Colorful Social Networks: inferring social ties in large networks

Jie Tang
Tsinghua University, China

Collaborate with
John Hopcroft , Jon Kleinberg (*Cornell*)
Jiawei Han, Chi Wang (*UIUC*)
Tiancheng Lou, Wenbin Tang, Honglei Zhuang (*Tsinghua*)

New Born Baby

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

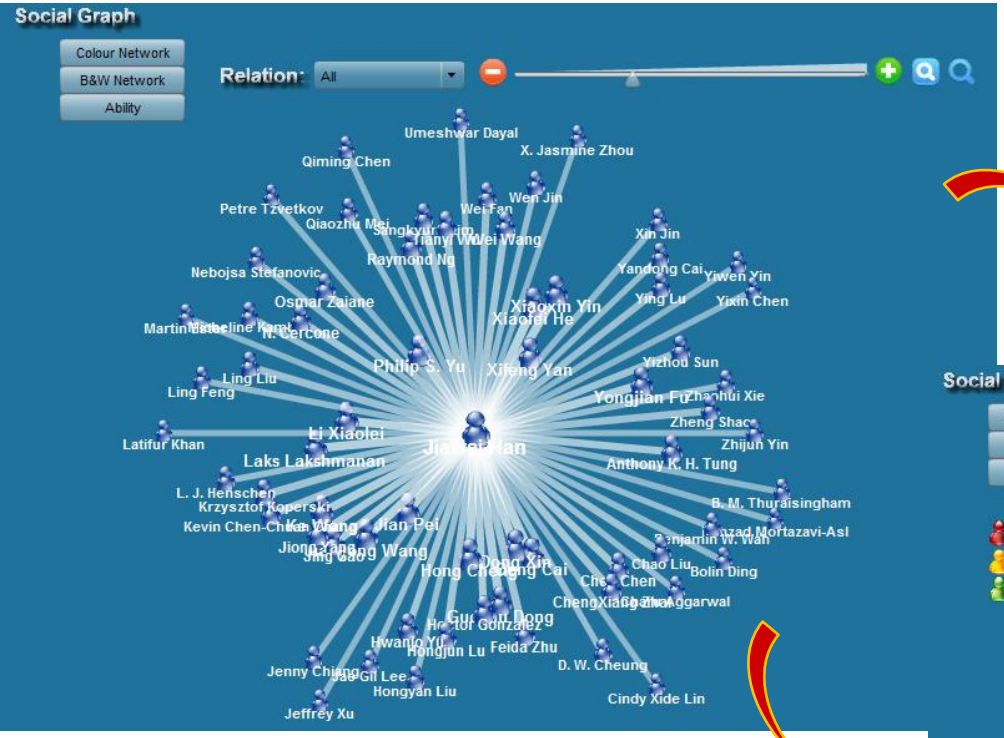
Even complex than we imaged!

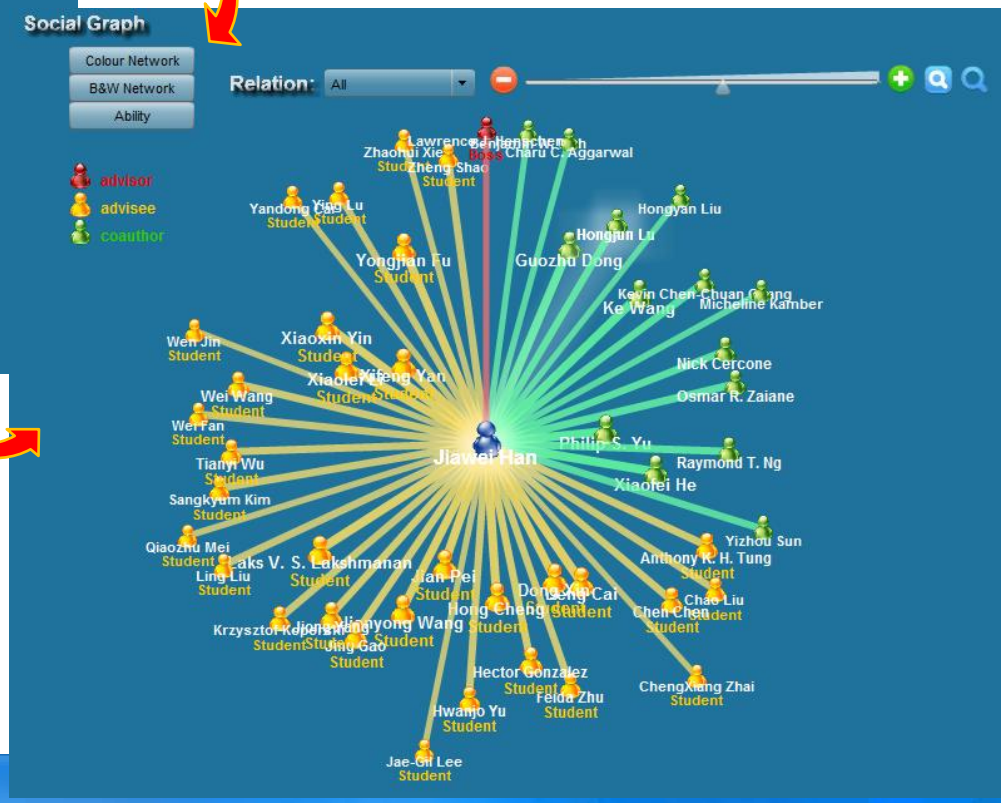
- Only 16% of mobile phone users in Europe have created custom contact groups
  - *users do not* take the time to create it
  - *users do not* know how to circle their friends

- 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)

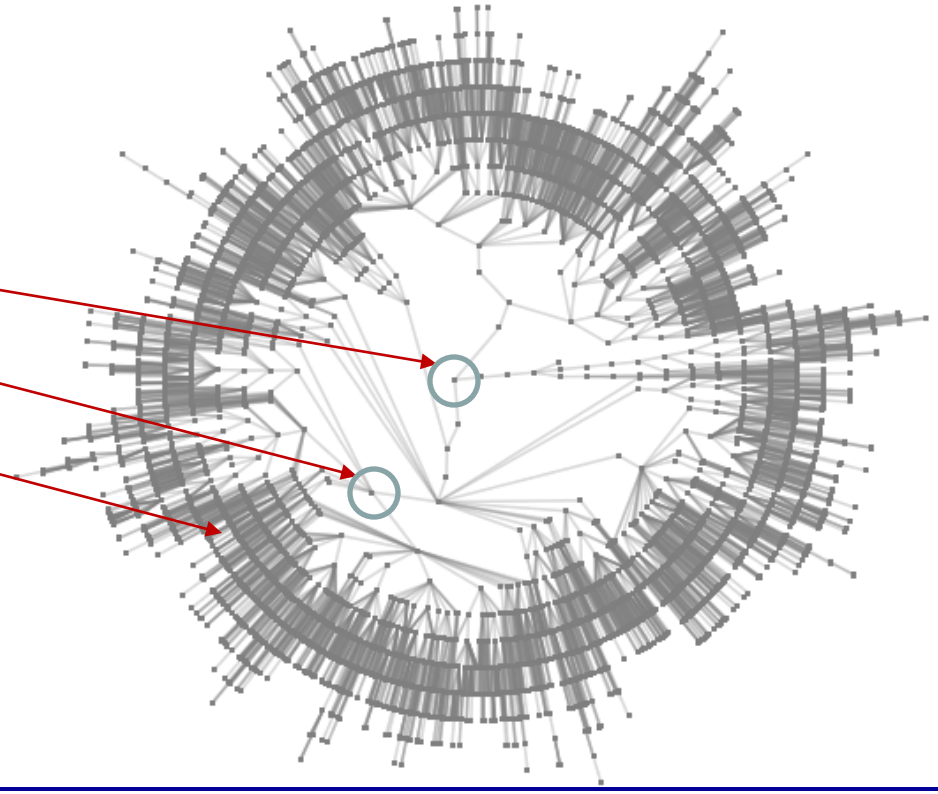
Enterprise email network

How to infer

CEO

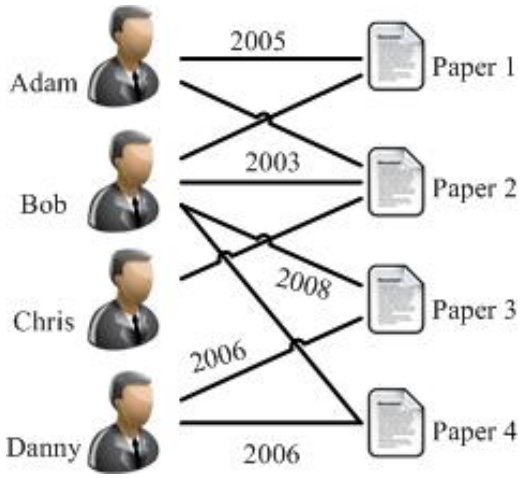
Manager

Employee

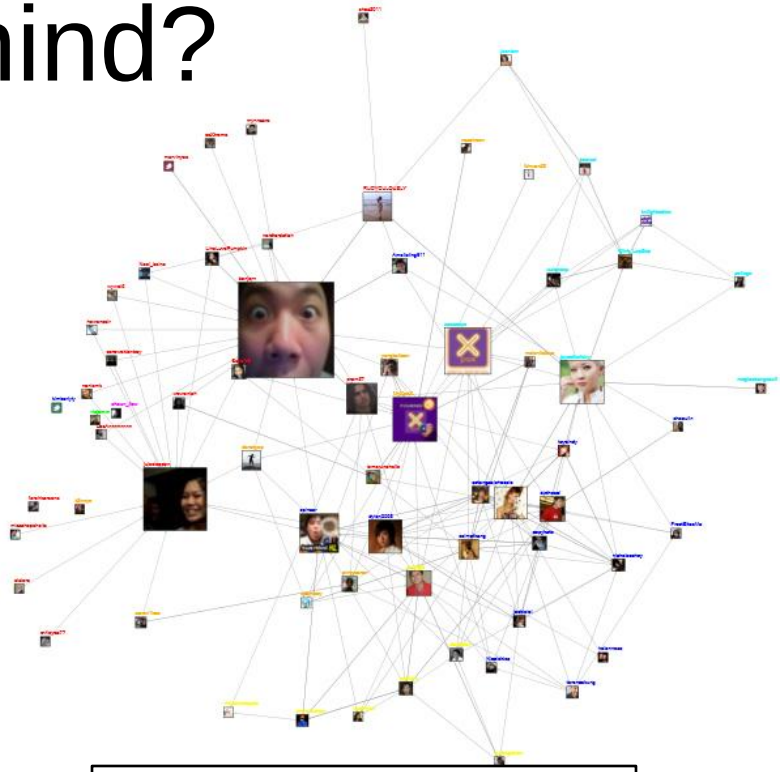


User interactions may form *implicit groups*

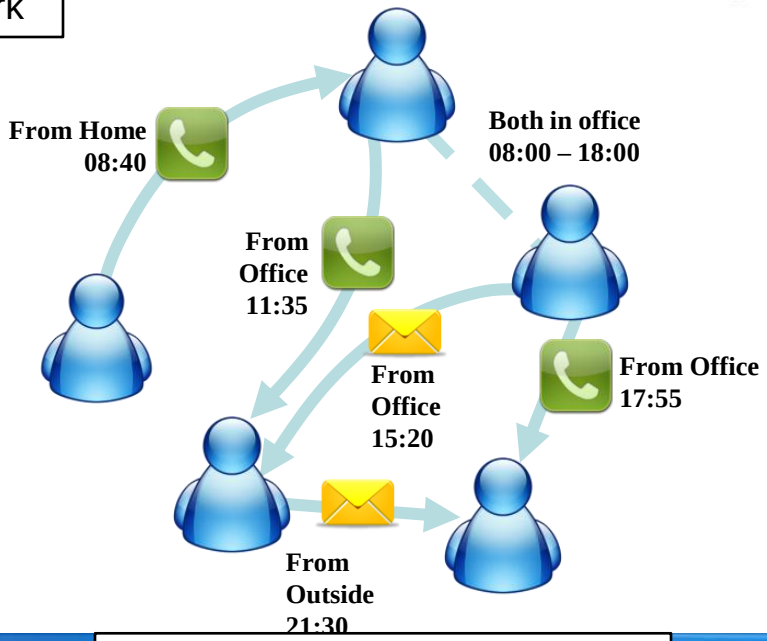
What is behind?



Publication network

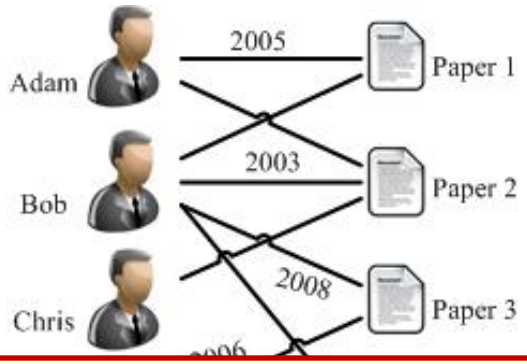


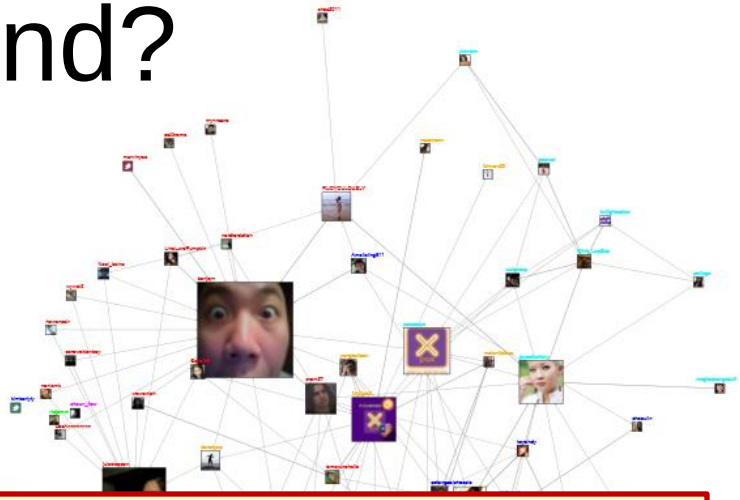
Twitter's following network



Mobile communication network

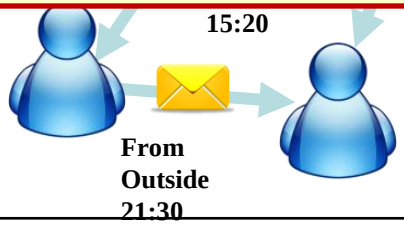
What is behind?





Questions:

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



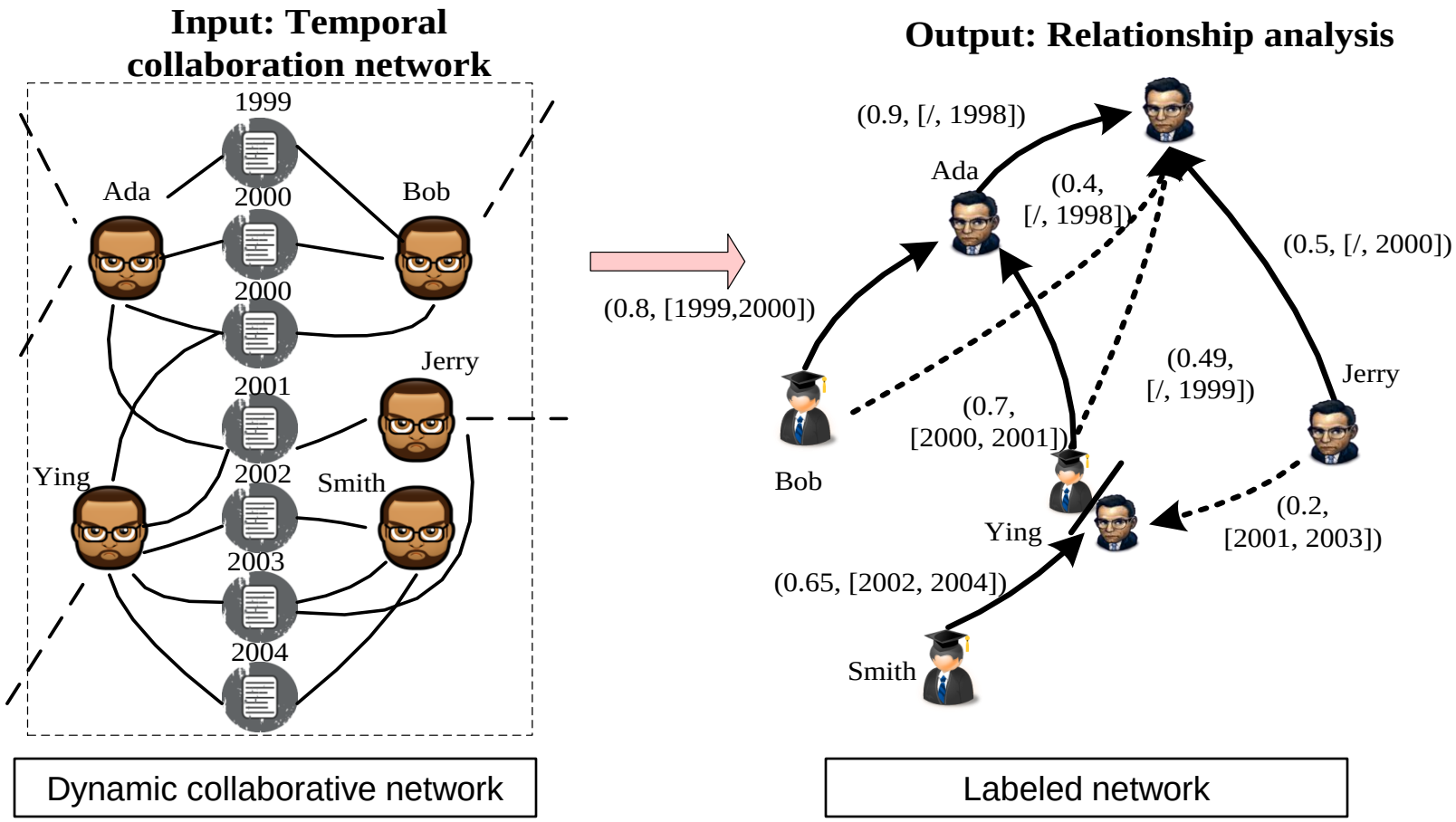
Mobile communication network

Learning Framework for Inferring Social Ties

KDD 2010, CIKM 2011,

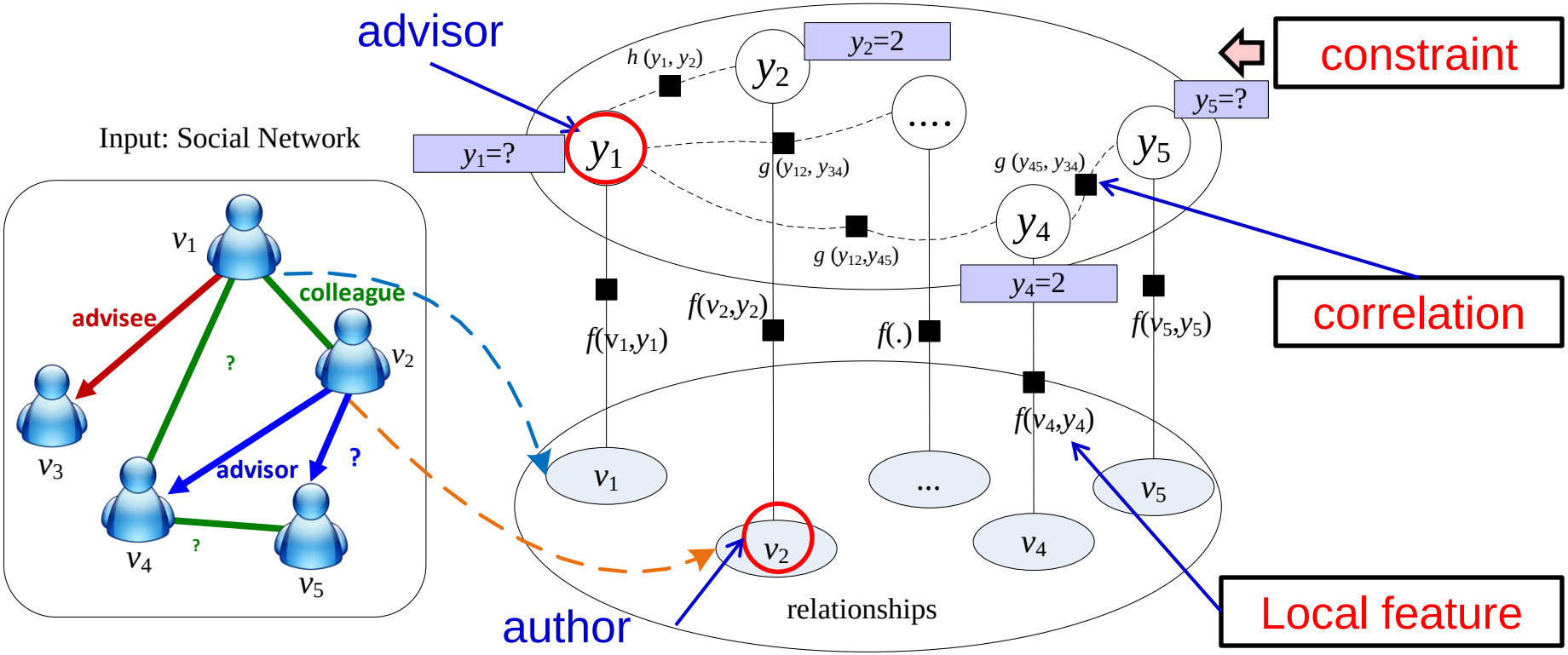
PKDD 2011 (Best Paper Runnerup)

Problem Analysis



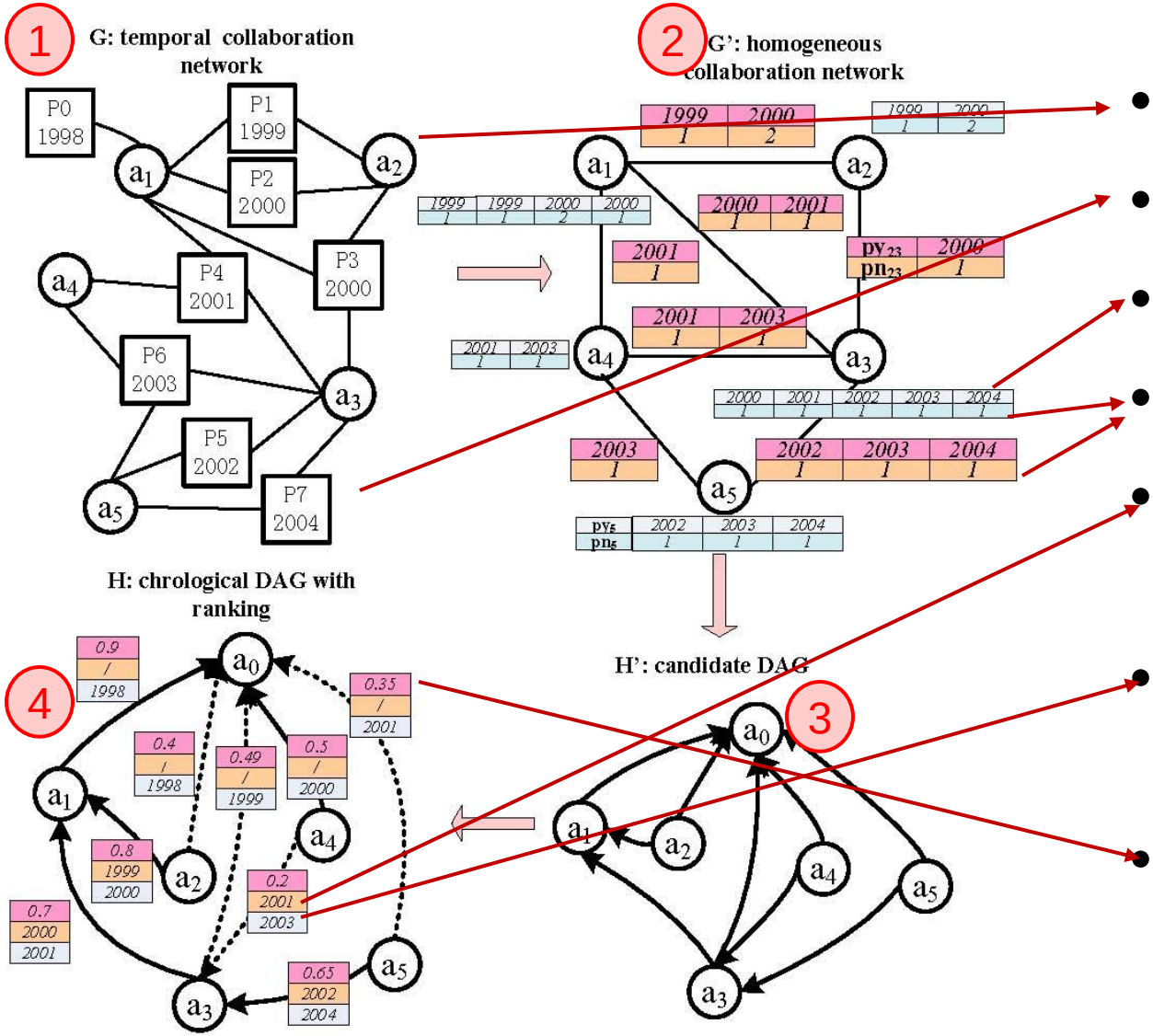
Output: potential types of relationships and their probabilities:
(type, prob, [s_time, e_time])

Factor Graph Model 1



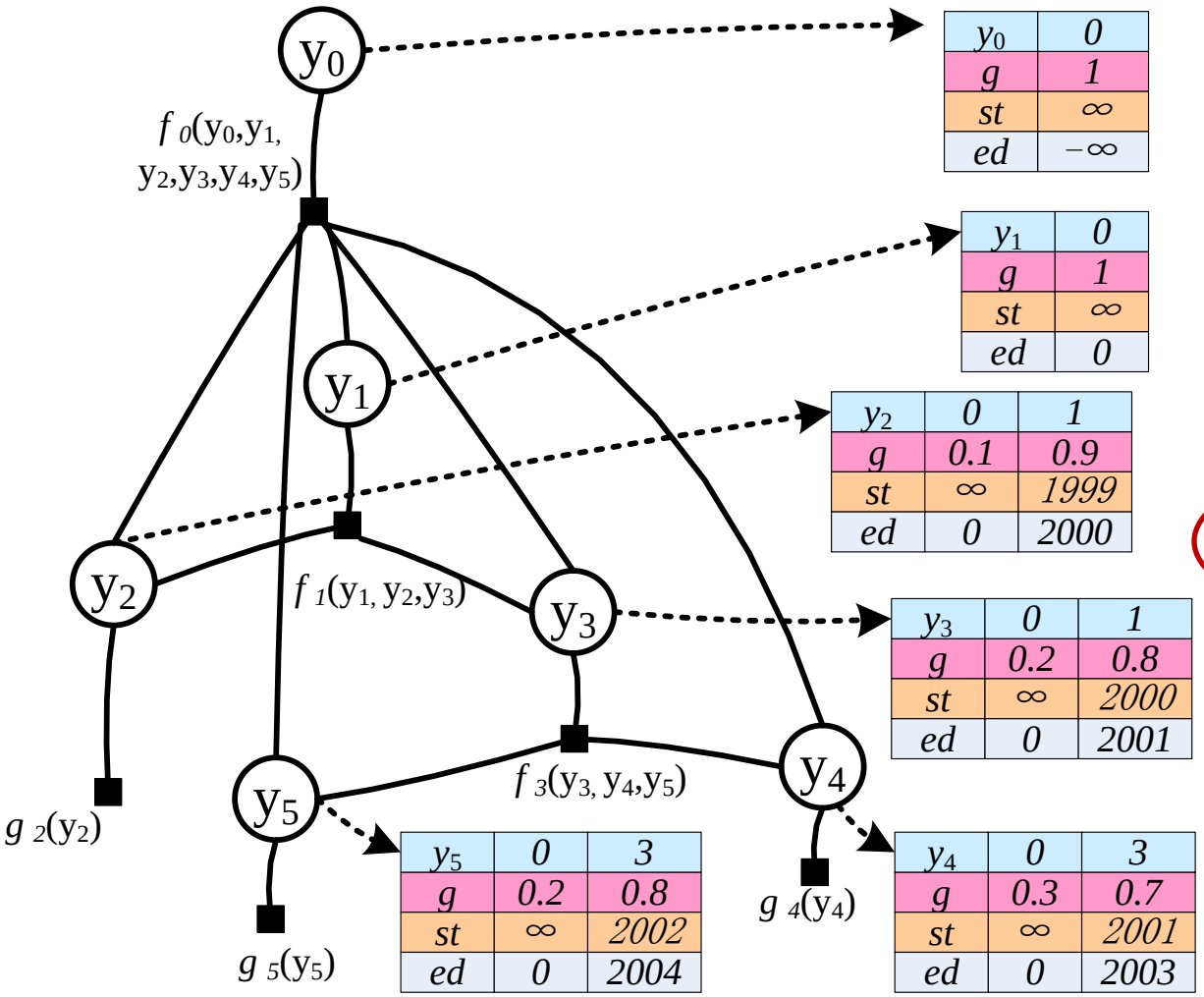
The problem is cast as, for each node, identifying which neighbor has the **highest probability** to be his/her advisor, i.e., $P(y_i=j \mid x_i, x_{\sim i}, \mathbf{y})$, where x_j and x_i are neighbors.

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

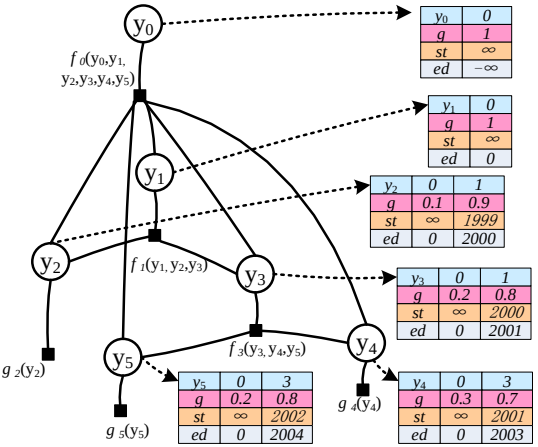
Time-constrained Probabilistic Factor Graph (TPFG)



- 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



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

with

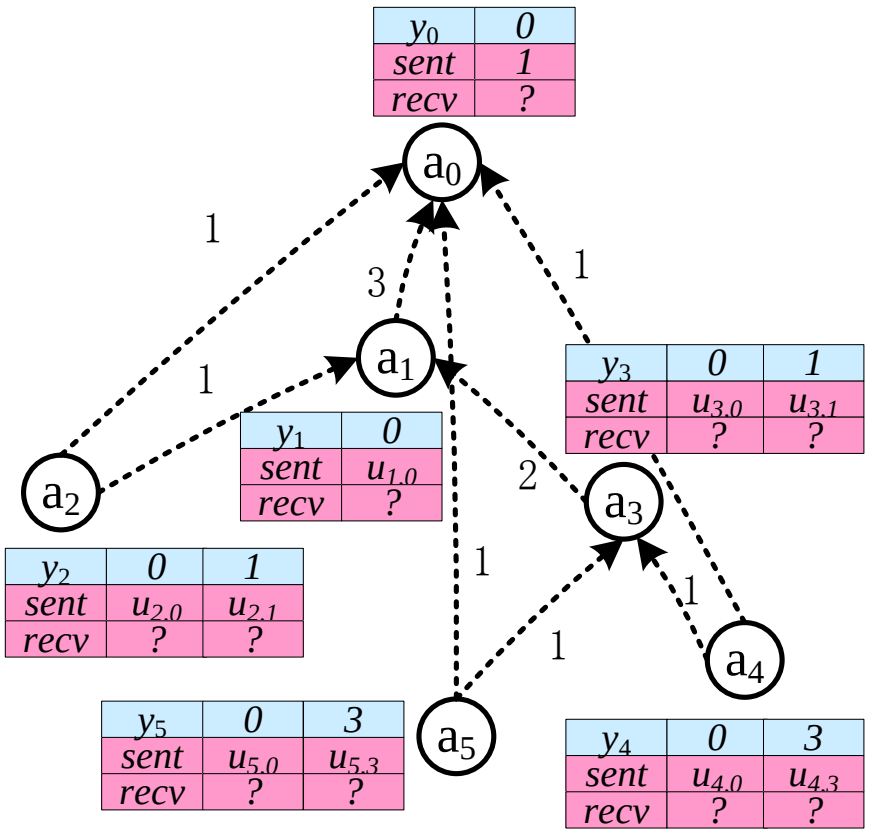
$$f_i(y_i \mid \{y_x \mid x\in Y_i^{-1}\}) = g(y_i, st_{ij}, ed_{ij})\prod_{x\in Y_i^{-1}} \phi(y_x, ed_{ij}, st_{xi})$$

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

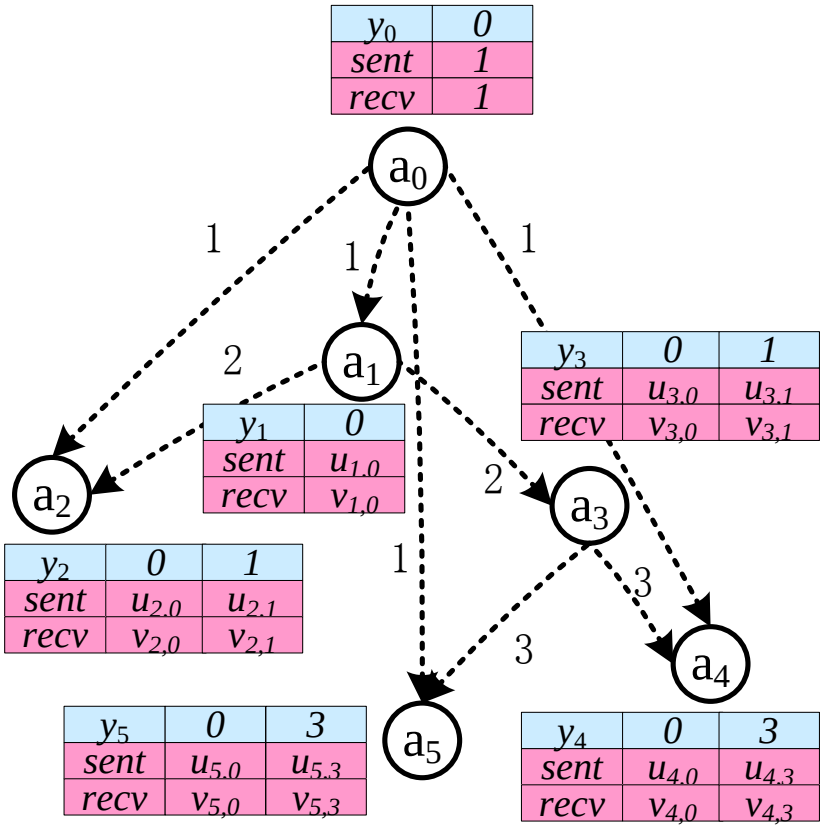
$$\phi(y_x, ed_{ij}, st_{xi}) = \begin{cases} 1, & y_x \neq i \vee ed_{ij} < st_{xi} \\ 0, & y_x = i \wedge ed_{ij} >= st_{xi} \end{cases}$$

Inference algorithm of TPFG

- $r_{ij} = \max P(y_1, \ldots, y_{na} | y_i = j) = \exp(sent_{ij} + recv_{ij})$



Phase 1



Phase 2

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%**

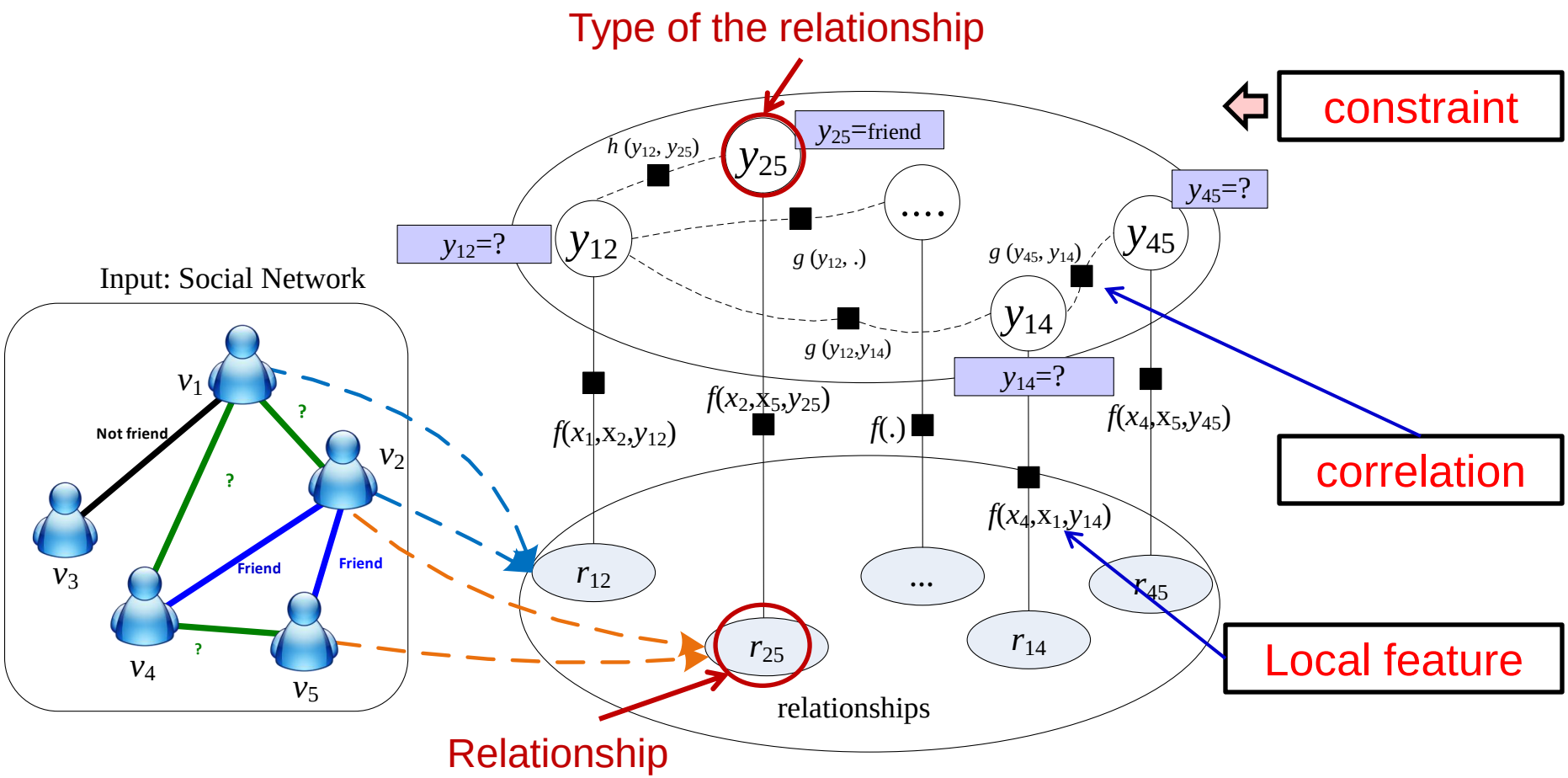
heuristics

Supervised learning

Empirical parameter

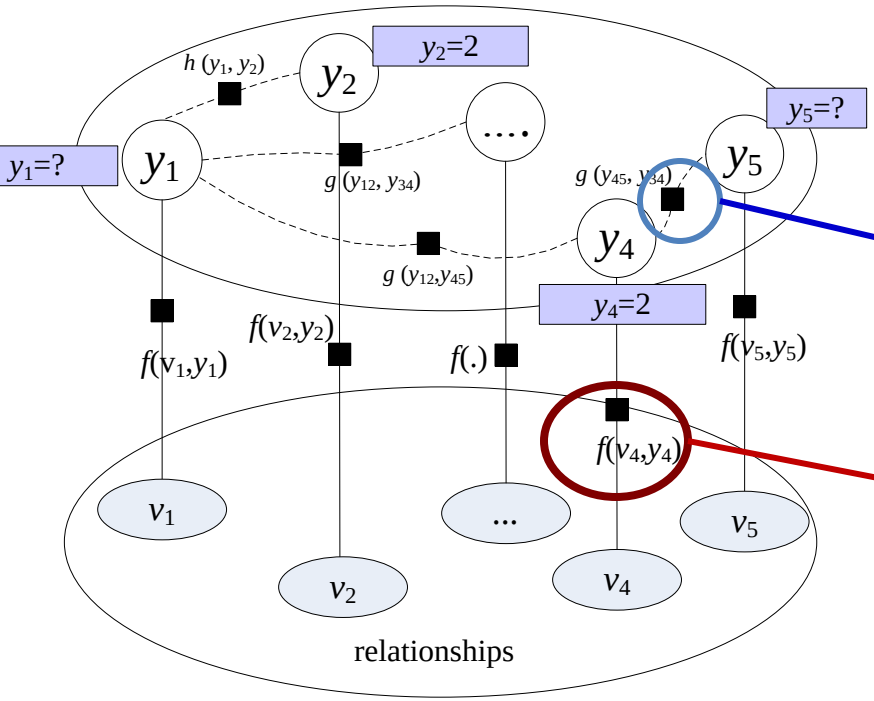
optimized parameter

Factor Graph Model 2



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

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

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

Learning Algorithm

- Maximize the log-likelihood of labeled relationships

$$O(\theta) = \sum_{i=1}^{|E|} \sum_{j=1}^{d} \alpha_j h_k(Y_{c_i}) + \sum_{c} \sum_{k} \mu_k h_k(Y_c) - \log Z$$

Input: learning rate η
Output: learned parameters θ

Initialize θ;
repeat
  Calculate $\mathbb{E}_{p_\theta(Y|Y^L,G)}\mathbf{S}$ using LBP ; → **Inferring unlabeled data**
  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

Variant gradient decent method

Results of Model 2

- **Email.** An email network comprised of 136,329 emails between 151 Enron employees.
  - To infer the manager-subordinate relationship
- **Mobile.** A communication network consisting of calling, bluetooth and location logs of 107 mobile users for 10 months
  - To infer friendships from the communication network
- **Twitter.** A following network including 112,044 Twitter users and about 500,000 following links among them for 2 months
  - To infer reciprocal relationships between users

Results of Model 2

Datasets	Method	Precision	Recall	F1-measure
Email	SVM	79.1	**88.6**	83.6
	Model 2	**88.6**	87.2	**87.9**
Mobile	SVM	**92.7**	64.9	76.4
	Model 2	89.4	**75.2**	**81.6**
Twitter	SVM	69.1	**61.0**	65.0
	Model 2	**97.0**	55.0	**70.0**

- Comparison methods:
  - SVM: with the same local features
  - Model 2: the proposed model with all features

Questions?

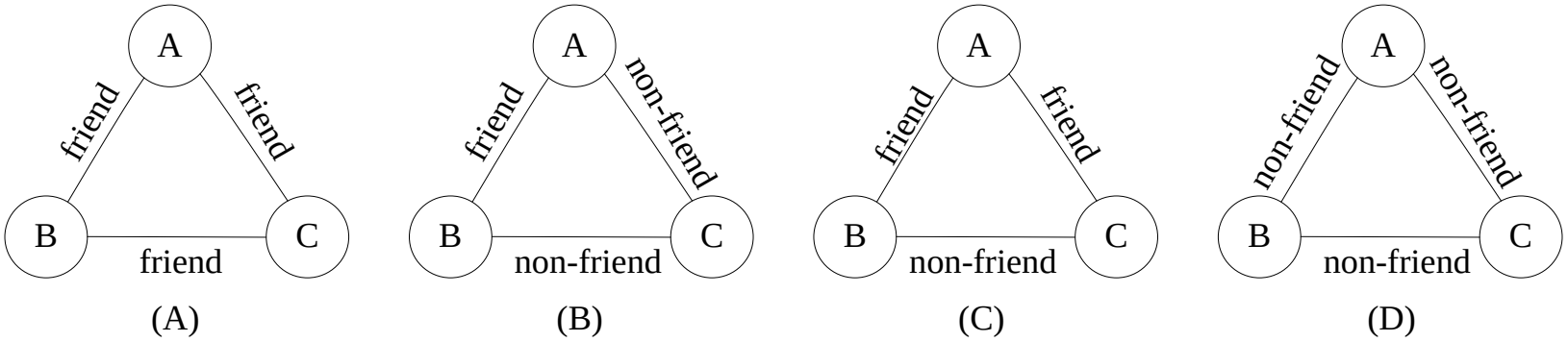
- Machine learning vs. social theories
- How to generalize the model to other networks?

Generalized framework by connecting to social theories

WSDM 2012

Social Theories

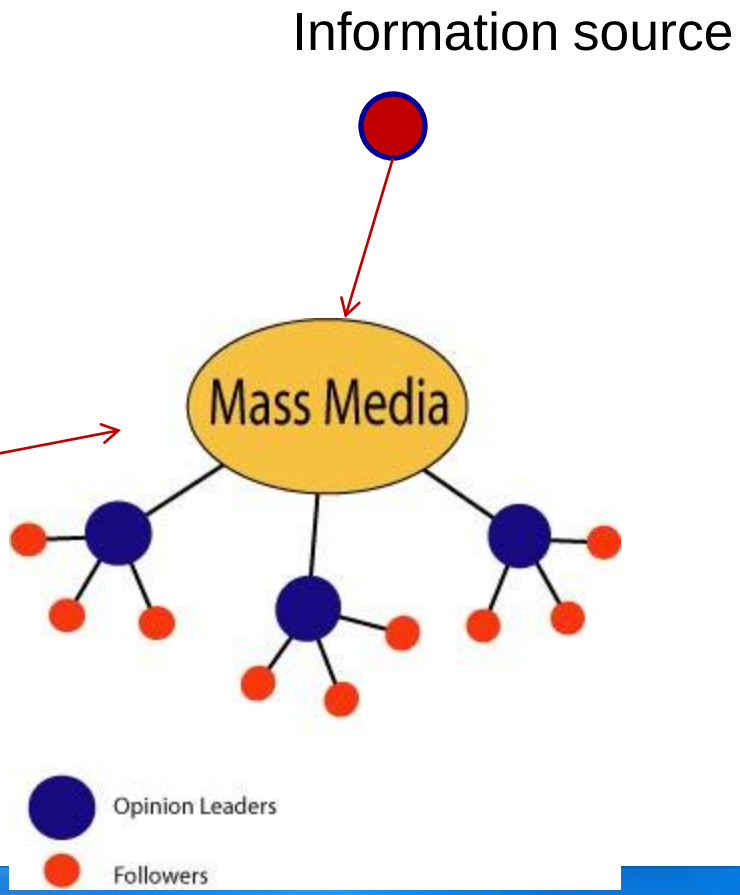
- Social balance theory
- Two-step flow theory
- Social status theory
- Structural hole theory



Social Theories—Two-step flow

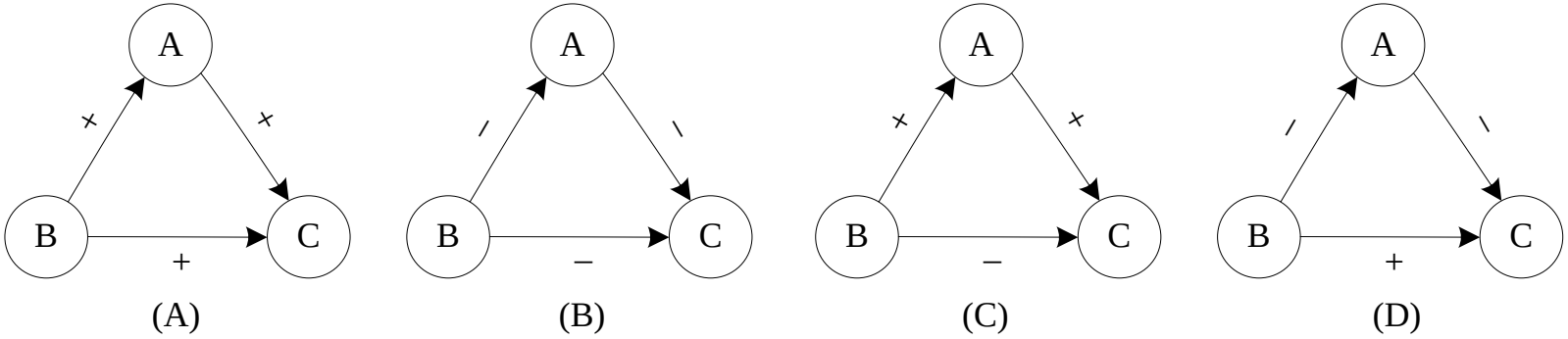
- Social balance theory
- Two-step flow theory
- Social status theory
- Structural hole theory

Two-step information propagation



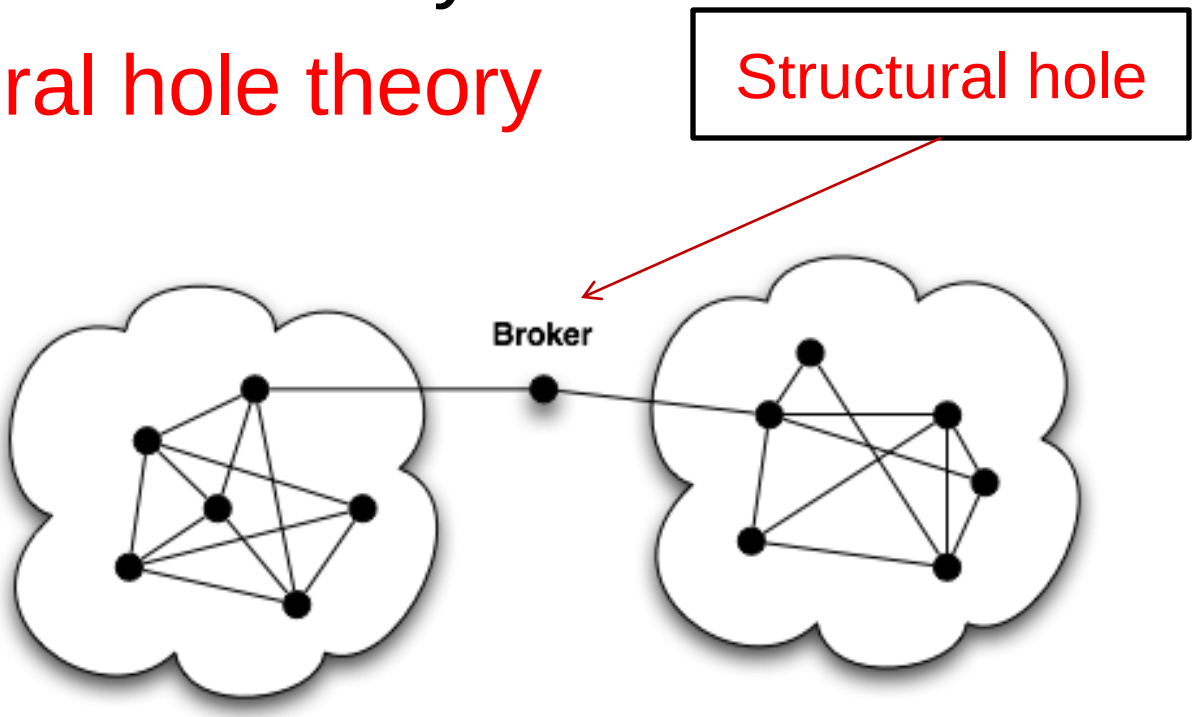
Social Theories—Social status

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



Social Theories—Structural hole

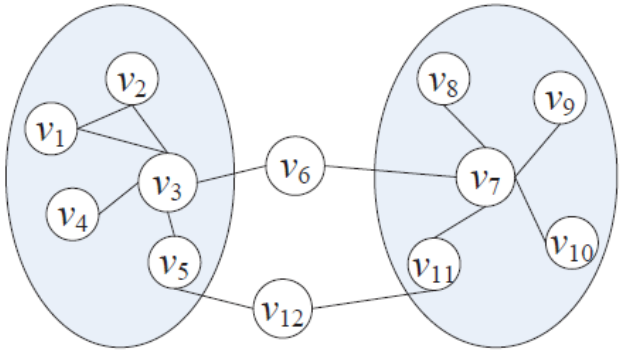
- Social balance theory
- Two-step flow theory
- Social status theory
- Structural hole theory

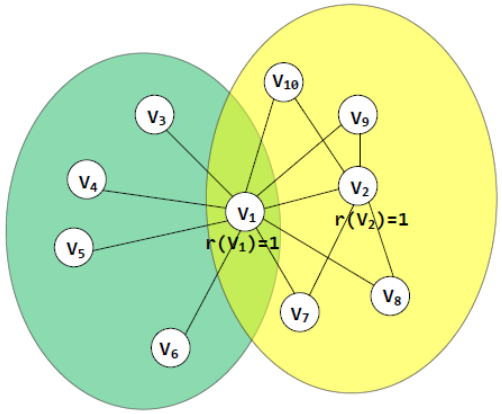


Formal definition for structural hole

Definition 1. **Structural Hole.** Let $G = (V, E)$ denote a social network, where $V = \{v_1, v_2, \cdots, v_n\}$ is a set of n users, and $E \subseteq V \times V$ is a set of undirected social relationships between users. Further assume that the vertices of the social network can be grouped into l subsets $\mathbf{C} = \{C_1, \cdots, C_l\}$ (called community). Then structural holes of the social network can be defined as a subset of k vertices in the network, which can maximize a quality function:

$$\max_{V_{SH}} Q(V_{SH}, \mathbf{C}), \text{with} \quad |V_{SH}| = k \tag{1}$$





Algorithm

$$I(v,i) = \max_{(u,v)\in E, i\in S} \{I(v,i), \alpha_i \cdot I(u,i) + \beta_S \cdot H(u,S)\}$$

$$H(v,S) = \min_{i\in S}\{I(v,i)\}$$

Theorem 1. Given α_i and β_S, the functions $I(v,i)$ and $H(v,S)$ exists for any graph $G=(V,E)$, if and only if,

$$\max_i\{\alpha_i + \max_{i\in S}\{\beta_S\}\} \leq 1$$

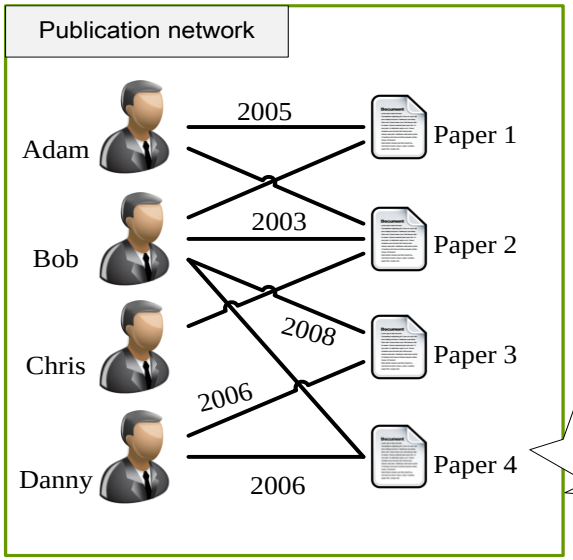
PROOF. On one hand, suppose there exists i and $i \in S$ such that $\alpha_i + \beta_S > 1$. Let $S = \{i, p_1, \cdots, p_{|S|-1}\}$. Consider the following graph $G=(V,E)$, there are two vertices v_1 and v_2 connected by one edge (v_1, v_2), and v_1 appears in communities C_i, $C_{p_1}, \cdots C_{p_{|S|-1}}$, and v_2 appears in communities C_i, and we have $r(v_1) = r(v_2) = 1$. Figure 2 shows the case for $|S| = 2$.

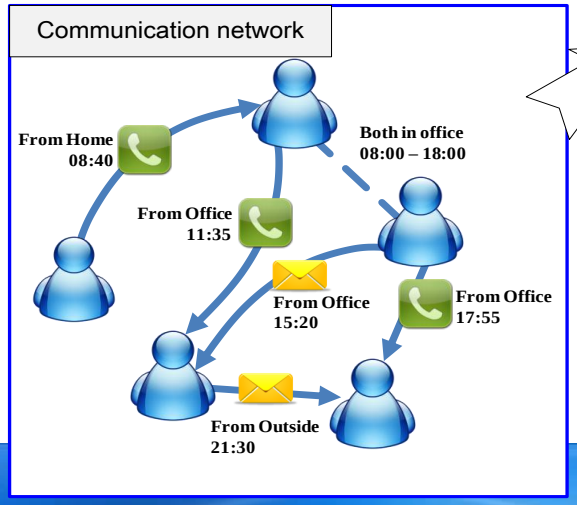
Thus, $I(v_1,i) = I(v_1,p_1) = \cdots = I(v_1,p_{|S|-1}) = r(v_1) = 1$. According to rule 3, we get $H(v_1,S) = 1$. According to rule 2, we have $I(v_2,i) \geq \alpha_i \cdot I(v_1,i) + \beta_S \cdot H(v_1,S) = \alpha_i + \beta_S > 1$, which is impossible.

On the other hand, if for each i and $i \in S$, we have $\alpha_i + \beta_S \leq 1$. We use induction to prove that, after infinite number of iterations, it also satisfies that $I(v,i) \leq 1$. In the first iteration, we have $I^{(0)}(v,i) \leq r(v) \leq 1$. Suppose after the k-th iteration, we have $I^{(k)}(v,i) \leq r(v) \leq 1$. Hence, in the $(k+1)$-th iteration, for each $i \in S$, we have $I^{(k+1)}(v,i) \leq \alpha_i \cdot I^{(k)}(u,i) + \beta_S \cdot H^{(k)}(u,S) \leq (\alpha_i + \beta_S) I^{(k)}(u,i) \leq I^{(k)}(u,i) \leq 1$. $\square$

Generalization—Redefining the problem across heterogeneous networks

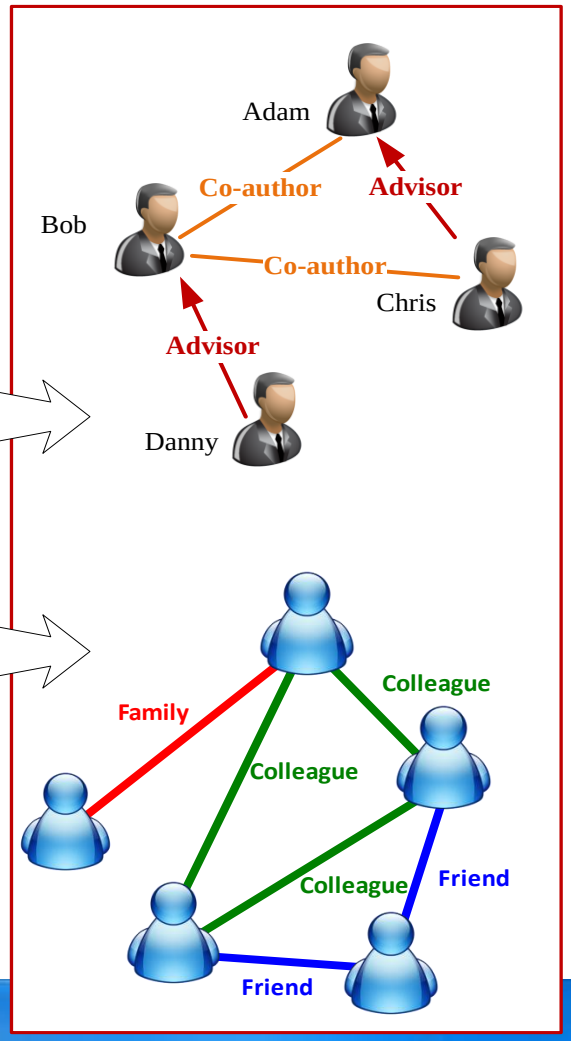
Input: Heterogeneous Networks





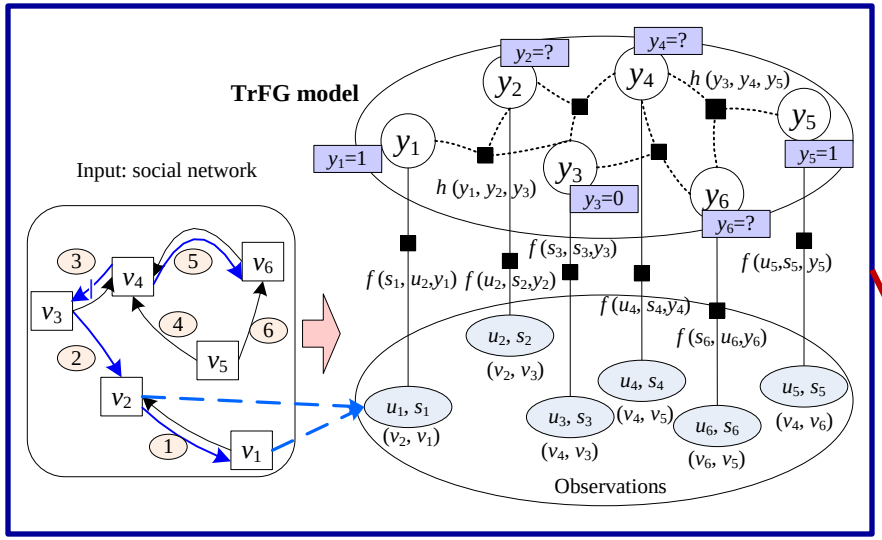
Knowledge Transfer for Inferring Social Ties

Output: Inferred social ties in different networks

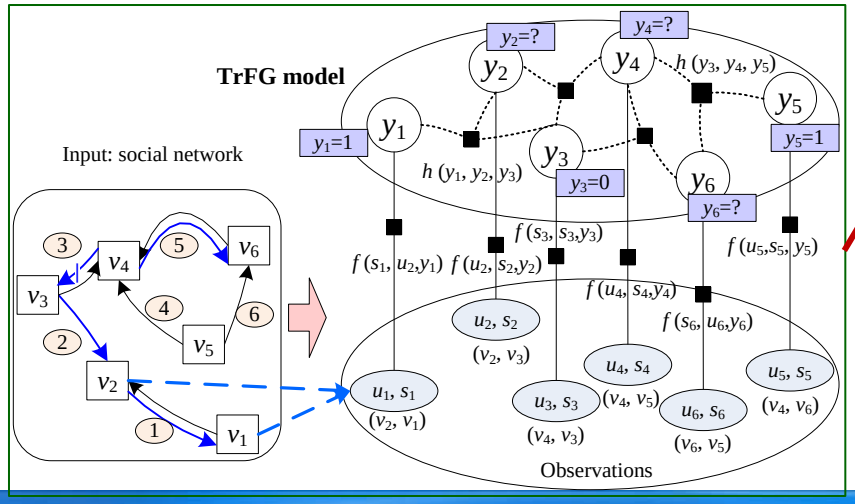


Transfer Factor Graph Model

Coauthor network



mobile



Bridge via social theories

Mathematical Formulation

Features defined in different networks

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

Triad-based features shared across networks

New Results

- **Epinions** a network of product reviewers: 131,828 nodes (users) and 841,372 edges
  - to infer the trust relationships between users
- **Slashdot**: 82,144 users and 59,202 edges
  - to infer “friend” relationships between users
- **Mobile**: 107 mobile users and 5,436 edges
  - to infer friendships between users
- **Coauthor**: 1,036,990 authors and 1,900,260 coauthorships
  - to infer advisor-advisee relationships
- **Enron**: 136,329 emails between 151 Enron employees
  - to infer manager-employee relationships

New Results

Datasets	Method	Precision	Recall	F1-measure
Epinions + Slashdot	SVM	78.69	**99.65**	87.94
	Model 2	93.00	64.36	76.07
	Triad Factor Graph	**94.14**	94.46	**94.30**
Epinions + Mobile	SVM	90.32	58.33	70.88
	Model 2	**100.00**	59.24	74.40
	Triad Factor Graph	82.39	**83.44**	**82.91**
Slashdot + Mobile	SVM	90.32	58.33	70.88
	Model 2	**100.00**	59.24	74.40
	Triad Factor Graph	72.58	**85.99**	**78.72**
Coauthor + Enron	SVM	60.23	**86.89**	71.14
	Model 1	**97.30**	65.45	78.26
	Triad Factor Graph	95.56	78.18	**86.00**

Two Concluding Remarks

- Connecting social theories to machine learning models can significantly improve the performance of inferring social ties

- Combining heterogeneous networks into one unified model can help social relationship analysis

ArnetMiner.org

- Academic research social network analysis and mining system

提供全面的研究者网络分析与挖掘功能

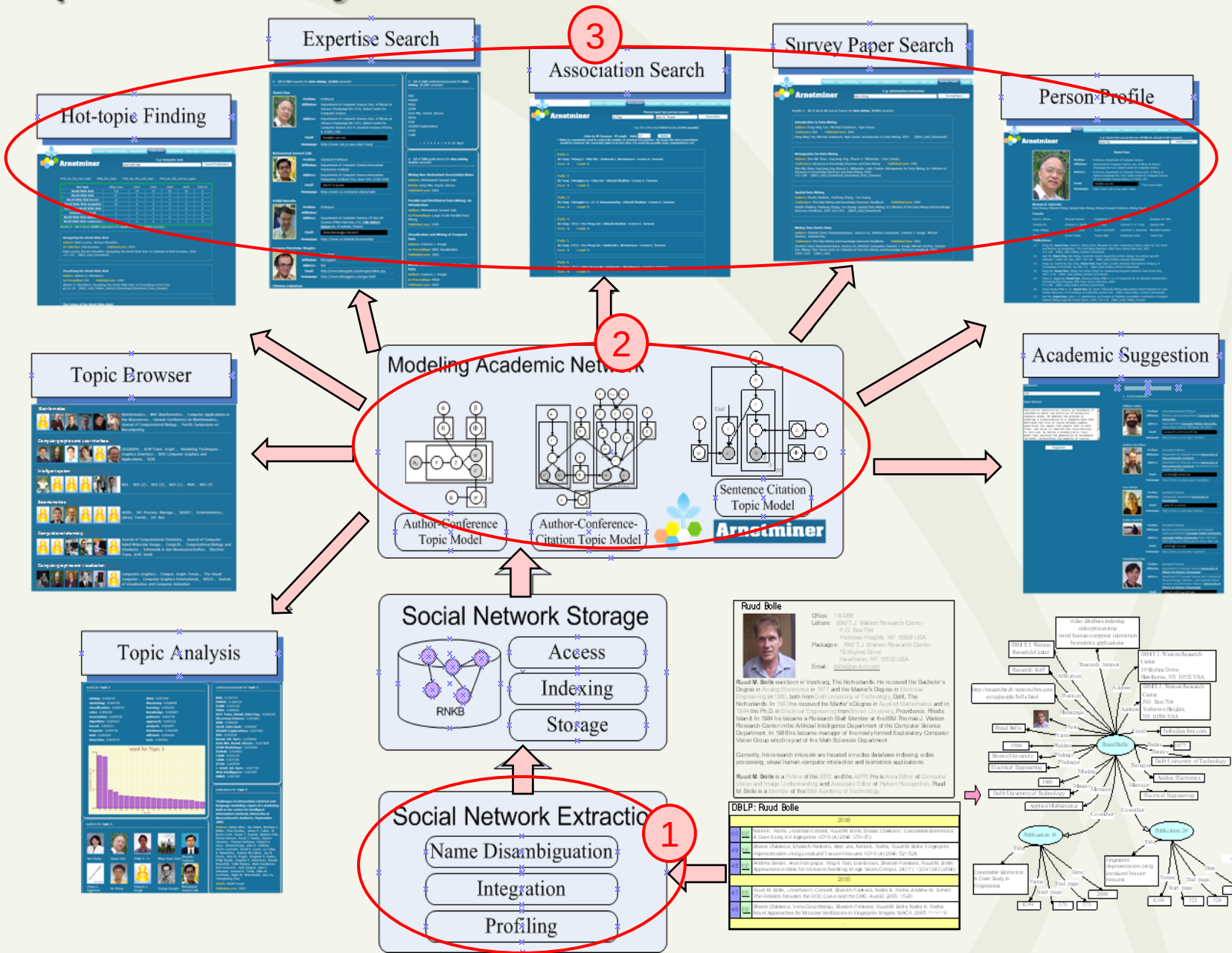
Papers published: ACM TKDD, MLJ, DKE, JIS; KDD’08-10, SDM’09, ICDM’07-10, CIKM’07-10

http://arnetminer.org/

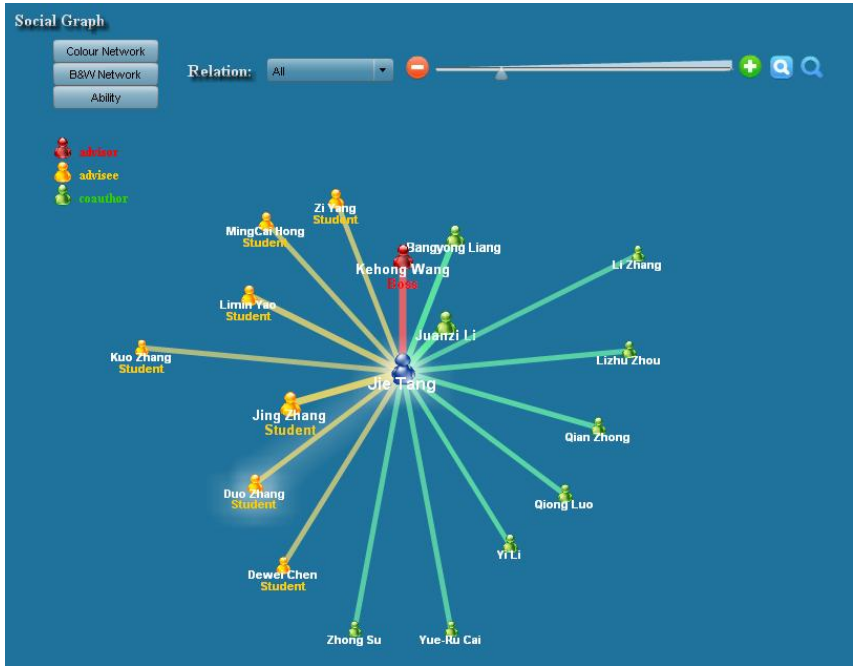
Arnetminer

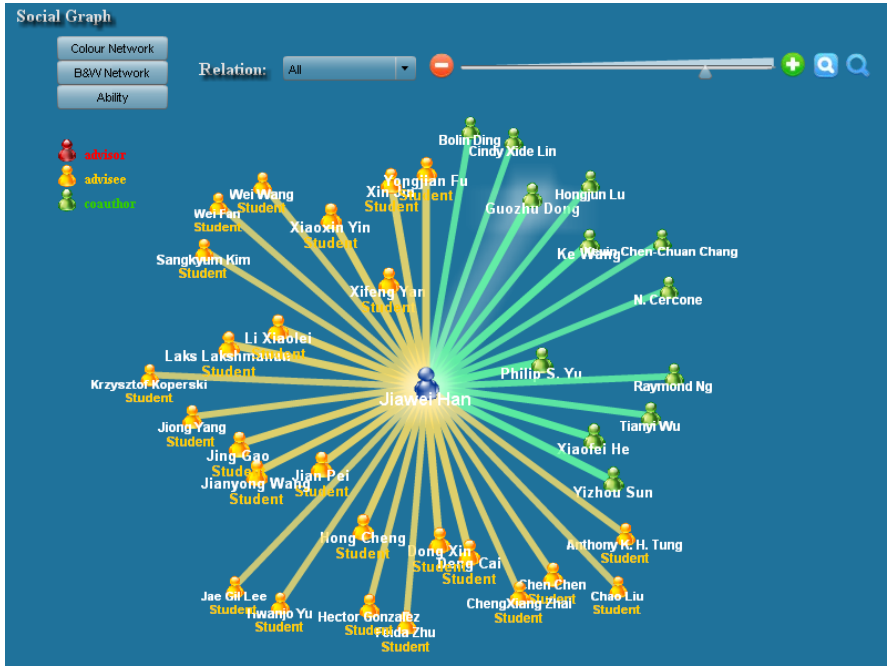
Http://www.arnetminer.org

Jie Tang, Jing Zhang, Feng Wang, Limin Yao, Duo Zhang, Liu Liu, and Juanzi Li
Knowledge Engineering Group, DCST, Tsinghua University
{tangjie, zhangjing, wangfeng, ylm, zhangduo, ljz}@keg.cs.tsinghua.edu.cn

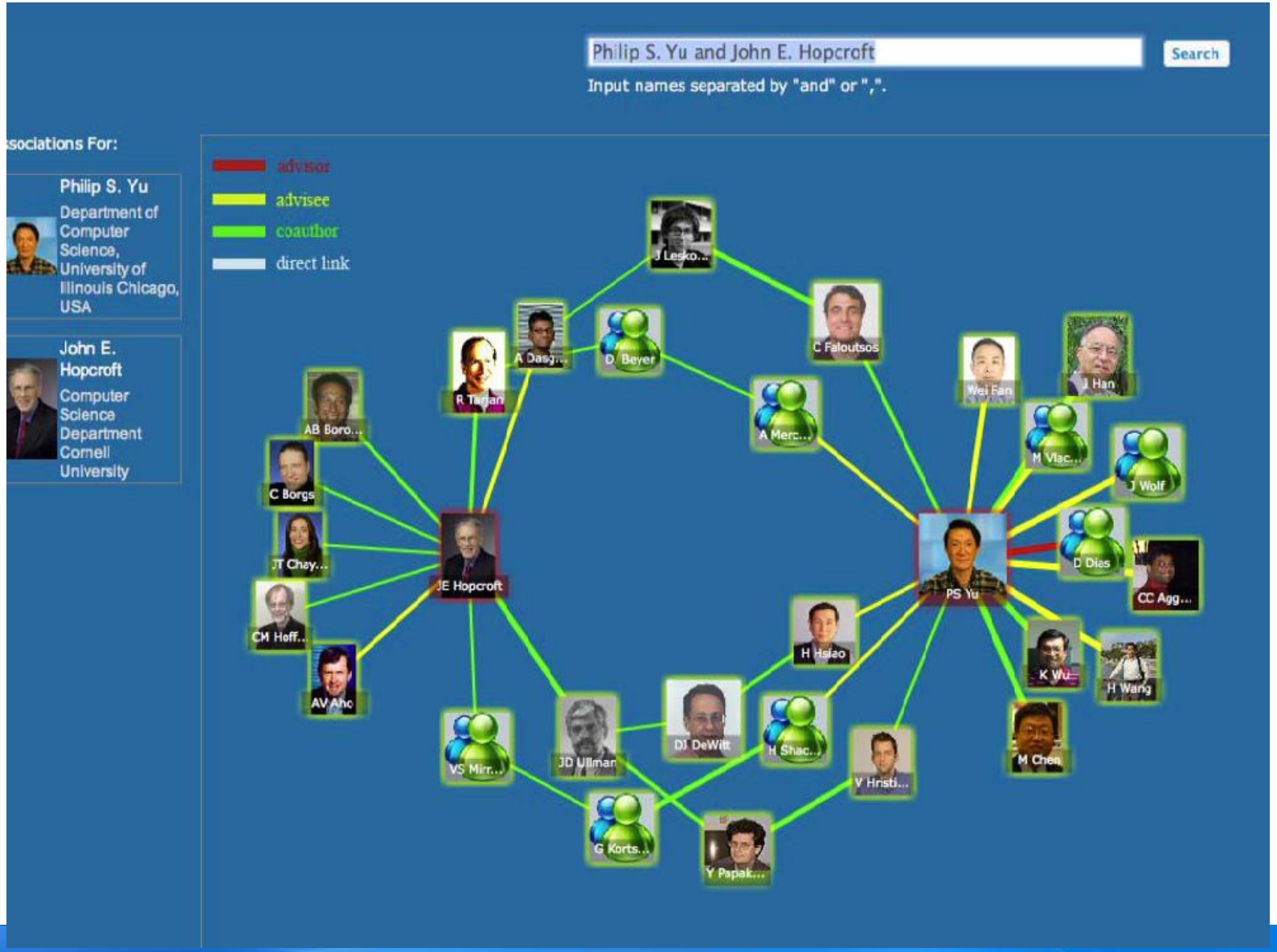


Colorful Social Graph In ArnetMiner





Instant Social Graph Search

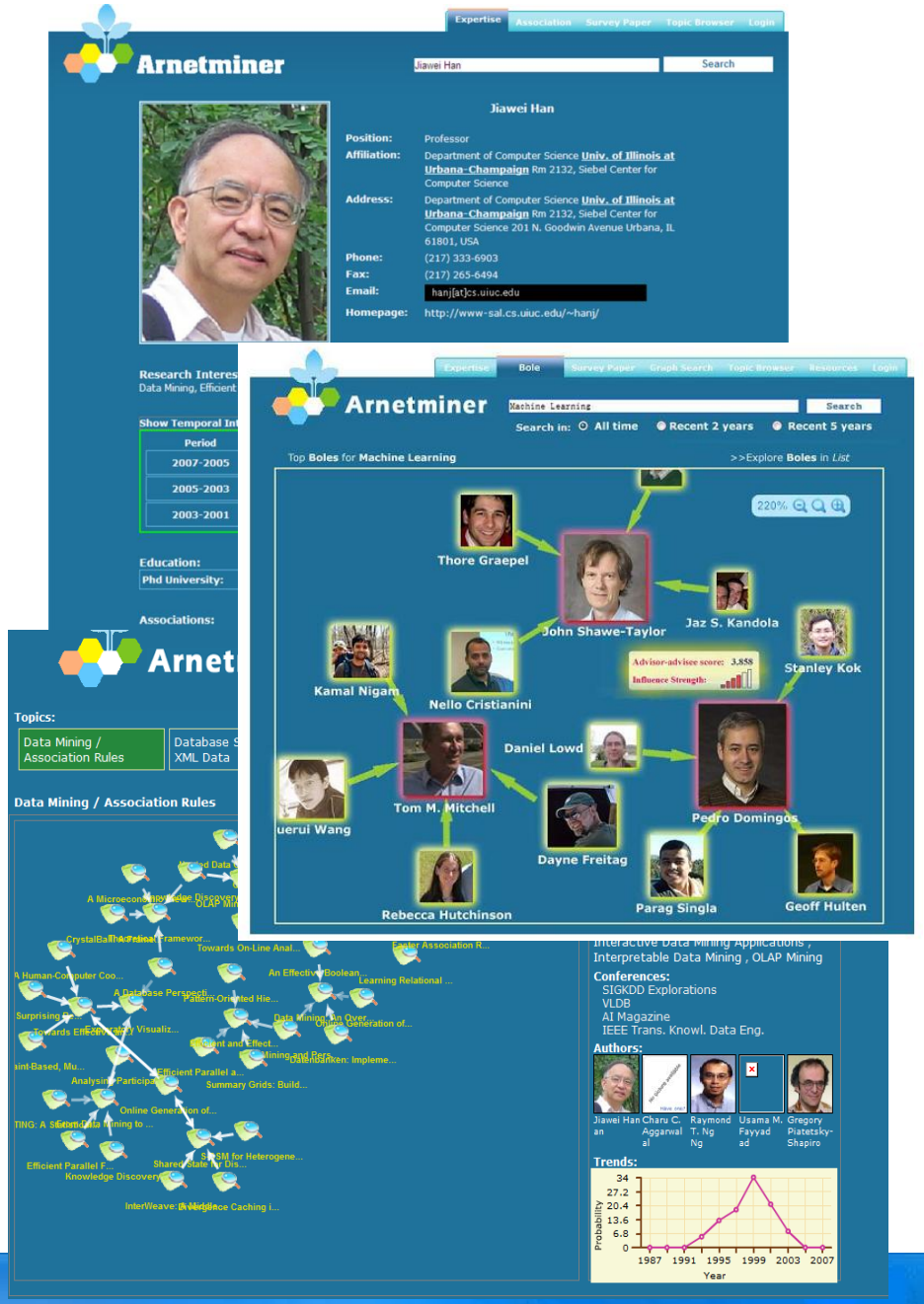


ArnetMiner Today

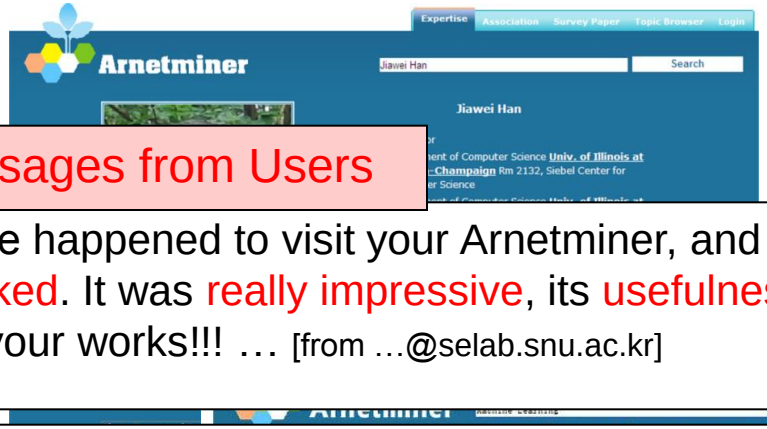
* Arnetminer data:
  > 1.5 M researcher profiles
  > 3M papers
  > 34M citation relationships
  > 8K conferences
  > 80M logs

* > **1,000,000 users** from more than **200** countries
* Continuously **20+%** increase of visits per month

* **>300,000 page views** per day



ArnetMiner Today



* Arnetminer data:
  > 1.5 M researcher profiles
  > 3M papers
  > 34M citation relationships
  > 8K conferences
  > 80M logs

Messages from Users

… I've happened to visit your Arnetminer, and shocked. It was really impressive, its usefulness and your works!!! … [from …@selab.snu.ac.kr]

…I would first of all congratulate you on the excellent work you have done in Arnetminer and I am much inspired… [from …@nu.edu.pk]

a survey by UK Southampton

Title: Semantic Technologies for Learning and Teaching in Web 2.0.

— Thanassis Tiropanis, Hugh Davis, Dave Millard, Mark Weal

…Exposing the expertise of the institution to the outside world in order to attract funding and students. ArnetMiner is the most representative example of such tools at the moment…

Contextualised queries and searches, searches across repositories potentially in different departments or institutions, and matching of people for collaborative activities. Best example of the surveyed technologies to this end is ArnetMiner.

Arnetminer's User Distribution

Top 10 countries

1. USA
2. China
3. Germany
4. India
5. UK
6. Canada
7. Japan
8. Spain
9. France
10. Italy

Arnetminer is widely used

- **The largest publisher:**

  Elsevier
- **Top confs:**

  KDD 2010-12

  PKDD 2011

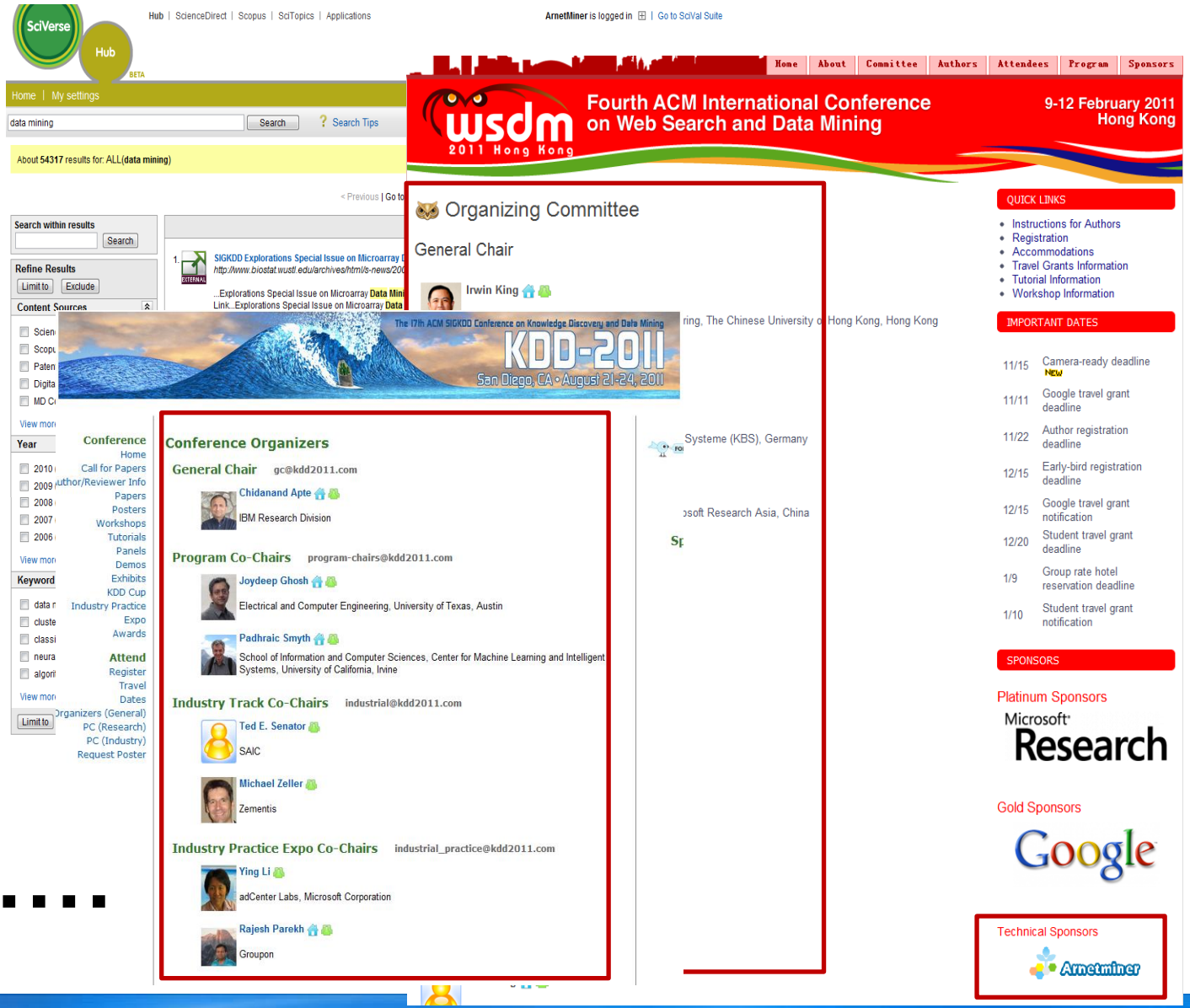
  WSDM 2011

  ICDM 2011

  ICMLA 2011

  ICTAI 2011

  etc.



……

Related Publications

- Jie Tang, Tiancheng Lou, and Jon Kleinberg. Inferring Social Ties across Heterogenous Networks. **WSDM'12**.
- Chi Wang, Jiawei Han, Yuntao Jia, Duo Zhang, Yintao Yu, Jie Tang, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**.
- Wenbin Tang, Honglei Zhuang, and Jie Tang. Learning to Infer Social Relationships in Large Networks. **PKDD'11.** **(Best Student Paper Runner-up)**
- John E. Hopcroft, Tiancheng Lou, and Jie Tang. Who Will Follow You Back? Reciprocal Relationship Prediction. **CIKM'2011.** (alphabetical author order)
- Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM **TKDD**, 2010.
- 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**, 2011.
- Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. User-level sentiment analysis incorporating social networks. **KDD'11**.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. **KDD'10**.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. **KDD'09**. (**Top 4** cited papers among KDD 2009's papers)
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. **KDD'08**.
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. **IJCAI'09**.

Others: **SIGMOD'09**, **ACL'07**, **SIGIR'11**, **ICDM'07-11**, **CIKM'07-11**, **SDM'09**, **ISWC'(06,09)**, **DKE**, **JIS**, etc.

Thanks!

Q&A

System: http://arnetminer.org

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

Basic Features/Constraints

- Features
  - First-paper-year-diff
  - Conference coverage
  - Coauthor ratio
  - Paper count
- Constraints
  - a_y can only advise a_x after he graduated
  - If a_y advises a_x since the year st_x, a_y must have a longer history of publication than a_x before st_x.

The model can incorporate other intuitions as factor functions.