Social Tie Analysis

—Computational aspect

Jie Tang

Tsinghua University, China

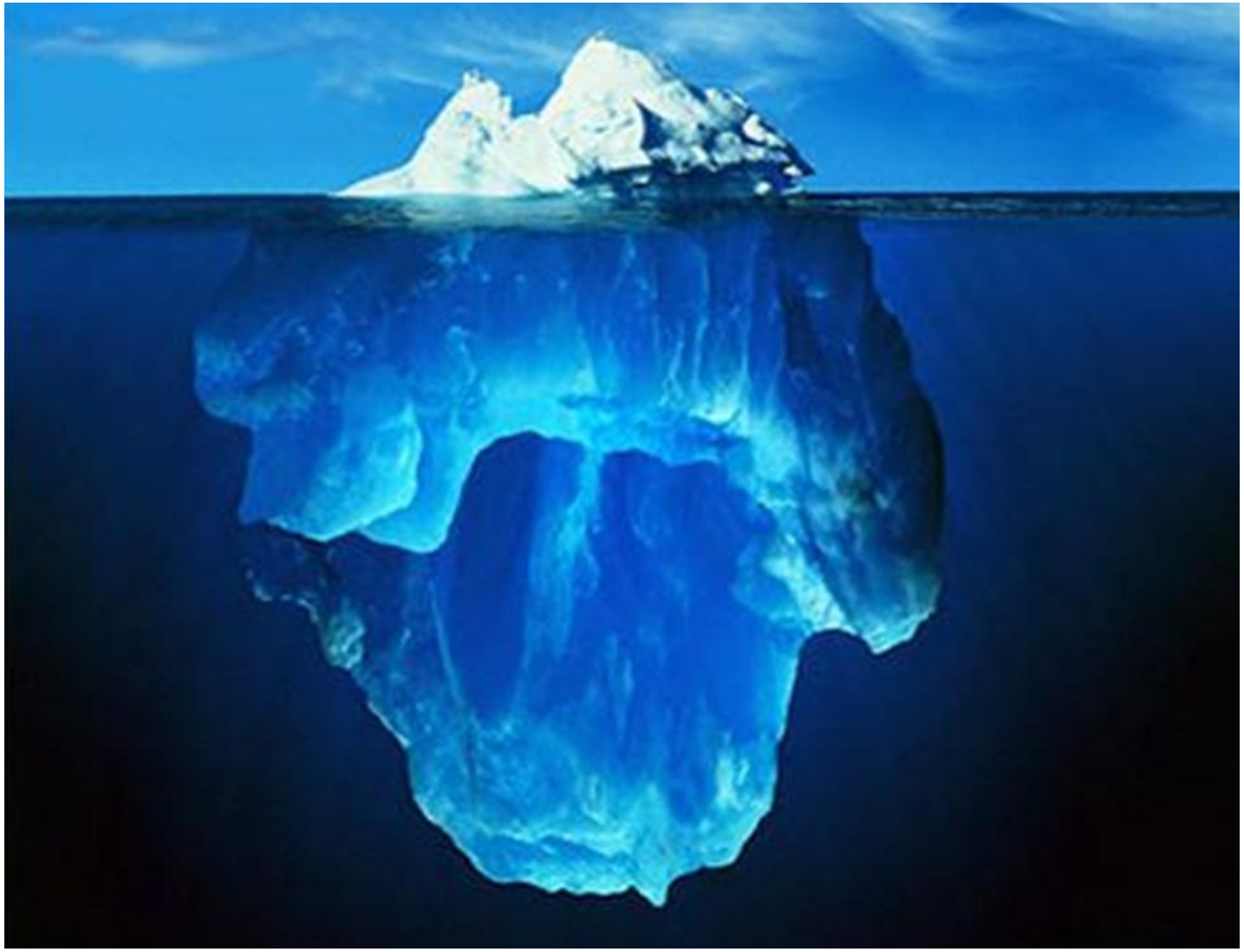
Collaborate with

Jon Kleinberg and John Hopcroft (*Cornell*)

Jiawei Han and Chi Wang (*UIUC*)

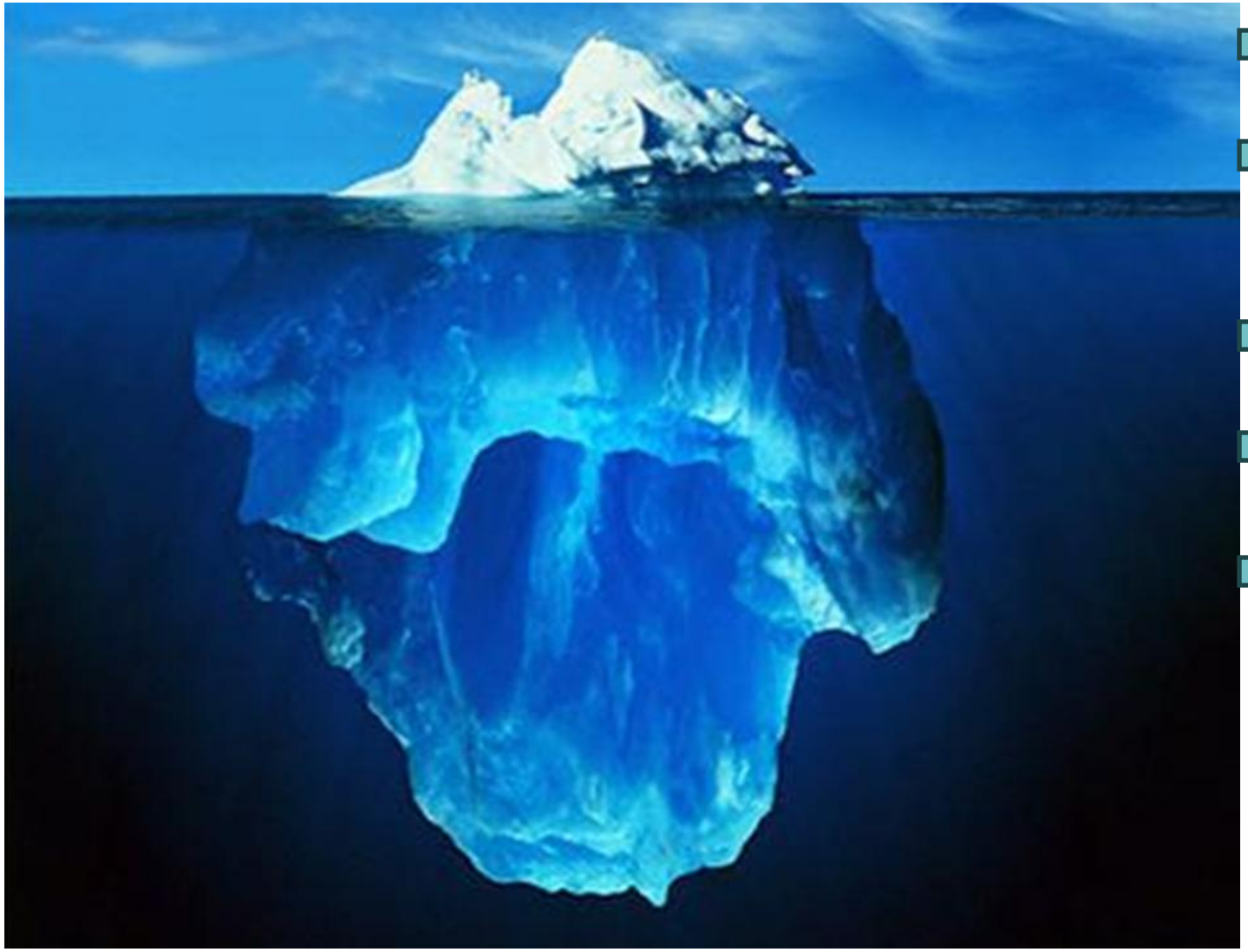
Tiancheng Lou, Wenbin Tang, Honglei Zhuang, and Jing Zhang (*THU*)

Iceberg Model for Social Network



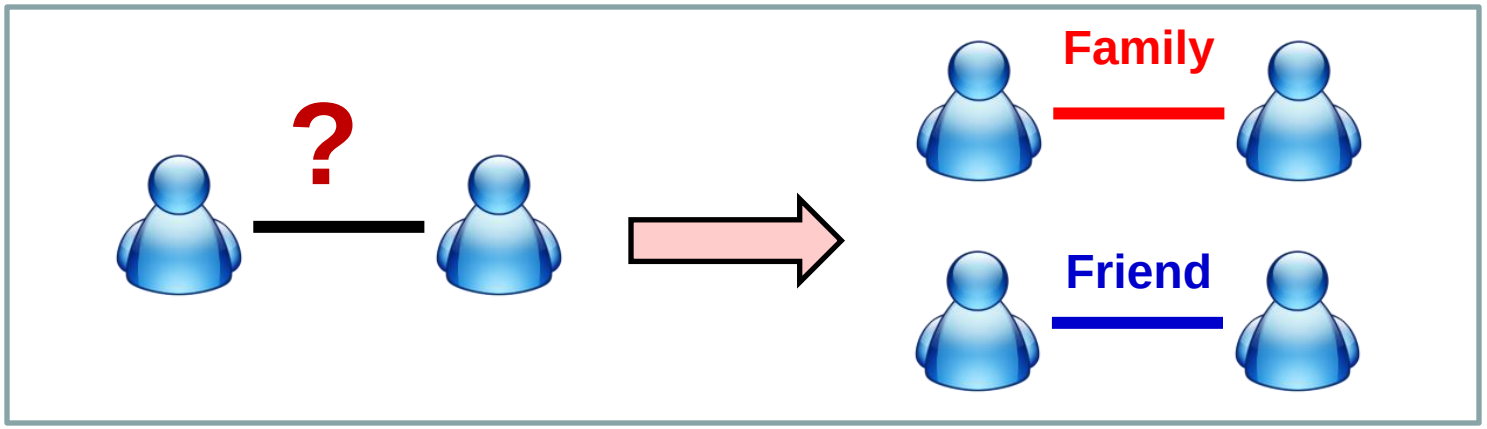
?

Iceberg Model for Social Network



- Information Difussion
- Collective Intelligence
- Tie
- Influence
- Traits and Motivates

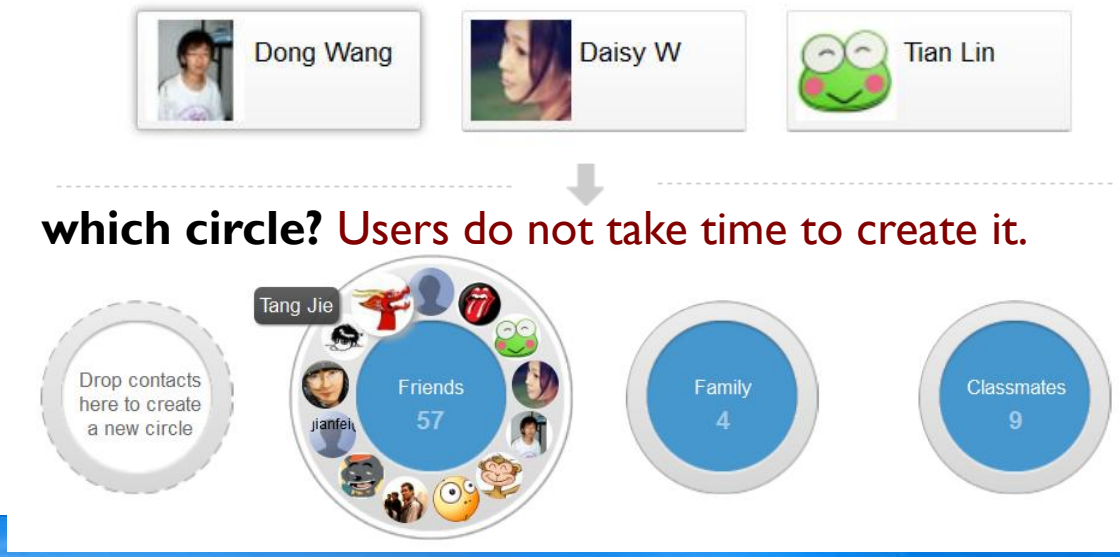
Inferring Social Ties



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

Real social networks are complex...

- Nobody exists only in one social network.
  - Public network vs. private network
  - Business network vs. family network
- However, existing networks (e.g., Facebook and Twitter) are trying to lump everyone into one big network
  - FB tries to solve this problem via lists/groups
  - However…
- Google+



Even complex than we imaged!

- Only 16% of mobile phone users in Europe have created custom contact groups
  - *users do not* take the time to create it
  - *users do not* know how to circle their friends

- The fact is that our social network is **black**-white…

Example 1: finding boss in email networks

(PKDD'11, Best Paper Runnerup)

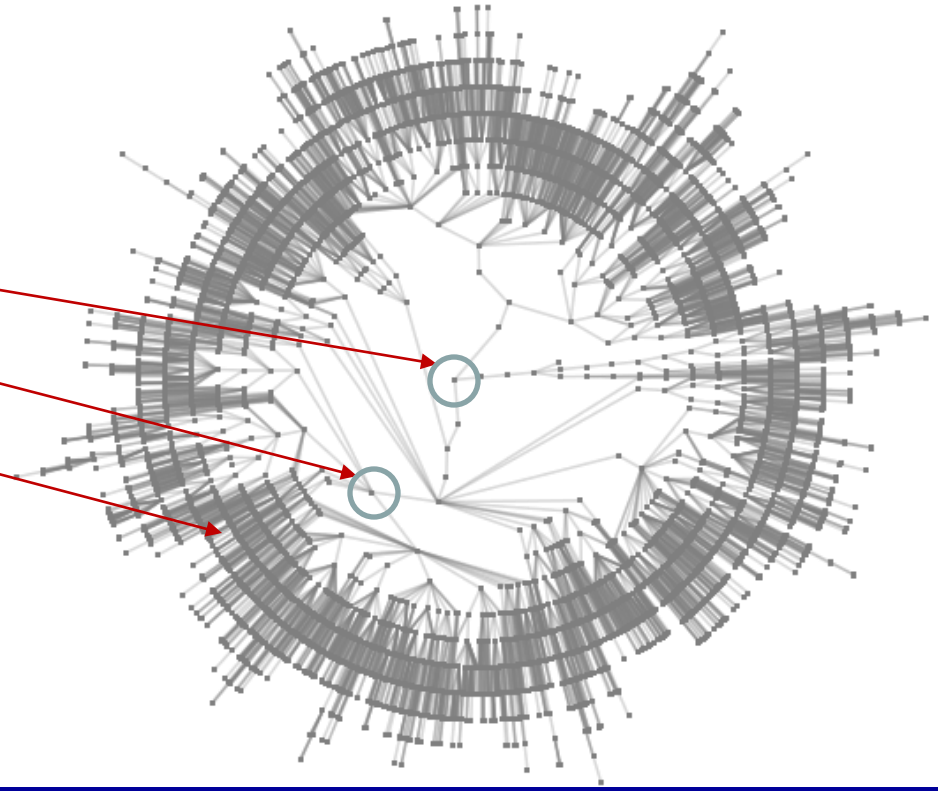
Enterprise email network

How to infer

CEO

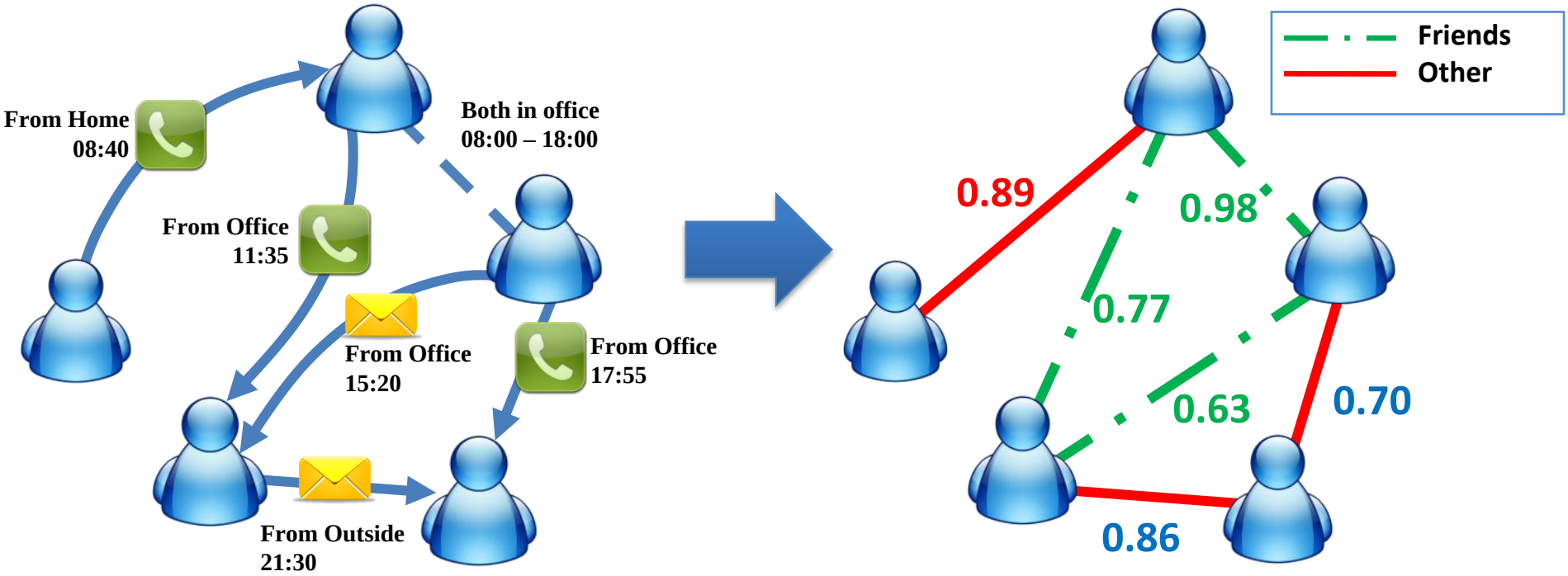
Manager

Employee



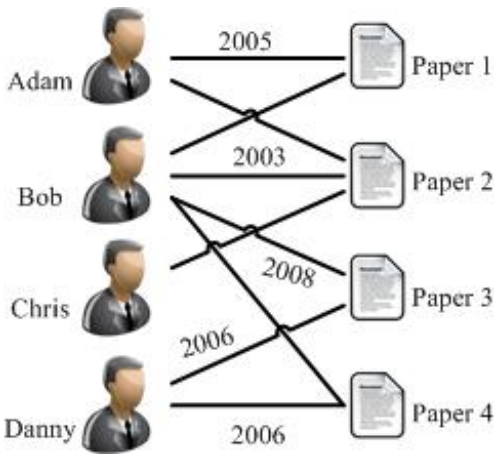
User interactions may form *implicit groups*

Example 2: finding friends in mobile networks

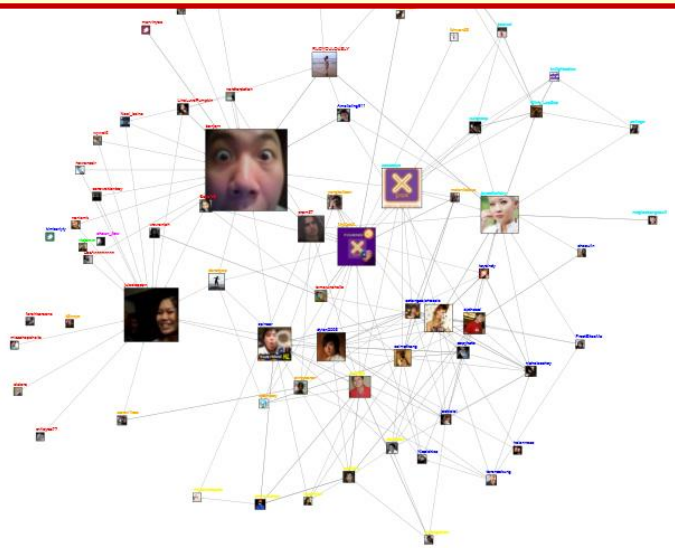


Challenges

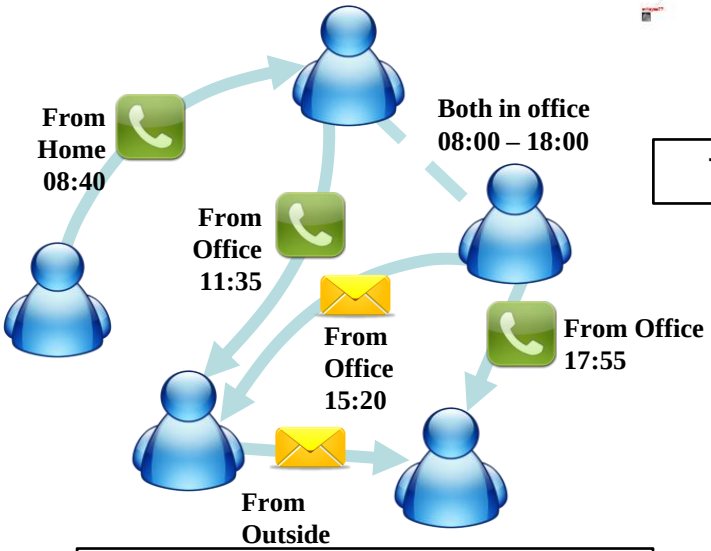
- What are the fundamental forces behind?
- Can we automatically infer the type of social ties?



Publication network



Twitter's following network



Mobile communication network

Networks

- **Epinions** a network of product reviewers: 131,828 nodes (users) and 841,372 edges
  - trust relationships between users
- **Slashdot**: 82,144 users and 59,202 edges
  - “friend” relationships between users
- **Mobile**: 107 mobile users and 5,436 edges
  - to infer friendships between users

Undirected network

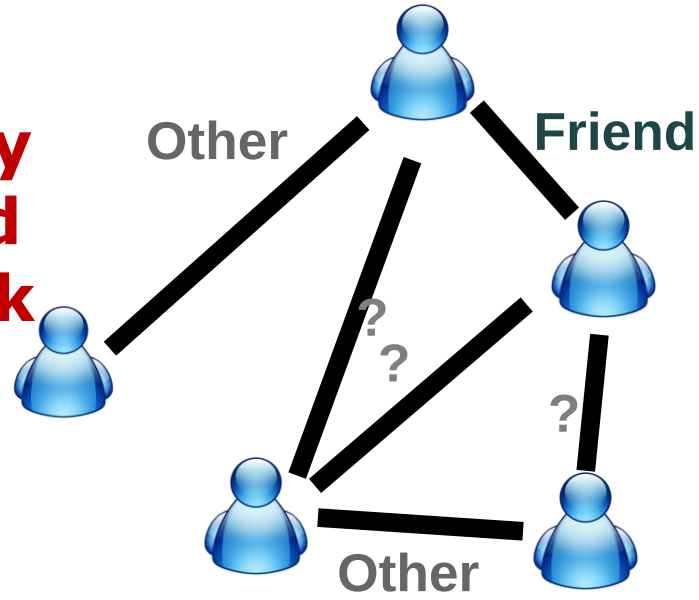
- **Coauthor**: 815,946 authors and 2,792,833 coauthor relationships
  - to infer advisor-advisee relationships between coauthors
- **Enron**: 151 Enron employees and 3572 edges
  - to infer manager-subordinate relationships betwee

Directed network

Problem Formulation

Input: $G=(V,E^L,E^U,R^L,W)$

Partially Labeled Network



V: Set of Users

E^L,R^L: Labeled relationships

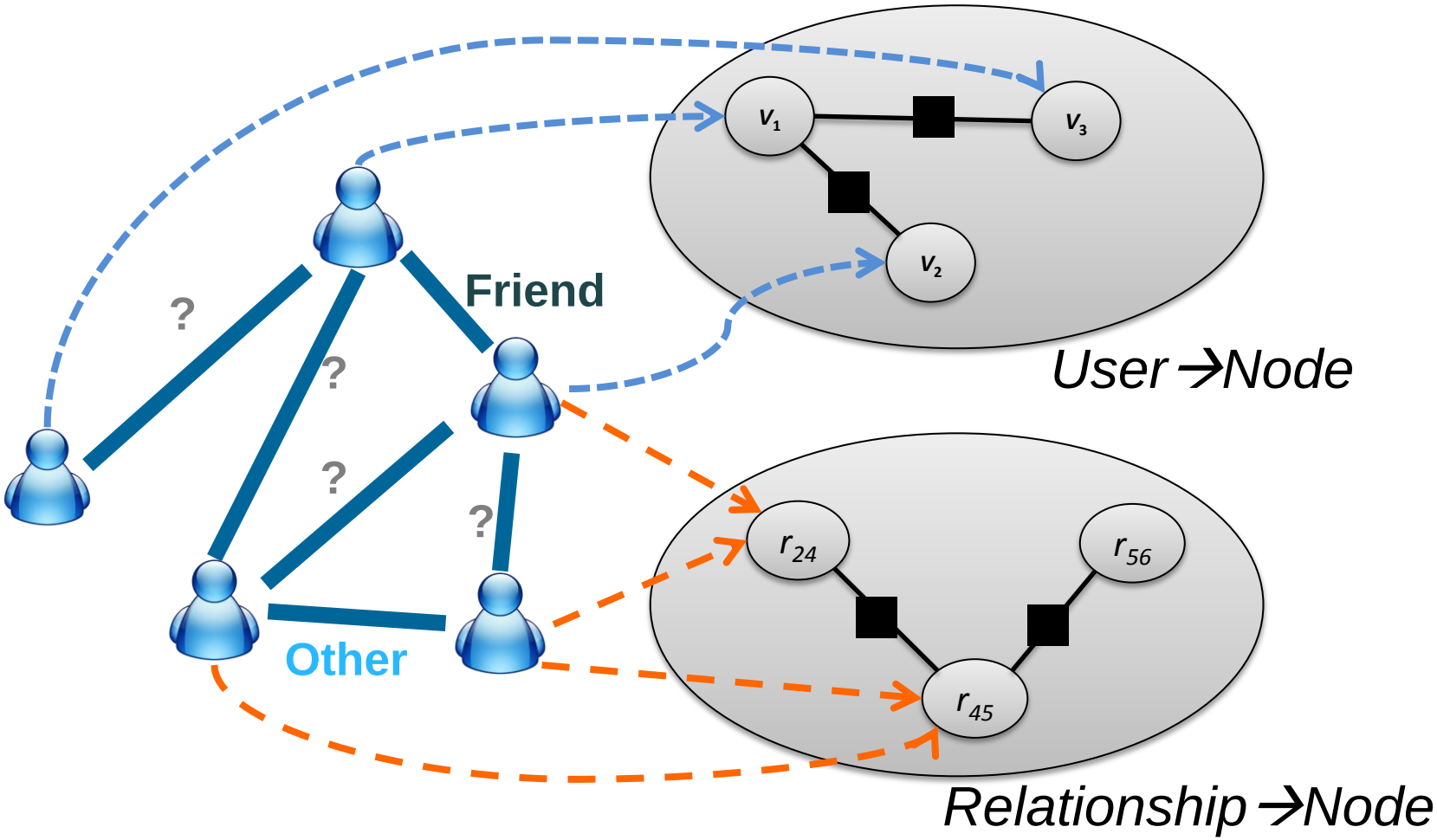
E^U: Unlabeled relationships

Input:
$G=(V,E^L,E^U,R^L,W)$

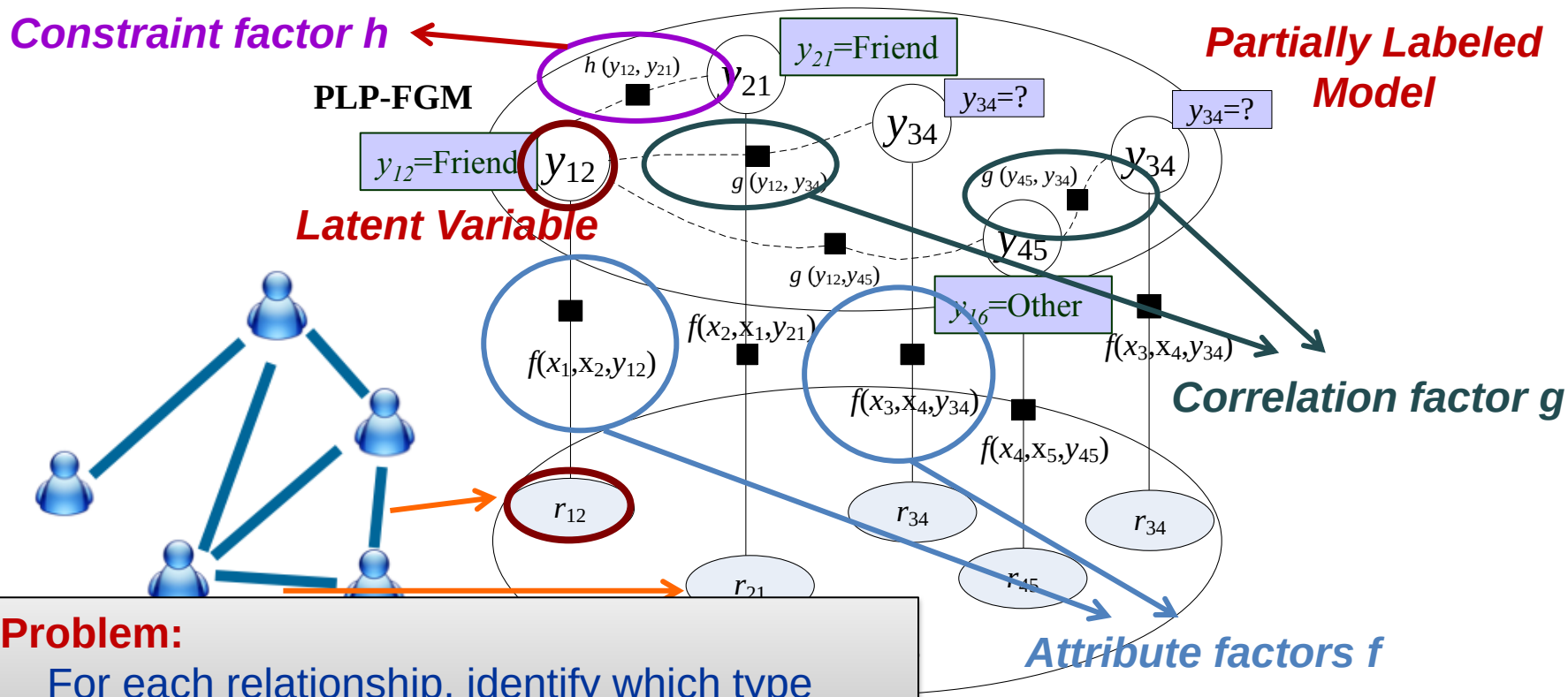
→

Output:
$f: G \rightarrow R$

Basic Idea


Friend
Other
?
?
?
?
v_1
v_3
v_2
User→Node
r_{24}
r_{56}
r_{45}
Relationship→Node

Partially Labeled Pairwise Factor Graph Model (PLP-FGM)



Problem:
For each relationship, identify which type has the highest probability?

Example:
A makes call to B immediately after the call to C.

Wenbin Tang, Honglei Zhuang, and Jie Tang. Learning to Infer Social Ties in Large Networks. In ECML/PKDD'2011. pp. 381-397. (**Best Student Paper Runner-up**)

Solutions(con't)

- Different ways to instantiate factors
  - We use exponential-linear functions
    - Attribute Factor:

$$f(y_i, \mathbf{x}_i) = \frac{1}{Z_\lambda} \exp\{\lambda^T \mathbf{\Phi}(y_i, \mathbf{x}_i)\}$$

    - Correlation / Constraint Factor:

$$g(y_i, G(y_i)) = \frac{1}{Z_\alpha} \exp\{\sum_{y_j \in G(y_i)} \alpha^T \mathbf{g}(y_i, y_j)\}$$

$$h(y_i, H(y_i)) = \frac{1}{Z_\beta} \exp\{\sum_{y_j \in H(y_i)} \beta^T \mathbf{h}(y_i, y_j)\}$$

  - $\theta = [\lambda, \alpha, \beta], s = [\Phi^T, g^T, h^T]^T$
  - Log-Likelihood of labeled Data:

$$\mathcal{O}(\theta) = \log \sum_{Y|Y^L} \exp\{\theta^T \mathbf{S}\} - \log \sum_{Y} \exp\{\theta^T \mathbf{S}\}$$

Learning Algorithm

- Maximize the log-likelihood of labeled relationships

Input: learning rate η
Output: learned parameters θ

Initialize θ;
repeat
  Calculate $\mathbb{E}_{p_\theta(Y|Y^L,G)}\mathbf{S}$ using LBP ;
  Calculate $\mathbb{E}_{p_\theta(Y|G)}\mathbf{S}$ using LBP ;
  Calculate the gradient of θ according to Eq. 7:

$$\nabla_\theta = \mathbb{E}_{p_\theta(Y|Y^L,G)}\mathbf{S} - \mathbb{E}_{p_\theta(Y|G)}\mathbf{S}$$

  Update parameter θ with the learning rate η:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_\theta$$

until *Convergence*;

Algorithm 1: Learning PLP-FGM.

Expectation Computing

Loopy Belief Propagation

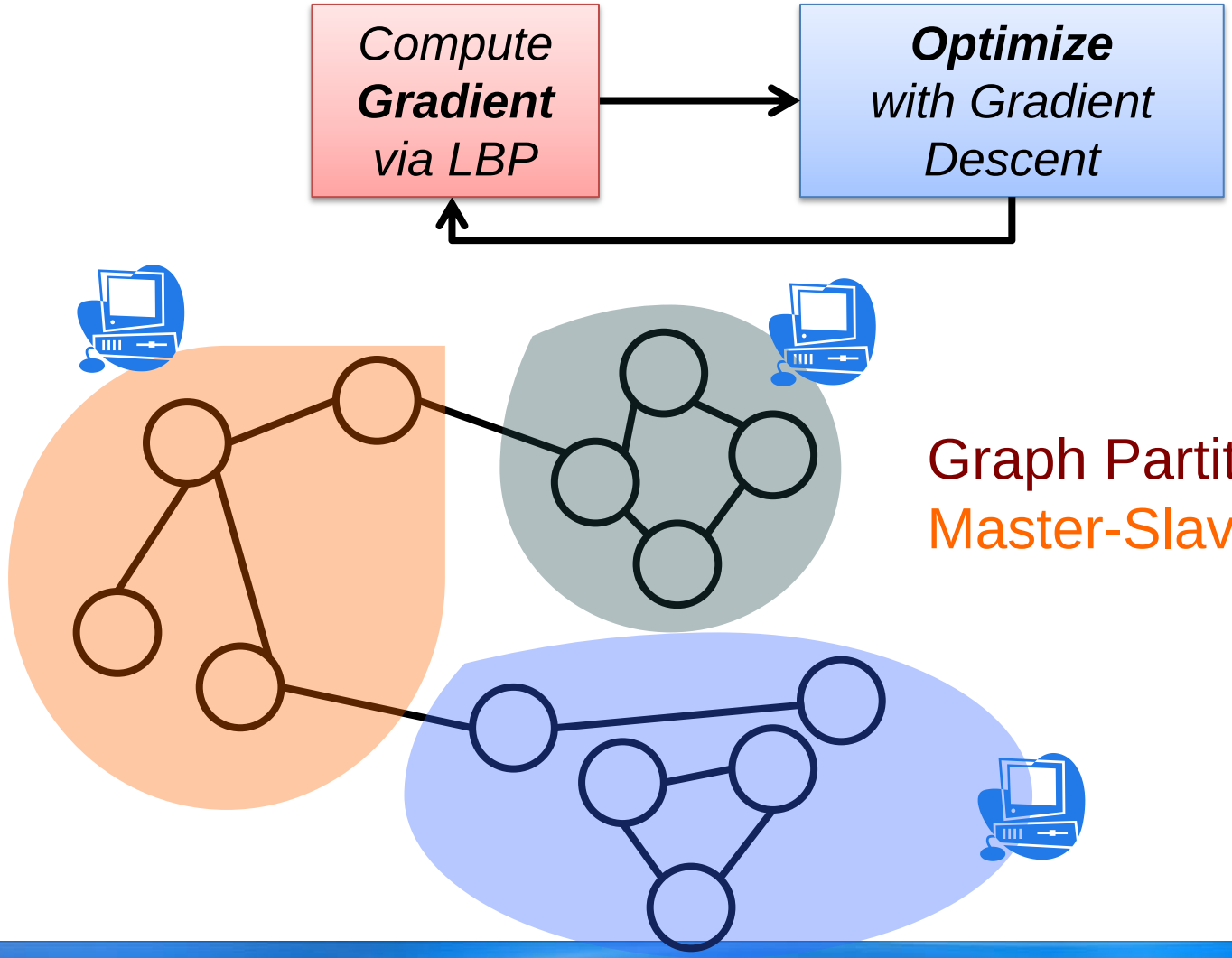
Gradient Ascent Method

Still Challenges?

Questions:

- How to obtain sufficiently training data?
- Can we leverage knowledge from other network?

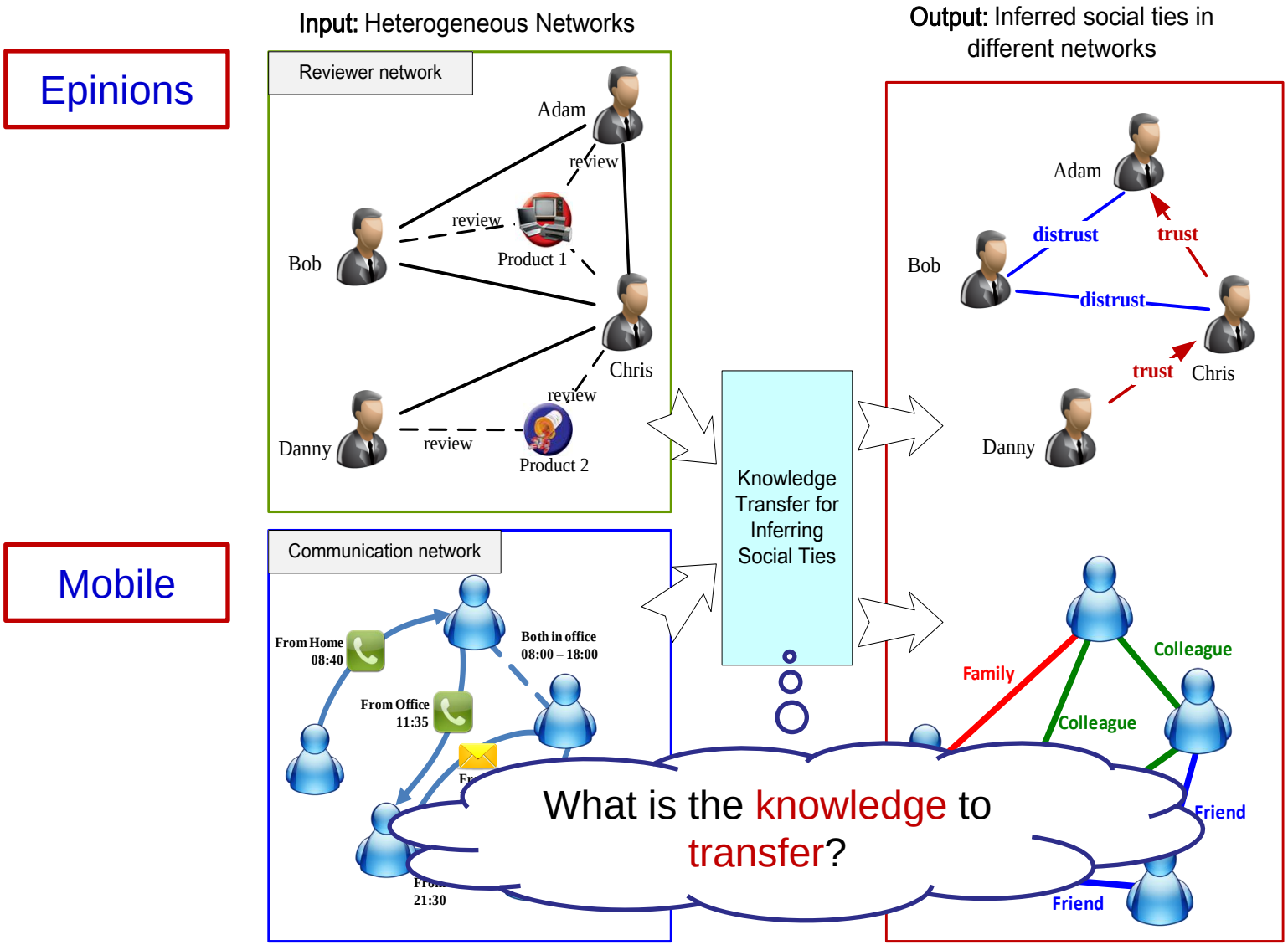
Distributed Learning



Graph Partition by Metis

Master-Slave Computing

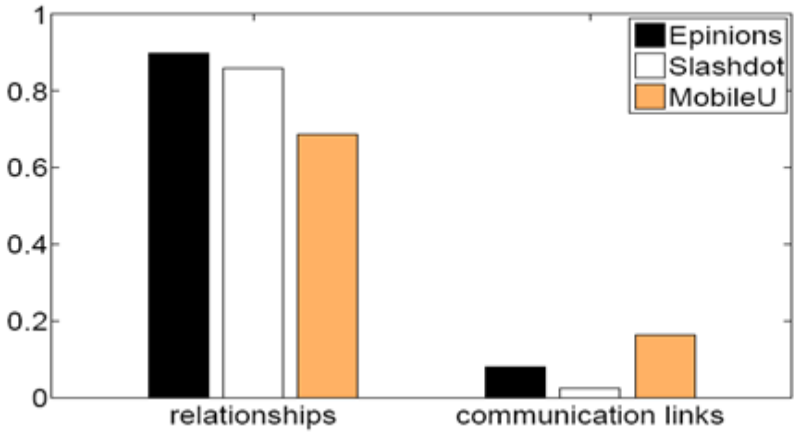
Inferring Social Ties Across Networks



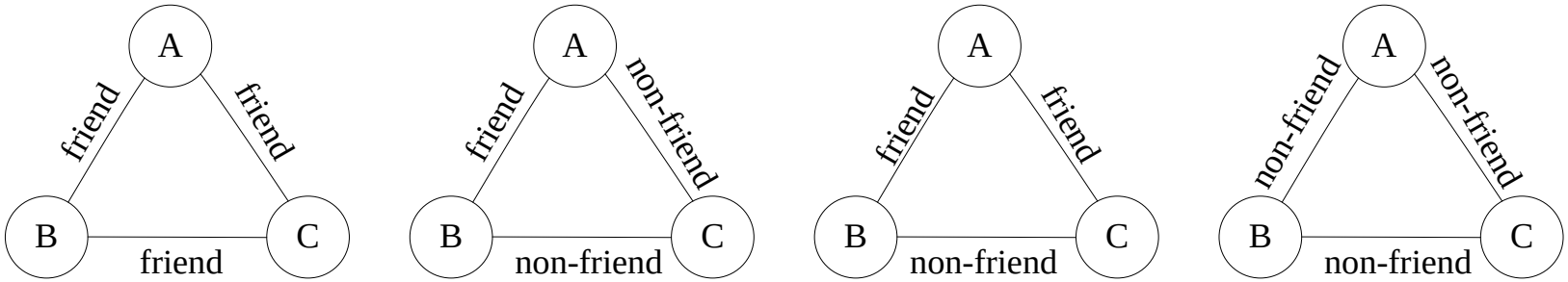
Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. In WSDM'2012. pp. 743-752.

Social Theories

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

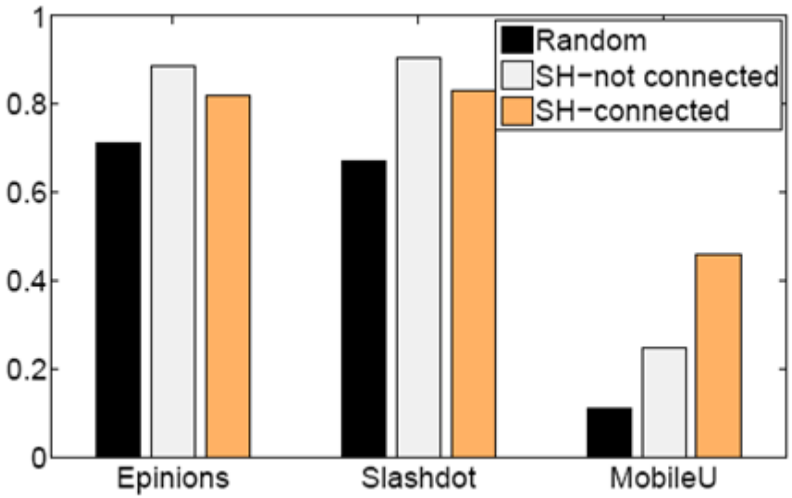


Observations:
(1) The underlying networks are unbalanced;
(2) While the friendship networks are balanced.

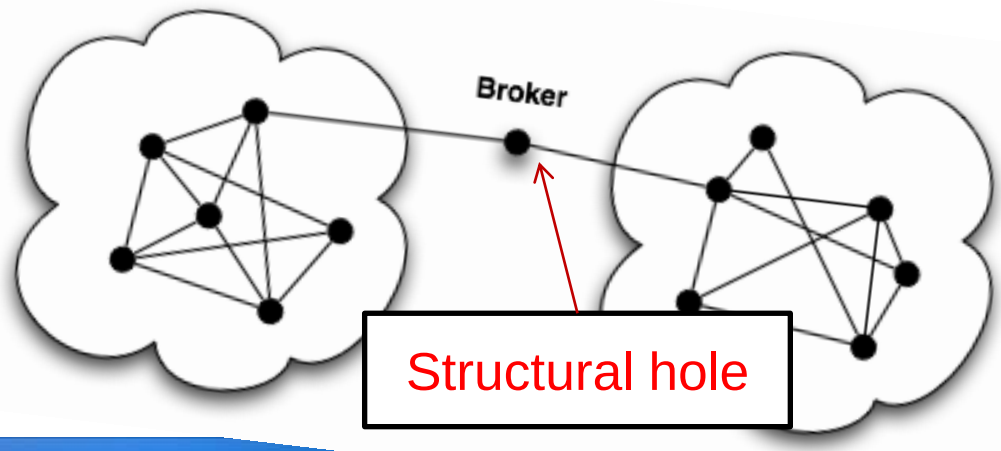


Social Theories—Structural hole

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

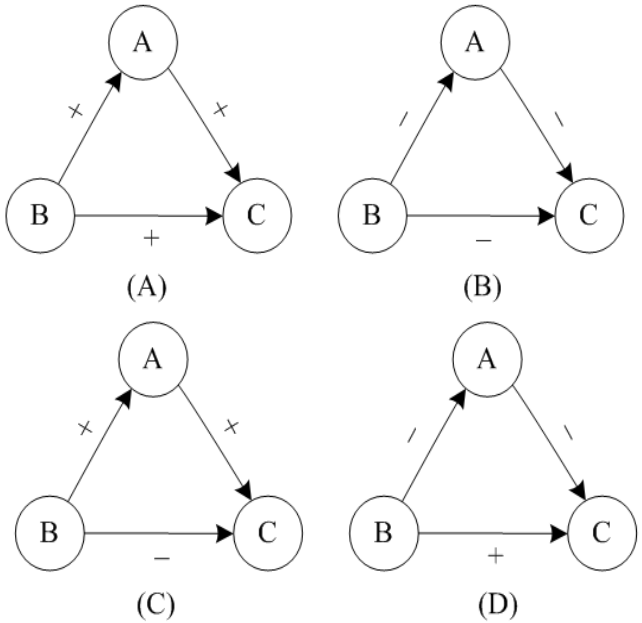


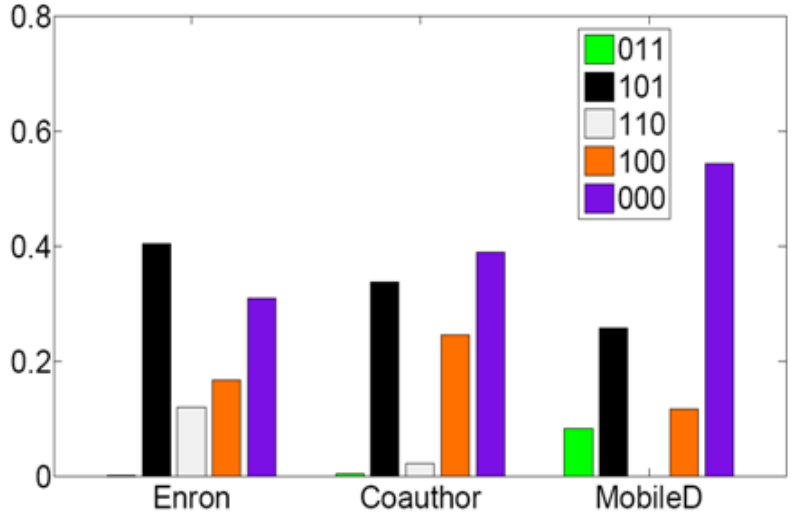
Observations: Users are more likely (+25-150% higher than change) to have the same type of relationship with C if C **spans structural holes**



Social Theories—Social status

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory



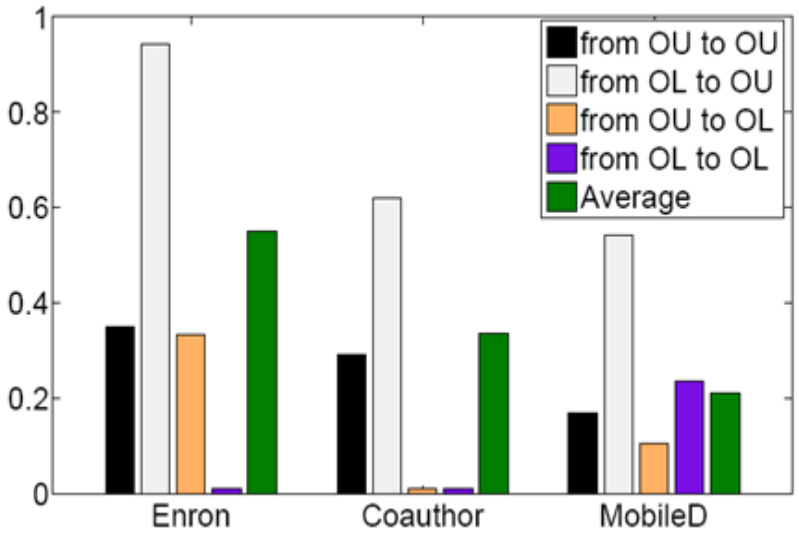


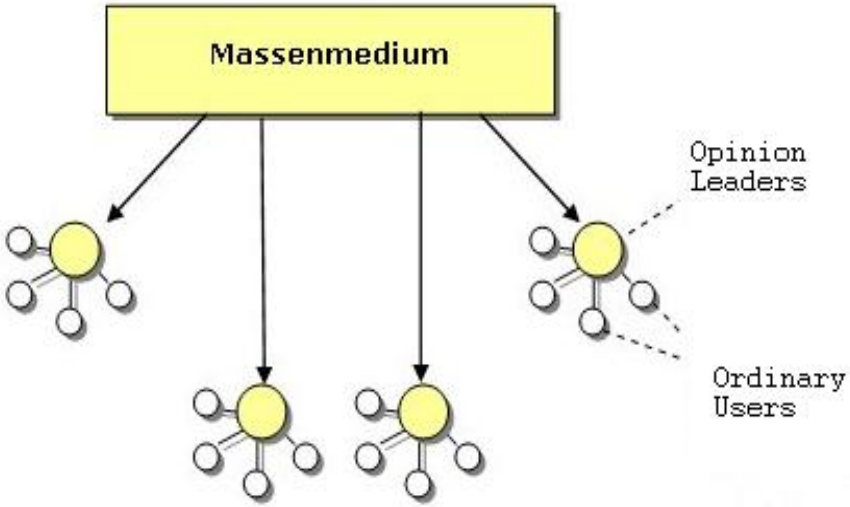
Observations: 99% of triads in the networks satisfy the social status theory

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.

Social Theories—Two-step-flow

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory



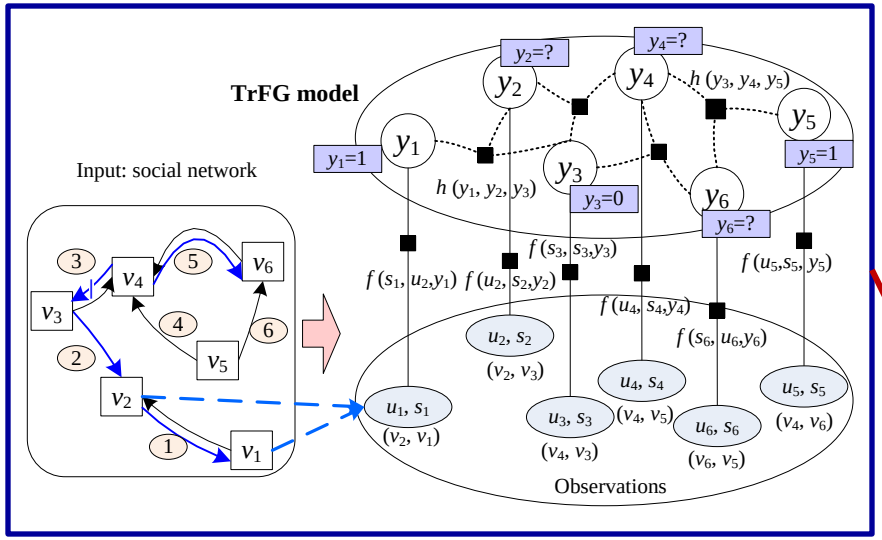


OL : Opinion leader;
OU : Ordinary user.

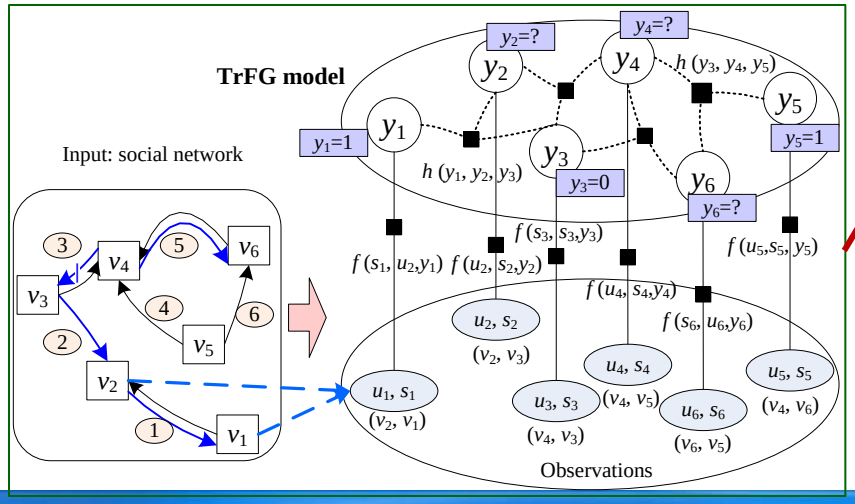
Observations: Opinion leaders are more likely (+71%-84% higher than chance) to have a higher social-status than ordinary users.

Transfer Factor Graph Model

Coauthor network



mobile



Bridge via social theories

Mathematical Formulation

Features defined in source network

Features defined in target network

$$\mathcal{O}(\alpha,\beta,\mu) = \mathcal{O}_S(\alpha,\mu) + \mathcal{O}_T(\beta,\mu)$$

$$= \sum_{i=1}^{|V_S|}\sum_{j=1}^{d} \boxed{\alpha_j g_j(x_{ij}^S, y_i^S)} + \sum_{i=1}^{|V_T|}\sum_{j=1}^{d'} \boxed{\beta_j g'_j(x_{ij}^T, y_i^T)}$$

$$+ \sum_k \mu_k \Big(\sum_{c\in G_S} \boxed{h_k(Y_c^S)} + \sum_{c\in G_T} \boxed{h_k(Y_c^T)} \Big)$$

$$- \log Z$$

Triad-based features shared across networks

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

Experiments

- **Data sets**
  - **Epinions:** 131,828 nodes (users) and 841,372 edges
  - **Slashdot**: 82,144 users and 59,202 edges
  - **Mobile**: 107 mobile users and 5,436 edges
  - **Coauthor**: 815,946 authors and 2,792,833 coauthor relationships
  - **Enron**: 151 Enron employees and 3572 edges
- Comparison methods
  - **SVM** and **CRF** are two baseline methods
  - **PFG** is the partially-labeled factor graph model
  - **TranFG** is the transfer–based factor graph model

Results – undirected networks

- **SVM** and **CRF** are two baseline methods
- **PFG** is the proposed partially-labeled factor graph model
- **TranFG** is the proposed transfer–based factor graph model.

Data Set	Method	Prec.	Rec.	F1-score
Epinions (S) to Slashdot (T) (40%)	SVM	0.7157	**0.9733**	0.8249
	CRF	0.8919	0.6710	0.7658
	PFG	0.9300	0.6436	0.7607
	TranFG	**0.9414**	0.9446	**0.9430**
Slashdot (S) to Epinions (T) (40%)	SVM	0.9132	**0.9925**	0.9512
	CRF	0.8923	0.9911	0.9393
	PFG	**0.9954**	0.9787	**0.9870**
	TranFG	**0.9954**	0.9787	**0.9870**
Epinions (S) to Mobile (T) (40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.8239	**0.8344**	**0.8291**
Slashdot (S) to Mobile (T) (40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.7258	**0.8599**	**0.7872**

Results – directed networks

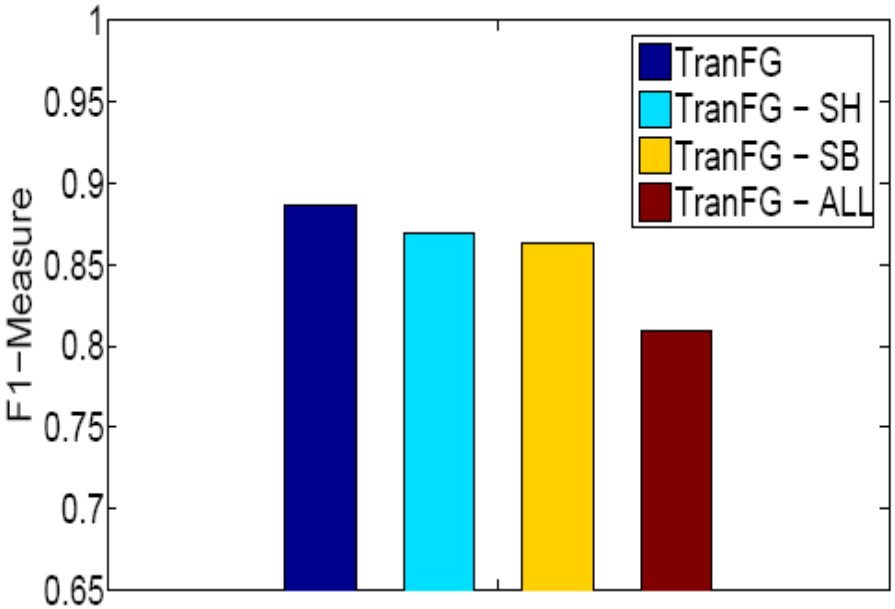
SVM and **CRF** are two baseline methods

PFG is the proposed partially-labeled factor graph model

TranFG is the proposed transfer–based factor graph model.

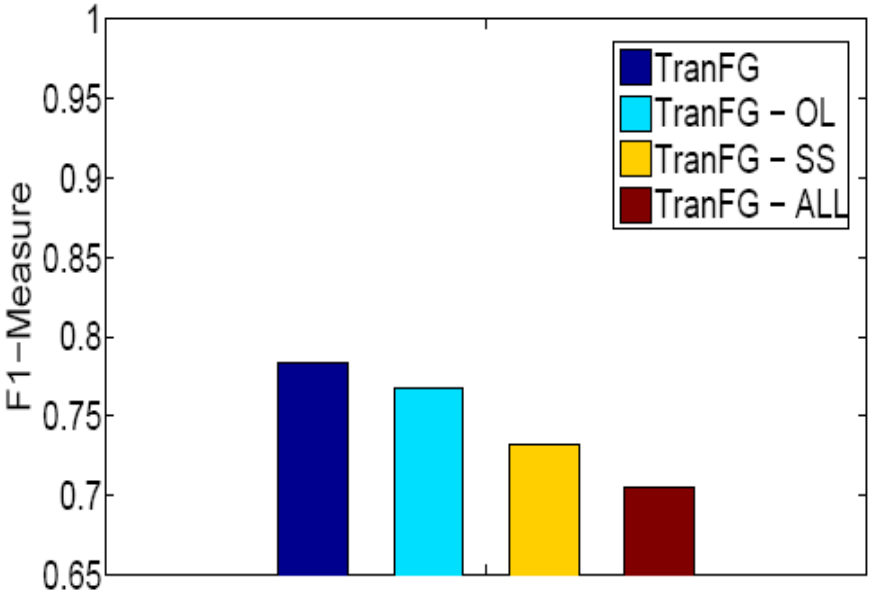
Data Set	Method	Prec.	Rec.	F1-score
Coauthor (S) to Enron (T) (40%)	SVM	0.9524	0.5556	0.7018
	CRF	0.9565	0.5366	0.6875
	PFG	**0.9730**	0.6545	0.7826
	TranFG	0.9556	**0.7818**	**0.8600**
Enron (S) to Coauthor (T) (40%)	SVM	0.6910	0.3727	0.4842
	CRF	**1.0000**	0.3043	0.4666
	PFG	0.9916	0.4591	0.6277
	TPFG	0.5936	**0.7611**	0.6669
	TranFG	0.9793	0.5525	**0.7065**

Factor Contribution Analysis



SH-Structural hole;
SB-Social balance.

Undirected Network



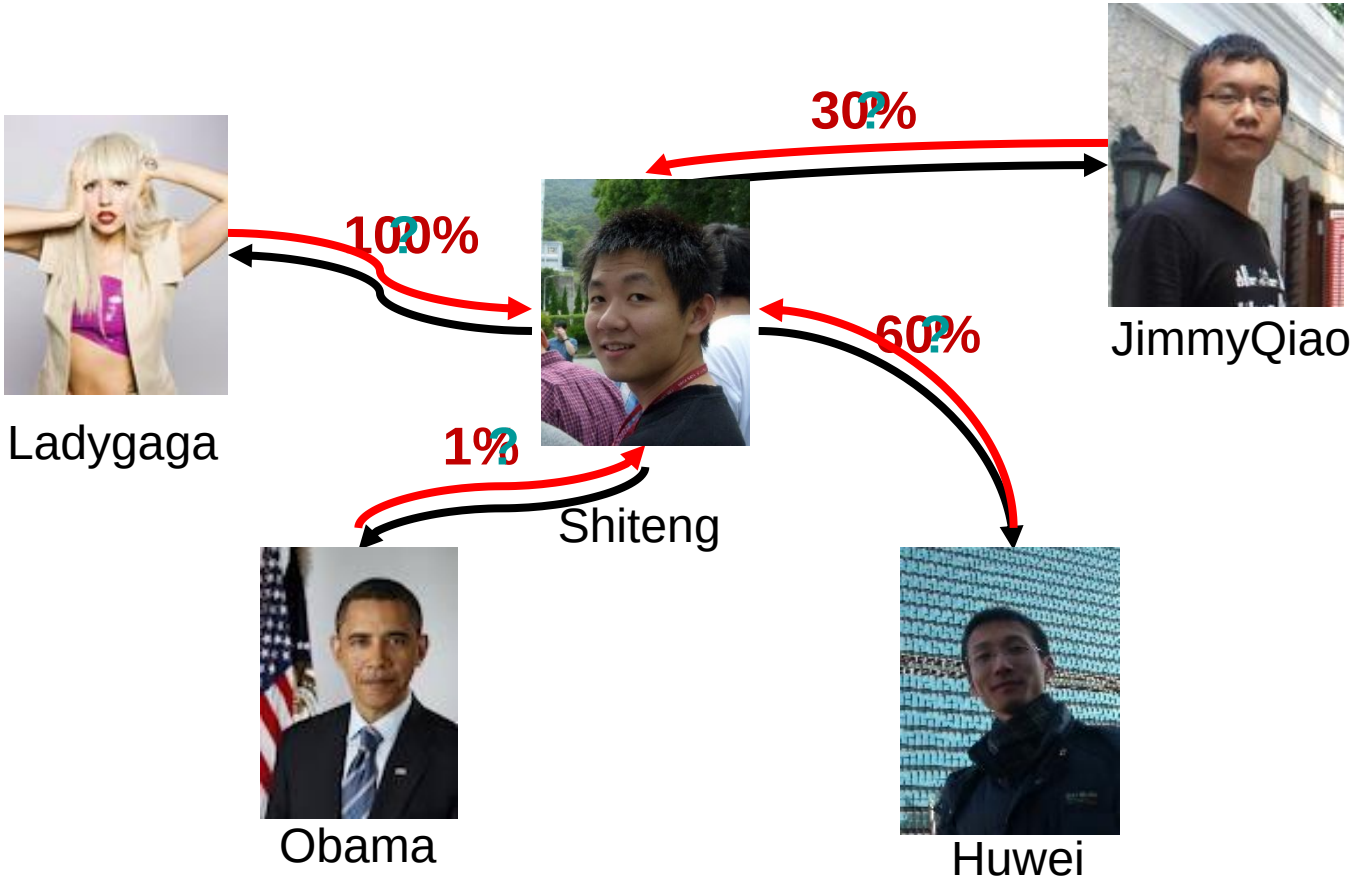
OL-Opinion leader;
SS-Social status.

Directed Network

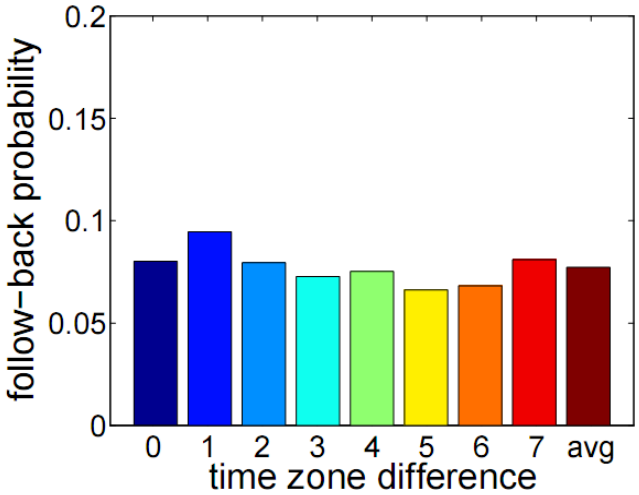
Parasocial vs. Reciprocal

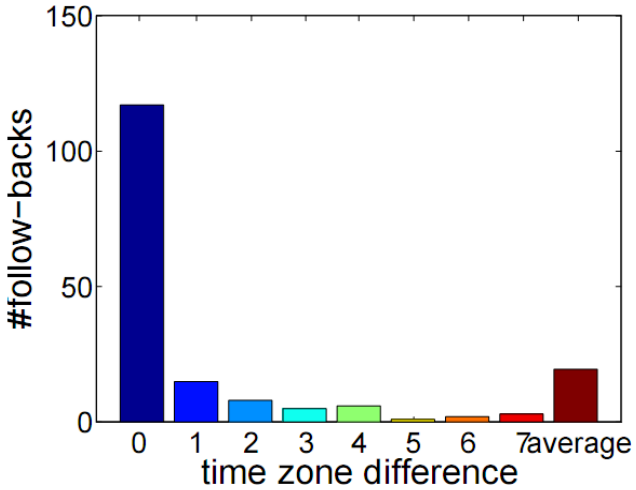
Who will follow you back?

On Twitter…



Geographic Distance

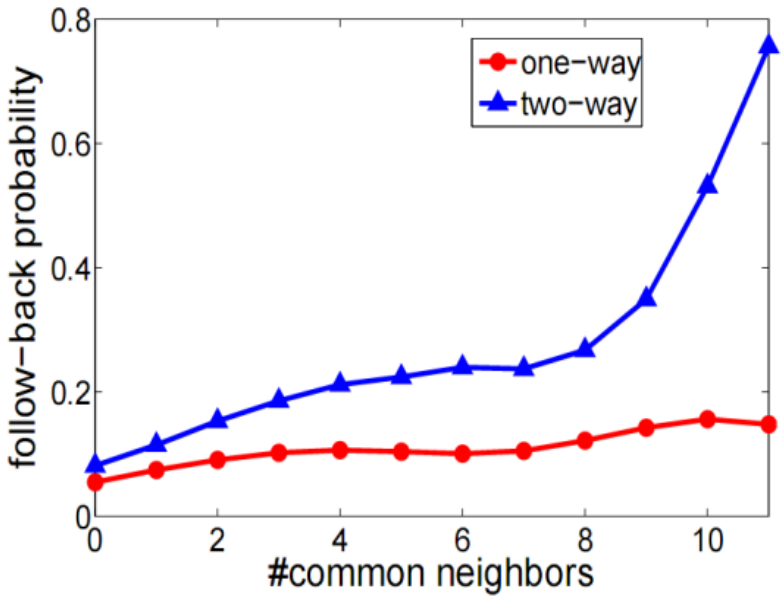


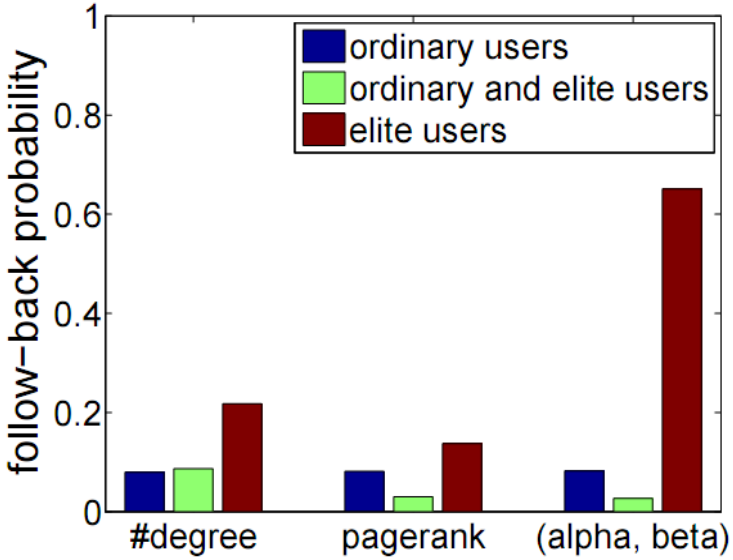


Global vs **Local**

John E. Hopcroft, Tiancheng Lou, and Jie Tang. Who Will Follow You Back? Reciprocal Relationship Prediction. In CIKM'2011. pp. 1137-1146. (alphabet author order)

Homophily

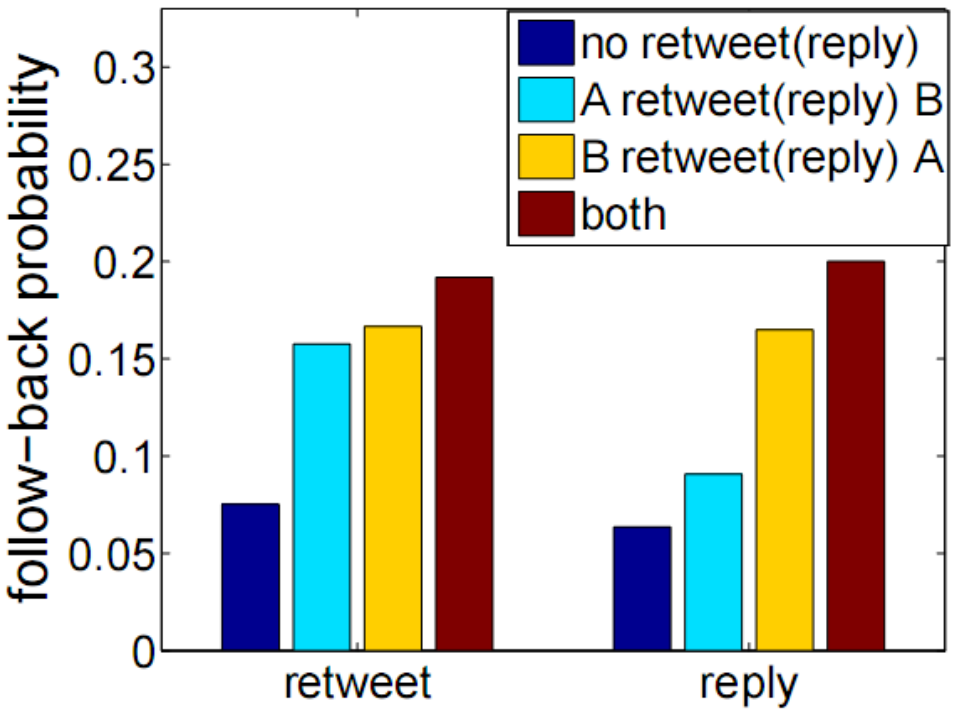




Link homophily: users who share common links will have a tendency to follow each other.

Status homophily: Elite users have a much stronger tendency to follow each other.

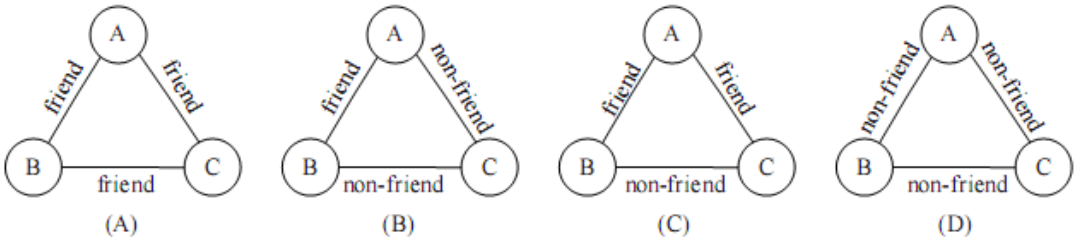
Interaction



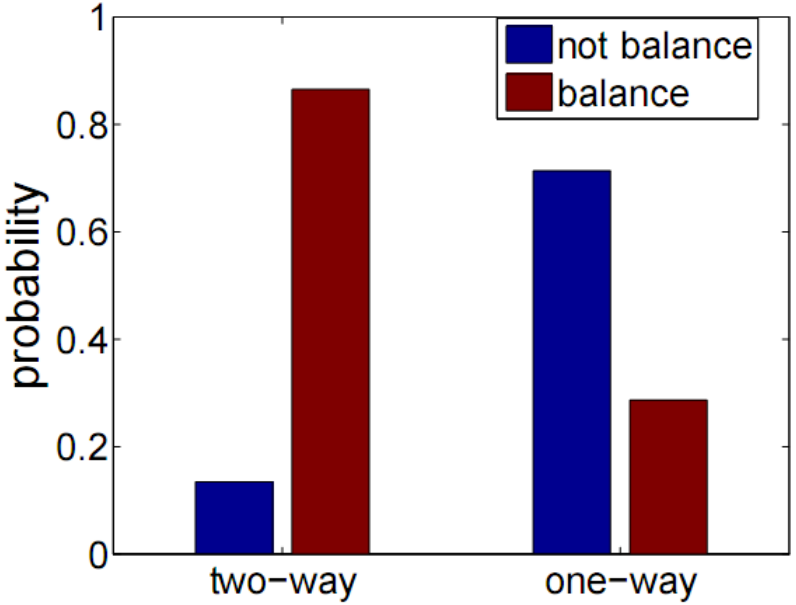
Retweet vs. reply

*Retweeting seems to be more helpful

Structural Balance

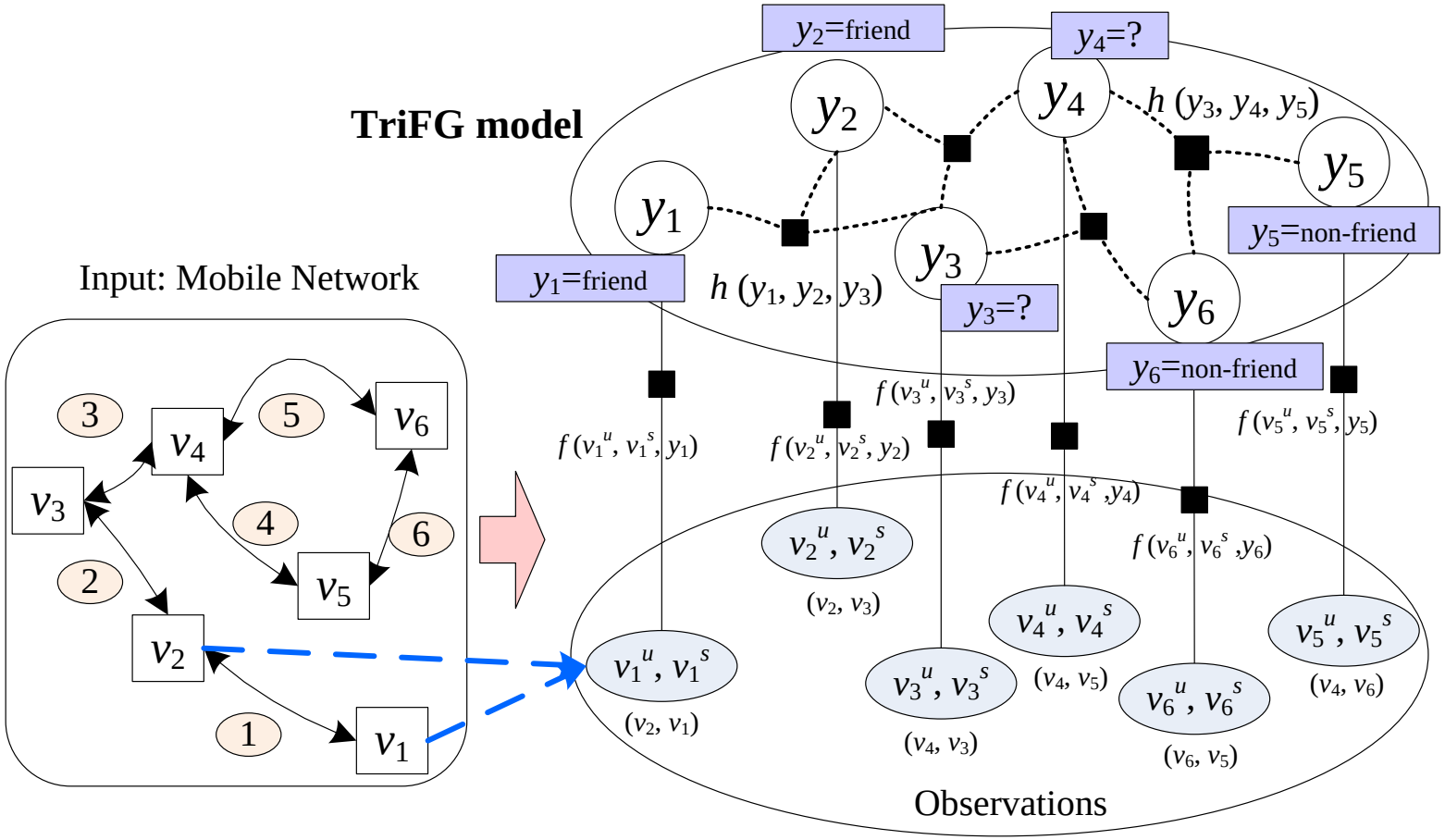


(A) and (B) are balanced, but (C) and (D) are not.



- Structural balance
  - Reciprocal relationships are balanced (**88%**);
  - Parasocial relationships are not (**only 29%**).

Triad Factor Graph (TriFG)



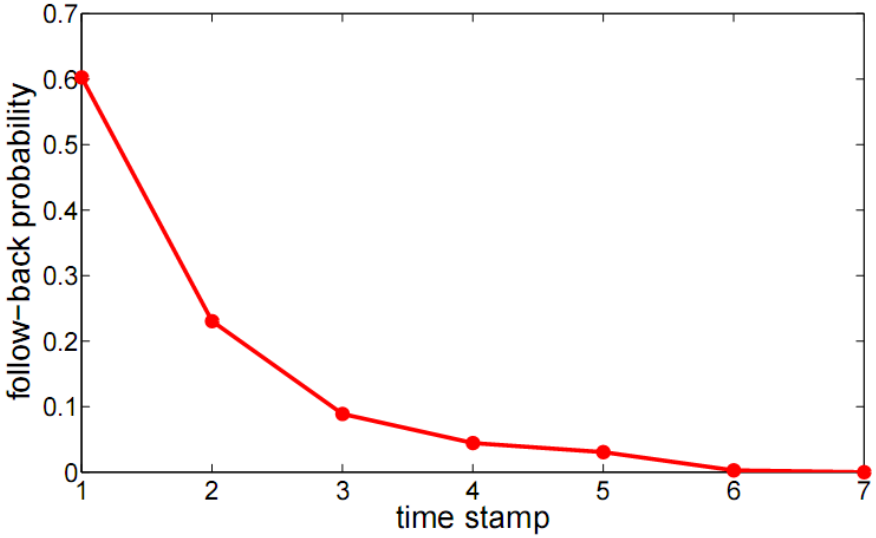
Experiments

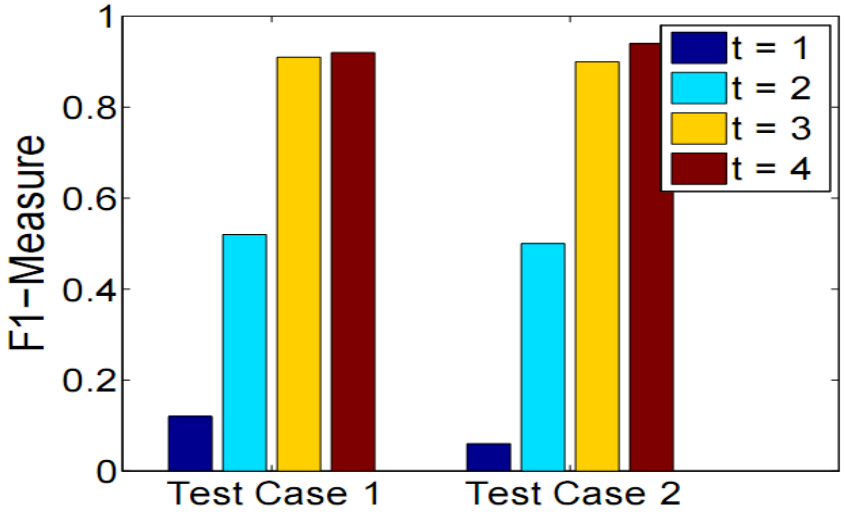
- Huge sub-network of twitter
  - 13,442,659 users and 56,893,234 following links.
  - Extracted 35,746,366 tweets.
- Dynamic networks
  - With an average of 728,509 new links per day.
  - Averagely 3,337 new follow-back links per day.
  - 13 time stamps by viewing every four days as a time stamp

Data	Algotithm	Precision	Recall	F1Measure	Accuracy
Test Case 1	SVM	0.6908	0.6129	0.6495	0.9590
	LRC	0.6957	0.2581	0.3765	0.9510
	CRF	1.0000	0.6290	0.7723	0.9770
	TriFG	1.0000	0.8548	**0.9217**	**0.9910**
Test Case 2	SVM	0.7323	0.6212	0.6722	0.9534
	LRC	0.8333	0.3030	0.4444	0.9417
	CRF	1.0000	0.6333	0.7755	0.9717
	TriFG	1.0000	0.8788	**0.9355**	**0.9907**

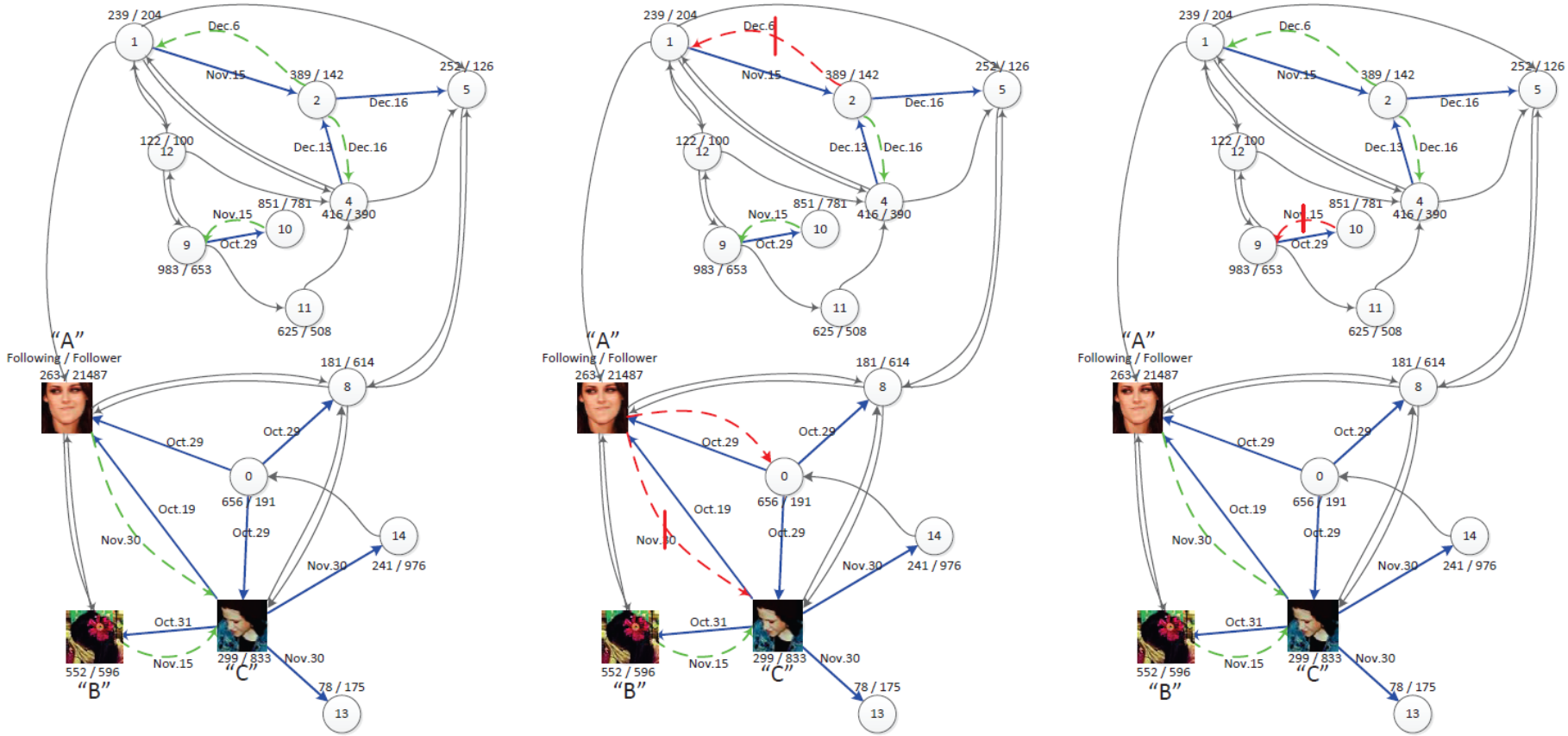
Effect of Time Span

- Distribution of follow back time
  - 60% for next-time stamp;
  - 37% for following 3 time stamps.
- Different settings of the time span
  - Performance drops sharply when two or less;
  - Acceptable for three time stamps.



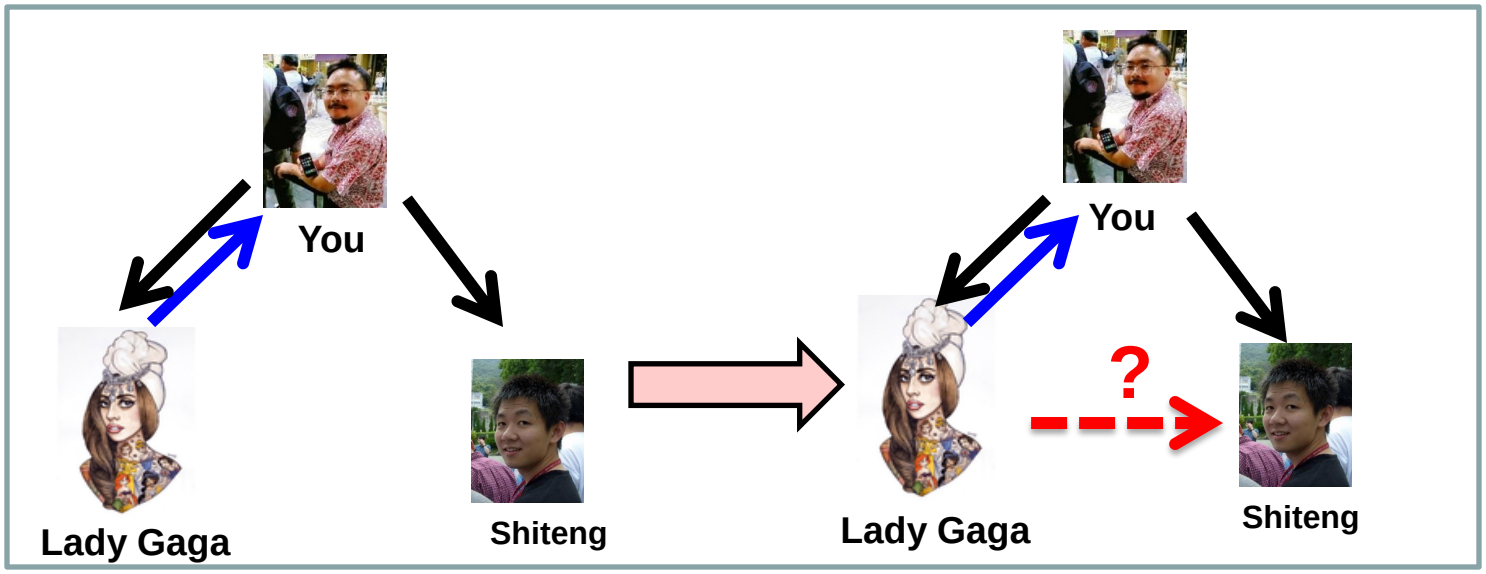


Case Study

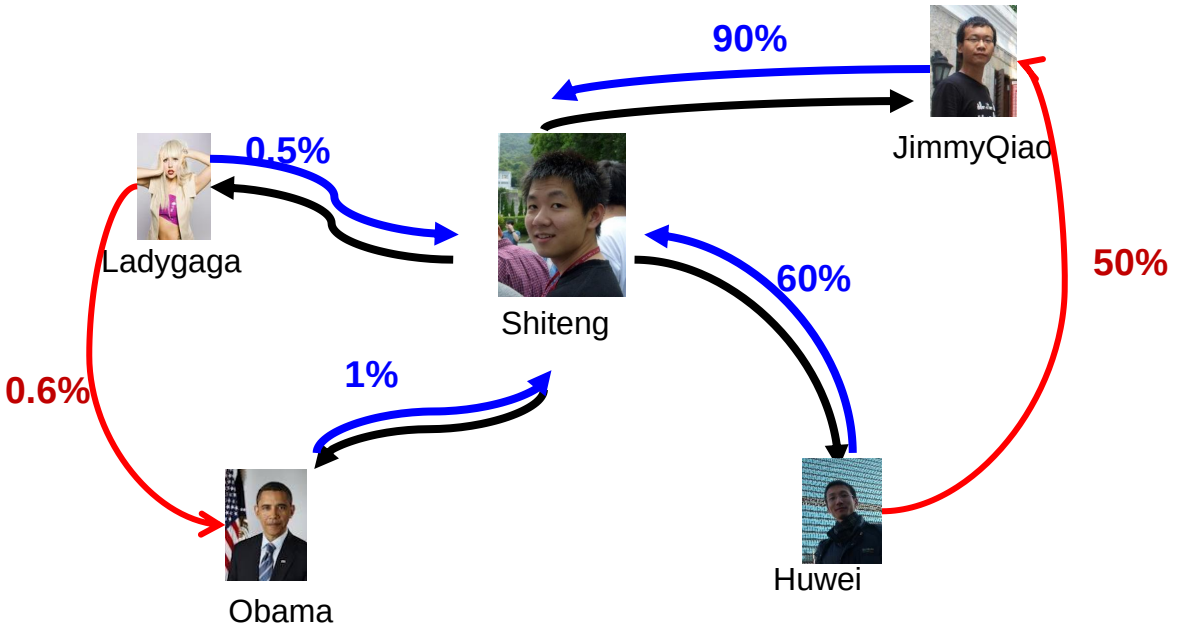


(a) Ground Truth (b) SVM (c) Our approach (TriFG)

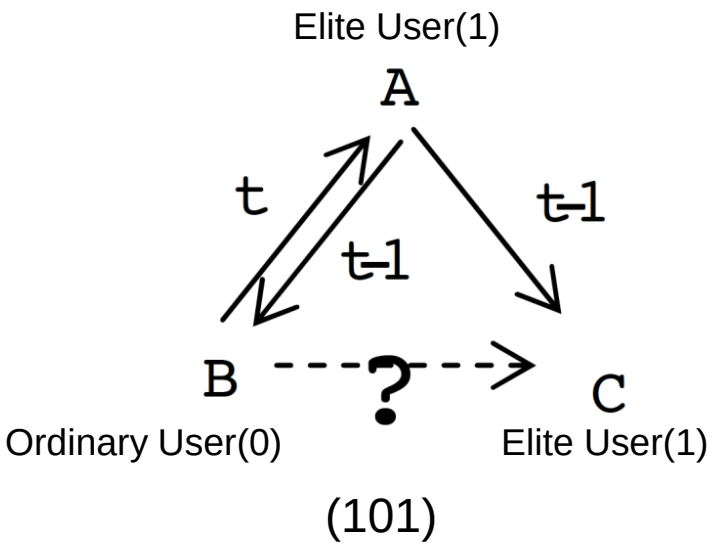
Triadic Closure

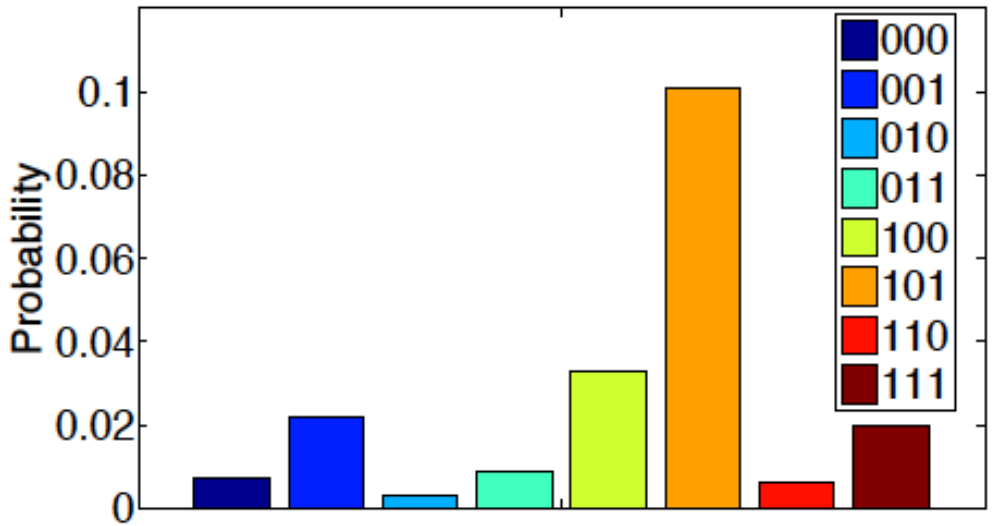


Triadic Closure



Triad Status



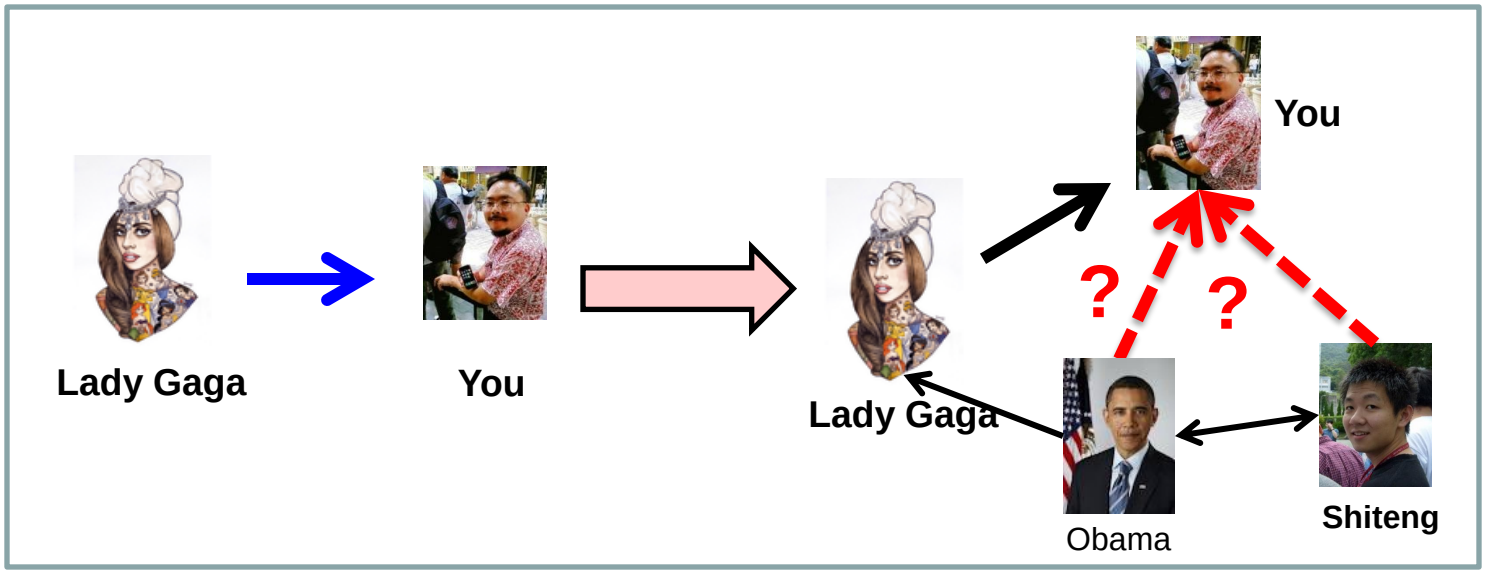


- **P(1XX) > P(0XX)**. Elites users play a more important role to form the triadic closure. The average probability of **1XX** is three times higher than that of **0XX**.
- **P(X0X) > P(X1X)**. Low-status users act as a bridge to connect users so as to form a closure triad. The likelihood of **X0X** is 2.8 times higher than **X1X**.
- **P(XX1) > P(XX0)**. The rich gets richer. This phenomenon validates the mechanism of preferential attachment [Newman 2001].

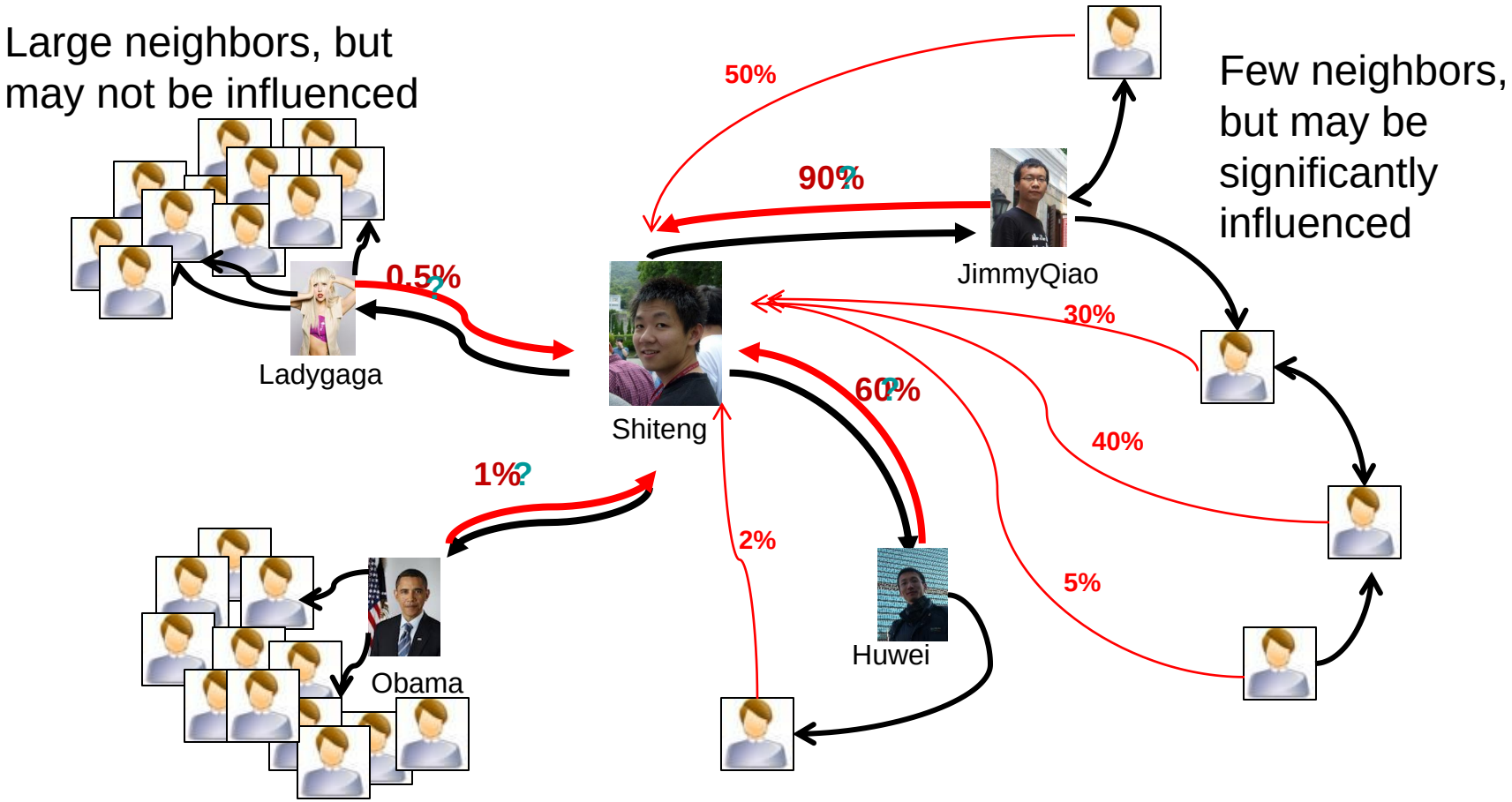
Triad Closure Prediction Result

Data	Algotithm	Precision	Recall	F1Measure
Test Case 1	SVM	0.0870	0.1429	0.1081
	LRC	0.0536	0.1404	0.0759
	CRF-balance	0.0208	0.0436	0.0282
	CRF	0.1111	0.0870	0.0976
	wTriFG	0.3333	0.0373	0.0671
	TriFG	**0.4545**	**0.2174**	**0.2941**
Test Case 2	SVM	0.2000	0.2222	0.2105
	LRC	0.1071	0.1667	0.1304
	CRF-balance	0.0909	0.0556	0.0690
	CRF	0.2222	0.2222	0.2222
	wTriFG	0.5000	0.0556	0.1000
	TriFG	**0.8571**	**0.3333**	**0.4800**

Follow Influence



Will the “following” be Influenced?



Influence Test

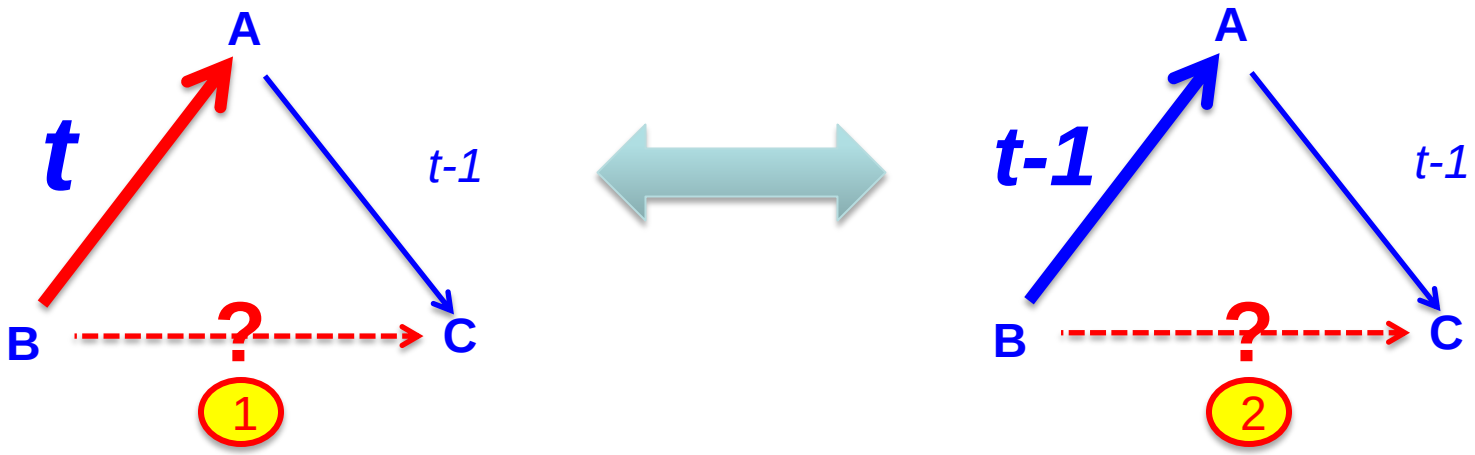
Question:

Whether there exist follow influence?

In which kind of triad the influence is significant?

Method:

Compare the same kind of triad with different timestamp.



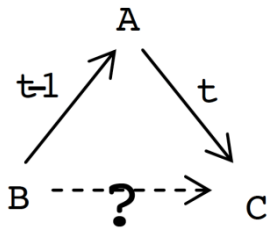
Assumption:

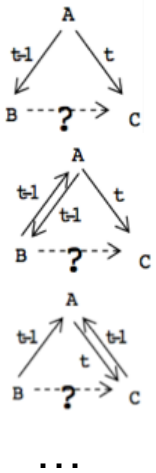
If P1(B->C) is much larger than P2(B->C), then influence exists.

Test Result

Two categories of triads have significant influence, compared with two other categories

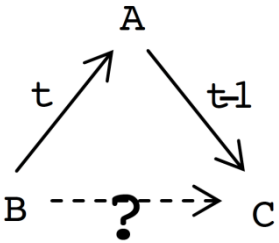
Attract more followers

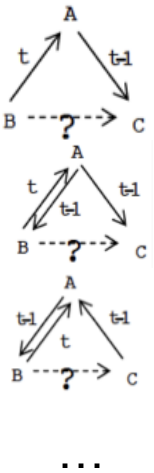




P1(B->C)=0.5%
P2(B->C)=0.1%

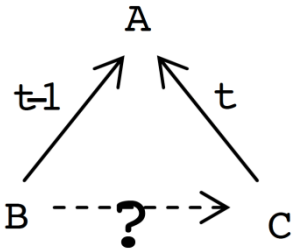
Follow More





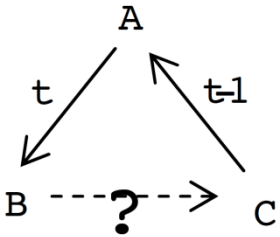
P1(B->C)=14.4%
P2(B->C)=0.1%

No influence



P1(B->C)=0.02%
P2(B->C)=0.02%

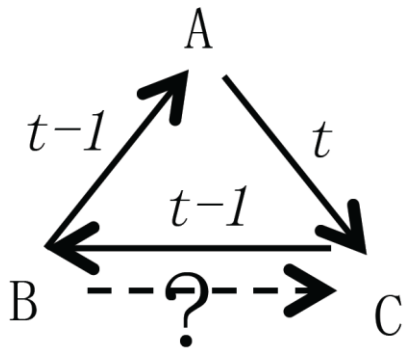
No influence



P1(B->C)=0.02%
P2(B->C)=0.02%

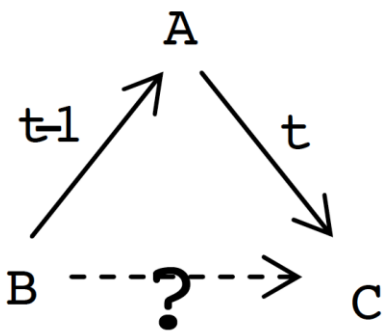
More…

P(B->C) is significantly boosted when the reversed follow link is pre-formed

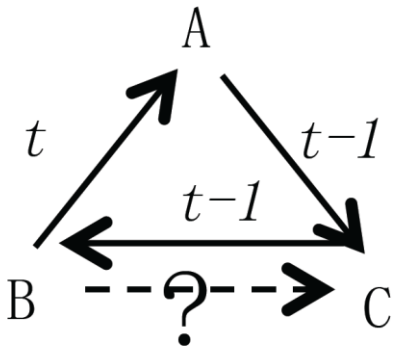


P1(B->C)=4.1%

>

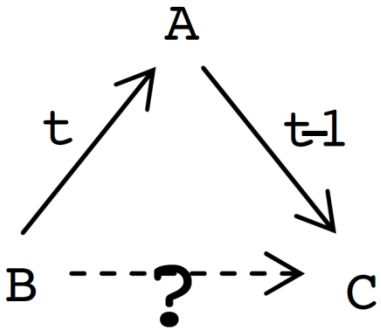


P1(B->C)=0.5%



P1(B->C)=81.7%

>

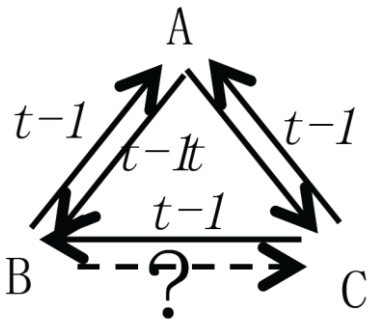


P1(B->C)=14.4%

Question: Are there any other factors that can boost P(B->C)?

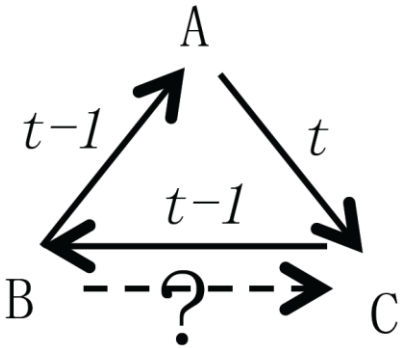
Structural Balance

P(B->C) is significantly boosted when the the resultant triad satisfies the balance theory



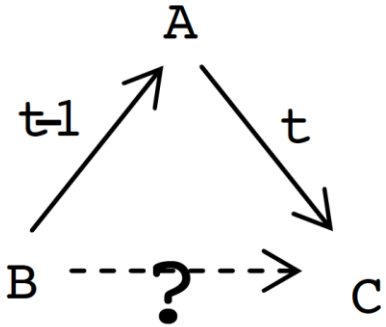
P1(B->C)=15.9%

>

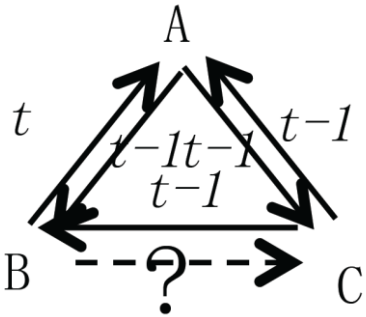


P1(B->C)=4.1%

>

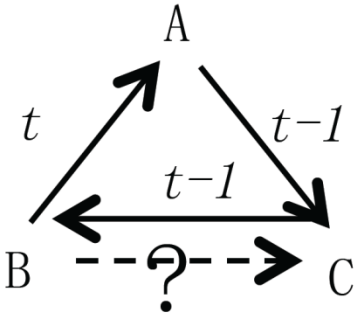


P1(B->C)=0.5%



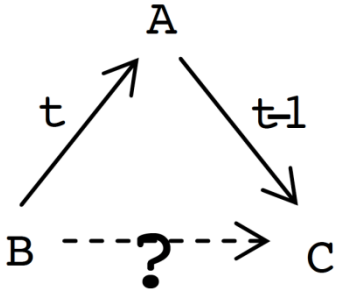
P1(B->C)=86.8%

>



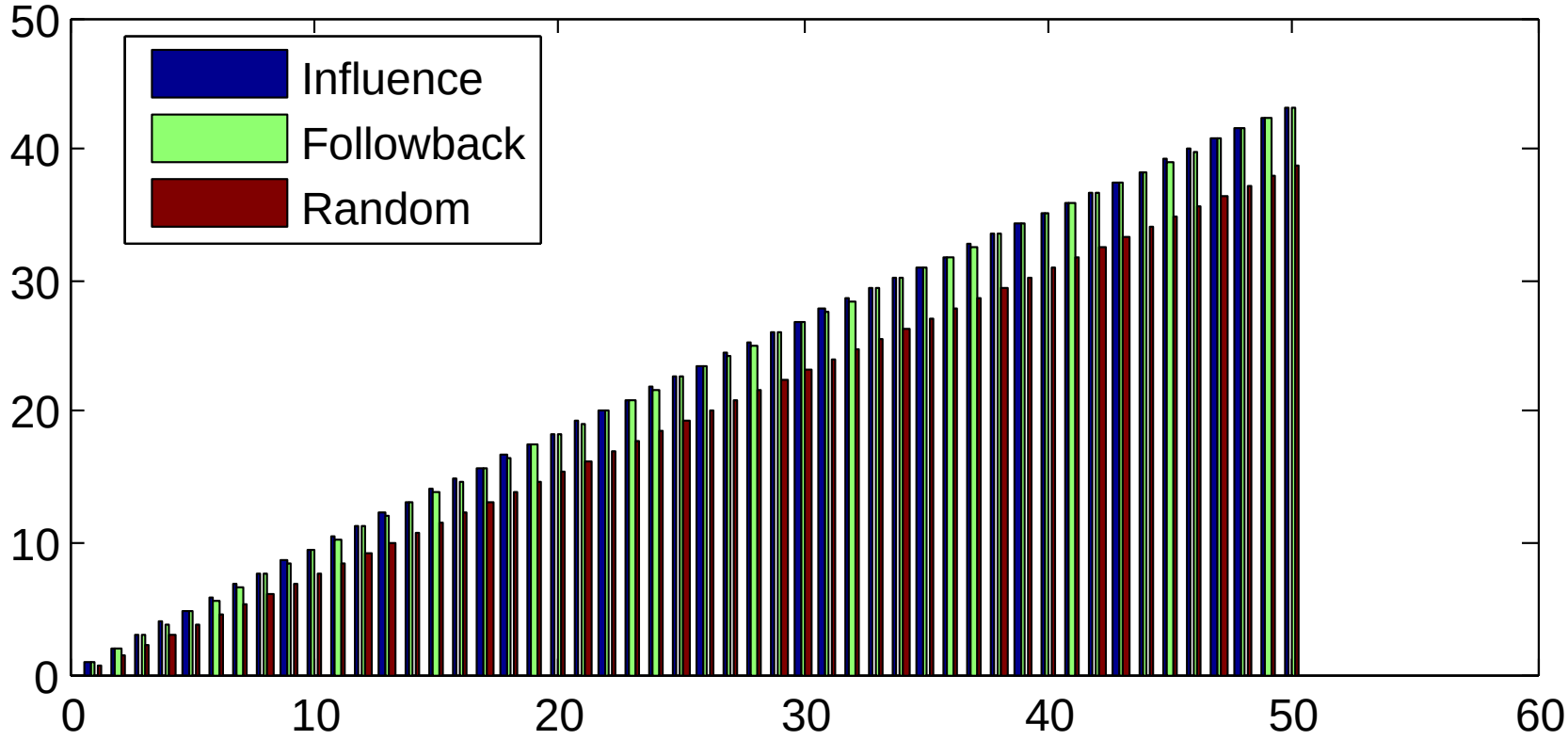
P1(B->C)=81.7%

>



P1(B->C)=14.4%

Application: Follow Influence Maximization



- Influence: Select seeds which can influence most users
- Followback: Select seeds which can follow back with the highest probabilities
- Random: Select seeds randomly

Summary

- Computational models for social tie analysis
  - Inferring social tie
  - Parasocial ->Reciprocal
  - Tradic closure
  - Follow influence
- This is just a start for social tie analysis
  - How social tie influences user behaviors?
  - How social tie influences the network structure?
  - …

Related Publications

- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure. **ACM Transactions on Knowledge Discovery from Data** (**TKDD**), (accepted).
- Honglei Zhuang, Jie Tang, Wenbin Tang, Tiancheng Lou, Alvin Chin, and Xia Wang. Actively Learning to Infer Social Ties. In **Data Mining and Knowledge Discovery** (**DMKD**), 2012, Volume 25, Issue 2, pages 270-297.
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. **KDD'12**. pp. 1285-1293.
- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogenous Networks. **WSDM'12**.
- Wenbin Tang, Honglei Zhuang, and Jie Tang. Learning to Infer Social Relationships in Large Networks. **PKDD'11.** **(Best Student Paper Runner-up)**
- John Hopcroft, Tiancheng Lou, and Jie Tang. Who Will Follow You Back? Reciprocal Relationship Prediction. **CIKM'11**. pp. 1137-1146.
- Chi Wang, Jiawei Han, Yuntao Jia, Duo Zhang, Yintao Yu, Jie Tang, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**.

Thank you !

QA?

Data & Code:

http://arnetminer.org/socialtieacross

http://arnetminer.org/socialtie

http://arnetminer.org/reciprocity