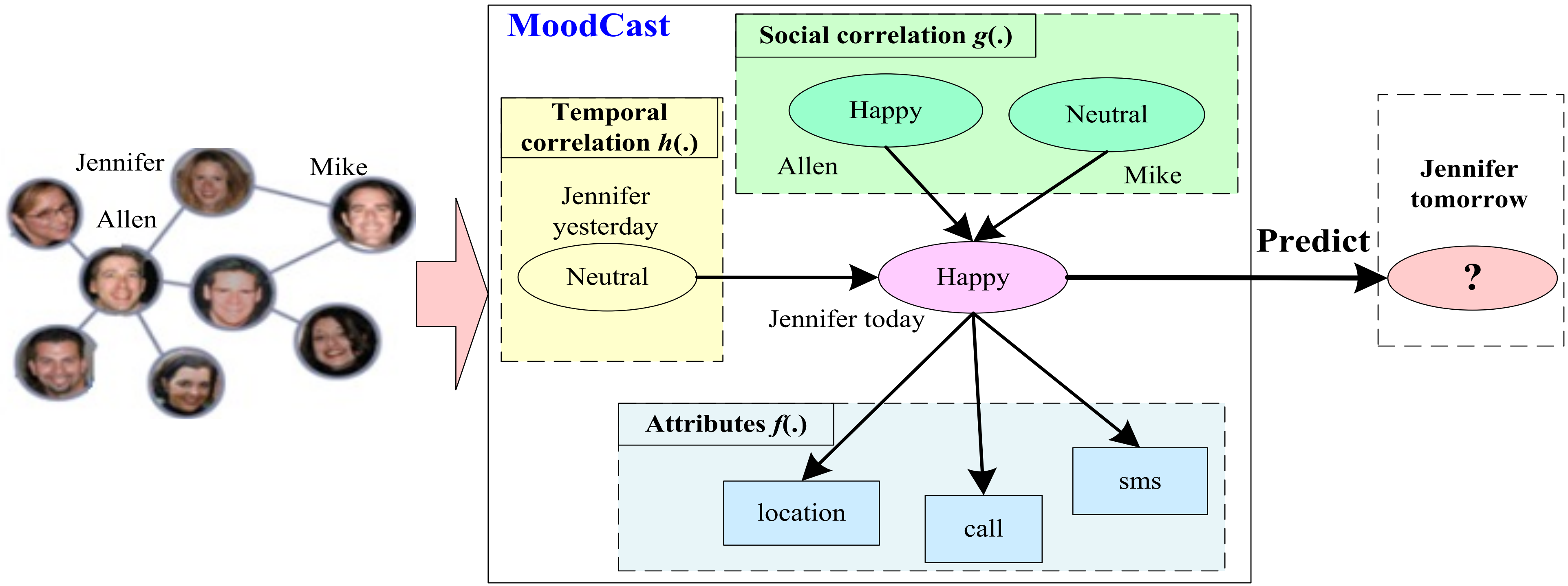
Slashdot

positive influence from friends / negative influence from foes

Case 4: Emotion Influence

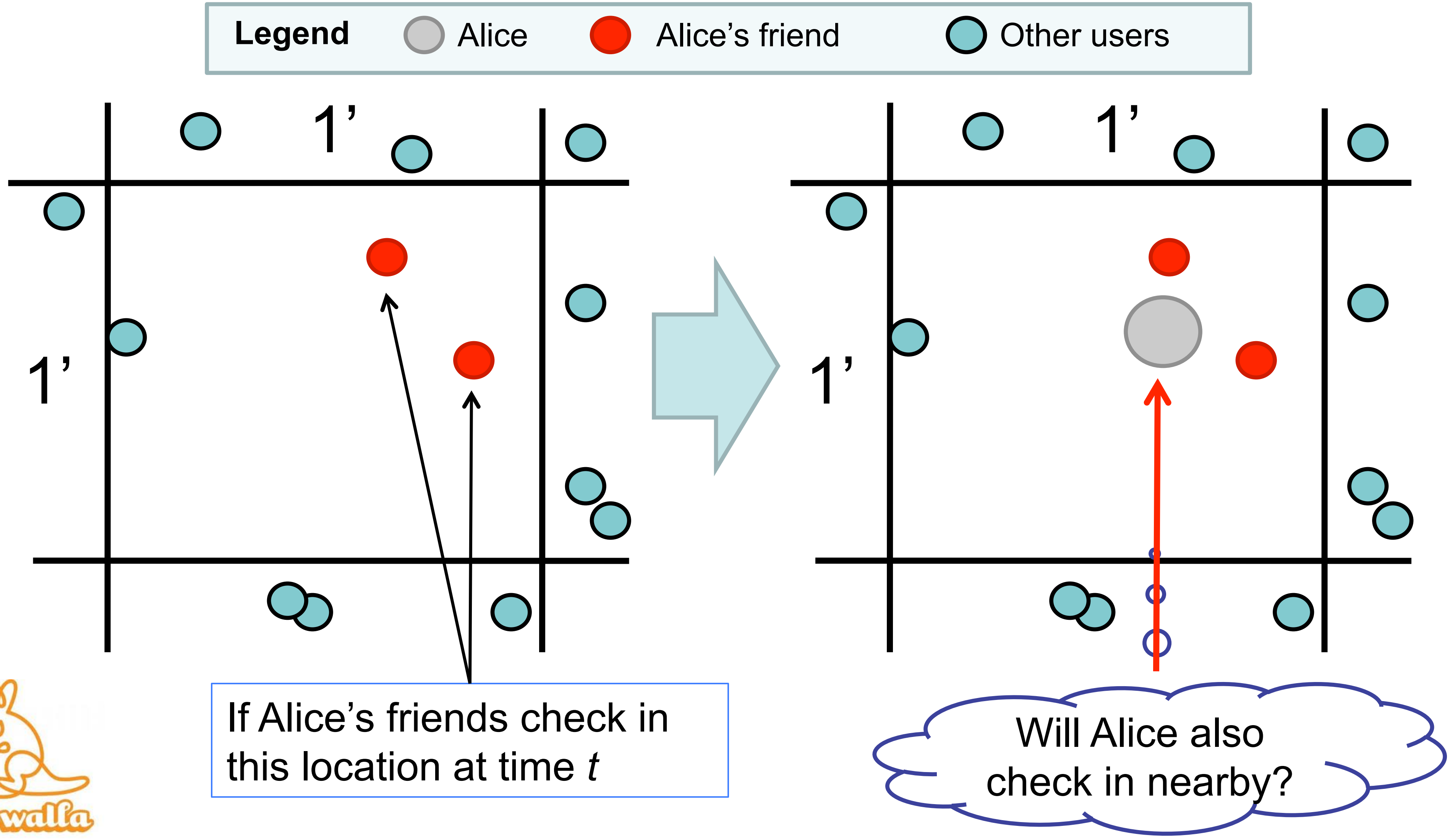


Case 4: Emotion Influence (cont.)

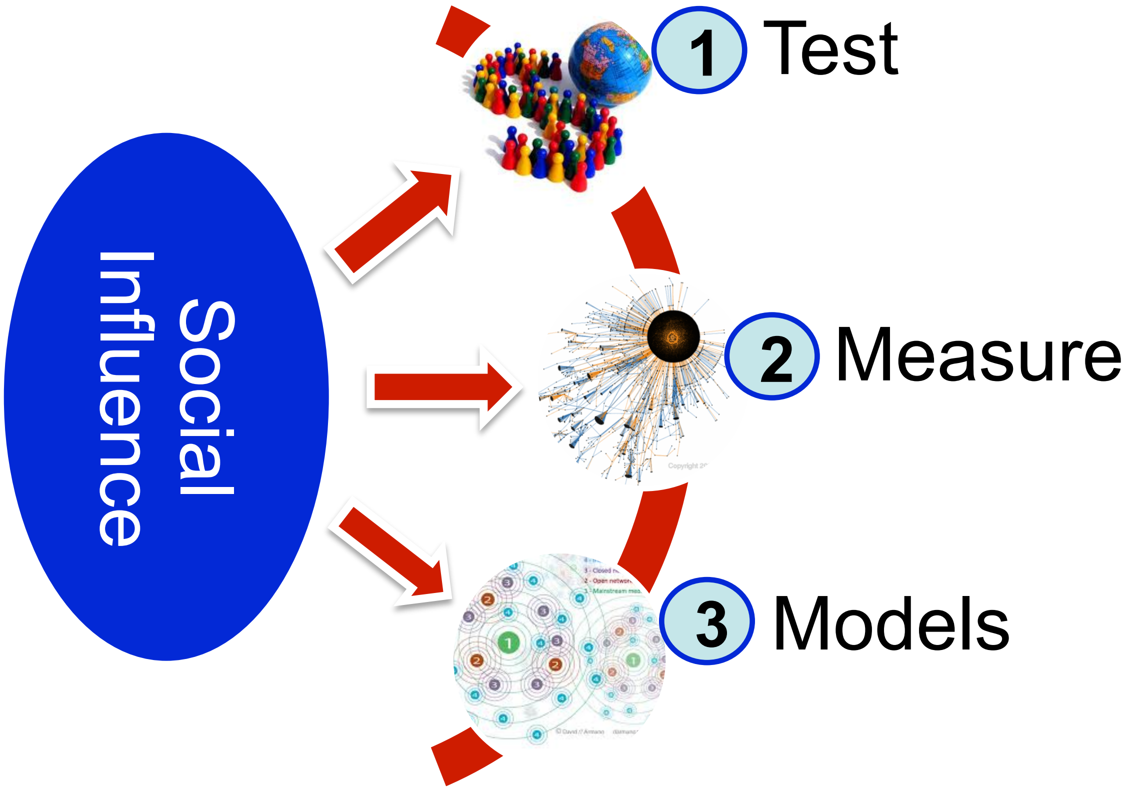


Can we predict users' emotion?

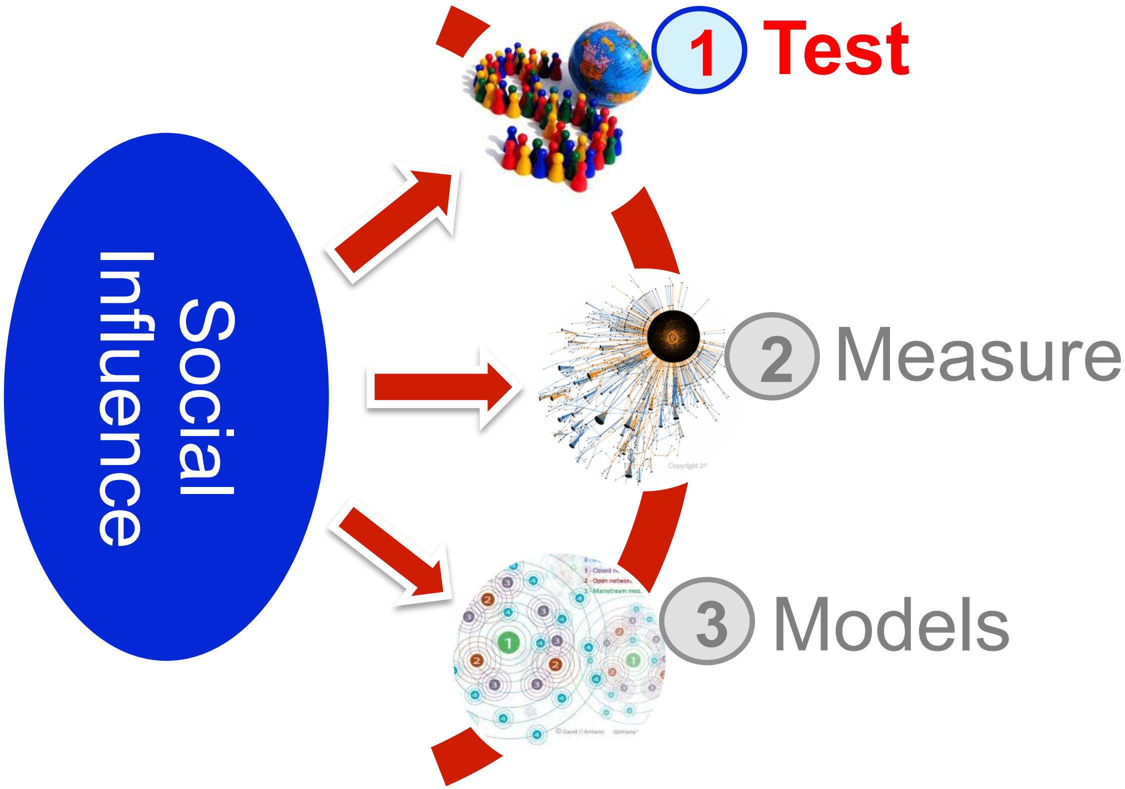
Case 5: Check-in Influence in Gowalla



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.

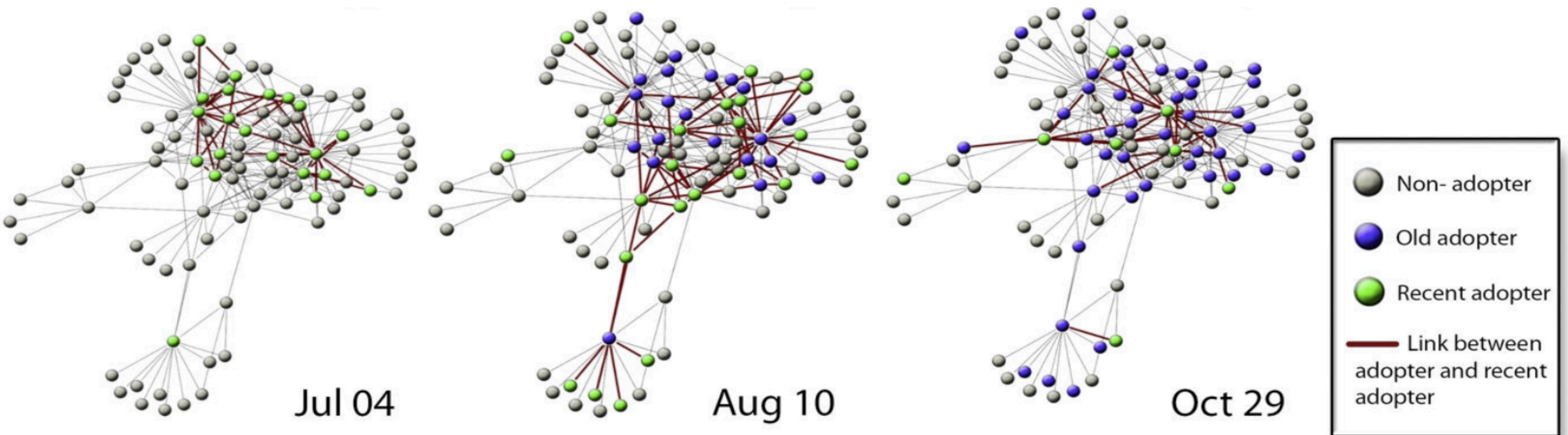
Treatment for Group 2

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

[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*.

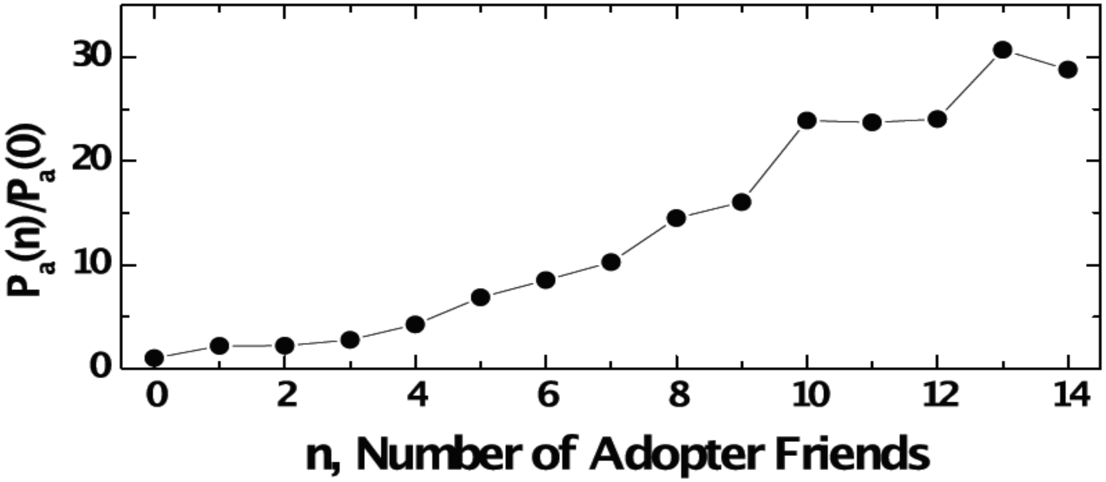
[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS, 106 (51):21544-21549, 2009.

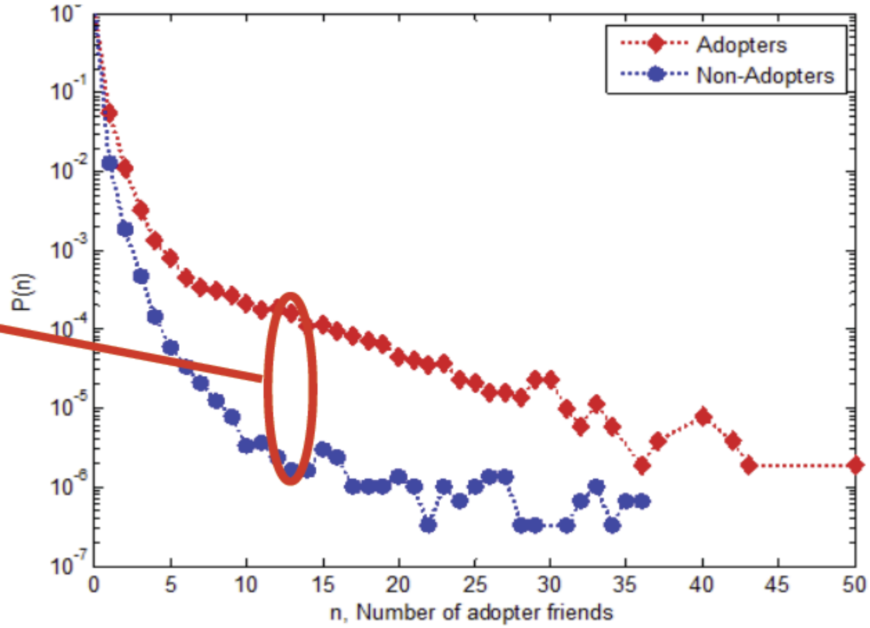
For an example

- 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

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

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

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

The new RCT:

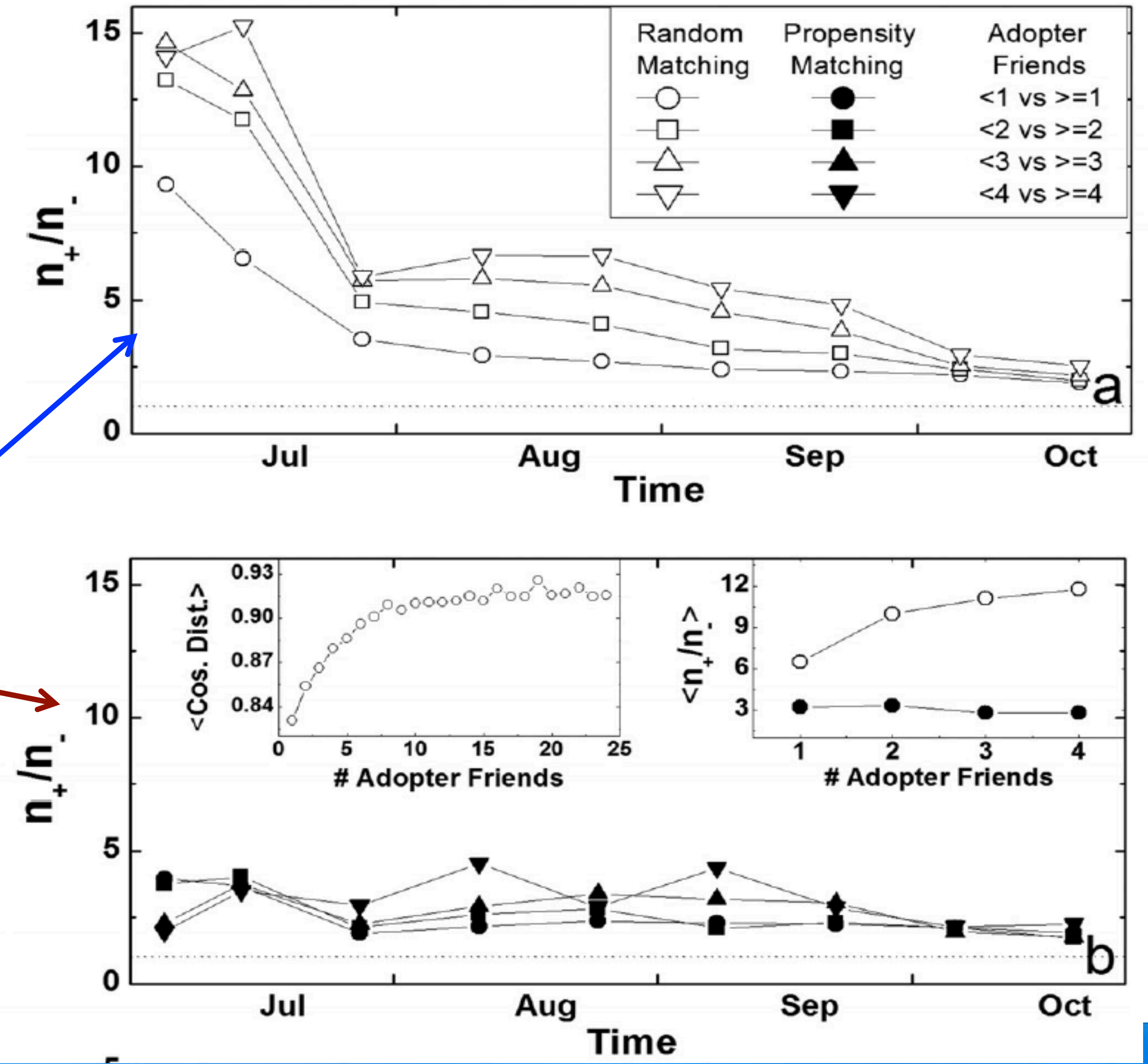
- **Treatment group:** a user *i* who have *k* friends have adopted the Y! Go at time *t*;
- **Controlled group:** a matched user *j* who do not have *k* friends adopt Y! Go at time *t*, but is very likely to have *k* friends to adopt Y!Go at time *t*, i.e., $|p_{it} - p_{jt}| < \sigma$

Results—Random sampling and Matched sampling

The fraction of observed treated to untreated adopters (n_+/n_-) under:

(a) Random sampling;

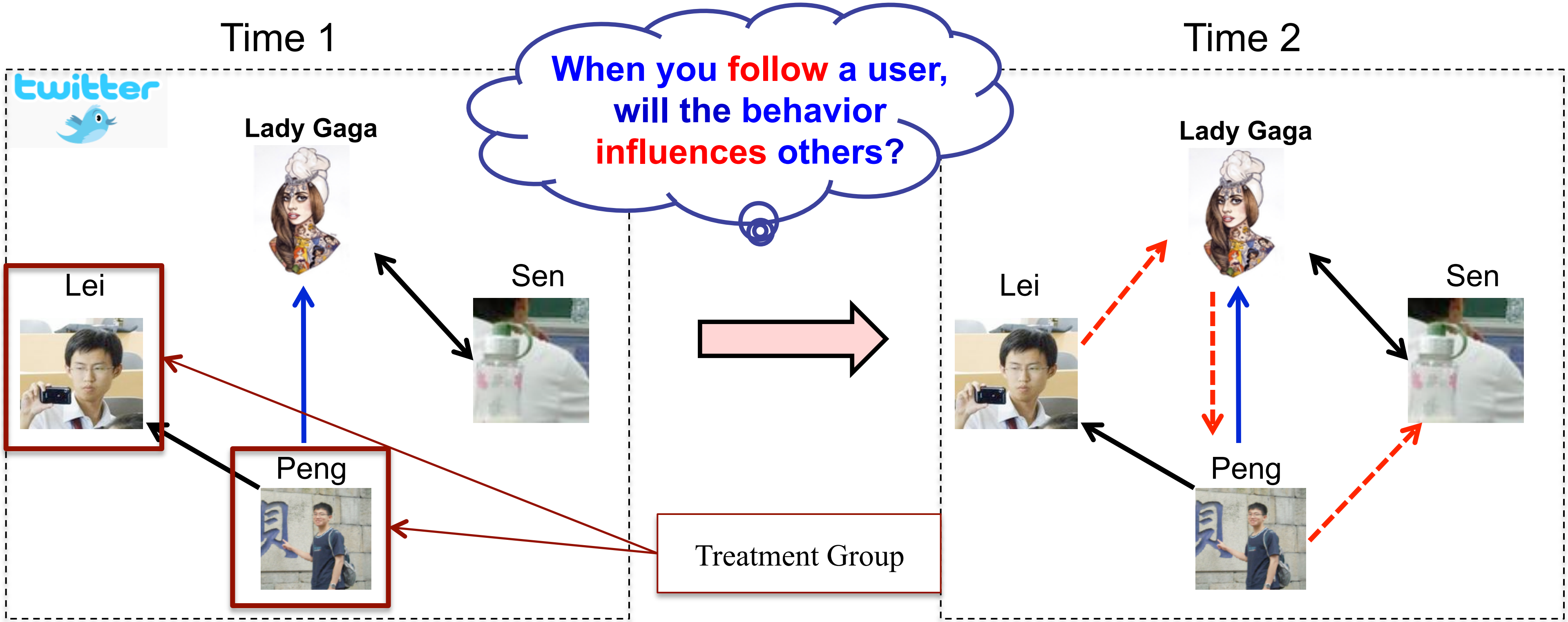
(b) Matched sampling.



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



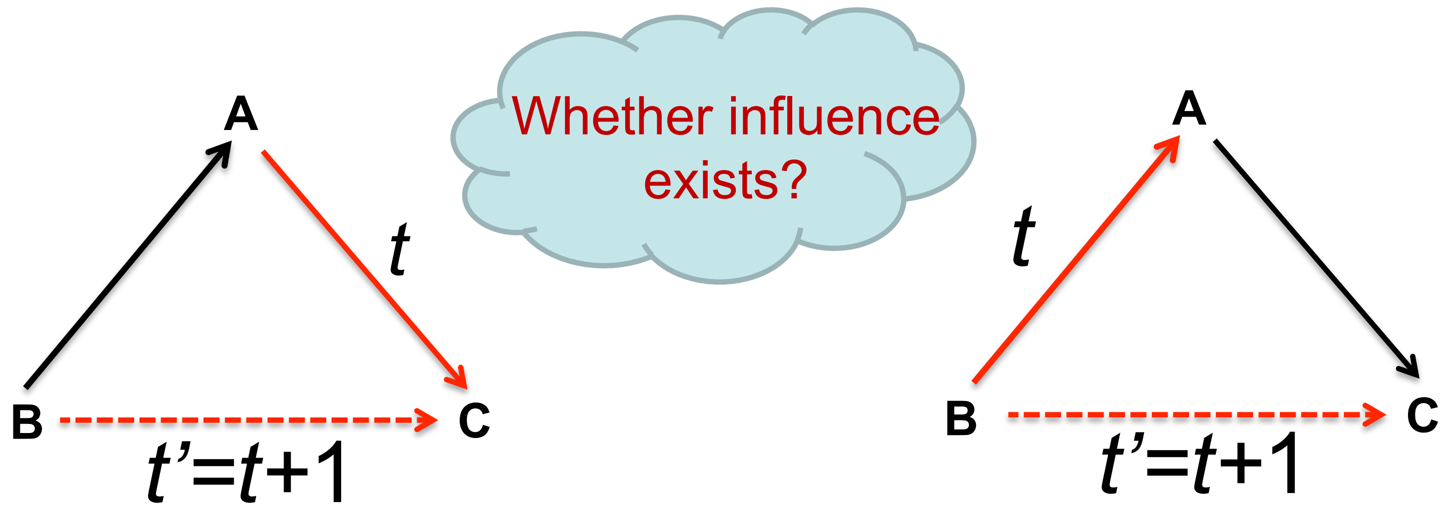
RCT:

- **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



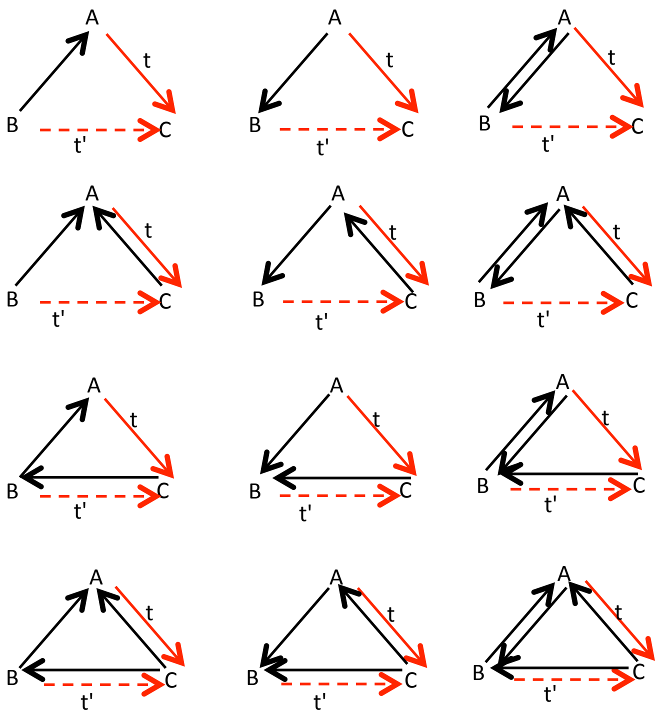
Follower diffusion

Followee diffusion

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

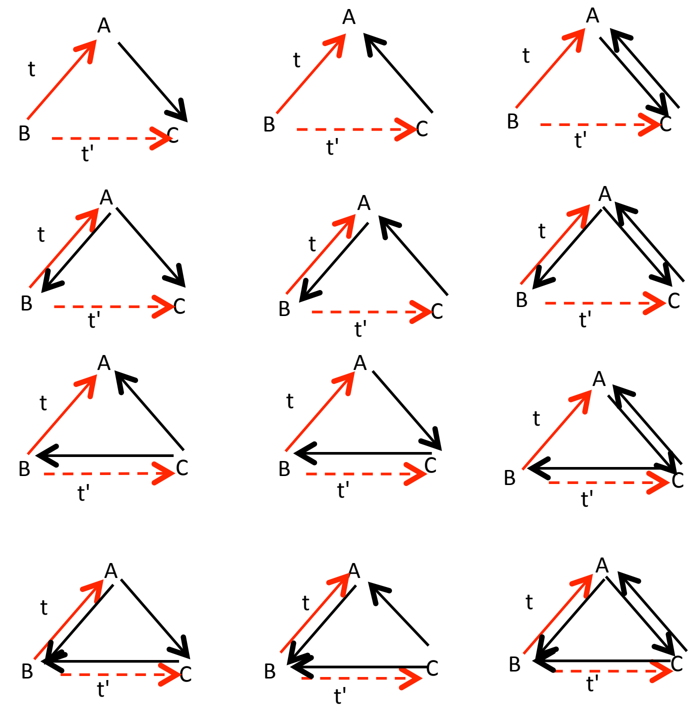
24 Triads in Following Influence

Follower diffusion



12 triads

Followee diffusion



12 triads

Twitter Data

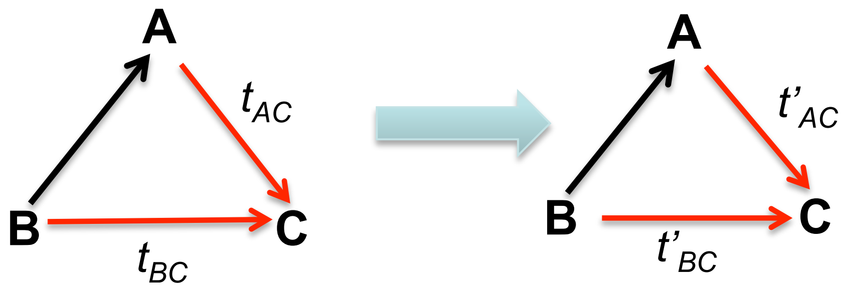
- Twitter data
  - “Lady Gaga” -> 10K followers -> millions of followers;
  - 13,442,659 users and 56,893,234 following links.
  - 35,746,366 tweets.

- 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

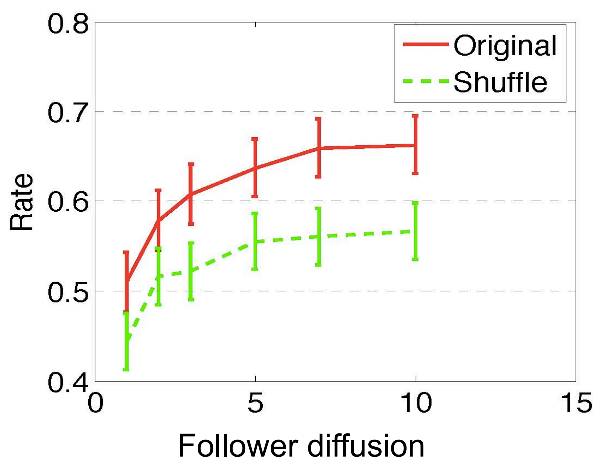


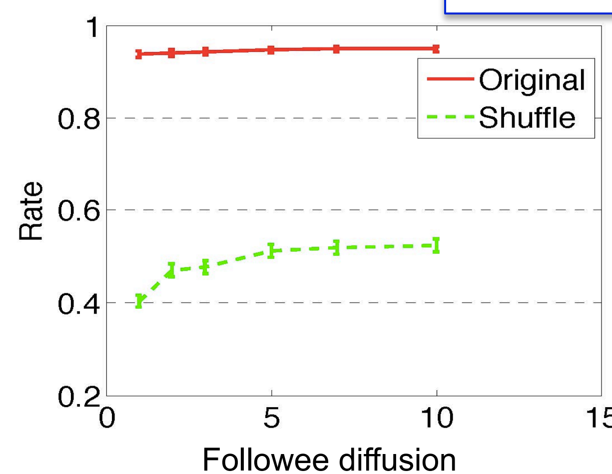
- Compare the rate under the original and shuffled dataset.

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

t-test, P<0.01

- Result

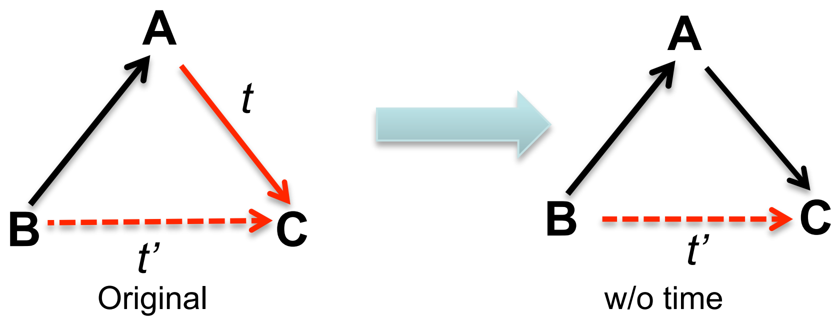




[1] A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks. In KDD, pages 7-15, 2008.

Test 2: Influence Decay Test

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



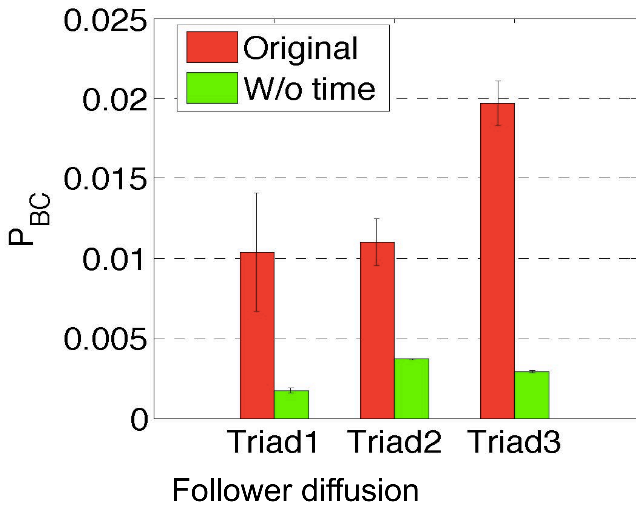
Shuffle test

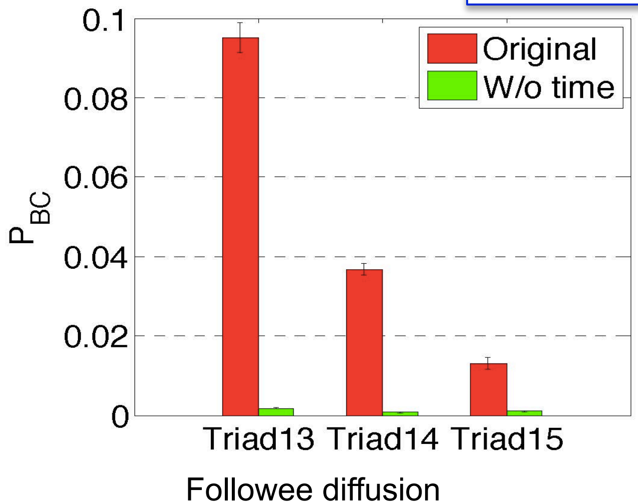
  - Compare the probability of B following C under the original and w/o time dataset.

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

t-test, *P*<0.01

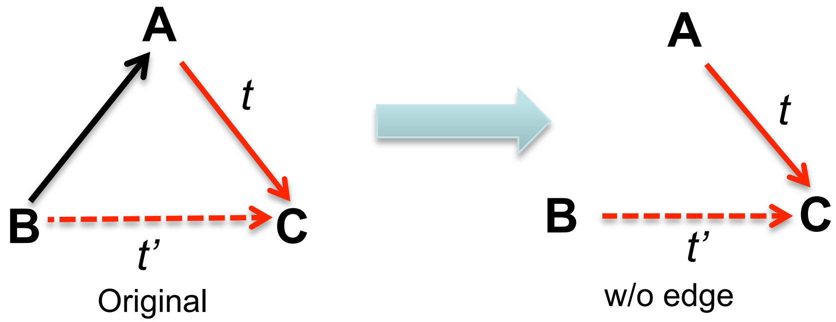
- Result





Test 3: Influence Propagation Test

- Method: Remove the relationship between A and B.



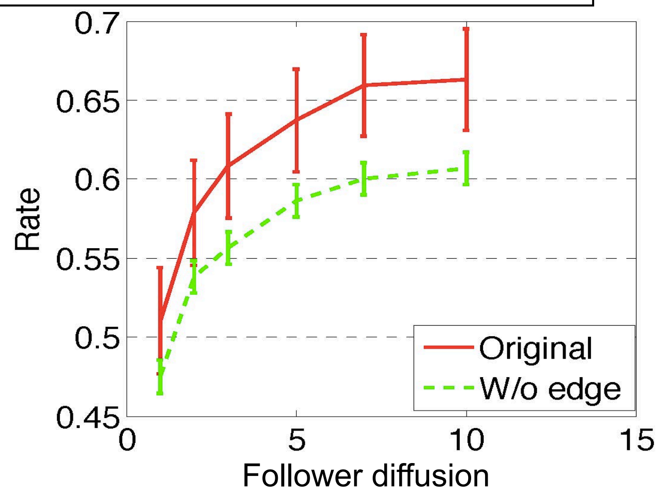
Reverse test

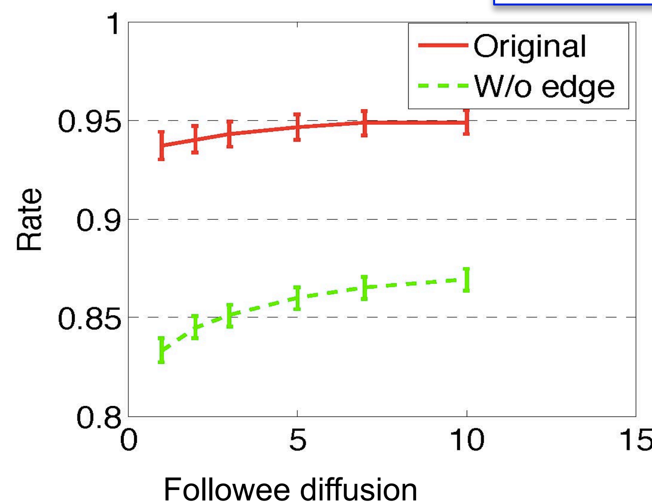
  - Compare the rate under the original and w/o edge dataset.

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

t-test, *P*<0.01

- Result

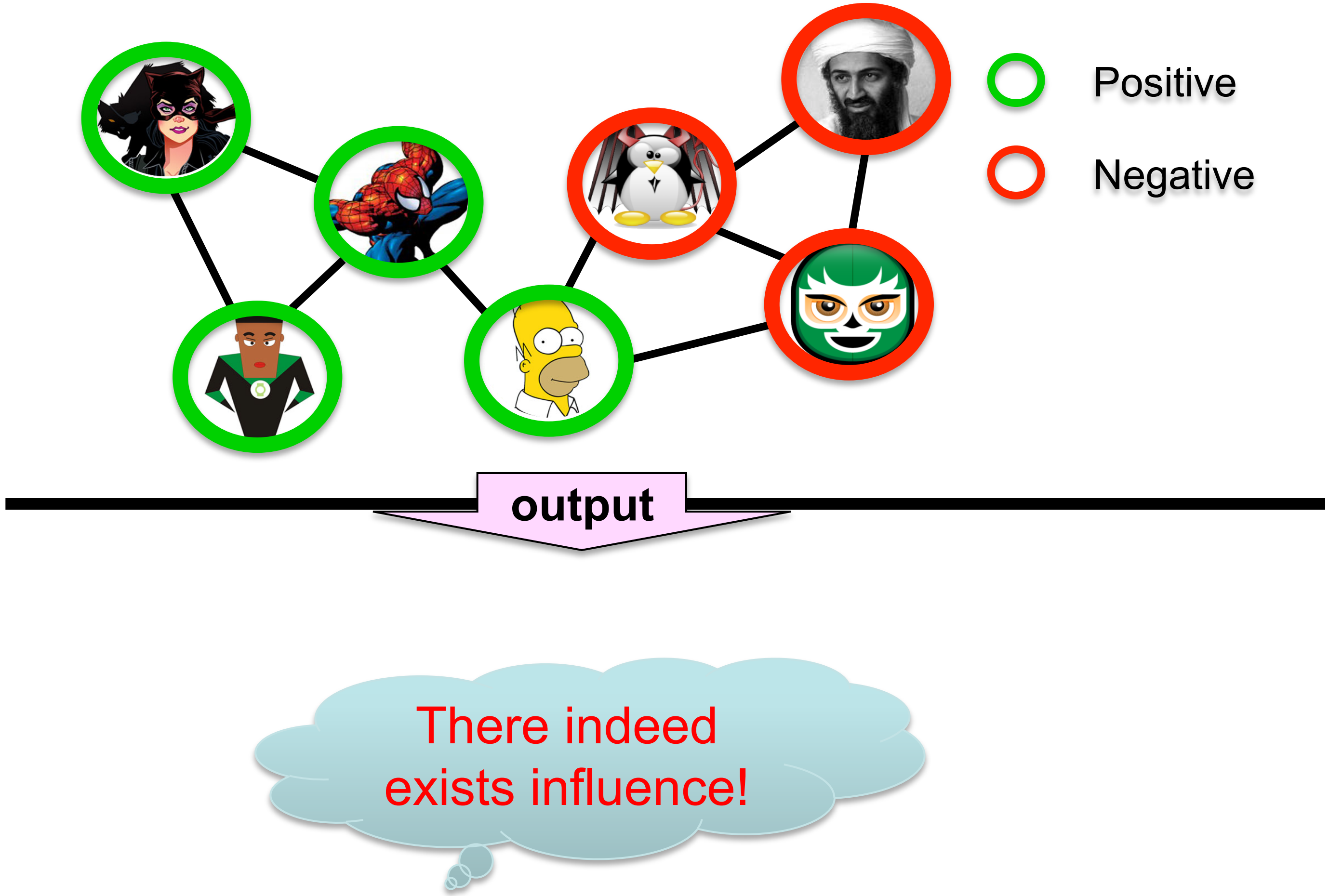




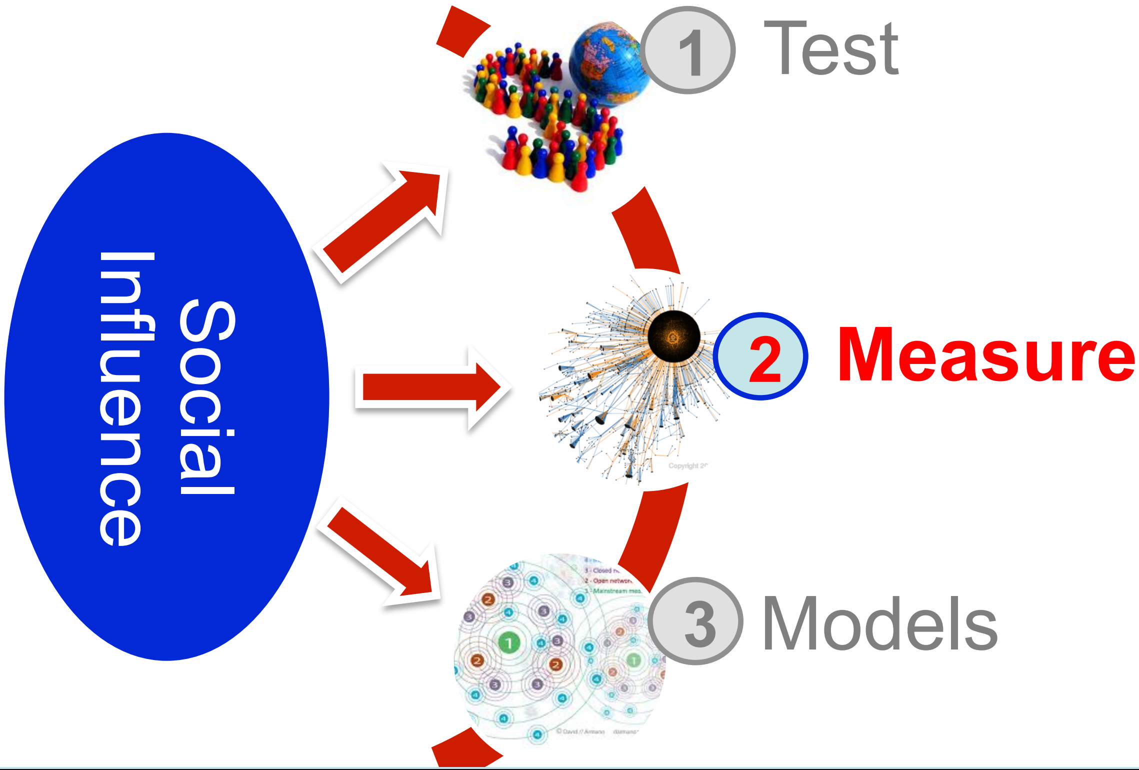
Summary

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

Output of Influence Test



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

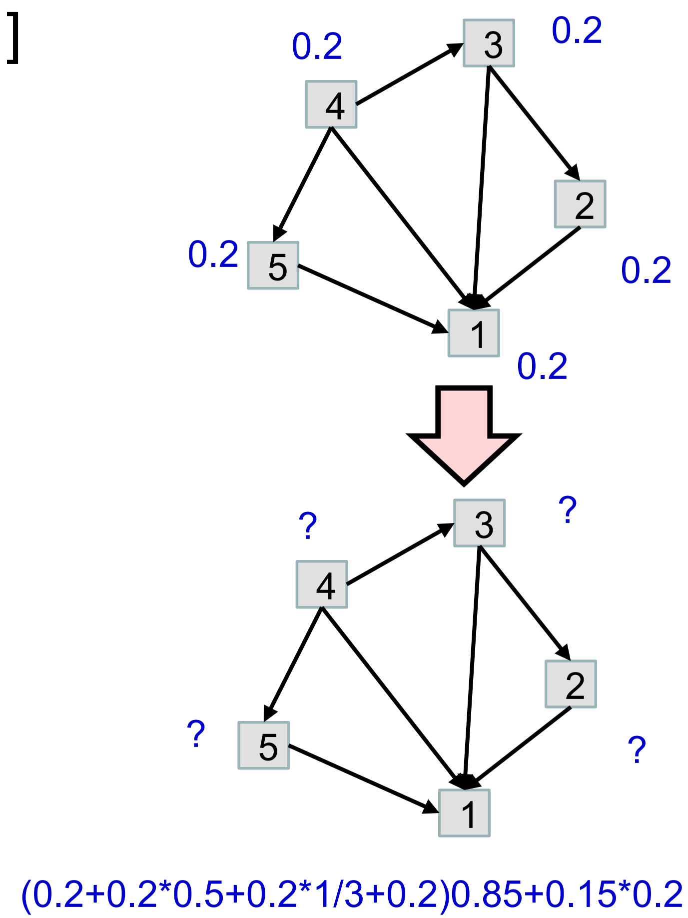
- Let us begin with PageRank[1]

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

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

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

$$\alpha = 0.15$$



[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.

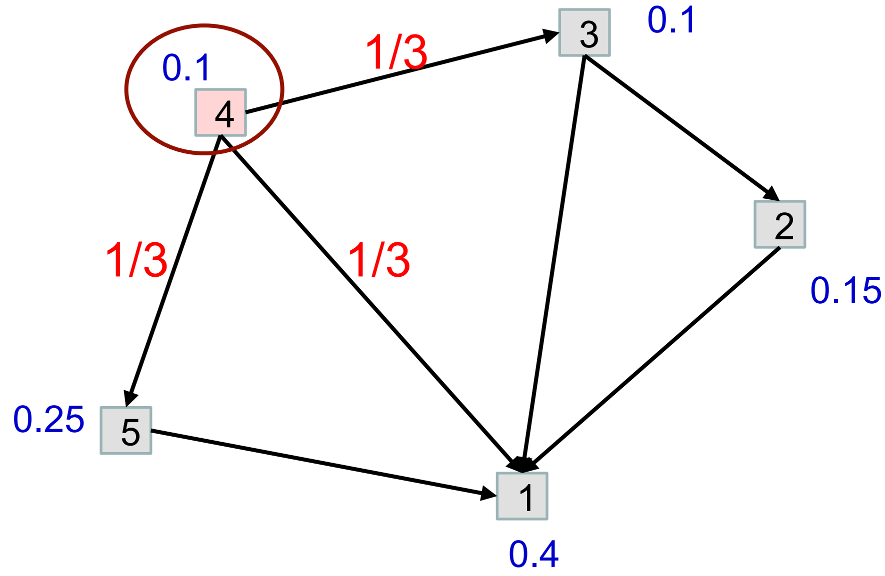
Random Walk Interpretation

- Probability distribution

$$P(t) = r$$

- Stationary distribution

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

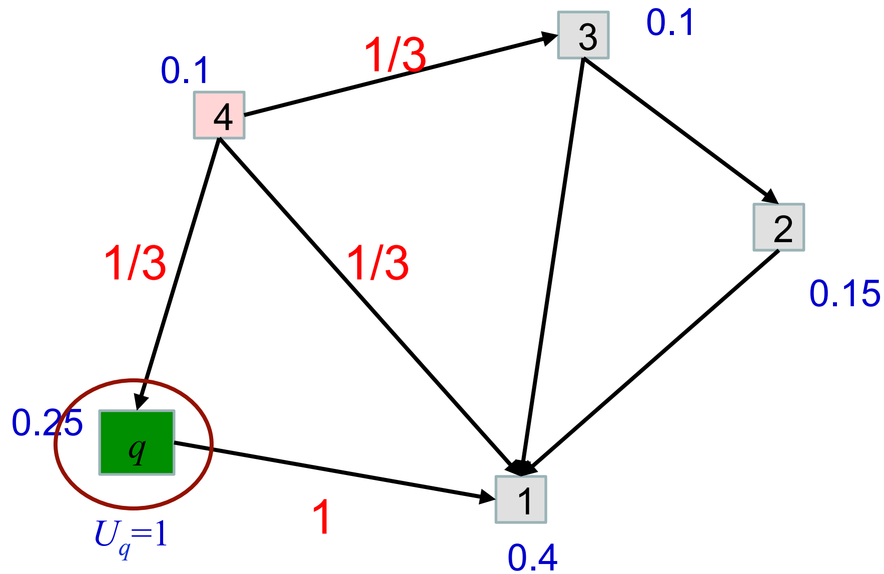


Random Walk with Restart[1]

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

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

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

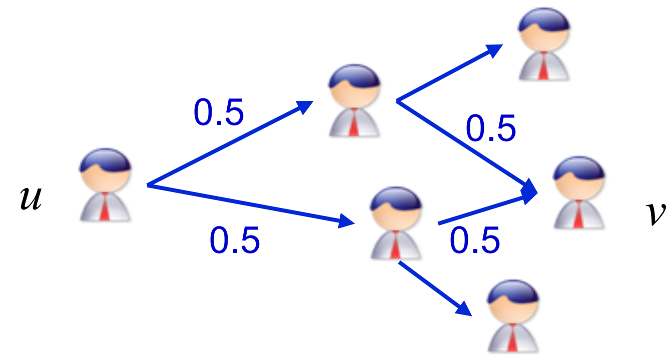


[1] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In ICDM'05, pages 418–425, 2005.

Measure Influence via Reachability[1]

- Influence of a path

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



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

- Influence of user u on v

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

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

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

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

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

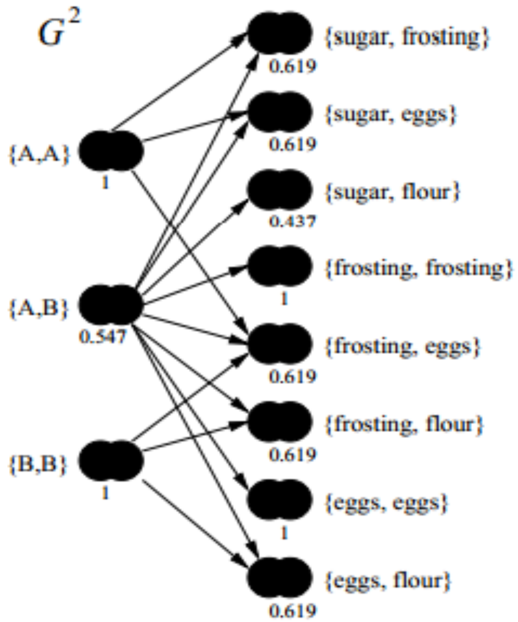
The set of pages which have inks pointing to u

[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.

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

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

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

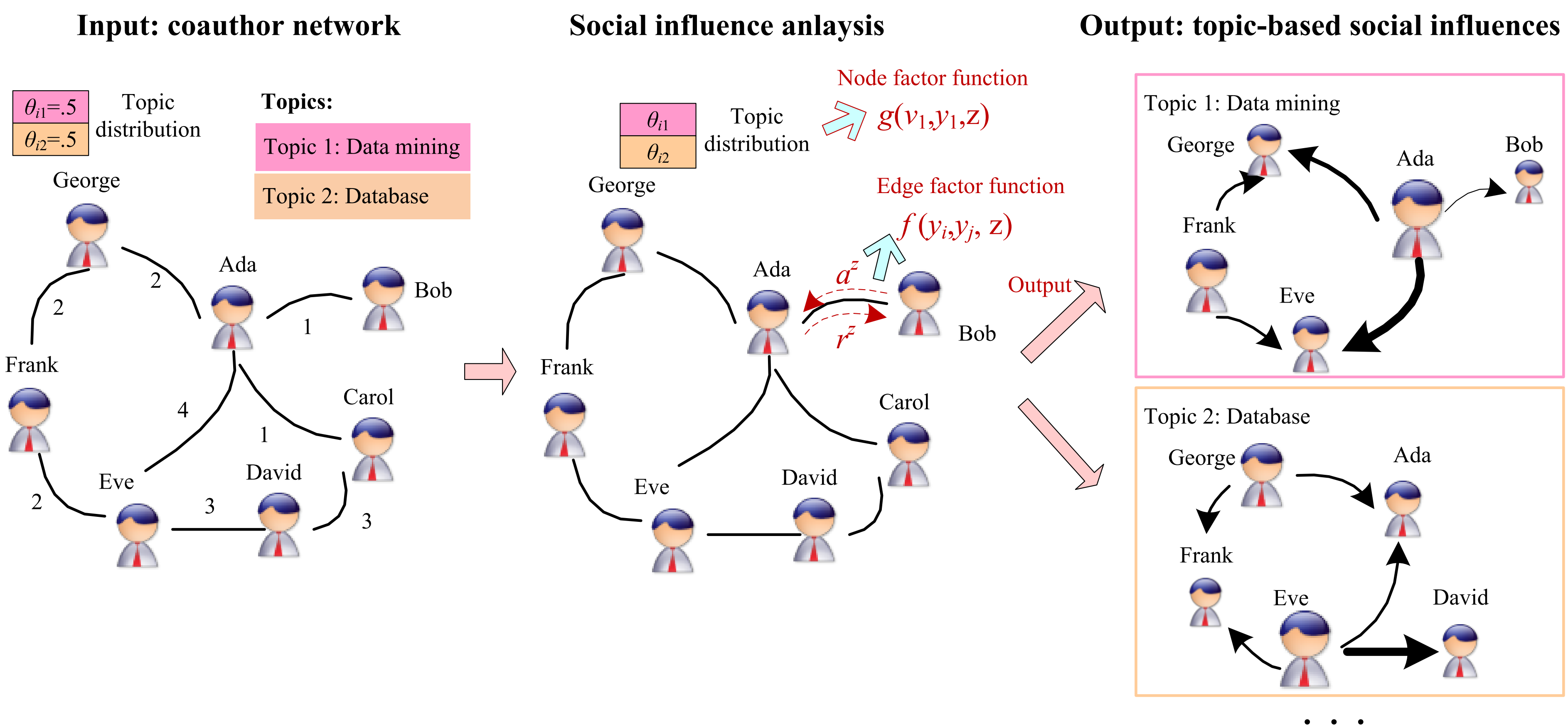
Note: Again, the method only considers the network information.

Methodologies

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

Topic-based Social Influence Analysis

- Social network -> Topical influence network



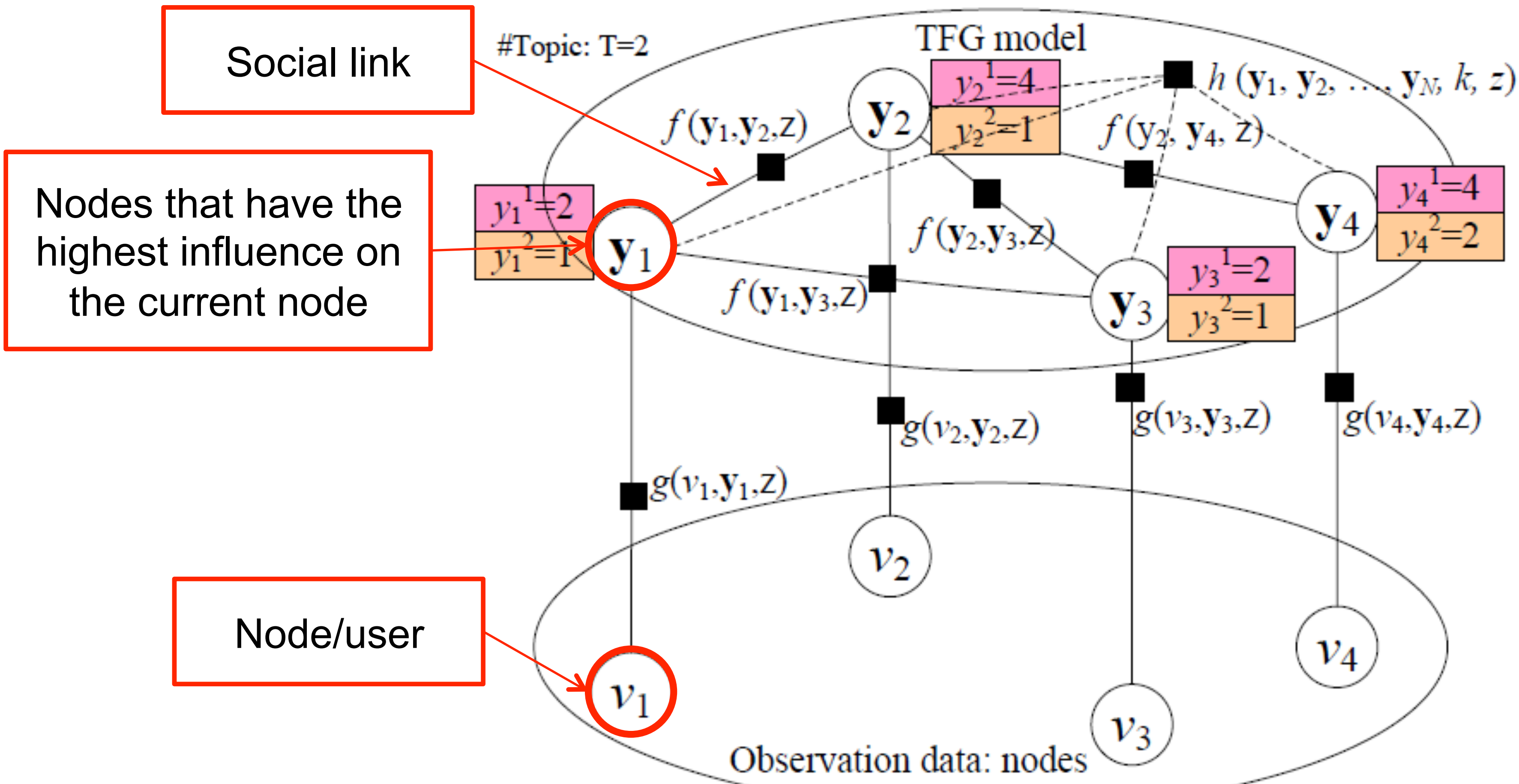
[1] J. Tang, J. Sun, C. Wang, and Z. Yang. Social Influence Analysis in Large-scale Networks. In KDD'09, pages 807-816, 2009.

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:

$$P(\mathbf{v}, \mathbf{Y}) = \frac{1}{Z} \prod_{k=1}^{N} \prod_{z=1}^{T} \boxed{h(\mathbf{y}_1, \ldots, \mathbf{y}_N, k, z)}$$

$$\prod_{i=1}^{N} \prod_{z=1}^{T} \boxed{g(v_i, \mathbf{y}_i, z)} \prod_{e_{kl} \in E} \prod_{z=1}^{T} \boxed{f(\mathbf{y}_k, \mathbf{y}_l, z)}$$

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

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

– 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

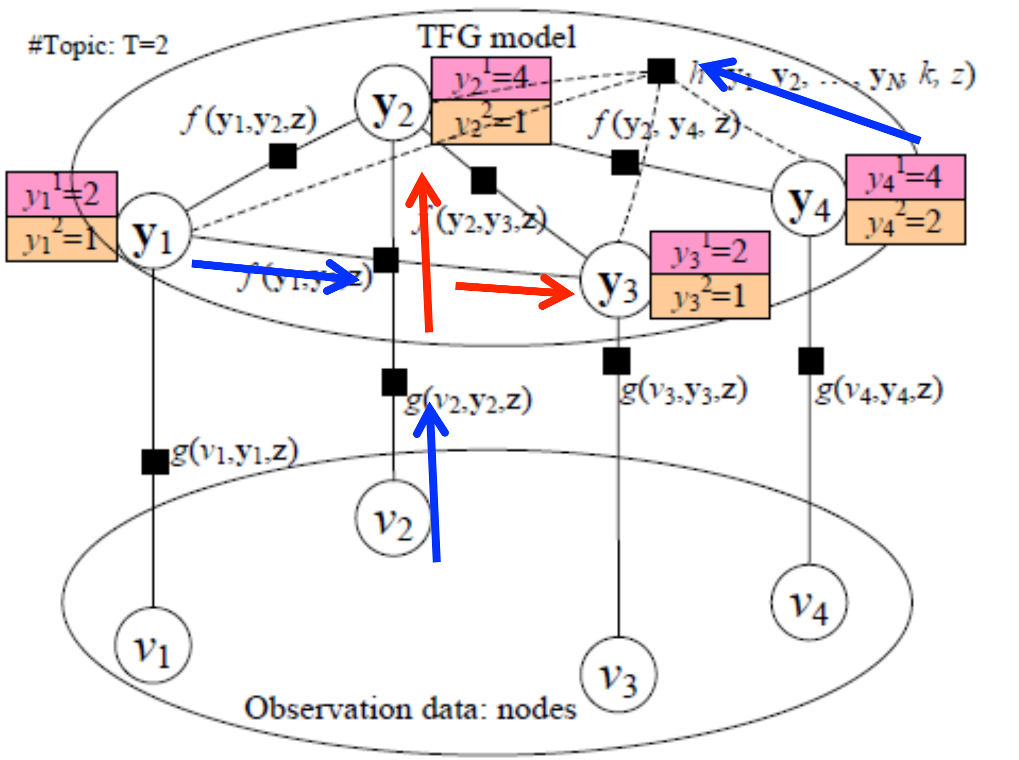
– Global feature function

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

Model Learning Algorithm

Sum-product:

$$
\begin{aligned}
\boxed{m_{y\to f}(y,z)} &= \prod_{f'\sim y\setminus f} m_{f'\to y}(y,z) \prod_{z'\neq z}\prod_{f'\sim y\setminus f} m_{f'\to y}(y,z')^{(\tau_{z'z})} \\
\boxed{m_{f\to y}(y,z)} &= \sum_{\sim\{y\}} \left(f(Y,z) \prod_{y'\sim f\setminus y} m_{y'\to f}(y',z) \right) \\
&+ \sum_{z'\neq z} \tau_{z'z} \sum_{\sim\{y\}} \left(f(Y,z') \prod_{y'\sim f\setminus y} m_{y'\to f}(y',z') \right) \qquad (4)
\end{aligned}
$$



- 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:

$\boxed{m_{ij}}$

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

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

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

[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 $\{G_z = (V_z, E_z)\}^T$

Calculate the node feature function $g(v_i, \mathbf{y}_i, z)$;
Calculate b^z_{ij} according to Eq. 8;

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

Initialize all $\{r^z_{ij}\} \leftarrow 0$;
repeat
    foreach *edge-topic pair* (e_{ij}, z) **do**
        Update r^z_{ij} according to Eq. 5;

$$r^z_{ij} = b^z_{ij} - \max_{k \in NB(j)} \{b^z_{ik} + a^z_{ik}\}$$

    end
    foreach *node-topic pair* (v_j, z) **do**
        Update a^z_{jj} according to Eq. 6;

$$a^z_{jj} = \max_{k \in NB(j)} \min\{r^z_{kj}, 0\}$$

    end
    foreach *edge-topic pair* (e_{ij}, z) **do**
        Update a^z_{ij} according to Eq. 7;

$$a^z_{ij} = \min(\max\{r^z_{jj}, 0\}, -\min\{r^z_{jj}, 0\} - \max_{k \in NB(j) \setminus \{i\}} \min\{r^z_{kj}, 0\}), i \in NB(j)$$

    end
until *convergence*;
foreach *node* v_t **do**
    foreach *neighboring node* $s \in NB(t) \cup \{t\}$ **do**
        Compute μ^z_{st} according to Eq. 9;

$$\mu^z_{st} = \frac{1}{1 + e^{-(r^z_{ts} + a^z_{ts})}}$$

    end
end
Generate $G_z = (V_z, E_z)$ for every topic z according to $\{\mu^z_{st}\}$;

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

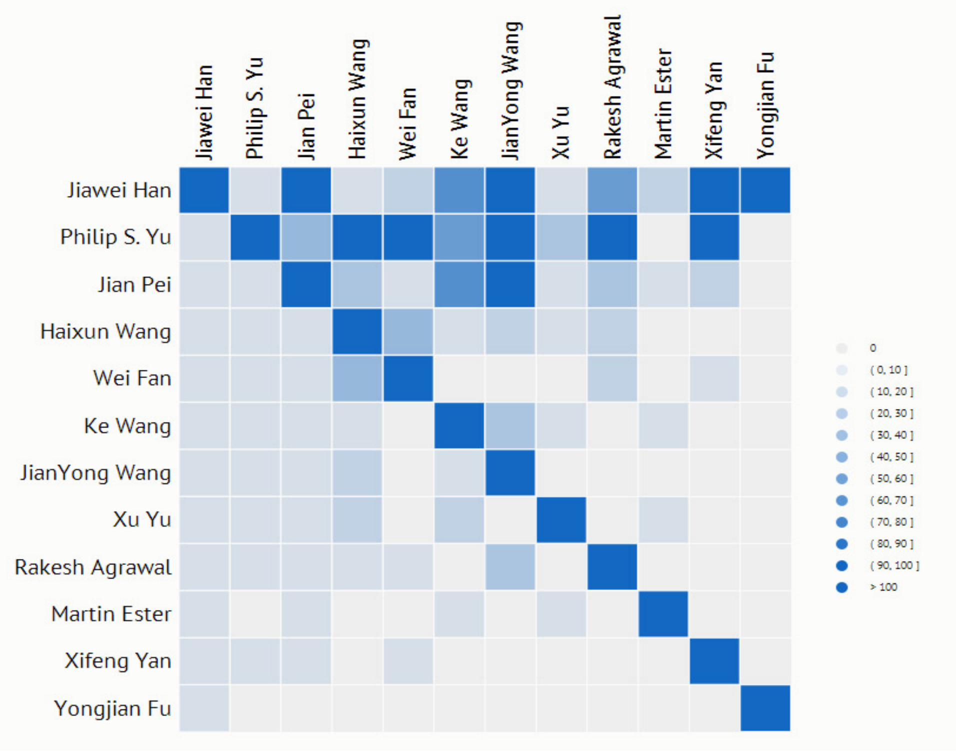
- Evaluation measures
  - CPU time
  - Case study
  - Application

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

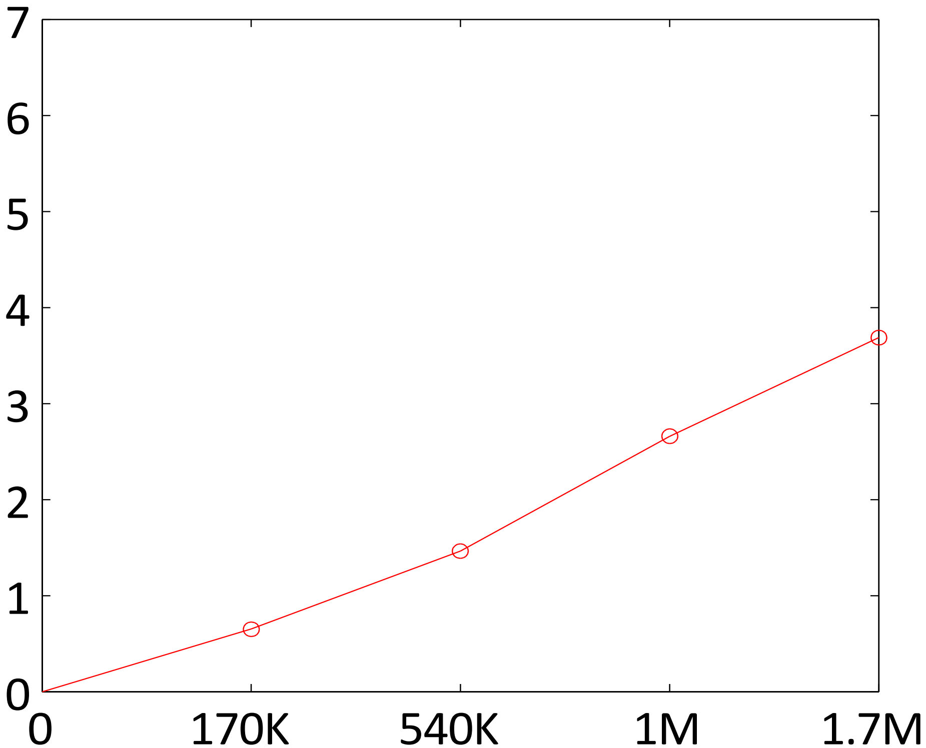
Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell, 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.

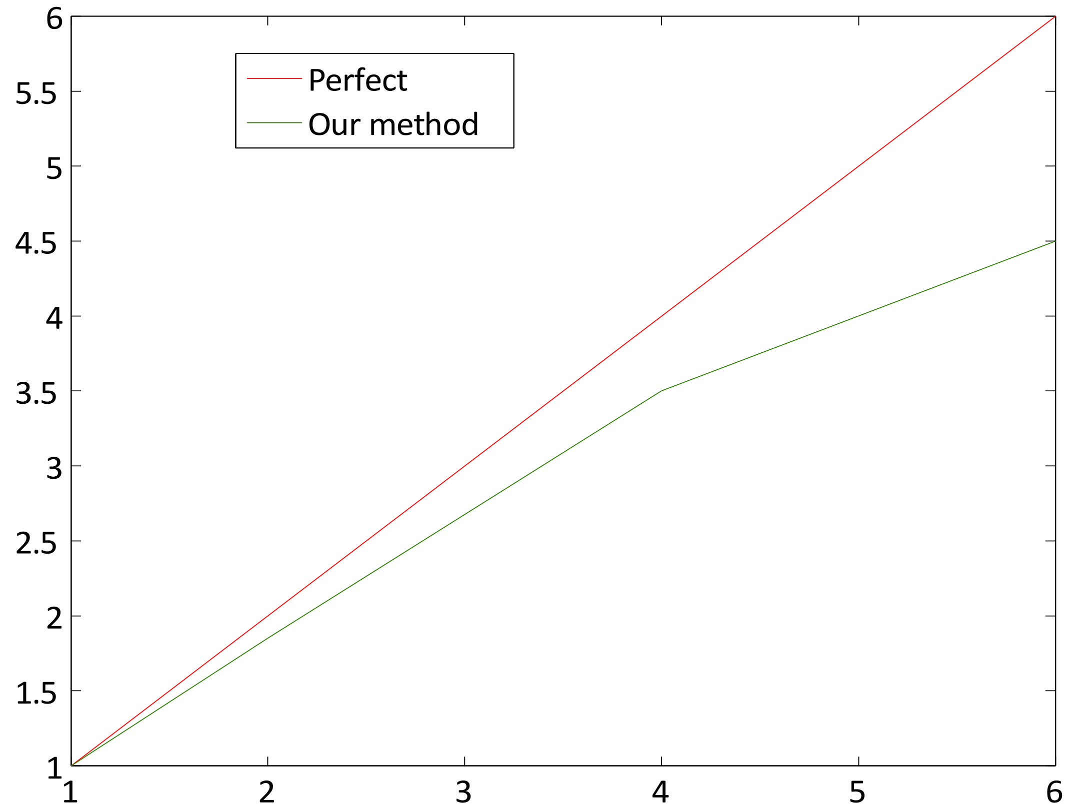
Methods	Citation	Coauthor	Film
Sum-Product	N/A	>10hr	1.8 hr
Basic TAP Learning	>10hr	369s	**57s**
Distributed TAP Learning	**39.33m**	**104s**	148s

Speedup results

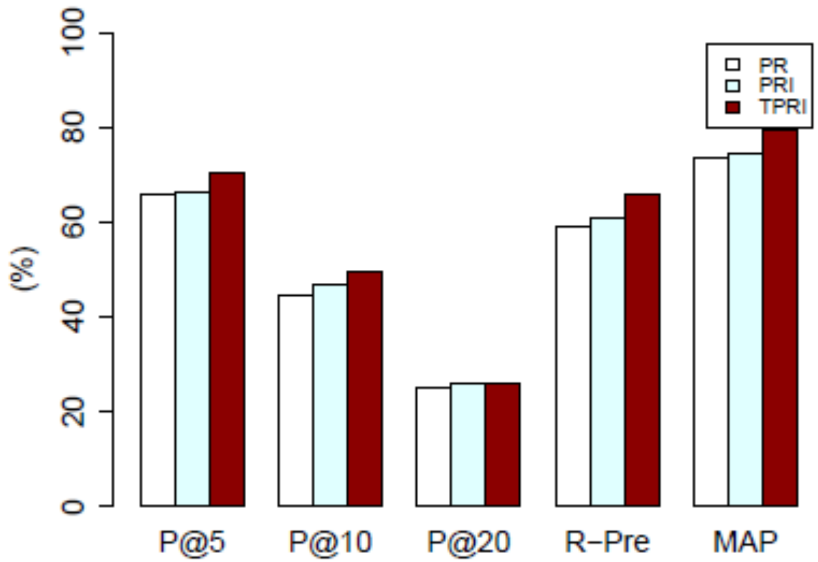


Speedup vs. Dataset size

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/

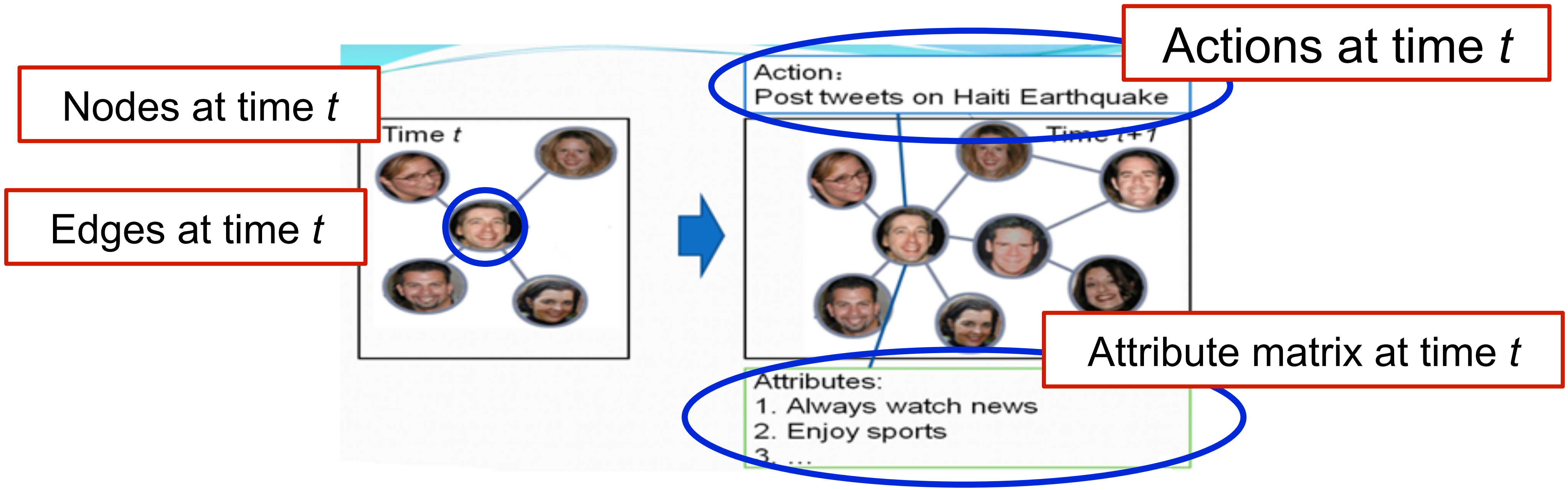
[1] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In KDD'08, pages 990-998, 2008.

Methodologies

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

Influence and Action

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



Input:

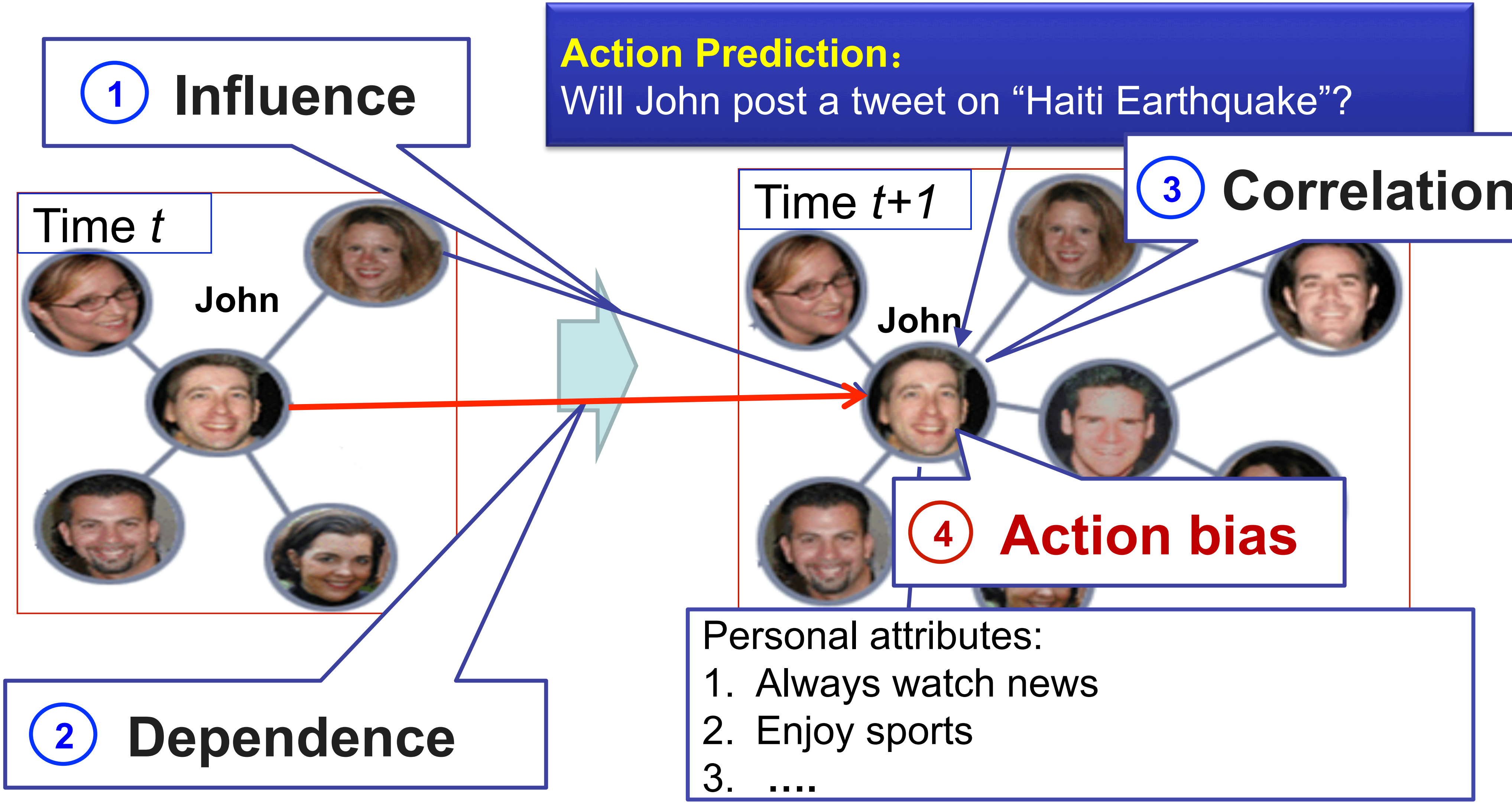
$G^t = (V^t, E^t, X^t, Y^t)$

$t = 1,2,\ldots T$

Output:

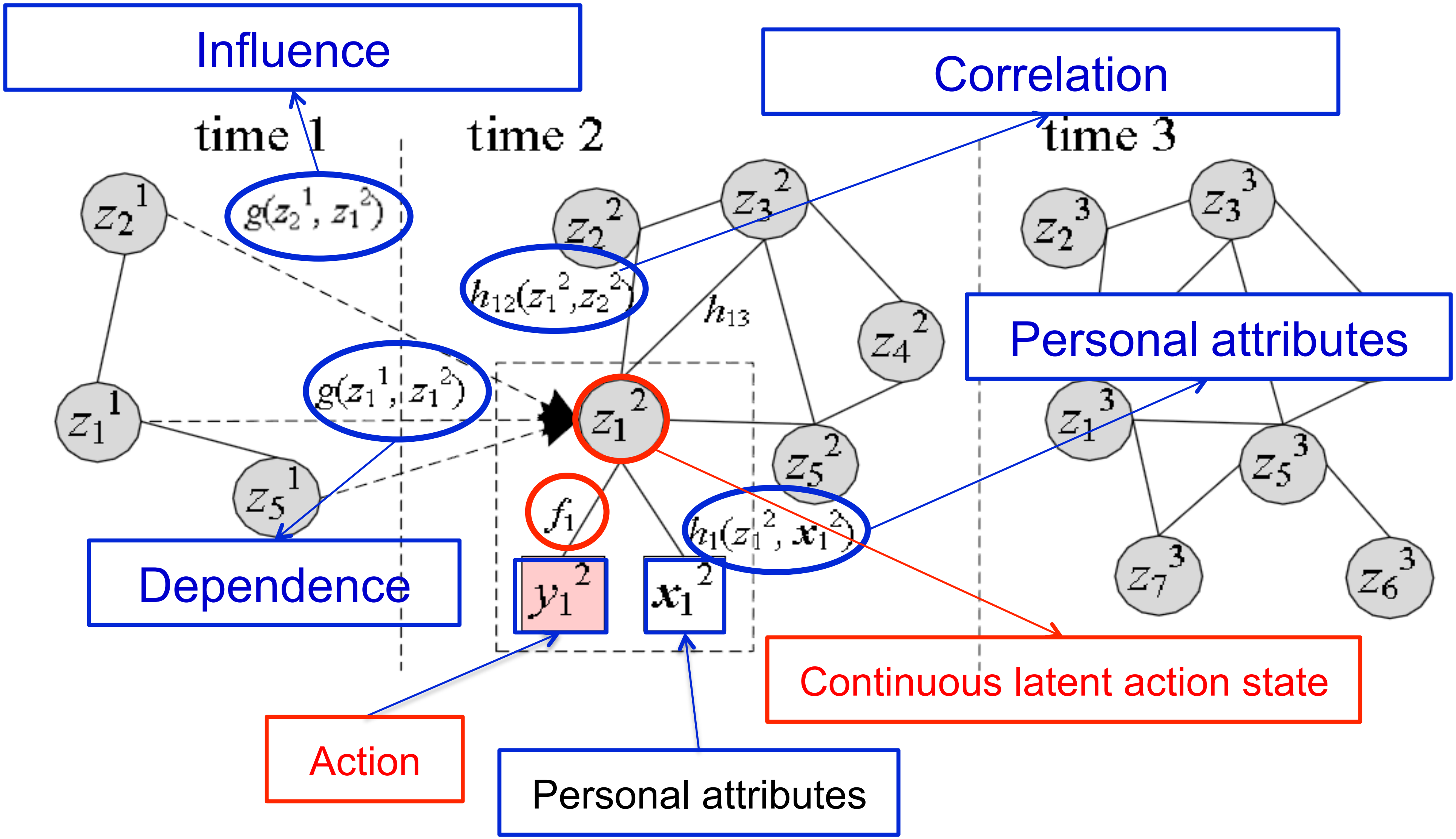
$F: f(G^t) \to Y^{(t+1)}$

Social Influence & Action Modeling[1]

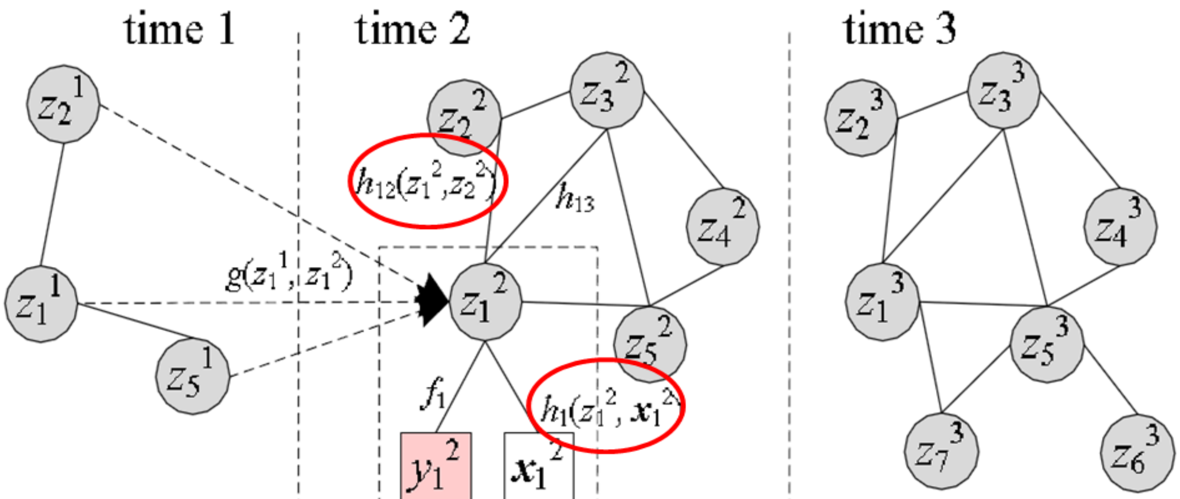


[1] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In KDD'10, pages 807–816, 2010.

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?

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

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

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:

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

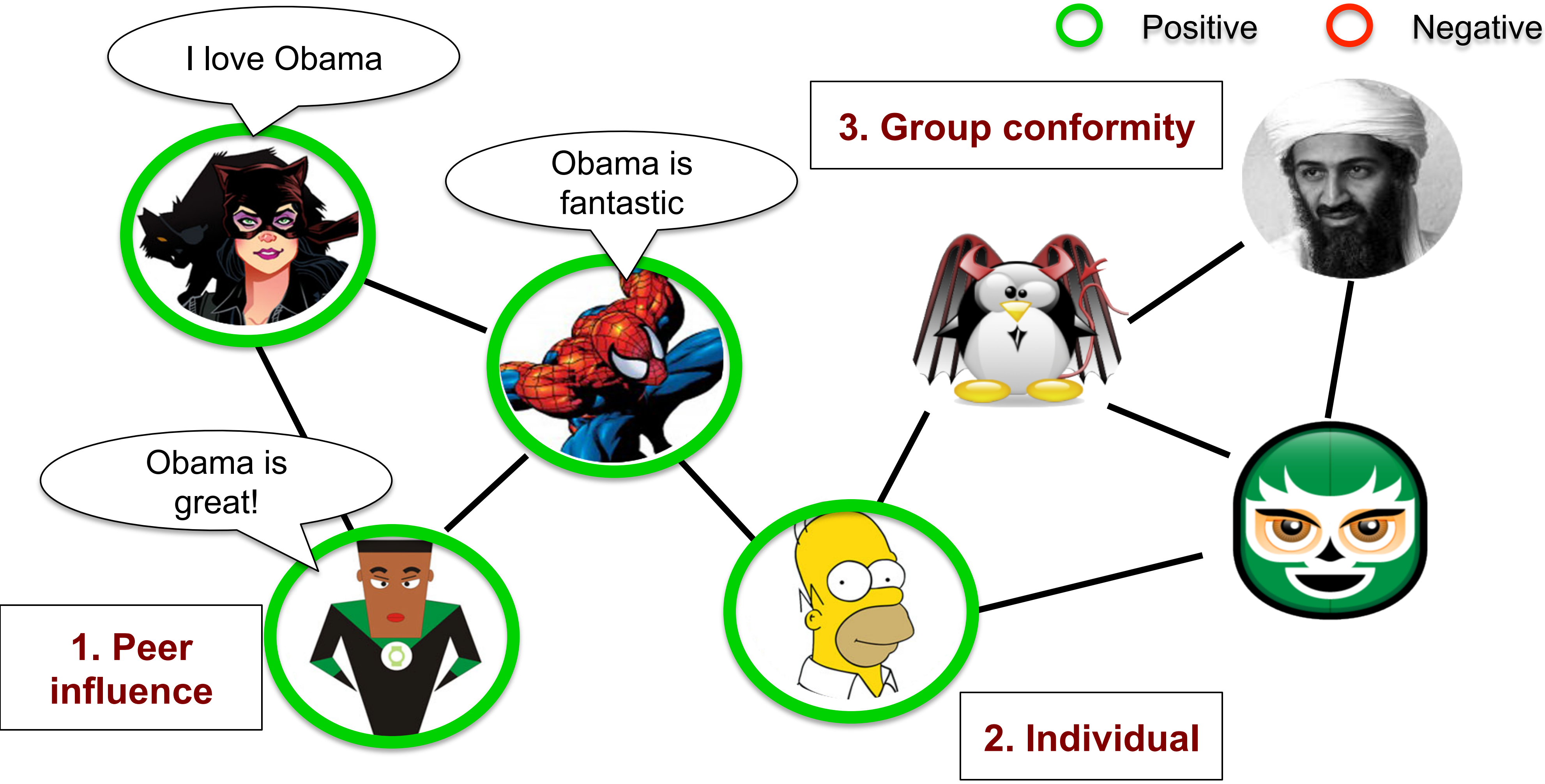
until *convergence*;

[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.

Still Challenges

- **Q1:** Are there any other social factor that may affect the prediction results?
- **Q2:** How to scale up the model to large networks?

Q1: Conformity Influence



[1] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'13**, 2013.

Conformity Factors

- Individual conformity

A specific action performed by user v at time t

$$icf(v) = \frac{|(a,v,t) \in A_v | \exists (a,v',t') : e_{vv'} \in E \wedge \epsilon \geq t - t' \geq 0|}{|A_v|}$$

All actions by user v

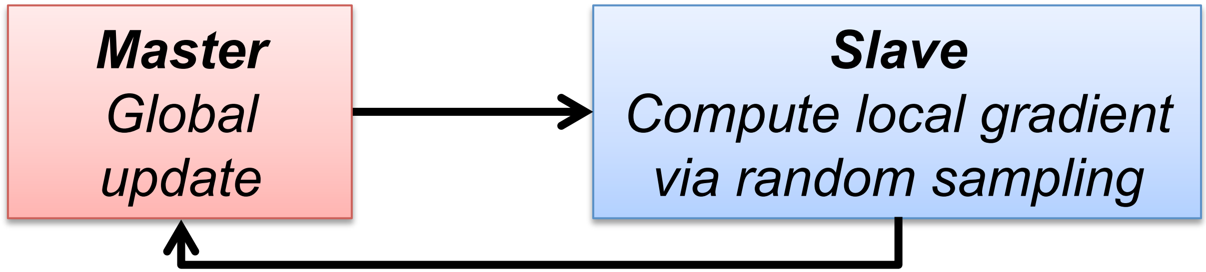
- Peer conformity

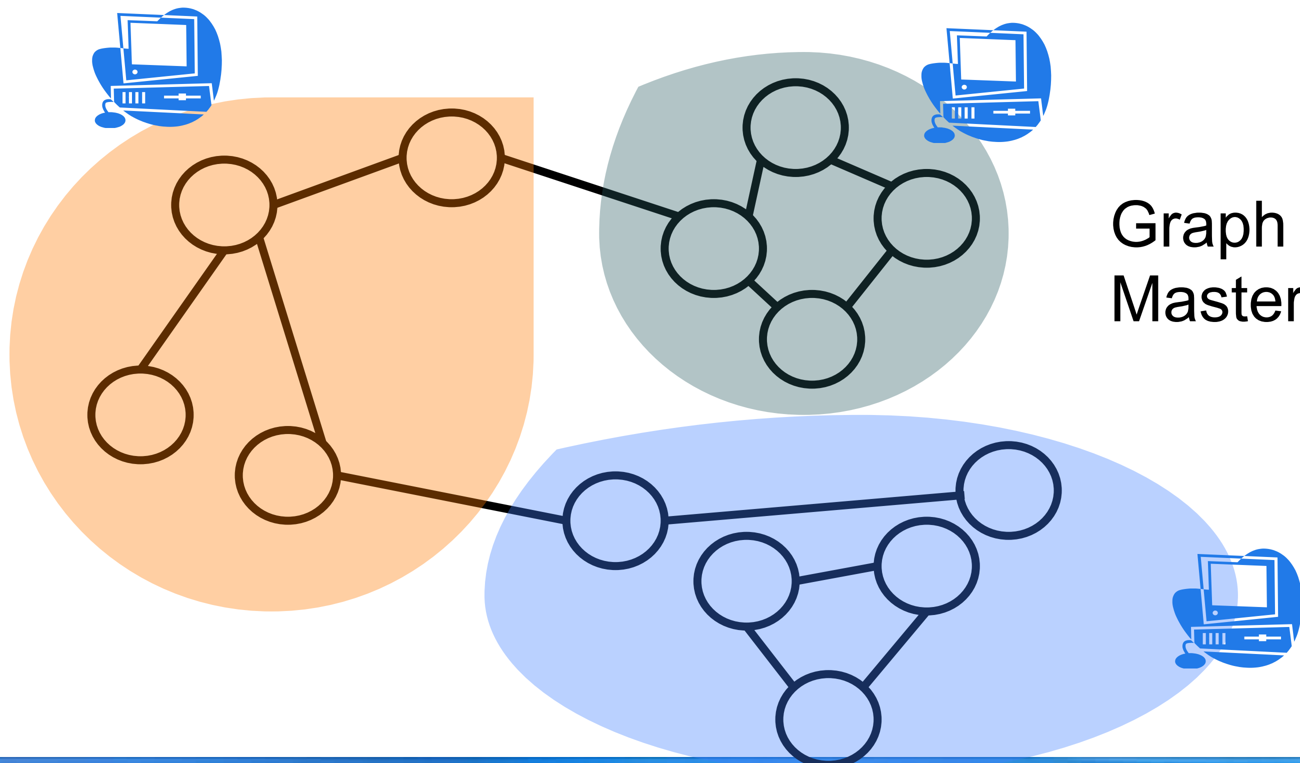
$$pcf(v,v') = \frac{|(a,v',t') \in A_{v'} | \exists (a,v,t) : e_{vv'} \in E \wedge \epsilon \geq t - t' \geq 0|}{|A_{v'}|}$$

- Group conformity

$$gcf^{\tau}(v, C_{vk}) = \frac{|(a,v',t') \in A^{\tau}_{C_k} | \exists (a,v,t) : \mathbb{I}[c_{ik}] \wedge \epsilon \geq t - t' \geq 0|}{|A^{\tau}_{C_k}|}$$

Q2: Distributed Learning

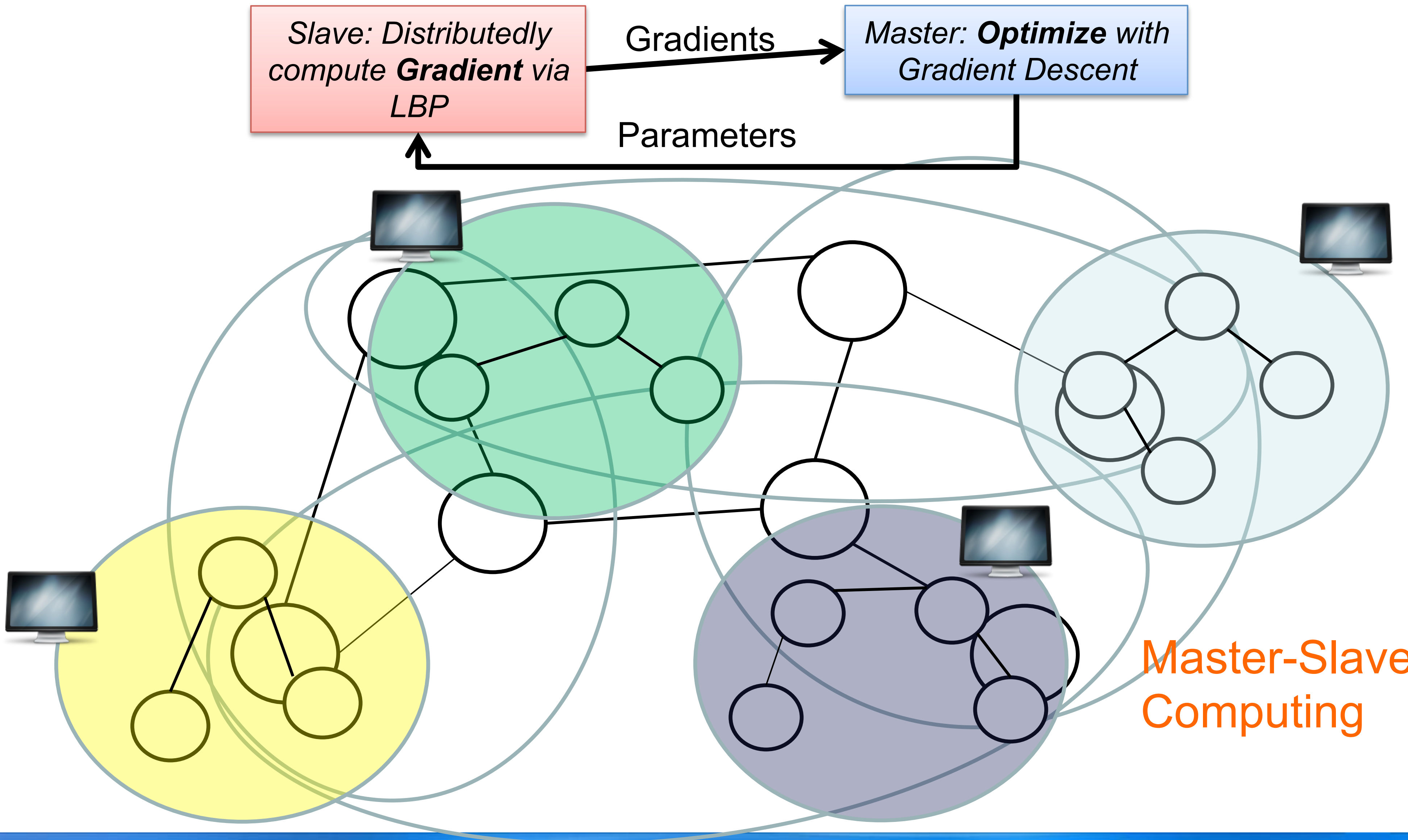




Graph Partition by Metis
Master-Slave Computing

Inevitable loss of correlation factors!

Random Factor Graphs



Model Inference

- Calculate marginal probability in each subgraph

- Aggregate the marginal probability and normalize

Theoretical Analysis

- Θ^*: Optional parameter of the complete graph
- Θ: Optional parameter of the subgraphs
- $P_{s,j}$: True marginal distributions on the complete graph
- $G^*_{s,j}$: True marginal distributions on subgraphs
- Let $E_{s,j} = \log G^*_{s,j} - \log P_{s,j}$, we have:

$$E_{s;j} \leq D(\theta||\theta^*) - \frac{\Delta_{s;j}}{G^*_{s;j}}$$

$$E_{s;j} \geq \log G^*_{s;j} - \log[1 - (1 - G^*_{s;j})exp\{-D(\theta||\theta^*) + \frac{\Delta_{s;j}}{1-G^*_{s;j}}\}]$$

$$\text{where } \Delta_{s;j} = \sum_{\alpha \in G \setminus G^*} \theta^*_\alpha cov_\theta\{\delta(x_s = j), \phi_\alpha(x)\}$$

$D(\theta||\theta^*)$ is the Kullback-Leibler divergence between $p(x;\theta)$ and $p(x;\theta^*)$

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

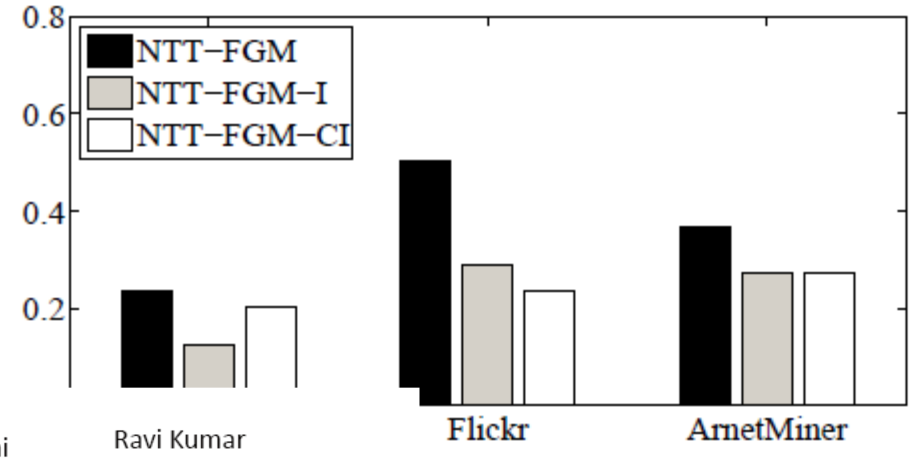
- Baseline
  - SVM
  - wvRN (Macskassy, 2003)
- Evaluation Measure:

  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-FG			
ArnetMiner	SVM			
	wvRN			
	NTT-FG			



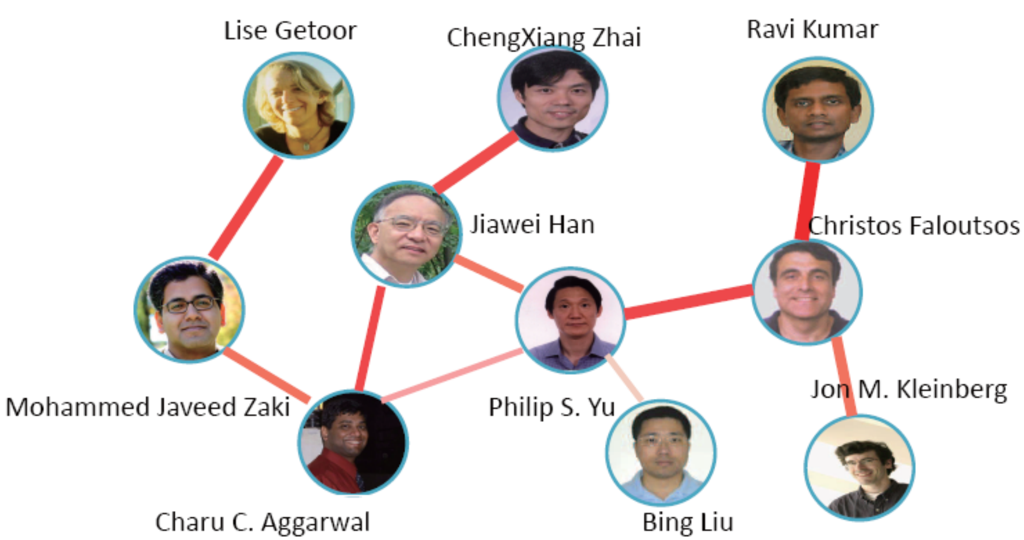
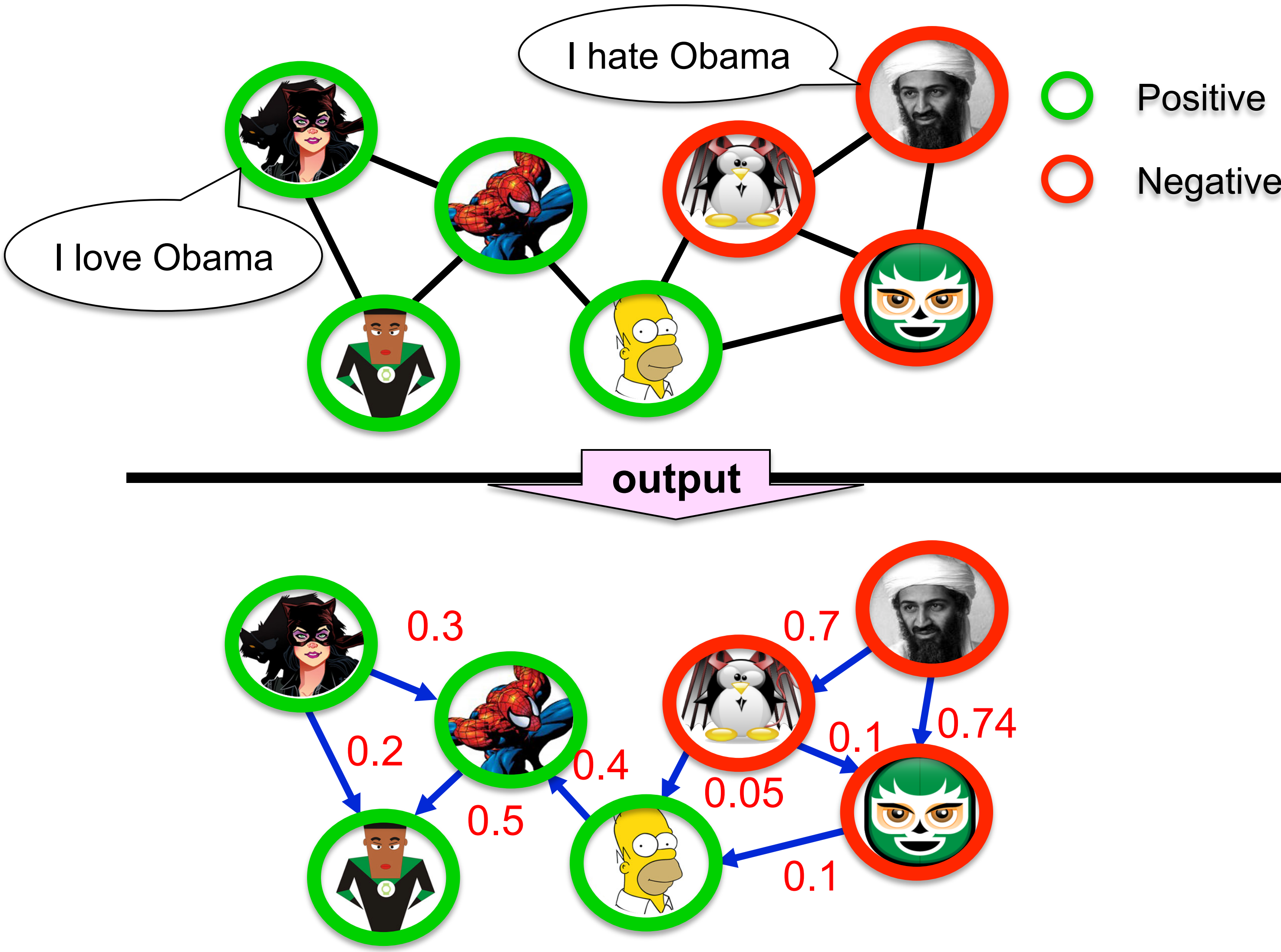


Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.

Summaries

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

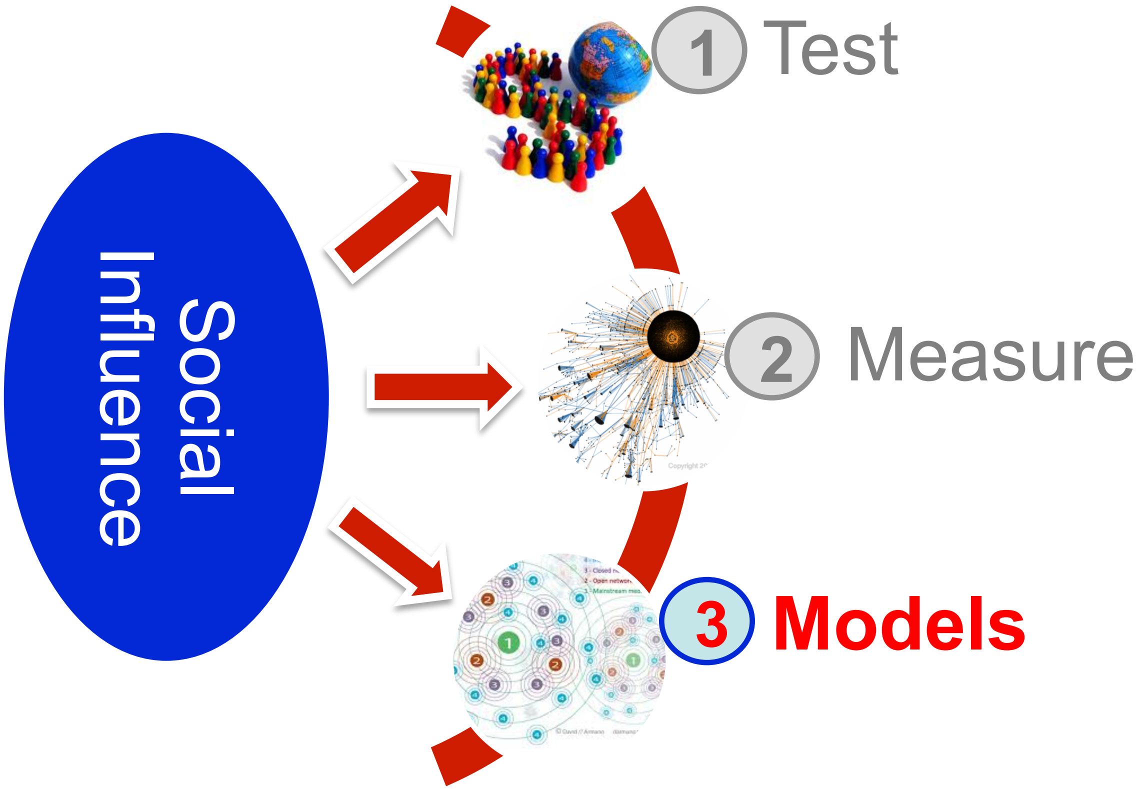
Output of Measuring Influence



Understanding the Emotional Impact in Social Networks

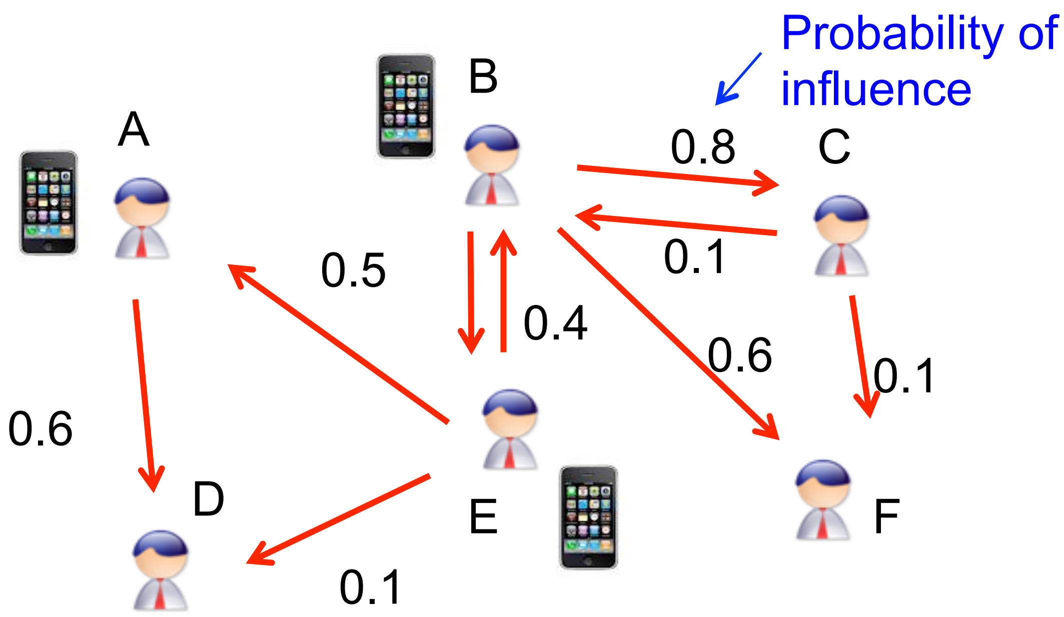
[1] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang. Can We Understand van Gogh's Mood? Learning to Infer Affects from Images in Social Networks. In ACM Multimedia, pages 857-860, 2012.

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

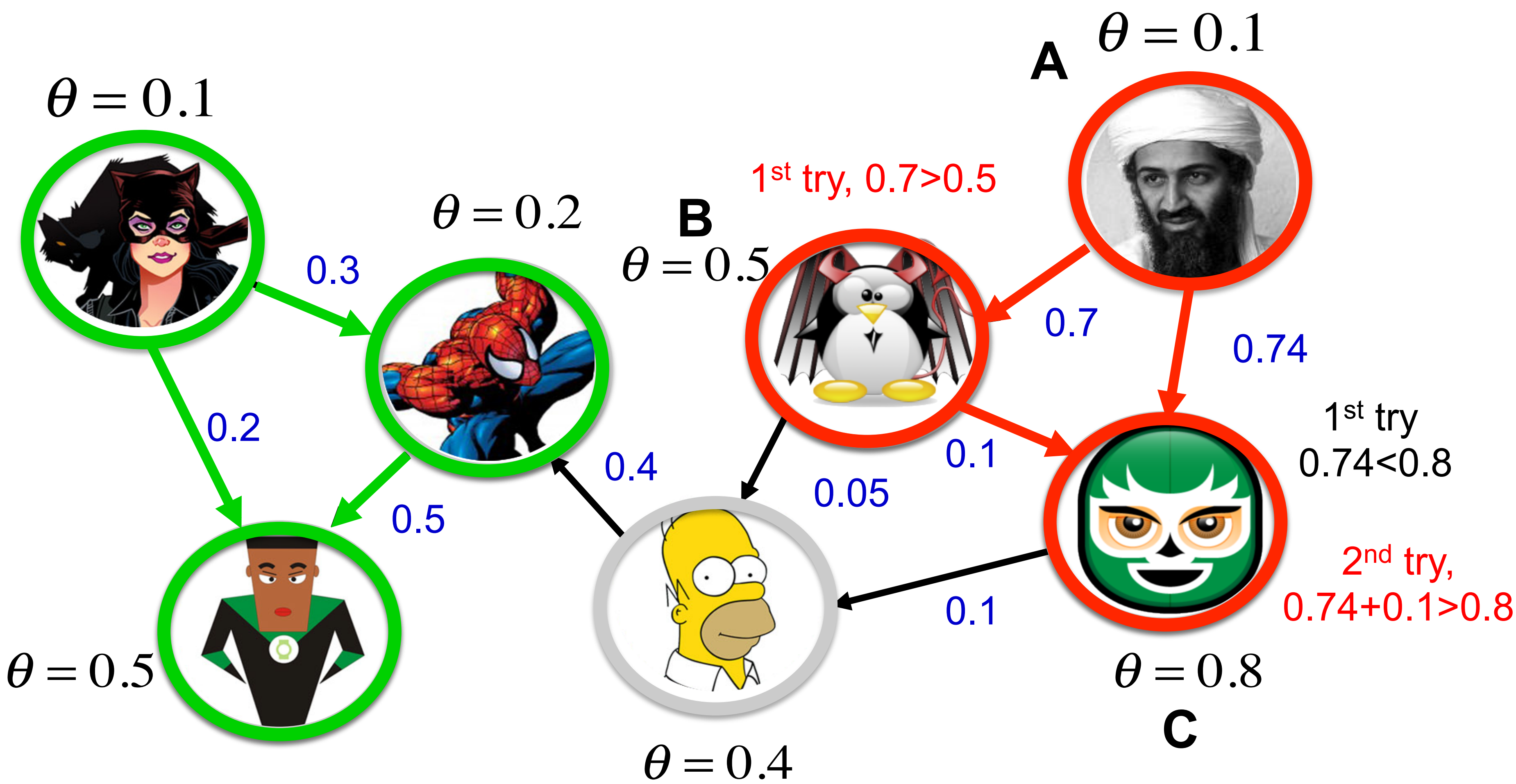
Linear Threshold Model

- 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)
  - $F(A)$ = 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 $\quad S \subseteq T.$

[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 $f(S) \geq \boxed{(1 - 1/e)} \cdot f(S^\star)$; 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.

Performance Guarantee

Let g_j be the j-th node selected by the greedy algorithm

- Let $G_j = \{g_1, \ldots, g_j\}$ and $G_0 = \emptyset$
- For $\forall S, |S| = k$ and $j = 0, 1, \ldots, k-1$

$$F(S) \underset{\text{monotonicity}}{\le} F(G_j \cup S) \underset{\text{greedy + submodularity}}{\le} F(G_j) + k g_{j+1}$$

- Let $\Delta_j = F(S^*) - F(G_j)$

where S^* is the optimal solution

- We have $g_{j+1} = \Delta_j - \Delta_{j+1}$

- Thus $\Delta_j \le k(\Delta_j - \Delta_{j+1})$

$$\Delta_k \le \left(1 - \frac{1}{k}\right)^k \Delta_0$$

$$\boxed{\text{Recall } e^x \ge 1 + x} \quad \le \frac{1}{e} F(S^*)$$

- Then

$$F(G_k) \ge \left(1 - \frac{1}{e}\right) F(S^*)$$

*The solution obtained by Greedy is better than **63%** of the optimal solution*

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)

```
initialize S = ∅ and R = 20000
for i = 1 to k do
    for each vertex v ∈ V \ S do
        s_v = 0.
        for i = 1 to R do
            s_v += |RanCas(S ∪ {v})|
        end for
        s_v = s_v / R
    end for
    S = S ∪ {arg max_{v∈V\S} {s_v}}
end for
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)

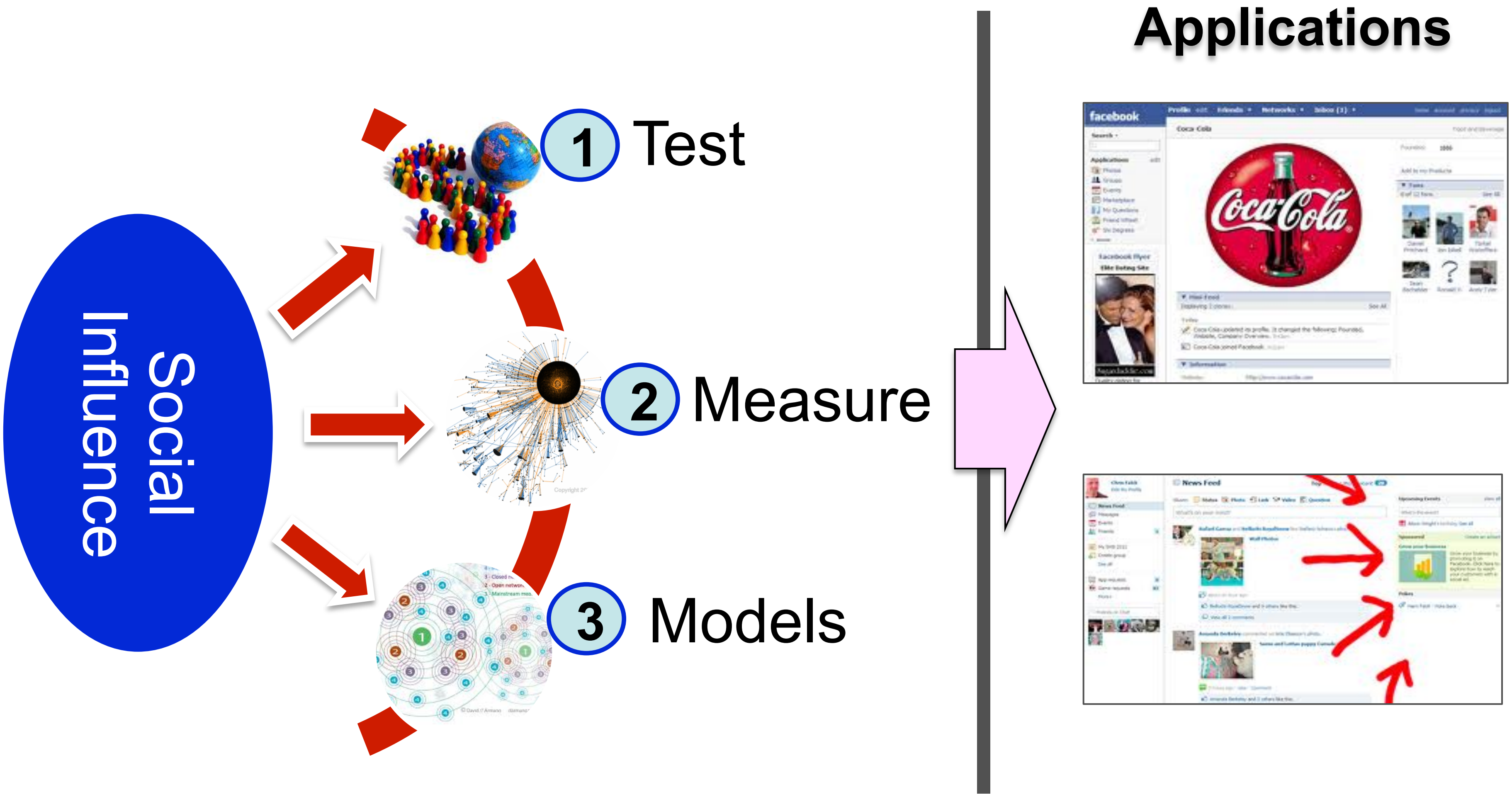
initialize $S = \emptyset$
for each vertex v **do**
    compute its degree d_v
    $dd_v = d_v$
    initialize t_v to 0
end for
for $i = 1$ to k **do**
    select $u = \arg\max_v \{dd_v \mid v \in V \setminus S\}$
    $S = S \cup \{u\}$
    for each neighbor v of u and $v \in V \setminus S$ **do**
        $t_v = t_v + 1$
        $dd_v = d_v - 2t_v - (d_v - t_v)t_v p$
    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.

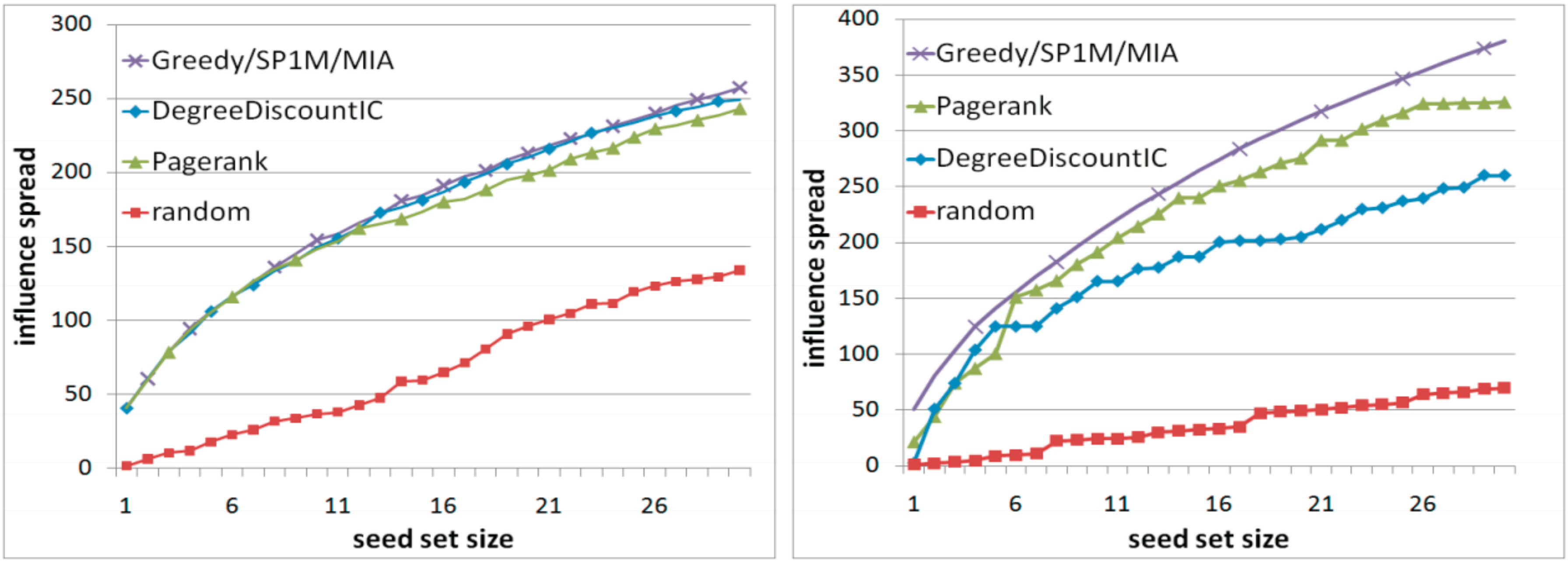
Example 1: Influence maximization with the learned influence probabilities

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

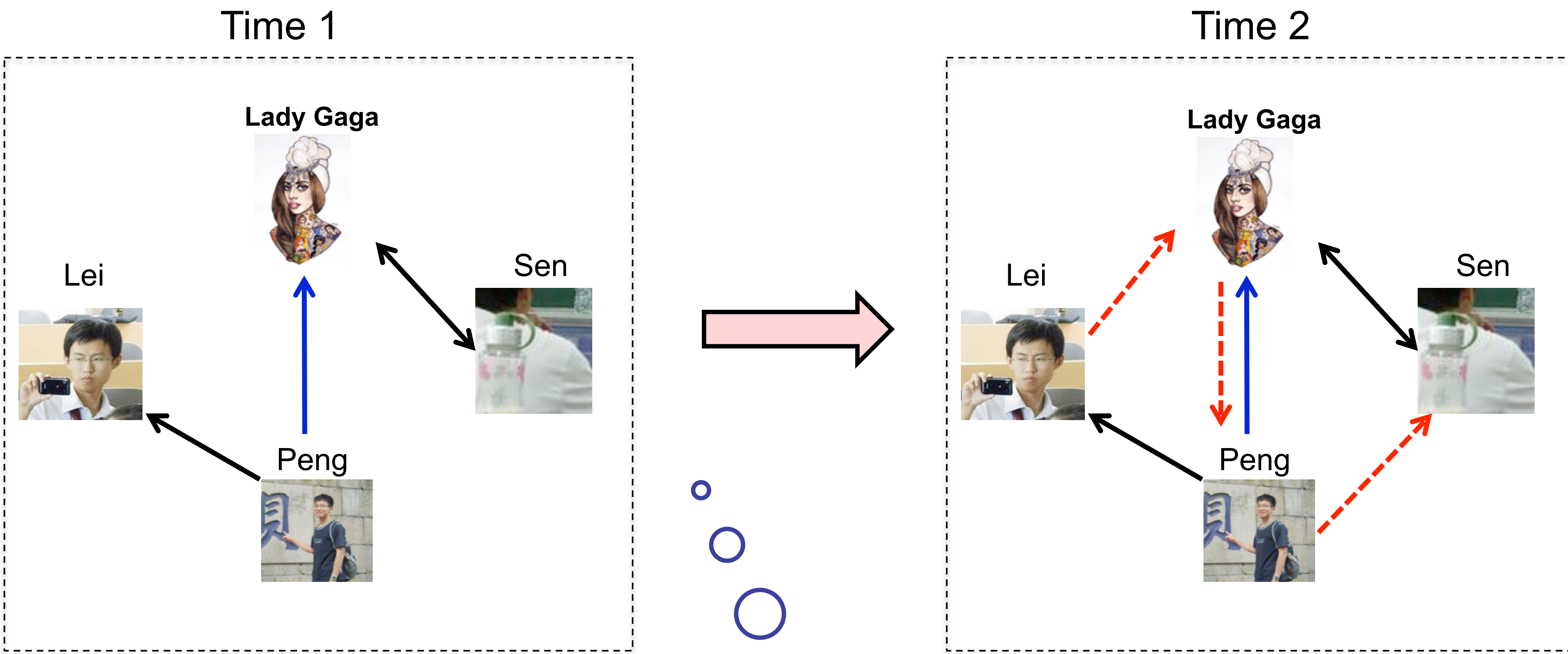
(b) With the learned influence

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.

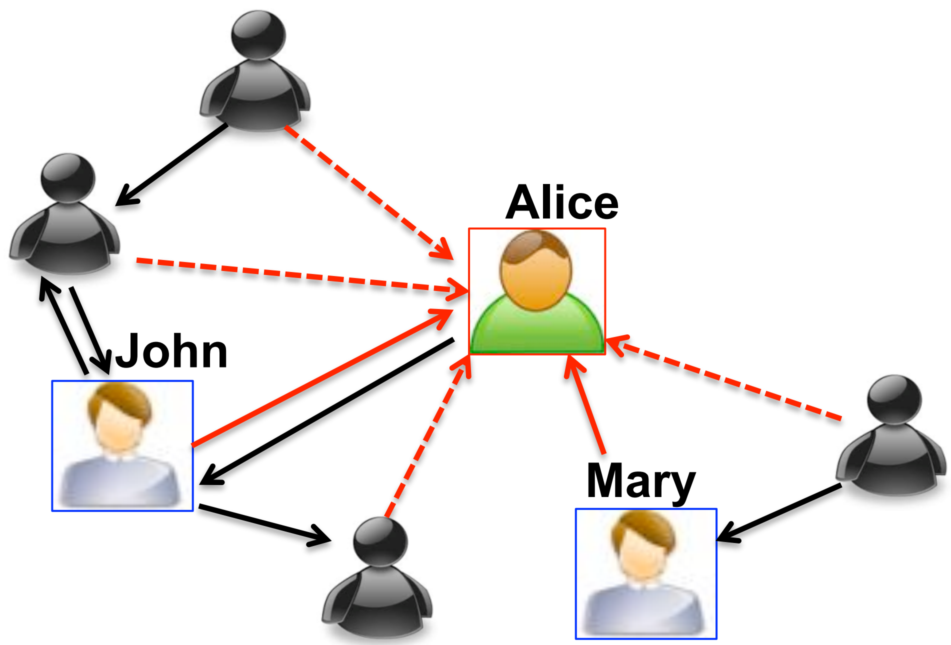
Example 2: Following Influence Applications

Following Influence Applications



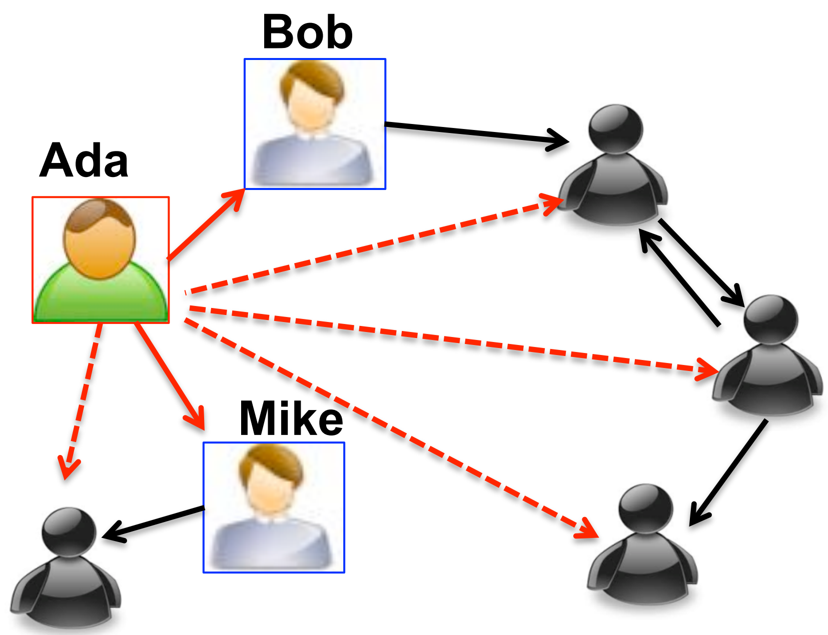
When you follow a user in a social network, will the behavior influences your friends to also follow her?

Applications: Influence Maximization



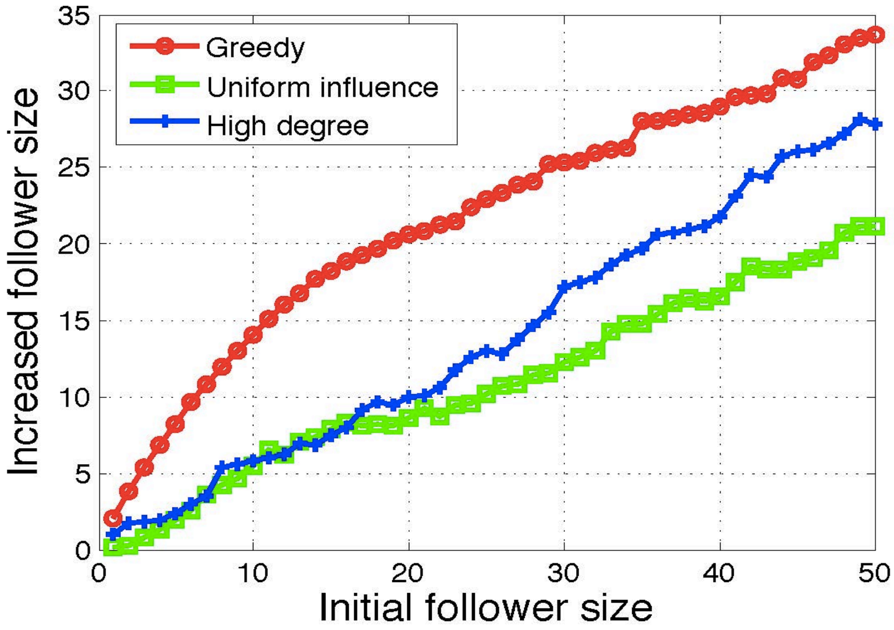
Find a set *S* of *k* initial followers to follow user *v* such that the number of newly activated users to follow *v* is maximized.

Applications: Friend Recommendation

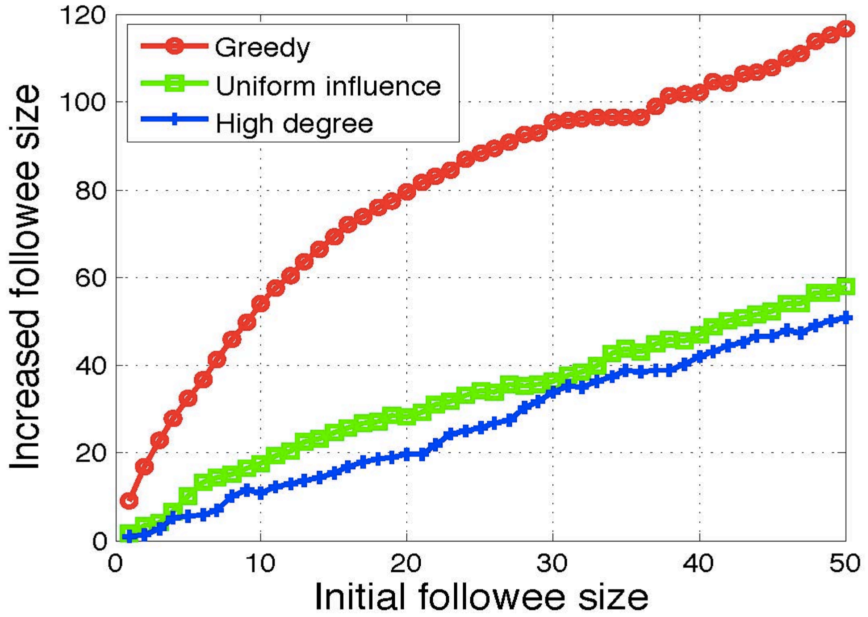


Find a set S of k initial followees for user v such that the total number of new followees accepted by v is maximized

Application Performance



Influence Maximization



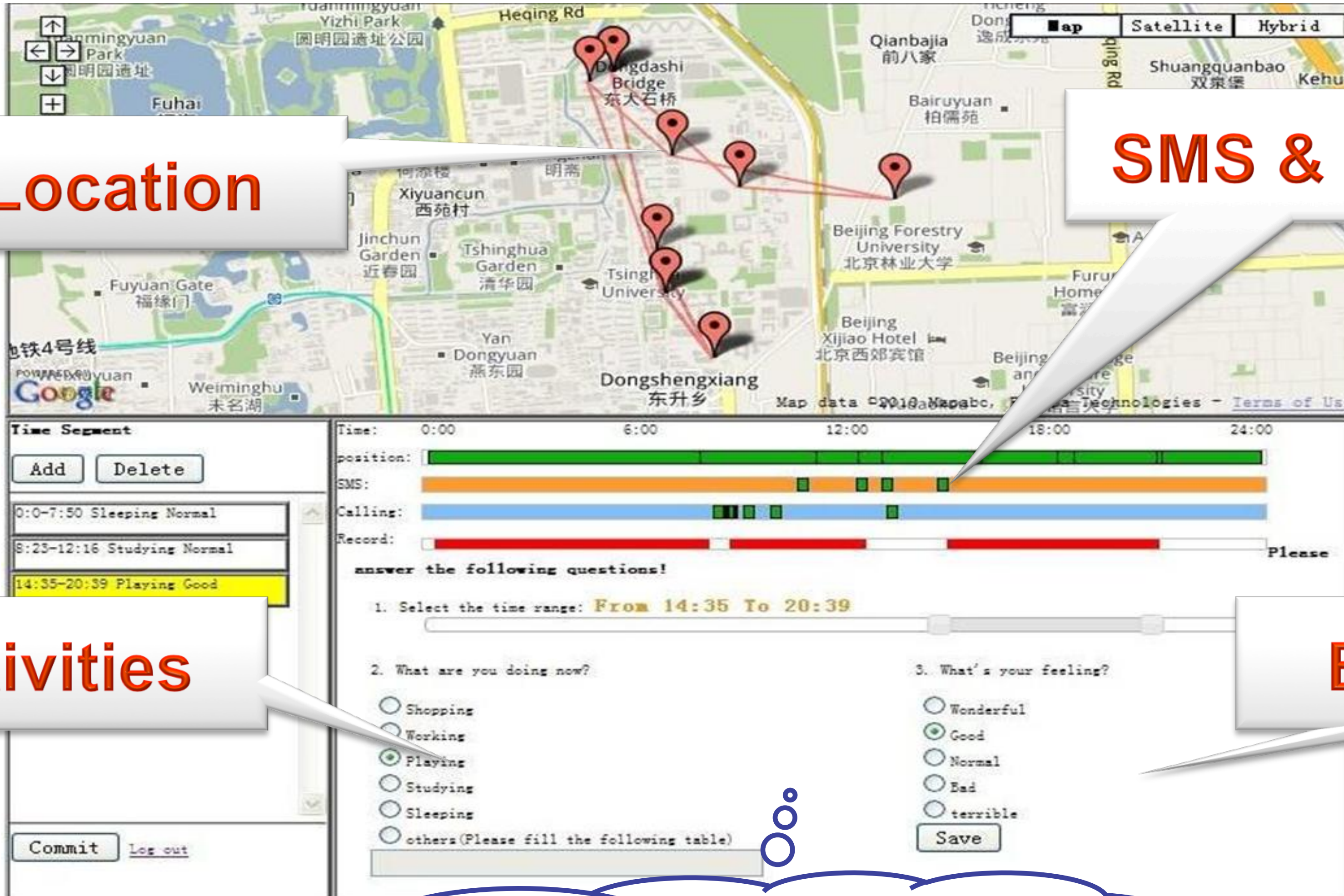
Recommendation

- 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



Location

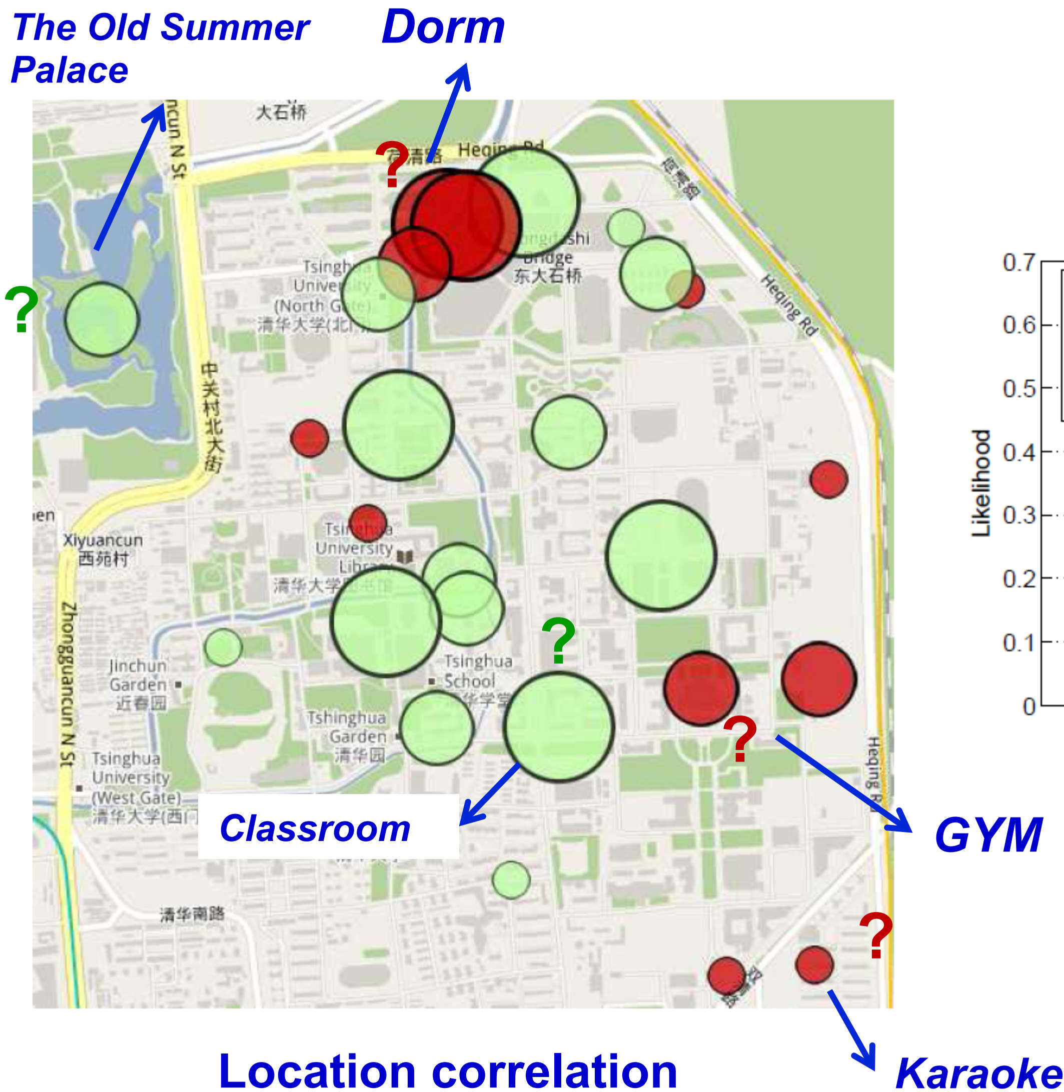
SMS & Calling

Activities

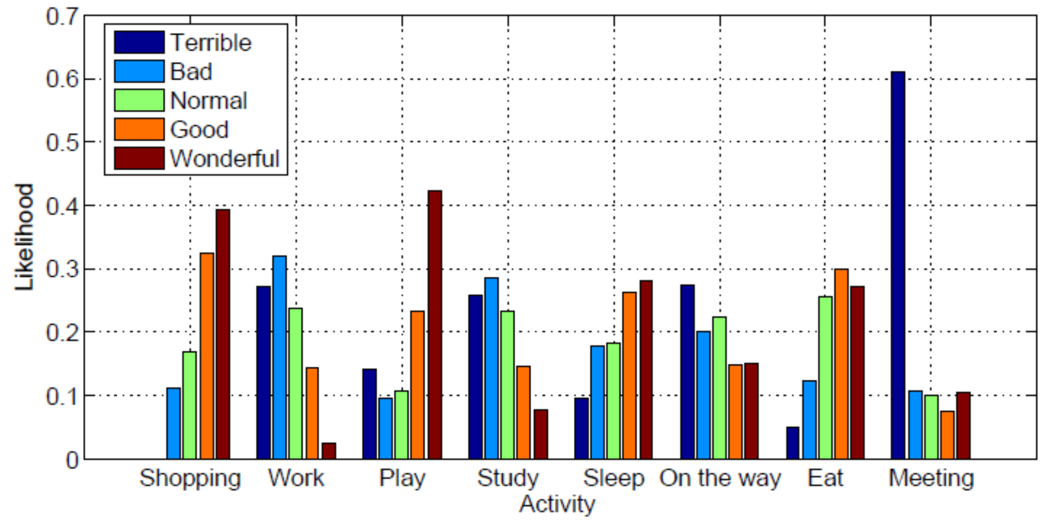
Emotion

Can we predict users' emotion?

Observations (cont.)

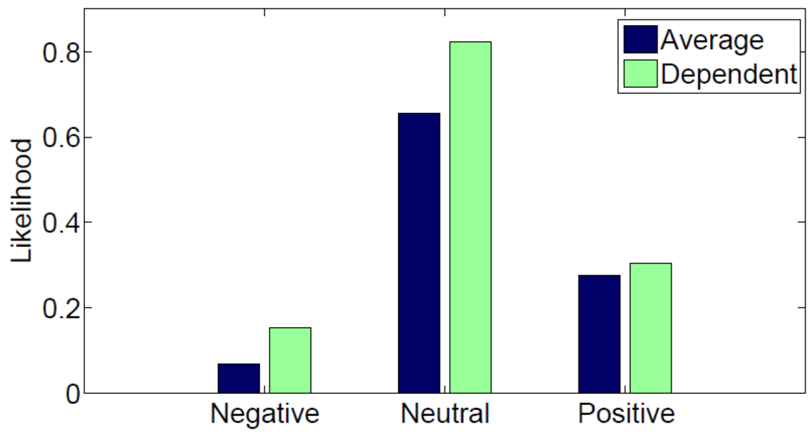


Location correlation
(Red-happy)



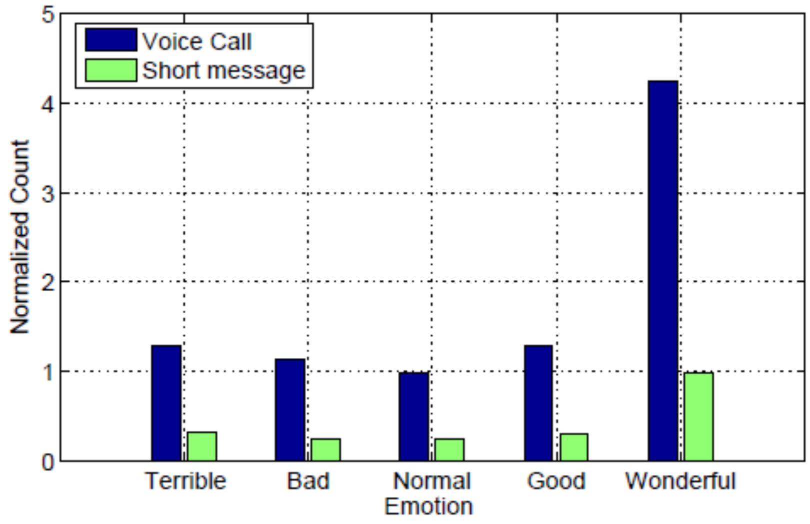
Activity correlation

Observations



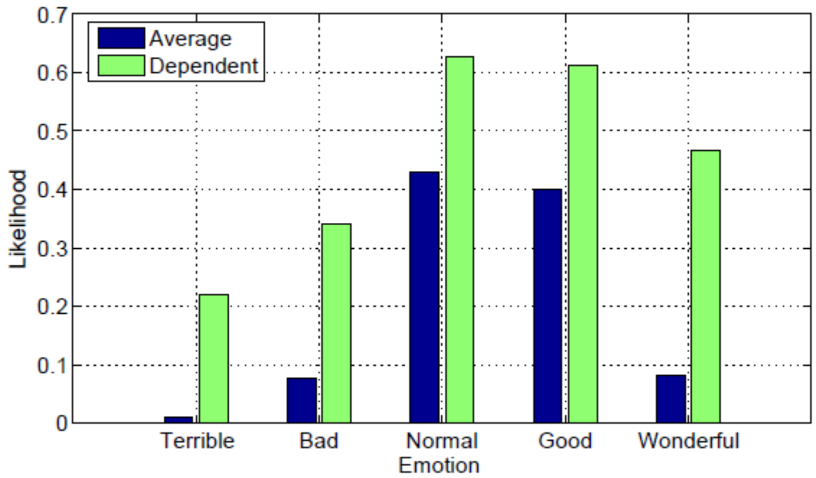
(a) Social correlation

(a) Implicit groups by emotions

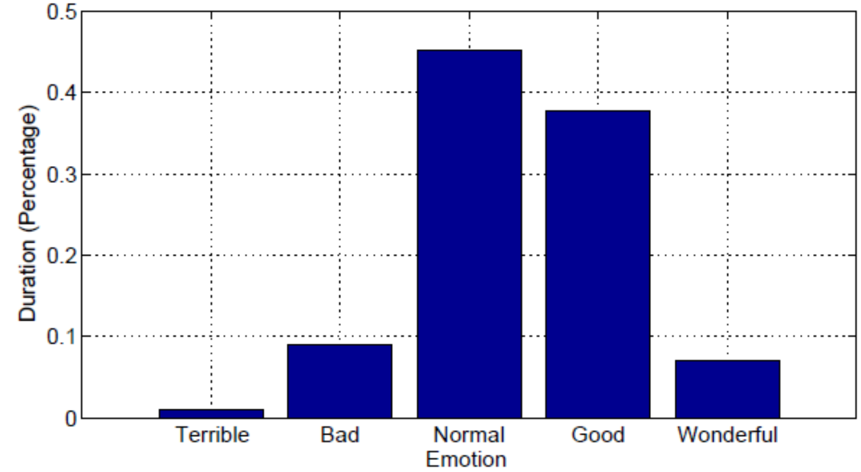


(c) Calling (SMS) correlation

Observations (cont.)

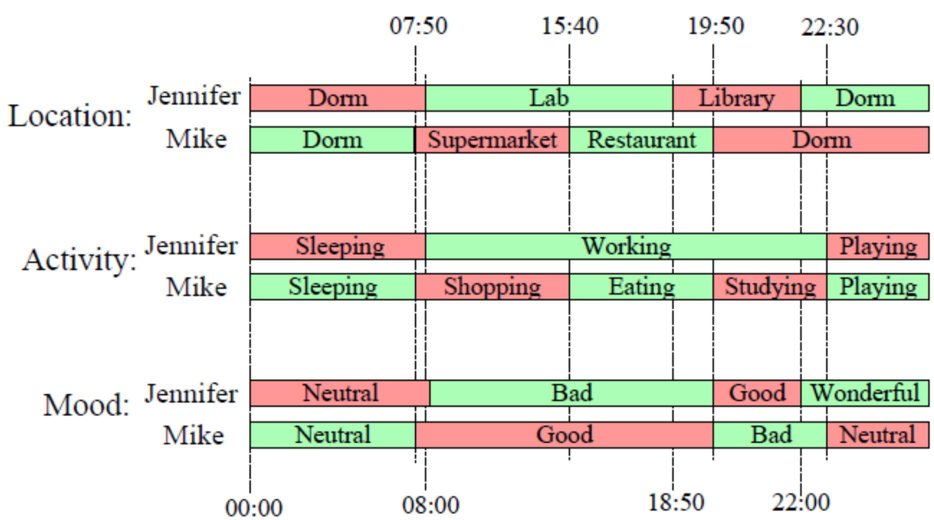


(a) Temporal correlation

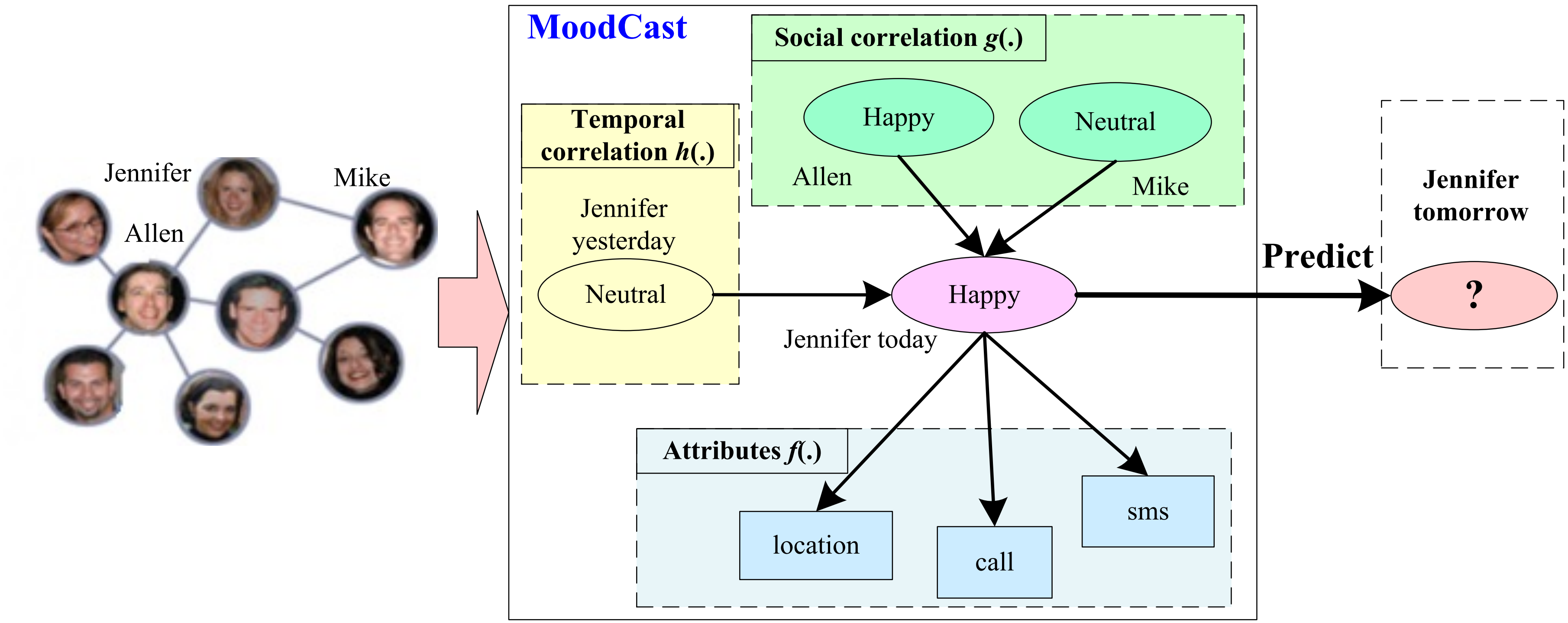


(b) Time duration

Temporal correlation



MoodCast: Dynamic Continuous Factor Graph Model

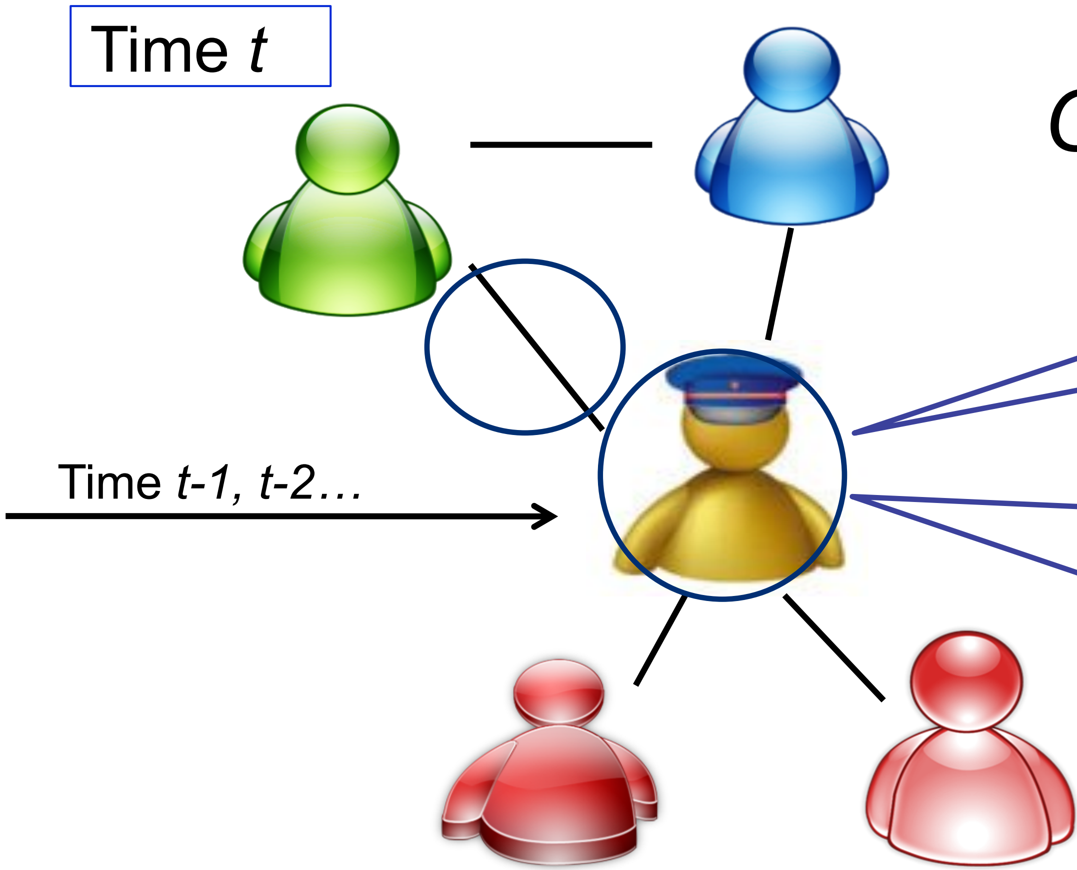


Our solution

1. We directly define continuous feature function;

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

Problem Formulation



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

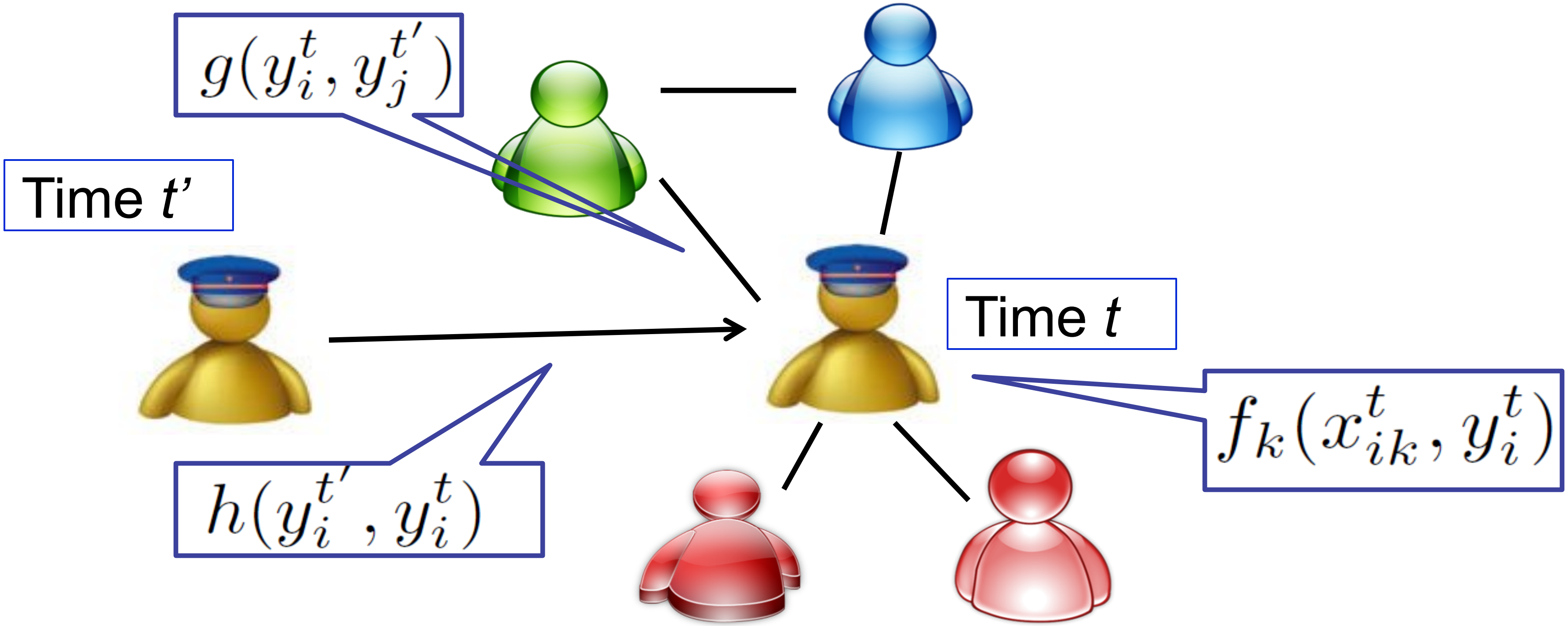
Emotion: Sad

Attributes:
- Location: Lab
- Activity: Working

Learning Task:

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

Dynamic Continuous Factor Graph Model

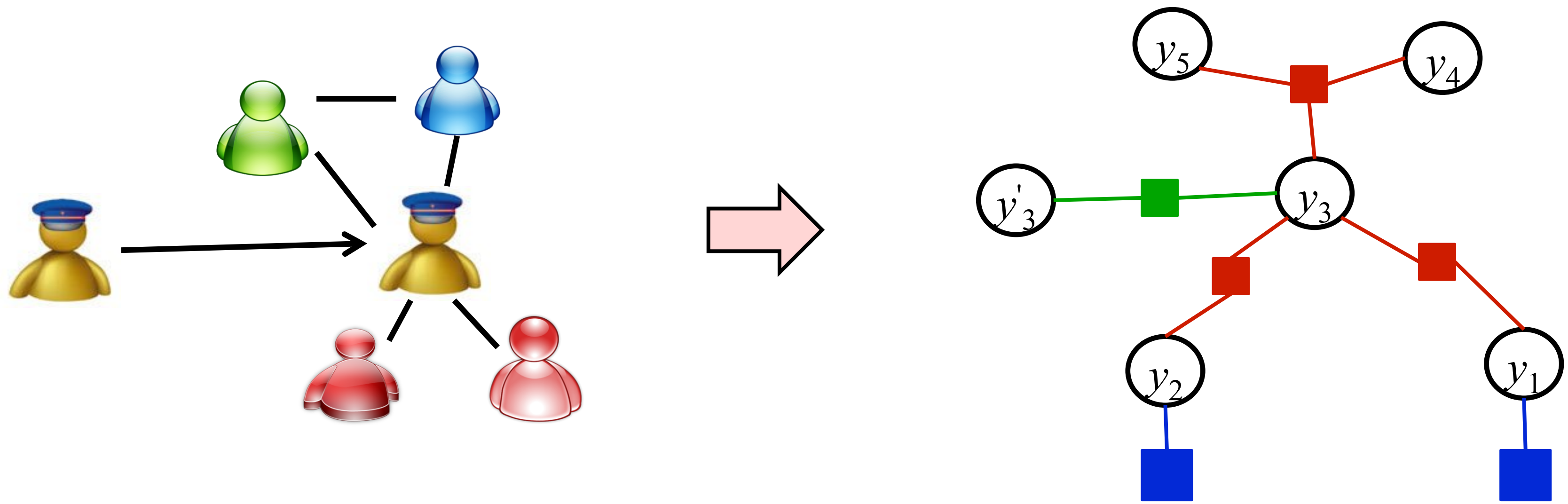


$f_k(x_{ik}^t, y_i^t)$: Binary function

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

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

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	

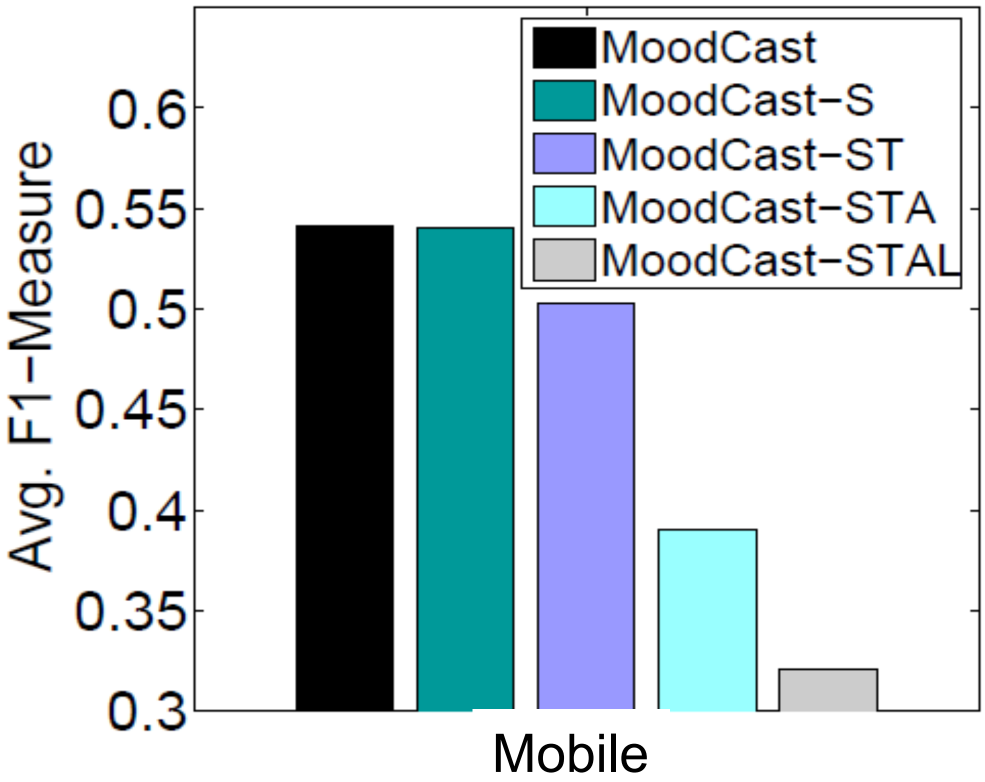
- Baseline
  - SVM
  - SVM with network features
  - Naïve Bayes
  - Naïve Bayes with network features
- Evaluation Measure:

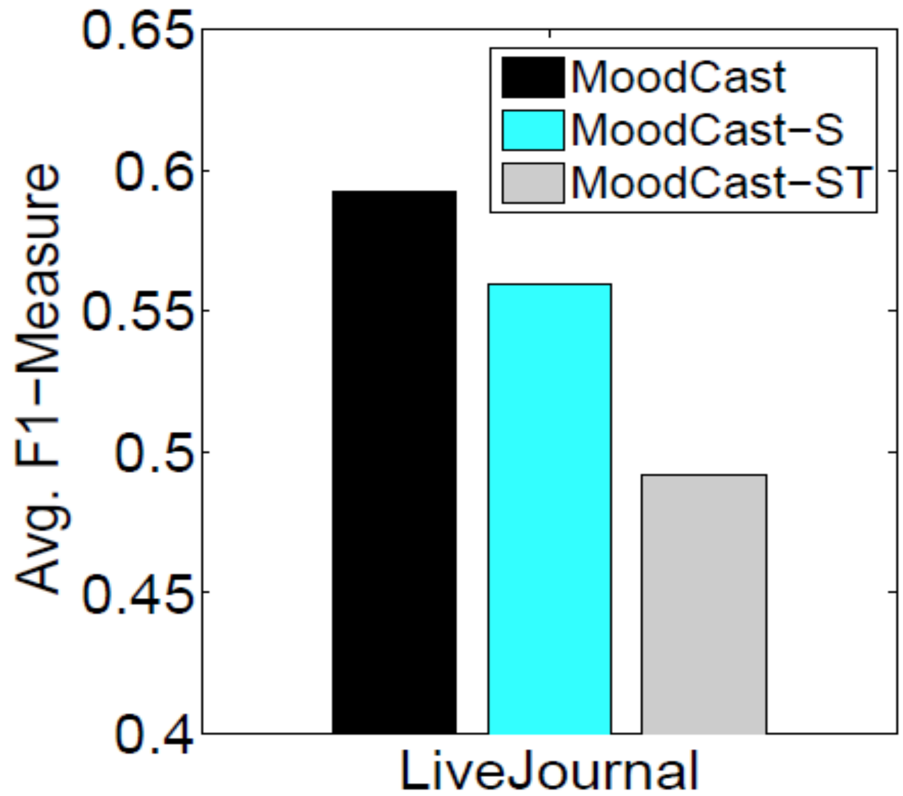
  Precision, Recall, F1-Measure

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



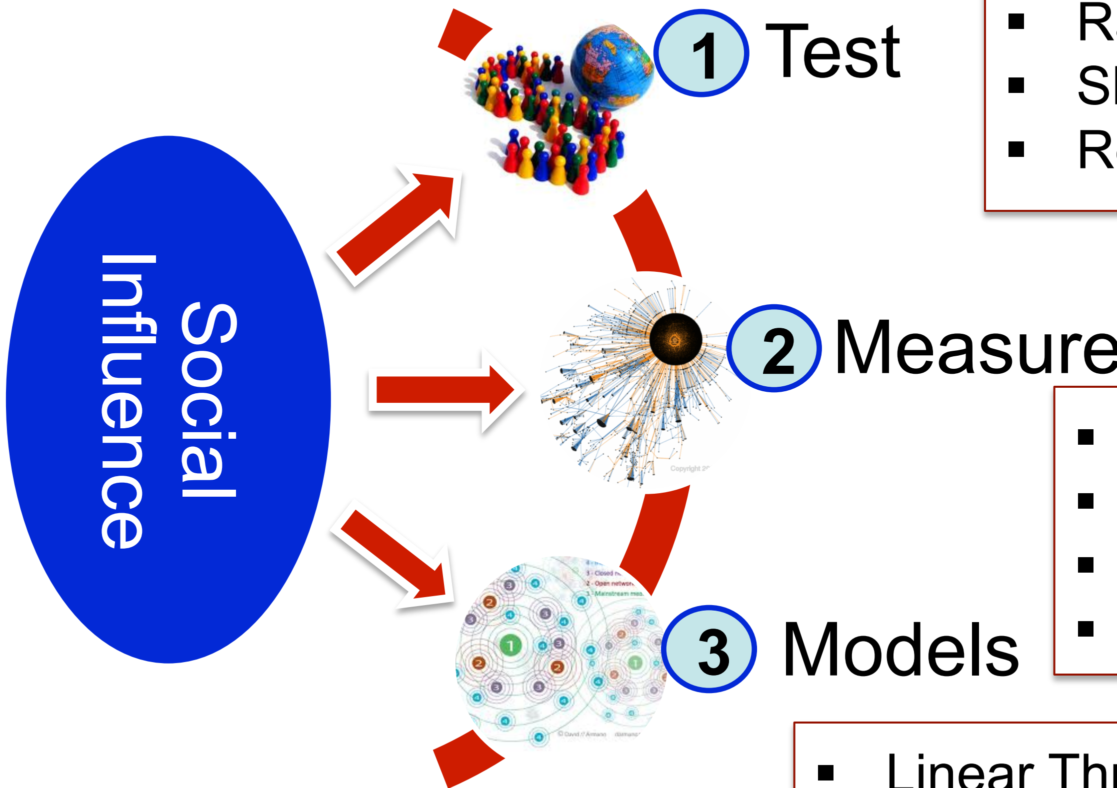


- All factors are important for predicting user emotions

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, 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.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In **KDD'09**, pages 807-816, 2009.
- 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.
- 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.
- 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.
- 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.
- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. In **WSDM'12**. pp. 743-752.
- 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, and Shiqiang Yang. Learning Influence from Heterogeneous Social Networks. In **DMKD**, 2012, Volume 25, Issue 3, pages 511-544.
- Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social Influence Locality for Modeling Retweeting Behaviors. In **IJCAI'13**.
- Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'2013**.
- Jimeng Sun and Jie Tang. Models and Algorithms for Social Influence Analysis. In **WSDM'13**. (Tutorial)
- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. In **TKDD**, 2013.

References

- N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In **WSDM’08**, pages 207–217, 2008.
- A. Anagnostopoulos, R. Kumar, M. Mahdian. Influence and correlation in social networks. In **KDD’08**, pages 7-15, 2008.
- S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. **PNAS**, 106 (51):21544-21549, 2009.
- S. Aral and D Walker. Identifying Influential and Susceptible Members of Social Networks. **Science**, 337:337-341, 2012.
- Barabasi and Albert (1999). Emergence of scaling n complex networks.
- 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.
- 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.
- P. Bonacich. Power and centrality: a family of measures. **American Journal of Sociology**, 92:1170–1182, 1987.
- 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.
- R. S. Burt. Structural holes and good ideas. American Journal of Sociology, 110:349–399, 2004.
- W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In **KDD'09**, pages 199-207, 2009.

References(cont.)

- 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. Domingos and M. Richardson. Mining the network value of customers. In **KDD'01**, pages 57–66, 2001.
- R. Dunbar. Neocortex size as a constraint on group size in primates. **Human Evolution**, 1992, 20: 469–493.
- 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.
- Erdős, P.; Rényi, A. (1959), "On Random Graphs.".
- T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In **WWW'10**, 2010.
- 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
- M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In **KDD'10**, pages 1019–1028, 2010.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In **CIKM'08**, pages 499–508, 2008.
- A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In **WSDM'10**, pages 207–217, 2010.

References(cont.)

- 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.
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In **KDD'03**, pages 137–146, 2003.
- J. Kleinberg. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, 46(5):604–632, 1999.
- Lazarsfeld et al. (1944). The people's choice: How the voter makes up his mind in a presidential campaign.
- 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.
- S. Milgram. The Small World Problem. **Psychology Today**, 1967, Vol. 2, 60–67
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. Science, 2004
- http://klout.com
- P. Moore. Why I Deleted My Klout Profile, at **Social Media Today**, originally published November 19, 2011; retrieved November 26 2011
- M. E. J. Newman. A measure of betweenness centrality based on random walks. **Social Networks**, 2005.
- 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.
- D. B. Rubin, 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology** 66, 5, 688–701.

References(cont.)

- 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.
- J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In **ICDM'05**, pages 418–425, 2005.
- J. Ugandera, L. Backstromb, C. Marlowb, and J. Kleinberg. Structural diversity in social contagion. **PNAS**, 109 (20):7591-7592, 2012.
- D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. **Nature**,393(6684), pages 440–442, Jun 1998.
- http://en.wikipedia.org/wiki/Randomized_experiment

Thank you！

Collaborators: John Hopcroft, Jon Kleinberg, Chenhao Tan (**Cornell**)

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

Tiancheng Lou (**Google**)

Jimeng Sun (**IBM**)

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