Cross-domain Link Prediction and Recommendation

Jie Tang

Department of Computer Science and Technology
Tsinghua University

Networked World

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

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

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

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

- **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

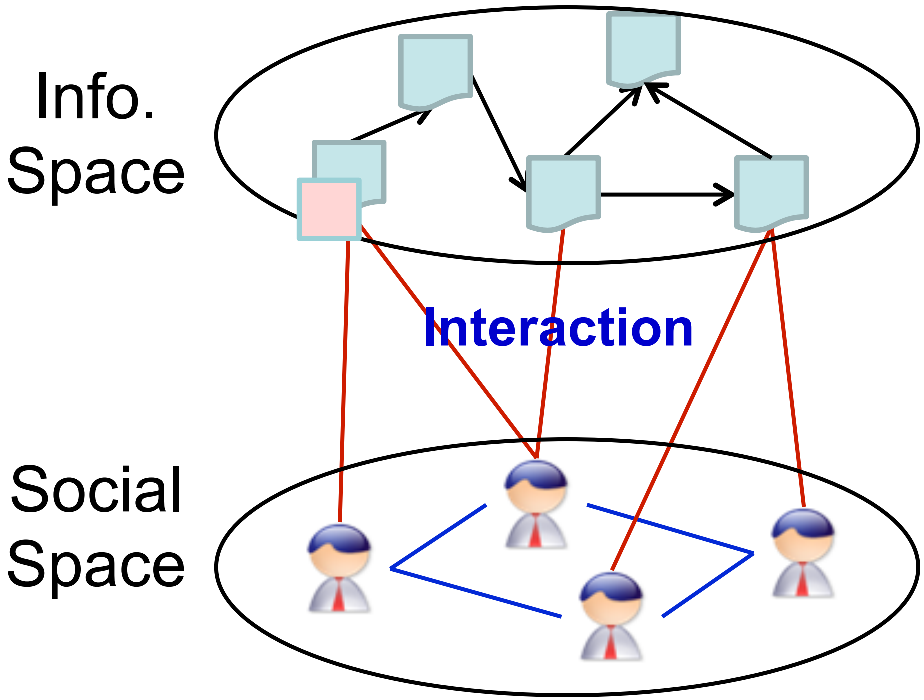
Challenge: Big Social Data

- We generate 2.5×10^{18} byte *big data* per day.

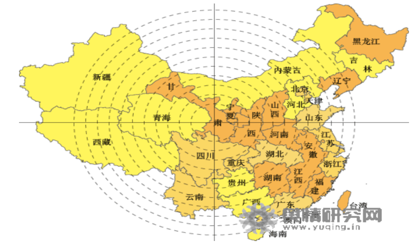
- Big social data:
  - 90% of the data was generated in the past 2 yrs
  - Mining in single data center → mining deep knowledge from multiple data sources

Social Networks

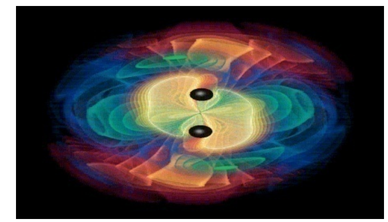
Info. Space vs. Social Space



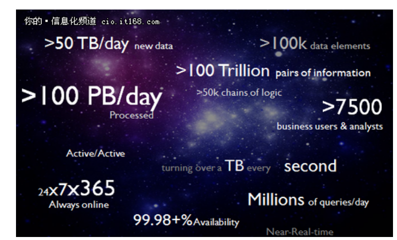
Understanding the mechanisms of interaction dynamics



Opinion Mining

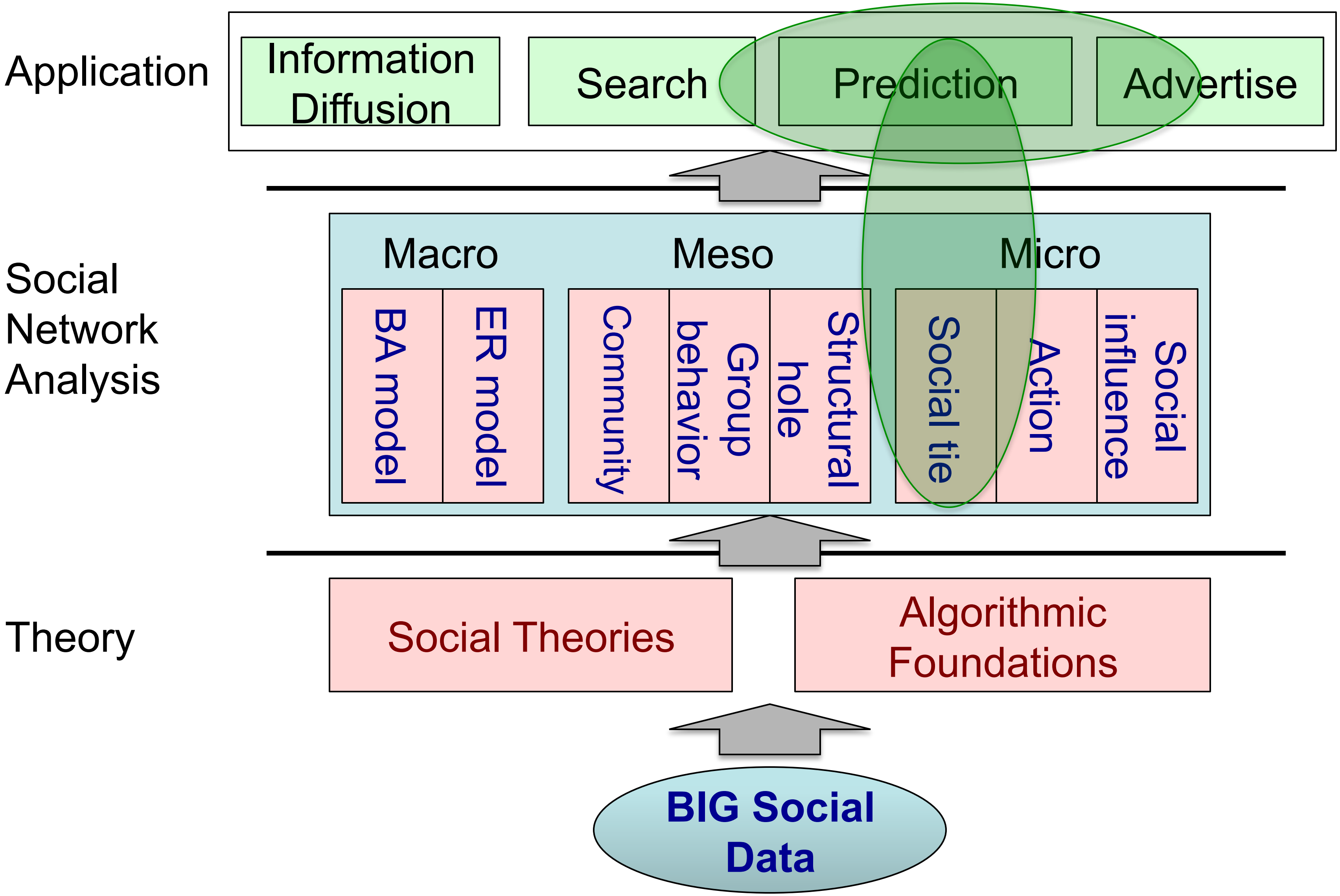


Innovation diffusion



Business intelligence

Core Research in Social Network



Part A:

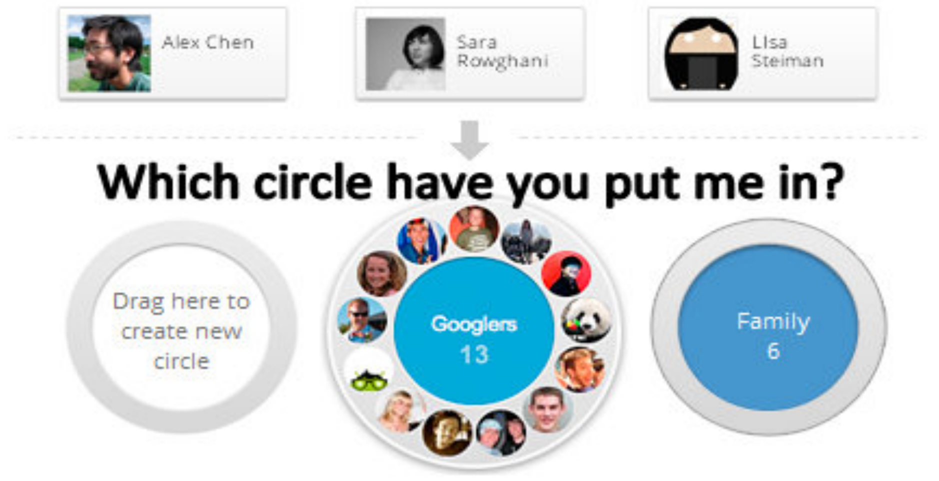
Let us start with a simple case

“inferring social ties in single network”

(KDD 2010, PKDD 2011 Best Runnerup)

Real social networks are complex...

- Nobody exists merely 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/QQ tries to solve this problem via lists/groups
  - however…
- Google circles



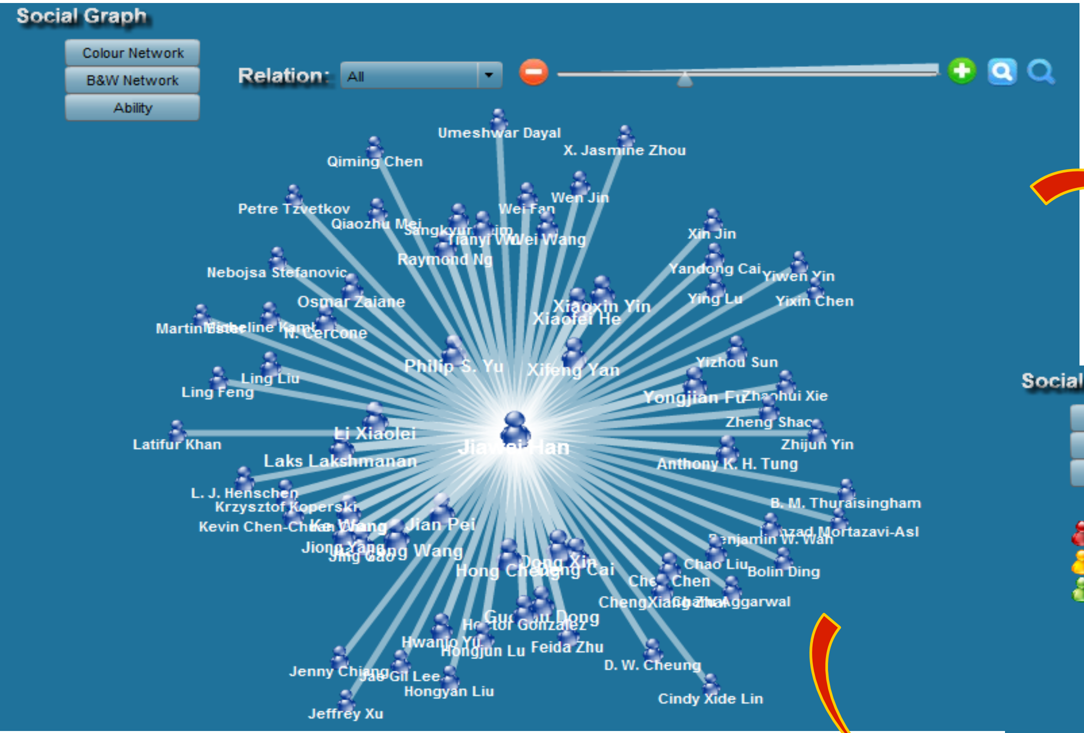
Even complex than we imaged!

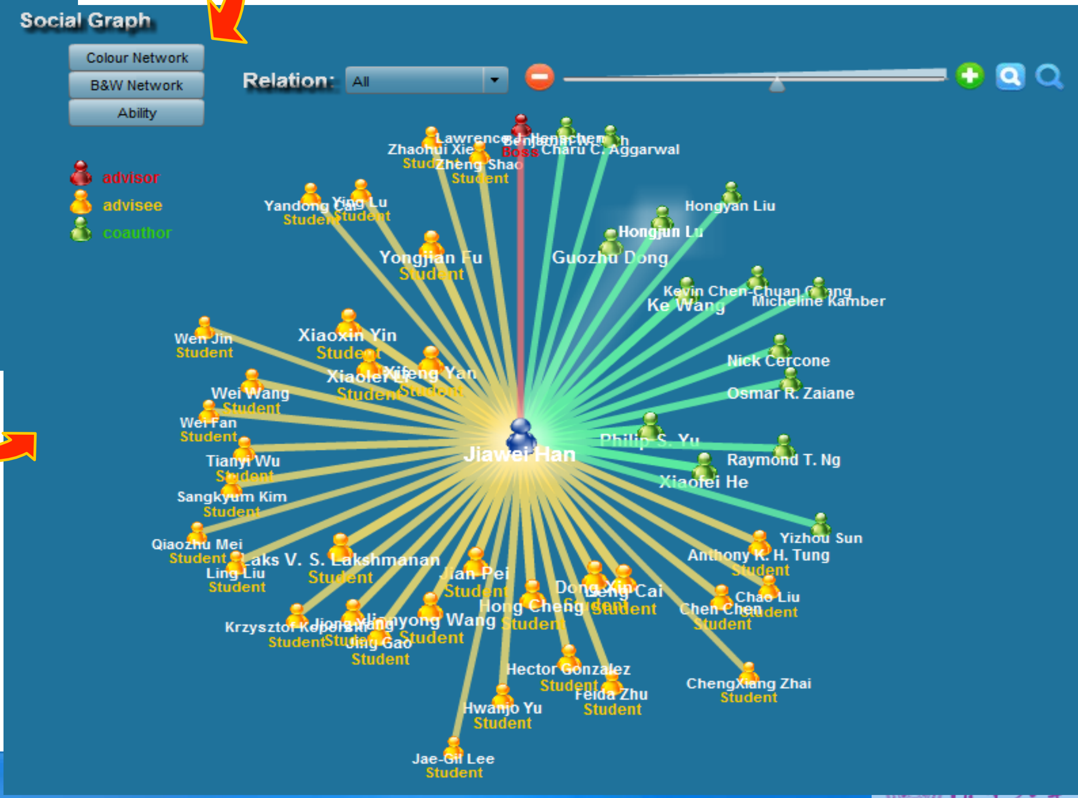
- Only 16% of mobile phone users in Europe have created custom contact groups
  - *users do not* take the time to create it
  - *users do not* know how to circle their friends

- The Problem is that online social network are black **white**…

Example 1. From BW to Color

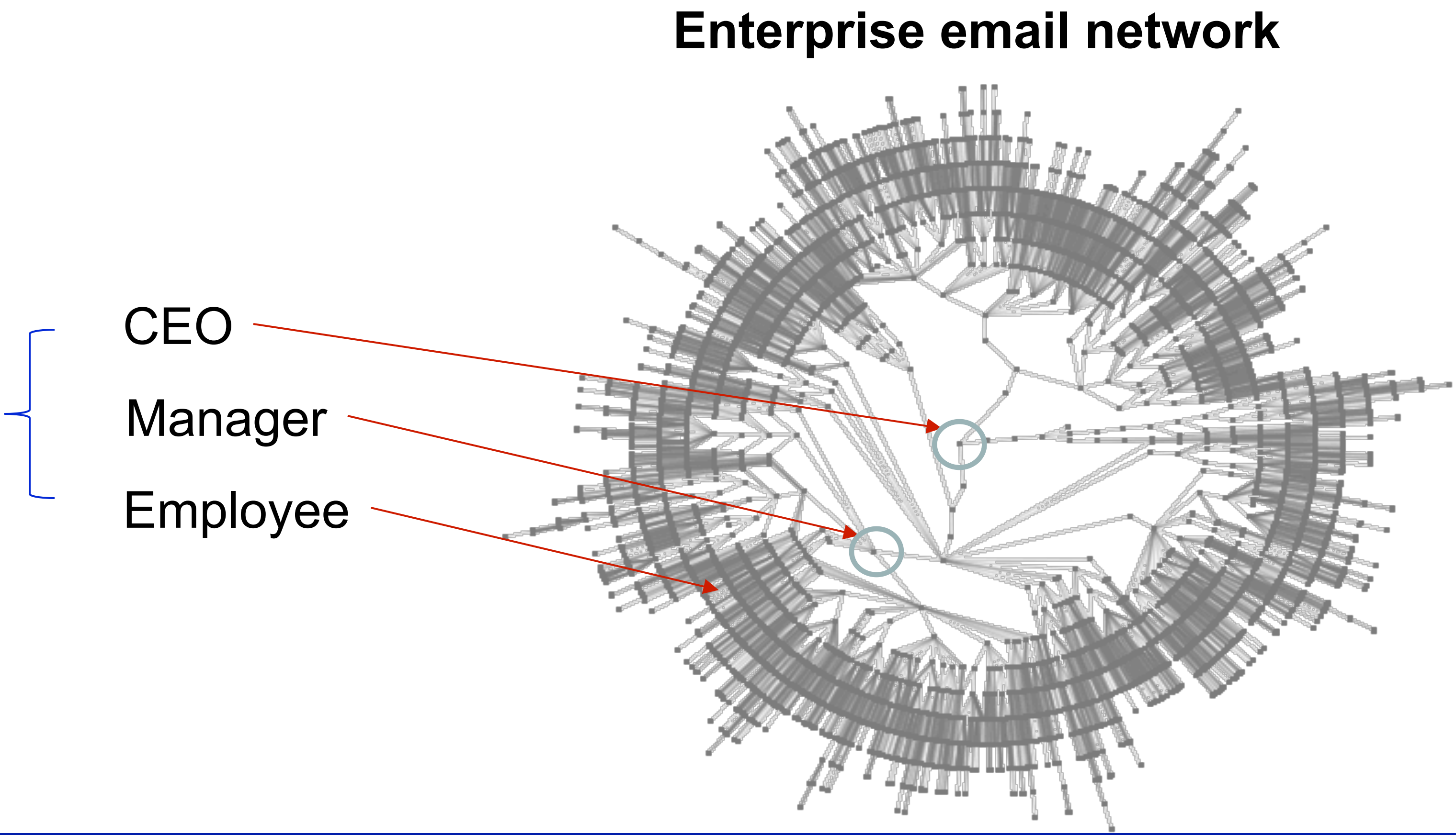
(KDD'10)





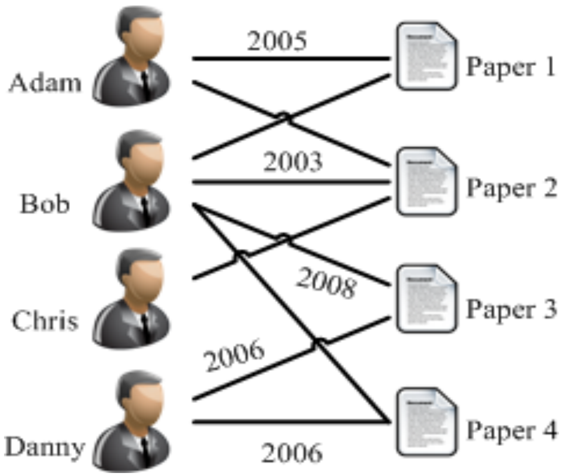
Example 2. From BW to Color
(PKDD'11, Best Paper Runnerup)

How to infer

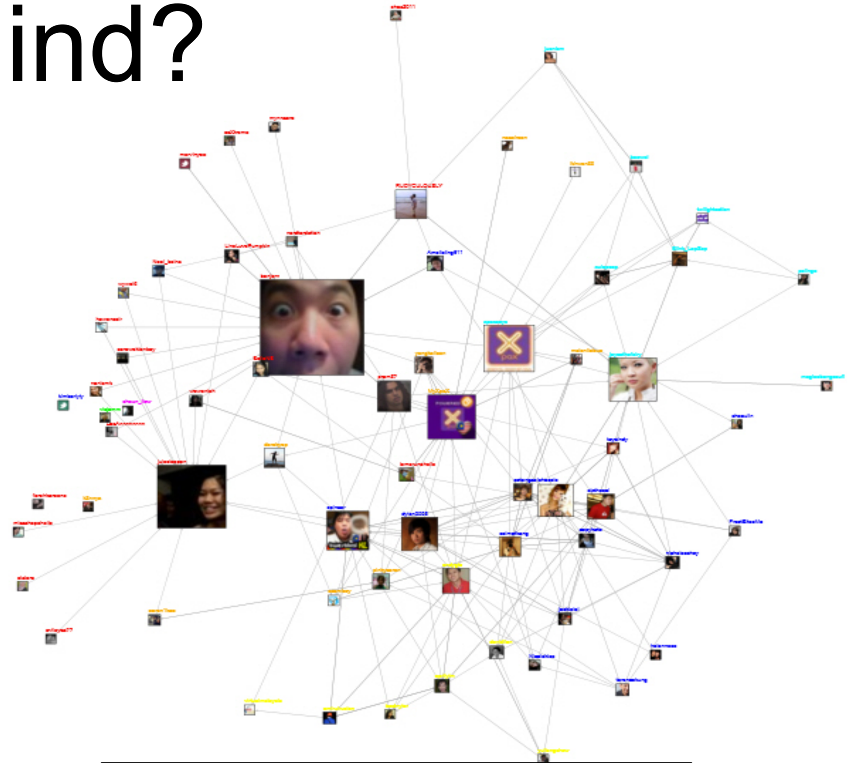


User interactions may form *implicit groups*

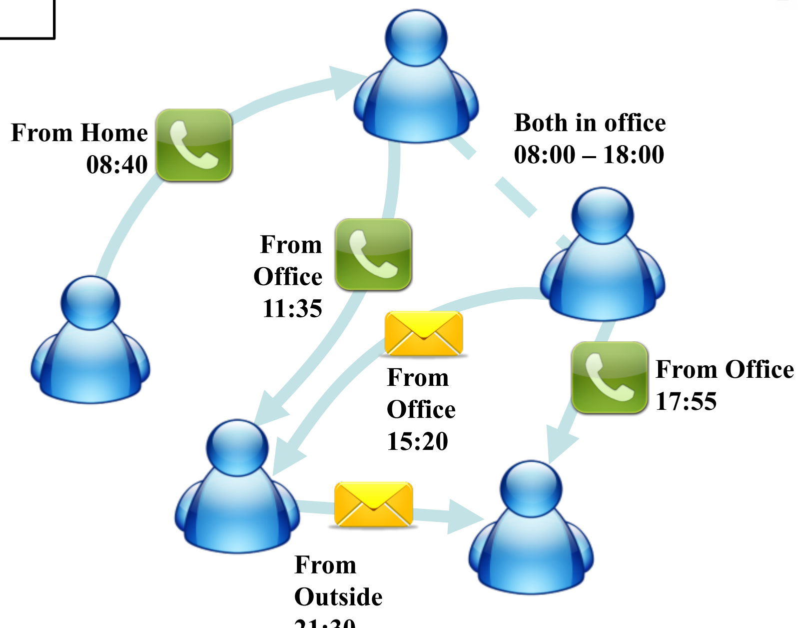
What is behind?



Publication network

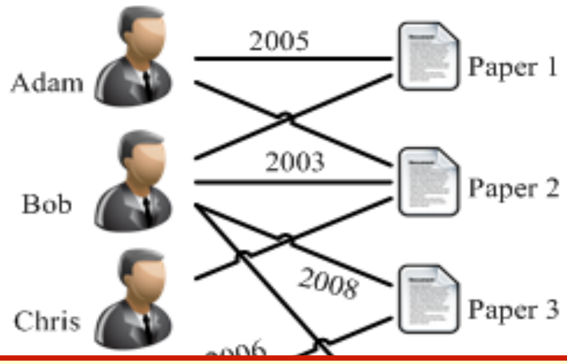


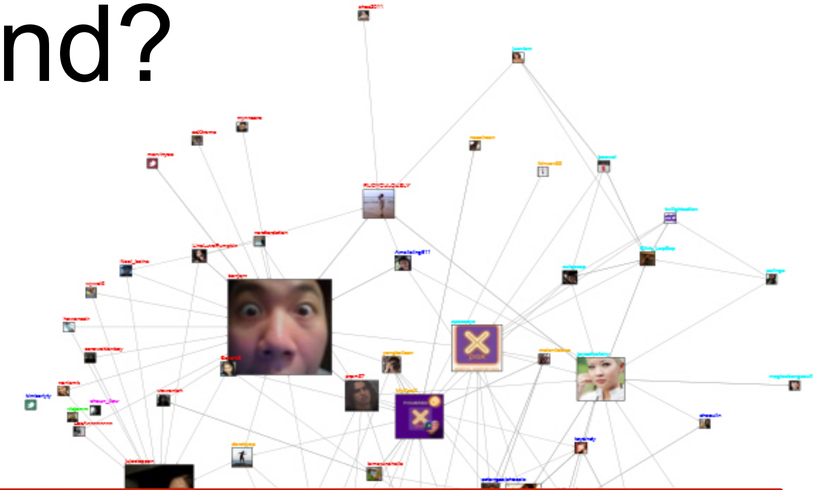
Twitter's following network



Mobile communication network

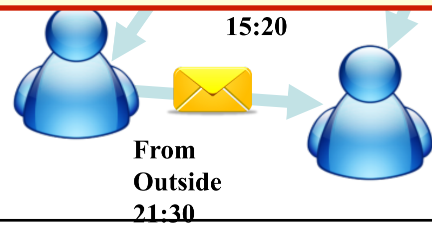
What is behind?





Questions:

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

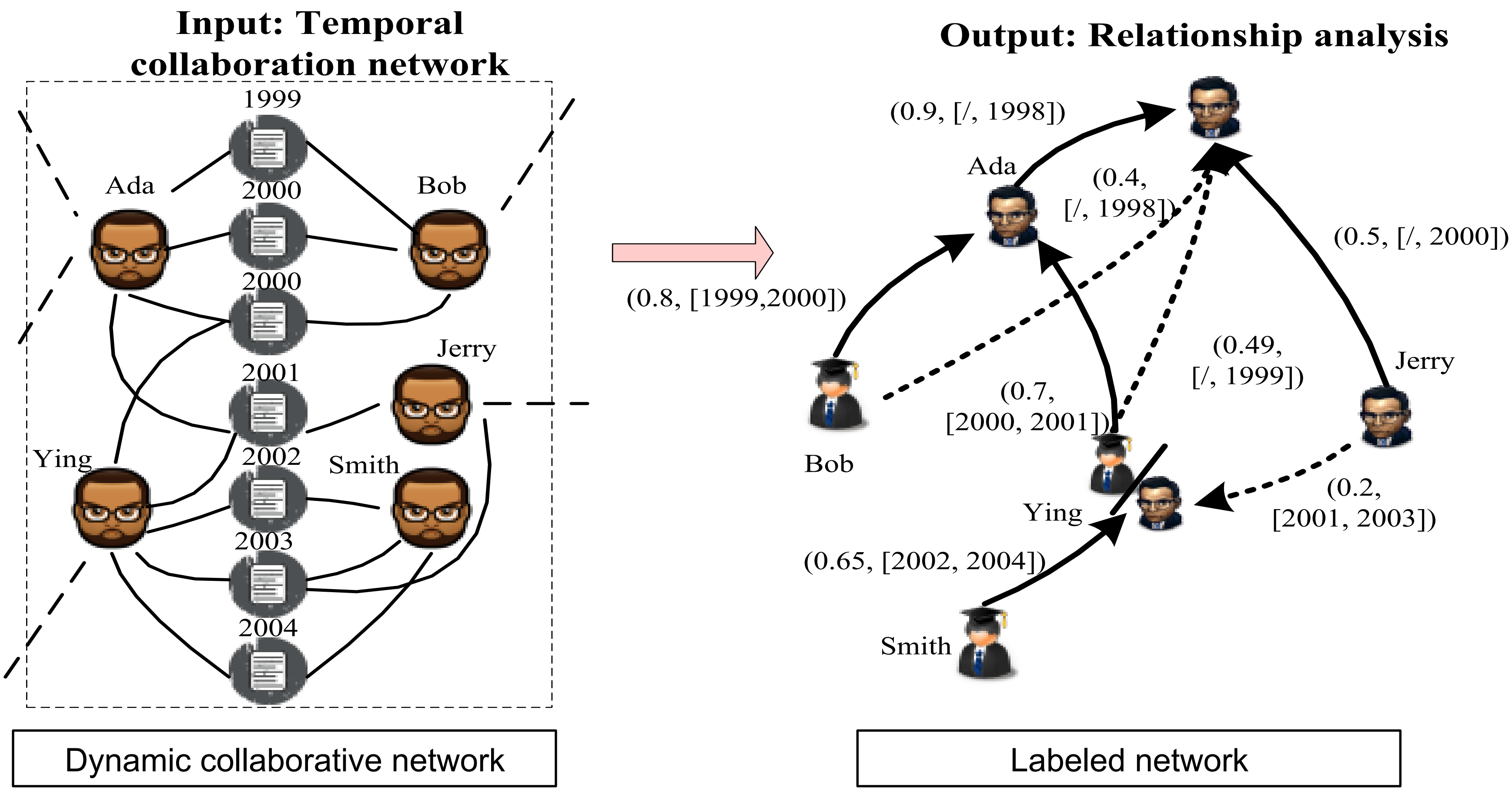


Mobile communication network

inferring social ties in single network

Learning Framework

Problem Analysis

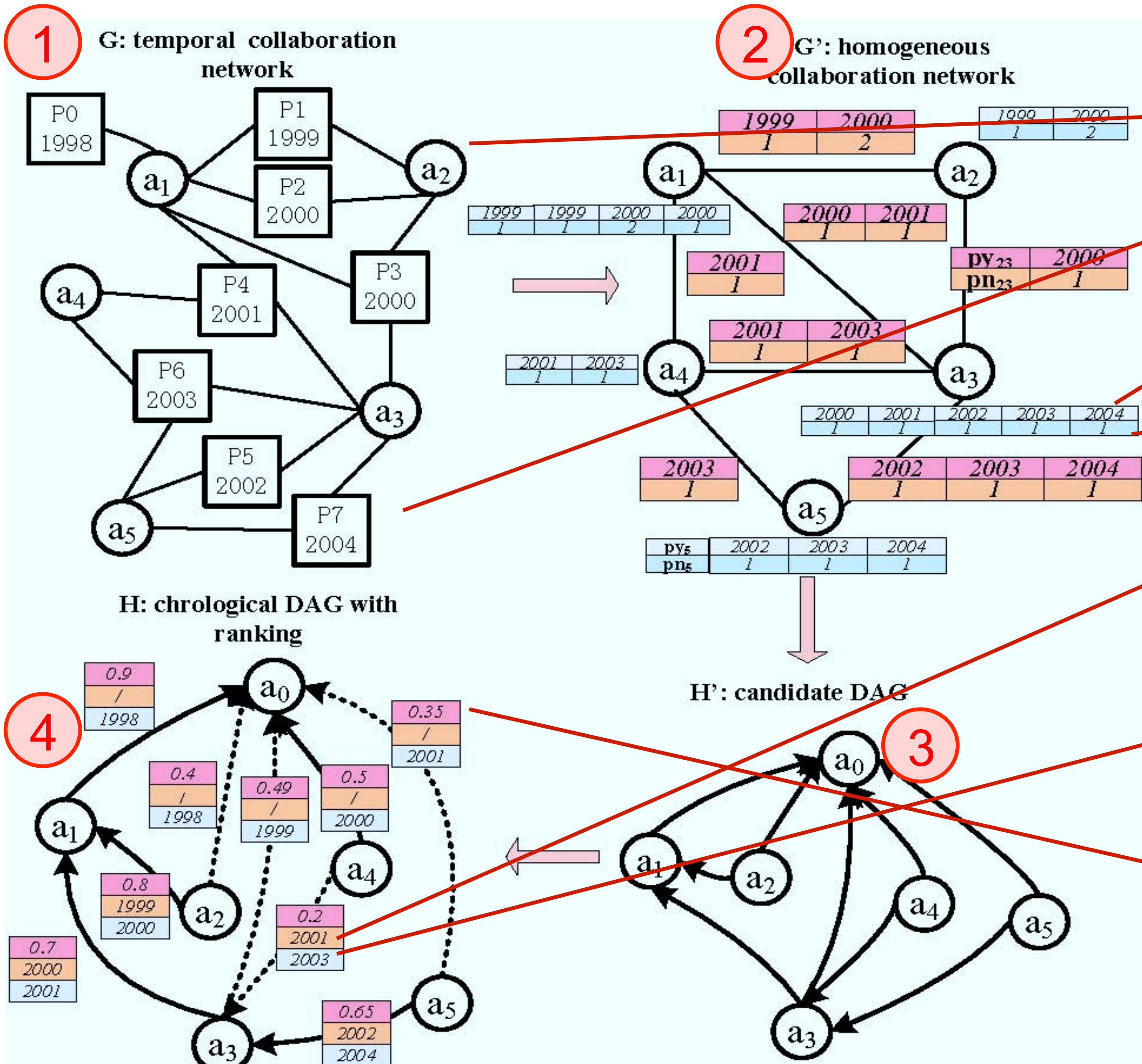


Output: potential types of relationships and their probabilities:

(type, prob, [s_time, e_time])

[1] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. KDD'10, pages 203-212.

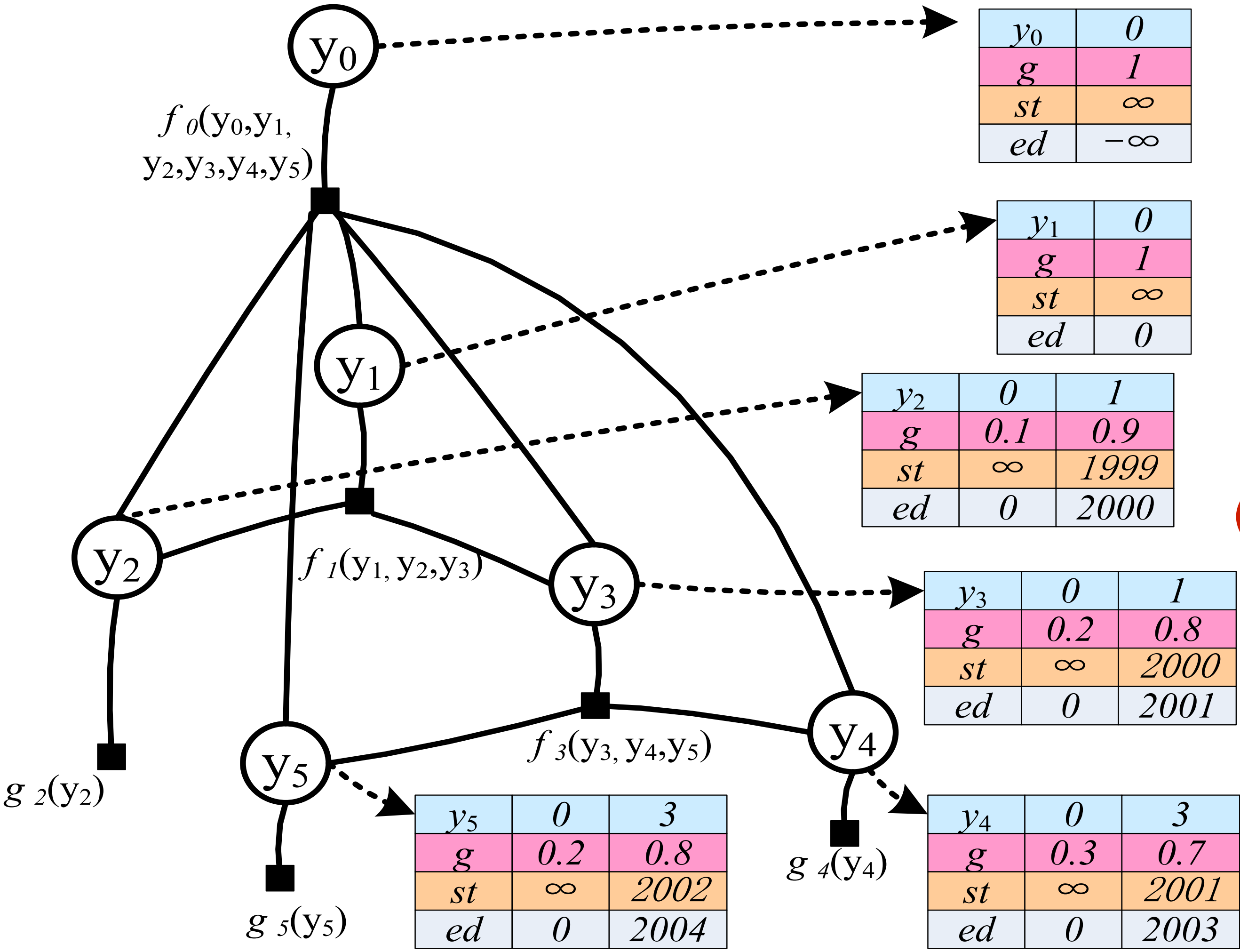
Overall Framework



- a_i: author *i*
- p_j: paper *j*
- py: paper year
- pn: paper#
- $st_{i,yi}$: starting time
- $ed_{i,yi}$: ending time
- $r_{i,yi}$: probability

The problem is cast as, for each node, identifying which neighbor has the **highest probability** to be his/her advisor, i.e., $P(y_i=j \mid x_i, x_{\sim i}, \mathbf{y})$, where x_j and x_i are neighbors.

Time-constrained Probabilistic Factor Graph (TPFG)



y_0	0
g	1
st	∞
ed	$-\infty$

y_1	0
g	1
st	∞
ed	0

y_2	0	1
g	0.1	0.9
st	∞	1999
ed	0	2000

y_3	0	1
g	0.2	0.8
st	∞	2000
ed	0	2001

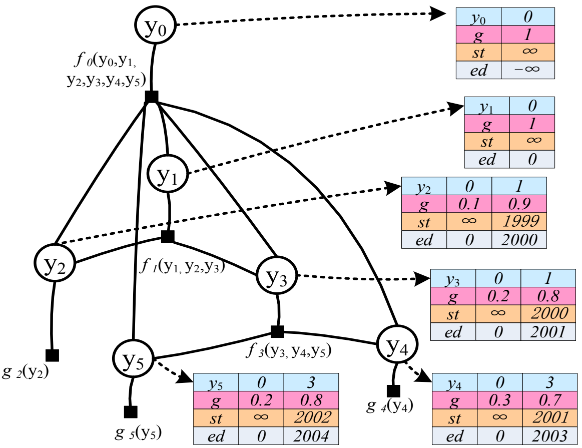
y_5	0	3
g	0.2	0.8
st	∞	2002
ed	0	2004

y_4	0	3
g	0.3	0.7
st	∞	2001
ed	0	2003

- Hidden variable y_x: a_x's advisor
- $st_{x,yx}$: starting time $ed_{x,yx}$: ending time
- $g(y_x, st_x, ed_x)$ is pairwise local feature
- $f_x(y_x, Z_x) = \max g(y_x, st_x, ed_x)$ *under time constraint*
- Y_x: set of potential advisors of a_x

Maximum likelihood estimation

- A general likelihood objective func can be defined as



$$P(y_1,\cdots,y_N)=\frac{1}{Z}\prod_{i=1}^{N} f_i(y_i\,|\,\{y_x\,|\,x\in Y_i^{-1}\})$$

with

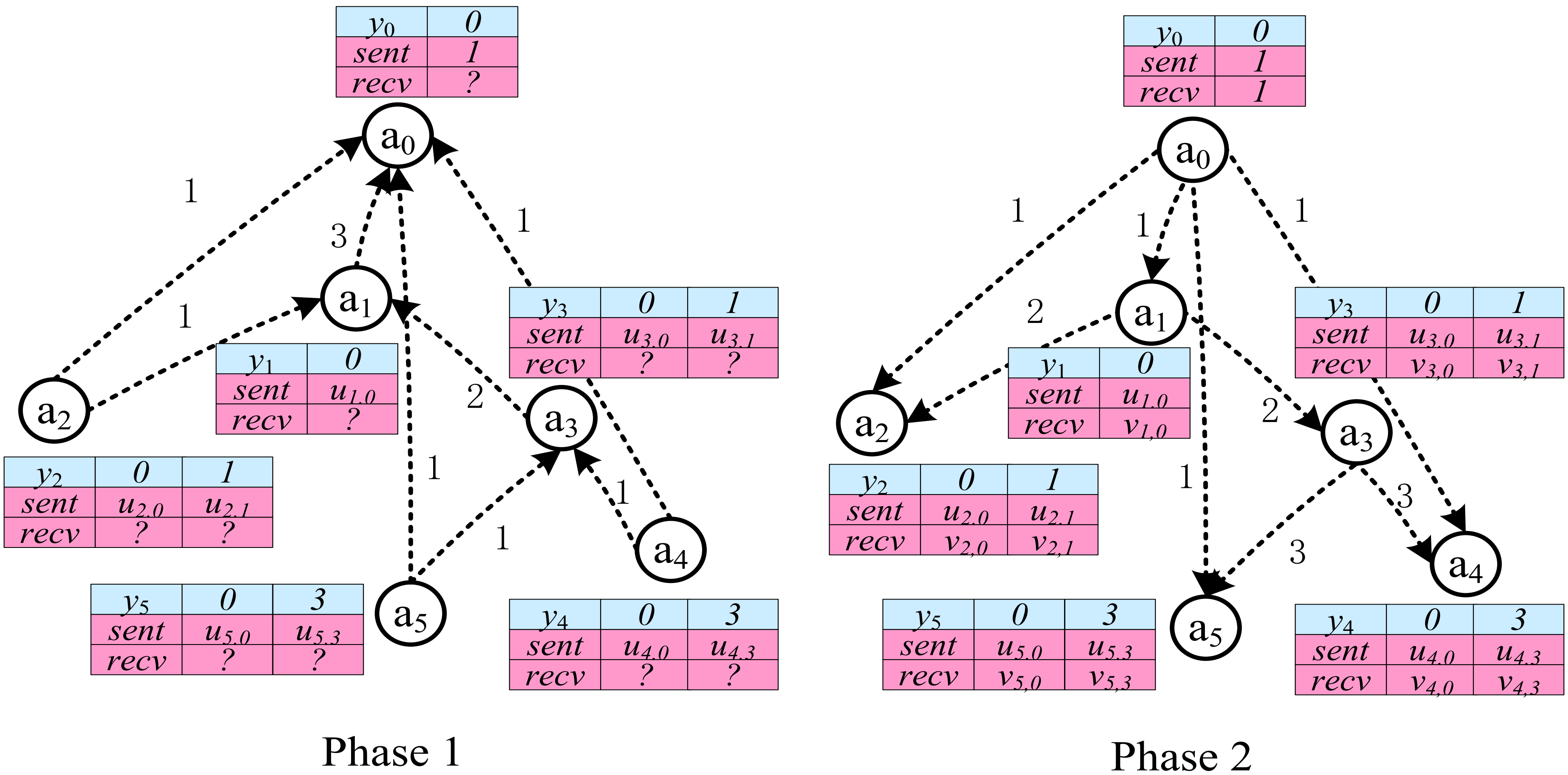
$$f_i(y_i\,|\,\{y_x\,|\,x\in Y_i^{-1}\})=g(y_i,st_{ij},ed_{ij})\prod_{x\in Y_i^{-1}}\phi(y_x,ed_{ij},st_{xi})$$

where $\Phi(.)$ can be instantiated in different ways, e.g.,

$$\phi(y_x,ed_{ij},st_{xi})=\begin{cases}1, & y_x\neq i\vee ed_{ij}<st_{xi}\\ 0, & y_x=i\wedge ed_{ij}>=st_{xi}\end{cases}$$

Inference algorithm of TPFG

- $r_{ij} = \max P(y_1, \ldots, y_{na} | y_i = j) = \exp(\mathrm{sent}_{ij} + \mathrm{recv}_{ij})$



Results of Model 1

- DBLP data: 654, 628 authors, 1,076,946 publications, years provided.
- Ground truth: MathGenealogy Project; AI Genealogy Project; Faculty Homepage

Datasets	RULE	SVM	IndMAX		Model 1	
TEST1	69.9%	73.4%	75.2%	78.9%	80.2%	**84.4%**
TEST2	69.8%	74.6%	74.6%	79.0%	81.5%	**84.3%**
TEST3	80.6%	86.7%	83.1%	90.9%	88.8%	**91.3%**

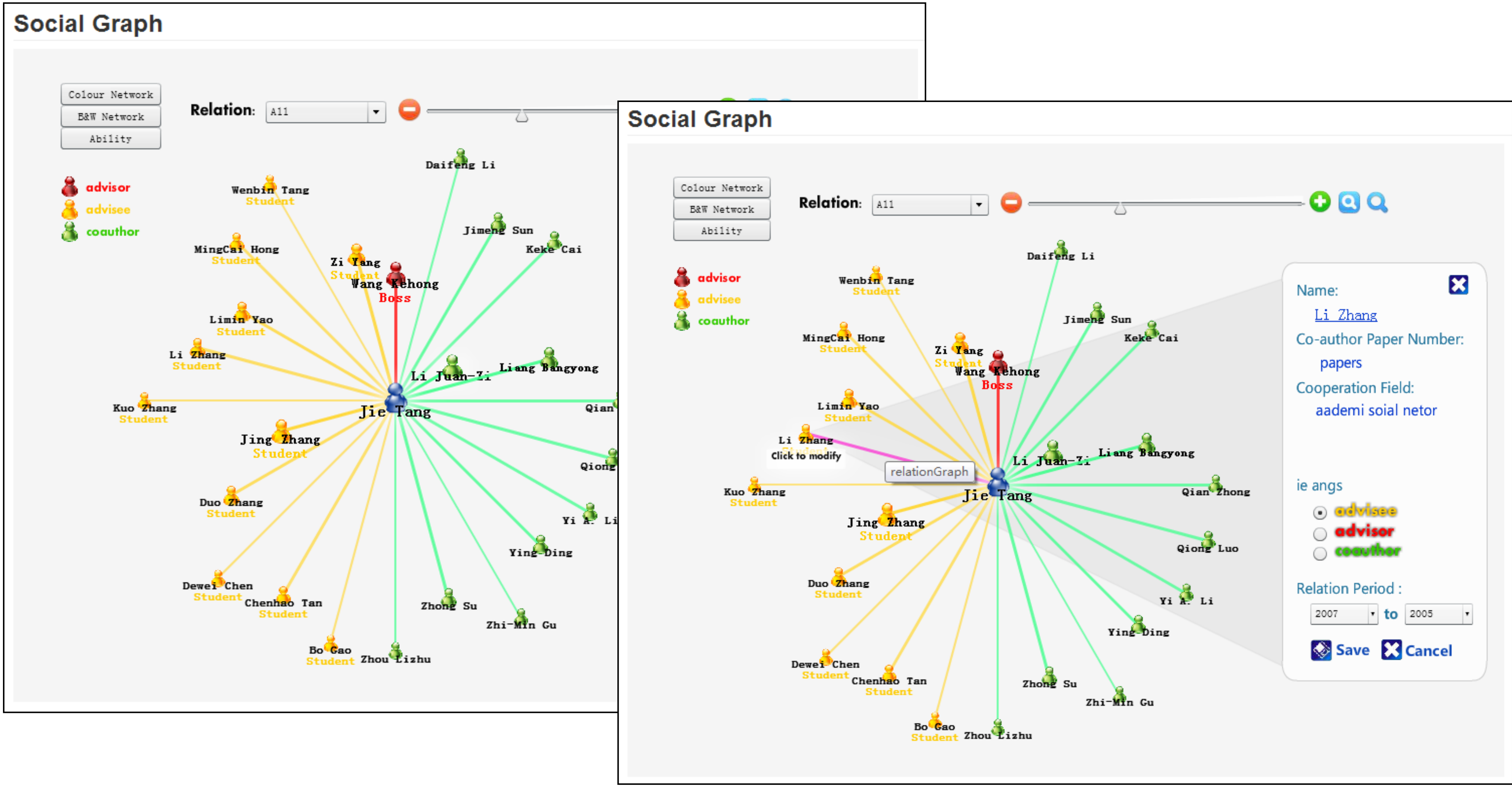
heuristics

Supervised learning

Empirical parameter

optimized parameter

Results



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

Part B:

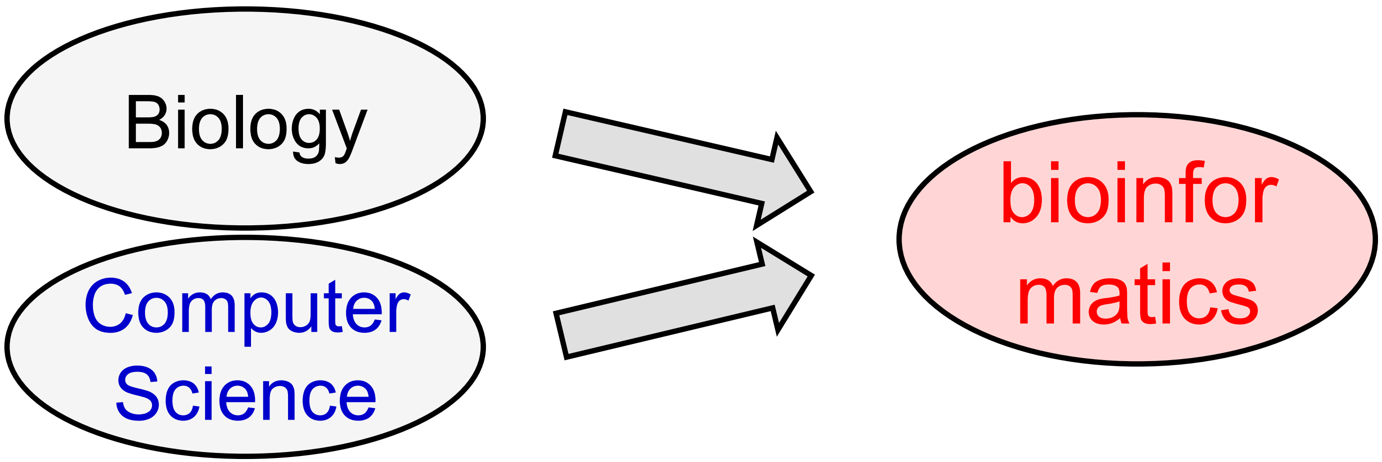
Extend the problem to cross-domain

"cross-domain collaboration recommendation"

(KDD 2012, WSDM

Cross-domain Collaboration

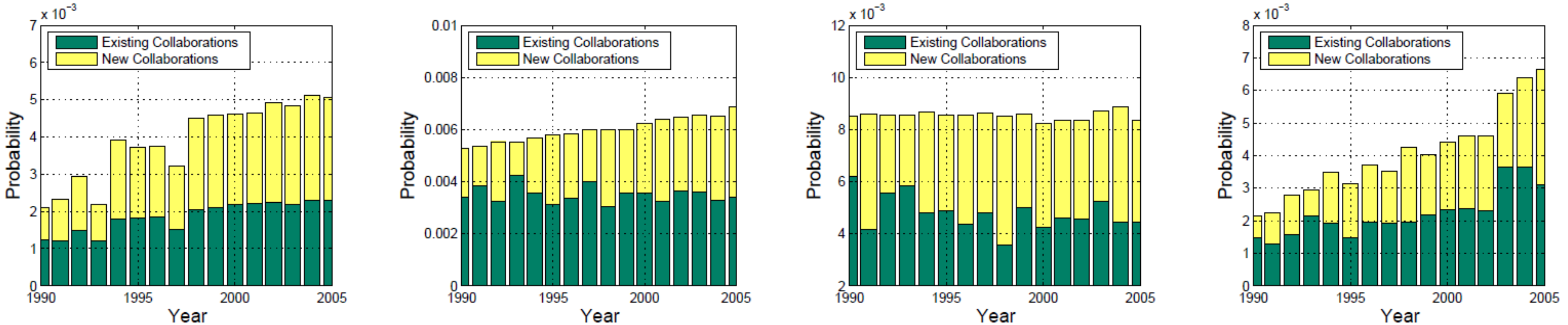
- Interdisciplinary collaborations have generated huge impact, for example,
  - 51 (>1/3) of the KDD 2012 papers are result of cross-domain collaborations between graph theory, visualization, economics, medical inf., DB, NLP, IR
  - Research field evolution



[1] J. Tang, S. Wu, J. Sun, and H. Su. Cross-domain Collaboration Recommendation. **KDD'12**, pages 1285-1293. (Full Presentation & Best Poster Award)

Cross-domain Collaboration (cont.)

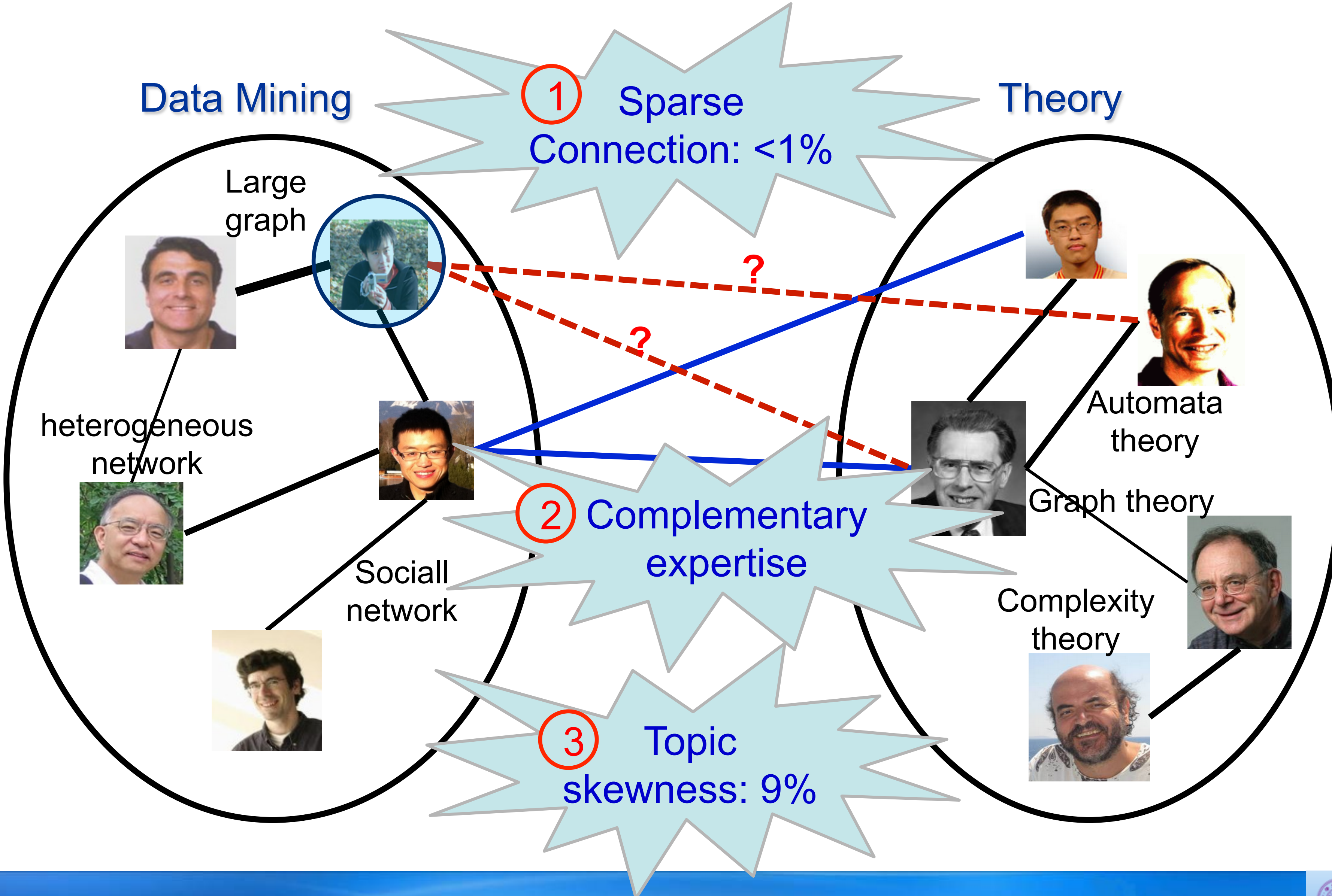
- Increasing trend of cross-domain collaborations



(a) DM - TH (b) DM - MI (c) DM - VIS (d) MI - DB

Data Mining(DM), Medical Informatics(MI), Theory(TH), Visualization(VIS)

Challenges



Data Mining
Theory
1 Sparse Connection: <1%
Large graph
heterogeneous network
Sociall network
?
?
Automata theory
Graph theory
Complexity theory
2 Complementary expertise
3 Topic skewness: 9%

Related Work-Collaboration recommendation

- Collaborative topic modeling for recommending papers
  - C. Wang and D.M. Blei. [2011]
- On social networks and collaborative recommendation
  - I. Konstas, V. Stathopoulos, and J. M. Jose. [2009]
- CollabSeer: a search engine for collaboration discovery
  - H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. [2007]
- Referral web: Combining social networks and collaborative filtering
  - H. Kautz, B. Selman, and M. Shah. [1997]
- Fab: content-based, collaborative recommendation
  - M. Balabanovi and Y. Shoham. [1997]

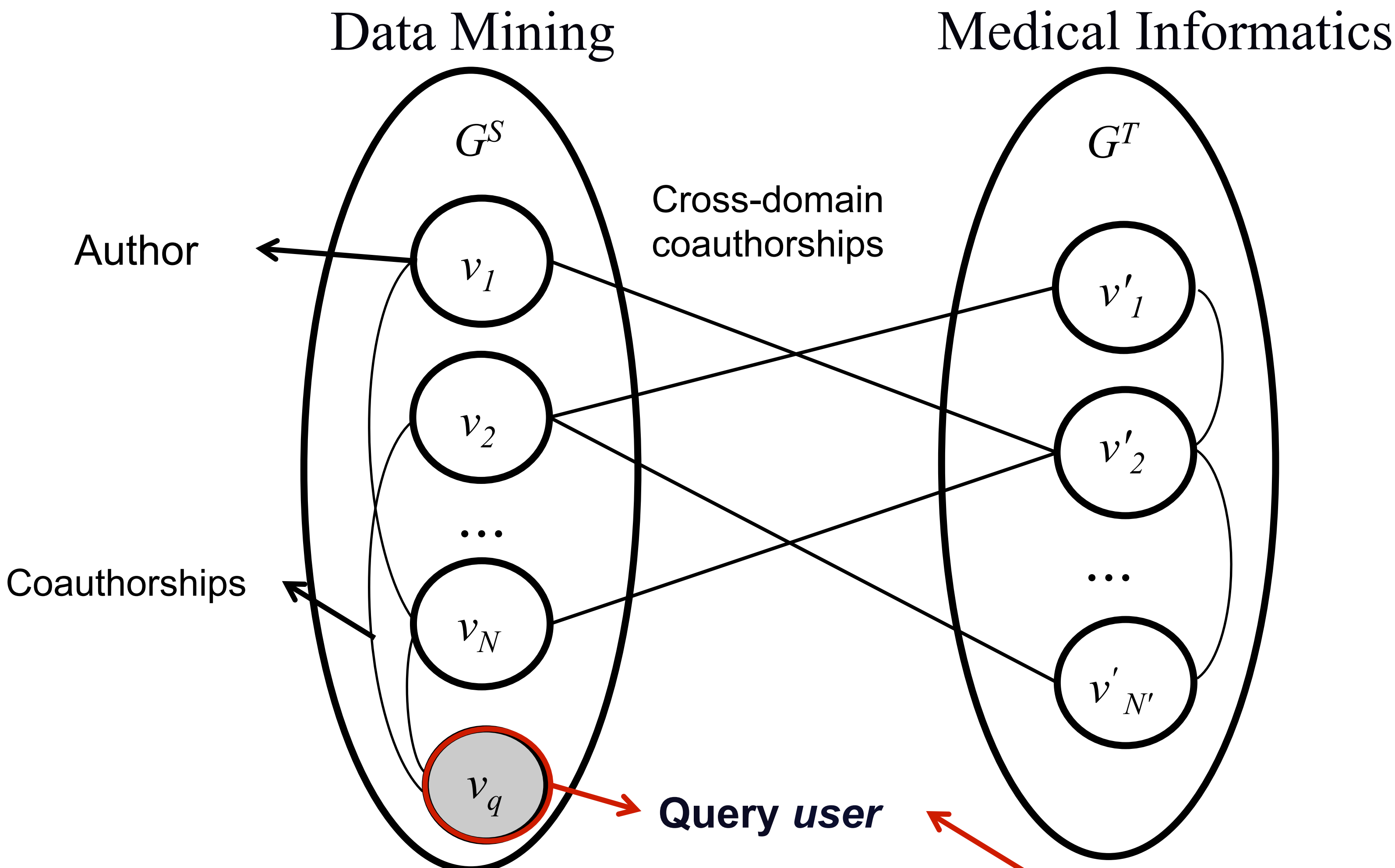
Related Work-Expert finding and matching

- Topic level expertise search over heterogeneous networks
  - J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. [2011]
- Formal models for expert finding in enterprise corpora
  - K. Balog, L. Azzopardi, and M.de Rijke. [2006]
- Expertise modeling for matching papers with reviewers
  - D. Mimno and A. McCallum. [2007]
- On optimization of expertise matching with various constraints
  - W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li. [2012]

cross-domain collaboration recommendation

Approach Framework
—Cross-domain Topic Learning

Author Matching



$$\mathbf{r}^{(t+1)} = (1-\tau)\mathbf{S}\cdot\mathbf{r}^{(t)} + \tau\boxed{\mathbf{q}}$$

Recall Random Walk

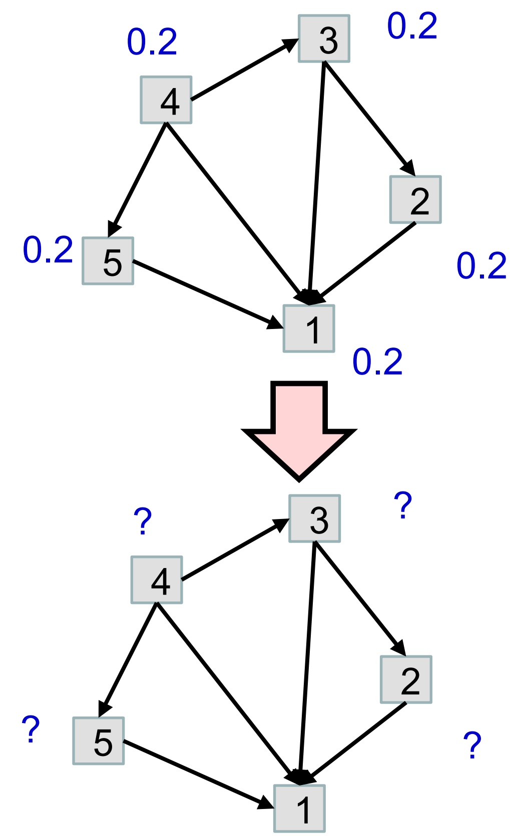
- Let us begin with PageRank[1]

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

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

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

$$\alpha = 0.15$$



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

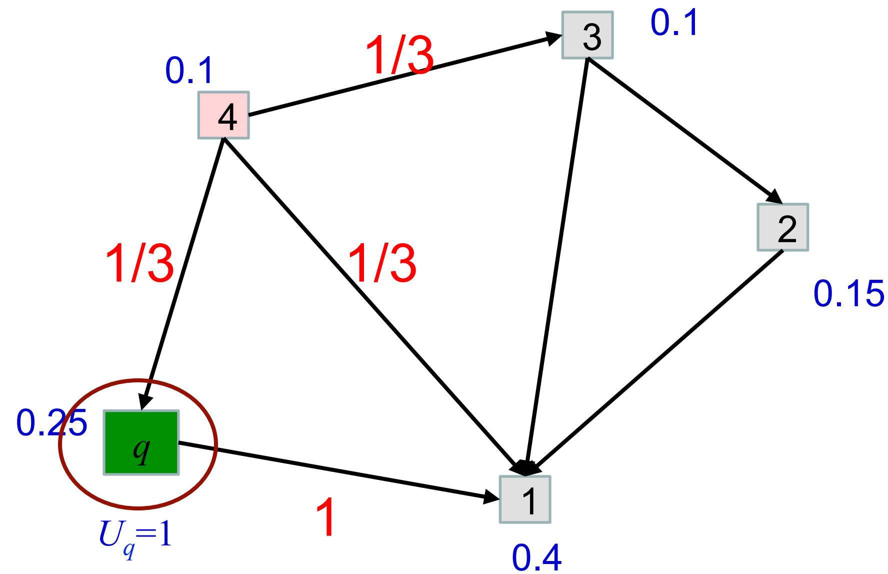
[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

Random Walk with Restart[1]

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

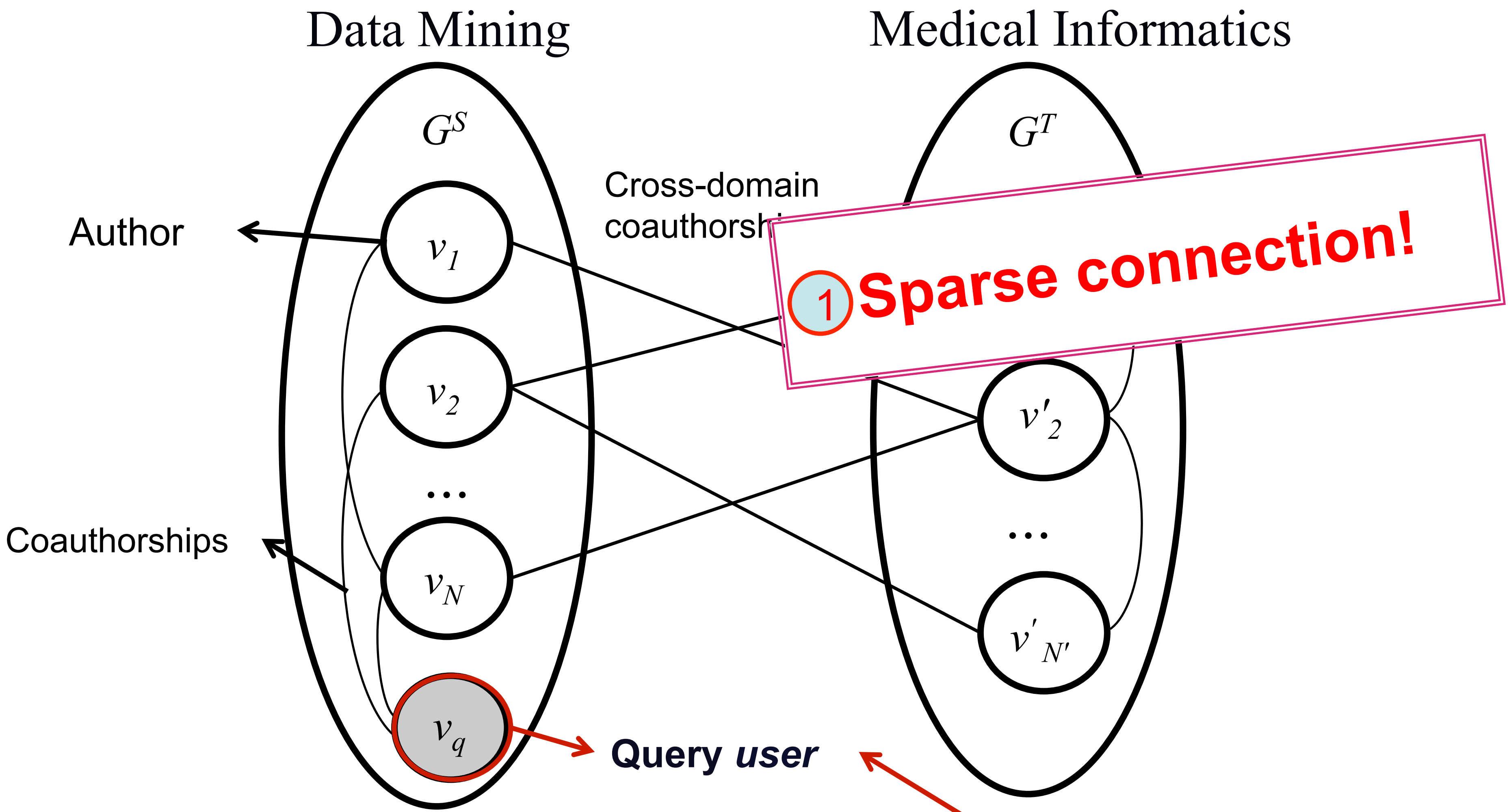
$$M_{ij} = \frac{1}{\text{outdeg}(v_i)}$$

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



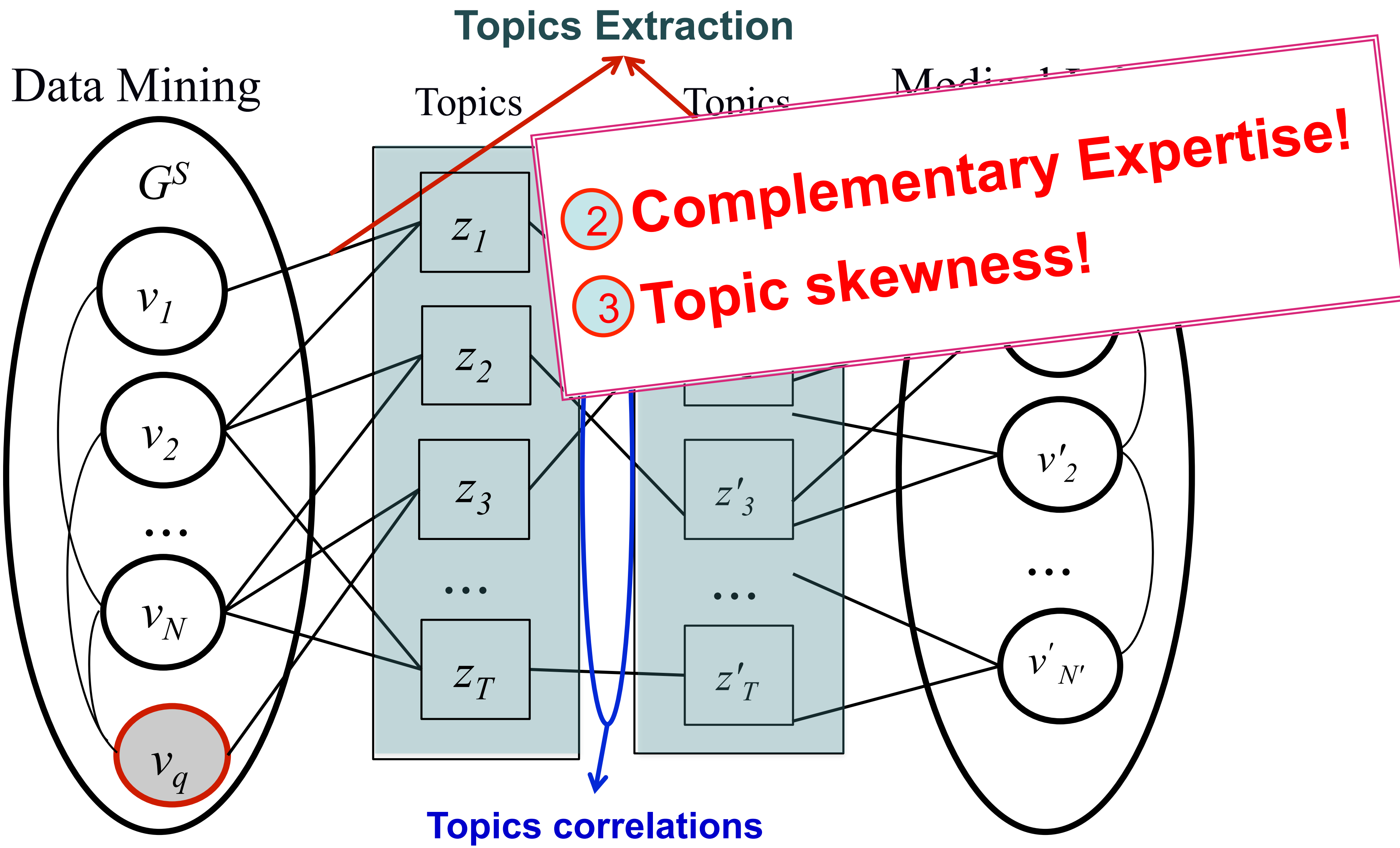
[1] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In ICDM'05, pages 418–425, 2005.

Author Matching

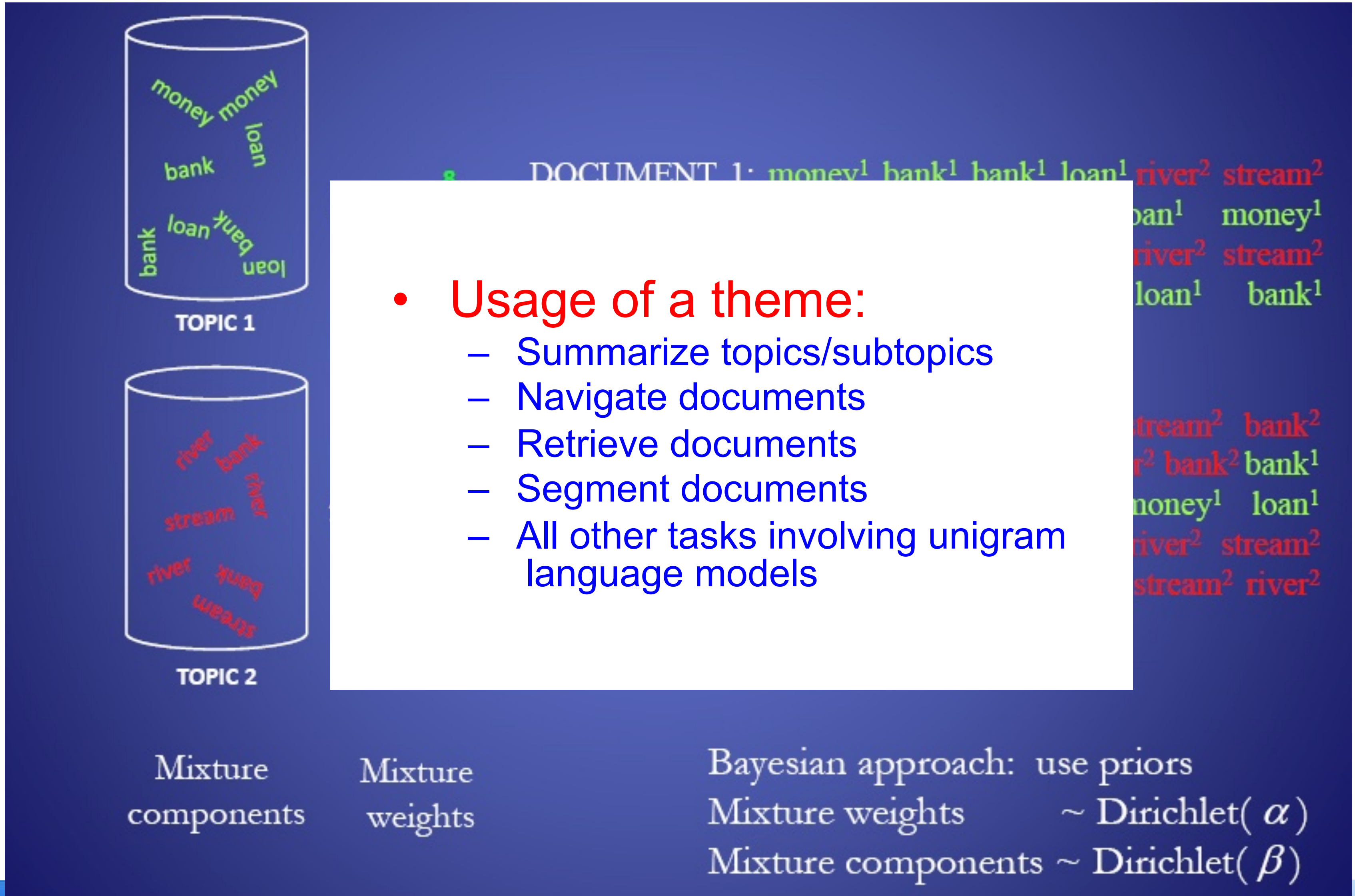


$$\mathbf{r}^{(t+1)} = (1-\tau)\mathbf{S}\cdot\mathbf{r}^{(t)} + \tau\boxed{\mathbf{q}}$$

Topic Matching

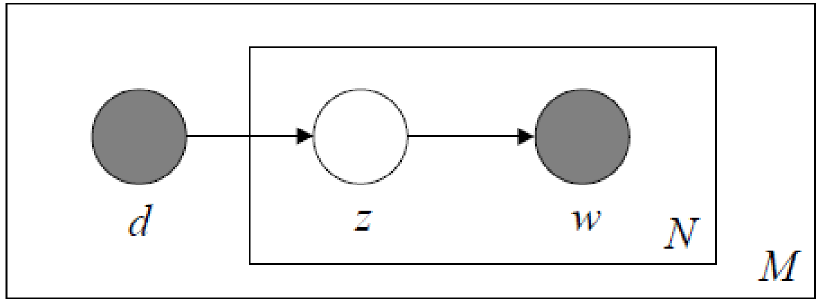


Recall Topic Model

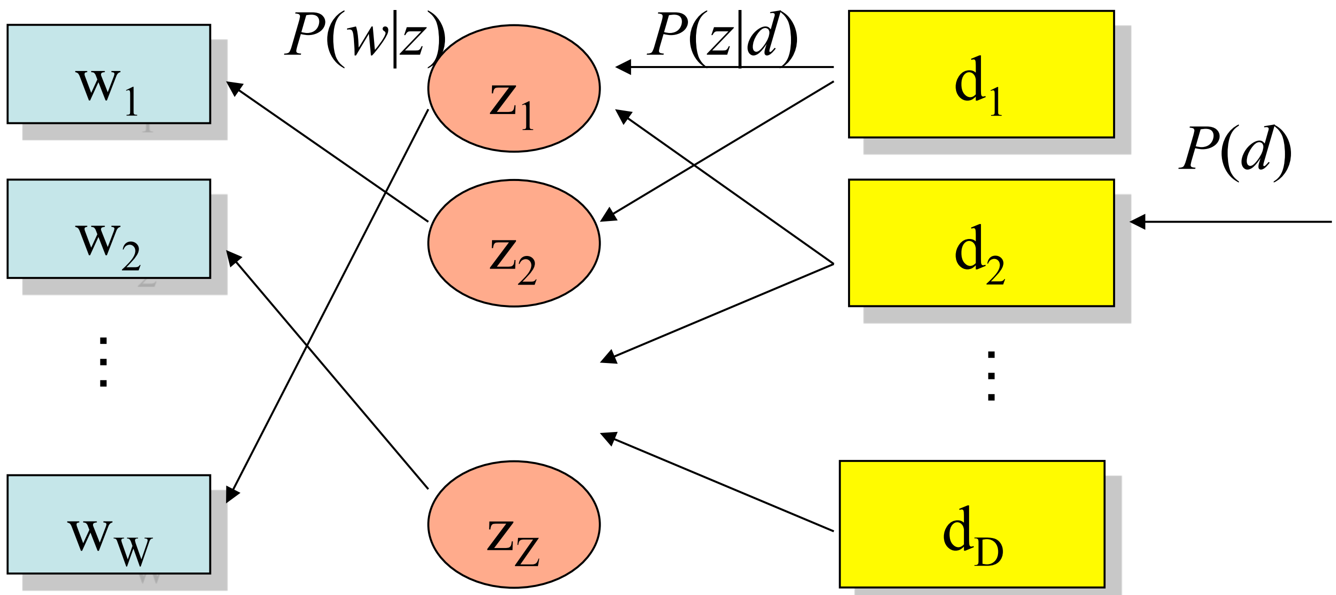


- Usage of a theme:
  - Summarize topics/subtopics
  - Navigate documents
  - Retrieve documents
  - Segment documents
  - All other tasks involving unigram language models

Topic Model

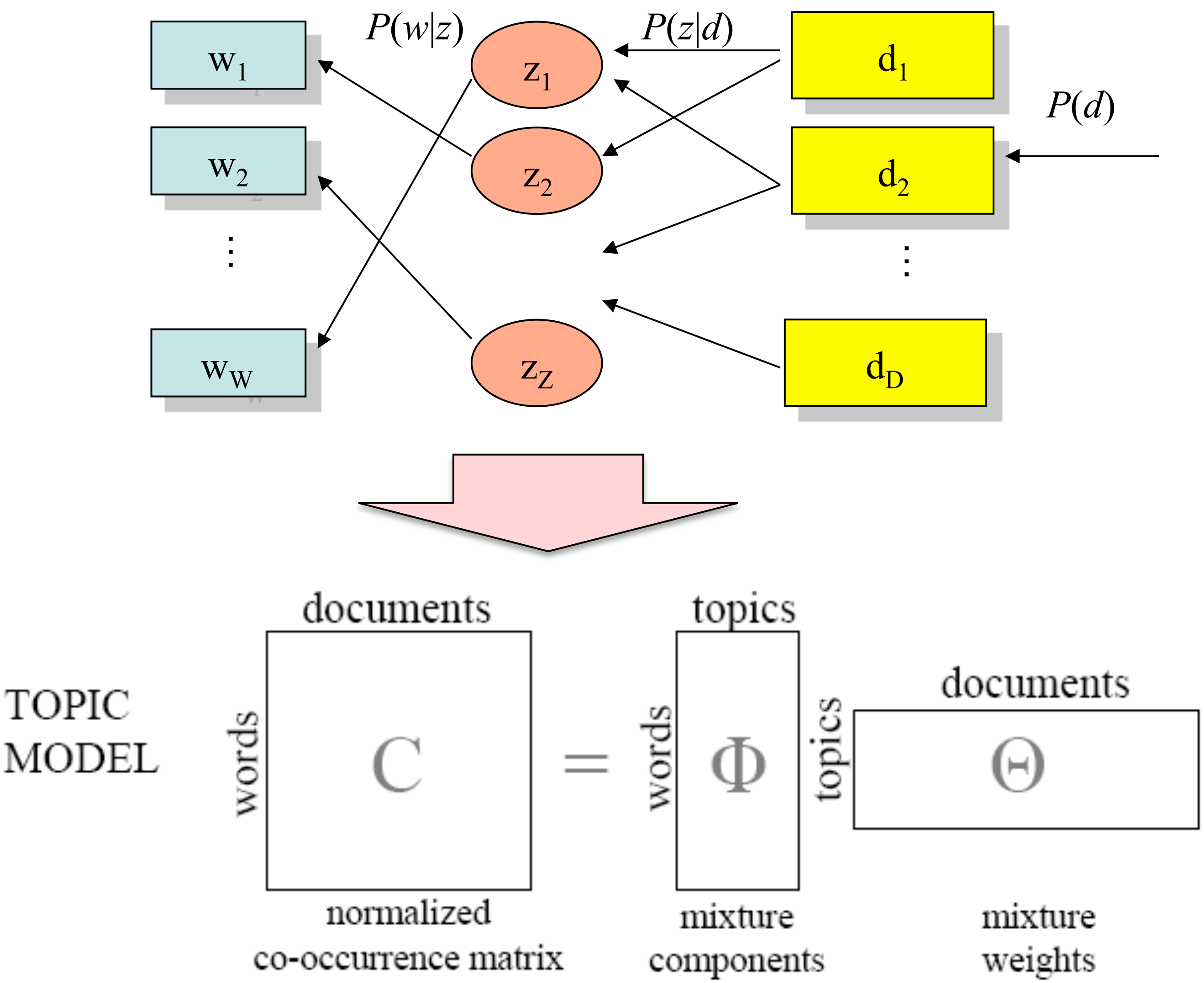


- A generative model for generating the co-occurrence of documents $d \in D=\{d_1,\ldots,d_D\}$ and terms $w \in W=\{w_1,\ldots,w_W\}$, which associates latent variable $z \in Z=\{z_1,\ldots,z_Z\}$.
- The generative processing is:



[1] T. Hofmann. Probabilistic latent semantic indexing. SIGIR'99, pages 50–57, 1999.

Topic Model



P(w|z)
P(z|d)
P(d)
w1
w2
wW
z1
z2
zZ
d1
d2
dD
TOPIC
MODEL
documents
words
C
=
topics
words
Φ
topics
documents
Θ
normalized
co-occurrence matrix
mixture
components
mixture
weights

Maximum-likelihood

- Definition
  - We have a density function $P(x|\Theta)$ that is govened by the set of parameters Θ, e.g., P might be a set of Gaussians and Θ could be the means and covariances
  - We also have a data set $X=\{x_1,\ldots,x_N\}$, supposedly drawn from this distribution *P,* and *a*ssume these data vectors are i.i.d. with *P.*
  - Then the log-likehihood function is:

$$L(\Theta \mid X) = \log p(X \mid \Theta) = \log \prod_i p(x_i \mid \Theta) = \sum_i \log p(x_i \mid \Theta)$$

  - The log-likelihood is thought of as a function of the parameters Θ where the data X is fixed. Our goal is to find the Θ that maximizes L. That is

$$\Theta^* = \arg\max_{\Theta} L(\Theta \mid X)$$

Topic Model

- Following the likelihood principle, we determines $P(d)$, $P(z|d)$, and $P(w|d)$ by maximization of the log-likelihood function

$P(d)$, $P(z|d)$, and $P(w|d)$

Unobserved data

co-occurrence times of *d* and *w*. Which is obtained according to the multi-distribution

$$L(\Theta \mid d, w, z) = \log \prod_{d} \prod_{w} P(d,w)^{n(d,w)}$$

Observed data

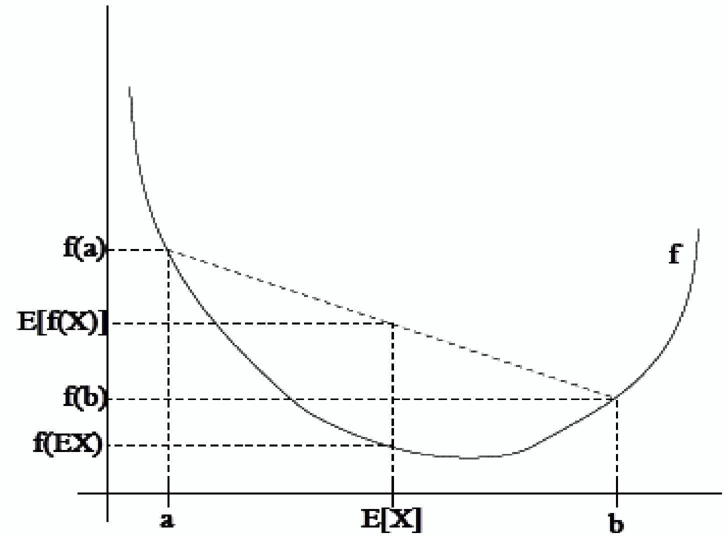
$$= \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w)$$

$$= \sum_{d \in D} \sum_{w \in W} n(d,w) \log \left(\sum_{z \in Z} P(w \mid z) P(d \mid z) P(z) \right)$$

Jensen's Inequality

- Recall that *f* is a convex function if *f* "(x)≥0, and *f* is strictly convex function if *f* "(x)>0
- Let *f* be a convex function, and let *X* be a random variable, then:

$$E[f(X)] \geq f(EX)$$



- Moreover, if *f* is strictly convex, then *E*[*f*(*X*)]=*f*(*EX*) holds true if and only if *X*=*E*[*X*] with probability 1 (i.e., if *X* is a constant)

Basic EM Algorithm

- However, Optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters:

$$L(\Theta \mid X) = \sum_i \log p(x_i \mid \Theta) = \sum_i \log \sum_z p(x_i, z \mid \Theta)$$

- Maximizing L(Θ) explicitly might be difficult, and the strategy is to instead repeatedly construct a lower-bound on L(E-step), and then optimize that lower bound (M-step).
  - For each i, let Q_i be some distribution over z ($\sum_z Q_i(z)=1$, $Q_i(z)\geq 0$), then

$$\sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \Theta) = \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})} \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \Theta)}{Q_i(z^{(i)})}$$

  - The above derivation used Jensen's inequality. Specifically, f(x) = logx is a concave function, since $f''(x)=-1/x^2<0$

Parameter Estimation-Using EM

- According to Basic EM:

$$Q_i(z^{(i)}) = p(z^{(i)} \mid x^{(i)}; \Theta)$$

- Then we define

$$Q_i(z^{(i)}) = p(z \mid d, w)$$

- Thus according to Jensen's inequality

$$L(\Theta) = \sum_{d \in D} \sum_{w \in W} n(d,w) \log \sum_{z \in Z} p(z \mid d,w) \frac{p(w \mid z) p(d \mid z) p(z)}{p(z \mid d,w)}$$

$$\geq \sum_{d \in D} \sum_{w \in W} n(d,w) \sum_{z \in Z} p(z \mid d,w) \log \frac{p(w \mid z) p(d \mid z) p(z)}{p(z \mid d,w)}$$

(1)Solve $P(w|z)$

- We introduce Lagrange multiplier λwith the constraint that $\sum_w P(w|z)=1$, and solve the following equation:

$$\frac{\partial}{\partial P(w|z)}\left\{\sum_{d\in D}\sum_{w\in W} n(d,w)\sum_{z\in Z} p(z|d,w)\log\frac{p(w|z)p(d|z)p(z)}{p(z|d,w)}+\lambda\left[\sum_{z} P(w|z)-1\right]\right\}=0$$

$$\therefore \frac{\sum_{d\in D} n(d,w)P(z|d,w)}{P(w|z)}+\lambda=0,$$

$$\therefore P(w|z)=-\frac{\sum_{d\in D} n(d,w)P(z|d,w)}{\lambda},$$

$$\sum_{w} P(w|z)=1,$$

$$\therefore \lambda=-\sum_{w\in W}\sum_{d\in D} n(d,w)P(z|d,w),$$

$$\therefore P(w|z)=\frac{\sum_{d\in D} n(d,w)P(z|d,w)}{\sum_{w\in W}\sum_{d\in D} n(d,w)P(z|d,w)}$$

The final update Equations

- E-step:

$$P(z|d,w)=\frac{P(w|z)P(d|z)P(z)}{\sum_{z\in Z}P(w|z)P(d|z)P(z)}$$

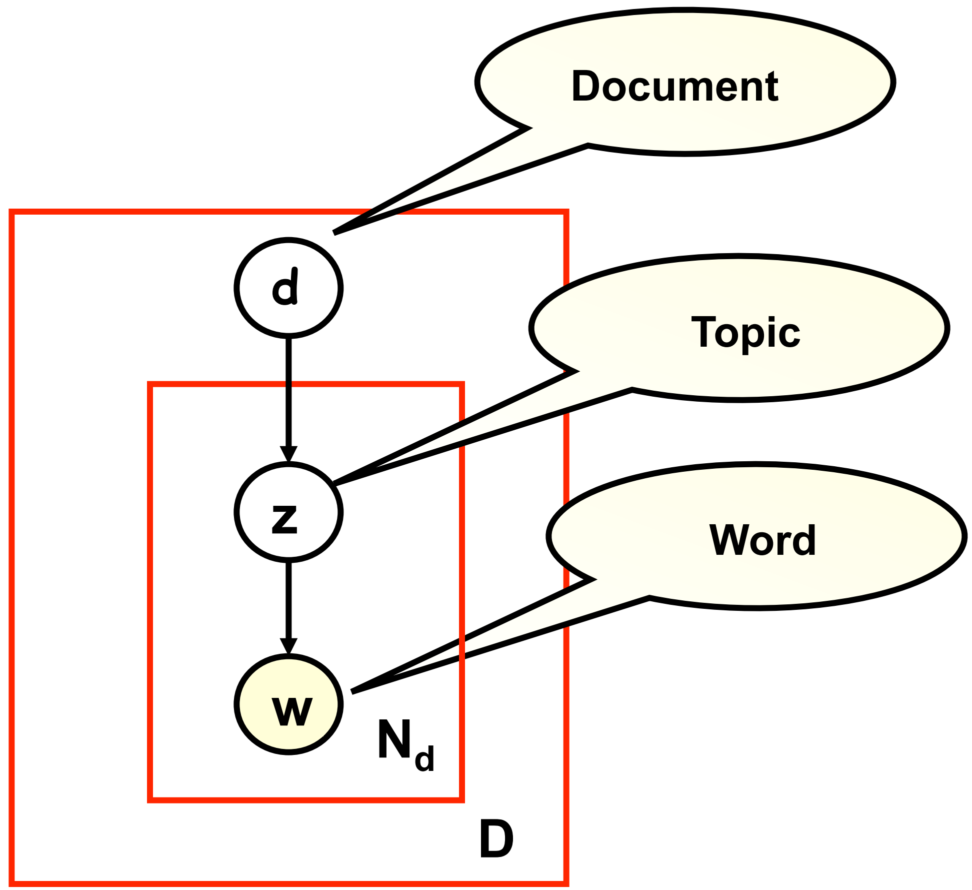
- M-step:

$$P(w|z)=\frac{\sum_{d\in D}n(d,w)P(z|d,w)}{\sum_{w\in W}\sum_{d\in D}n(d,w)P(z|d,w)}$$

$$P(d|z)=\frac{\sum_{w\in W}n(d,w)P(z|d,w)}{\sum_{d\in D}\sum_{w\in W}n(d,w)P(z|d,w)}$$

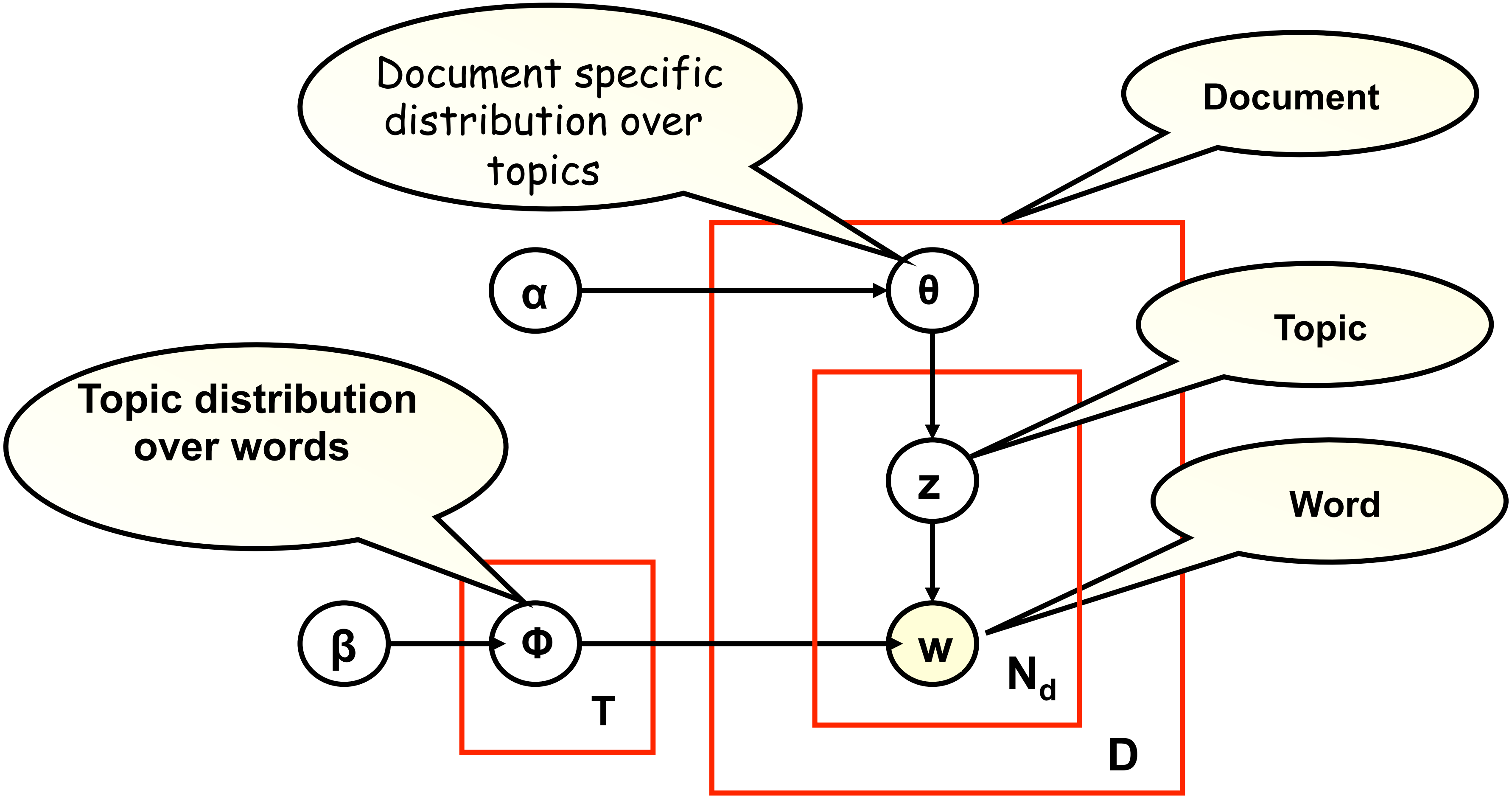
$$P(z)=\frac{\sum_{d\in D}\sum_{w\in W}n(d,w)P(z|d,w)}{\sum_{w\in W}\sum_{d\in D}n(d,w)}$$

PLSI(SIGIR'99)



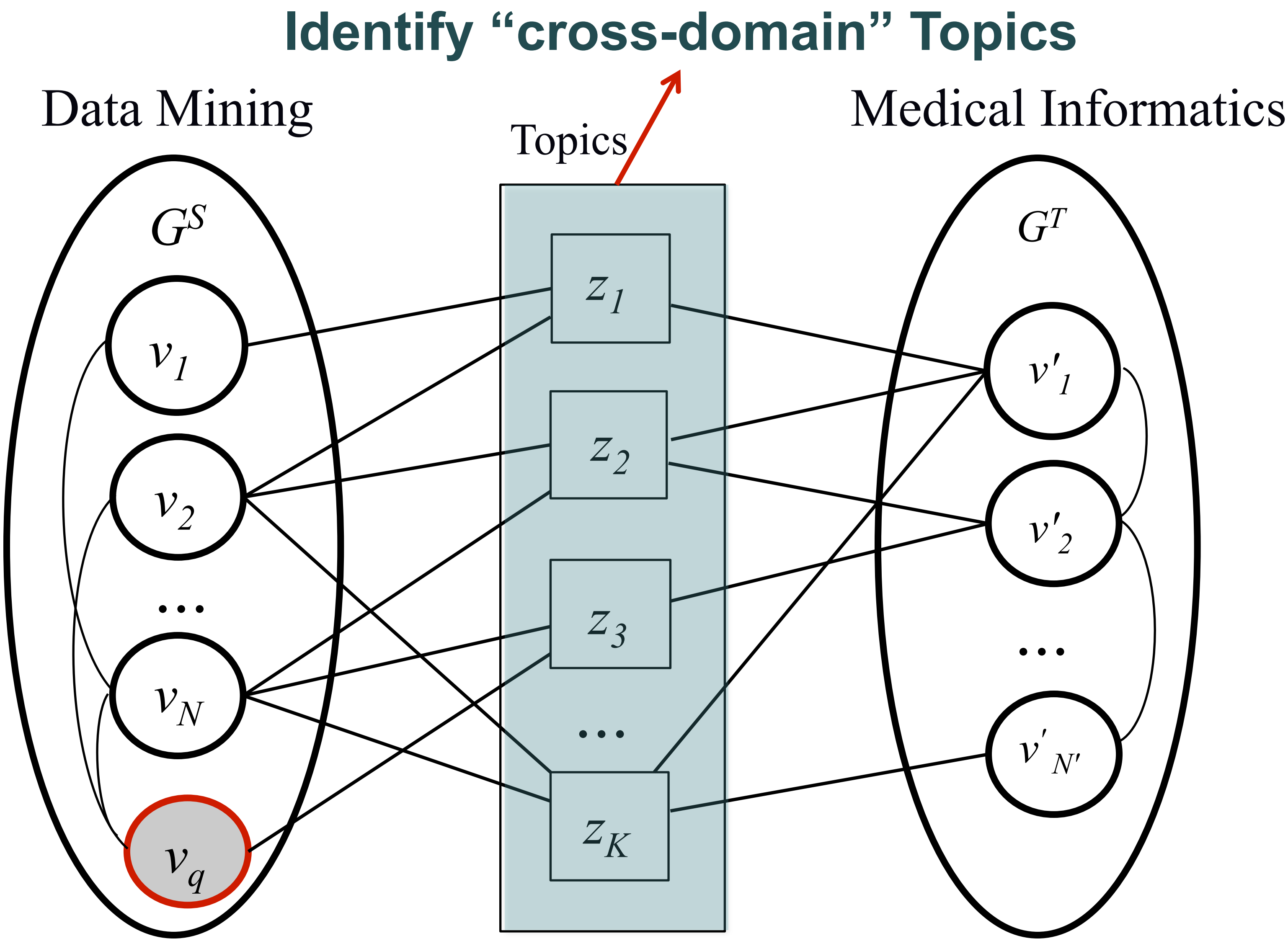
[1] T. Hofmann. Probabilistic latent semantic indexing. SIGIR'99, pages 50–57, 1999.

LDA (JMLR'03)

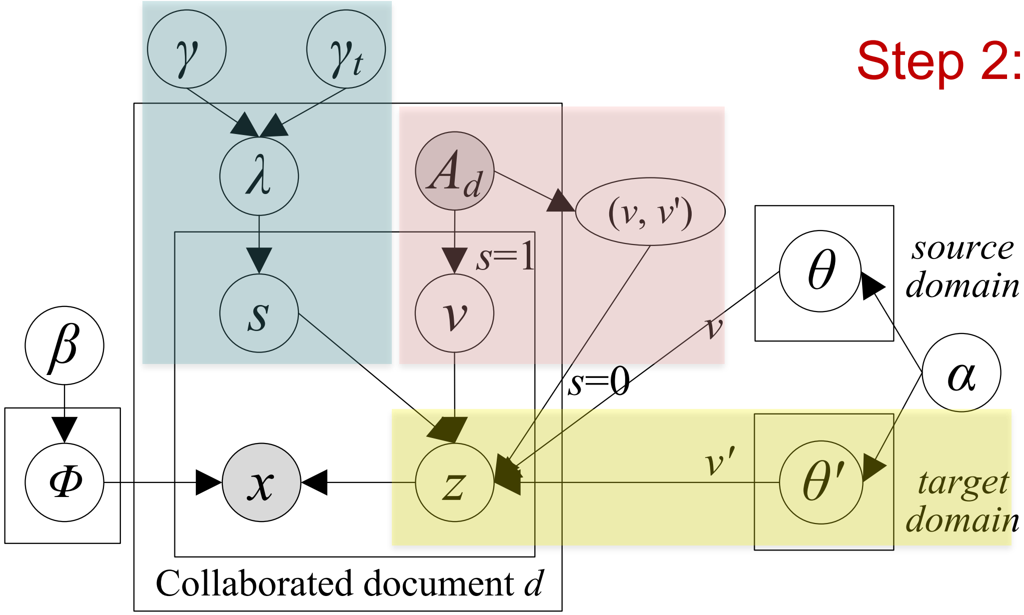


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR, 3:993–1022, 2003.

Cross-domain Topic Learning



Collaboration Topics Extraction



Input: a source domain G^S and a target domain G^T
Output: estimated parameters θ,θ',ϕ, ϑ, and λ

Step 1:

Initialize an ACT model in G^S by learning from documents written by authors only from G^S;
Similarly, initialize an ACT model for target domain G^T;

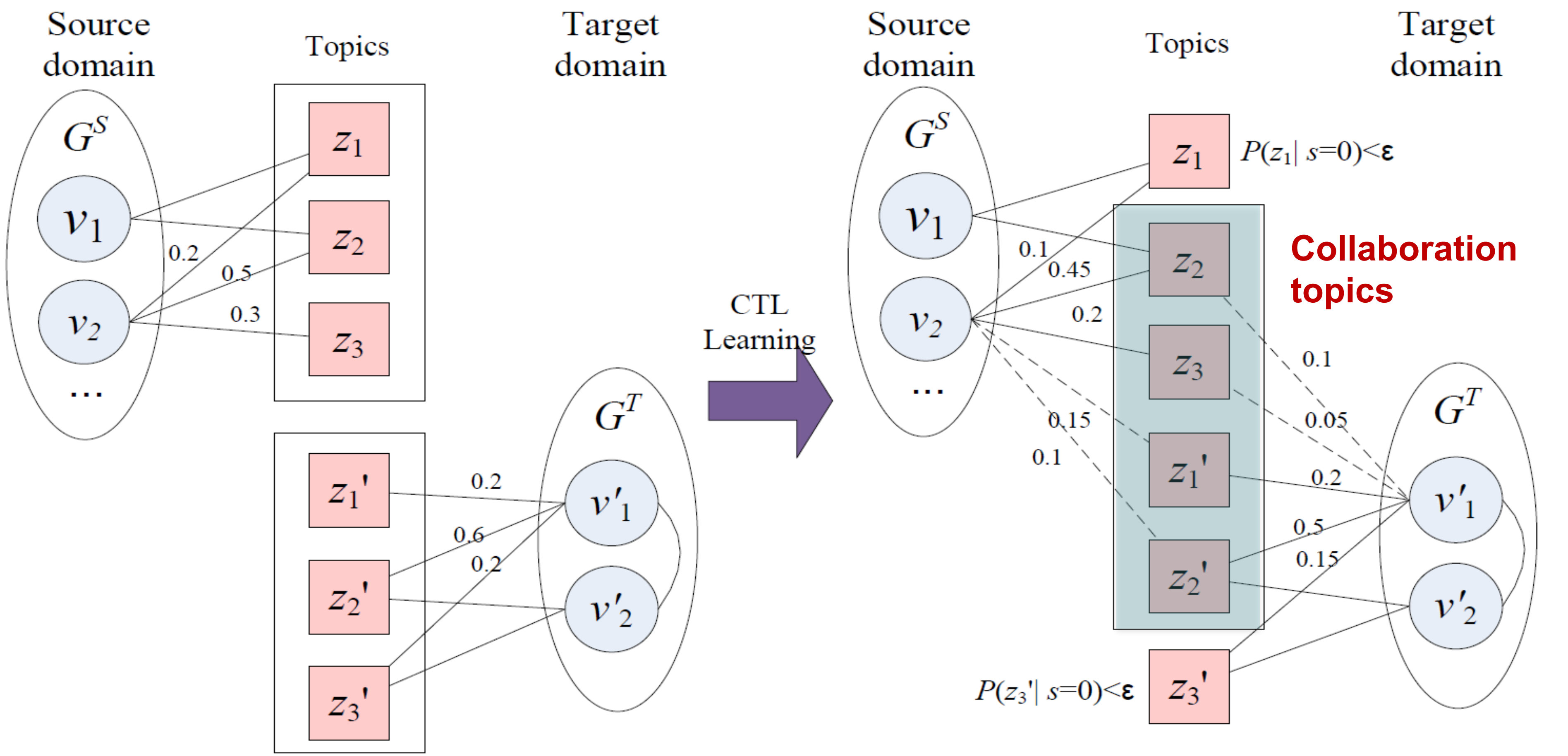
foreach *collaborated document* d **do**

Step 2:

- **foreach** *word* $x_{di} \in d$ **do**
  - Toss a coin s_{di} according to $bernoulli(s_{di}) \sim beta(\gamma_t, \gamma)$, where $beta(.)$ is a Beta distribution, and γ_t and γ are two parameters;
  - **if** $s_{di} = 0$ **then**
    - Randomly select a pair (v, v') from d's authors, where v is an author from G^S and v' from G^T;
    - Draw a topic $z_{di} \sim multi(\vartheta_{vv'})$ from the topic mixture $\vartheta_{vv'}$ specific to (v, v');
  - **end**
  - **if** $s_{di} = 1$ **then**
    - Randomly select a user v;
    - Draw a topic $z_{di} \sim multi(\theta_v)$ from the topic model of user v;
  - **end**
- **end**
- Draw a word $x_{di} \sim multi(\phi_{z_{di}})$ from z_{di}-specific word distribution;

end

Intuitive explanation of Step 2 in CTL



cross-domain collaboration recommendation

Experiments

Data Set and Baselines

- Arnetminer (available at http://arnetminer.org/collaboration)

Domain	Authors	Relationships	Source
Data Mining	6,282	22,862	KDD, SDM, ICDM, WSDM, PKDD
Medical Informatics	9,150	31,851	JAMIA, JBI, AIM, TMI, TITB
Theory	5,449	27,712	STOC, FOCS, SODA
Visualization	5,268	19,261	CVPR, ICCV, VAST, TVCG, IV
Database	7,590	37,592	SIGMOD, VLDB, ICDE

- Baselines
  - Content Similarity(Content)
  - Collaborative Filtering(CF)
  - Hybrid
  - Katz
  - Author Matching(Author), Topic Matching(Topic)

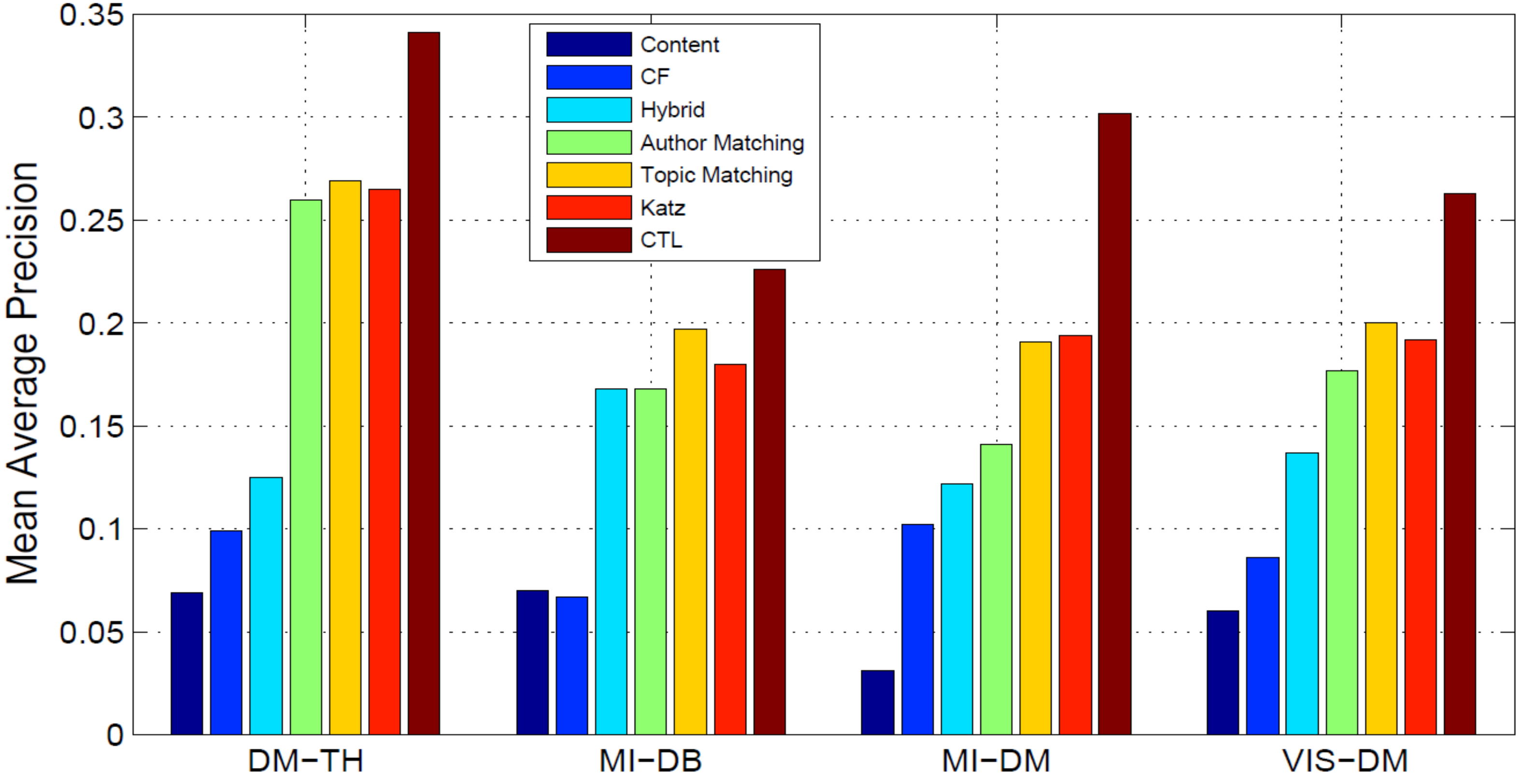
Performance Analysis

Training: collaboration before 2001 Validation: 2001-2005

Cross Domain	ALG	P@10	P@20	MAP	R@100	ARHR -10	ARHR -20
Data Mining(S) to Theory(T)	Content	10.3	10.2	10.9	31.4	4.9	2.1
	CF	15.6	13.3	23.1	26.2	4.9	2.8
	Hybrid	17.4	19.1	20.0	29.5	5.0	2.4
	Author	27.2	22.3	25.7	32.4	10.1	6.4
	Topic	28.0	26.0	32.4	33.5	13.4	7.1
	Katz	30.4	29.8	21.6	27.4	11.2	5.9
	CTL	37.7	36.4	40.6	35.6	14.3	7.5

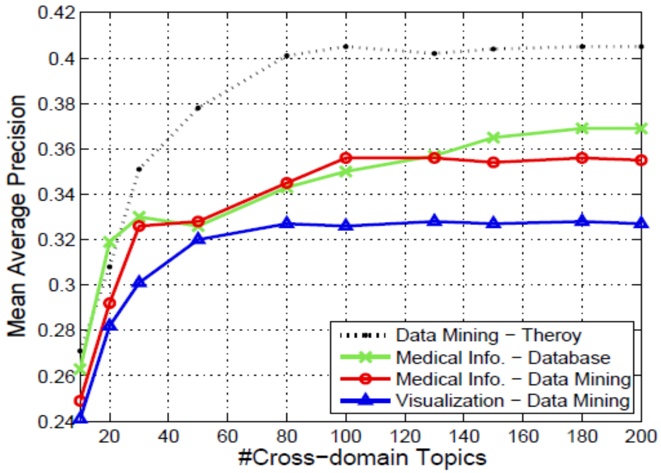
Content Similarity(Content): based on similarity between authors' publications
Collaborative Filtering(CF): based on existing collaborations
Hybrid: a linear combination of the scores obtained by the Content and the CF methods.
Katz: the best link predictor in link-prediction problem for social networks
Author Matching(Author): based on the random walk with restart on the collaboration graph
Topic Matching(Topic): combining the extracted topics into the random walking algorithm

Performance on New Collaboration Prediction

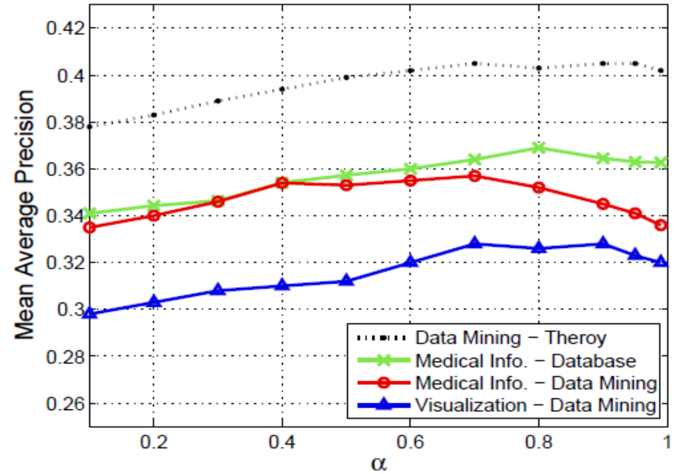


CTL can still maintain about 0.3 in terms of MAP which is significantly higher than baselines.

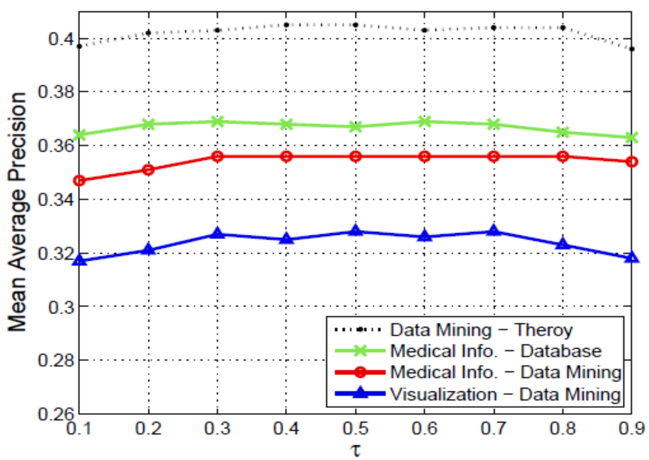
Parameter Analysis



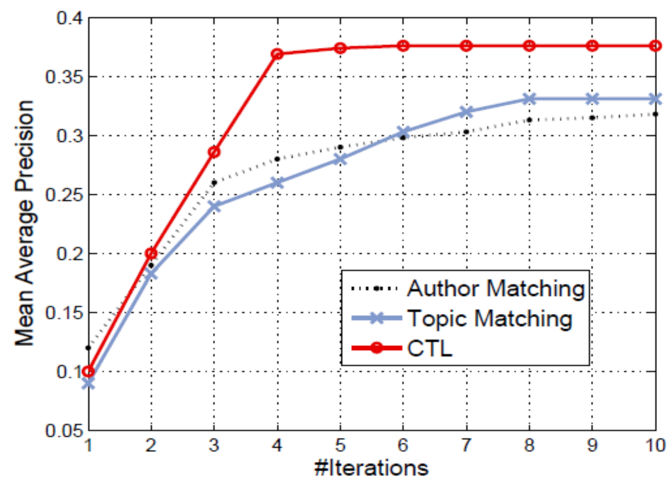
(a) number of topics T



(b) Hyperparameter α



(c) RWR parameter τ



(d) Convergence analysis

(a) varying the number of topics T

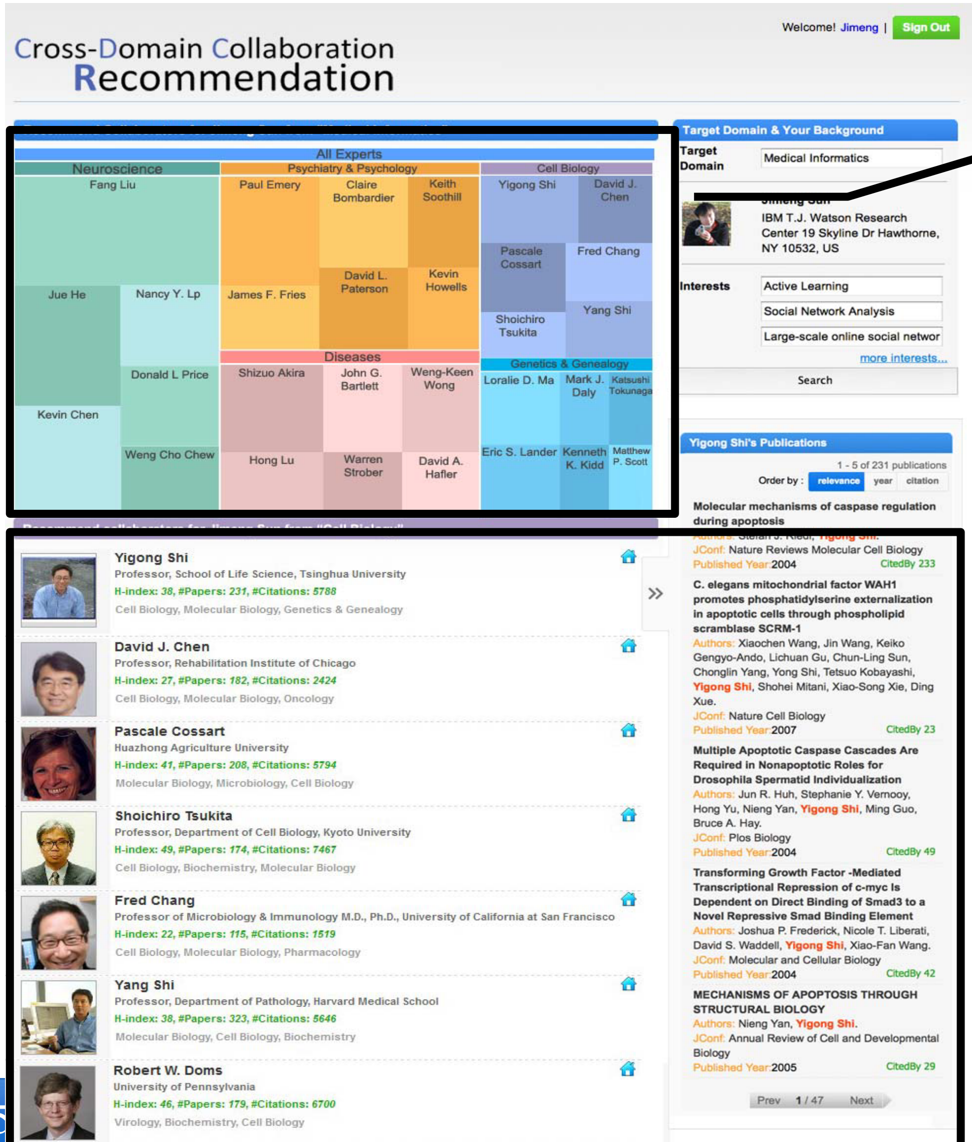
(b) varying α parameter

(c) varying the restart parameter τ in the random walk

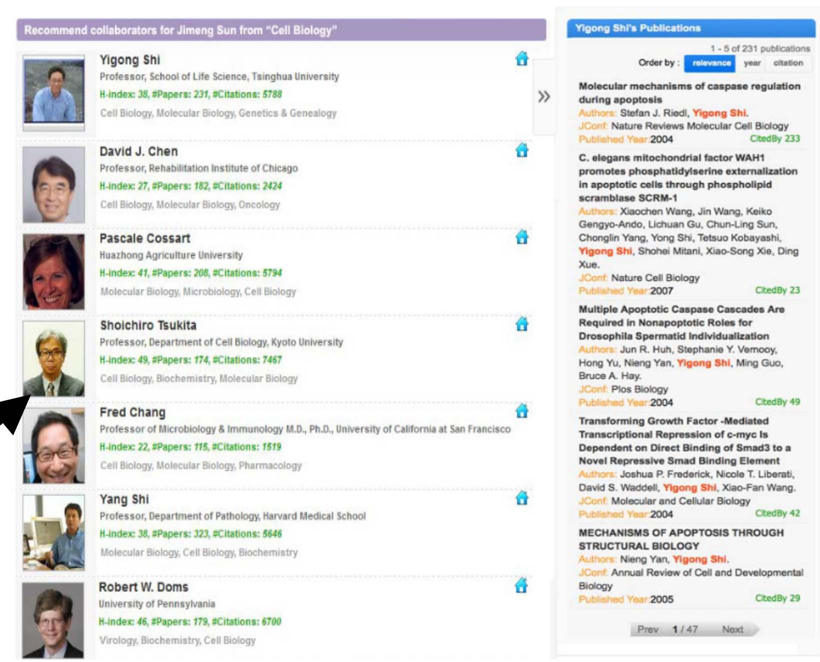
(d) Convergence analysis

Prototype System

http://arnetminer.org/collaborator



Treemap: representing subtopic in the target domain



Recommend Collaborators & Their relevant publications

Part C:

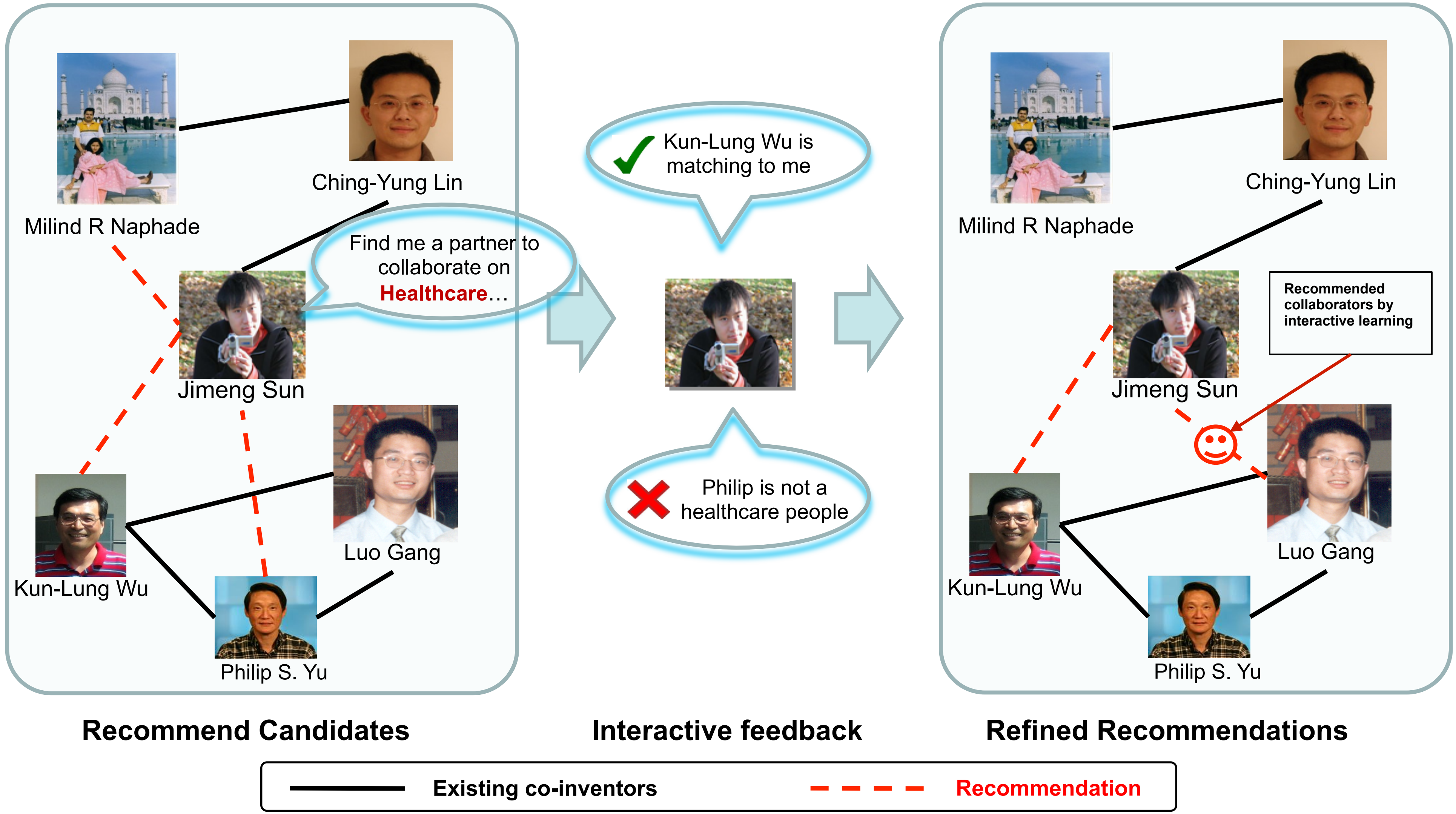
Further incorporate user feedback

"interactive collaboration recommendation"

(ACM TKDD, TIST, WSDM 2013-14)

Example

Finding co-inventors in IBM (>300,000 employers)



[1] S. Wu, J. Sun, and J. Tang. Patent Partner Recommendation in Enterprise Social Networks. WSDM'13, pages 43-52.

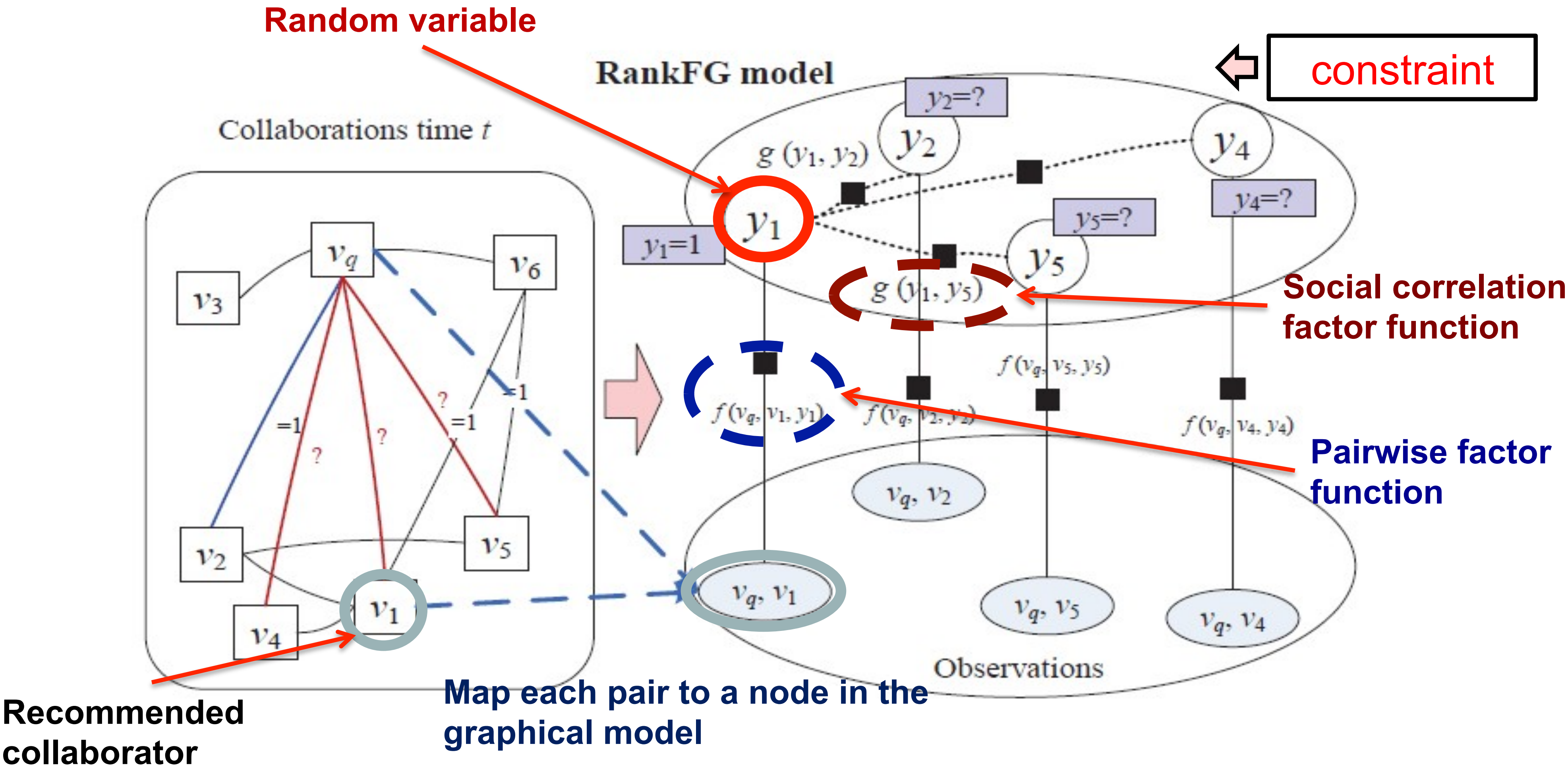
Challenges

- What are the **fundamental factors** that influence the formation of co-invention relationships?
- How to design an **interactive mechanism** so that the user can provide feedback to the system to refine the recommendations?
- How to learn the interactive recommendation framework in an **online** mode?

interactive collaboration recommendation

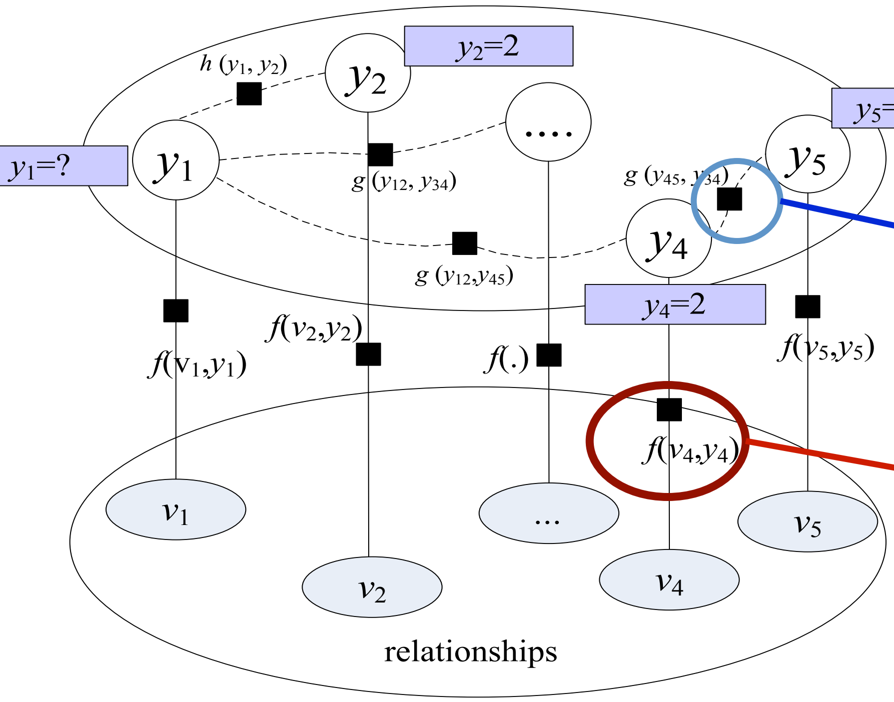
Learning framework

RankFG Model



The problem is cast as, for each relationship, identifying which type has the **highest probability**.

Modeling with exponential family



Partially Labeled Model

$$P(y_i \mid Y_{-i}) \propto \exp\{\sum_{c_i} \sum_k \mu_k h_k(Y_{c_i})\}$$

$$P(x_i \mid y_i) \propto \exp\{\sum_{j=1}^{d} \alpha_j g_j(x_{ij}, y_i)\}$$

Likelihood objective function

$$P(Y \mid X, G) = \frac{P(X, G \mid Y) P(Y)}{P(X, G)}$$

$$\propto P(X \mid Y) \cdot P(Y \mid G) = P(Y \mid G) \prod_i P(x_i \mid y_i)$$

Ranking Factor Graphs

- Pairwise factor function:

$$f(v_q, v_i, y_i) = \frac{1}{Z_a} \exp\{\sum_k \alpha_k \psi_k(\mathbf{x}_q, \mathbf{x}_i, y_i)\}$$

- Correlation factor function:

$$g(y_i, y_j) = \frac{1}{Z_b} \exp\{\sum_l \beta_l \phi_l(y_i, y_j)\}$$

- Log-likelihood objective function:

$$\log P(Y|X,\theta) = \sum_{y_i \in Y} \sum_k \alpha_k \psi_k(\mathbf{x_q}, \mathbf{x_i}, y_i) + \sum_{v_i \sim v_j} \sum_l \beta_l \phi_l(y_i, y_j) - \log Z$$

- Model learning

$$\theta^\star = \arg\max_\theta \log P(Y|X,\theta)$$

Learning Algorithm

Input: Query inventors $Q = \{v_q\}$ with corresponding topics $\{q\}$, $G = (V, E, X)$, and the learning rate η;
Output: learned parameters θ;

$\theta \leftarrow \mathbf{0}$;
repeat
  foreach $v_q \in Q$ *and* q **do**
    //Initialization;
    $L \leftarrow$ initialization list;
    Factor graph $FG \leftarrow BuildFactorGraph(L)$;
    // Learn the parameter θ for factor graph model;
    repeat
      foreach $v_i \in order$ **do**
        Update the messages of v_i by Eqs. 8 and 9;
      end
    until *(all messages μ do not change)*;
    foreach $\theta_i \in \theta$ **do**
      Calculate gradient ∇_i according to Eq. 7;
      Update $\theta^{new} = \theta^{old} + \eta \cdot \nabla_i$;
    end
  end
until *converge*;

Expectation Computing

Loopy Belief Propagation

Algorithm 1: Learning algorithm for RankFG.

Still Challenge

How to incrementally incorporate users' feedback?

Learning Algorithm

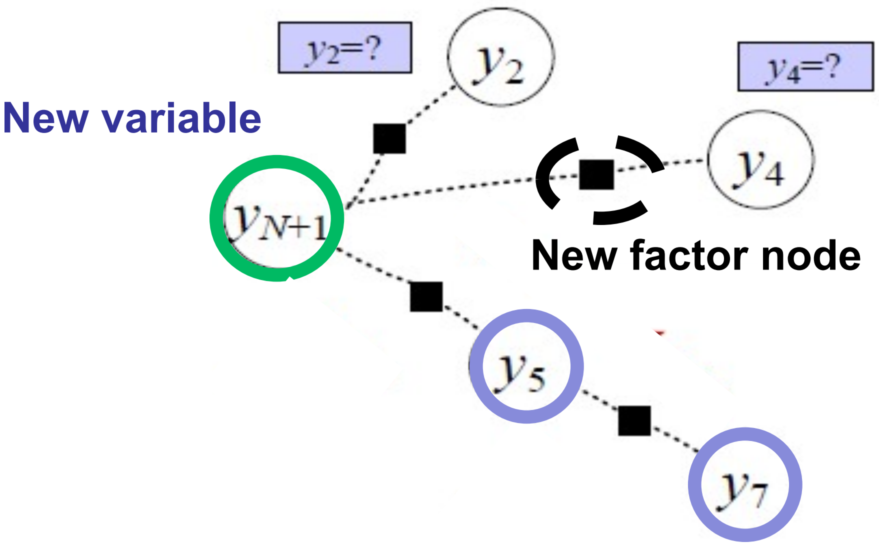
Input: Query inventors $Q = \{v_q\}$ with corresponding topics $\{q\}$, $G = (V, E, X)$, and the learning rate η;
Output: learned parameters θ;

$\theta \leftarrow \mathbf{0}$;
repeat
    foreach $v_q \in Q$ *and* q **do**
        //Initialization;
        $L \leftarrow$ initialization list;
        Factor graph $FG \leftarrow BuildFactorGraph(L)$;
        // Learn the parameter θ for factor graph model;
        repeat
            foreach $v_i \in order$ **do**
                Update the messages of v_i by Eqs. 8 and 9;
            end
        until *(all messages μ do not change)*;
        foreach $\theta_i \in \theta$ **do**
            Calculate gradient ∇_i according to Eq. 7;
            Update $\theta^{new} = \theta^{old} + \eta \cdot \nabla_i$;
        end
    end
until *converge*;

Algorithm 1: Learning algorithm for RankFG.

Incremental estimation

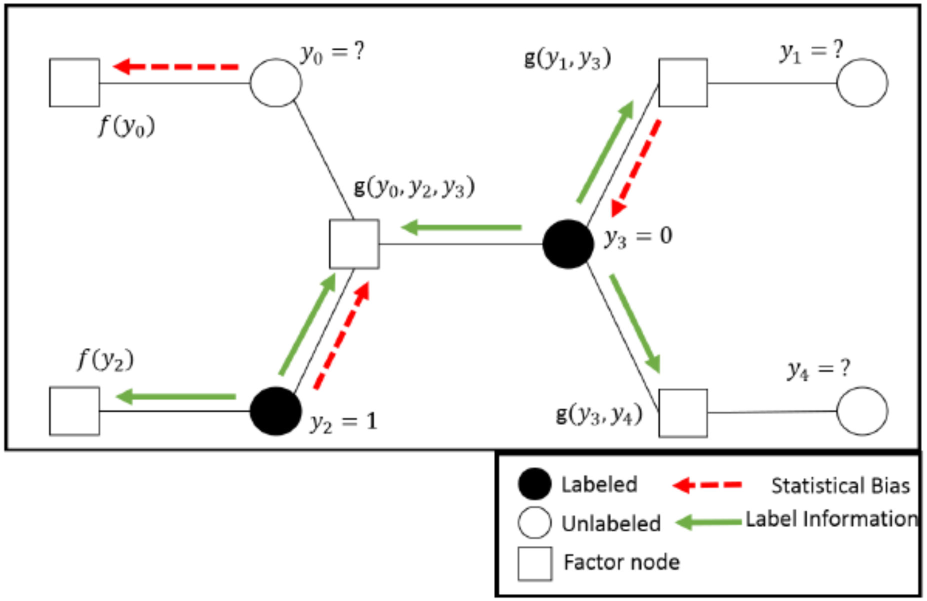
Interactive Learning



1) add new factor nodes to the factor graph built in the model learning process.
2) l-step message passing:
   - Start from the new variable node y_{N+1}ot node).
   - Send messages to all of its neighborhood factors.
   - Propagate the messages up to l-step
   - Perform a backward messages passing.
3) Calculate an approximate value of the marginal probabilities of the newly factors.

$$\mathbb{E}^{new}[.] = \frac{N}{N+1}\mathbb{E}^{old}[.] + \frac{1}{N+1}\sum_k \theta_k \phi_k(\mathbf{x}_{N+1}, \mathbf{y}_{N+1})$$

From passive interactive to active



- Entropy

$$\mu(v) = \sum_{y \in \mathcal{Y}} \mathcal{B}_v(y) \log \frac{1}{\mathcal{B}_v(y)}$$

- Threshold

$$t(v) = \min\{\lceil \eta(\mu\max - \mu(v))d(v) \rceil, d(v)\}$$

- Influence model

$$f_\tau(v) = \begin{cases} 1 & \text{if } \sum_{u \in \mathrm{NB}(v)} f_{\tau-1}(u) \geq t(v) \\ 0 & \text{if } \sum_{u \in \mathrm{NB}(v)} f_{\tau-1}(u) < t(v) \end{cases}$$

[1] Z. Yang, J. Tang, and B. Xu. Active Learning for Networked Data Based on Non-progressive Diffusion Model. WSDM'14.

[2] L. Shi, Y. Zhao, and J. Tang. Batch Mode Active Learning for Networked Data. ACM Transactions on Intelligent Systems and Technology (TIST), Volume 3, Issue 2 (2012), Pages 33:1--33:25.

Active learning via Non-progressive diffusion model

- Maximizing the diffusion

$$\max_{V_S \subseteq V_U} \{ \max_{V_T \subseteq V_U} |V_T| \}, \quad |V_S| \leq k$$

with the constraints:

$$f_0(v) = 1 \iff v \in V_S \tag{2}$$

$$\exists \tau_M \text{ s.t. } \forall v \in V_T \ \forall \tau > \tau_M \ f_\tau(v) = 1 \tag{3}$$

$$\forall v \in V_U \backslash V_T, \forall u \in V_T, \mu(v) \leq \mu(u) \tag{4}$$

$$f_\tau(v) = 1 \iff \sum_{u \in \mathrm{NB}(v)} f_{\tau-1}(u) \geq t(v) \tag{5}$$

NP-hard!

MinSS

- Greedily expand V_p

$V_P \leftarrow V_U \backslash V_\tau$

sort nodes in V_P in descending order of $t(v)$ as $v_1, v_2, \ldots, v_p$

foreach $v \in V_\tau$ **do**

  $w(v) \leftarrow 0$

for $i \leftarrow 1$ **to** p **do**

  if $w(u) < d(u) - t(u)\ \forall u \in NB(v_i) \cap V_\tau$ **then**

    foreach $u \in NB(v_i) \cap V_\tau$ **do**

      $w(u) \leftarrow w(u) + 1$

    $V_P \leftarrow V_P \backslash \{v_i\}$

MinSS(cont.)

if $V_P = \emptyset$ **then**
    sort nodes in V_τ in ascending order of $d(v)$ as $v_1, v_2, ..., v_m$
    foreach $v \in V_\tau$ **do**
        $w(v) \leftarrow 0$
    foreach $i \leftarrow 1$ **to** m **do**
        if $\exists u \in NB(v_i) \cap V_\tau$ *st.* $w(u) = d(u) - t(u)$ **then**
            $V_S \leftarrow V_S \cup \{v_i\}$
        else
            foreach $u \in NB(v_i) \cap V_\tau$ **do**
                $w(u) \leftarrow w(u) + 1$

Lower Bound and Upper Bound

Theorem 3. **Lower Bound.** Let $D(V) = \sum_{v \in V} d(v)$, $T(V) = \sum_{v \in V} t(v)$, and suppose $t(v) \leq \beta d(v)$ for all $v \in V$. If $2T(V_U) - D(V_U) > 0$ and $V_T = V_U$, we have an lower bound for optimal solution $\left|V_{S,\text{opt}}\right|$ to problem 2.

$$\left|V_{S,\text{opt}}\right| \geq \frac{2T(V_U) - D(V_U)}{\beta \Delta} \tag{9}$$

Theorem 4. **Upper Bound.** Suppose $t(v) \leq \beta d(v)$ for all $v \in V$, we can derive an upper bound for MinSS algorithm.

$$|V_S| \leq \frac{\beta \Delta}{1 - \beta + \beta \Delta} |V_U|$$

Approximation Ratio

$$\left|V_{S,\mathrm{opt}}\right| \geq \frac{2T(V_U) - D(V_U)}{\beta\Delta}$$

$$|V_S| \leq \frac{\beta\Delta}{1-\beta+\beta\Delta}|V_U|$$

COROLLARY 2. ***Approximation Ratio.*** *Let $V_{S,g}$ denote the solution given by MinSS algorithm, $V_{S,opt}$ represent the optimal solution and Δ be the maximum degree in the graph. Suppose $t(v) \leq \beta d(v)$ for all $v \in V$, if $V_T = V_U$ and $2T(V_U) > D(V_U)$, we have*

$$\frac{|V_{S,g}|}{\left|V_{S,opt}\right|} \leq \frac{(\beta\Delta)^2}{(1-\beta+\beta\Delta)\cdot\mathbb{E}[2t(v)-d(v)]} \qquad (10)$$

where $\mathbb{E}[.]$ represents the expectation over all samples in the network.

interactive collaboration recommendation

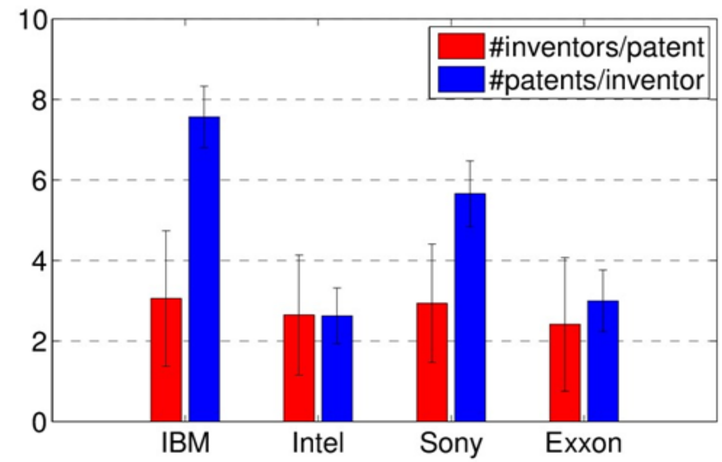
Experiments

Data Set

- PatentMiner (pminer.org)

DataSet	Inventors	Patents	Average increase #patent	Average increase #co-invention
IBM	55,967	46,782	8.26%	11.9%
Intel	18,264	54,095	18.8%	35.5%
Sony	8,505	31,569	11.7%	13/0%
Exxon	19,174	53,671	10.6%	14.7%

- Baselines:
  - Content Similarity (Content)
  - Collaborative Filtering (CF)
  - Hybrid
  - SVM-Rank



Statistics of four test data

[1] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, and A. K. Usadi. PatentMiner: Topic-driven Patent Analysis and Mining. **KDD'12**, pages 1366-1374.

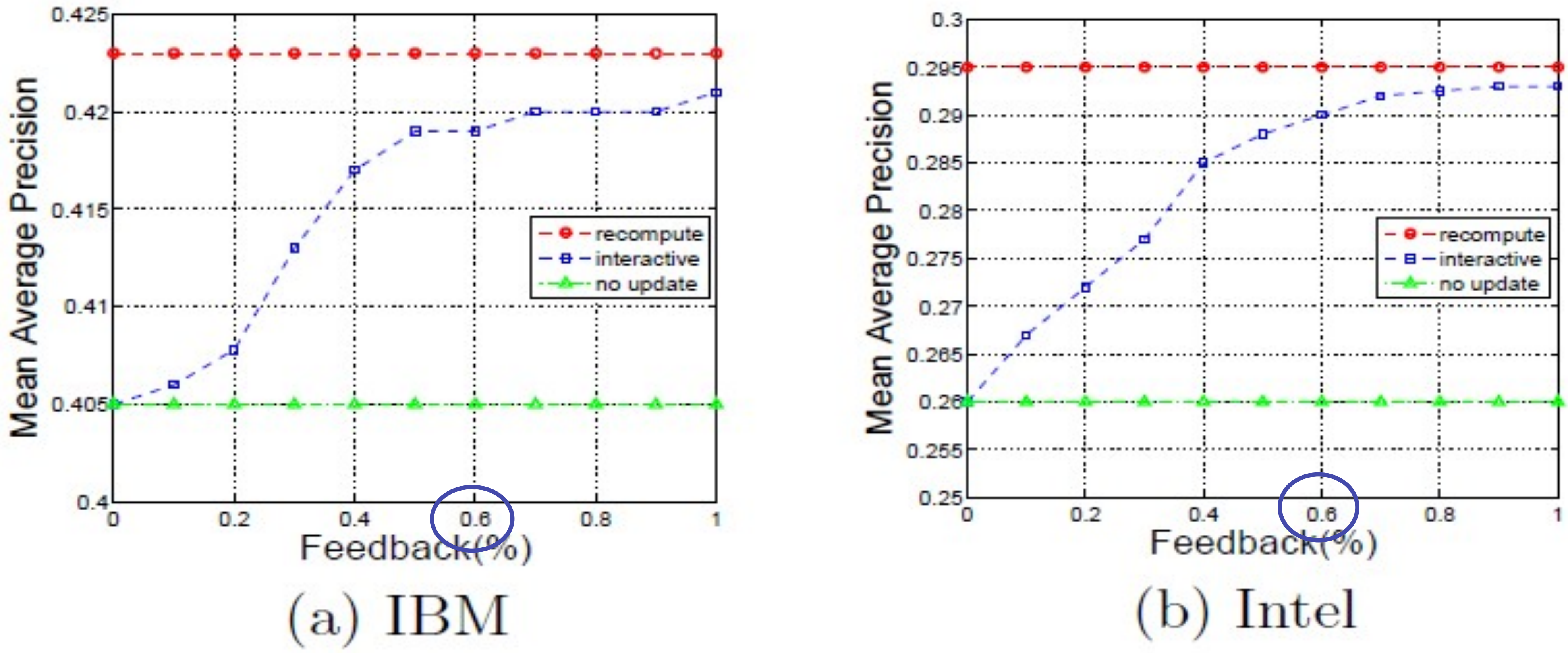
Performance Analysis-IBM

Training: collaboration before 2000 Validation: 2001-2010

Data	ALG	P@5	P@10	P@15	P@20	MAP	R@100
IBM	Content	23.0	23.3	18.8	15.6	24.0	33.7
	CF	13.8	12.8	11.3	11.5	21.7	36.4
	Hybrid	13.9	12.8	11.5	11.5	21.8	36.7
	SVMRank	13.3	11.9	9.6	9.8	22.2	43.5
	RankFG	31.1	27.5	25.6	22.4	40.5	46.8
	RankFG+	31.2	27.5	26.6	22.9	42.1	51.0

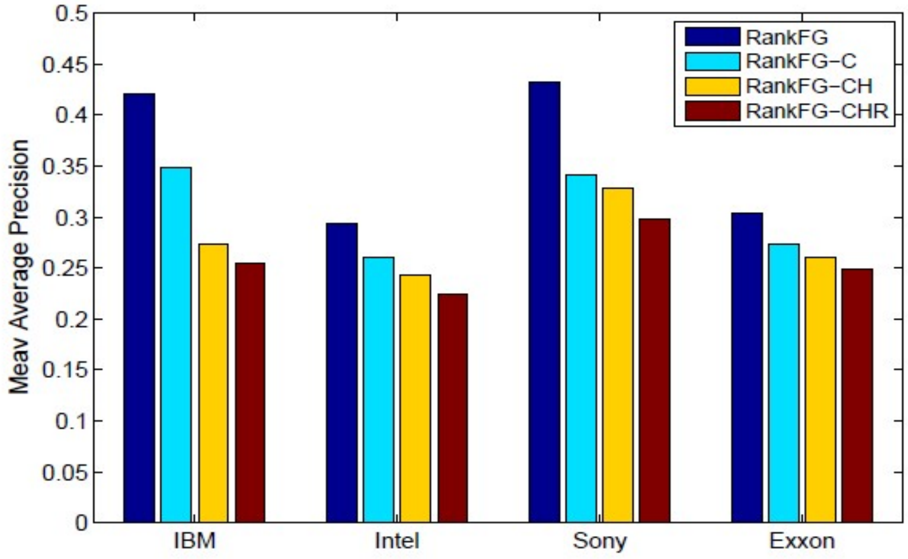
RankFG+: it uses the proposed RankFG model with 1% interactive feedback.

Interactive Learning Analysis

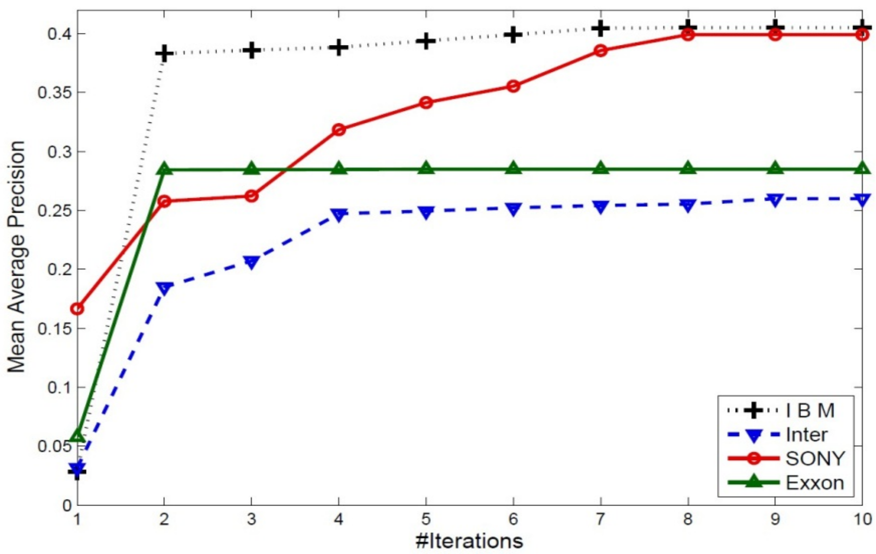


Interactive learning achieves a close performance to the complete learning with only **1/100** of the running time used for complete training.

Parameter Analysis



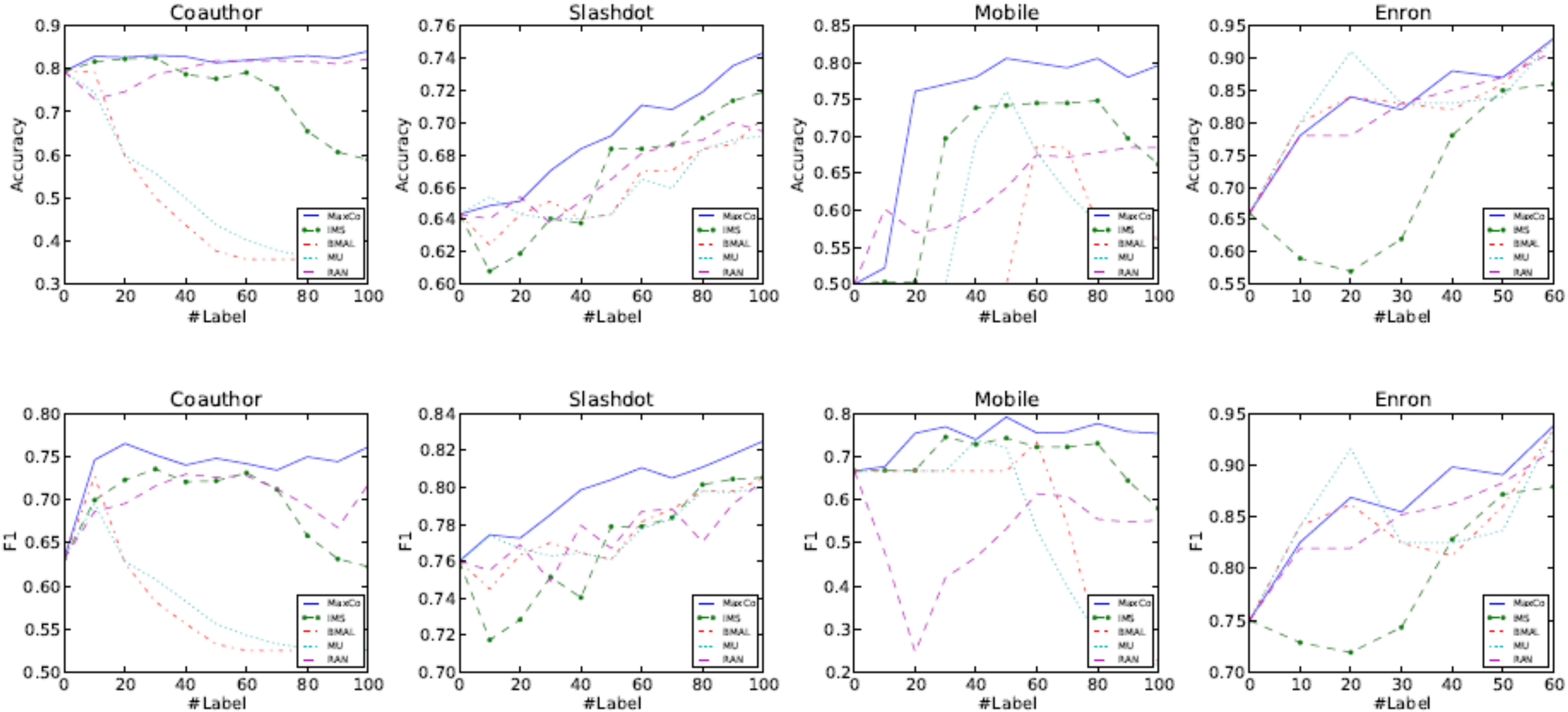
Factor contribution analysis



Convergence analysis

RankFG-C: stands for ignoring referral chaining factor functions.
RankFG-CH: stands for ignoring both referral chaining and homophily.
RankFG-CHR: stands for further ignoring recency.

Results of Active Learning

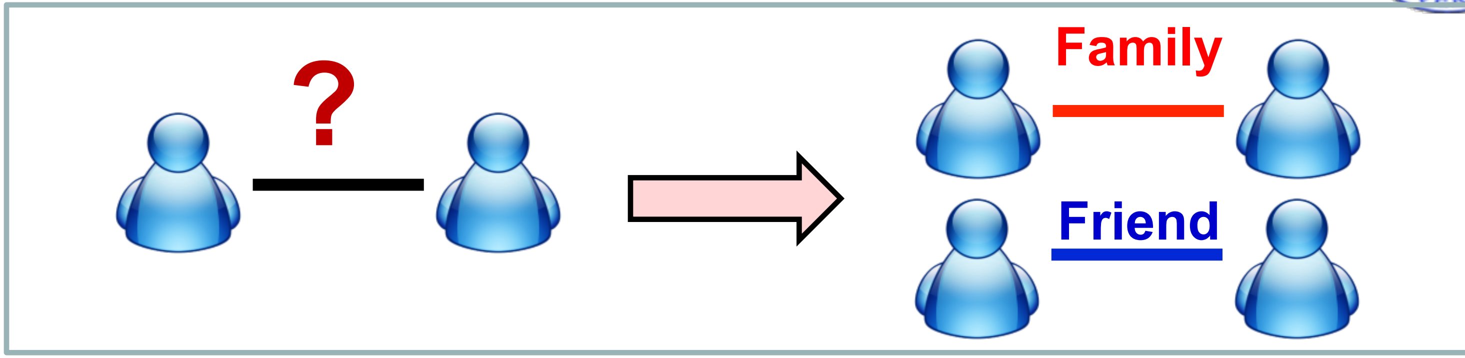


Summaries

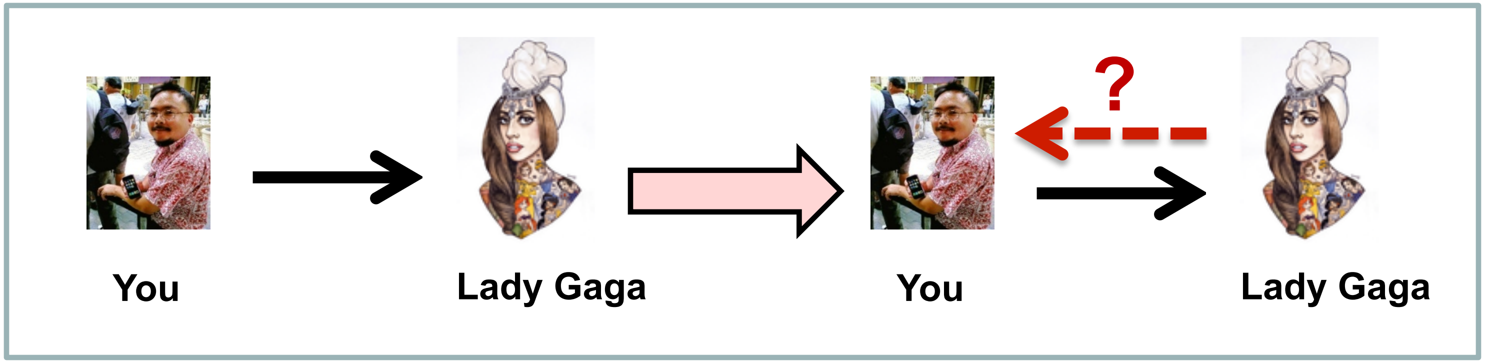
- Inferring social ties in single network
  - Time-dependent factor graph model
- Cross-domain collaboration recommendation
  - Cross-domain topic learning
- Interactive collaboration recommendation
  - Ranking factor graph model
  - Active learning via non-progressive diffusion

Future Work

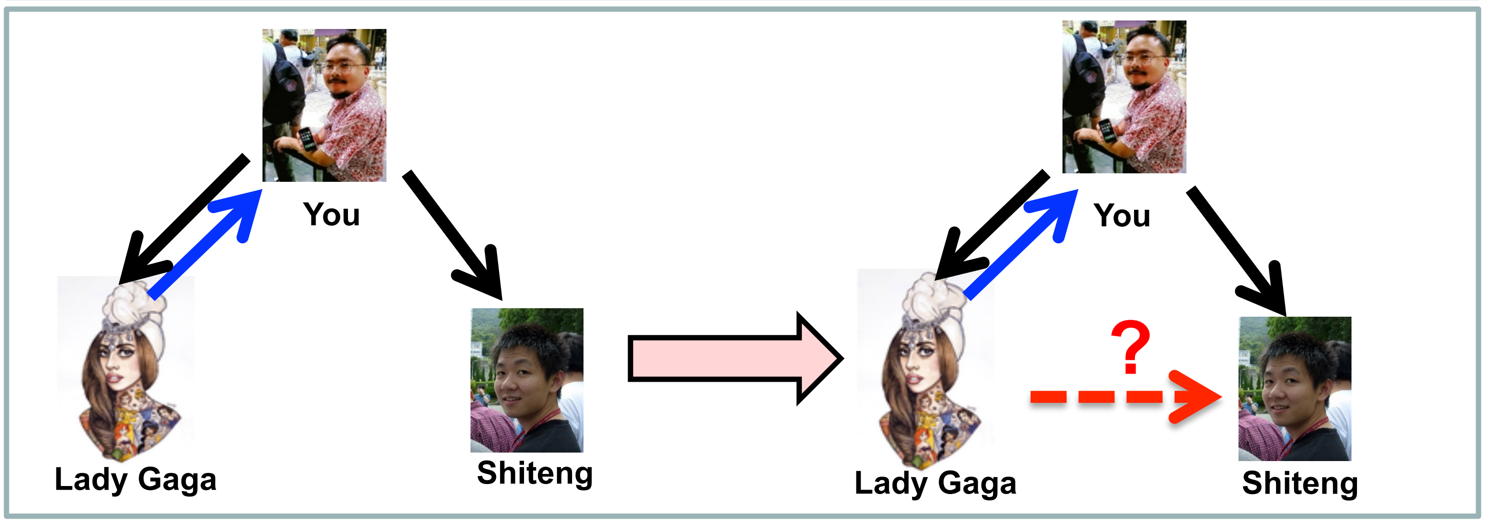
Inferring social ties



Reciprocity



Triadic Closure



References

- Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. In **TKDD**, 2013.
- Yi Cai, Ho-fung Leung, Qing Li, Hao Han, Jie Tang, Juanzi Li. Typicality-based Collaborative Filtering Recommendation. IEEE Transaction on Knowledge and Data Engineering (**TKDE**).
- Honglei Zhuang, Jie Tang, Wenbin Tang, Tiancheng Lou, Alvin Chin, and Xia Wang. Actively Learning to Infer Social Ties. **DMKD**, Vol. 25, Issue 2 (2012), pages 270-297.
- Lixin Shi, Yuhang Zhao, and Jie Tang. Batch Mode Active Learning for Networked Data. ACM Transactions on Intelligent Systems and Technology (**TIST**), Volume 3, Issue 2 (2012), Pages 33:1--33:25.
- Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. **Machine Learning Journal**, Vol. 82, Issue 2 (2011), pages 211-237.
- Zhilin Yang, Jie Tang, and Bin Xu. Active Learning for Networked Data Based on Non-progressive Diffusion Model. **WSDM'14**.
- Sen Wu, Jimeng Sun, and Jie Tang. Patent Partner Recommendation in Enterprise Social Networks. WSDM'13, pages 43-52.
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. **KDD'12**, pages 1285-1293. (Full Presentation & Best Poster Award)
- Jie Tang, Bo Wang, Yang Yang, Po Hu, Yanting Zhao, Xinyu Yan, Bo Gao, Minlie Huang, Peng Xu, Weichang Li, and Adam K. Usadi. PatentMiner: Topic-driven Patent Analysis and Mining. **KDD'12**, pages 1366-1374.
- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogeneous Networks. **WSDM'12**, pages 743-752.
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**, pages 203-212.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. **KDD'08**, pages 990-998.

Thank you！

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

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

Tiancheng Lou (**Google**)

Jimeng Sun (**IBM**)

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

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

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