Colorful Social Networks: Relationship Mining in Social Networks

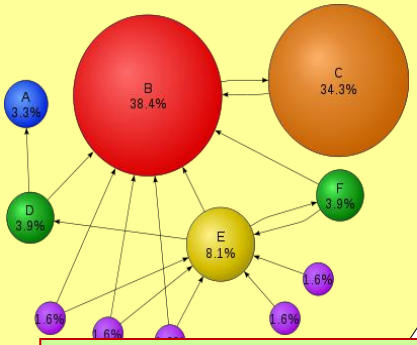
Jie Tang
Tsinghua University, China

Collaborate with
Chi Wang and Jiawei Han (UIUC), Jeffrey Xu Yu (CUHK)
Zhong Su, et al. (IBM)

Social Networks

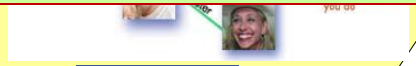
Web 1.0 (1989)

Pages, hyperlinks
Relevance search



Web 2.0 (2004)

social networks



facebook

Mobile Web (2008-20)

Connecting via mobiles…

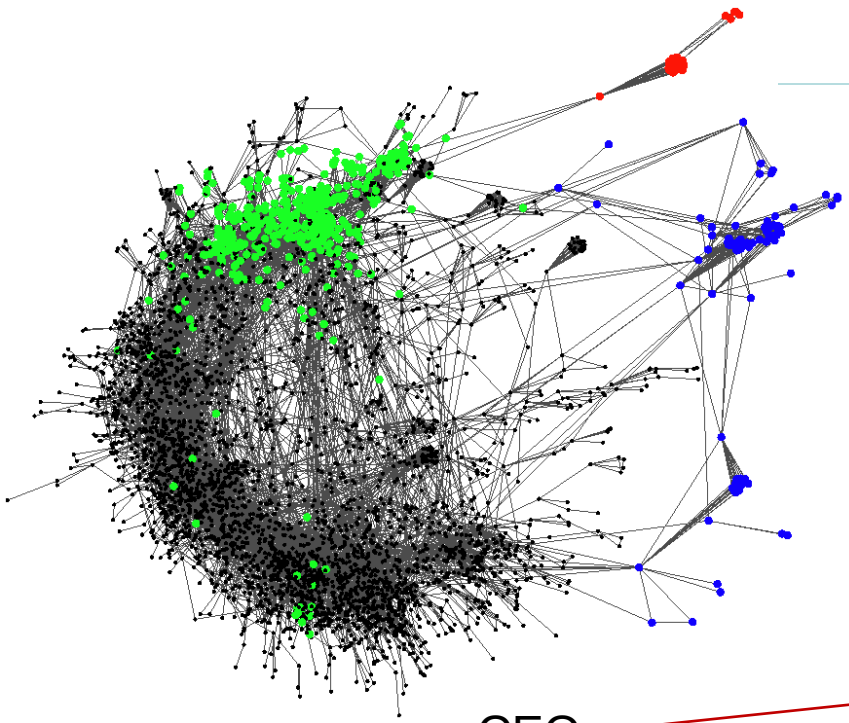
Web-based (or mobile-based) social networks already become a bridge to connect our real daily life and the virtual web space

Colorful Social Relationship?

- Social link (relationship) is one of the most important information in the social network.
- However, although one's relationship should be *colorful*, e.g., "family", "co-workers", "classmates", *users do not* take the time to create it.
  - E.g., only 16% of mobile phone users in Europe have created custom contact groups

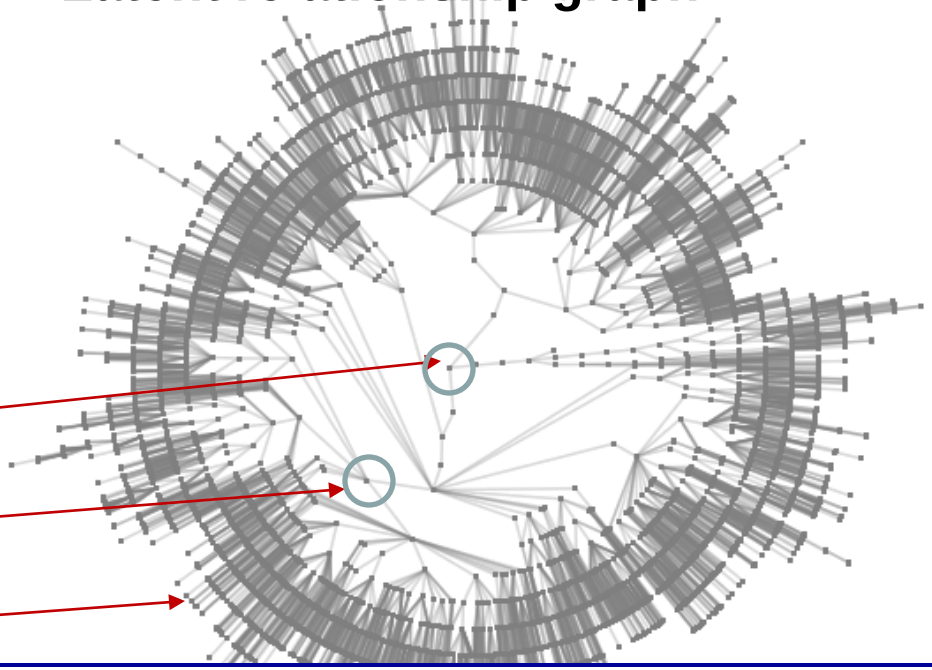
- The fact is that our social network is **black** white…

Our Social Network is Black White



Social network without role/relationship info, e.g. a company's email network

Latent relationship graph

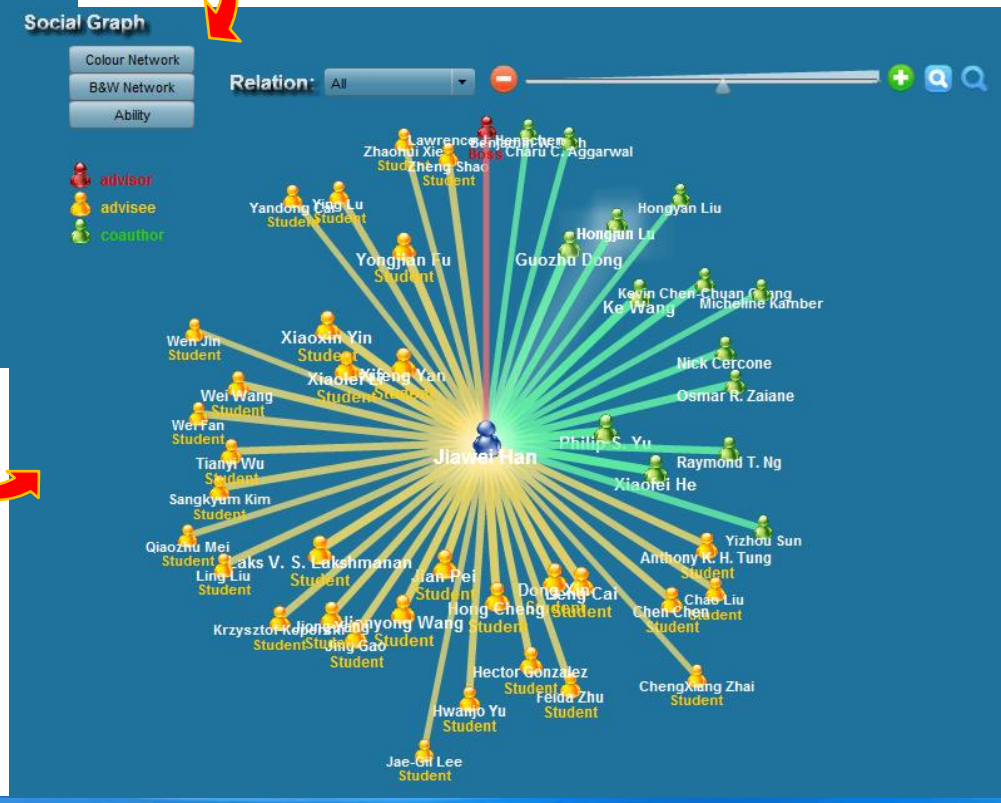


How to infer

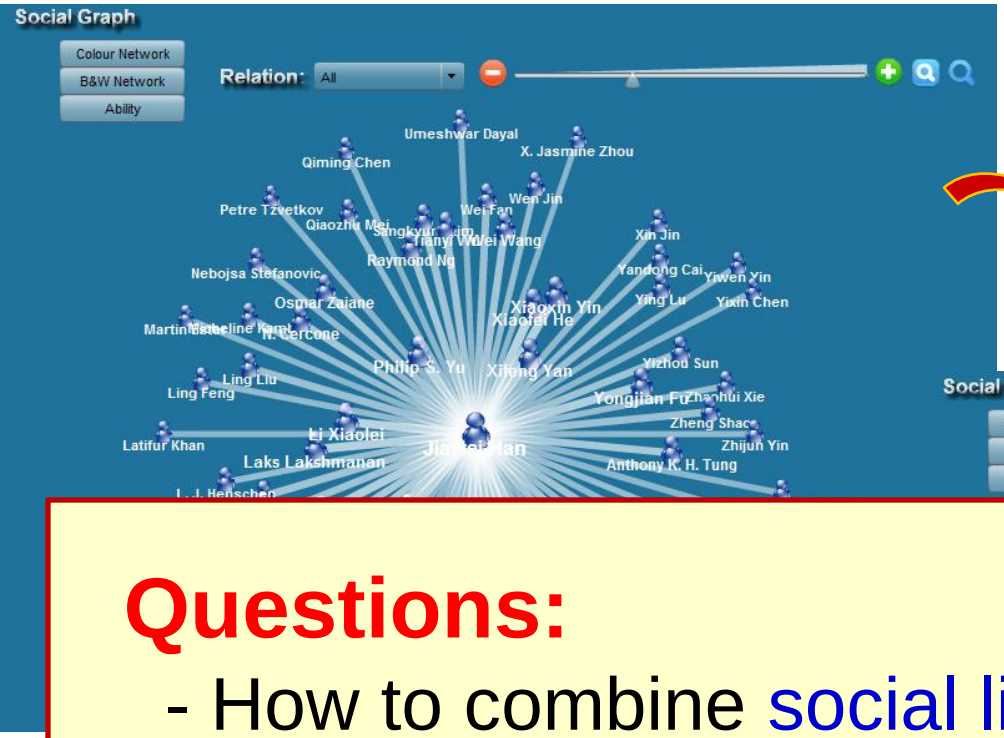
- CEO
- Manager
- Employee

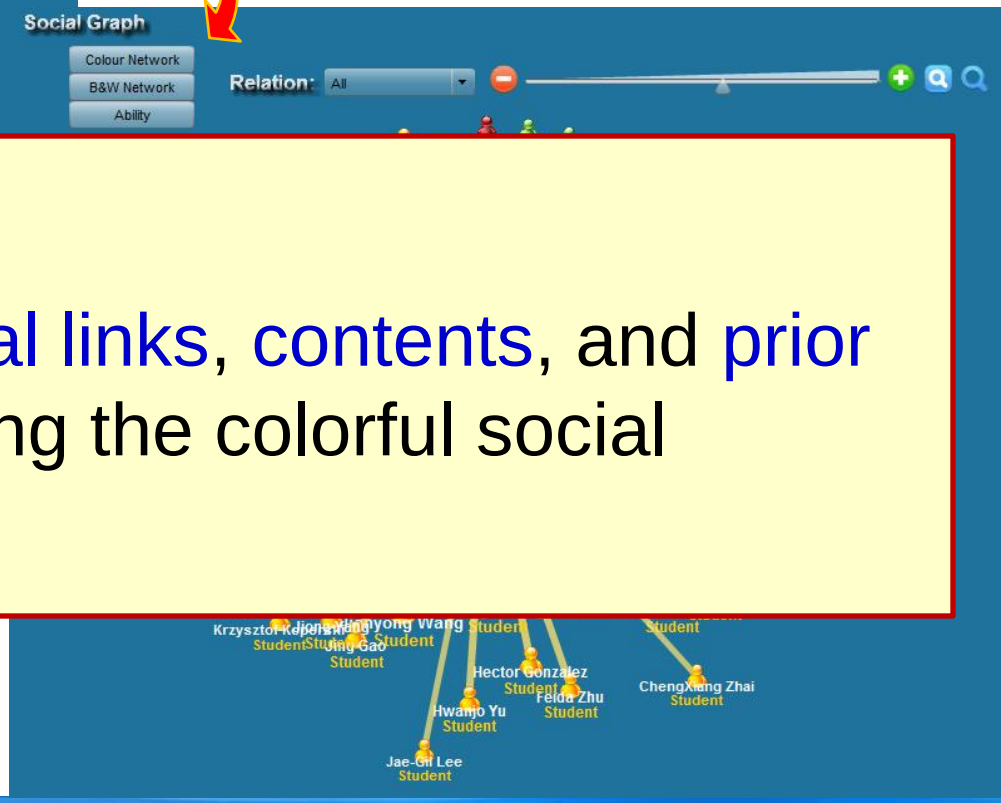
Fortunately, user interactions form *implicit groups*

From BW to Color



From BW to Color





Questions:

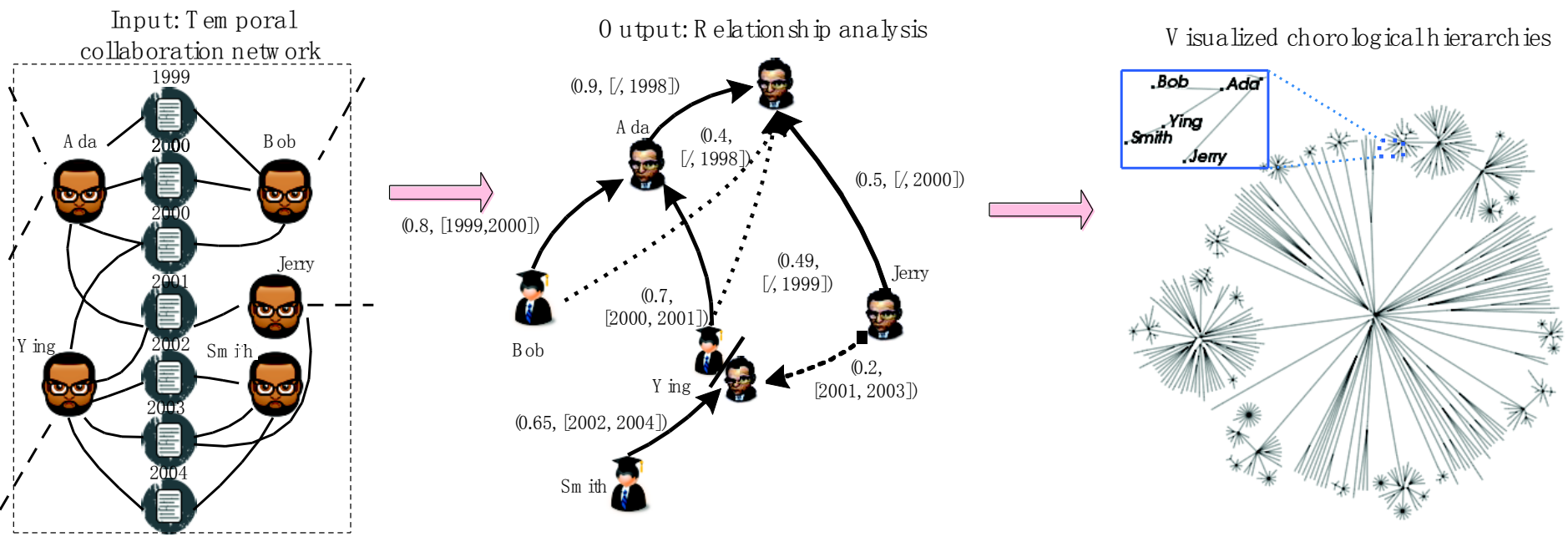
- How to combine social links, contents, and prior knowledge for discovering the colorful social relationship?

Mining Advisor-advisee Relationship from Research Publication Network

This Work: Advisor-advisee

- Input: research publication network.

  Output: potential advising relationships and their probabilities– (prob, [s_time, e_time])



Problem Analysis

Objective: extract semantic meaning from plain links to finely model and better organize information networks

Challenge?

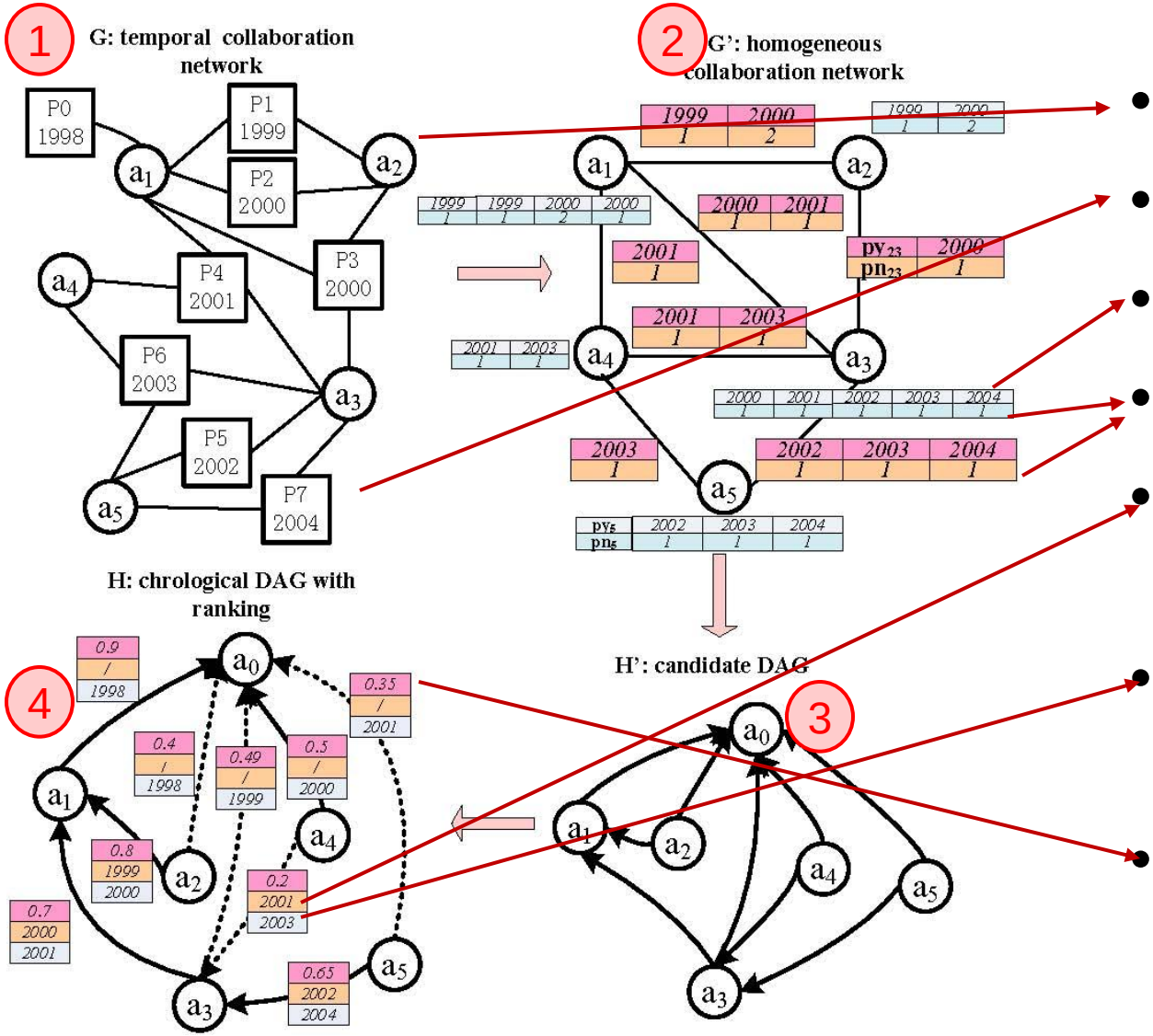
- Latent semantic knowledge
- Volatility
- Interdependency
- Scalability

Opportunity?

- Human intuition
- Propagate constraint
- Crosscheck using network
- Collective intelligence

Methodology: propagate simple intuitive rules and constraints over the whole network

Overall Framework



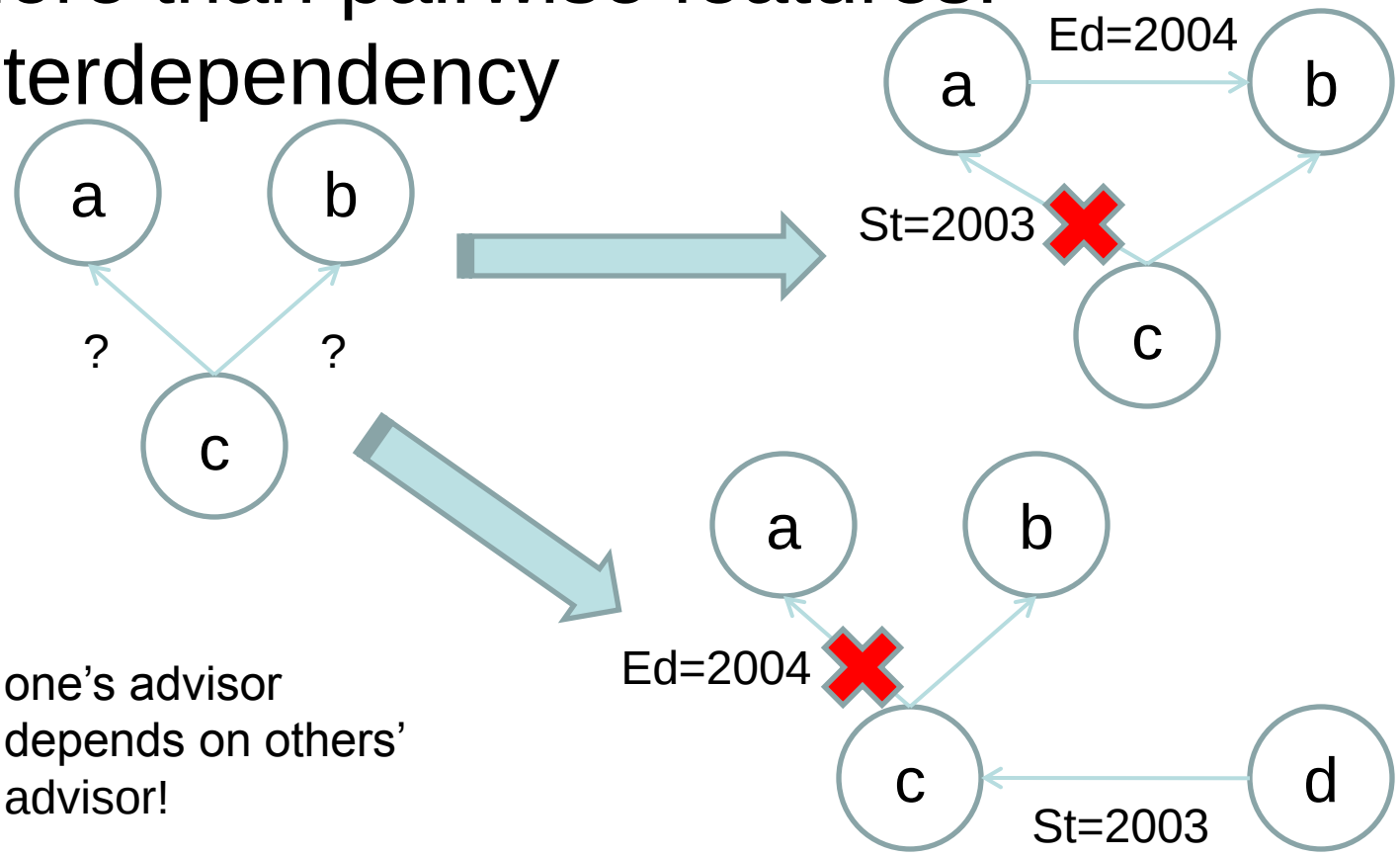
- a_i: author *i*
- p_j: paper *j*
- py: paper year
- pn: paper#
- $st_{i,yi}$: starting time
- $ed_{i,yi}$: ending time
- $r_{i,yi}$: probability

Local features - preprocess

- For every pair of coauthors a_i and a_j
  - Create a potential link from a_i to a_j if a_j has a longer publication history than a_i
  - Compute Kulczynski and Imbalance Ratio measure for the coauthored publications at different time t
  - Estimate the advising time
    - St_{ij} = the start time of coauthorship
    - Ed_{ij} = the time point when correlation drops YEAR1: $Kul_{ij}^{t} > Kul_{ij}^{t+1}$, YEAR2: $\max(Kul_{ij}^{t} - Kul_{ij}^{t+1})$
  - Remove the link if certain rules apply, o.w. sum average Kul and IR as a rough likelihood

Why is network structure helpful?

- More than pairwise features: interdependency



one's advisor depends on others' advisor!

Basic constraints

- a_y can only advise a_x after it graduated

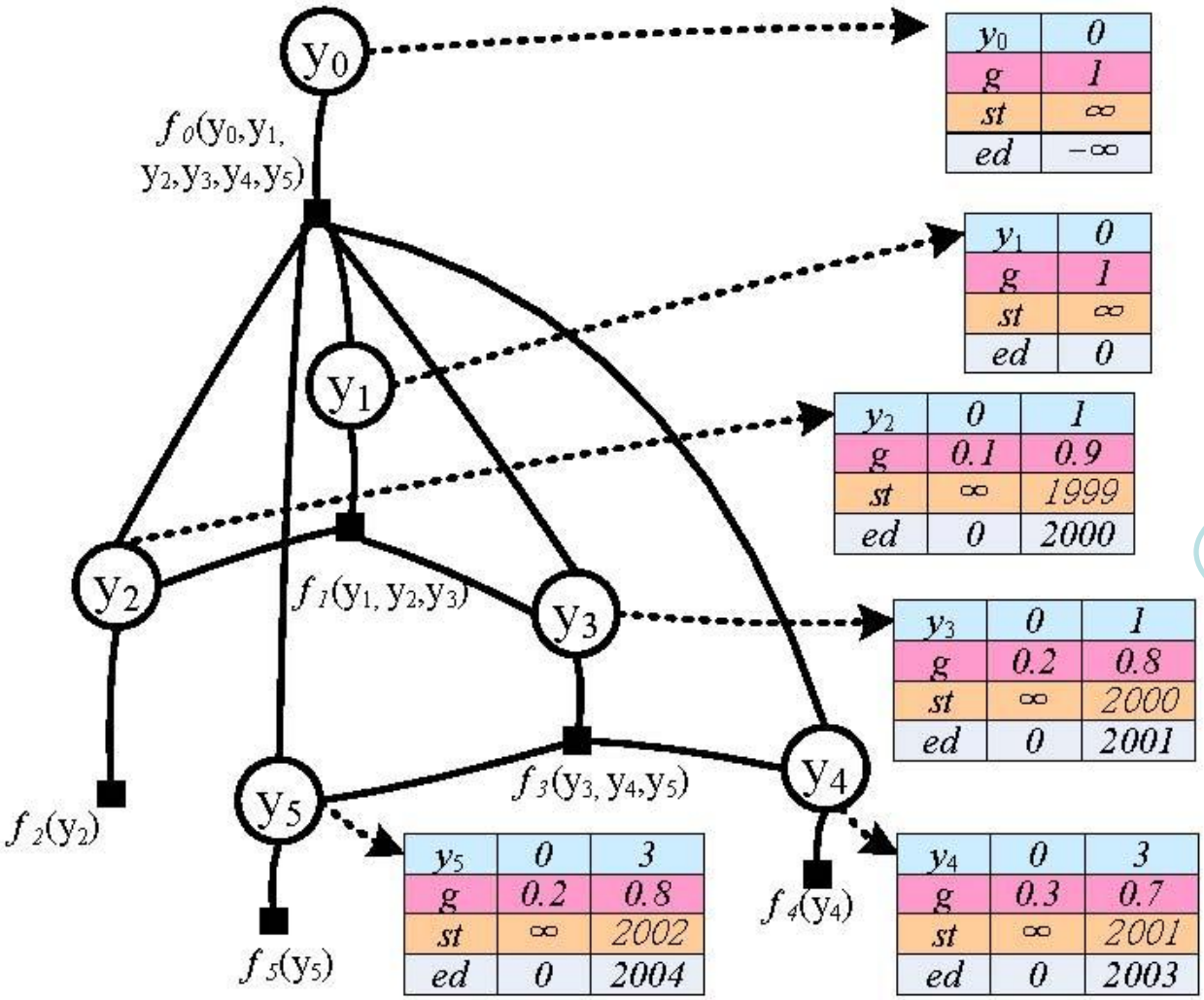
➔ $ed_y < st_x < ed_x$ If a_y advises a_x

- 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 candidate graph H' is a DAG.

The model can incorporate other intuitions as factor functions.

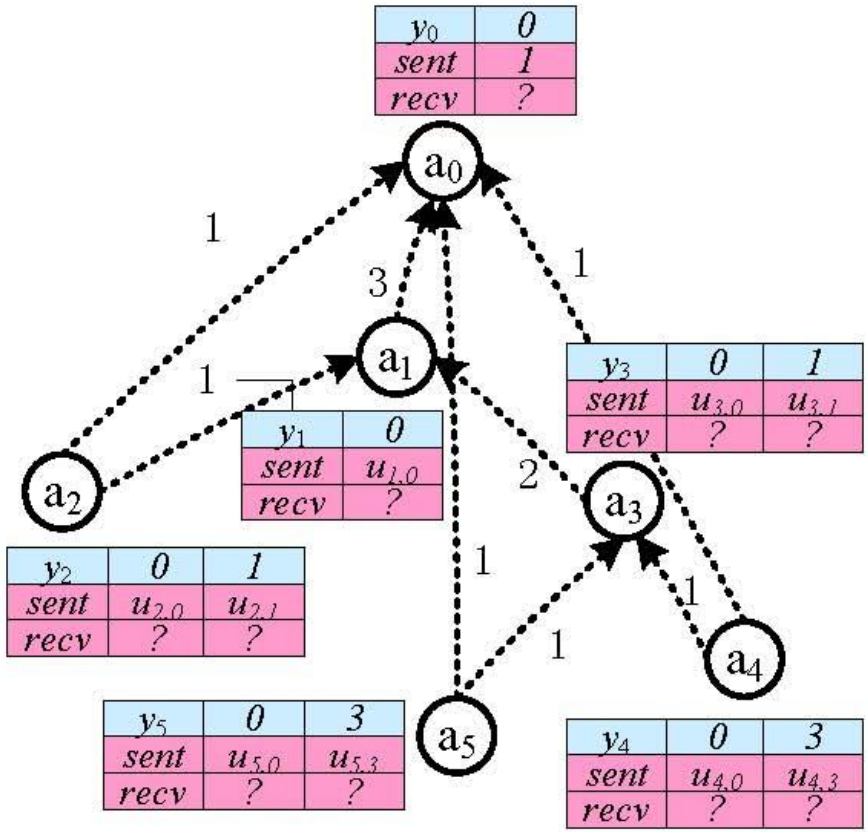
Time-constrained Probabilistic Factor Graph (TPFG)



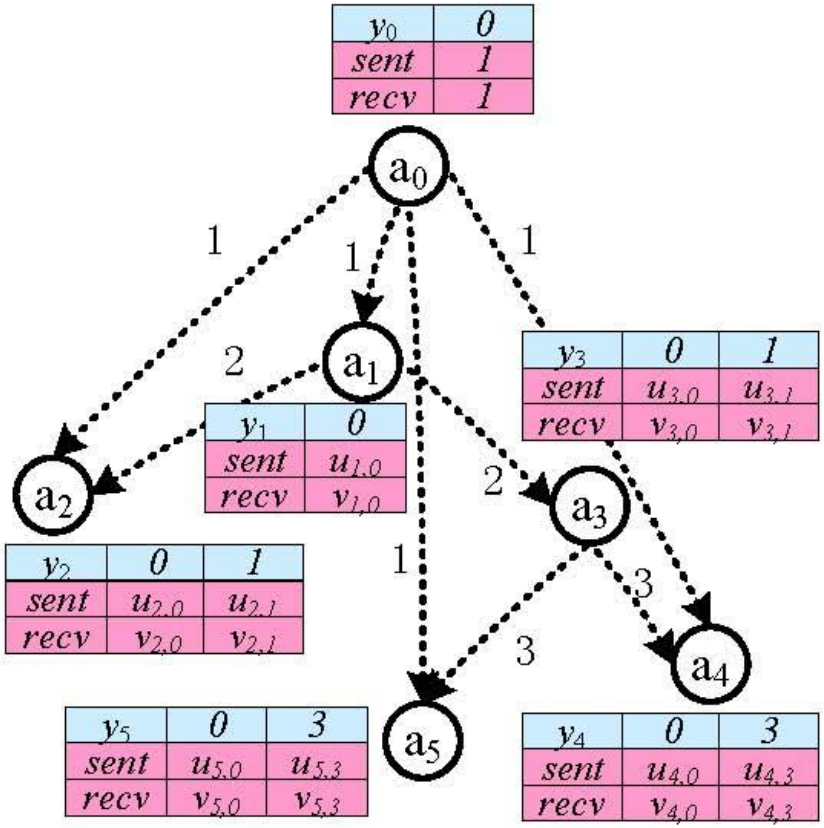
- Hidden variable y_x: a_x's advisor
- $st_{x,yx}$: starting time
  $ed_{x,yx}$: ending time
- $g(y_x, st_x, ed_x)$ is pairwise local feature
- $f_x(y_x, Z_x) = \max g(y_x, st_x, ed_x)$ *under time constraint*
- Objective function $P(\{y_x\}) = \prod_x f_x(y_x, Z)$
- $Z = \{z \mid x \in Y_z\}$
- Y_x: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

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

Experiment Results

- DBLP data: 654, 628 authors, 1076,946 publications, years provided.
- Labeled data: MathGenealogy Project; AI Gealogy Project; Faculty Homepage

Datasets	RULE	SVM	IndMAX		TPFG	
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

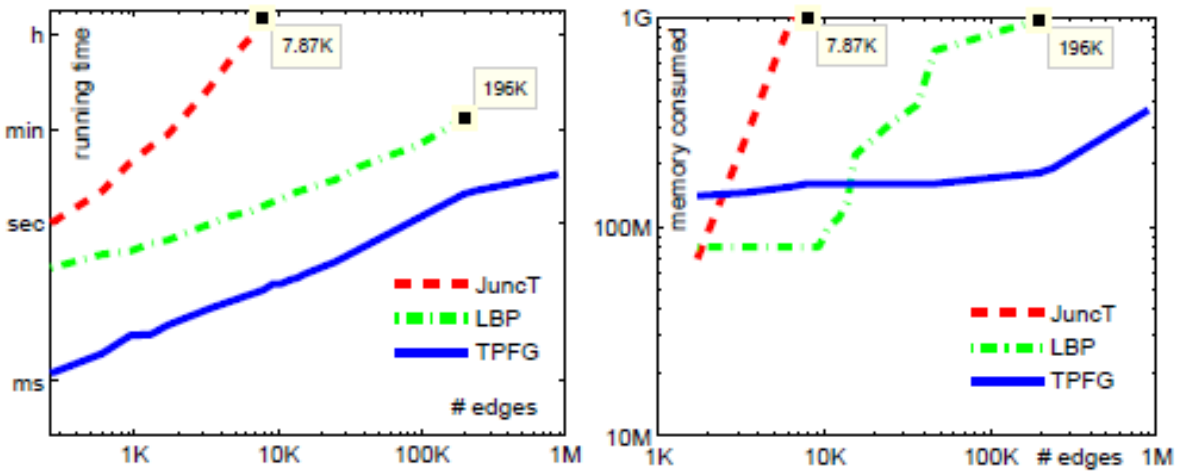
Supervised learning

Empirical parameter

optimized parameter

Case Study & Scalability

Advisee	Top Ranked Advisor	Time	Note
David M. Blei	1. Michael I. Jordan	01-03	PhD advisor, 2004 grad
	2. John D. Lafferty	05-06	Postdoc, 2006
Hong Cheng	1. Qiang Yang	02-03	MS advisor, 2003
	2. Jiawei Han	04-08	PhD advisor, 2008
Sergey Brin	1. Rajeev Motawani	97-98	“Unofficial advisor”



(a) Time (b) Space

Effect of rules - ROC curve

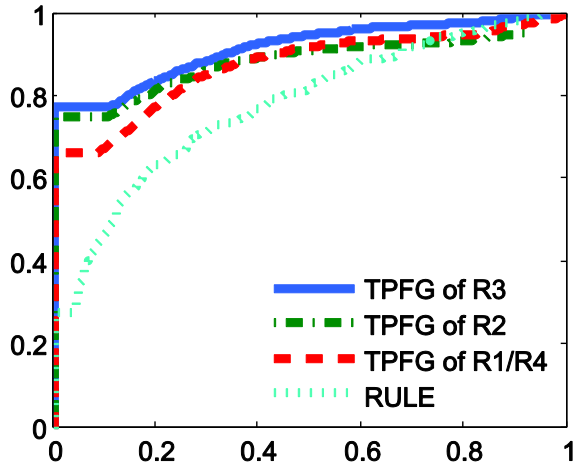
- Filtering rules in TPFG

$R1$: $IR_{ij}^t < 0$ in the sequence $\{IR_{ij}^t\}_t$ during the collaboration period of a_i and a_j,

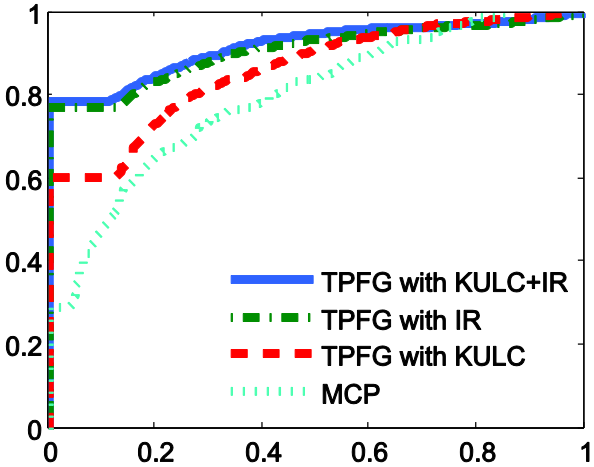
$R2$: there is no increase in the sequence $\{kulc_{ij}^t\}_t$ during the collaboration period,

$R3$: the collaboration period of a_i and a_j lasts only for one year,

$R4$: $py_j^1 + 2 > py_{ij}^1$,

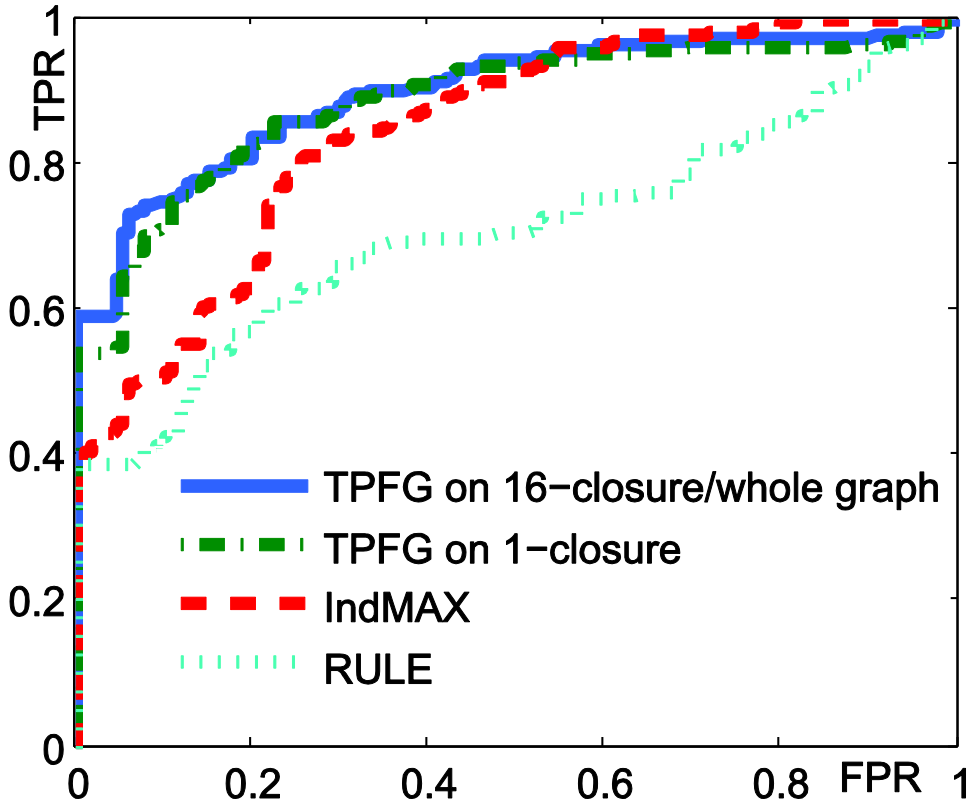


❑Local feature measure: KULC and IR



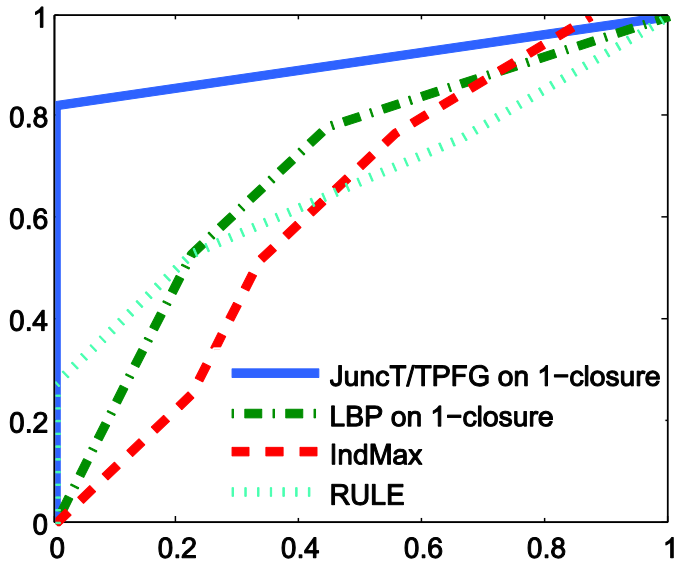
Effect of network depth

- Different closures of given set of nodes
  - DFS with bounded maximal depth d: d-closure



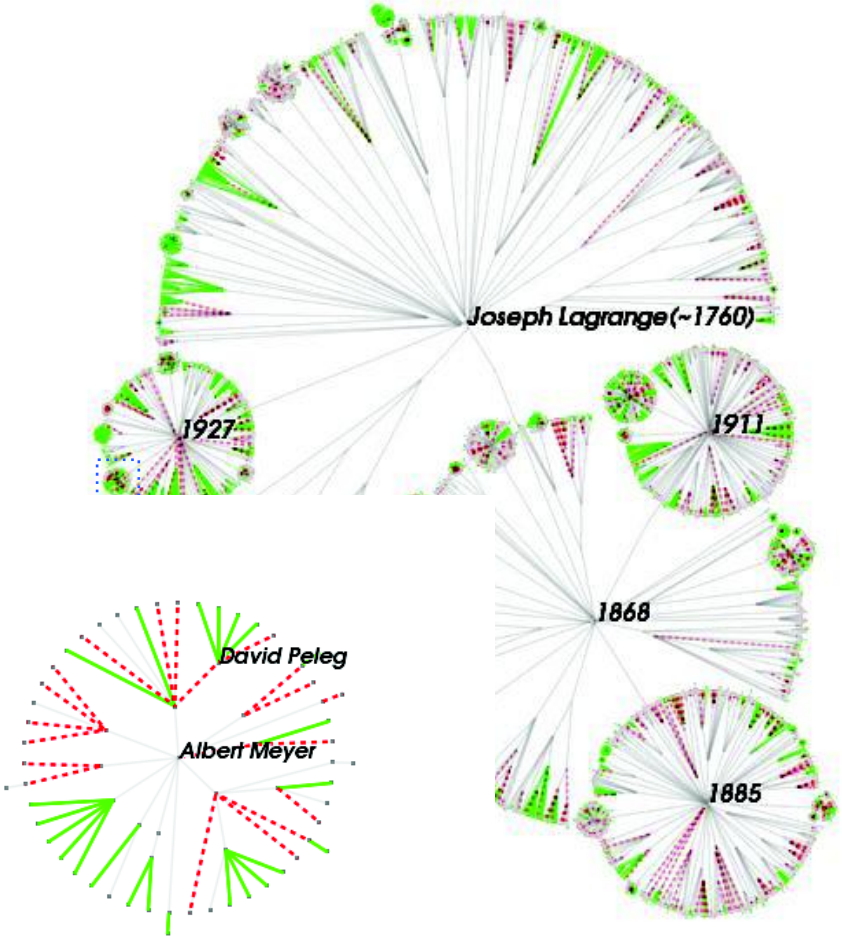
Exact VS approximate inference

- Exact inference of TPFG
  - JuncT: Junction Tree + Sum-Product
- Approximate inference
  - LBP: Loopy Belief Propagation
  - TPFG: the proposed message passing algorithm

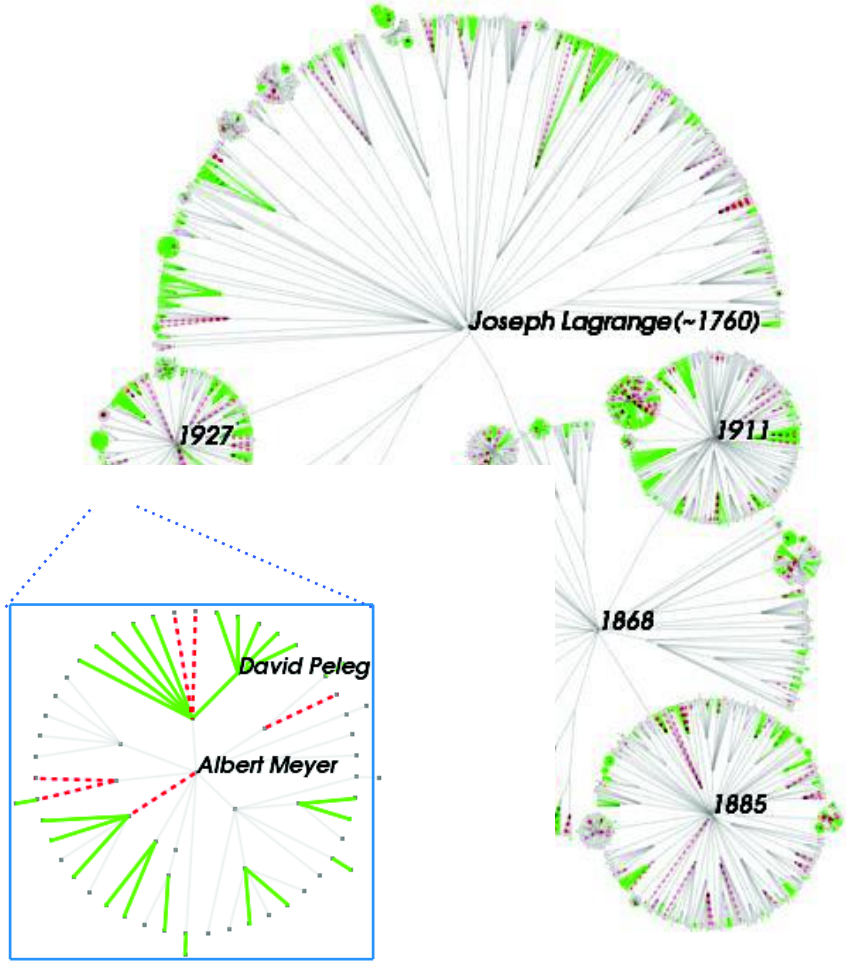


Application: visualization

RULE

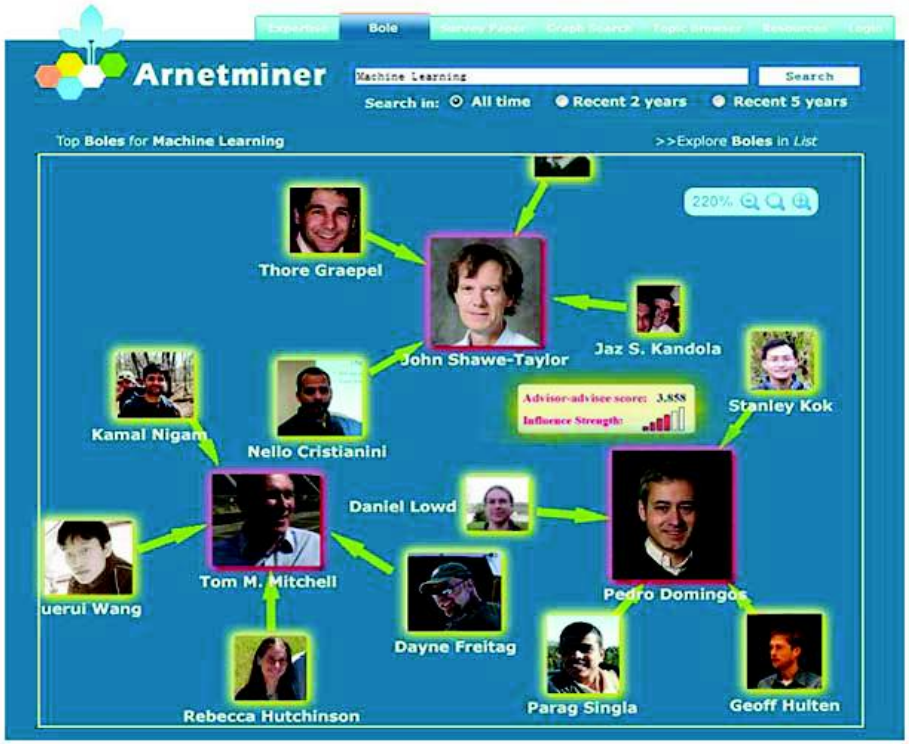


TPFG

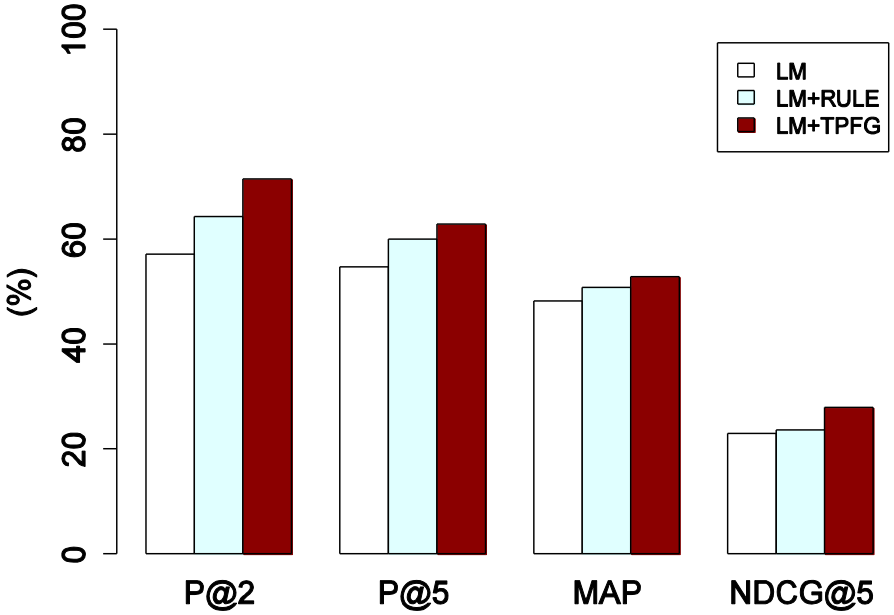


Bole Search In Arnetminer

An example on a real system: Arnetminer



Performance improvement



Ongoing and future work

- “Replying” relationship of forum posts
  - Input: forum posts without replying relation.
  - Output: a thread tree organized by their replying hierarchy
- More General: hierarchical relationship
  - Automatic feature/rule extraction
  - Generic form of constraints
  - Supervised information incorporation
- Applying to other problems
  - Hierarchical clustering; network summarization; etc.

Topic Distributions over Links on Web

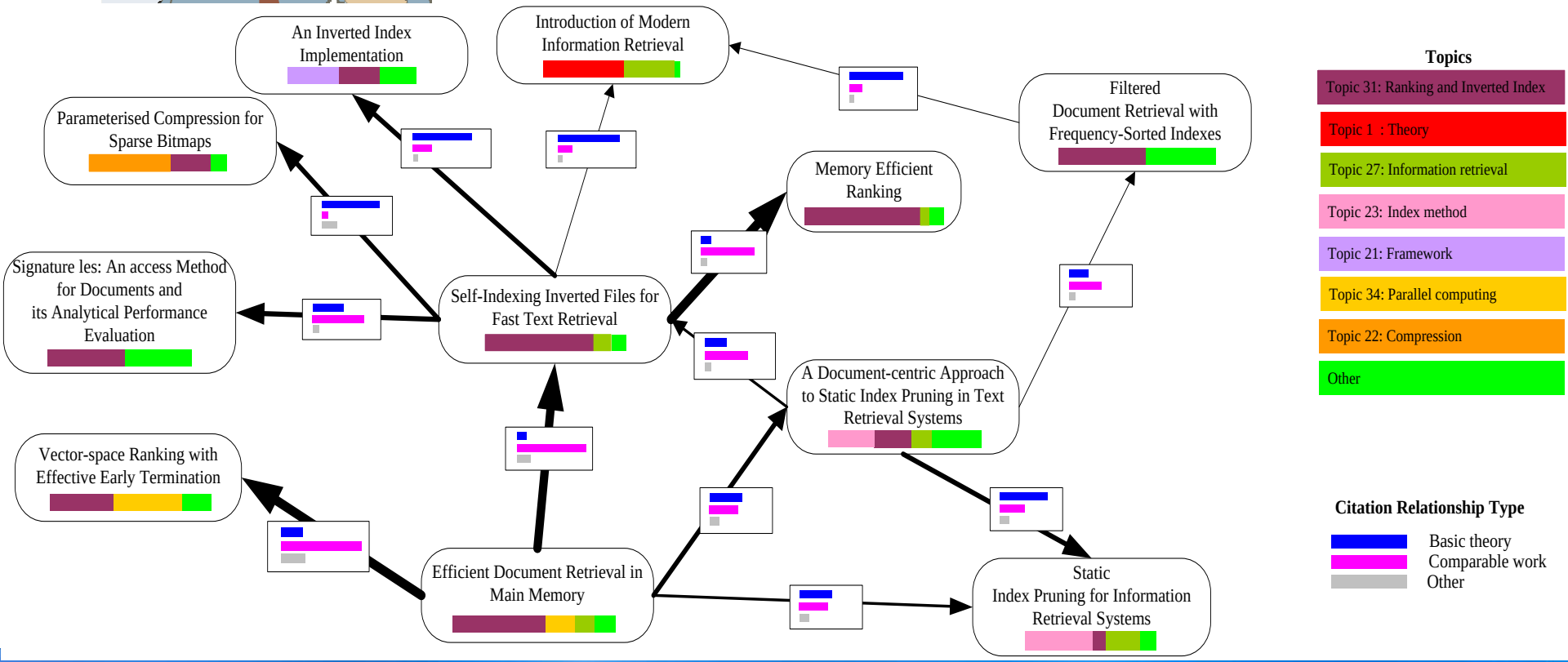
Motivation

- Web users create links with significantly different intentions
- Understanding of the category and the influence of each link can benefit many applications, e.g.,
  - Expert finding
  - Collaborator finding
  - New friends recommendation
  - …

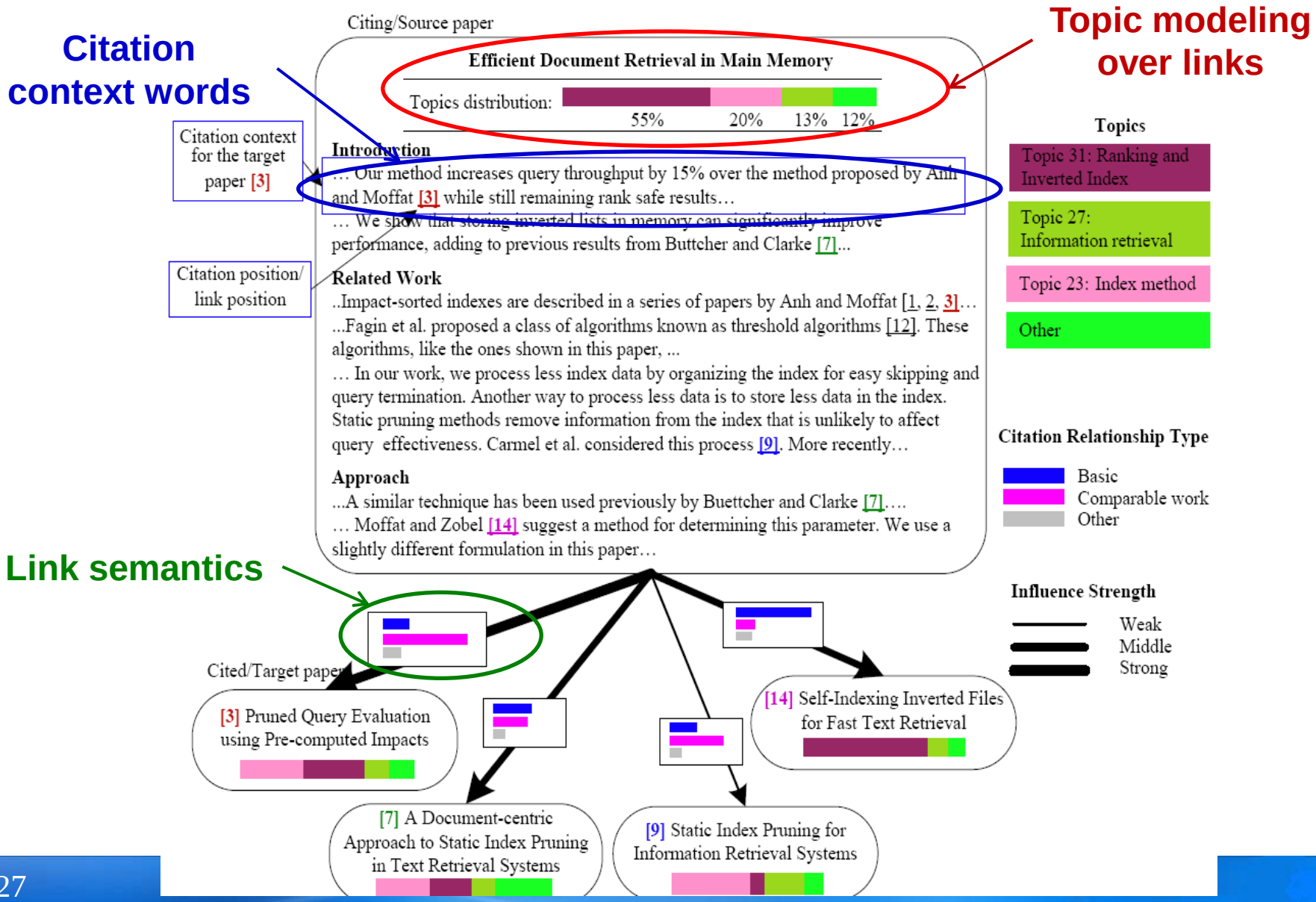
Examples – Topic distribution analysis over citations

Researcher A

- an in-depth understanding of the research field?



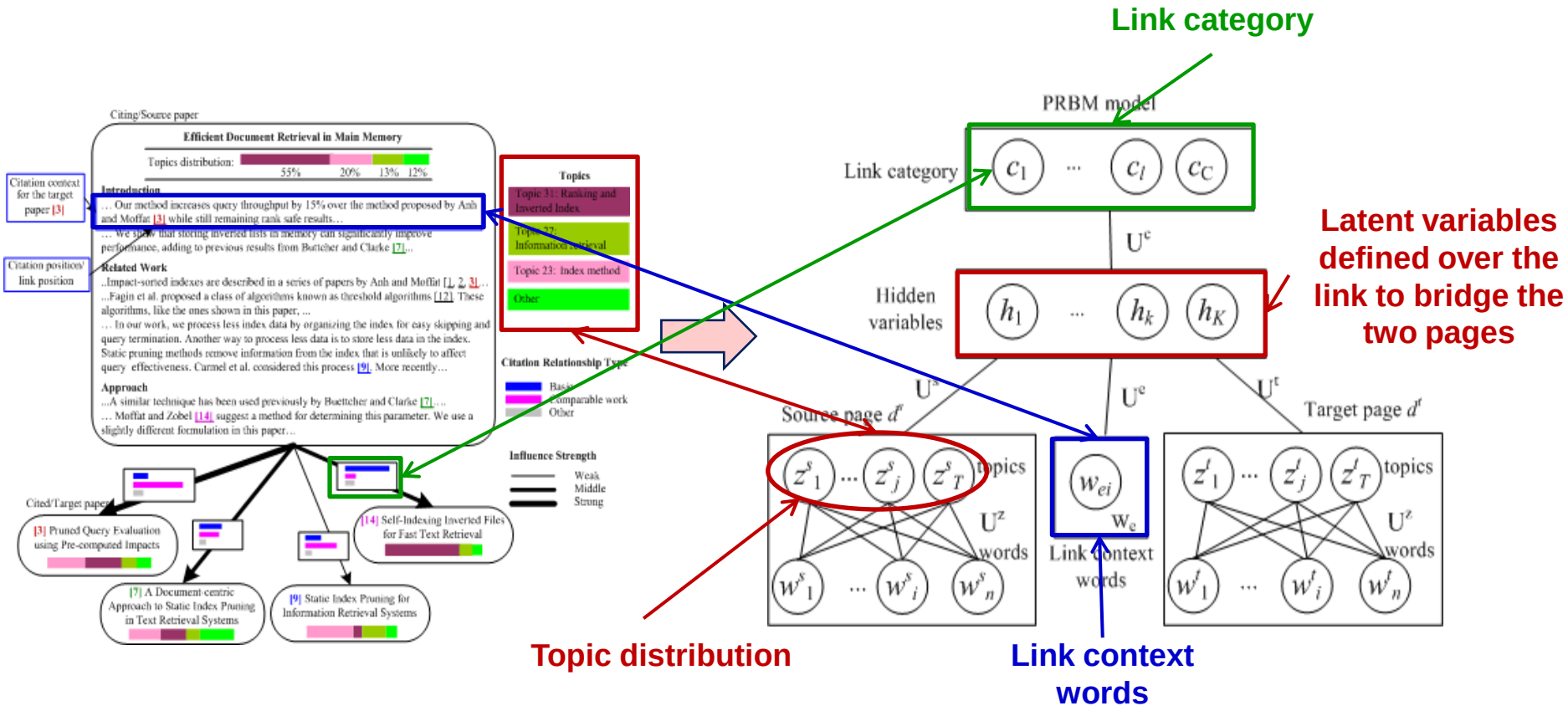
Problem: Link Semantic Analysis



Outline

- Our Approach
  - Pairwise Restricted Boltzmann Machines (PRBMs)
- Experimental Results
- Conclusion & Future Work

Pairwise Restricted Boltzmann Machines (PRBMs)

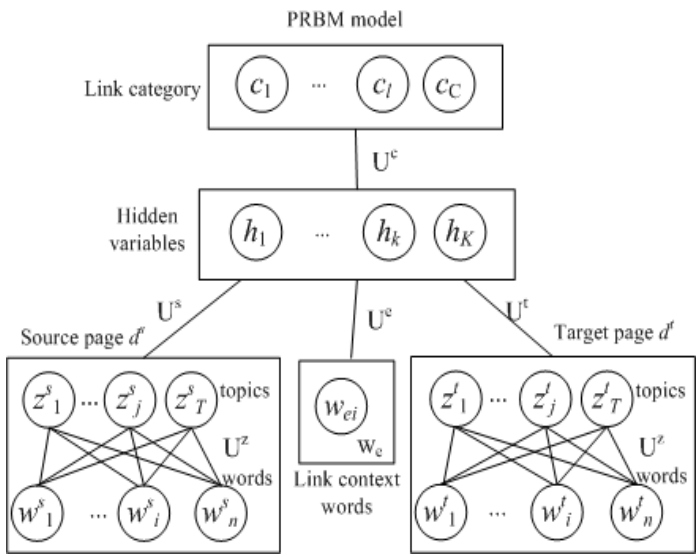


Example

Pairwise Restricted Boltzmann Machines (PRBMs)

Formalization of PRBMs

Formalization



PRBMs

Obj. Func: $$\mathcal{L}_{gen}(D) = -\sum_{e=1}^{N} \log \boxed{p(\mathbf{w}_e, \mathbf{c}_e)}$$

with

$$p(\mathbf{w}_e, \mathbf{c}_e) = \frac{1}{Z} \sum_{\mathbf{h}_e} \sum_{\mathbf{z}^s} \sum_{\mathbf{z}^t} \exp^{(-\boxed{E(\mathbf{z}^s_e, \mathbf{z}^t_e, \mathbf{w}_e, \mathbf{c}_e, \mathbf{h}_e)})}$$

$$E(\mathbf{z}^s_e, \mathbf{z}^t_e, \mathbf{w}_e, \mathbf{c}_e, \mathbf{h}_e) = -\sum_{j=1}^{T} \sum_{k=1}^{K} U^s_{jk} z^s_{ej} h_{ek}$$
$$-\sum_{j=1}^{T} \sum_{k=1}^{K} U^t_{jk} z^t_{ej} h_{ek} - \sum_{i=1}^{W} \sum_{k=1}^{K} U^e_{ik} w_{ei} h_{ek} - \sum_{l=1}^{C} \sum_{k=1}^{K} U^c_{lk} c_{el} h_{ek}$$
$$-\sum_{j=1}^{T} b_j z^s_{ej} - \sum_{j=1}^{T} b_j z^t_{ej} - \sum_{i=1}^{W} o_i w_{ei} - \sum_{k=1}^{K} g_k h_{ek} - \sum_{l=1}^{C} s_l c_{el}$$

Model Learning

Generative learning

Expectation w.r.t. the data distribution

Expectation w.r.t. the distribution defined by the model

$$\frac{\partial L_{gen}(D)}{\partial \theta} = \boxed{\mathbb{E}_{P_0}[\frac{\partial}{\partial \theta} E(D)]} - \boxed{\mathbb{E}_{P_M}[\frac{\partial}{\partial \theta} E(\hat{D})]}$$

We use the Contrast Divergence to learn the model distribution P_M

Discriminative learning

Obj. Func:

$$\mathcal{L}_{dis}(D) = -\sum_{d=1}^{M} \ln\, p(\mathbf{w}_d) - \sum_{e=1}^{N} \log\, p(\mathbf{c}_e | \mathbf{z}_e^s, \mathbf{z}_e^t, \mathbf{w}_e)$$

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U_{jk}^s} = \mathbb{E}_{P_0}[z_{ej}^s h_{ek}] - \mathbb{E}_{P_T}[z_{ej}^s \hat{h_{ek}}]$$

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U_{jk}^t} = \mathbb{E}_{P_0}[z_{ej}^t h_{ek}] - \mathbb{E}_{P_T}[z_{ej}^t \hat{h_{ek}}]$$

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U_{jk}^e} = \mathbb{E}_{P_0}[w_{ei} h_{ek}] - \mathbb{E}_{P_T}[w_{ei} \hat{h_{ek}}]$$

Hybrid learning

$$\mathcal{L}_{hybrid}(D) = \alpha \mathcal{L}_{gen}(D) + \mathcal{L}_{dis}(D)$$

Link Semantic Analysis

- Link category annotation
  - First we calculate $p(z_j = 1|\mathbf{w}) = \sigma(\sum_i U_{ij}^z w_i + a_j)$
  - Then we estimate the probability *p*(*c*|*e*) by a mean field algorithm
- Link influence estimation
  - Estimate influence by KL divergence

$$f = KL(\phi^{d^s} \| \phi^{d^t}) = \sum_{k=1}^{K} \phi_k^{d^s} \log \frac{\phi_k^{d^s}}{\phi_k^{d^t}}$$

  - An alternative way is to generate the influence score by a Gaussian distribution, thus

$$p(f|\mathbf{h}_e) = Gaussian(f| \sum_{k=1}^{K} U_k^f h_{ek} + r, 1)$$

Experimental Setting

- Data sets
  - Arnetminer data: 978,504papers, 14Mcitations
  - Wikipedia: 14K “article” pages and 25 K links
- Evaluation measures
  - Link categorization accuracy
  - Topical analysis
- Baselines:
  - SVM+LDA
  - SVM+RBM

Accuracy of Link Categorization

Approach	Type	Precision	Recall	F1-measure	Accuracy
SVM	drill down	64.09	53.60	58.24	88.50
	similar	77.66	83.34	80.33	76.59
	other	64.63	59.58	61.77	79.71
	Avg.	68.79	65.51	66.78	81.60
SVM+LDA	drill down	66.31	58.56	61.71	89.24
	similar	80.69	85.47	82.93	79.71
	other	69.58	64.36	66.75	82.42
	Avg.	72.19	69.46	70.46	83.79
SVM+RBM	drill down	68.63	53.0	59.66	89.40
	similar	76.47	91.74	83.36	78.89
	other	76.65	53.28	62.54	82.58
	Avg.	73.92	66.02	68.52	83.62
gPRBM	drill down	69.77	65.22	67.42	88.07
	similar	77.33	87.22	81.98	79.01
	other	78.00	60.94	68.42	85.19
	Avg.	75.03	71.12	72.61	84.09
dPRBM	drill down	96.00	44.44	61.54	95.88
	similar	76.85	96.89	85.71	78.60
	other	75.00	37.50	50.00	80.25
	Avg.	**82.62**	59.61	65.75	84.91
hPRBM	drill down	65.38	68.00	66.67	93.00
	similar	83.54	92.31	87.71	84.77
	other	88.14	69.33	77.61	87.65
	Avg.	79.02	**76.55**	**77.33**	**88.48**

gPRBM: our approach with generative learning

dPRBM: our approach with discriminative learning

hPRBM: our approach with hybrid learning

Category-Topic Mixture

Topics (z^s) of Citing Papers	Topics (z^t) of Cited Papers	Influential Strength
Drill down (Basic theory)		
Topic 37 (Probabilistic model)	Topic 51 (Maximum entropy)	5.05
Topic 37 (Probabilistic model)	Topic 14 (Theory)	4.38
Topic 13 (Semantic Web)	Topic 26 (Ontology)	4.36
Similar (Comparable work)		
Topic 7 (Recommendation)	Topic 7 (Recommendation)	0.00
Topic 21 (Frequent pattern learning)	Topic 21 (Frequent pattern learning)	0.00
Topic 26 (Ranking & Inverted Index)	Topic 26 (Ranking & Inverted Index)	0.00
Other		
Topic 26 (Ranking & Inverted Index)	Topic 43 (Information retrieval)	4.87
Topic 26 (Ranking & Inverted Index)	Topic 48 (Web mining)	4.24
Topic 9 (Clustering)	Topic 18 (Natural Language Processing)	4.69

Example Analysis

Source Paper	Target Paper	Influence
Drill Down (Basic theory)		
Latent dirichlet allocation	A variational bayesian framework for graphical models	2.36
A sentiment-aware model for predicting sales	Probabilistic latent semantic analysis	1.93
Mining and summarizing customer reviews	Foundations of statistical natural language processing	3.25
Similar (Comparable work)		
Constraint-driven clustering	Max-min d-cluster formation in wireless ad hoc networks	2.00
Hmms and coupled hmms for multi-channel eeg classification	Gaussian observation hidden markov models for eeg analysis	0.51
Scalable collaborative filtering using cluster-based smoothing	Evaluating collaborative filtering recommender systems	1.51
Other		
Latent dirichlet allocation	Overview of the first text retrieval conference	3.93
Utility scoring of product reviews	Introduction to modern information retrieval	3.16
Sentiment analyzer: extracting sentiments about a given topic	Finding parts in very large corpora	2.71

Conclusion & Future Work

- Concluding remarks
  - Investigate the problem of quantifying link semantics on the Web
  - Propose a Pairwise Restricted Boltzmann Machines to solve this problem
- Future Work
  - Correlation between the link semantics and the information propagation
  - Apply the link semantics to social prediction

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

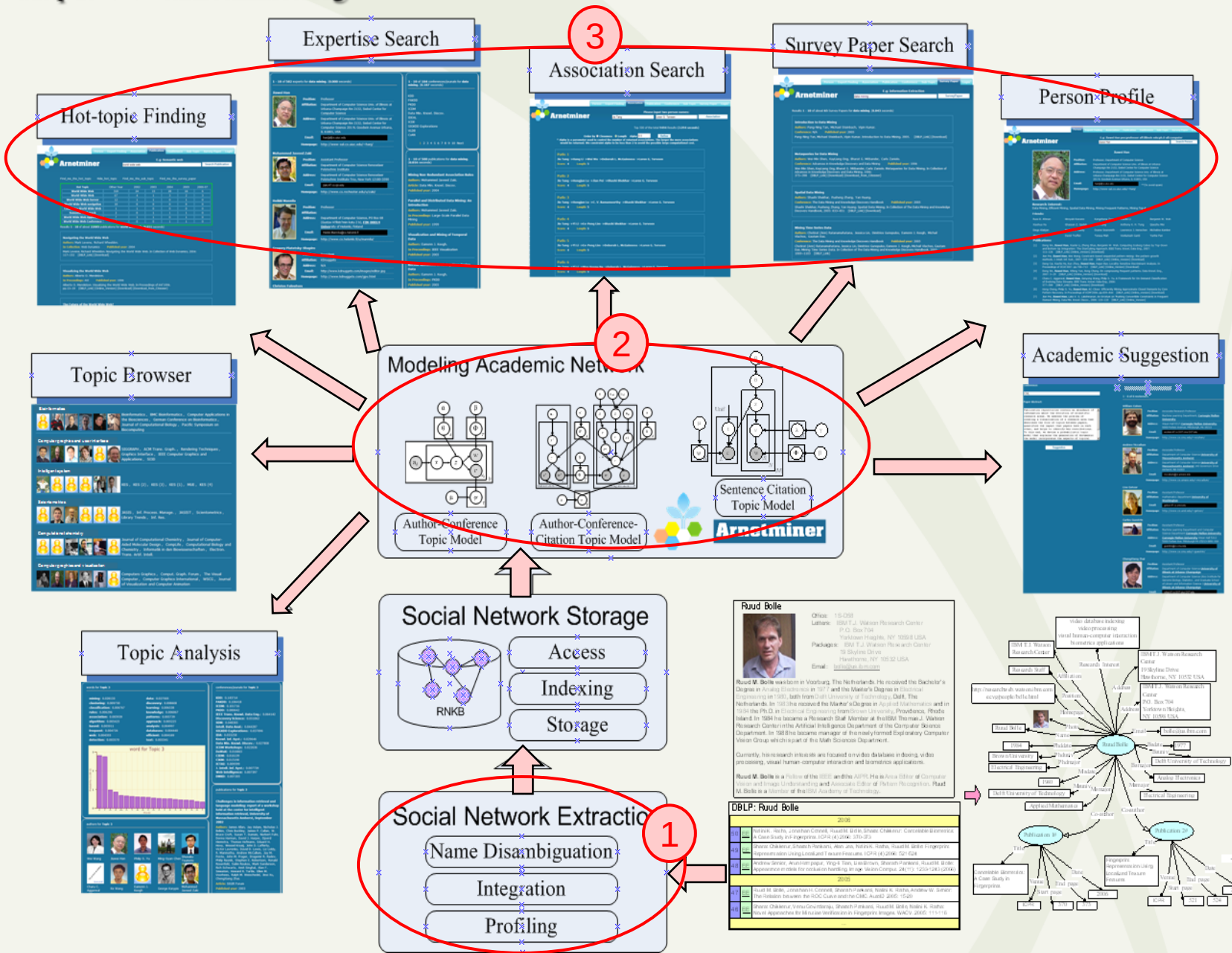
Papers published: ACM TKDD, MLJ, DKE, JIS; KDD’08-10, SDM’09, ICDM’07-09, CIKM’07-10

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



ArnetMiner's History

- 2006/5, V0.1 Perl-based CGI version
  - Profile extraction, person/paper/conf. search
- 2006/8, V1.0 Java (Demo @ ASWC)
  - Rewrite the above functions
- 2007/7, V2.0 (Demo @ KDD, ISWC)
  - New: survey search, research interest, association search
- 2008/4, V3.0 (Demo @ WWW)
  - Query understanding, New search GUI, log analysis
- 2008/11, V4.0 (Demo @ KDD, ICDM)
  - Graph search, topic mining, NSFC/NSF
- 2009/4, V5.0 (Demo @ KDD)
  - Bole/course search, profile editing, open resources, #citation
- 2009/12, V6.0
  - Academic statistics, user feedbacks, refined ranking
- V7.0, coming soon
  - Name disambiguation, reviewer assignment, supervisor suggestion, open API

ArnetMiner Today

* Ar
  >
  >
  >
  >
  >

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.

Opportunity: exploiting semantic web and social network in the real-world

Web, relational data, ontological data, social data

Data Mining and Social Network techniques

Scientific Literature

Users cover >180 countries
>600K researcher
>3M papers

Arnetminer.org
(**NSFC**, **863**)

Social search & mining

Social network Extraction
Social network Mining

IBM

Advertisement

Advertisement Recommendation

Sohu

Mobile Context

Mobile search & recommendation

Nokia

Energy trend analysis

Energy product Evolution
Techniques
Trend

Oil Company

Large-scale Mining

Scalable algorithms for message tagging and community Discovery

Google

科技信息资源内容监测与分析服务平台 **(中国科技部信息情报研究所)**

Search, browsing, complex query, integration, collaboration, trustable analysis, decision support, intelligent services,

Representative Publications

- Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM **TKDD**, 2010.
- Juanzi Li, Jie Tang, Yi Li, Qiong Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE **TKDE**, 2009.
- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. **KDD'10**.
- Chi Wang, Jiawei Han, Yuntao Jia, Duo Zhang, Yintao Yu, Jie Tang, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. **KDD'10**.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. **KDD'09**.
- 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, Hang Li, Yunbo Cao, and Zhaohui Tang. Email Data Cleaning. **KDD'05**.
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. **IJCAI'09**.
- Qian Zhong, Hanyu Li, Juanzi Li, Guotong Xie, Jie Tang, Lizhu Zhou. A Gauss Function based Approach for Unbalanced Ontology Matching. **SIGMOD'09**.
- Feng Shi, Juanzi Li, Jie Tang. Actively Learning Ontology Matching via User Interaction. **ISWC'09**.
- Chonghui Zhu, Jie Tang, Hang Li, Hwee Tou Ng, and Tiejun Zhao. A Unified Tagging Approach to Text Normalization. **ACL'07**.

Others: **ICDM'07-09**, **CIKM'07-09**, **SDM'09**, **ISWC'06**, **DKE**, **JIS**, etc.

Thanks!

Q&A

System: http://arnetminer.org

HP: http://keg.cs.tsinghua.edu.cn/persons/tj/