Inferring Social Ties across Heterogeneous Networks

Jie Tang*, Tiancheng Lou*, and Jon Kleinberg+

*Tsinghua University

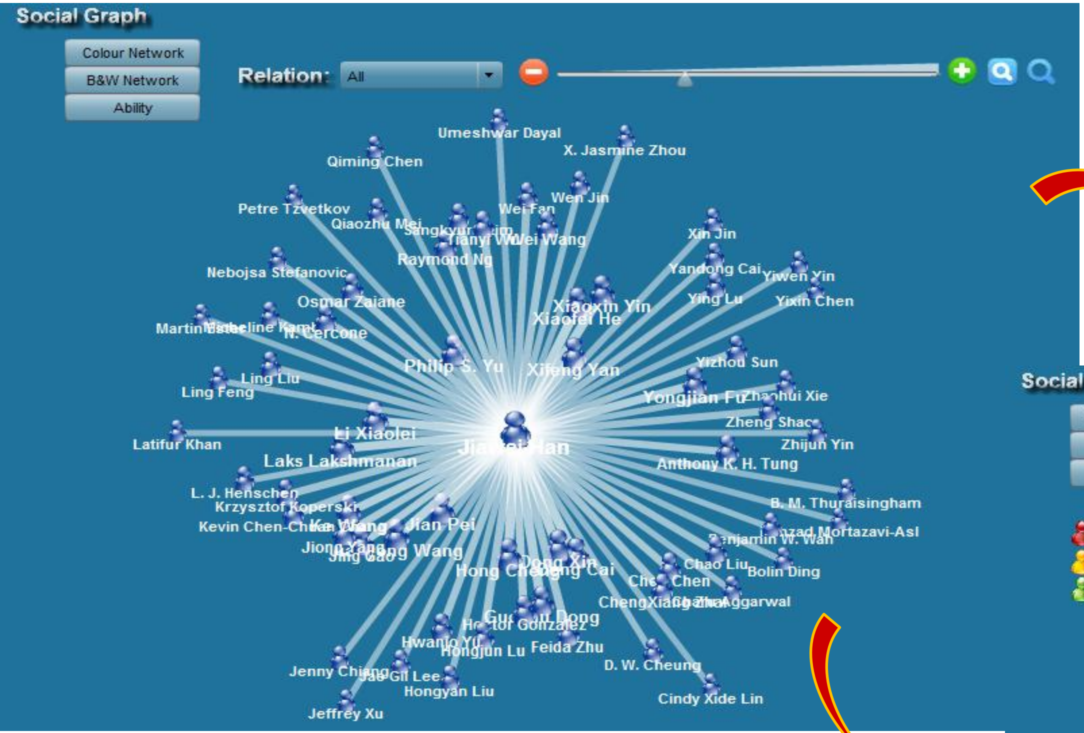
+Cornell University

Real social networks are complex...

- Different social ties have different influence on people
  - Close friends vs. Acquaintances
  - Colleagues vs. Family members
- 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+

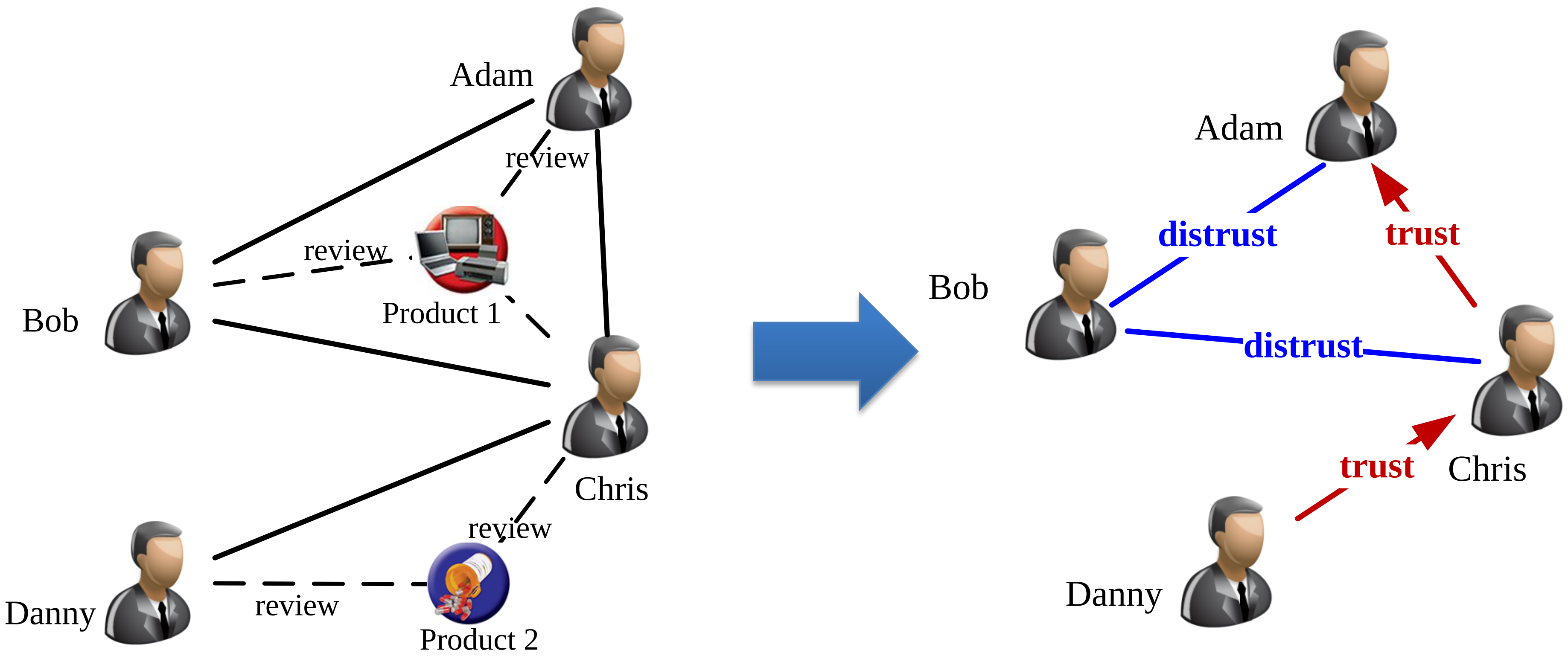
which circle? Users do not take time to create it.

Example 1. Advisor-advisee relationship



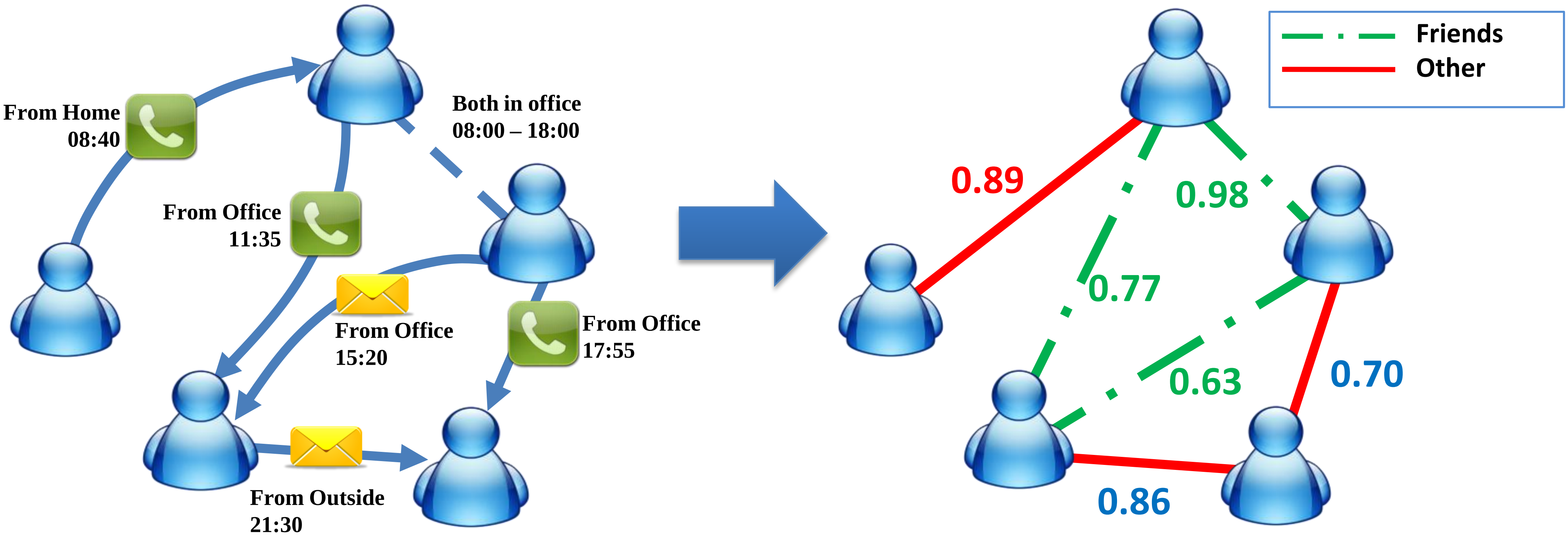
Arnetminer

Example 2. Trustful relationship



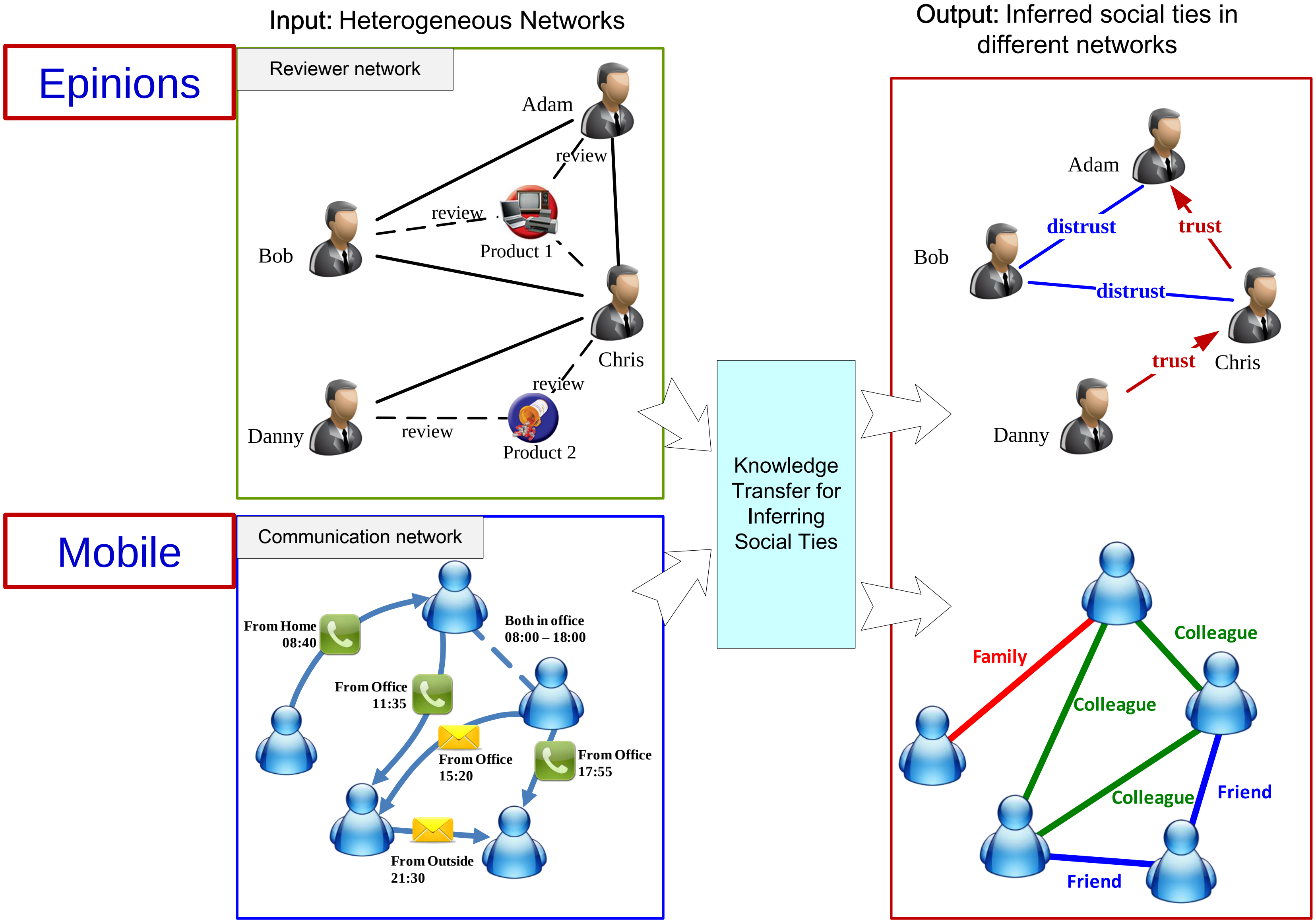
Epinions

Example 3: Friendship in mobile network



Mobile

Inferring Social Ties Across Networks

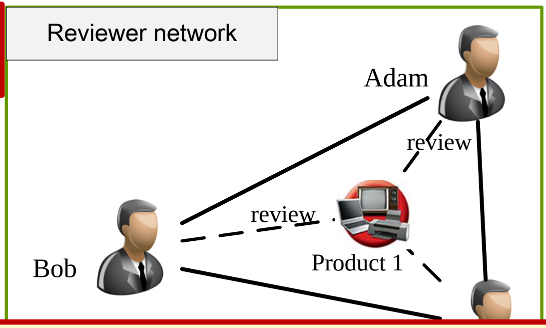


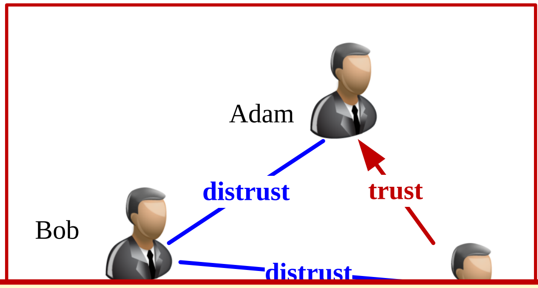
Inferring Social Ties Across Networks

Input: Heterogeneous Networks

Output: Inferred social ties in different networks

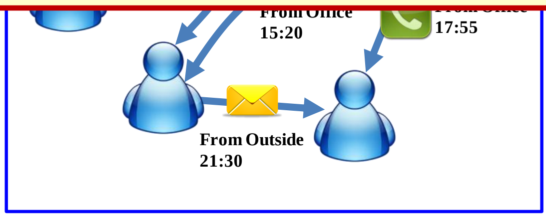
Epinions

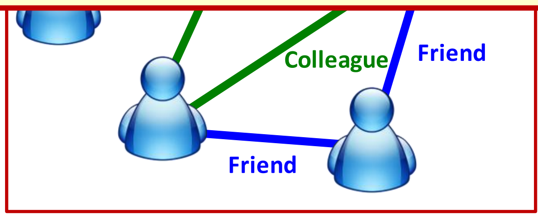




Questions:

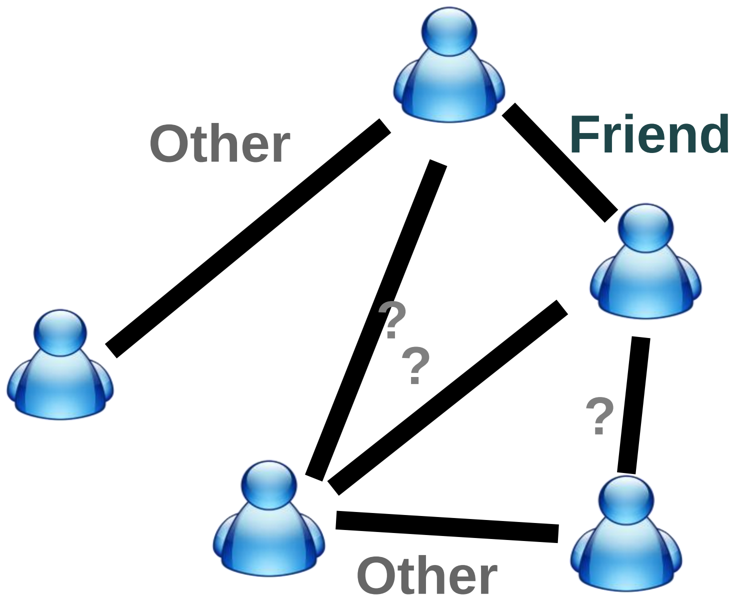
- What are the fundamental forces behind?
- A generalized framework for inferring social ties?
- How to connect the different networks?





Problem Formulation in a Single Network

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



V: Set of Users

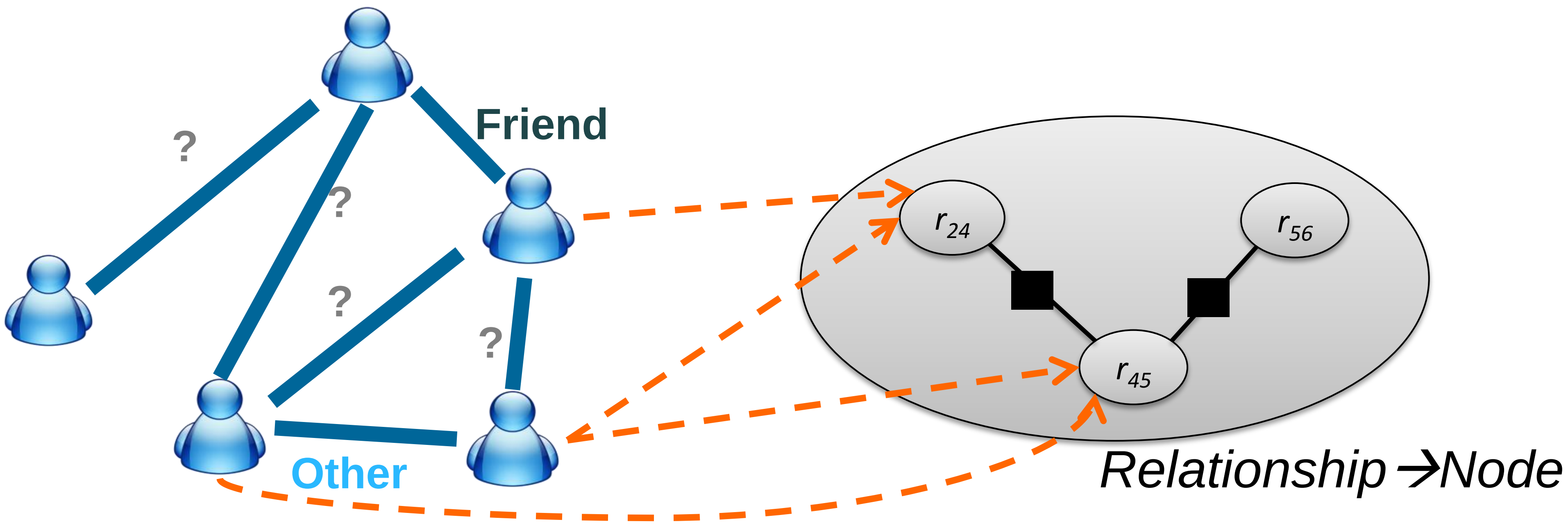
E^L,R^L: Labeled relationships

E^U: Unlabeled relationships

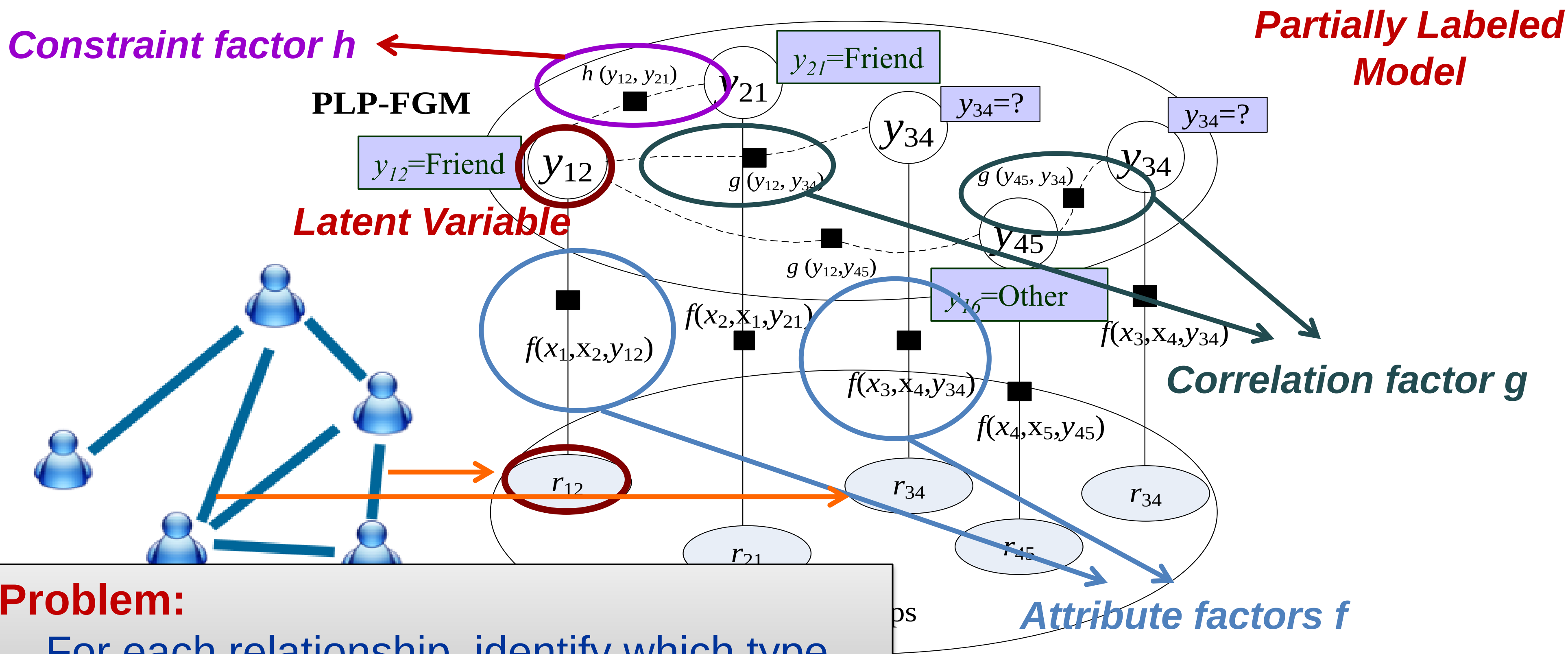
Input:
$G=(V,E^L,E^U,R^L,W)$

Output:
$f: G \rightarrow R$

Basic Idea



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.

Solutions$_{\text{(con't)}}$

- 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:

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

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

  - Log-Likelihood of labeled Data:

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

Parameters to estimate $\theta = [\lambda, \alpha, \beta], s = [\Phi^T, g^T, h^T]^T$

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

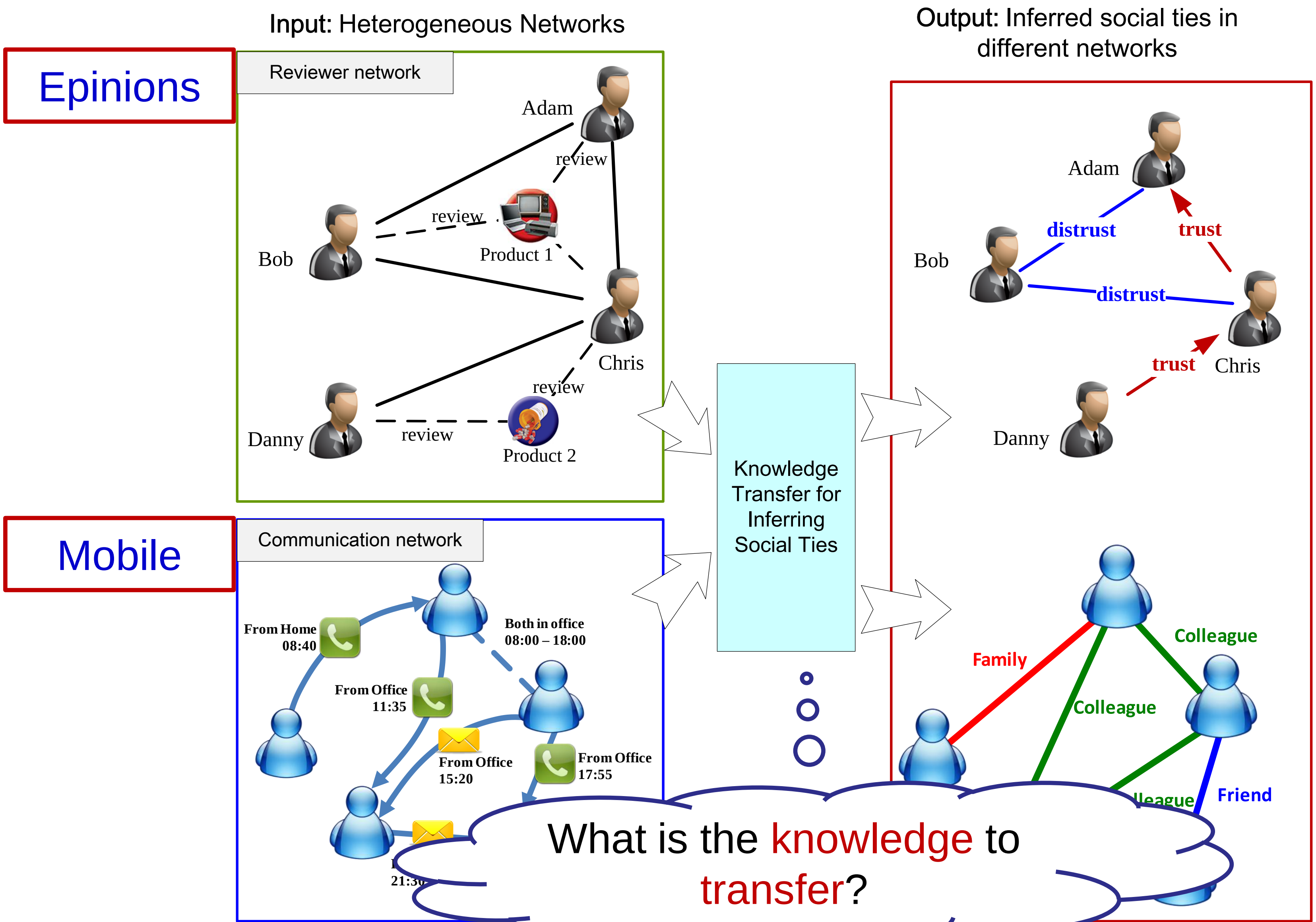
Gradient Ascent Method

Still Challenges?

Questions:

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

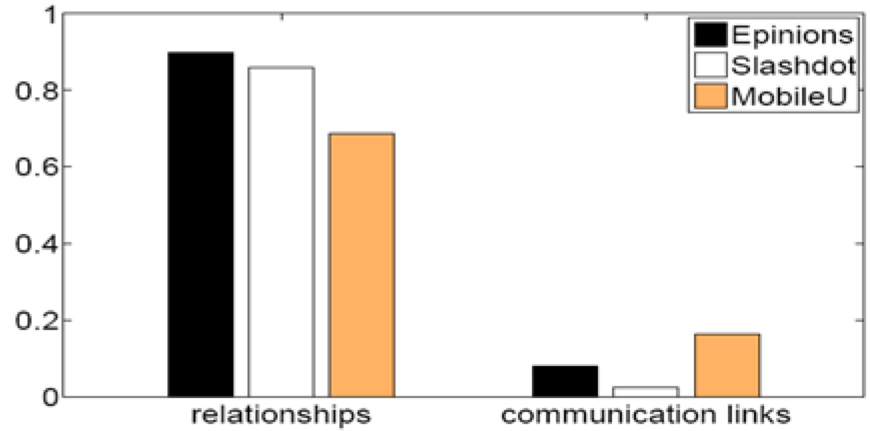
Inferring Social Ties Across Networks



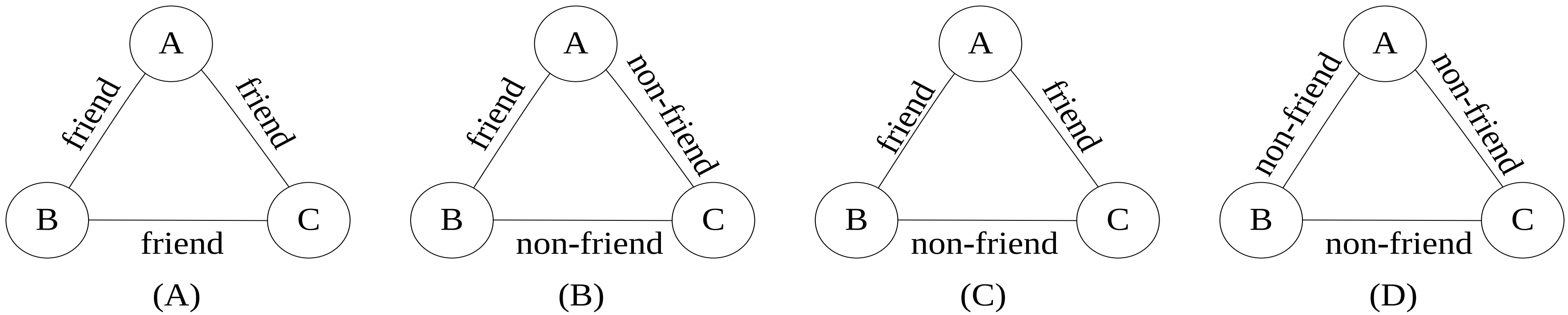
What is the knowledge to transfer?

Social Theories

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

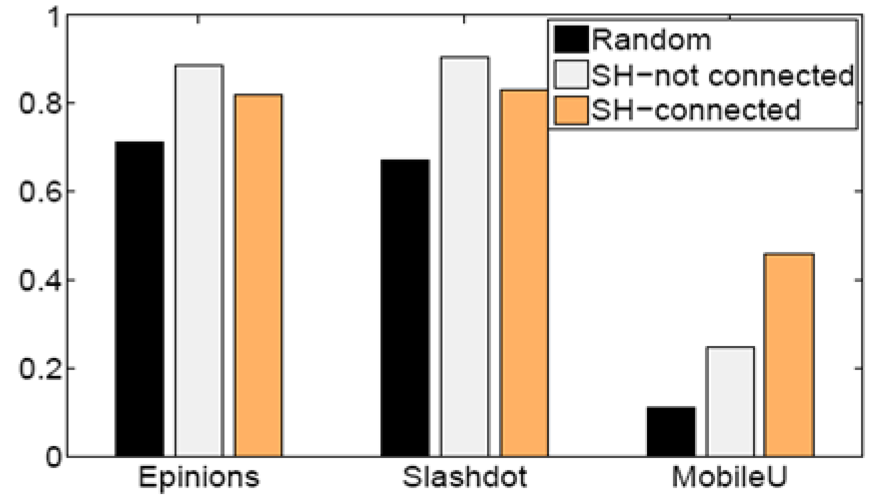


Observations:
(1) The underlying networks are unbalanced;
(2) While the friendship networks are balanced.

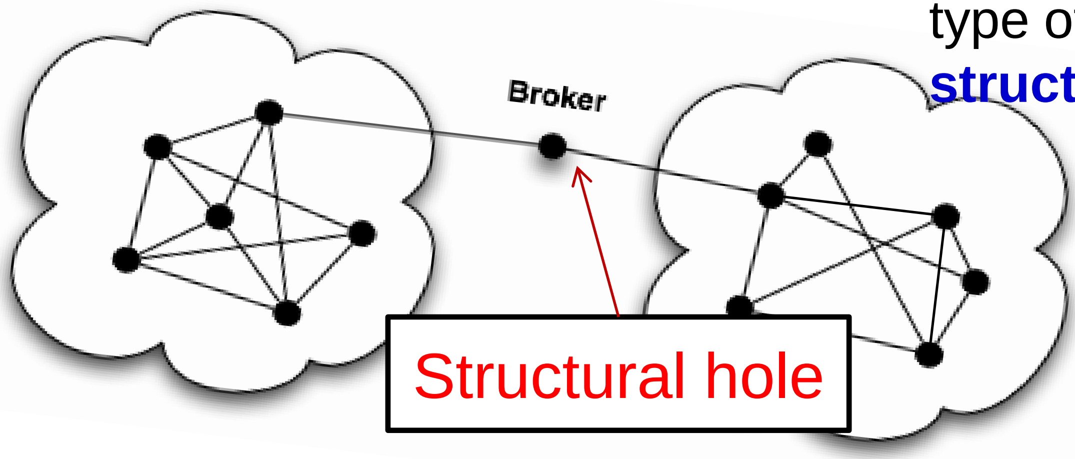


Social Theories—Structural hole

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

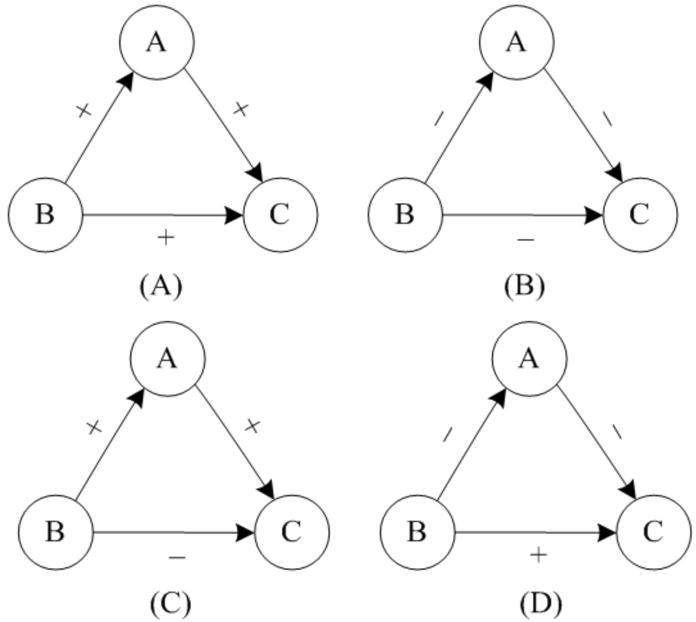


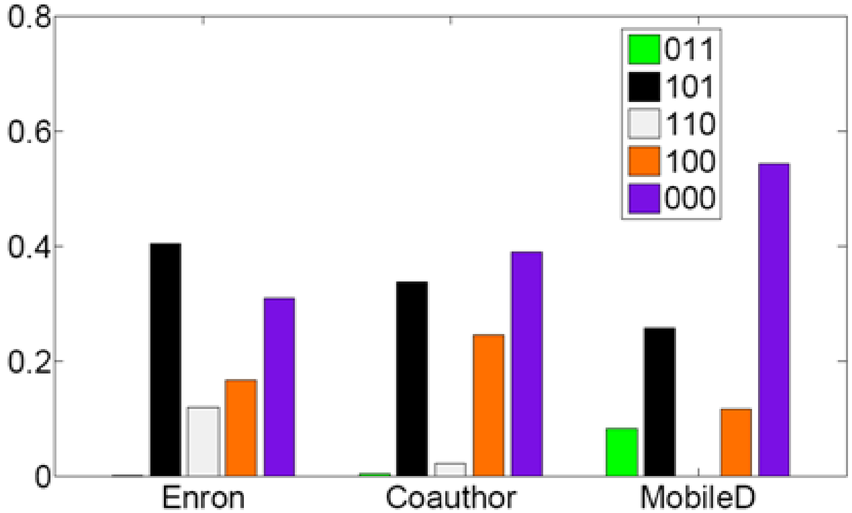
Observations: Users are more likely (+25-150% higher than change) to have the same type of relationship with C if C **spans structural holes**



Social Theories—Social status

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



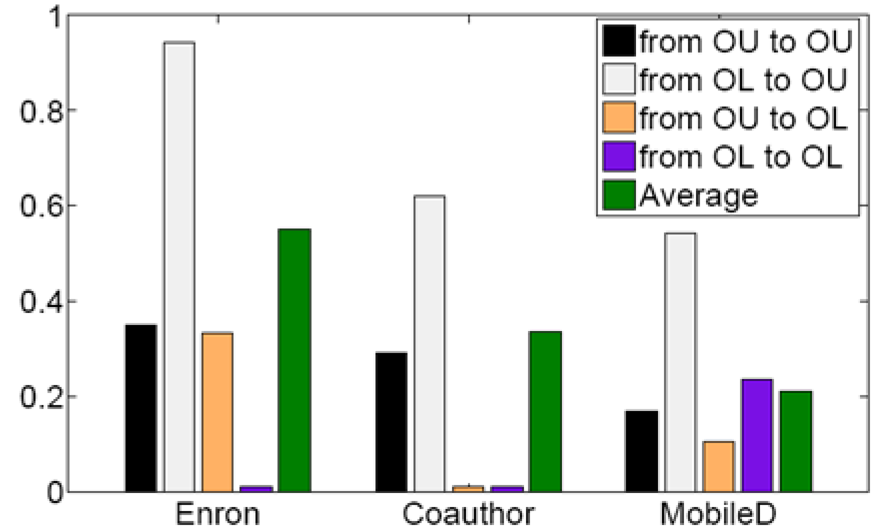


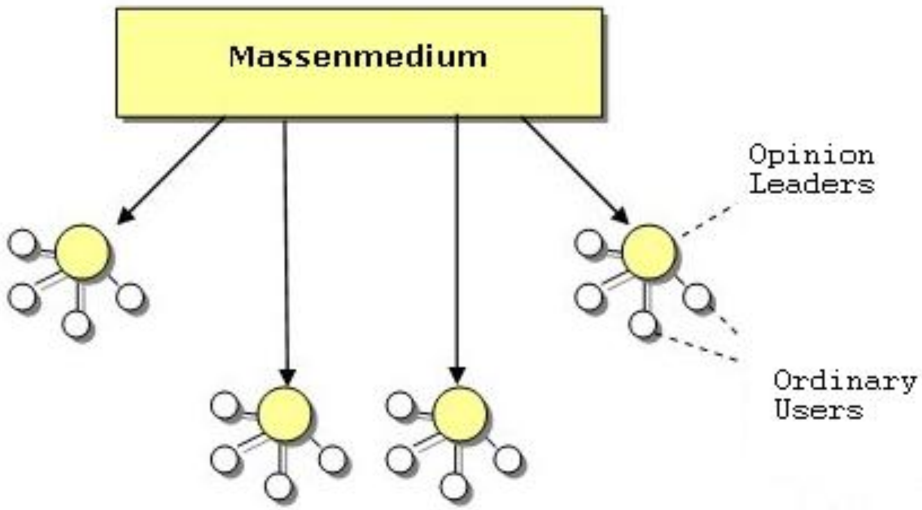
Observations: 99% of triads in the networks satisfy the social status theory

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



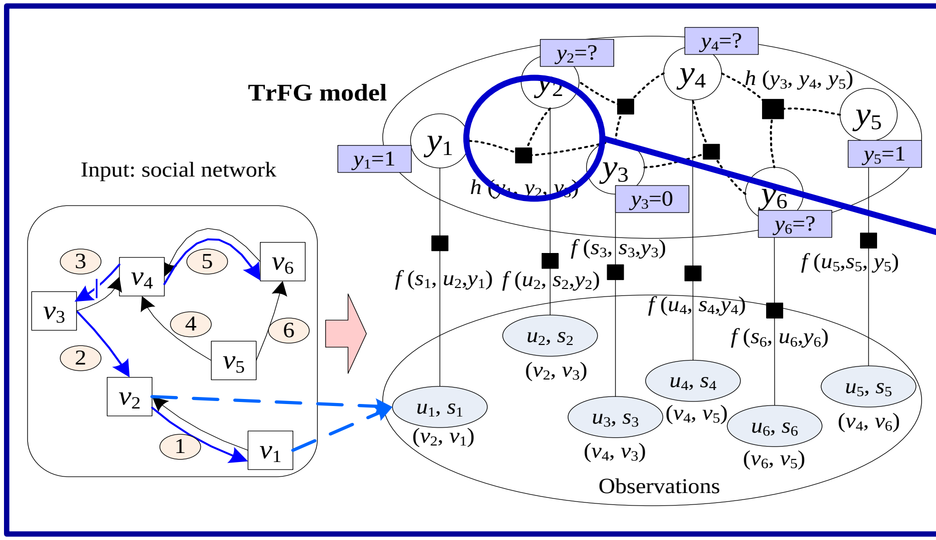


OL : Opinion leader;
OU : Ordinary user.

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

Transfer Factor Graph Model

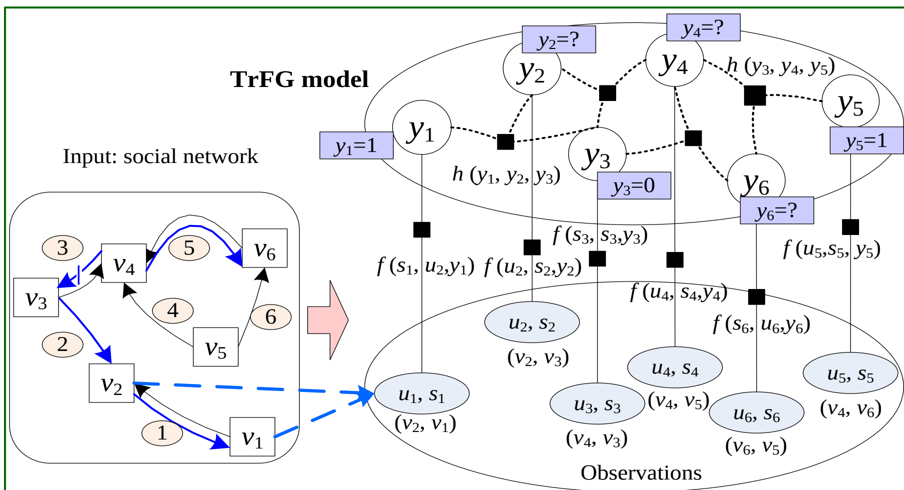
Coauthor network



Triad-based factor

Bridge via social theories

mobile



Mathematical Formulation

Features defined in source network

Features defined in target network

$$
\begin{aligned}
\mathcal{O}(\alpha, \beta, \mu) &= \mathcal{O}_S(\alpha, \mu) + \mathcal{O}_T(\beta, \mu) \\
&= \sum_{i=1}^{|V_S|} \sum_{j=1}^{d} \boxed{\alpha_j g_j(x_{ij}^S, y_i^S)} + \sum_{i=1}^{|V_T|} \sum_{j=1}^{d'} \boxed{\beta_j g'_j(x_{ij}^T, y_i^T)} \\
&\quad + \sum_k \mu_k \Big(\sum_{c \in G_S} \boxed{h_k(Y_c^S)} + \sum_{c \in G_T} \boxed{h_k(Y_c^T)} \Big) \\
&\quad - \log Z
\end{aligned}
$$

Triad-based features shared across networks

Data Sets

- **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 between users

Directed network

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

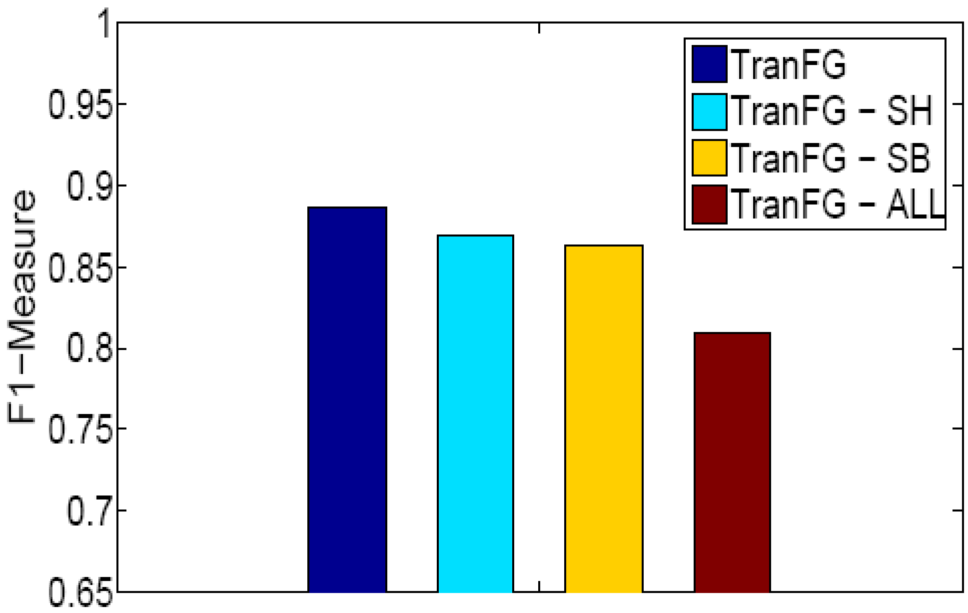
SVM and **CRF** are two baseline methods

PFG is the proposed partially-labeled factor graph model

TranFG is the proposed transfer–based factor graph model.

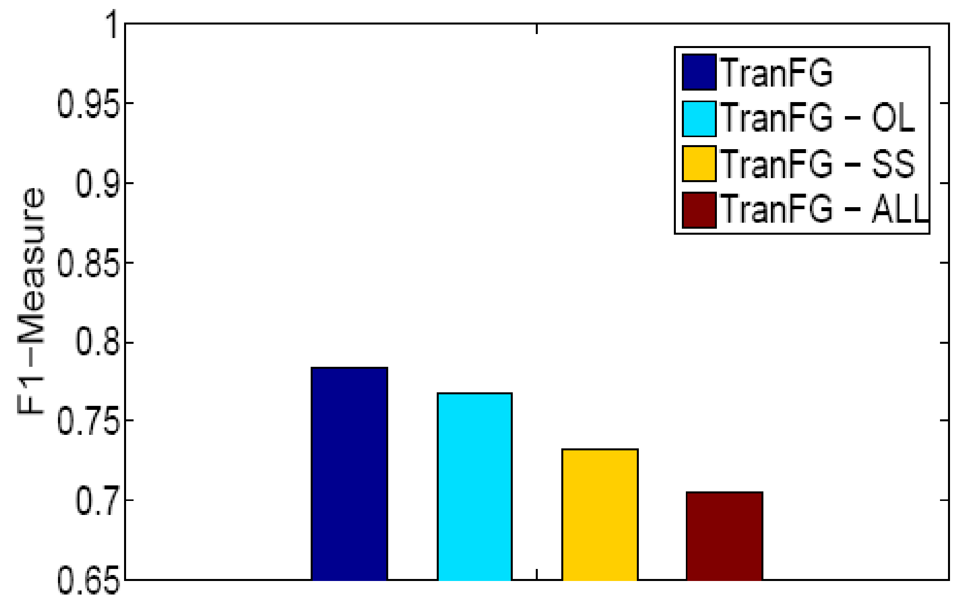
Data Set	Method	Prec.	Rec.	F1-score
Coauthor (S) to Enron (T) (40%)	SVM	0.9524	0.5556	0.7018
	CRF	0.9565	0.5366	0.6875
	PFG	**0.9730**	0.6545	0.7826
	TranFG	0.9556	**0.7818**	**0.8600**
Enron (S) to Coauthor (T) (40%)	SVM	0.6910	0.3727	0.4842
	CRF	**1.0000**	0.3043	0.4666
	PFG	0.9916	0.4591	0.6277
	TPFG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

Conclusions and Future Work

- **Conclusions**
  - different types of social ties have essentially different structural patterns in social networks;
  - By incorporating social theories, our proposed model can significantly improve (+4-14%) the inferring accuracy.

- **Future work**
  - Inferring complex relationships between users, e.g., family, colleague, manager-subordinate;
  - Active learning for inferring social ties.

Thanks!

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

System: http://arnetminer.org

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

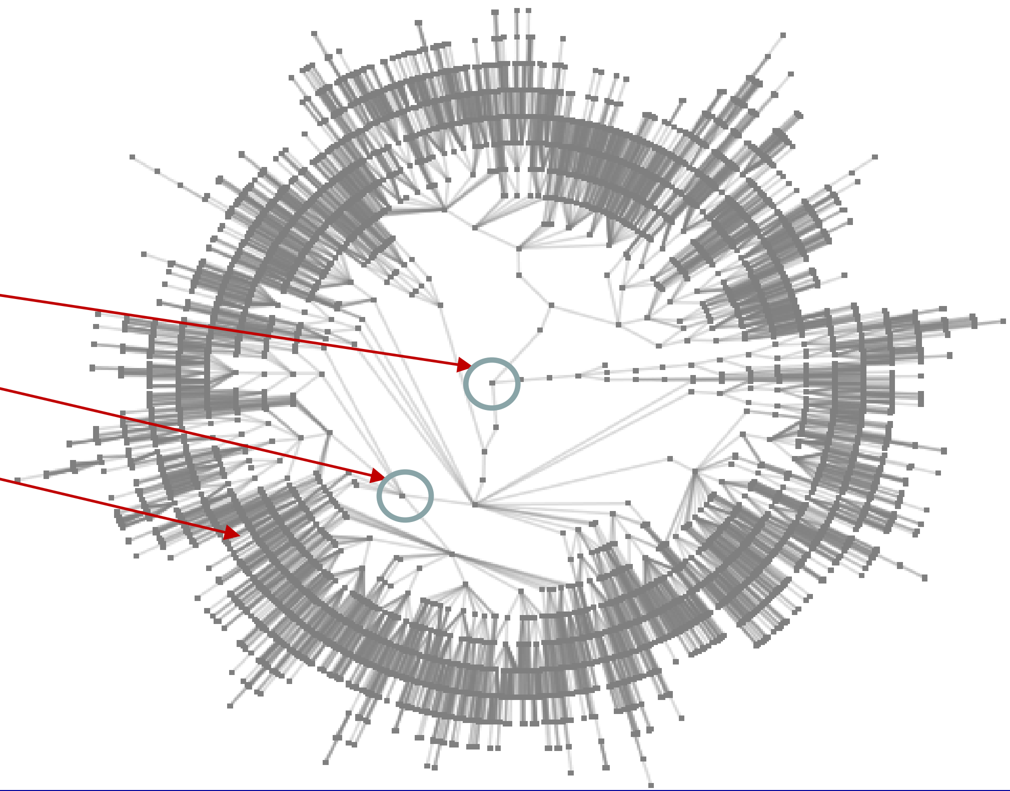
- The fact is that our social network is **black**-white…

Example 2. Manager-employee relationship

Enterprise email network

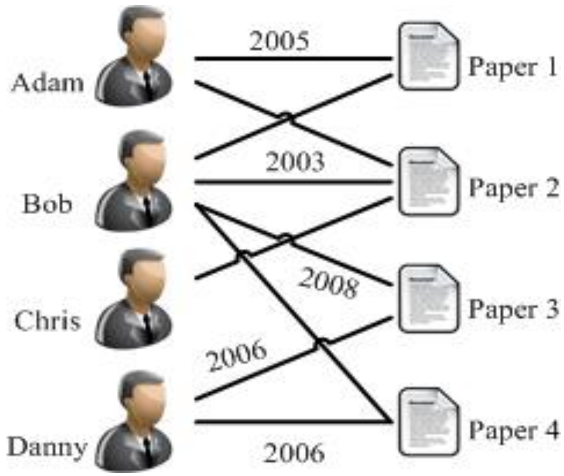
How to infer

- CEO
- Manager
- Employee

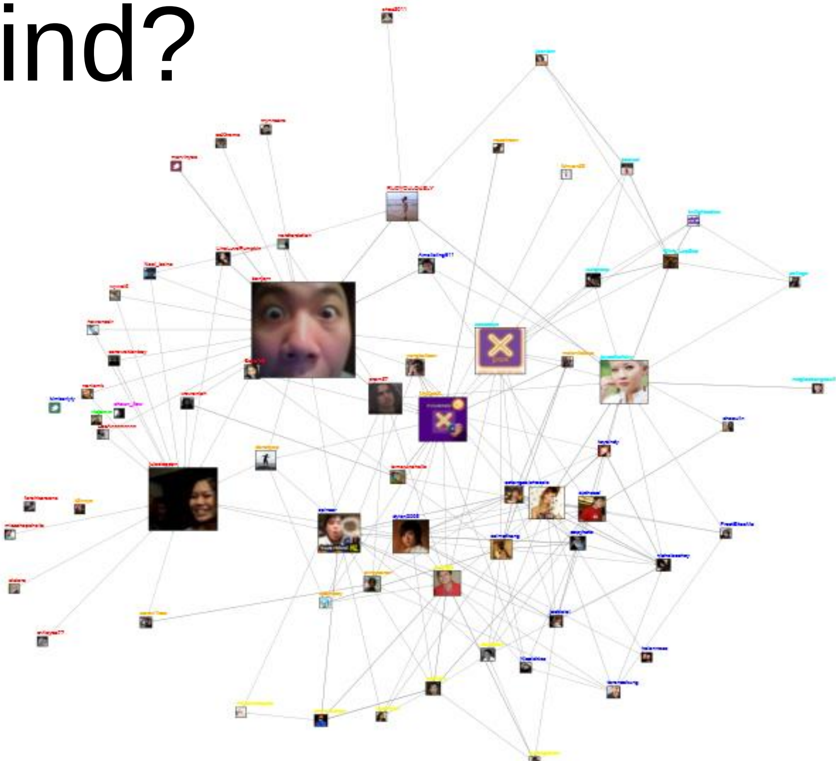


User interactions may form *implicit groups*

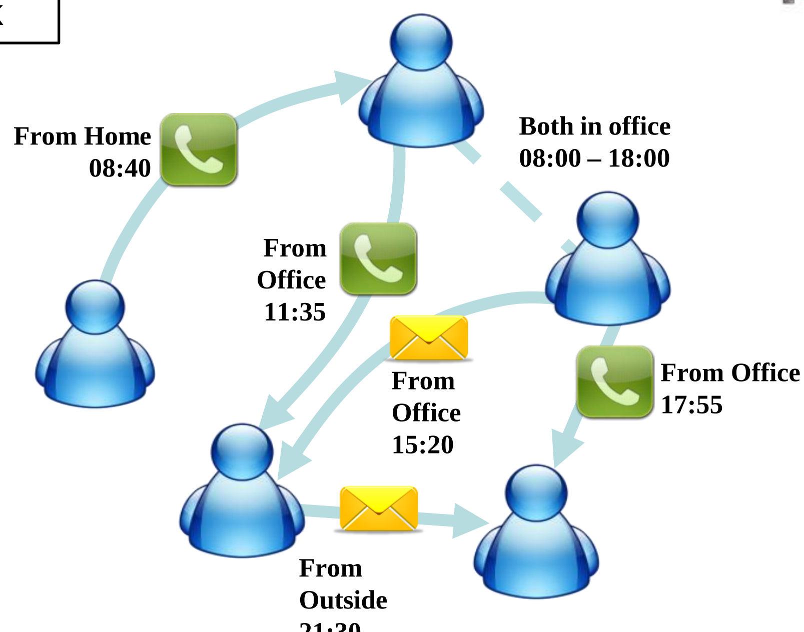
What is behind?



Publication network

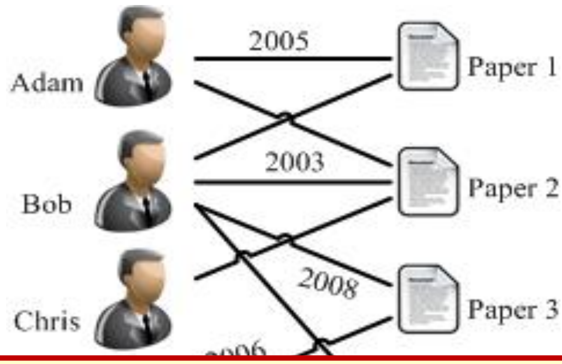


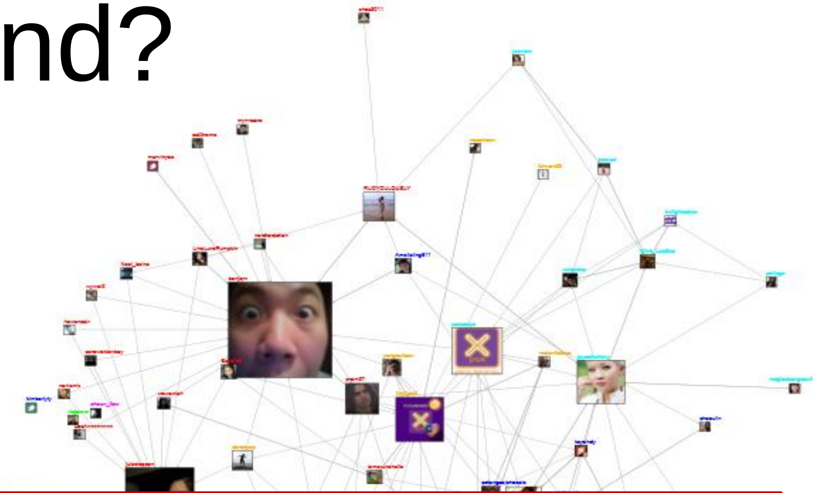
Twitter's following network



Mobile communication network

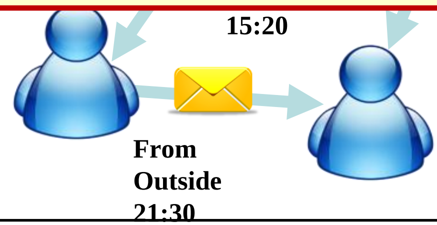
What is behind?





Questions:

- What are the fundamental forces behind?
- A generalized framework for inferring social ties?
- How to connect the different networks?

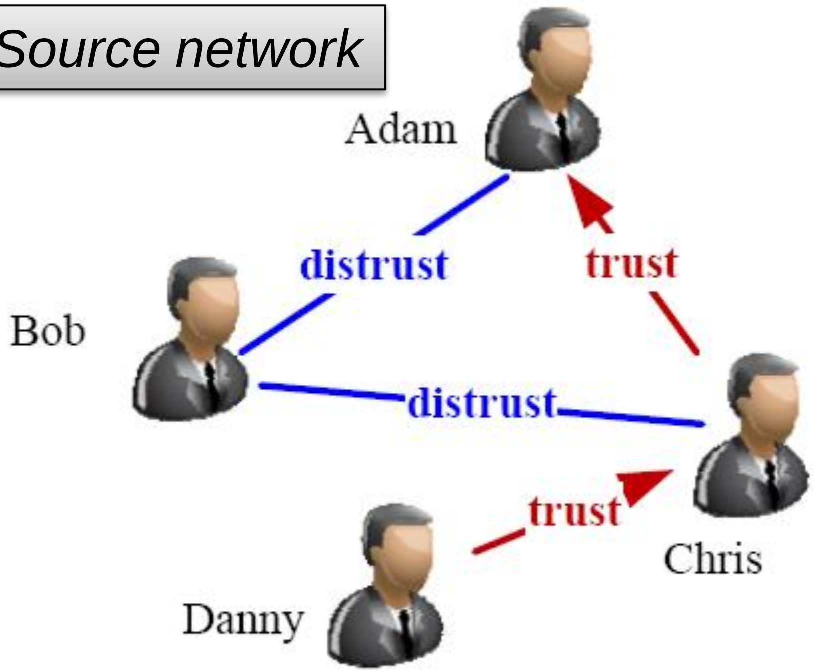


Mobile communication network

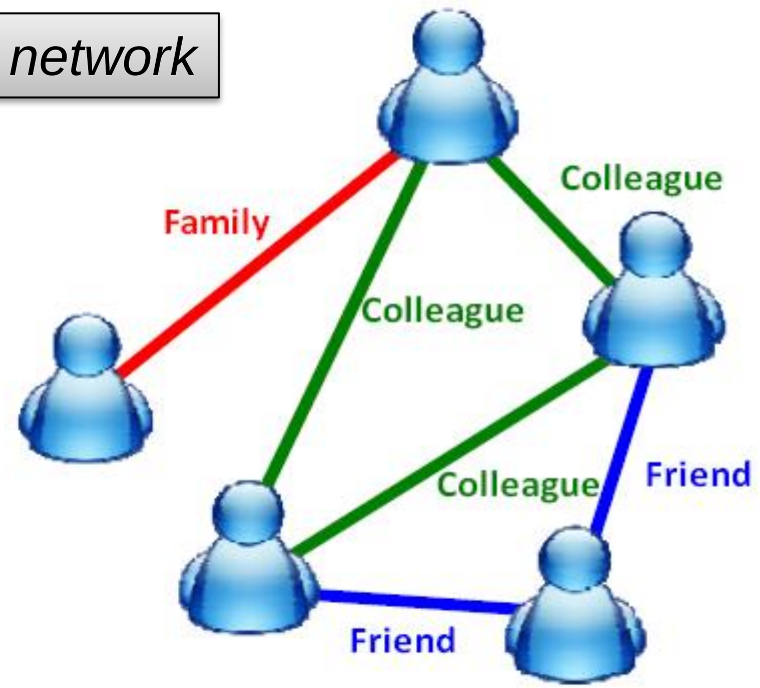
Problem : Transfer Learning

Input: *two networks* G_S *and* Gt
with $|E_S^L| >> |E_T^L|$

Source network



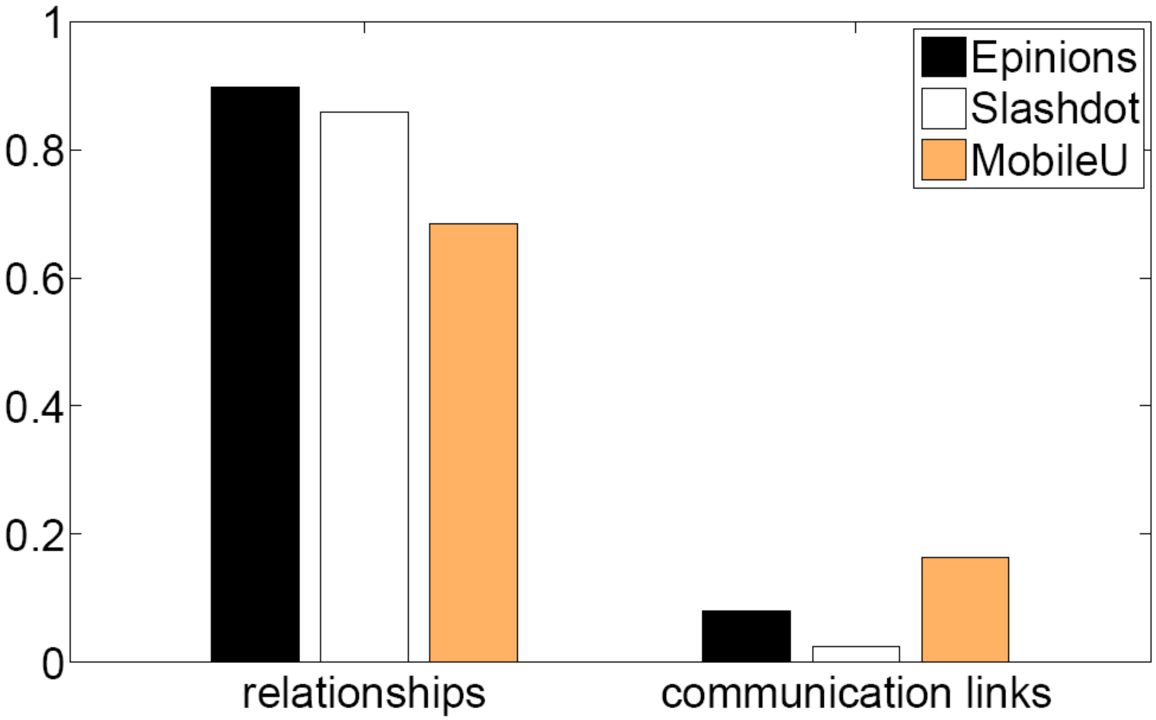
Target network



Input:
G_S, G_T

Output:
f: $(G_T|G_S) \rightarrow R$

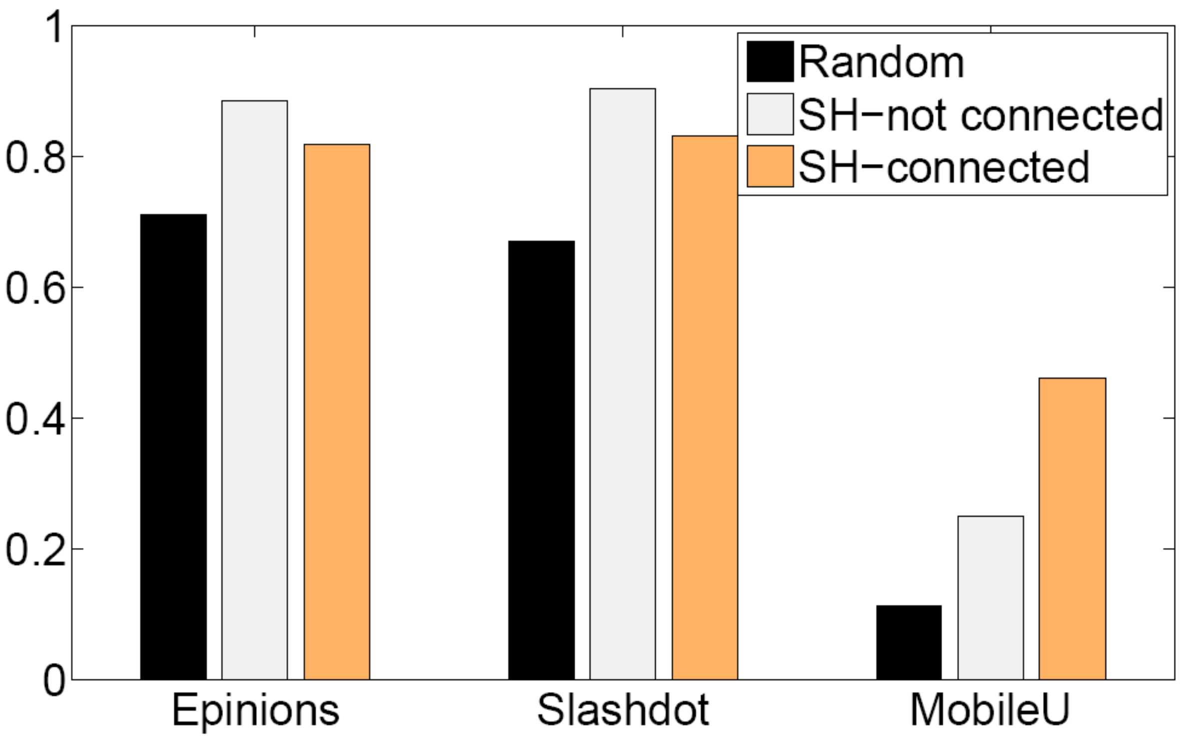
Observation – Social balance



Different networks have very different balance probabilities.

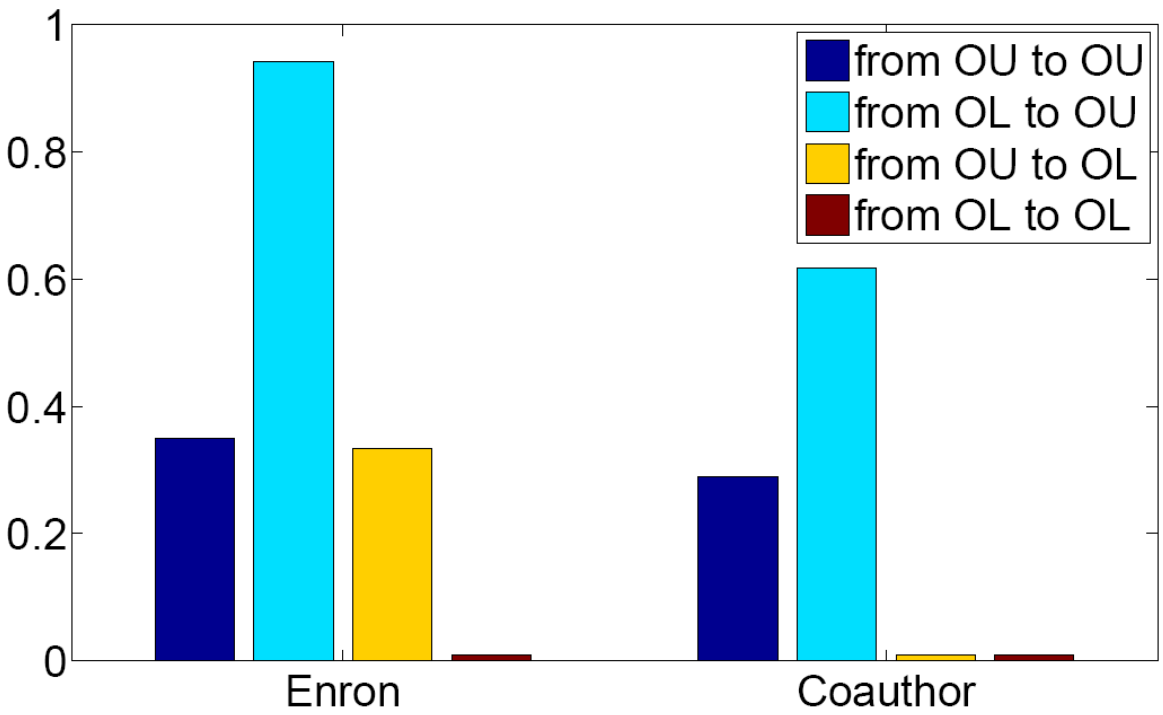
friendships of the three networks have a relatively similar probability.

Observation—Structural hole



Users are more likely (average +70% higher than change) to have the same type of relationship with C if C spans a structural hole

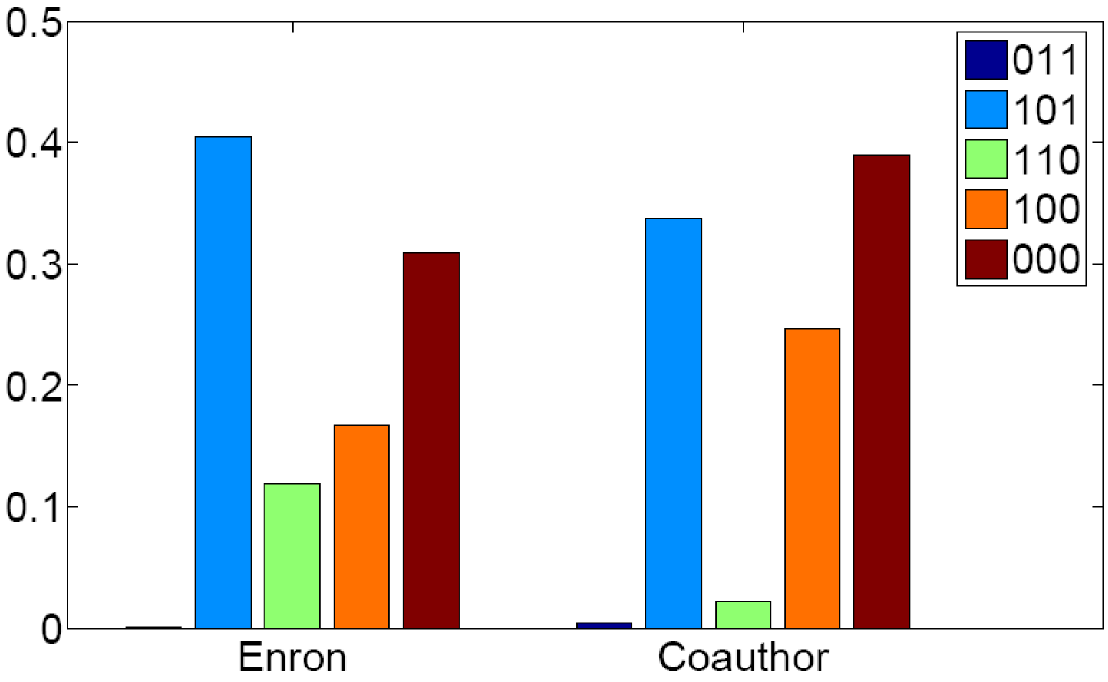
Observation—Two-step-flow



OL : Opinion leader.
OU : Ordinary user.

Opinion leaders are more likely to have a higher social-status than ordinary users.

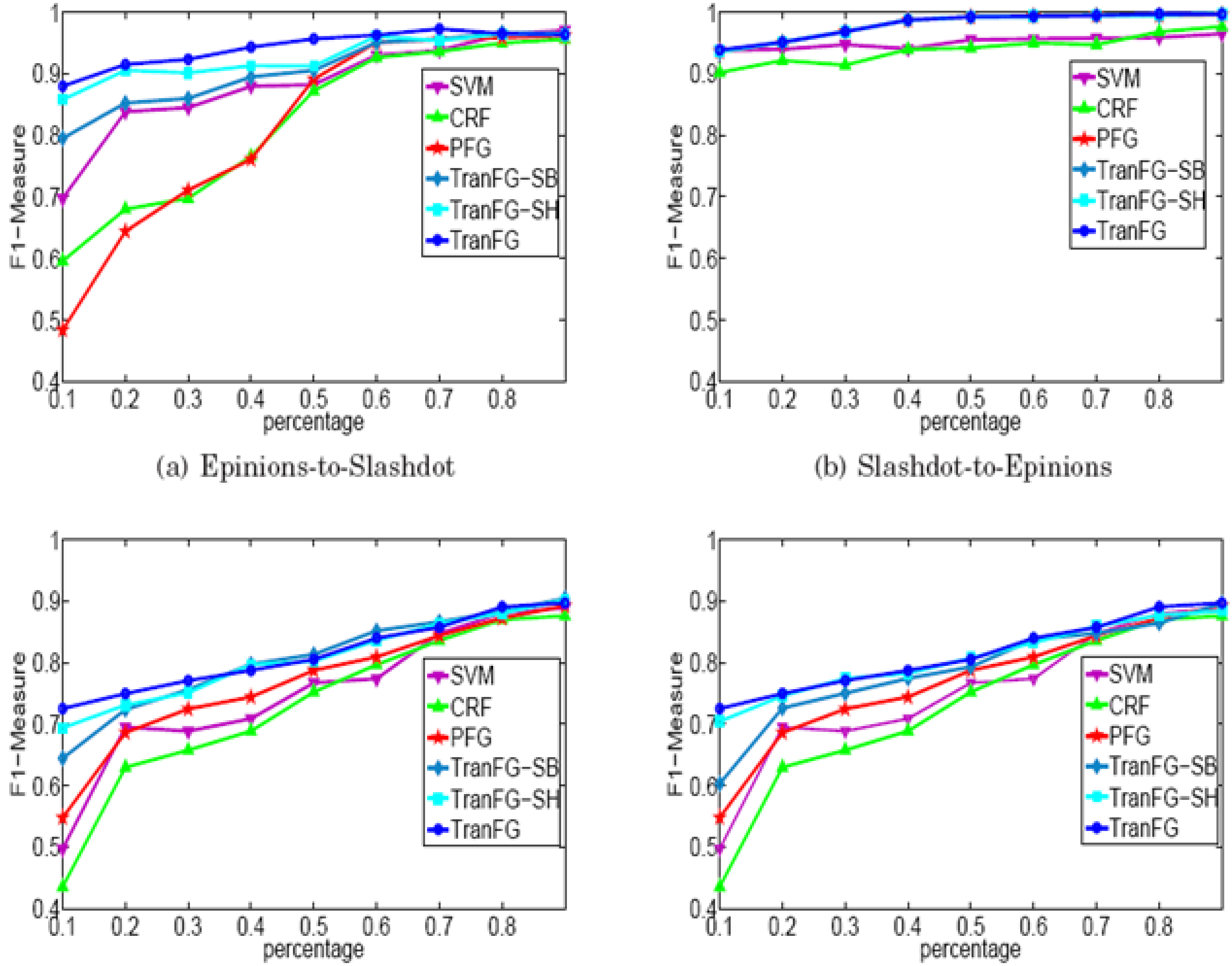
Observation—Social status



99% of triads in the two networks satisfy the social status theory

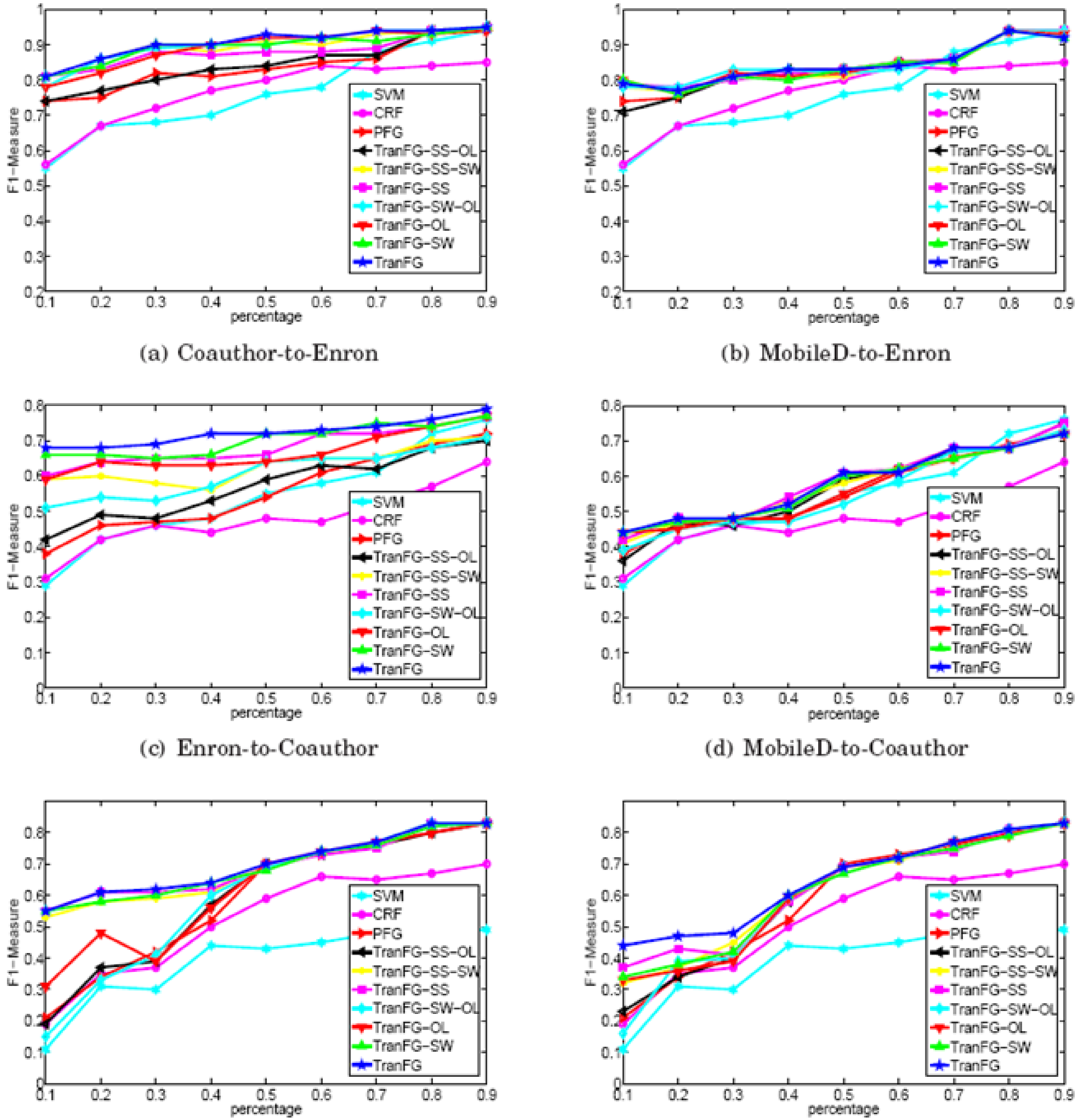
The two networks share a similar distribution on the five frequent forms of triads.

Undirected networks



(a) Epinions-to-Slashdot
(b) Slashdot-to-Epinions
(c) Epinions-to-MobileU
(d) Slashdot-to-MobileU

Directed network



(a) Coauthor-to-Enron

(b) MobileD-to-Enron

(c) Enron-to-Coauthor

(d) MobileD-to-Coauthor

(e) Coauthor-to-MobileD

(f) Enron-to-MobileD