



Computational Models for Social Networks

—**Social Influence** and **Structural Hole**

Jie Tang

Department of Computer Science and Technology
Tsinghua University

Networked World

facebook

- **1.26 billion** users
- **700 billion** minutes/month



- **280 million** users
- **80% of users** are 80-90's

twitter



- **555 million** users
- **.5 billion** tweets/day



- **560 million** users
- **influencing** our daily life

amazon.com

- **79 million** users per month
- **9.65 billion** items/year



- **500 million** users
- **35 billion** on 11/11



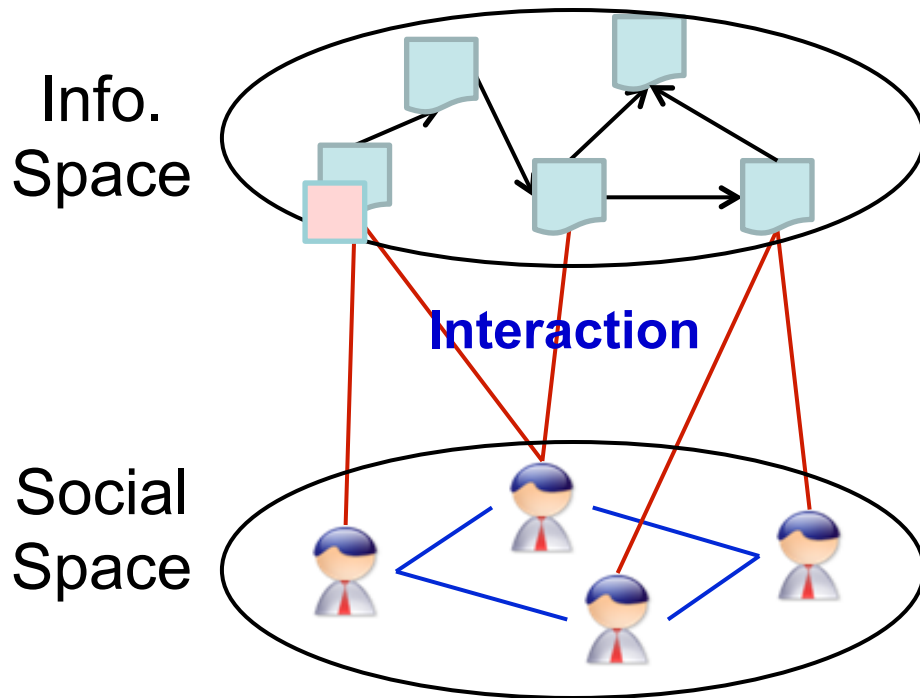
- **800 million** users
- **~50% revenue** from network life

Challenge: Big Social Data

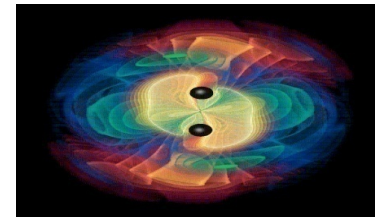
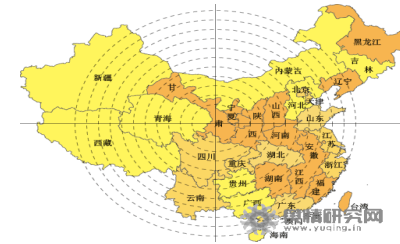
- We generate 2.5×10^{18} byte *big data* per day.
- Big social data:
 - 90% of the data was generated in the past 2 yrs
 - How to mine deep knowledge from the big social data?

Social Networks

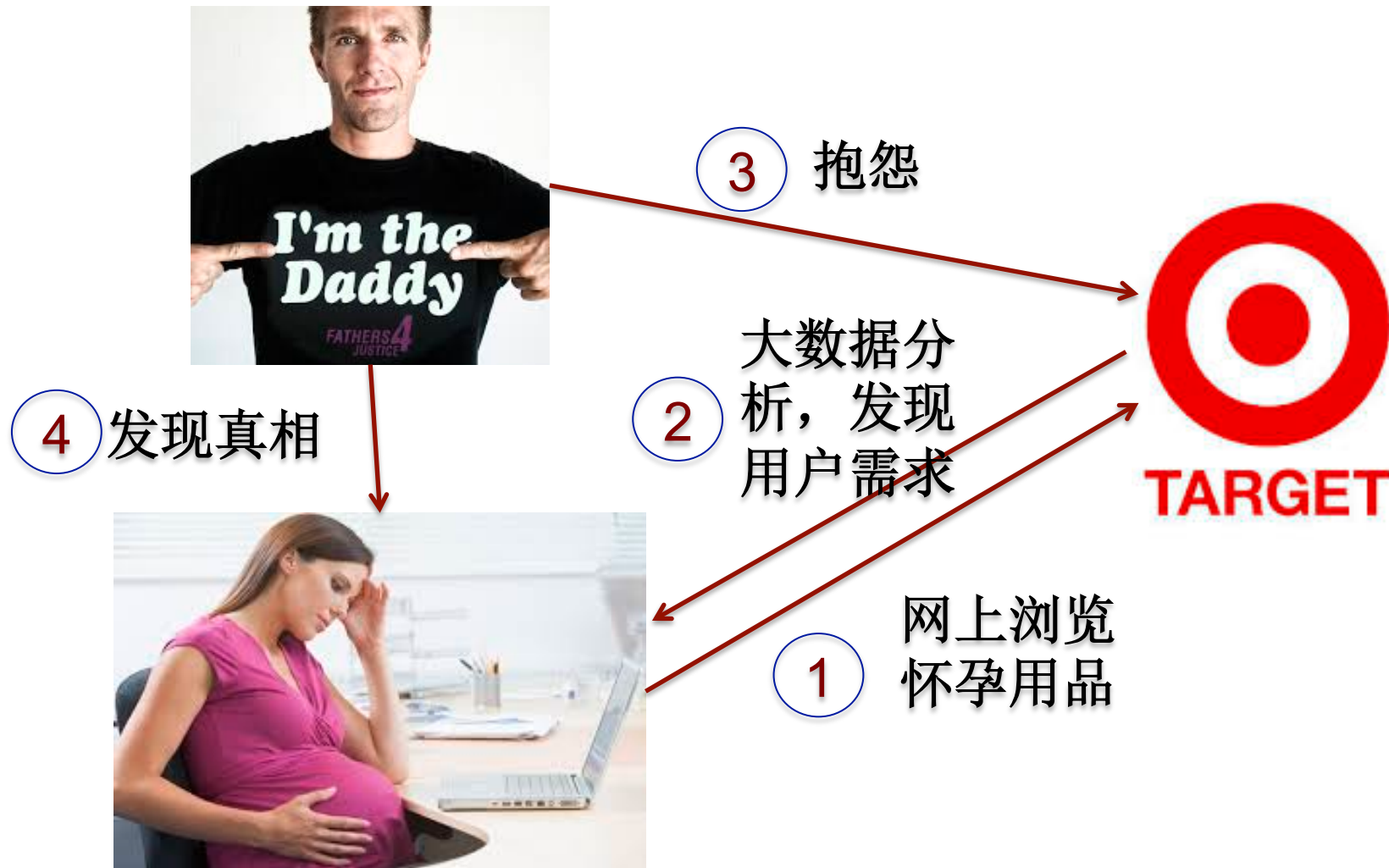
Info. Space vs. Social Space



Understanding the
mechanisms of interaction dynamics

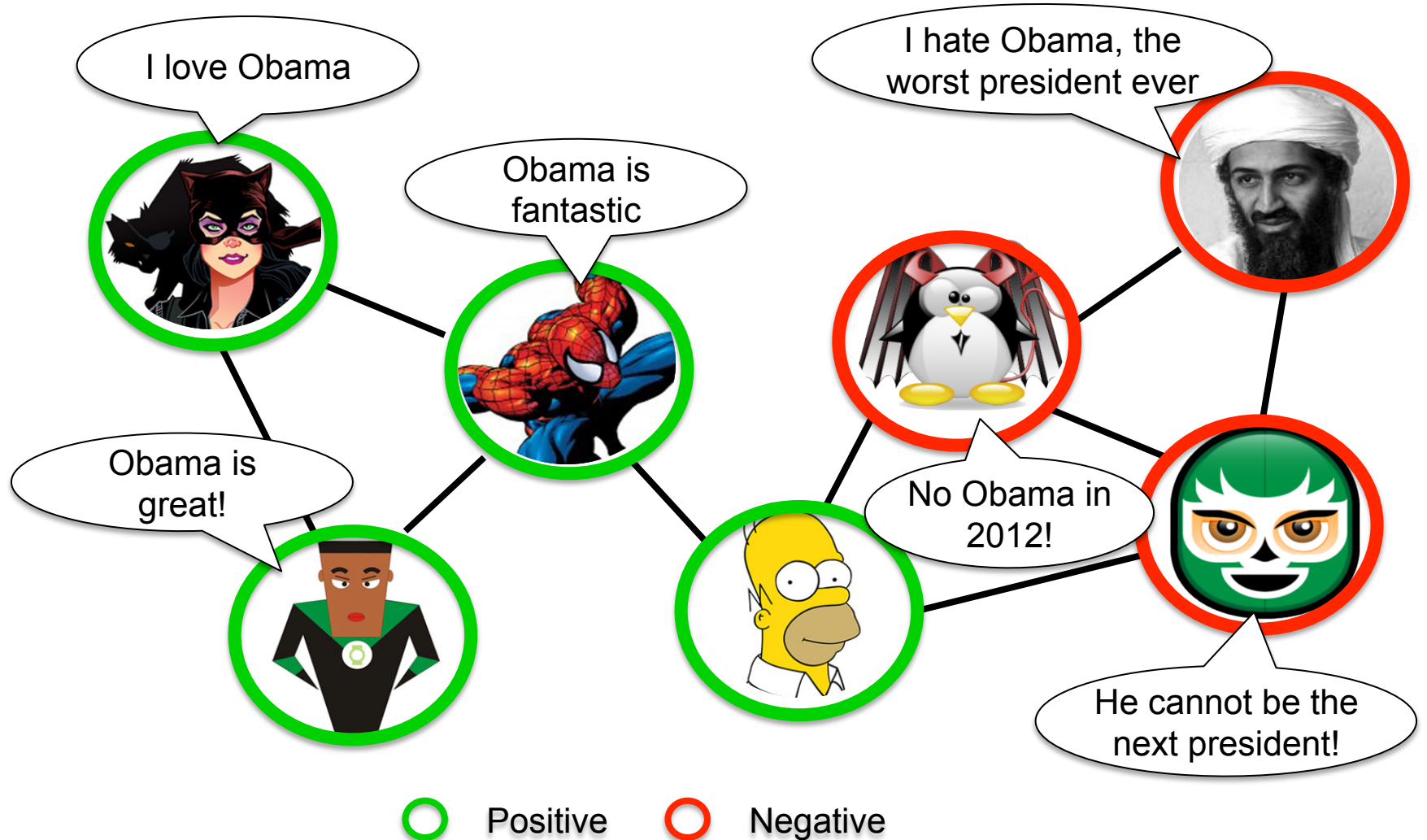


Big (Social) Data Analysis

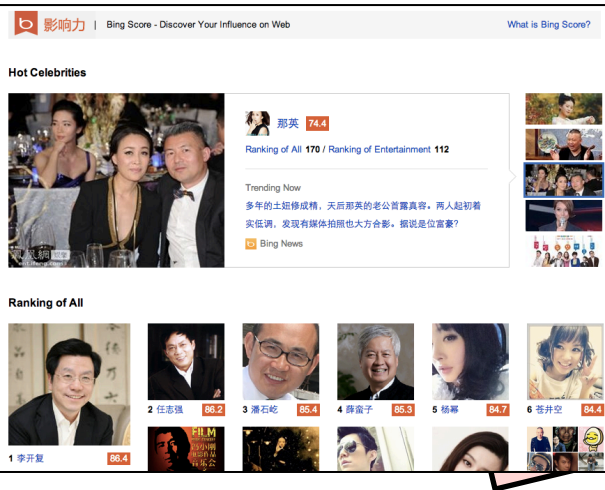


“Love Obama”

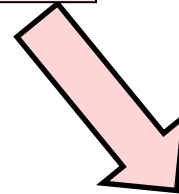
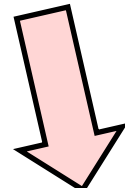
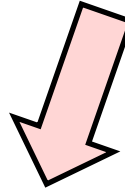
—social influence in online social networks



Revolutionary Changes



Social Networks



Search

Embedding social in search:

- Google plus
- FB graph search
- Bing's influence

Education

Human Computation:

- CAPTCHA + OCR
- MOOC
- Duolingo (Machine Translation)

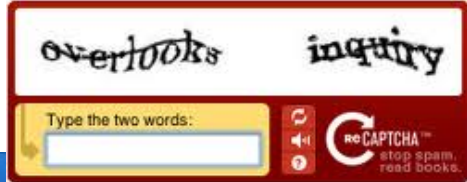
O2O

The Web knows you than yourself:

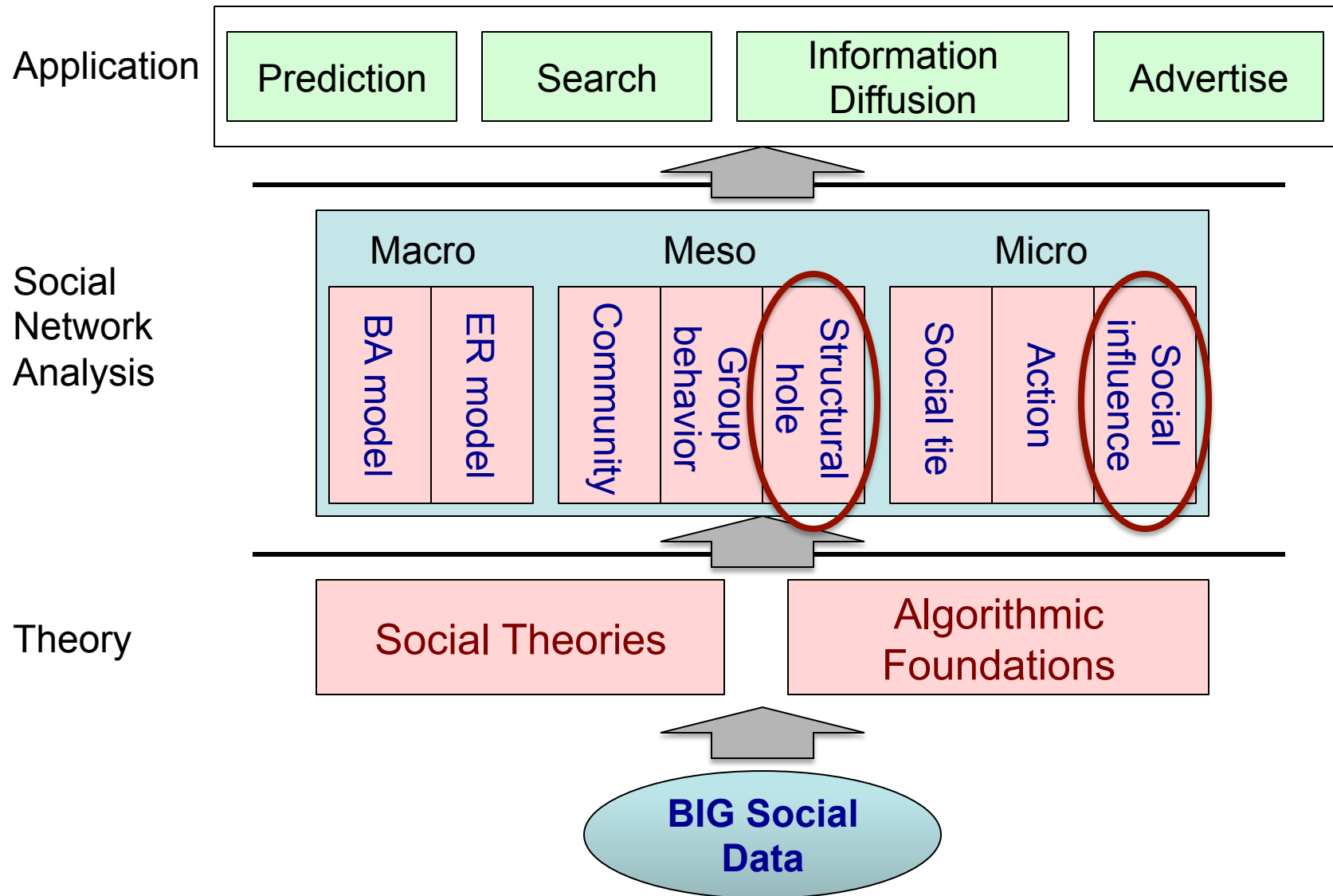
- Contextual computing
- Big data marketing

...

More ...



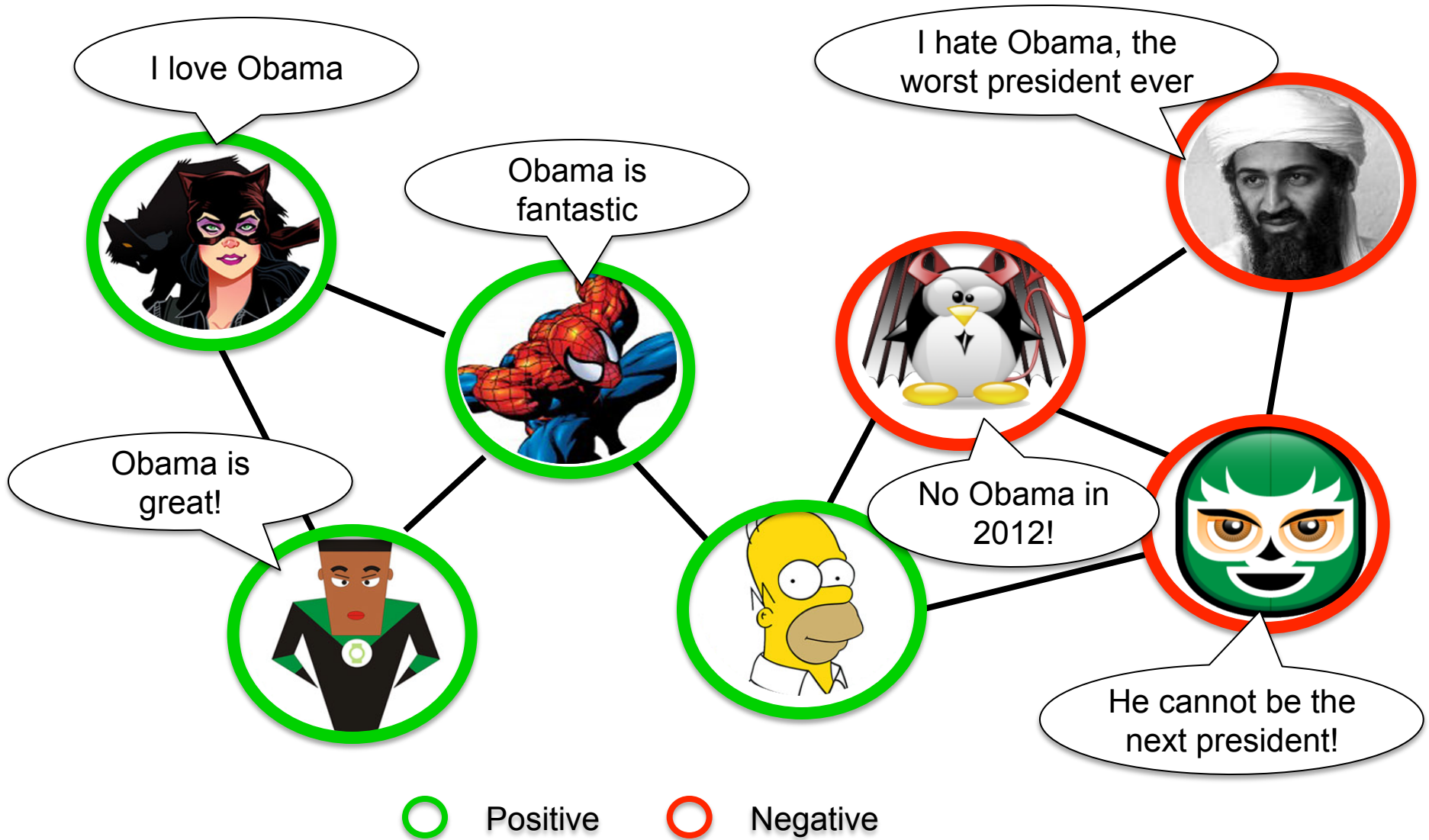
Core Research in Social Network





Social Influence

“Love Obama”



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

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.

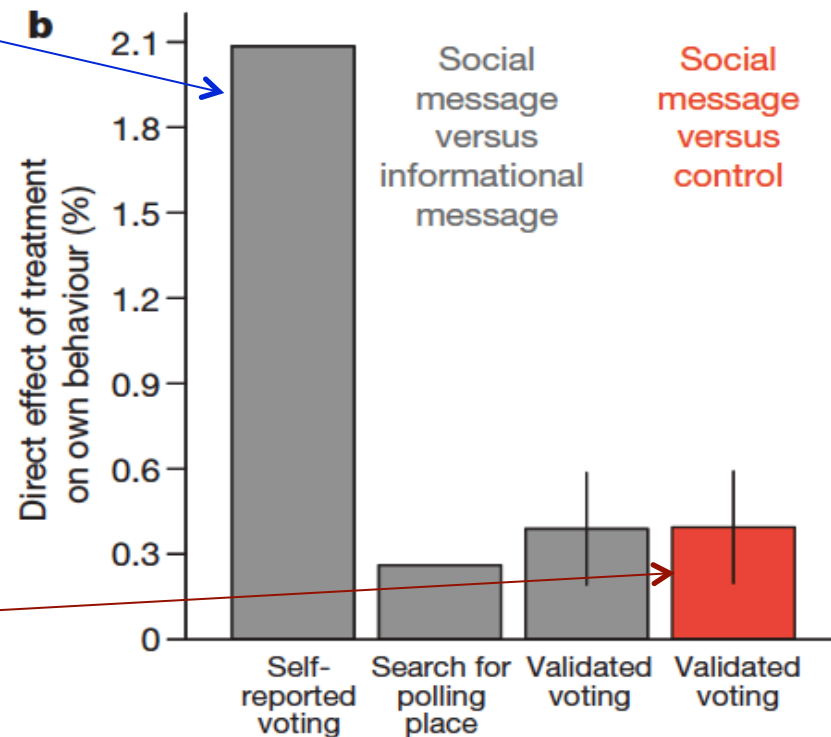
Case 1: Social Influence and Political Mobilization

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)



Our Case: Influence in Game Social Networks

- Online gaming is one of the largest industries on the Internet...
- Facebook
 - 250 million users play games monthly
 - 200 games with more than 1 million active users
 - 12% of the company's revenue is from games
- Tencent (Market Cap: ~150B \$)
 - More than 400 million gaming users
 - 50% of Tencent's overall revenue is from games

Two games: QQ Speed

- QQ Speed
 - A racing game that users can partake in competitions to play against other users
 - 200+ million users
 - Users can race against other users by individuals or form a group to race together



Task

- Given behavior log data and paying logs of online game users, predict

Free users -> Paying users

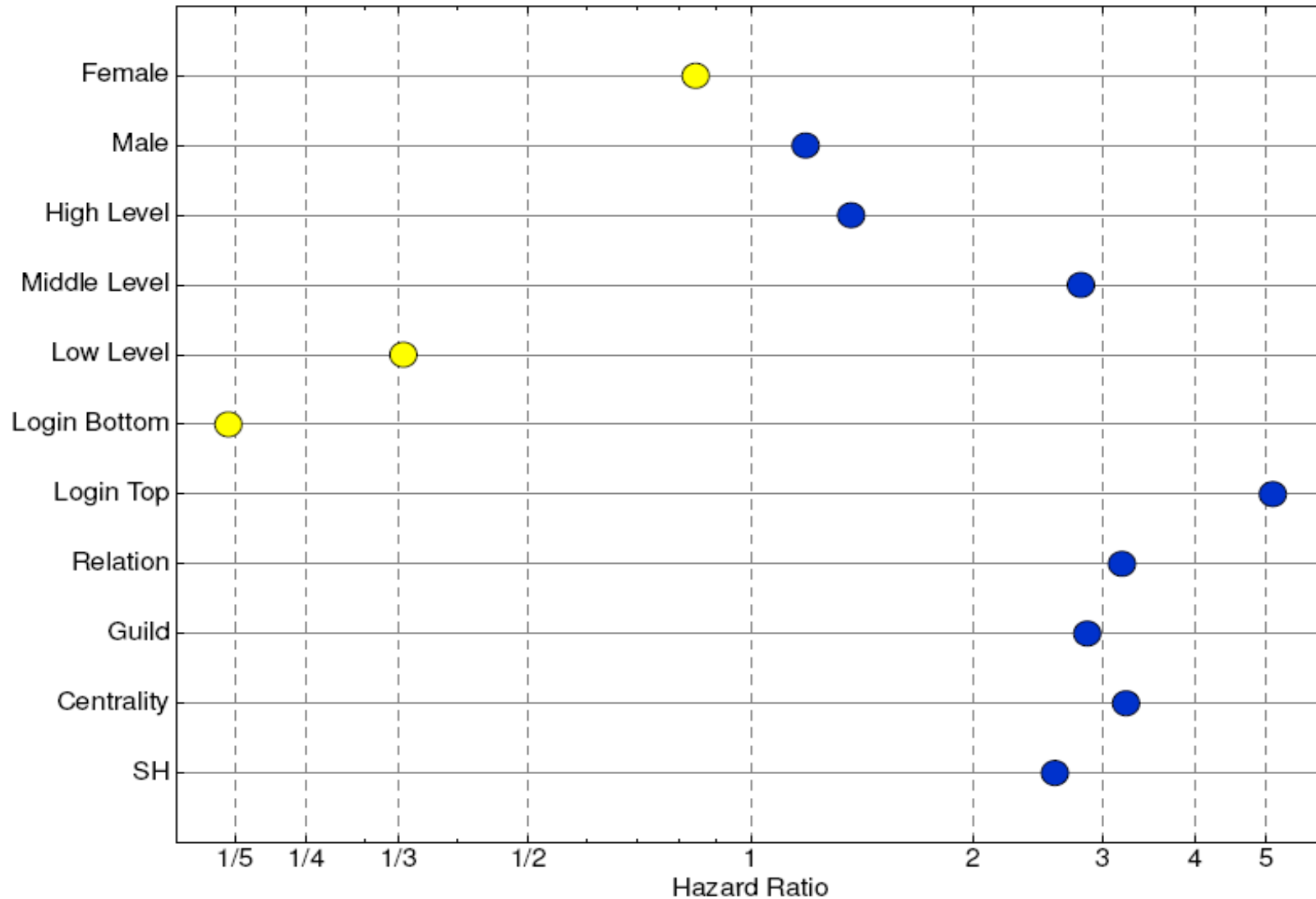
- Whether social influence will play an important role in this task?

The Big social data

- Statistics of the datasets

Category	Type	QQSpeed	DNF
User	free users	5,812,894	204,112
	paying users	1,394,630	109,099
	new payers	399,747	34,568
Relationship	co-playing	134,812,639	7,306,265
Guild	guilds	600,860	49,680
	co-guild	66,740,051	51,792,212
Activity	activity types	58	64
	activity logs	44,742,907,507	5,716,434,808
Date span	from	2013.6.20	2013.4.1
	to	2013.8.20	2013.6.30

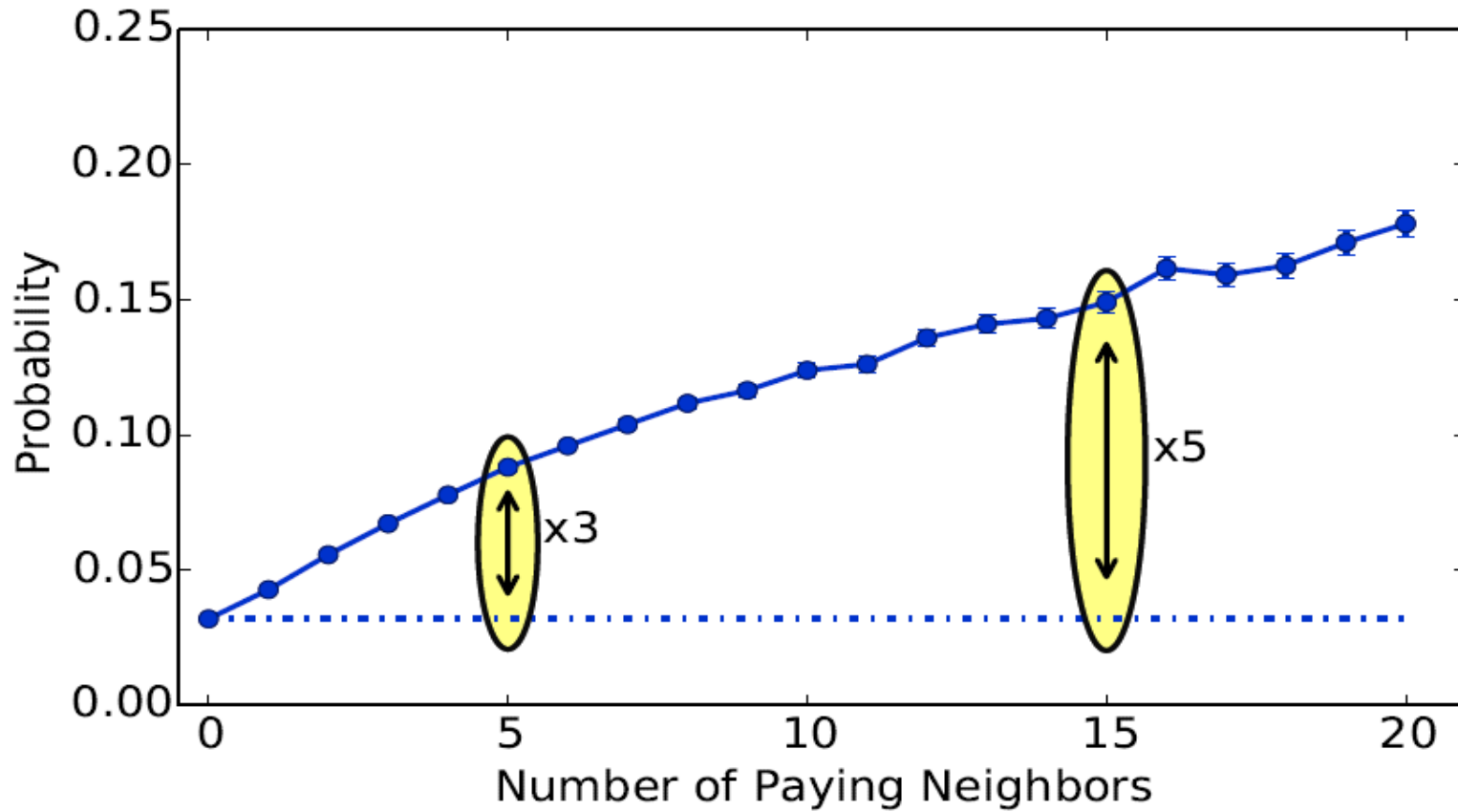
Demographics Analysis



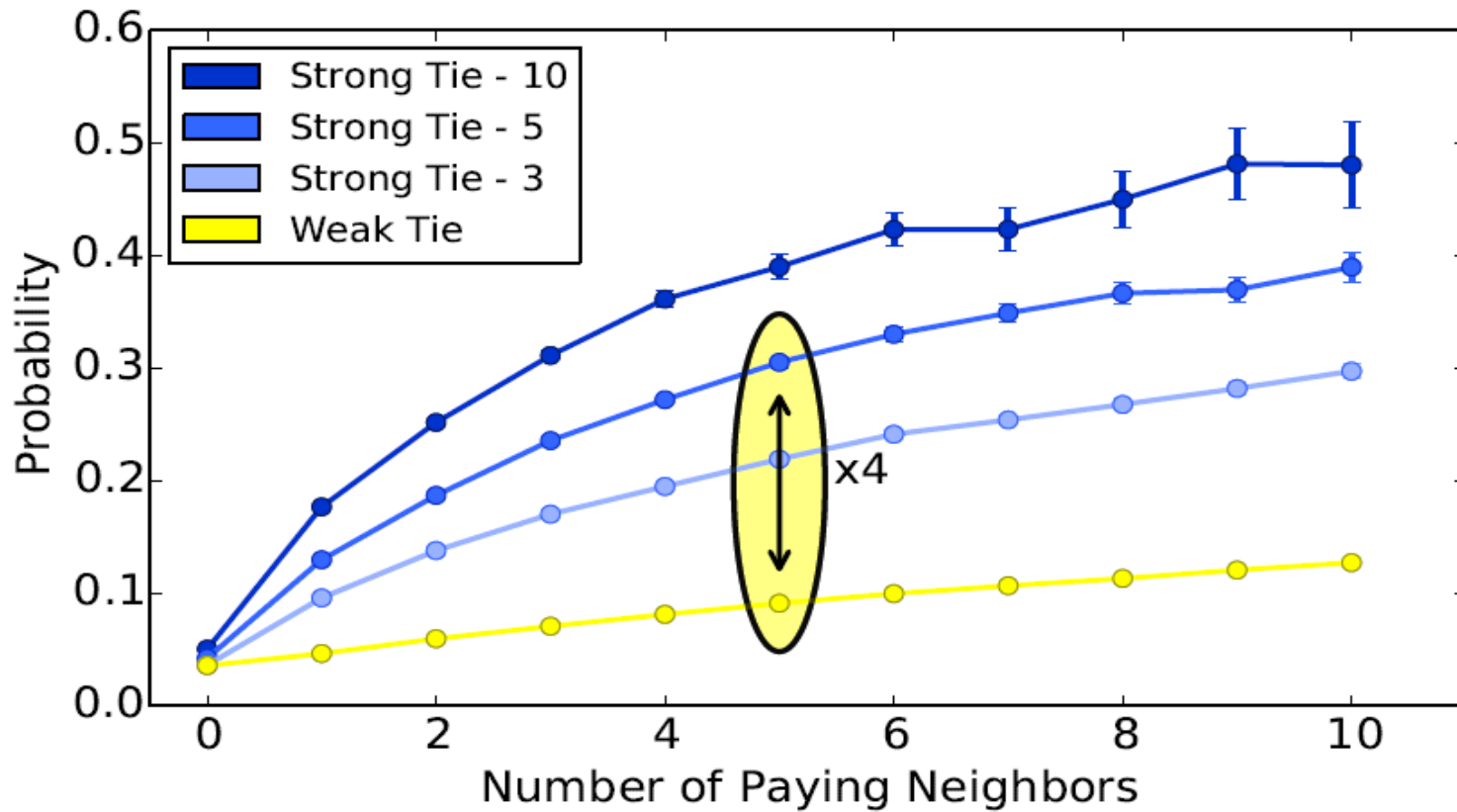
Analysis – Social influence

- Social network construction
 - Co-playing network
- Social relationship
 - Social influence
 - Strong/Weak tie
 - Status
- Structural influence

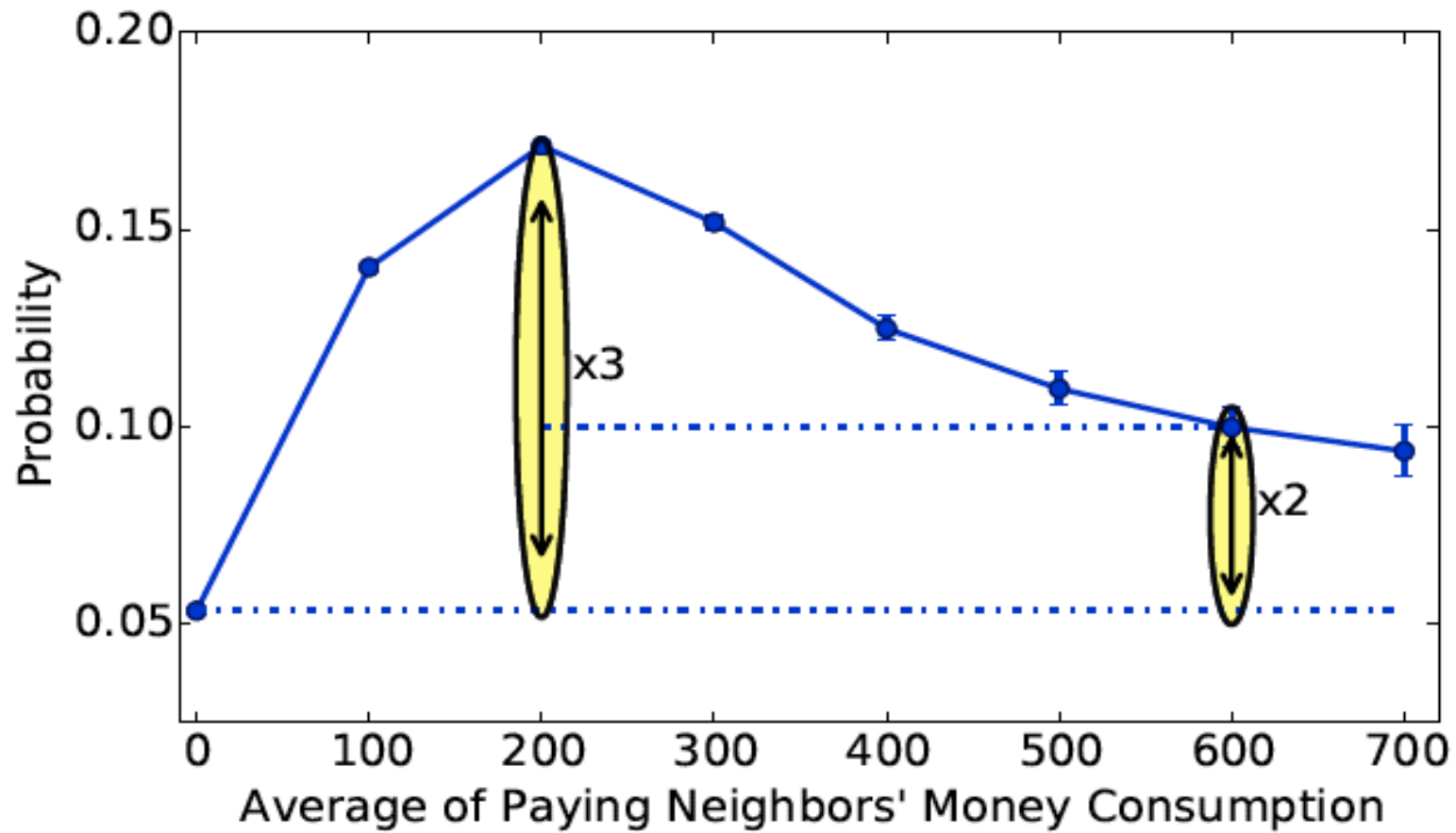
Social Influence



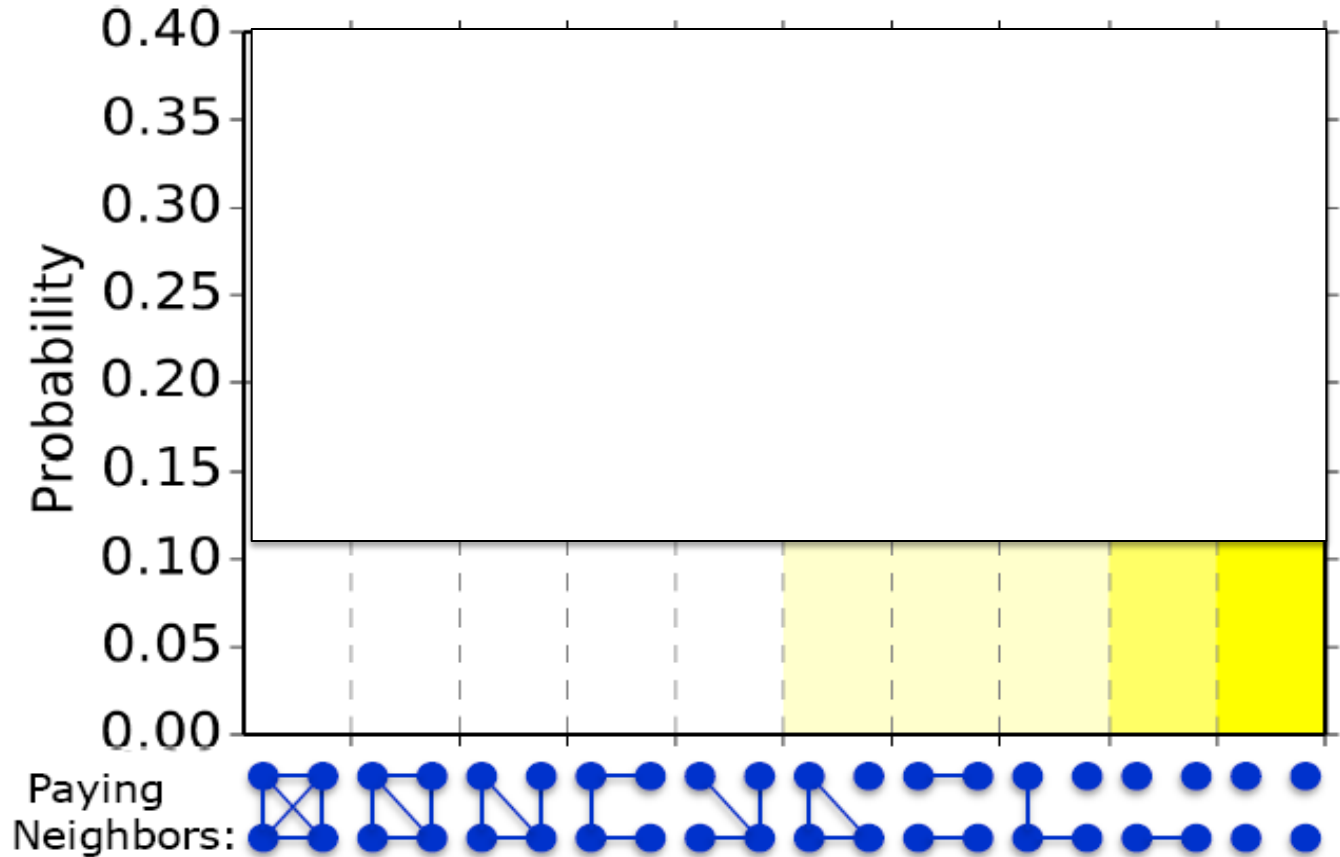
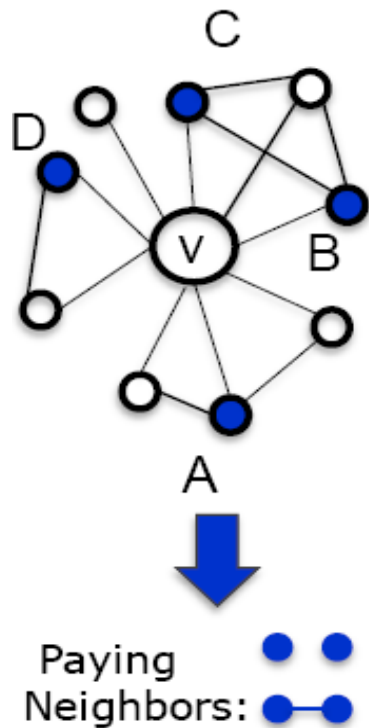
Influence + Tie Strength



Influence + Friends' Status



Structural Influence



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

Factorization Machines

- The prediction of feature vector \mathbf{x}_i :

$$y(\mathbf{x}_i) = w_0 + \sum_{j=1}^d w_j x_{ij} + \sum_{j=1}^{d-1} \sum_{j'=j+1}^d x_{ij} x_{ij'} \langle \mathbf{p}_j, \mathbf{p}_{j'} \rangle$$

- and the corresponding objective function:

$$O = \sum_{x_i} (y(\mathbf{x}_i) - y_i)^2 + \lambda \sum_{i=1}^d \|\mathbf{p}_i\|^2$$

Add the influence patterns as features into the factorization model for prediction.

Online Test

- Test setting
 - Two groups: *test group* and *control group*
 - Send msgs to invite the user to attend a promotion activity.

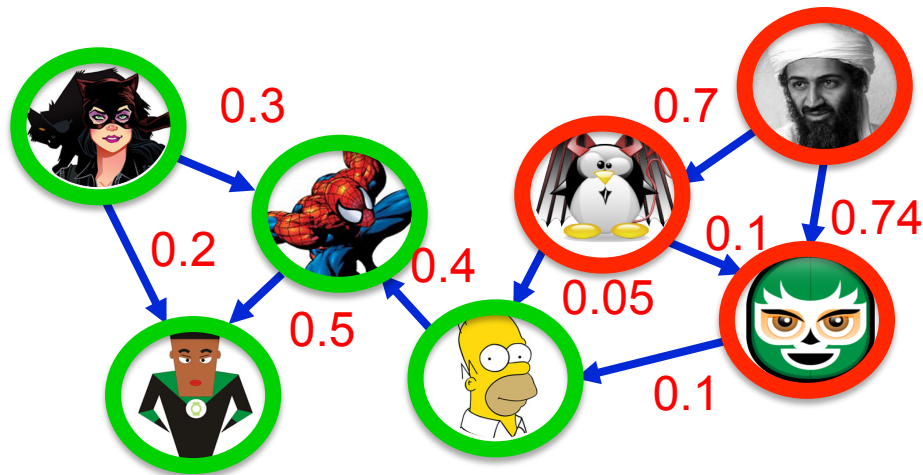
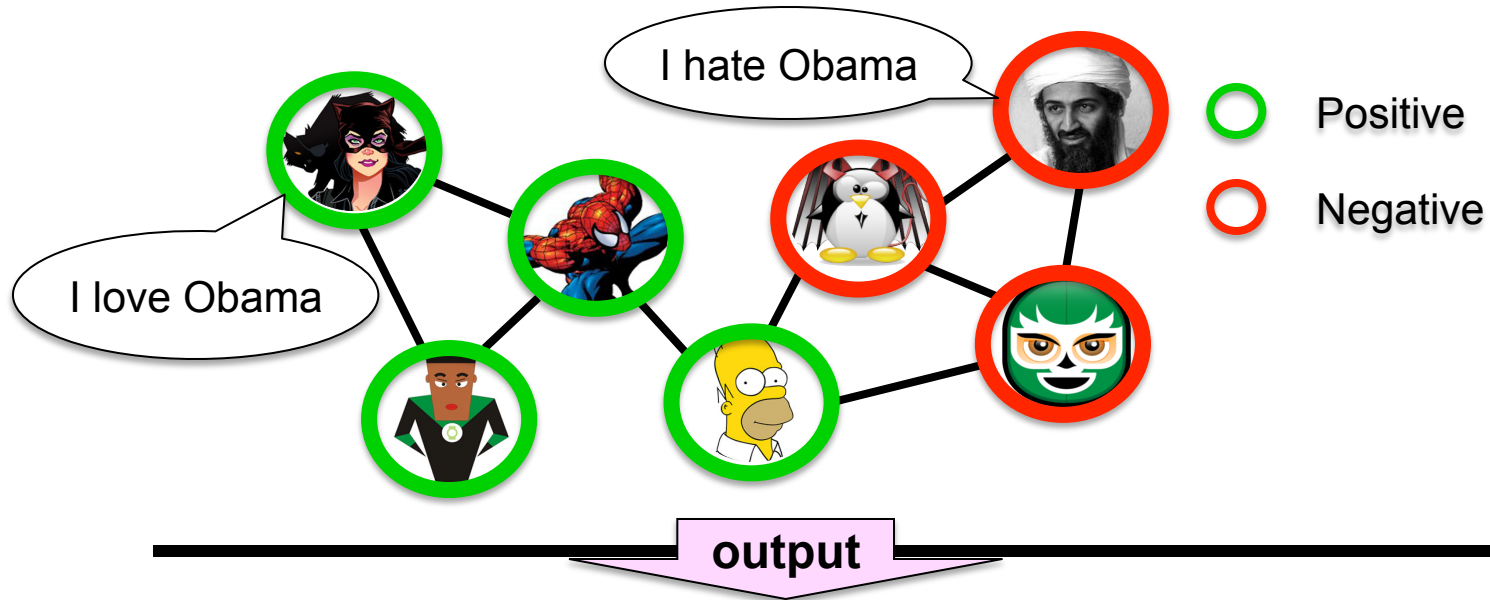


	Online Test 1 2013.12.27 - 2014.1.3		Online Test 2 2014.1.24 - 2014.1.27		
Group name	test group	control group	test group	control group	random
Group size	600K	200K	400K	400K	200K
#Message read	345K	106K	229K	215K	106K
Message read rate	57.50%	53.00%	57.25%	53.75%	53.00%
#Message clicked	47584	7466	23325	20922	6299
Message clicked rate	7.93%	3.73%	5.83%	5.23%	3.15%
Lift_Ratio	196.87%	0%	123.63%	73.40%	0%

Challenges

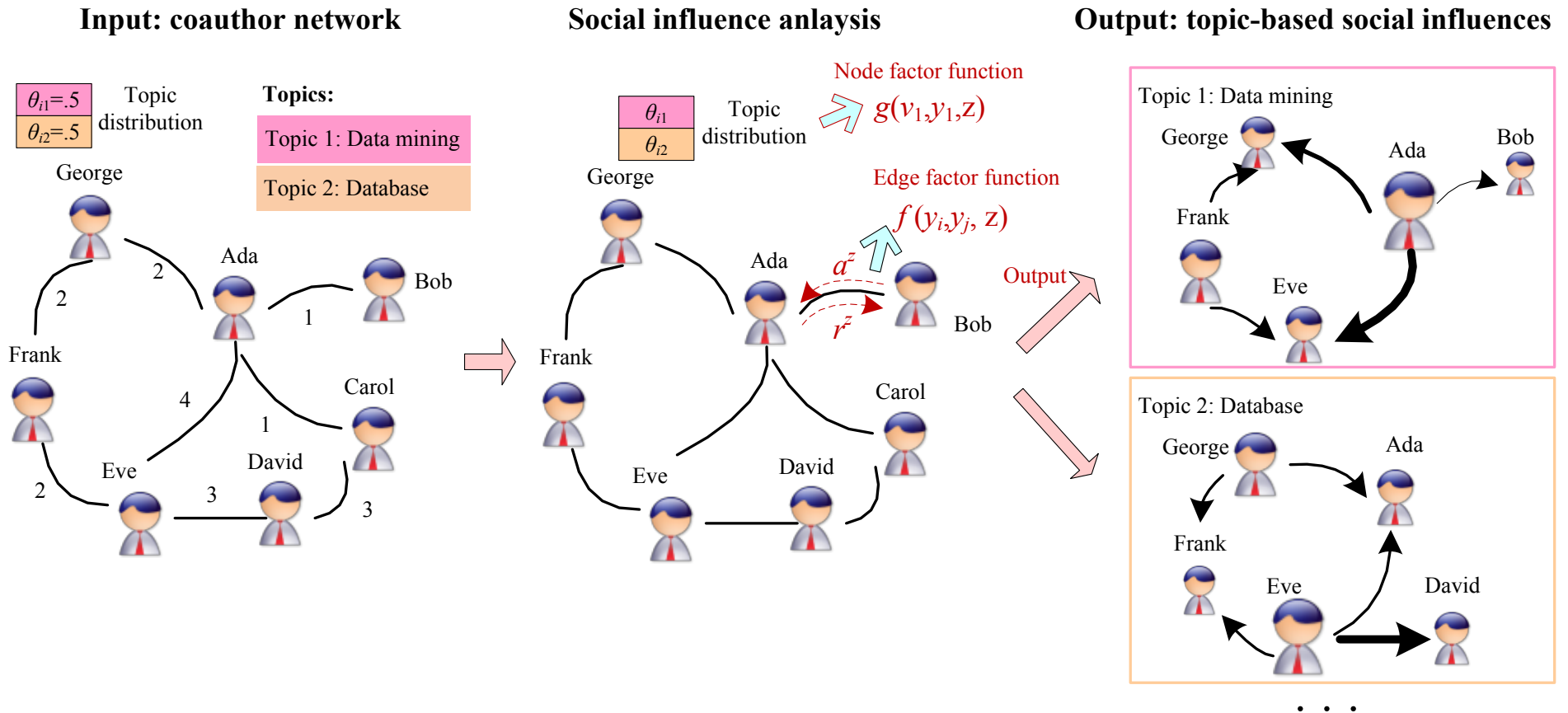
1. **How** to **measure** influence?
2. **How** to **model** influence?

Measuring Influence



Topic-based Social Influence Analysis

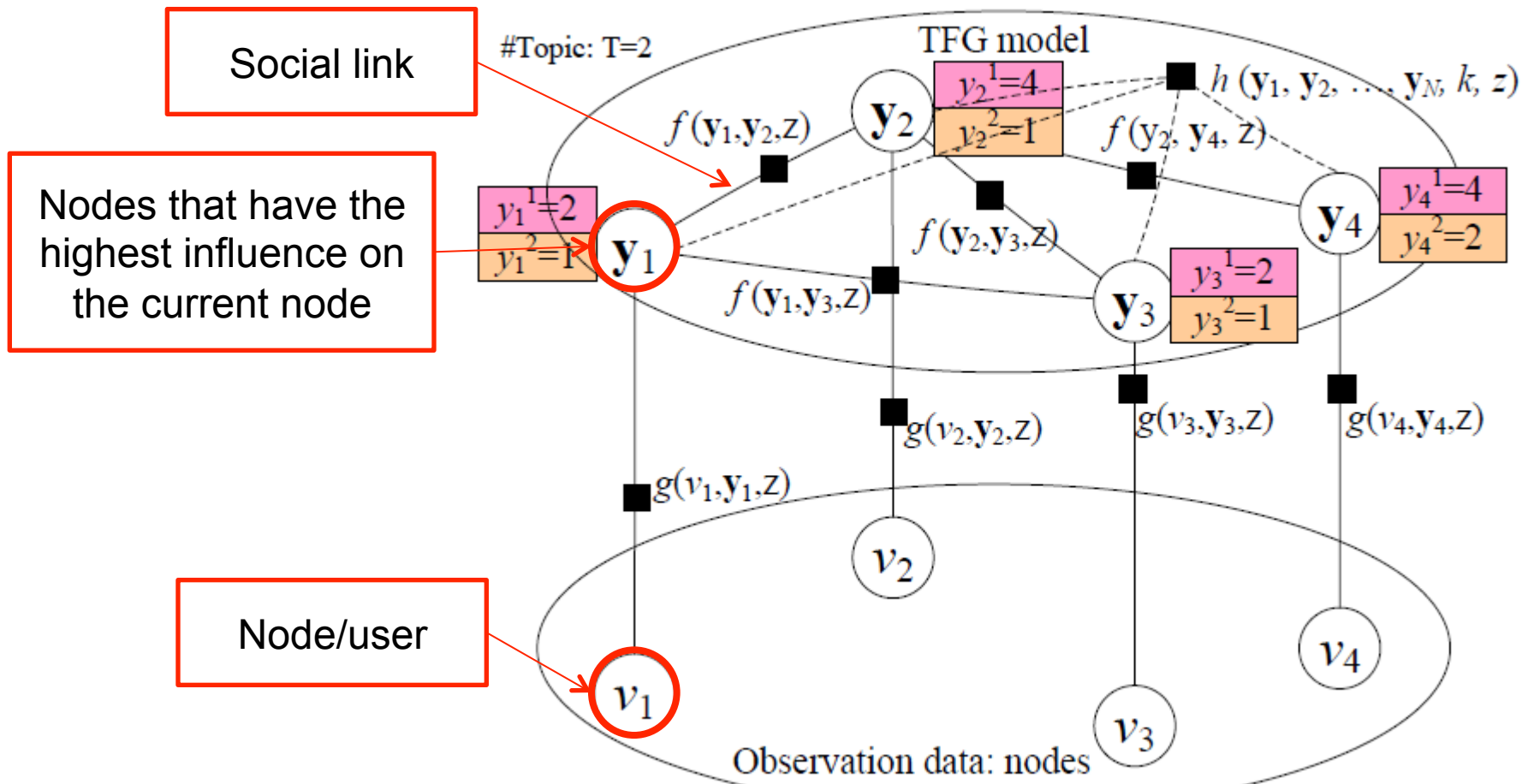
- Social network -> Topical influence network



The Solution: Topical Affinity Propagation

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

Topical Factor Graph (TFG) Model



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

Topical Factor Graph (TFG)

Objective function:

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

1. How to define?
2. How to optimize?


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

How to define (topical) feature functions?

- Node feature function

$$g(v_i, \mathbf{y}_i, z) = \begin{cases} \frac{w_{iy_i}^z}{\sum_{j \in NB(i)} (w_{ij}^z + w_{ji}^z)} & y_i^z \neq i \\ \frac{\sum_{j \in NB(i)} w_{ji}^z}{\sum_{j \in NB(i)} (w_{ij}^z + w_{ji}^z)} & y_i^z = i \end{cases}$$

similarity



- 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

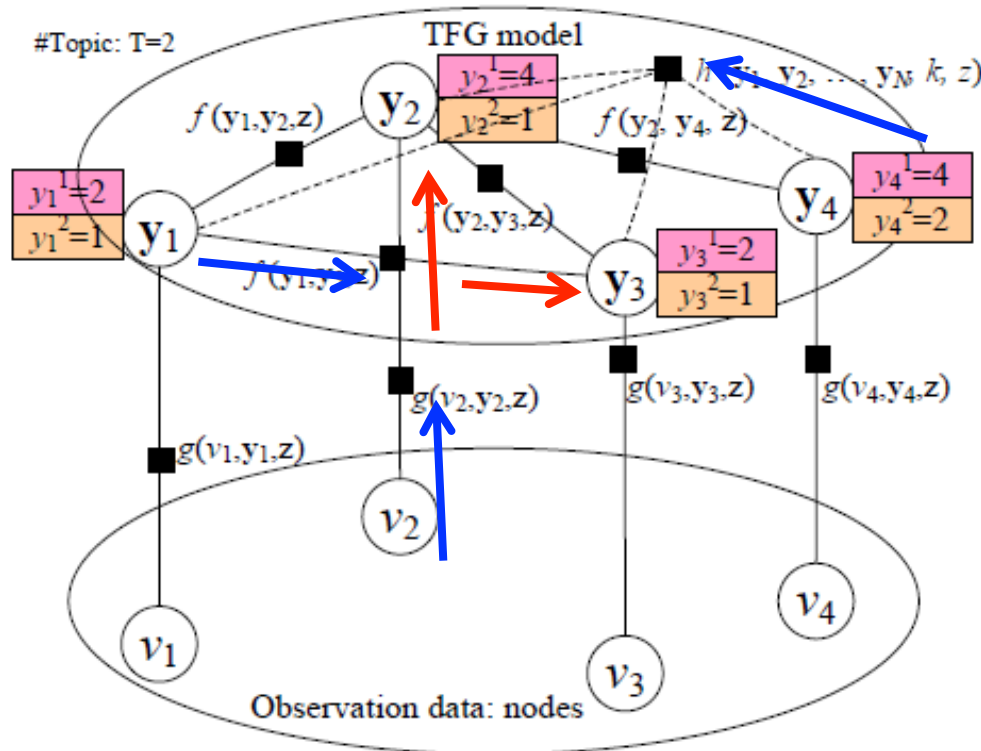
- Global feature function

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

Model Learning Algorithm

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


Sum-product: $m_{f \rightarrow y}(y, z) = \sum_{\sim \{y\}} \left(f(Y, z) \prod_{y' \sim f \setminus y} m_{y' \rightarrow f}(y', z) \right) + \sum_{z' \neq z} \tau_{z'z} \sum_{\sim \{y\}} \left(f(Y, z') \prod_{y' \sim f \setminus y} m_{y' \rightarrow f}(y', z') \right)$ (4)



- Low efficiency!
- Not easy for distributed learning!

New TAP Learning Algorithm

1. Introduce two new variables r and a , to replace the original message m .
2. Design new update rules:



A diagram showing a blue square box containing the message m_{ij} . Two blue arrows originate from the right side of the box. The top arrow points to the equation $r_{ij}^z = b_{ij}^z - \max_{k \in NB(j)} \{b_{ik}^z + a_{ik}^z\}$. The bottom arrow points to the equation $a_{jj}^z = \max_{k \in NB(j)} \min \{r_{kj}^z, 0\}$.

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

The TAP Learning Algorithm

Input: $G = (V, E)$ and topic distributions $\{\theta_v\}_{v \in V}$

Output: topic-level social influence graphs $\{G_z = (V_z, E_z)\}^T$

1.1 Calculate the node feature function $g(v_i, y_i, z)$;

1.2 Calculate b_{ij}^z according to Eq. 8;

1.3 Initialize all $\{r_{ij}^z\} \leftarrow 0$;

1.4 repeat

1.5 foreach *edge-topic pair* (e_{ij}, z) do

1.6 | Update r_{ij}^z according to Eq. 5;

1.7 end

1.8 foreach *node-topic pair* (v_j, z) do

1.9 | Update a_{jj}^z according to Eq. 6;

1.10 end

1.11 foreach *edge-topic pair* (e_{ij}, z) do

1.12 | Update a_{ij}^z according to Eq. 7;

1.13 end

1.14 until *convergence*;

1.15 foreach *node* v_t do

1.16 foreach *neighboring node* $s \in NB(t) \cup \{t\}$ do

1.17 | Compute μ_{st}^z according to Eq. 9;

1.18 end

1.19 end

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

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

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

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

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

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

Distributed TAP Learning

- Map-Reduce
 - Map: (key, value) pairs
 - $e_{ij}/a_{ij} \rightarrow e_{i^*}/a_{ij}; e_{ij}/b_{ij} \rightarrow e_{i^*}/b_{ij}; e_{ij}/r_{ij} \rightarrow e_{j^*}/r_{ij}.$
 - Reduce: (key, value) pairs
 - $e_{ij} / * \rightarrow \text{new } r_{ij}; e_{ij} / * \rightarrow \text{new } a_{ij}$
- 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 $i \in \{1, \dots, N\}$, $y_i \neq k, y'_i \neq k$, $h_k(y_1, \dots, y_i, \dots, y_N) = h_k(y_1, \dots, y'_i, \dots, y_N)$, 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

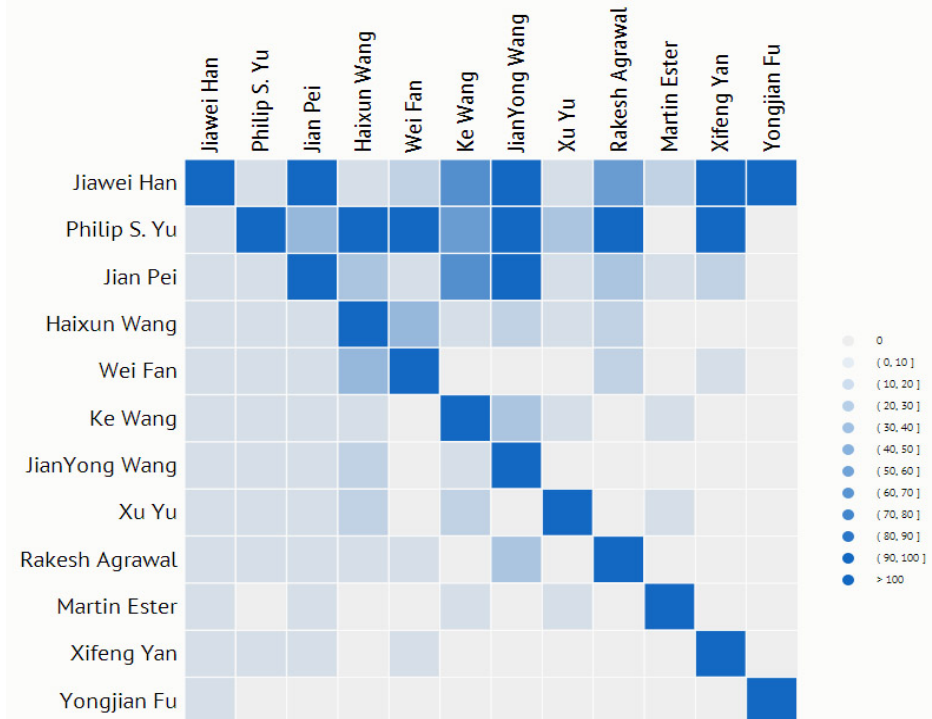
- Evaluation measures
 - CPU time
 - Case study
 - Application

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	Influence on Dr. Pei	Jiawei Han (0.4961)
-	Influenced by Dr. Pei	Jiawei Han (0.0082)
2002	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	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	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	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

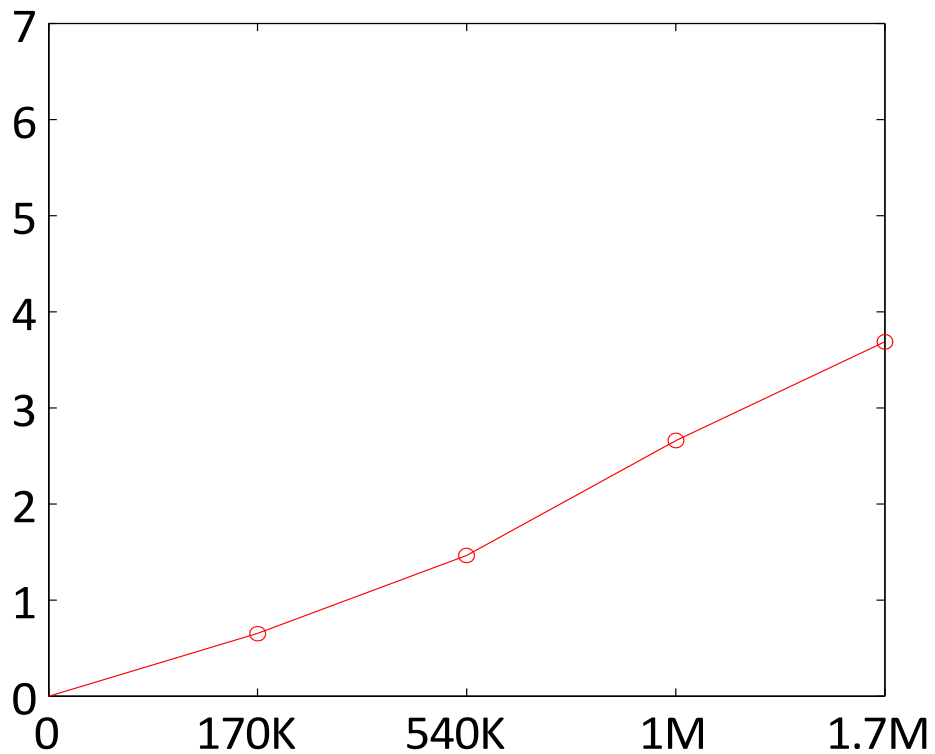
Dataset	Topic	Representative Nodes
Author	Data Mining	Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane
	Machine Learning	Pat Langley, Alex Waibel, Trevor Darrell,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

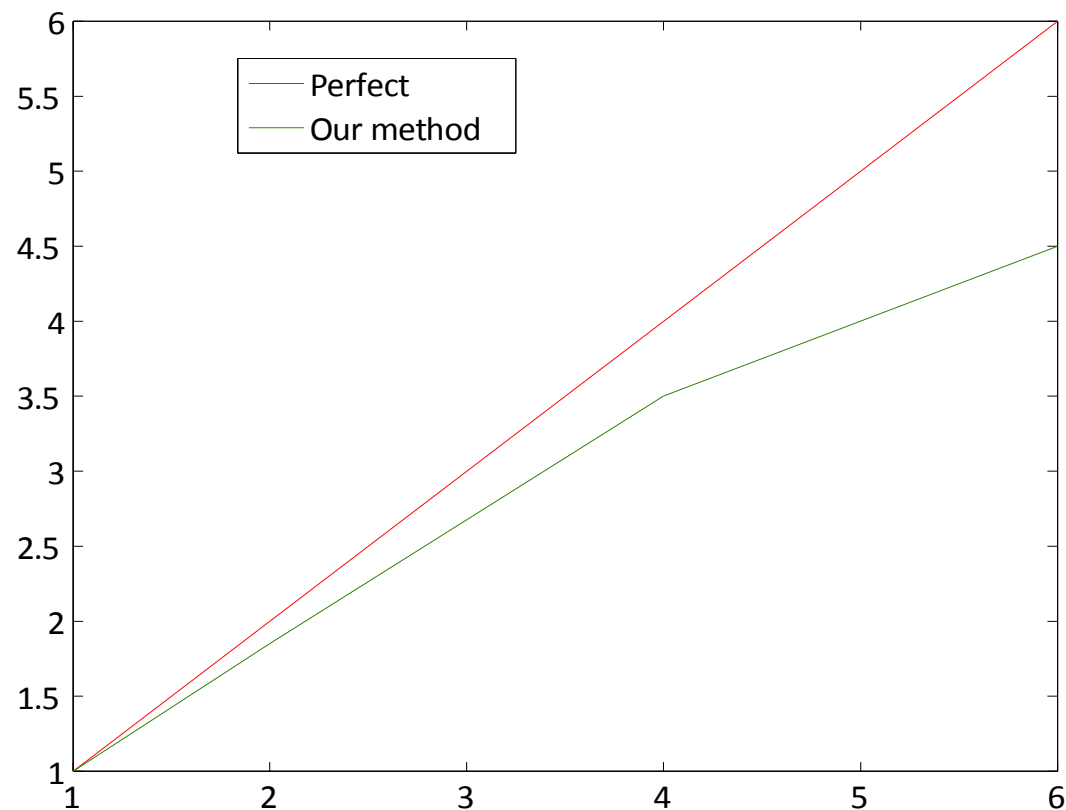
Methods	Citation	Coauthor	Film
Sum-Product	N/A	>10hr	1.8 hr
Basic TAP Learning	>10hr	369s	57s
Distributed TAP Learning	39.33m	104s	148s

Speedup results



Speedup vs. Dataset size

Speedup vs. #Computer nodes

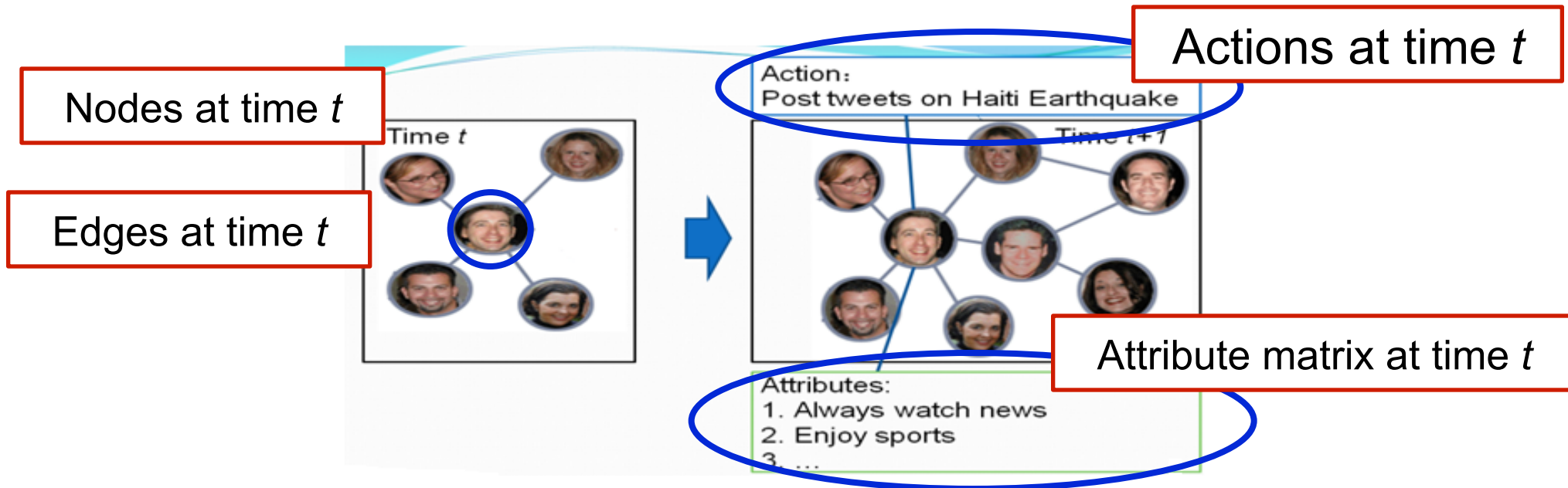


Still Challenges

1. **How** to **model** dynamic influence?
2. **How** to **distinguish** influence from other social factors?

Dynamic Influence Analysis

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



Input:

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

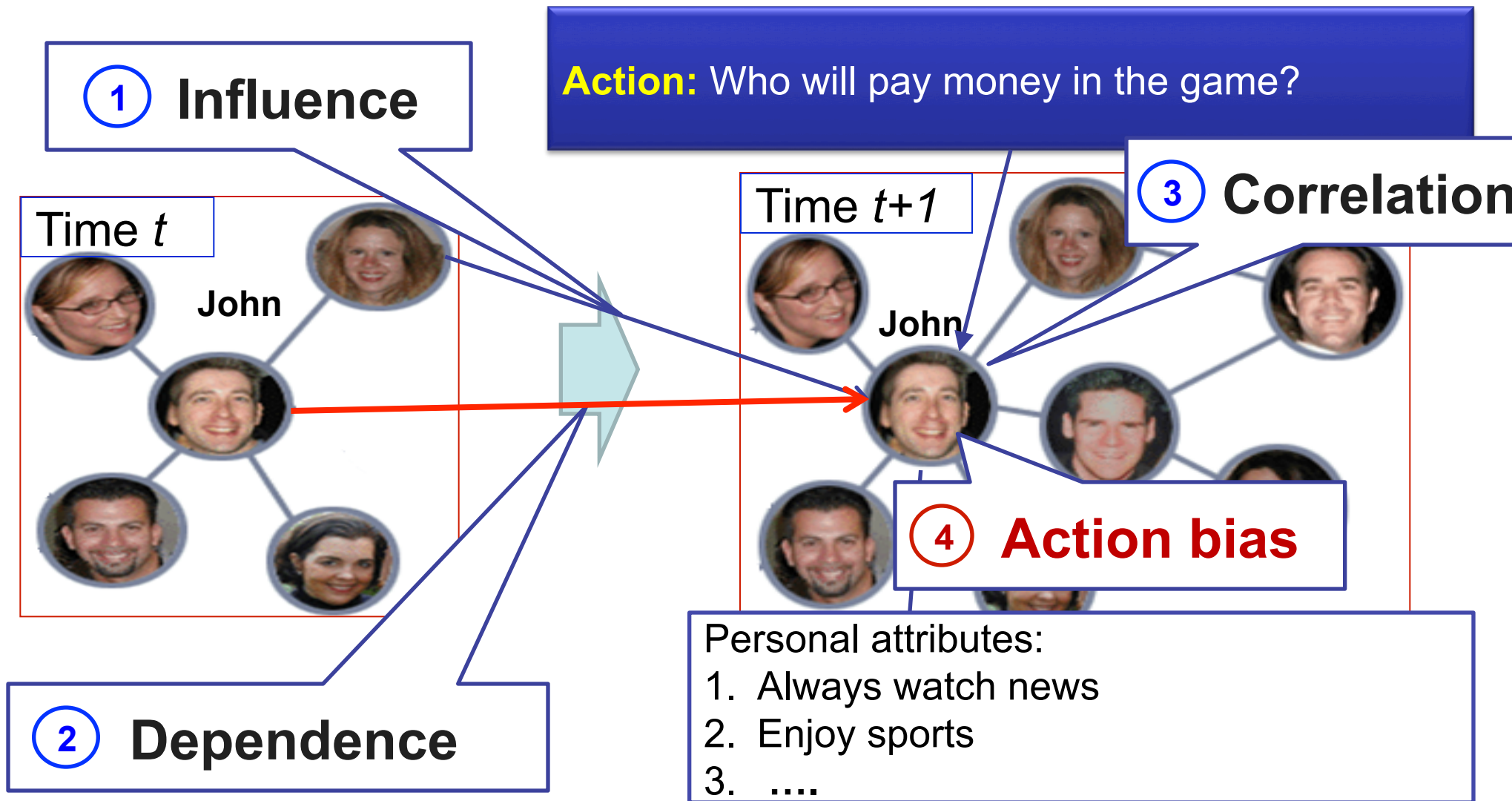
$t = 1, 2, \dots, T$



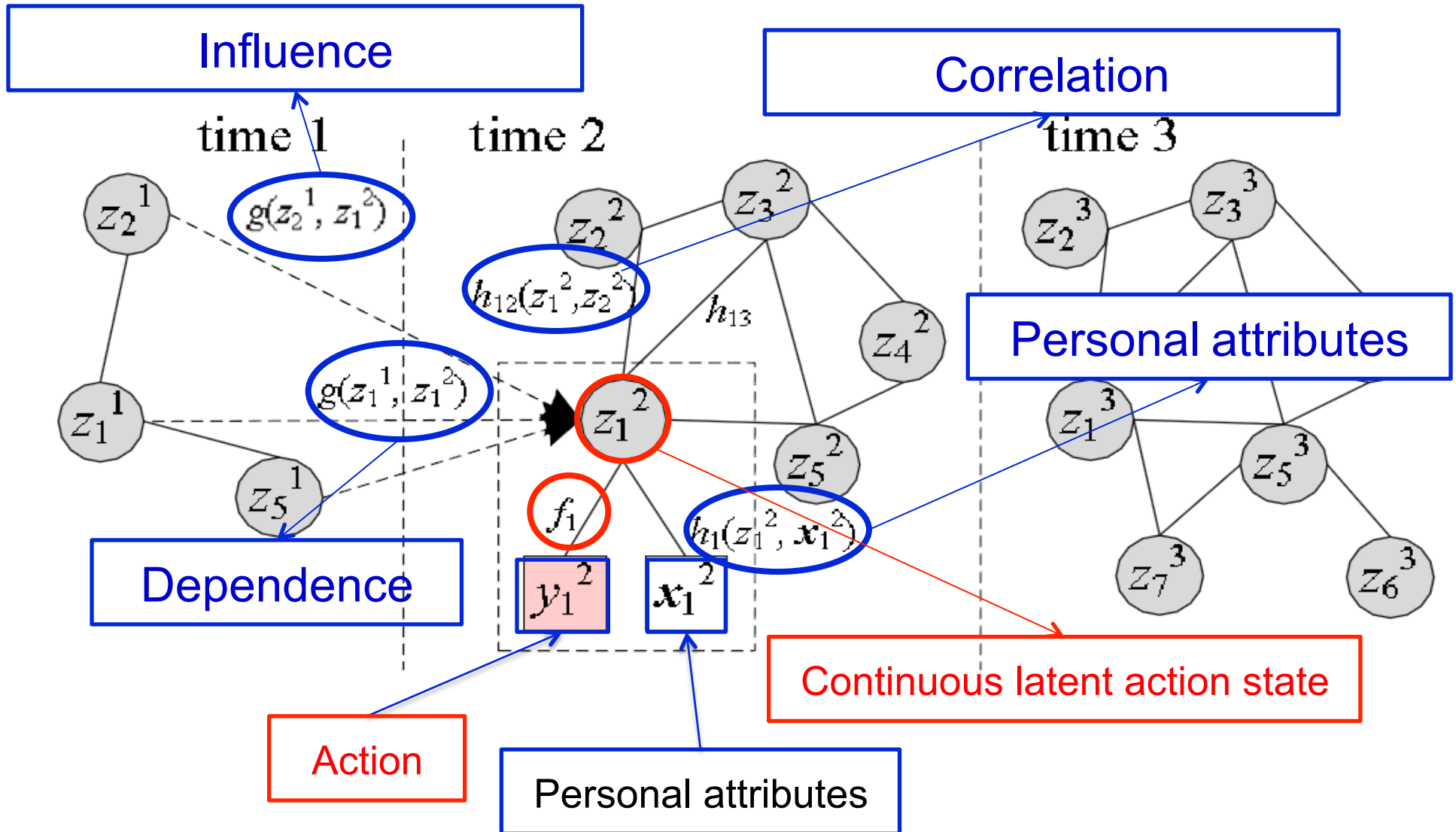
Output:

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

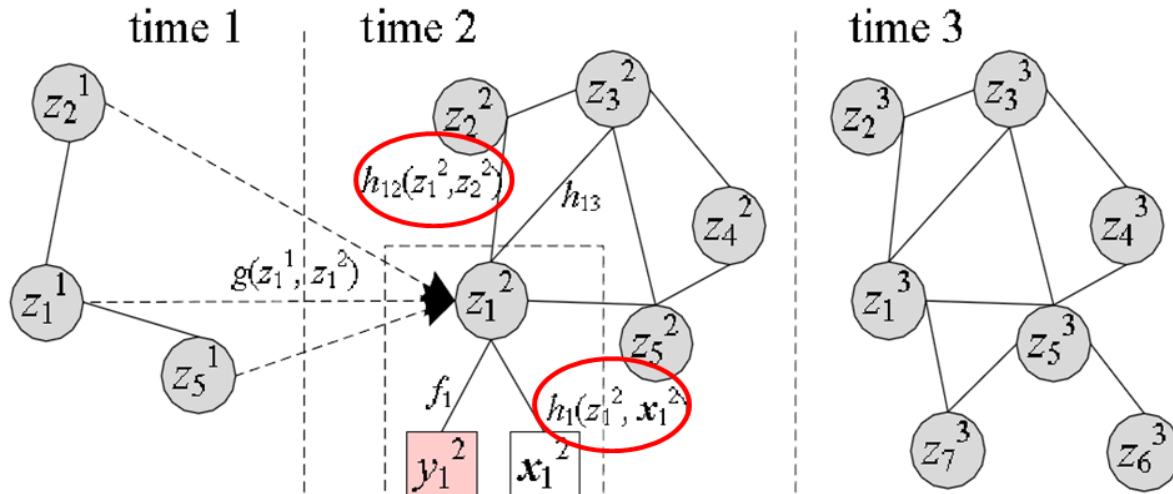
Social Influence & Action Modeling^[1]



A Discriminative Model: NTT-FGM



Model Instantiation



$$g_{ji}(z_i^t, z_j^{t-1}) = -(z_i^t - z_j^{t-1})^2$$

$$h_{ij}(z_i^t, z_j^t) = -(z_i^t - z_j^t)^2$$

$$h_k(z_i^t, x_{ik}^t) = -(z_i^t - x_{ik}^t)^2$$

How to estimate the parameters?

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

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

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:

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

until *convergence*;

Experiment

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

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

- Baseline
 - SVM
 - wvRN (Macskassy, 2003)
- Evaluation Measure:
Precision, Recall, F1-Measure

Results

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

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

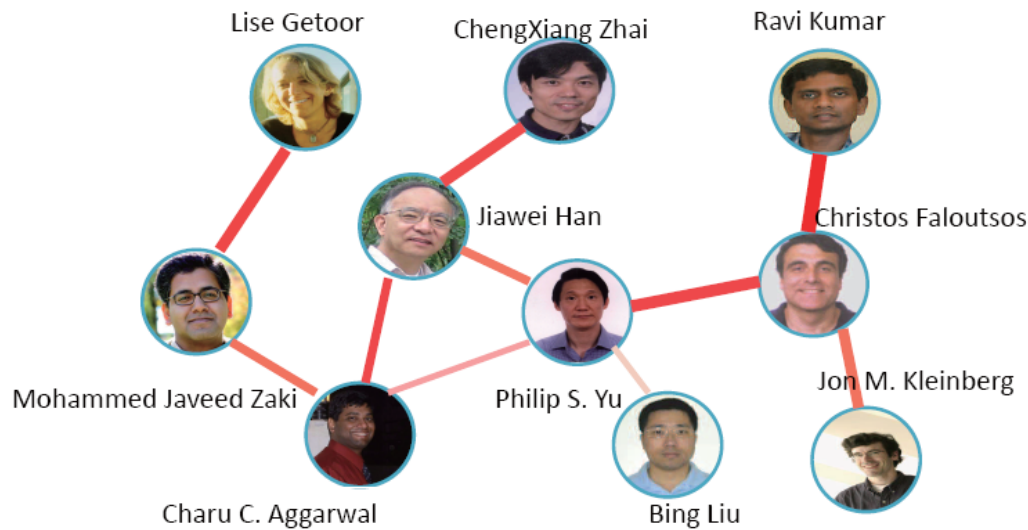
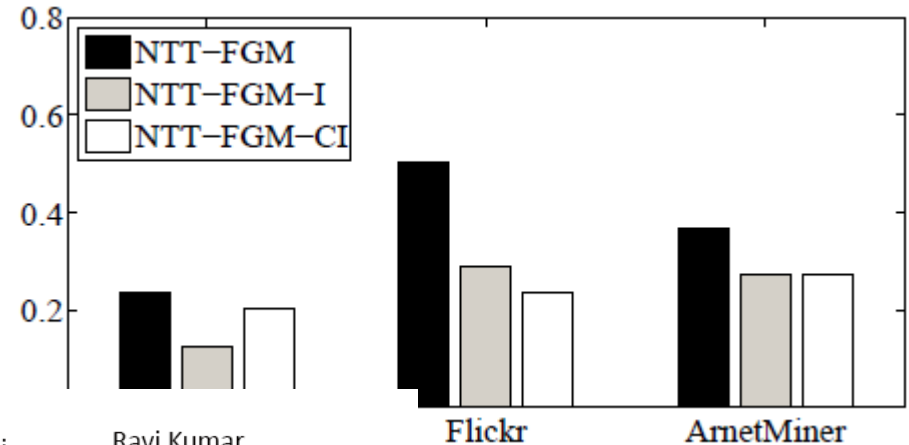


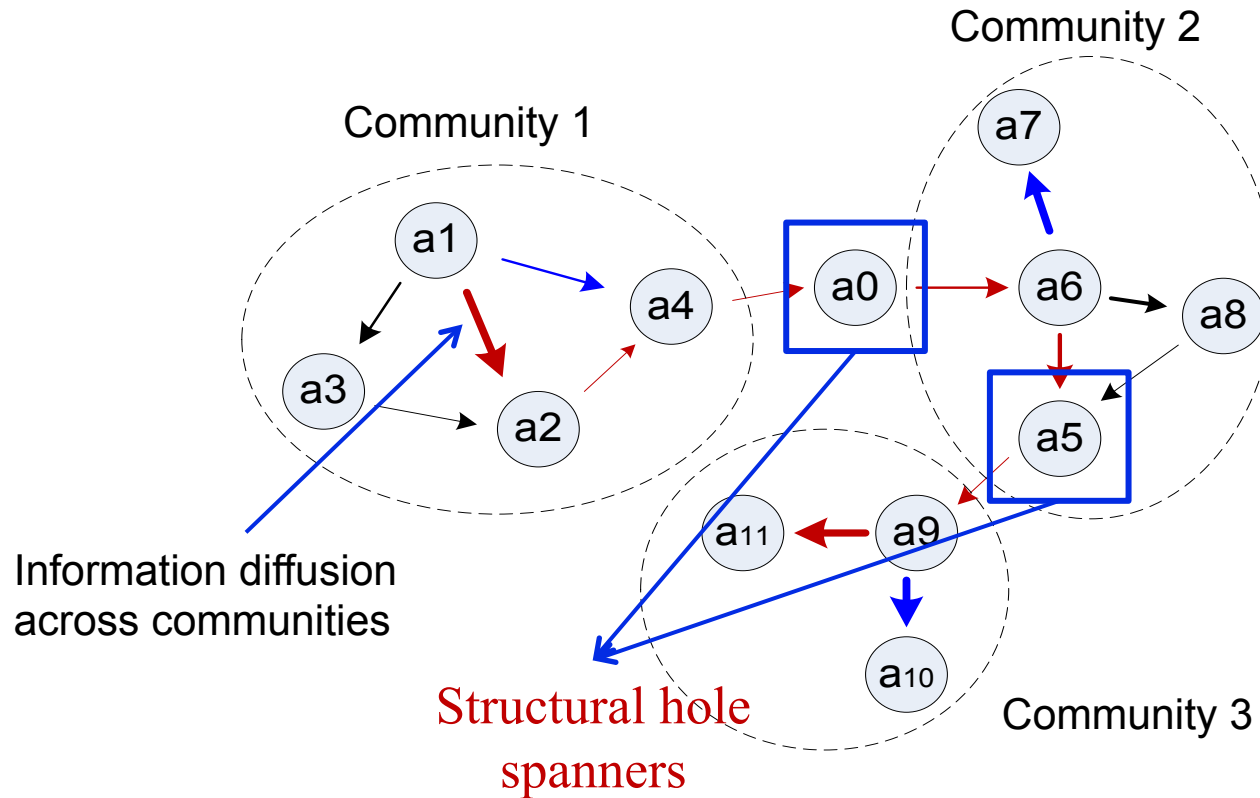
Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.



Mining structural hole spanners

Social Role:

Opinion leader vs. Structural hole spanner



twitter

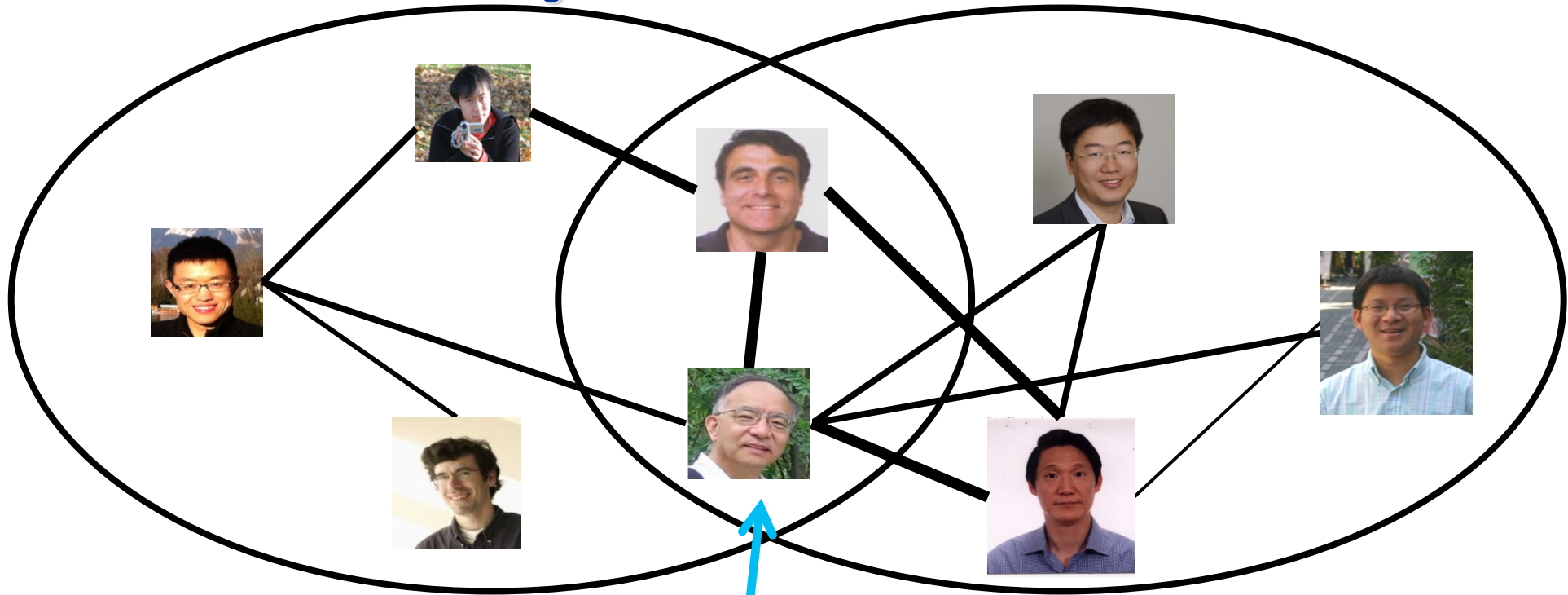
1% twitter users control
25% retweeting behaviors
between communities.

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

Examples of DBLP & Challenges

Data Mining

Database



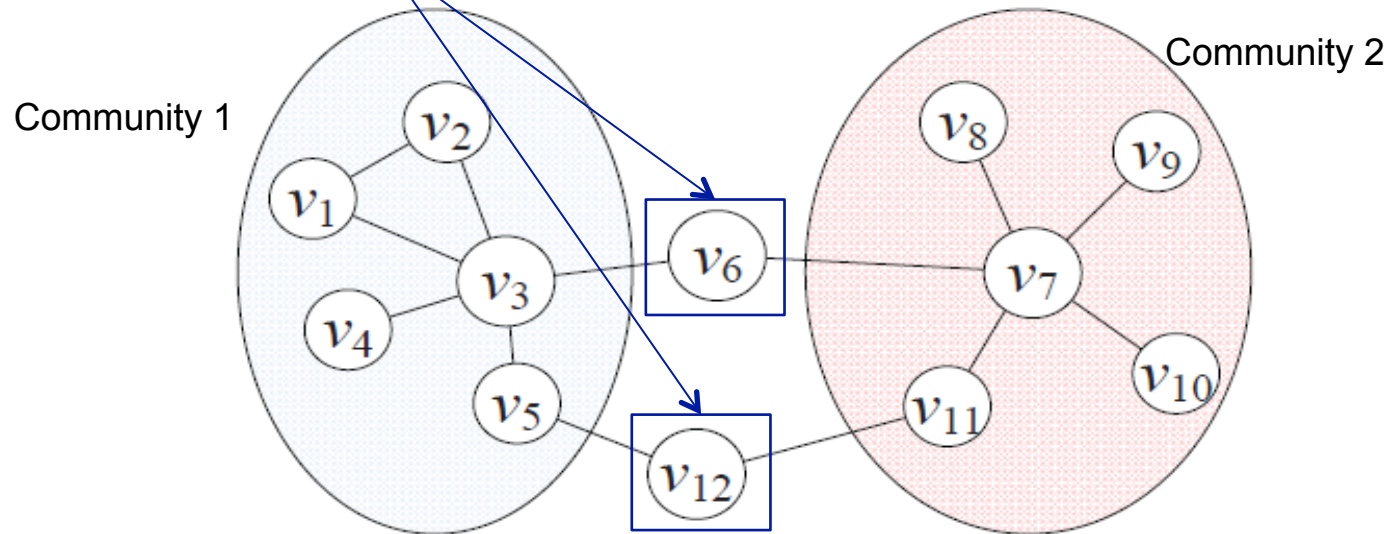
Challenge 1 : Struct
spanner vs Opinio

**82 overlapped PC members of
SIGMOD/ICDT/VLDB and
SIGKDD/ICDM during years 2007
– 2009.**

**Who control the
diffusion?**

Problem Definition

Which node is the best structural hole spanner?



Well, **mining top-k structural hole spanners** is more complex...

Problem definition

- INPUT :
 - A social network, $G = (V, E)$ and L communities $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_L)$
- Identifying top-k structural hole spanners.

$$\max Q(\mathbf{V}_{\text{SH}}, \mathbf{C}), \text{ with } |\mathbf{V}_{\text{SH}}| = k$$

Utility function $Q(\mathbf{V}^*, \mathbf{C})$:
measure \mathbf{V}^* 's degree to span
structural holes.

\mathbf{V}_{SH} : Top-k structural holes
spanners as a subset of k nodes

Data

	#User	#Relationship	#Messages
Coauthor	815,946	2,792,833	1,572,277 papers
Twitter	112,044	468,238	2,409,768 tweets
Inventor	2,445,351	5,841,940	3,880,211 patents

- In **Coauthor**, we try to understand how authors bridge different research fields (e.g., DM, DB, DP, NC, GV);
- In **Twitter**, we try to examine how structural hole spanners control the information diffusion process;
- In **Inventor**, we study how technologies spread across different companies via inventors who span structural holes.

Our first questions

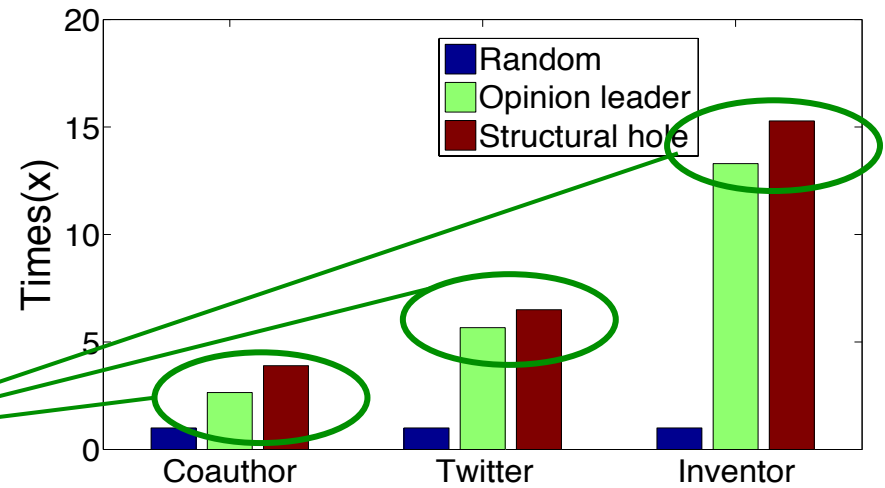
- Observable analysis
 - How likely would **structural hole spanners** connect with “**opinion leaders**” ?
 - How likely would **structural hole spanners** influence the “**information diffusion**”?

Structural hole spanners vs Opinion leaders

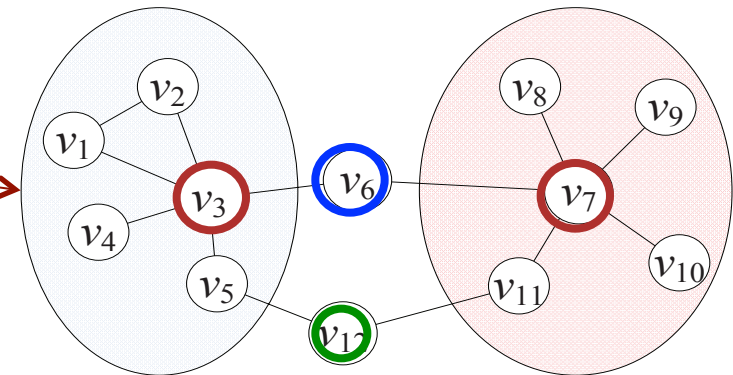
**Structural hole vs.
Opinion leader vs. Random**

Result: Structural hole spanners
are more likely to connect
important nodes

+15% - 50%

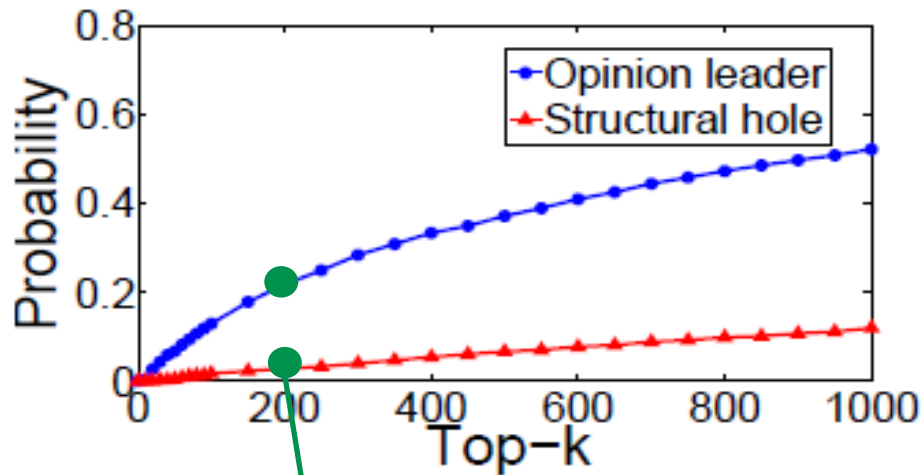


The two-step information flow theory^[1] suggests structural hole spanners are connected with many “opinion leaders”

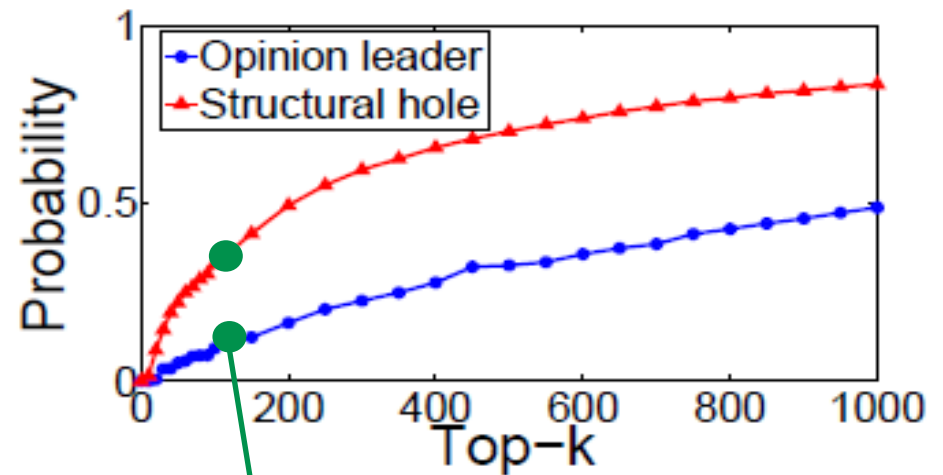


[1] E. Katz. The two-step flow of communication: an up-to-date report of an hypothesis. In Enis and Cox(eds.), Marketing Classics, pages 175–193, 1973.

Structural hole spanners control the information diffusion



(a) Inner domain



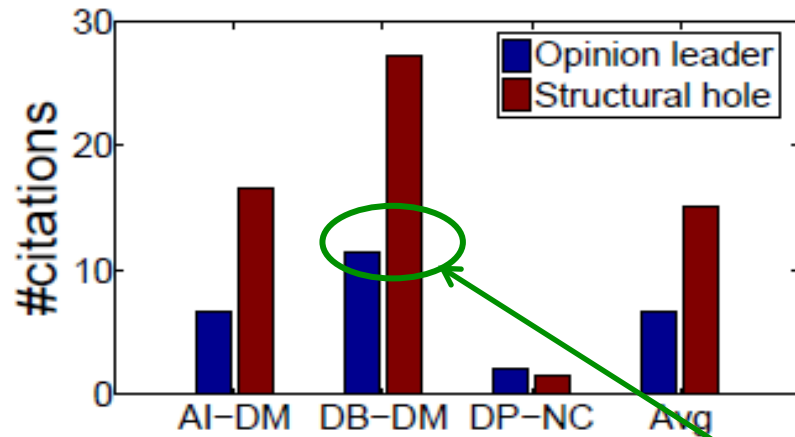
(b) Cross domain

Opinion leaders 5 times higher

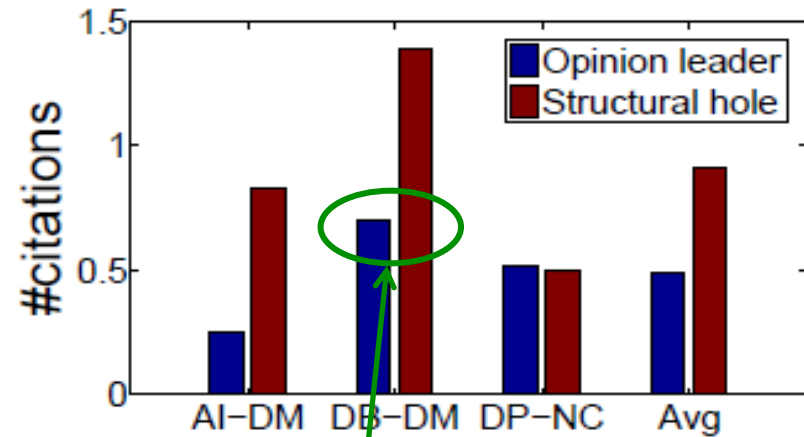
Structural hole spanners 3 times higher

Results: Opinion leaders controls information flows within communities, while Structural hole spanners dominate information spread across communities.

Structural hole spanners influence the information diffusion



(a) Cross domain



(b) Outer domain

In the **Coauthor** network :

Structural hole spanners almost **double** opinion leaders on number of **cross** domain (and **outer** domain) citations.

Intuitions

- Structural hole spanners are more likely to **connect important nodes** in different communities.



Model 1 : HIS

- Structural hole spanners **control the information diffusion** between communities.



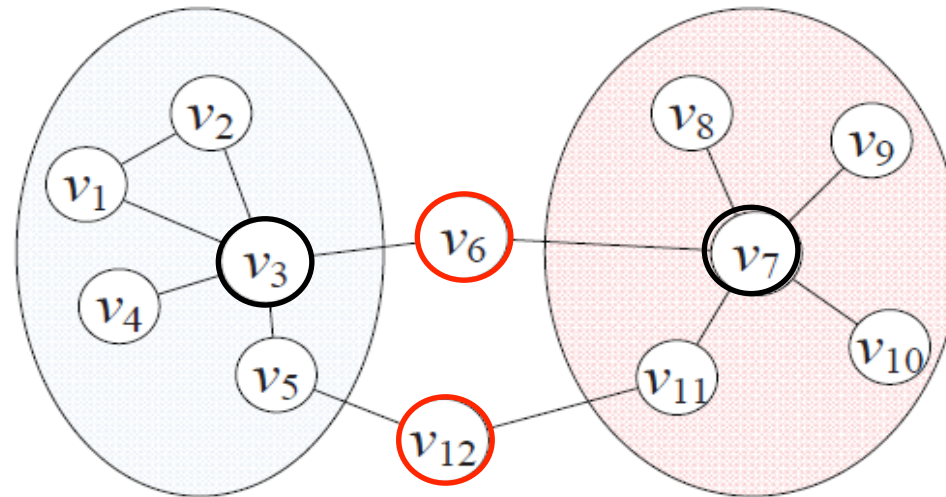
Model 2 : MaxD

Mining structural hole spanners

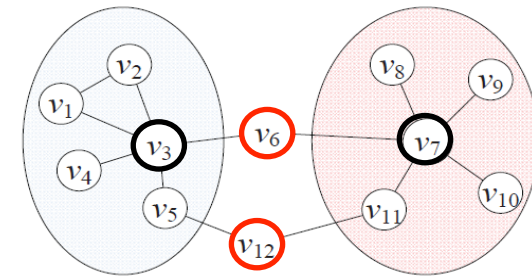
- How to design effective **models and algorithms** for **detecting structural hole spanners**?

Model One : HIS

- Structural hole spanners are more likely to connect important nodes in different communities.
 - If a user is connected with many opinion leaders in different communities, more likely to span **structural holes**.
 - If a user is connected with **structural hole spanners**, more likely to act as an opinion leader.



Model One : HIS



- Structural hole spanners are more likely to connect important nodes in different communities.
 - If a user is connected with many opinion leaders in different communities, more likely to span **structural holes**.
 - If a user is connected with **structural hole spanners**, more likely to act as an opinion leader.
- Model

$$I(v, C_i) = \max \{ I(v, C_i), \alpha_i I(v, C_i) + \beta_S H(v, S) \}$$
$$H(v, S) = \min \{ I(v, C_i) \}$$

$I(v, C_i)$: importance of v in community C_i .
 $H(v, S)$: likelihood of v spanning structural holes across S (subset of communities).

α and β are two parameters

Algorithm for HIS

Input: $G = (V, E)$, parameters α_i, β_S , and convergence threshold ϵ

Output: Importance I and structural hole score H

Initialize $I(v, C_i)$ according to Eq. 4;

repeat

 foreach $v \in V$ do

 foreach $C_i \in \mathbf{C}$ do

$P(v, C_i) =$

$\max_{S \subseteq \mathbf{C} \wedge C_i \in S} \{ \alpha_i I(v, C_i) + \beta_S H(v, S) \};$

 end

 end

 foreach $v \in V$ do

 foreach $C_i \in \mathbf{C}$ do

$I'(v, C_i) = \max \{ I(v, C_i), \max_{e_{uv} \in E} P(u, C_i) \};$

 end

 foreach $S \subseteq \mathbf{C}$ do

$H'(v, S) = \min_{C_i \in S} I'(v, C_i);$

 end

 end

 Check the ϵ -convergence condition by

$$\max_{v \in V, C_i \in \mathbf{C}} |I'(v, C_i) - I(v, C_i)| \leq \epsilon$$

 Update $I = I'$ and $H = H'$;

until Convergence;

$$I(v, C_i) = \underline{r(v)}, \quad v \in C_i$$
$$I(v, C_i) = 0, \quad v \notin C_i$$

By PageRank or
HITS

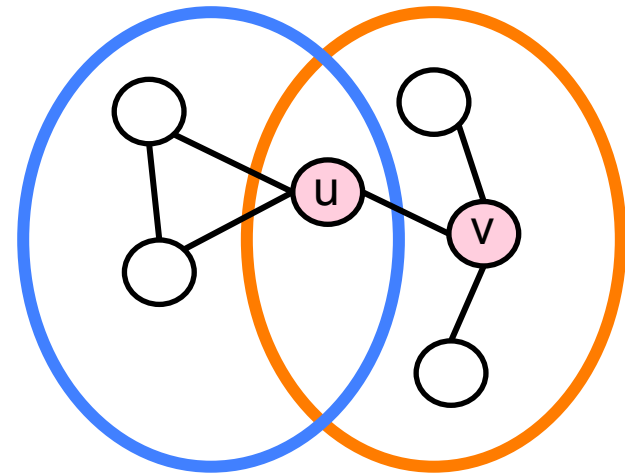
Parameter to control
the convergence

Theoretical Analysis—Existence

- Given α_i and β_S , solution exists ($I(v, C_i), H(v, S) \leq 1$) for any graph, **if and only if, $\alpha_i + \beta_S \leq 1$.**

– For the *only if* direction

- Suppose $\alpha_i + \beta_S > 1, S = \{C_{\text{blue}}, C_{\text{yellow}}\}$
- $r(u) = r(v) = 1;$
- $I(u, C_{\text{blue}}) = I(u, C_{\text{yellow}}) = 1;$
- $H(u, S) = \min \{ I(u, C_{\text{blue}}), I(u, C_{\text{yellow}}) \} = 1;$
- $I(v, C_{\text{yellow}}) \geq \alpha_i I(u, C_i) + \beta_S H(u, S) = \alpha_i + \beta_S > 1$



$$I(v, C_i) = \max \{ I(v, C_i), \alpha_i I(u, C_i) + \beta_S H(u, S) \}$$

$$H(v, S) = \min \{ I(v, C_i) \}$$

Theoretical Analysis—Existence

- Given α_i and β_S , solution exists ($I(v, C_i), H(v, S) \leq 1$) for any graph, **if and only if, $\alpha_i + \beta_S \leq 1$.**
 - For the *if* direction
 - If $\alpha_i + \beta_S \leq 1$, we use induction to prove $I(v, C_i) \leq 1$;
 - Obviously $I^{(0)}(v, C_i) \leq r(v) \leq 1$;
 - Suppose after the k -th iteration, we have $I^{(k)}(v, C_i) \leq 1$;
 - Hence, in the $(k + 1)$ -th iteration, $I^{(k+1)}(v, C_i) \leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k)}(u, S) \leq (\alpha_i + \beta_S) I^{(k)}(u, C_i) \leq 1$.

Theoretical Analysis—Convergence

- Denote $\gamma = \alpha_i + \beta_S \leq 1$, we have

$$|I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$$

- When $k = 0$, we have $I^{(1)}(v, C_i) \leq 1$, thus

$$|I^{(1)}(v, C_i) - I^{(0)}(v, C_i)| \leq 1$$

- Assume after k -th iteration, we have

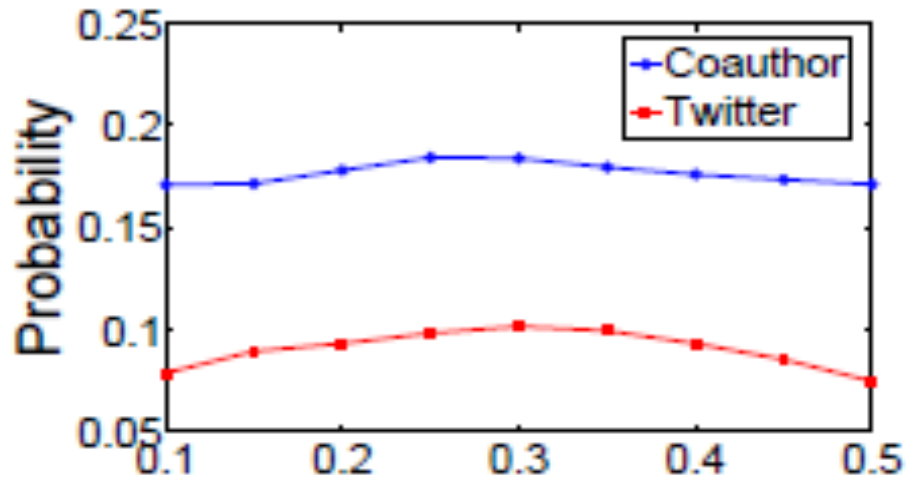
$$|I^{(k+1)}(v, C_i) - I^{(k)}(v, C_i)| \leq \gamma^k$$

- After $(k+1)$ -th iteration, we have

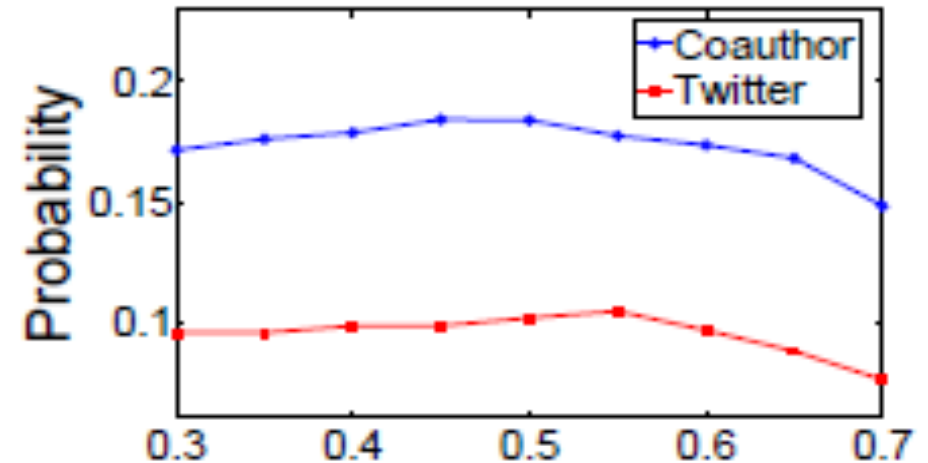
$$\begin{aligned} I^{(k+2)}(v, C_i) &= \alpha_i I^{(k+1)}(u, C_i) + \beta_S H^{(k+1)}(u, S) \\ &\leq \alpha_i [I^{(k)}(u, C_i) + \gamma^k] + \beta_S [H^{(k+1)}(u, S) + \gamma^k] \\ &\leq \alpha_i I^{(k)}(u, C_i) + \beta_S H^{(k+1)}(u, S) + \gamma^{k+1} \\ &\leq I^{(k+1)}(u, C_i) + \gamma^{k+1} \end{aligned}$$

Convergence Analysis

- Parameter analysis.
 - The performance is insensitive to the different parameter settings.



(a) α

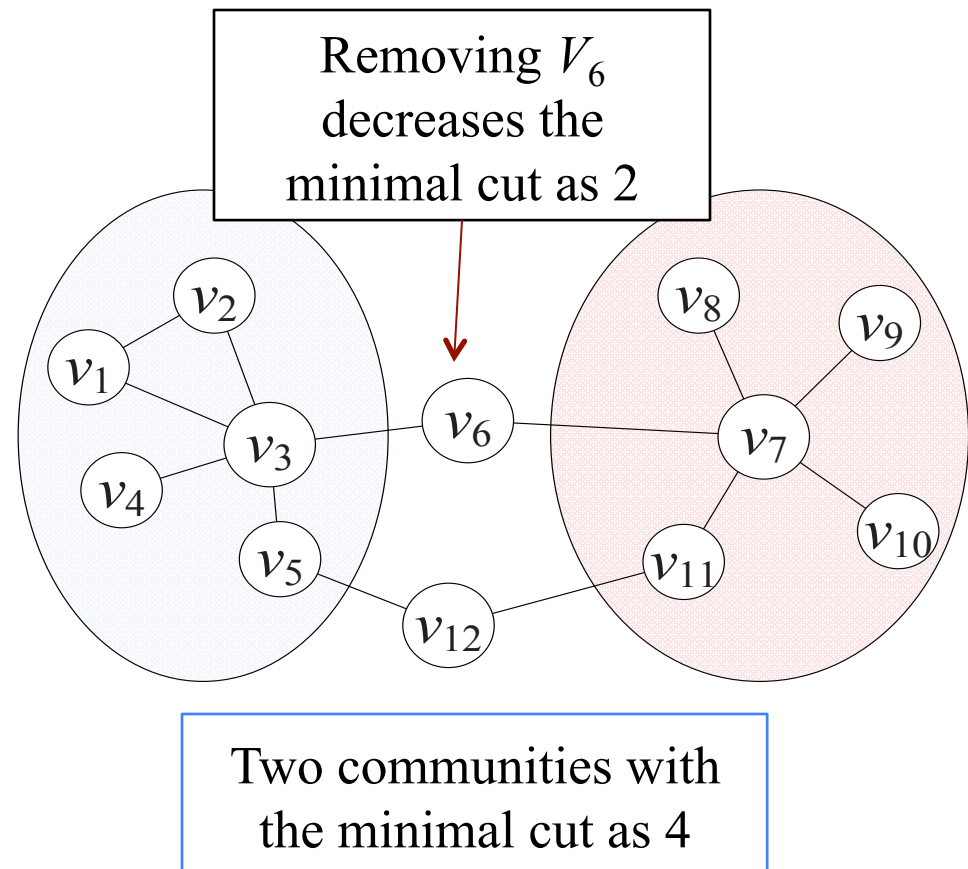


(b) β

Model Two: MaxD

- **Structural hole spanner detection:** finding top- k nodes such that after removing these nodes, the **decrease** of the **minimal cut** will be maximized.

Minimal cut: the minimal number of edges to *separate* nodes in different communities.



Model Two: MaxD

- Structural holes spanners play an important role in **information diffusion**

$$Q(V_{SH}, C) = \boxed{MC}(G, C) - \boxed{MC}(G \setminus V_{SH}, C)$$

$MC(G, C)$ = the minimal cut of communities C in G.

Shapley value is defined to evaluate the contribution of each user in a cooperative game theory
—Lloyd Shapley, 1953

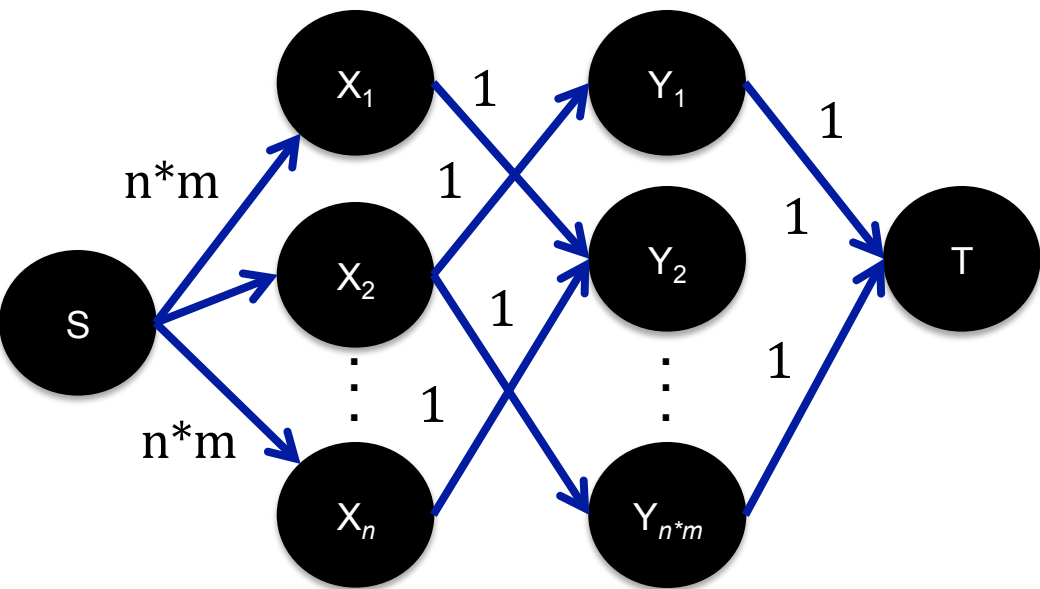
Hardness Analysis

$$Q(V_{SH}, C) = MC(G, C) - MC(G \setminus V_{SH}, C)$$

- Hardness analysis
 - If $|V_{SH}| = 2$, the problem can be viewed as **minimal node-cut problem**
 - We already have NP-Hardness proof for **minimal node-cut problem**, but the graph is exponentially weighted.
 - Proof NP-Hardness in an un-weighted (polybounded -weighted) graph, by reduction from **k-DENSEST-SUBGRAPH** problem.

Hardness Analysis

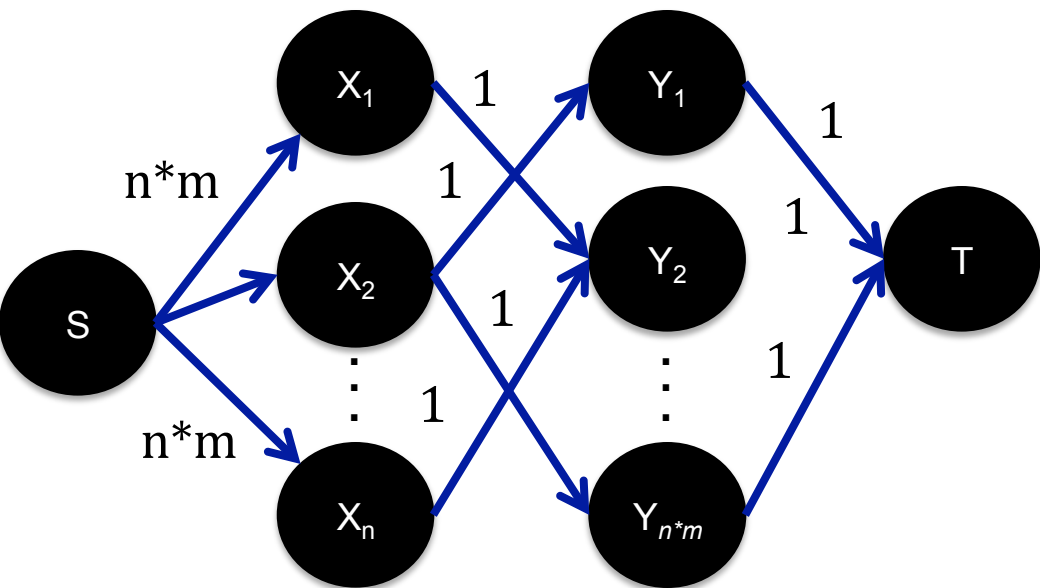
- Let us reduce the problem to an instance of the k -DENSEST SUBGRAPH problem



- Given an instance $\{G' = \langle V, E \rangle, k, d\}$ of the k -DENSEST SUBGRAPH problem, $n = |V|$, $m = |E|$;
- Build a graph G with a source node S and target node T ;
- Build n nodes connecting with S with capacity $n \cdot m$;
- Build n nodes for each edge in G' , connect each of them to T with capacity 1 ;

Hardness Analysis (cont.)

- Build a link from x_i to y_j with capacity 1 if the x_i in G' appears on the j -th edge;
- $MC(G) = n * m$;



- The instance is satisfiable, **if and only if** there exists a subset

$$|V_{SH}| = k$$

such that

$$MC(G \setminus V_{SH}) \leq n(m - d)$$

Proof: NP-hardness (cont.)

- For the *only if* direction
 - Suppose we have a sub-graph consists of k nodes $\{x'\}$ and at least d edges;
 - We can choose $V_{SH} = \{x\}$;
 - For the k -th edge y in G' , if y exists in the sub-graph, two nodes appearing on y are removed in G ;
 - Thus y cannot be reached and we lost n flows for y ;
 - Hence, we have $MC(G \setminus V_{SH}) \leq n^*(m-d)$.

Proof: NP-hardness (cont.)

- For the *if* direction
 - If there exists a k -subset V_{SH} such that $\text{MC}(G \setminus V_{SH}) \leq n^*(m-d)$;
 - Denote $V_{SH}' = V_{SH} \wedge \{x\}$, the size of V_{SH}' is at most k , and $\text{MC}(G \setminus V_{SH}') \leq n^*(m-d)$;
 - Let the node set of the sub-graph be V_{SH}' , thus there are at least d edges in that sub-graph.

Approximation Algorithm

- Two approximation algorithms:
 - Greedy: in each iteration, select a node which will result in a **max-decrease of $Q(\cdot)$** when removed it from the network.
 - Network-flow: for any possible partitions E_S and E_T , we call a network-flow algorithm to compute the minimal cut.

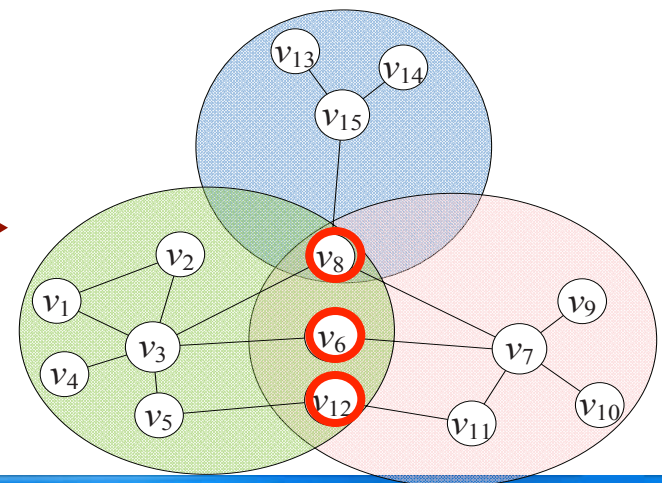
Complexity: $O(nkT_l(n))$; $T_l(n)$ —the complexity for computing min-cut

Complexity: $O(k^2T_l(n))$;

Approximation Ratio: $O(n^{1/4+\epsilon})$

An example: finding top 3 structural holes

- Step 1: select V8 and decrease the minimal cut from 7 to 4
Step 2: select V6 and decrease the minimal cut from 4 to 2
Step 3: select V12 and decrease the minimal cut from 2 to 0



Approximation Algorithm

Greedy : In each round, choose the node which results in the max-decrease of Q .

Input: $G = (V, E), k, l, \mathbf{C} = \{C_i\}$

Output: Top- k structural hole nodes V_{SH}

Initialize $V_{SH} = \emptyset$;

while $|V_{SH}| < k$ **do**

Initialize $f(v) = 0$, for each $v \in V$;

foreach non empty $S \subset \{1, \dots, l\}$ **do**

$E_S = \cup_{i \in S} C_i$ and $E_T = \cup_{i \notin S} C_i$;

 Compute the maximal flow with source E_S and sink E_T on the induced graph $G \setminus V_{SH}$;

foreach $v \in V$ **do**

 | Add $f(v)$ by the flow through node v ;

end

end

Choose $O(k)$ nodes with the largest f as candidates D ;

Compute $p^* = \arg \min_{p \in D} MC(G \setminus (V_{SH} \cup \{p\}), \mathbf{C})$;

Update $V_{SH} = V_{SH} \cup \{p^*\}$

end

Step 1: Consider top $O(k)$ nodes with maximal sum of flows through them as candidates.

Step 2: Compute $MC(*, *)$ by trying all possible partitions.

Complexity: $O(2^l T_2(n))$; $T_2(n)$ —the complexity for computing min-cut

Approximation ratio: $O(\log l)$

Mining structural hole spanners

- Evaluate the **performance** of the proposed models.

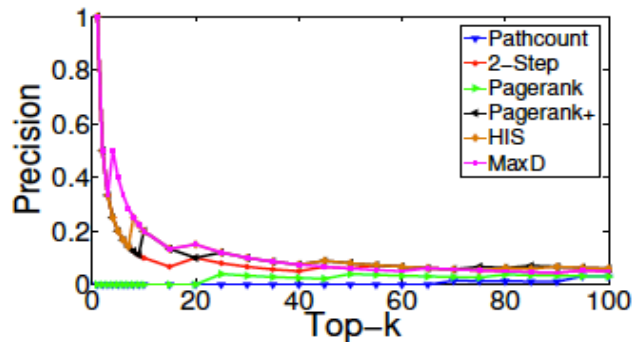
Experiment

	#User	#Relationship	#Messages
Coauthor	815,946	2,792,833	1,572,277 papers
Twitter	112,044	468,238	2,409,768 tweets
Inventor	2,445,351	5,841,940	3,880,211 patents

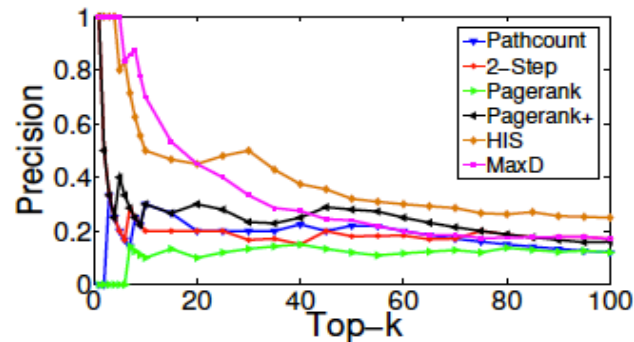
- Evaluation metrics
 - Accuracy (Overlapped PC members in the Coauthor network)
 - Information diffusion on Coauthor and Twitter.
- Baselines
 - Pathcount: #shortest path a node lies on
 - 2-step connectivity: #pairs of disconnected neighbors
 - Pagerank and PageRank+: high PR in more than one communities

Experiments

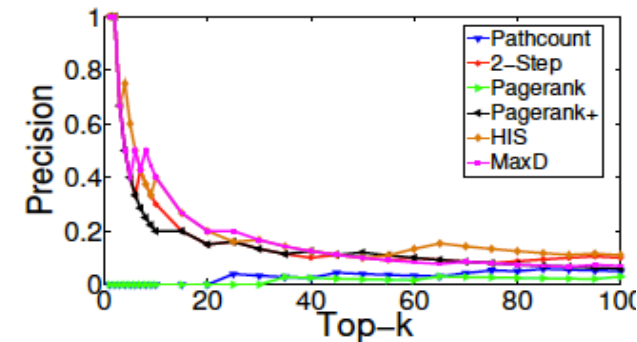
- Accuracy evaluation on Coauthor network



(a) AI-DM



(b) DB-DM



(c) DP-NC

- Predict overlapped PC members on the Coauthor network.
 - +20 – 40% on precision of AI-DM, DB-DM and DP-NC
- What happened to AI-DM?

Experiment results (accuracy)

- What happened to AI-DB?
 - Only 4 overlapped PC members on AI and DB during 2007 – 2009, but 40 now.
 - Our conjecture : **dynamic of structural holes.**

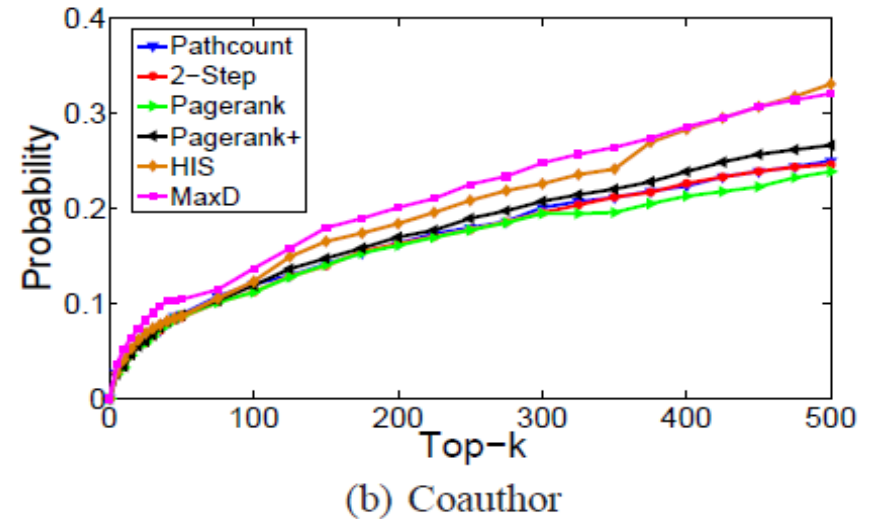
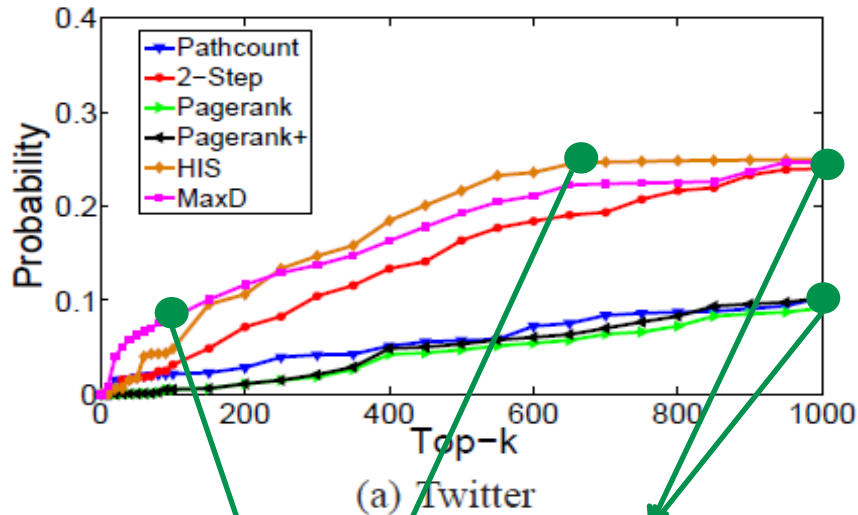
Structural holes spanners of AI and DB form the new area DM.

Similar pattern for
1) Collaborations between experts in AI and DB.
2) Influential of **DM** papers.

Significantly increase of coauthor links of AI and DB around year **1994**.

Most overlapped PC members on AI and DB are also PC of **SIGKDD**

Maximization of Information Spread



Clear improvement. **(2.5 times)**

Top 0.2% - 10 %

Top 1% - 25 %

Improvement is limited, due to top a few authors dominate.

Improvement is statistically significant ($p \ll 0.01$)

Case study on the inventor network

- Most structural holes have more than one jobs.
- Mark * on inventors with highest PageRank scores.
 - HIS selects people with highest PageRank scores,
 - MaxD tends to select people how have been working on more jobs.

Inventor	HIS	MaxD	Title
E. Boyden		√	Professor (MIT Media Lab)
			Associate Professor (MIT McGovern Inst.)
			Group Leader (Synthetic Neurobiology)
A.A. Czarnik		√	Founder and Manager (Protia, LLC)
			Visiting Professor (University of Nevada)
			Co-Founder (Chief Scientific Officer)
A. Nishio		√	Director of Operations (WBI)
			Director of Department Responsible (IDA)
E. Nowak*	√		Senior vice President (Walt Disney)
A. Rofougaran	√		Secretary of Trustees (The New York Eye)
			Consultant (various wireless companies)
			Co-founder (Innovent System Corp.)
			Leader (RF-CMOS)
S. Yamazaki*	√		President and majority shareholder (SEL)

Efficiency

- Running time of different algorithms in three data sets

Data Set	Pathcount	2-Step	PageRank	HIS	MaxD
Coauthor	350.66s	4.71s	0.20s	0.60s	189.78m
Twitter	32.03m	12.09s	0.67s	3.87s	602.37m
Inventor	494.3 hr	98.96s	3.61s	26.11s	370.8hr

Inefficient!!

Summaries

- **Models for social influence and diffusion**
 - Learning social influence
 - Distinguish influence from other factors
 - Cases: Game
- **This is just a start for social influence analysis**
 - How influence correlates with social relationships?
 - How social influence correlates with the network structure (e.g., personal social circles)?
 - ...

Related Publications

- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. In **KDD'09**, pages 807-816, 2009.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social action tracking via noise tolerant time-varying factor graphs. In **KDD'10**, pages 807–816, 2010.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. In **KDD'11**, pages 1397–1405, 2011.
- Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity Influence in Large Social Networks. In **KDD'13**, pages 347-355, 2013.
- Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social Influence Locality for Modeling Retweeting Behaviors. In **IJCAI'13**, pages 2761-2767, 2013.
- Jing Zhang, Jie Tang, Honglei Zhuang, Cane Wing-Ki Leung, and Juanzi Li. Role-aware Conformity Influence Modeling and Analysis in Social Networks. In **AAAI'14**, 2014.
- 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.
- Tiancheng Lou and Jie Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. In **WWW'13**, pages 837-848, 2013.
- 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.
- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. In **TKDD**, Vol 7(2), 2013.
- 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 and Jimeng Sun. Models and Algorithms for Social Influence Analysis. In **WWW'14**. (Tutorial)

References

- S. Milgram. The Small World Problem. **Psychology Today**, 1967, Vol. 2, 60–67
- J.H. Fowler and N.A. Christakis. The Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. **British Medical Journal** 2008; 337: a2338
- R. Dunbar. Neocortex size as a constraint on group size in primates. **Human Evolution**, 1992, 20: 469–493.
- R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. **Nature**, 489:295-298, 2012.
- <http://klout.com>
- Why I Deleted My Klout Profile, by Pam Moore, at **Social Media Today**, originally published November 19, 2011; retrieved November 26 2011
- S. Aral and D Walker. Identifying Influential and Susceptible Members of Social Networks. **Science**, 337:337-341, 2012.
- J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. **PNAS**, 109 (20):7591-7592, 2012.
- S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. **PNAS**, 106 (51):21544-21549, 2009.
- J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In **KDD'09**, pages 747–756, 2009.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. **Journal of Educational Psychology** 66, 5, 688–701.
- http://en.wikipedia.org/wiki/Randomized_experiment

References(cont.)

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

References(cont.)

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

Thank you !

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

Jimeng Sun (**IBM**) Tiancheng Lou (**Google**)

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

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

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

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,

<http://keg.cs.tsinghua.edu.cn/jietang>
<http://arnetminer.org/download>