Exploring the Power of Heterogeneous Information Networks in Data Mining

Jiawei Han
University of Illinois at Urbana-Champaign

Collaborated with many students in my group, especially Yizhou Sun, Ming Ji, Chi Wang and Xiaoxin Yin

Acknowledgements: NSF, ARL, NASA, AFOSR (MURI), Microsoft, IBM, Yahoo!, Google, HP Lab & Boeing

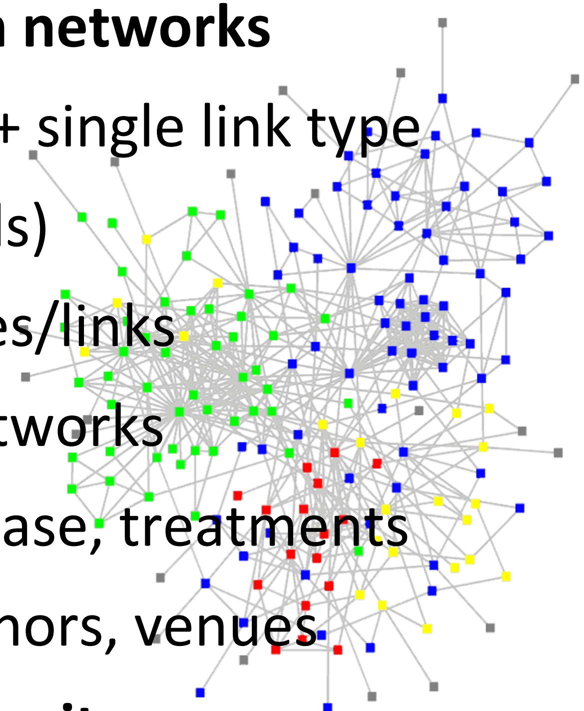
April 29, 2011

Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

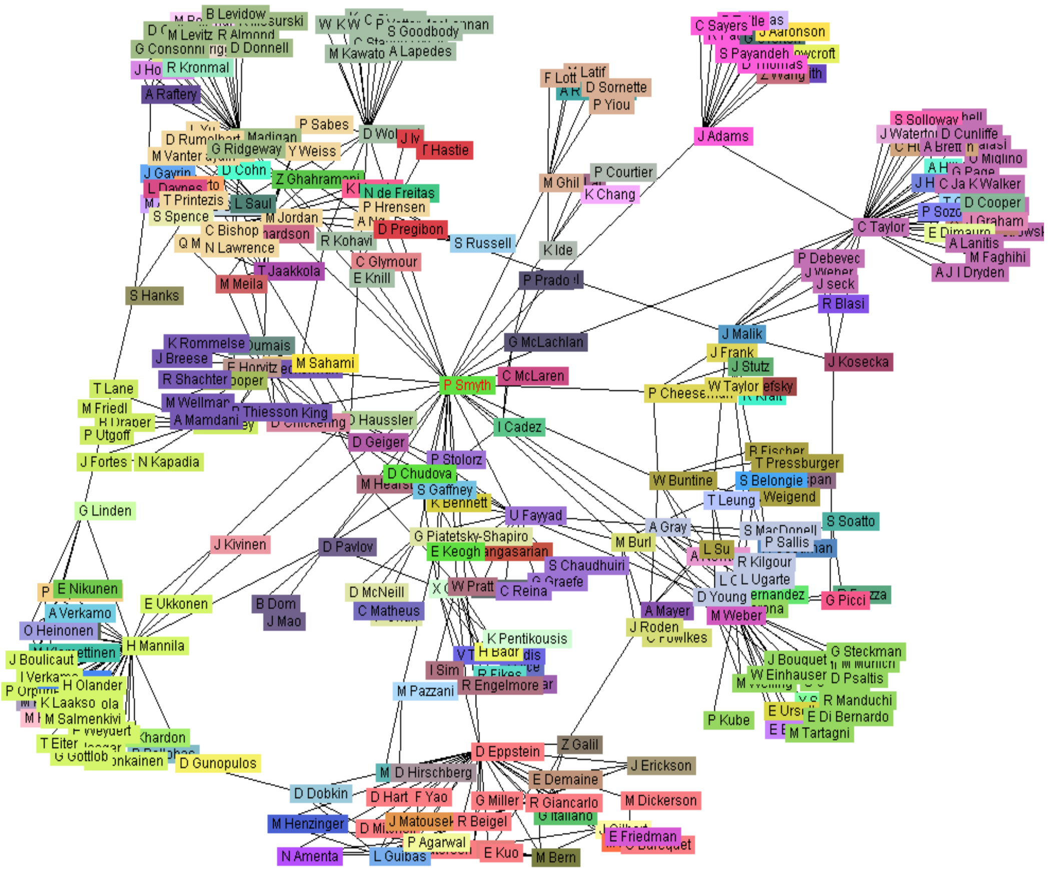
Why Mining with Heterogeneous Info. Networks?

- **Homogeneous vs. heterogeneous information networks**
  - Homogeneous network: Single object type + single link type
    - Single mode social networks (e.g., friends)
    - WWW viewed as collection of Web pages/links
  - Multi-typed, structured, heterogeneous networks
    - Medical network: patients, doctors, disease, treatments
    - Bibliographic network: publications, authors, venues

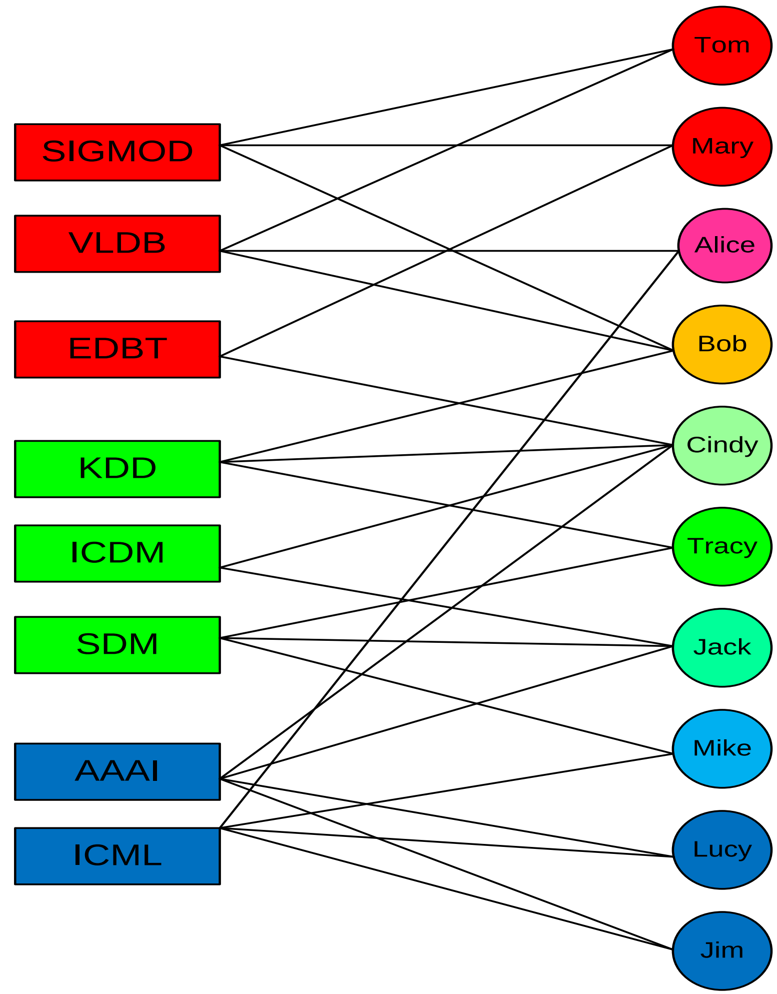


- **Heterogeneous information networks are ubiquitous**
  - Different from unorganized, multiple kinds of nodes and links
  - Typed nodes and links carry rich structural information
  - Power of mining may come from such structures and links

Homogeneous vs. Heterogeneous Networks



Co-author Network



Conference-Author Network

DBLP: An Interesting and Familiar Network

- DBLP: A computer science publication bibliographic database
  - 1.4 M records (papers), 0.7 M authors, 5 K conferences, …
- Will this database disclose interesting knowledge about us?
  - How are CS research forums structured?
  - Who are the leading researchers on Web search?
  - How do the authors in this subfield collaborate and evolve?
  - How many Wei Wang's in DBLP, which papers by which one?
  - Who is Sergy Brin's supervisor and when?
  - Can you predict which topics Faloutsos will work on? ……
- All these kinds of questions, and potentially much more, can be nicely answered by the DBLP-InfoNet
  - How? Exploring the power of structures and links in networks!

Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

RankClus: Clustering and Ranking in Heterogeneous Information Networks

- Ranking & clustering: Each provides a structured view on data
- Ranking globally without considering clusters?
  - Dumb!! One cannot rank chicken and ducks together!
- Clustering authors in one huge cluster without distinction?
  - Dull!! 30000 entries found? (this is why PageRank!)
- RankClus: Integrates clustering with ranking
  - Ranking is conditional (i.e., relative) to a specific cluster
  - Better clustering? Using highly ranked objects!
- RankClus: Clustering and ranking are mutually enhanced
- *RankClus: Integrating Clustering with Ranking for Heterog. Information Network Analysis* (Y. Sun, J. Han, et al.) EDBT'09.

Global Ranking vs. Cluster-Based Ranking

- A toy example: One cannot rank chicken and ducks together!
  - Two areas with 10 conf.s and 100 authors in each area

Table 1: A set of conferences from two research areas

DB/DM	{SIGMOD, VLDB, PODS, ICDE, ICDT, KDD, ICDM, CIKM, PAKDD, PKDD}
HW/CA	{ASPLOS, ISCA, DAC, MICRO, ICCAD, HPCA, ISLPED, CODES, DATE, VTS }

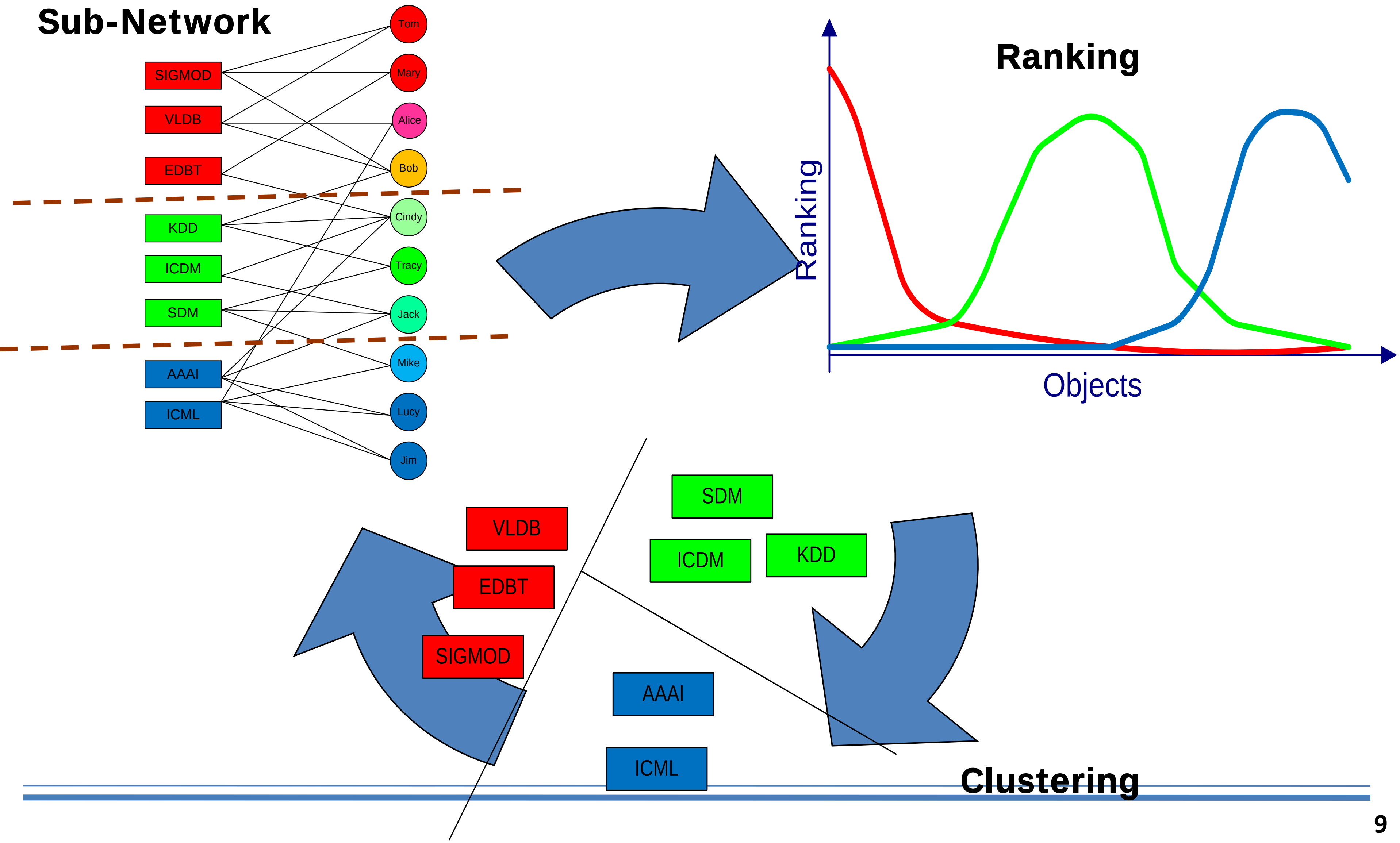
Table 2: Top-10 ranked conferences and authors in the mixed conference set

Rank	Conf.	Rank	Authors
1	DAC	1	Alberto L. Sangiovanni-Vincentelli
2	ICCAD	2	Robert K. Brayton
3	DATE	3	Massoud Pedram
4	ISLPED	4	Miodrag Potkonjak
5	VTS	5	Andrew B. Kahng
6	CODES	6	Kwang-Ting Cheng
7	ISCA	7	Lawrence T. Pileggi
8	VLDB	8	David Blaauw
9	SIGMOD	9	Jason Cong
10	ICDE	10	D. F. Wong

Table 3: Top-10 ranked conferences and authors in DB/DM set

Rank	Conf.	Rank	Authors
1	VLDB	1	H. V. Jagadish
2	SIGMOD	2	Surajit Chaudhuri
3	ICDE	3	Divesh Srivastava
4	PODS	4	Michael Stonebraker
5	KDD	5	Hector Garcia-Molina
6	CIKM	6	Jeffrey F. Naughton
7	ICDM	7	David J. DeWitt
8	PAKDD	8	Jiawei Han
9	ICDT	9	Rakesh Agrawal
10	PKDD	10	Raghu Ramakrishnan

RankClus: An Integrated Framework



The RankClus Philosophy

- Why integrated Ranking and Clustering?
  - Ranking and clustering can be mutually improved
  - Ranking: Once a cluster becomes more accurate, ranking will be more reasonable for such a cluster and will be the distinguished feature of the cluster
  - Clustering: Once ranking is more distinguished from each other, the clusters can be adjusted and get more accurate results
- Not every object should be treated equally in clustering!
- Objects preserve similarity under new measure space
  - E.g., VLDB vs. SIGMOD

RankClus: Algorithm Framework

- Step 0. Initialization
  - Randomly partition target objects into K clusters
- Step 1. Ranking
  - Ranking for each sub-network induced from each cluster, which serves as feature for each cluster
- Step 2. Generating new measure space
  - Estimate mixture model coefficients for each target object
- Step 3. Adjusting cluster
- Step 4. Repeating Steps 1-3 until stable

Focus on a Bi-Typed Network Case

- Conference-author network, links can exist between
  - Conference (X) and author (Y)
  - Author (Y) and author (Y)

DEFINITION 1. ***Bi-type Information Network.*** *Given two types of object sets X and Y, where $X = \{x_1, x_2, \ldots, x_m\}$, and $Y = \{y_1, y_2, \ldots, y_n\}$, graph $G = \langle V, E \rangle$ is called a bi-type information network on types X and Y, if $V(G) = X \cup Y$ and $E(G) = \{\langle o_i, o_j \rangle\}$, where $o_i, o_j \in X \cup Y$.*

- Use W to denote the links and there weights

$$W = \begin{pmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{pmatrix}$$

Ranking: Feature Extraction

- Simple ranking vs. authority ranking
- Simple Ranking
  - Proportional to degree counting for objects, e.g., # of publications of an author
  - Considers only immediate neighborhood in the network
- Authority Ranking: Extension to HITS in weighted bi-type network
  - Rule 1: Highly ranked authors publish *many* papers in highly ranked conferences
  - Rule 2: Highly ranked conferences attract *many* papers from *many* highly ranked authors
  - Rule 3: The rank of an author is enhanced if he or she co-authors with many authors or many highly ranked authors

Encoding Rules in Authority Ranking

- Rule 1: Highly ranked authors publish *many* papers in highly ranked conferences

$$\vec{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i)\vec{r}_X(i).$$

- Rule 2: Highly ranked conferences attract *many* papers from *many* highly ranked authors

$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j)$$

- Rule 3: The rank of an author is enhanced if he or she co-authors with many authors or many highly ranked authors

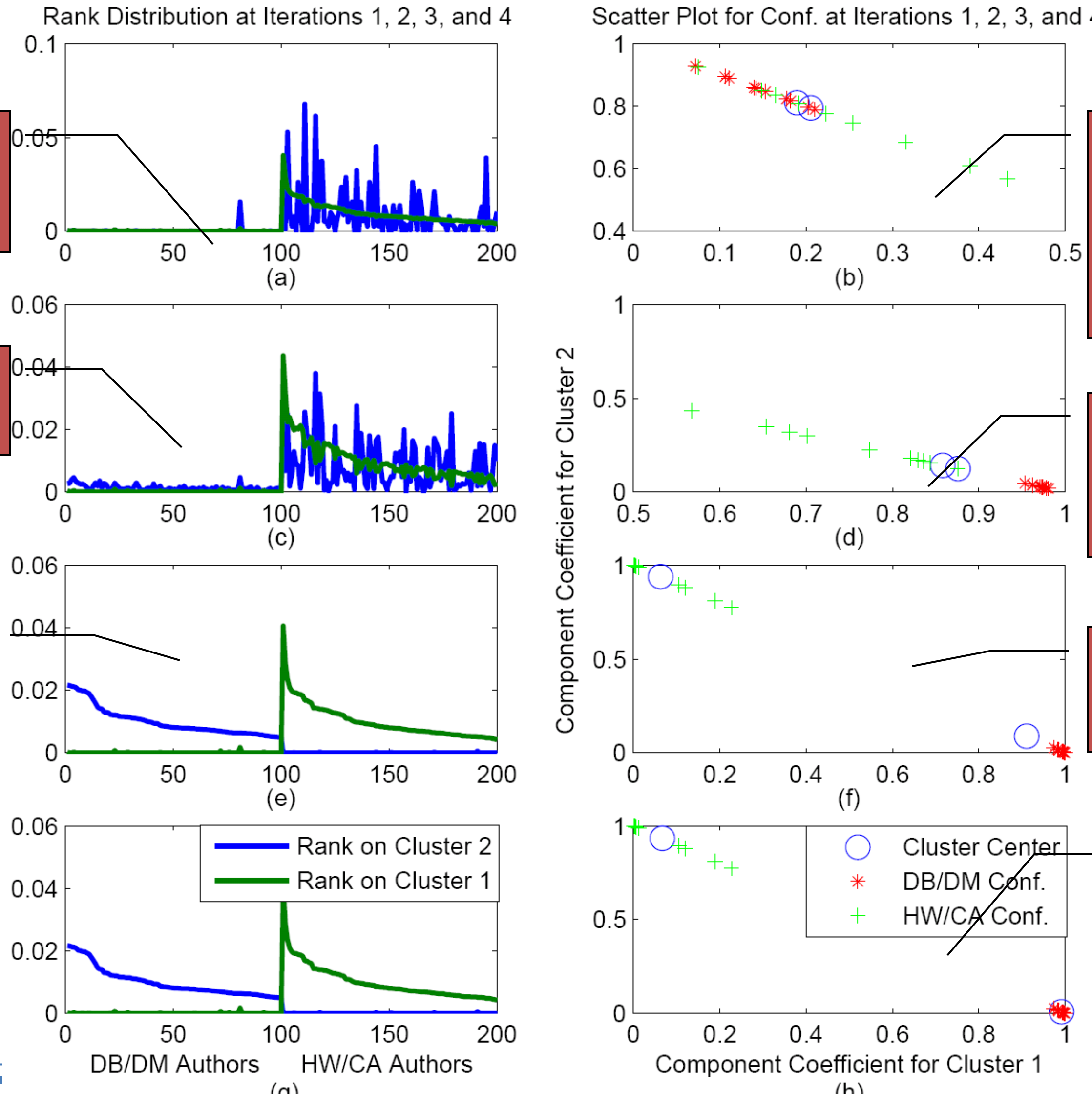
$$\vec{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\vec{r}_X(j) + (1-\alpha)\sum_{j=1}^{n} W_{YY}(i,j)\vec{r}_Y(j)$$

Step-by-Step Running of RankClus

Initially, ranking distributions are mixed together

Improved a little

Improved significantly



Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable

Case Study: Dataset: DBLP

- All the 2676 conferences and 20,000 authors with most publications, from the time period of year 1998 to year 2007
- Both conference-author relationships and co-author relationships are used
- K=15 (select only 5 clusters here)

Table 5: Top-10 Conferences in 5 Clusters Using RankClus

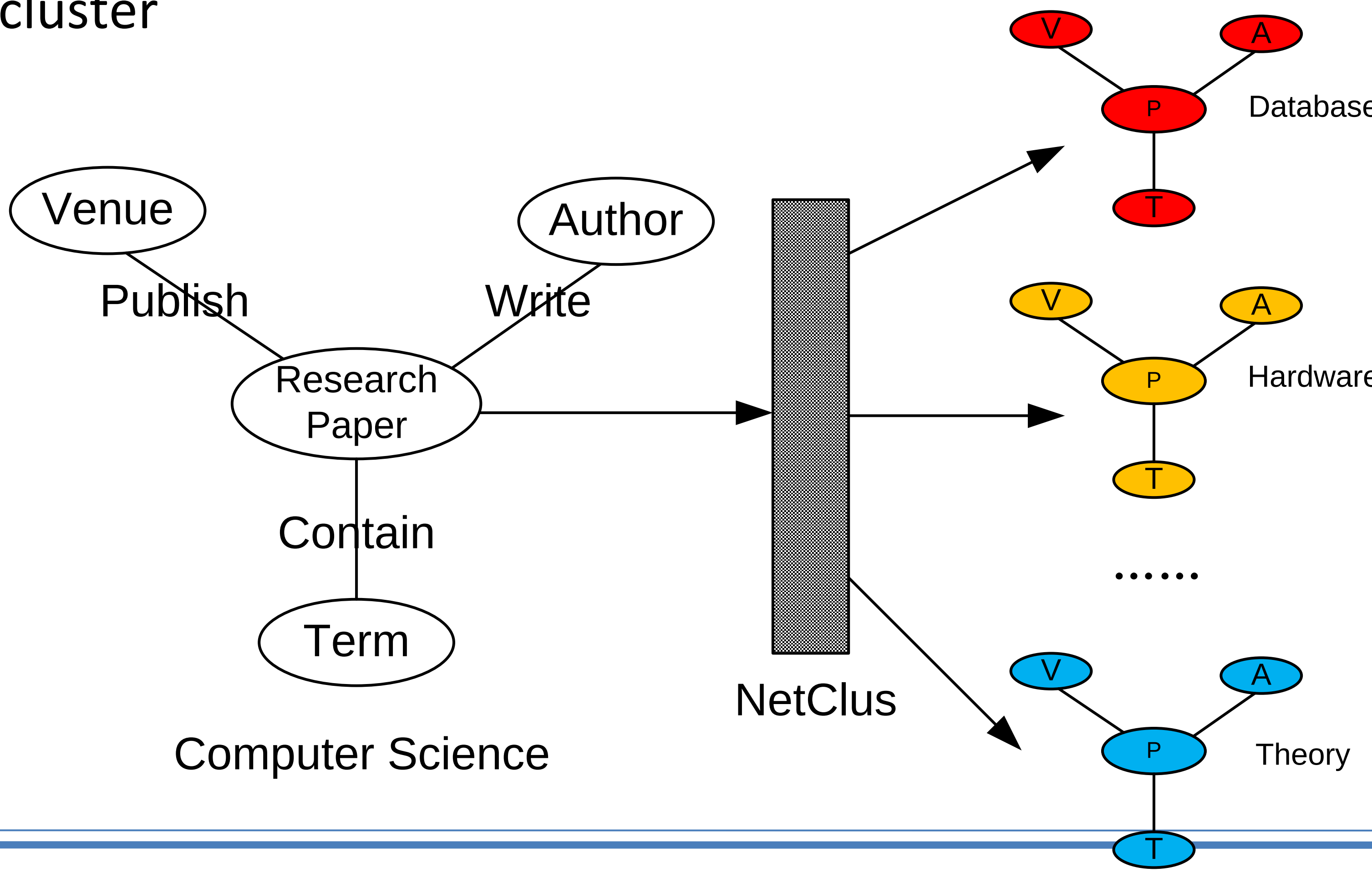
	DB	Network	AI	Theory	IR
1	VLDB	INFOCOM	AAMAS	SODA	SIGIR
2	ICDE	SIGMETRICS	IJCAI	STOC	ACM Multimedia
3	SIGMOD	ICNP	AAAI	FOCS	CIKM
4	KDD	SIGCOMM	Agents	ICALP	TREC
5	ICDM	MOBICOM	AAAI/IAAI	CCC	JCDL
6	EDBT	ICDCS	ECAI	SPAA	CLEF
7	DASFAA	NETWORKING	RoboCup	PODC	WWW
8	PODS	MobiHoc	IAT	CRYPTO	ECDL
9	SSDBM	ISCC	ICMAS	APPROX-RANDOM	ECIR
10	SDM	SenSys	CP	EUROCRYPT	CIVR

Time Complexity: Linear to # of Links

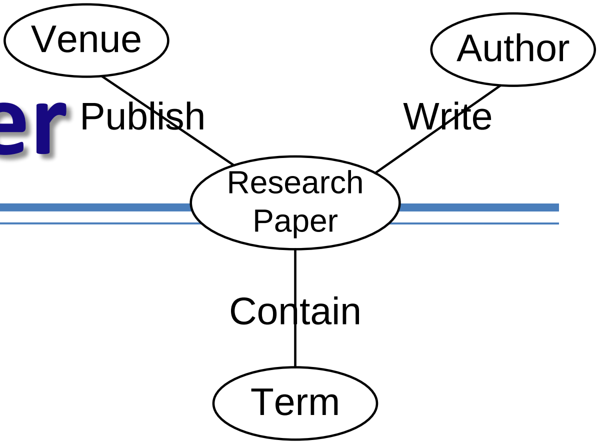
- At each iteration, |E|: edges in network, m: number of target objects, K: number of clusters
  - Ranking for sparse network
    - ~O(|E|)
  - Mixture model estimation
    - ~O(K|E|+mK)
  - Cluster adjustment
    - ~O(mK^2)
- In all, linear to |E|
  - ~O(K|E|)
- Note: SimRank will be at least quadratic at each iteration since it evaluates distance between every pair in the network

NetClus: Ranking & Clustering with Star Network Schema [KDD'09]

- Beyond bi-typed information network: A Star Network Schema
- Split a network into different layers, each representing by a net-cluster

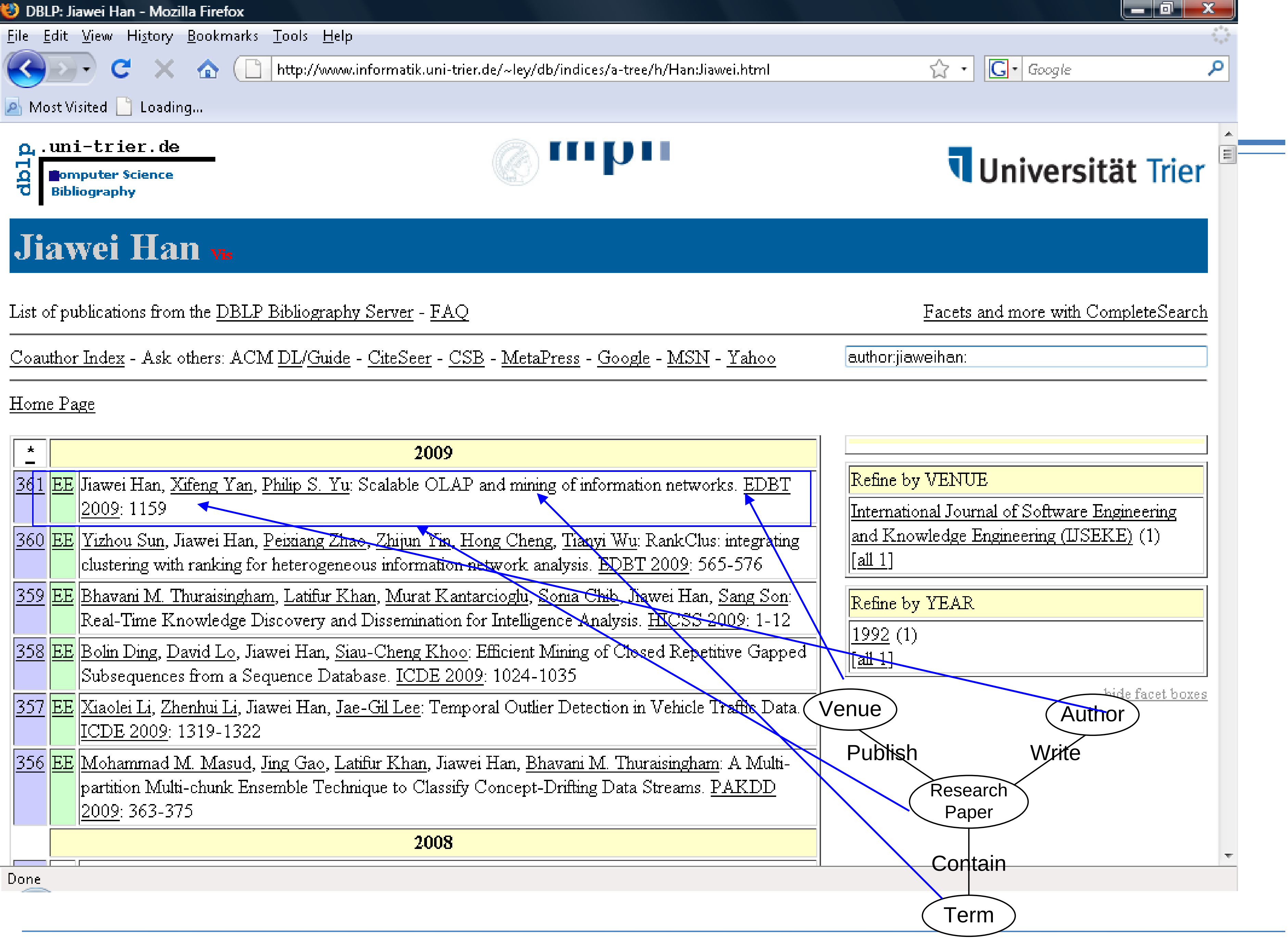


StarNet: Schema & Net-Cluster

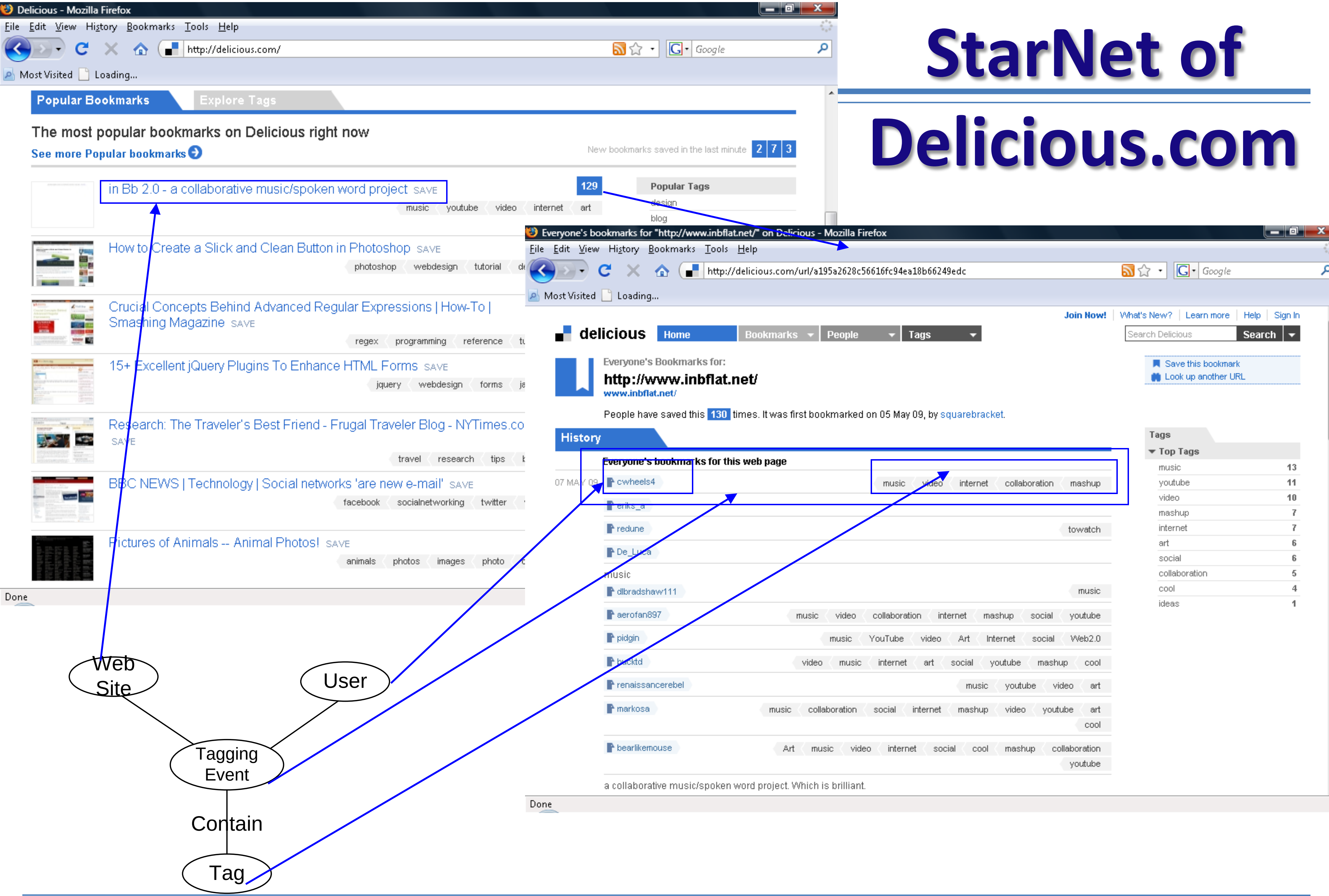


DBLP

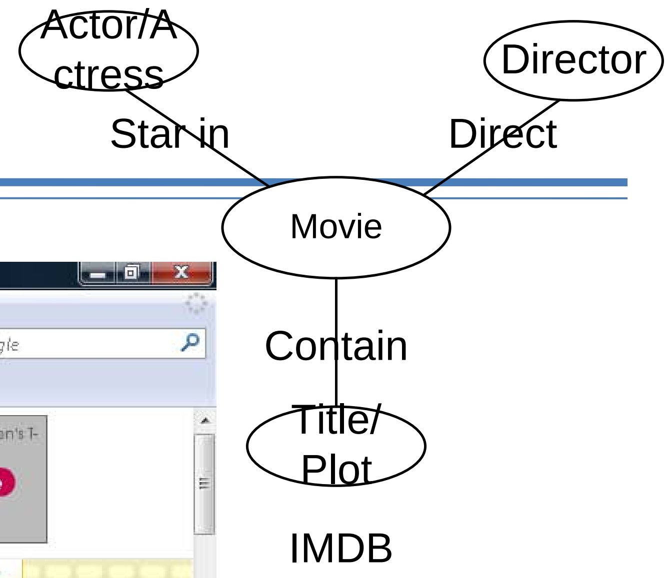
- Star Network Schema
  - **Center type**: Target type
    - E.g., a paper, a movie, a tagging event
    - A **center object** is a co-occurrence of a bag of different types of objects, which stands for **a multi-relation among different types of objects**
  - **Surrounding types**: Attribute (property) types
- **NetCluster**
  - Given a information network G, a net-cluster C contains two pieces of information:
    - Node set and link set as a sub-network of G
    - Membership indicator for each node x: P(x in C)

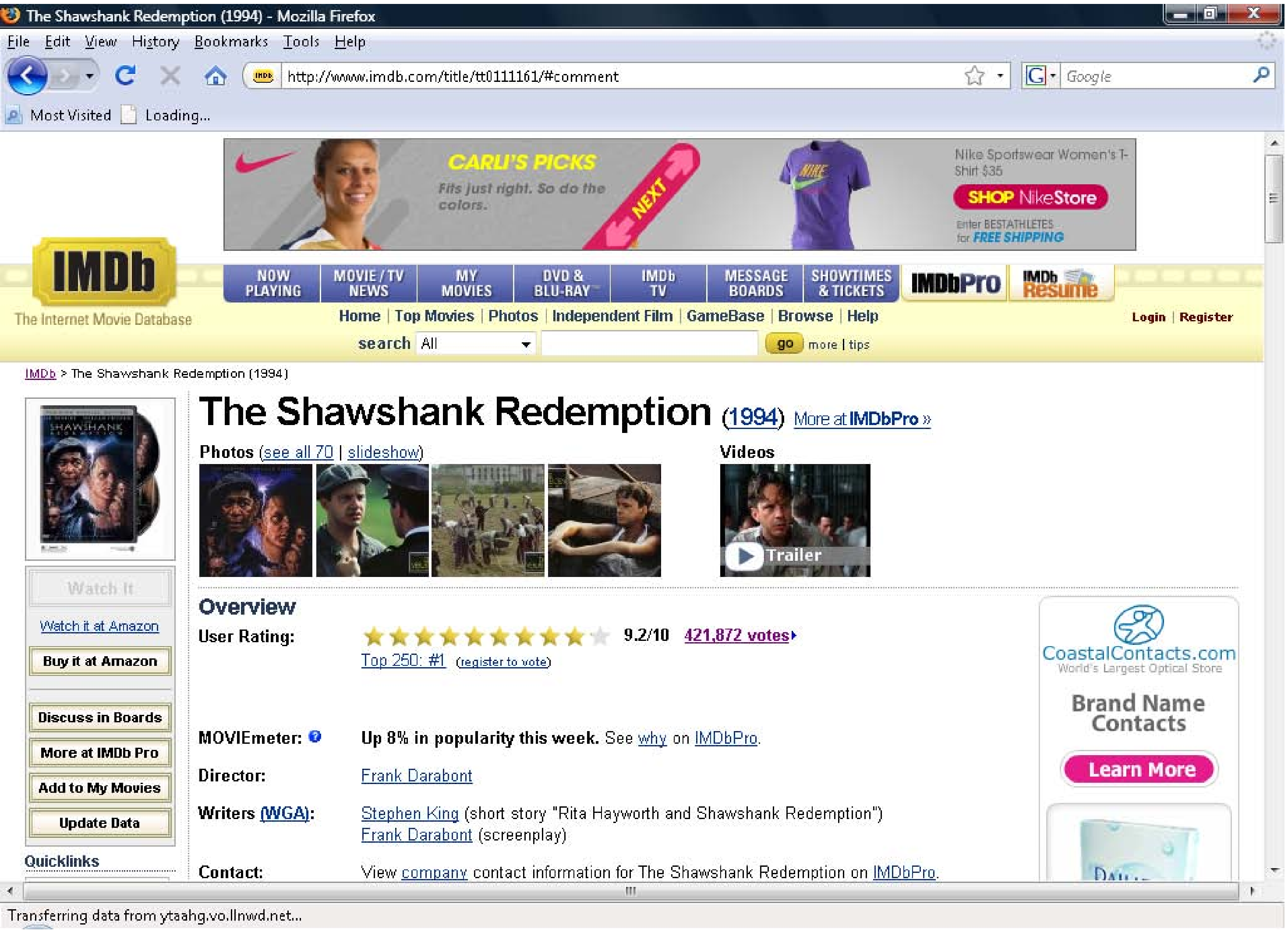

DBLP: Jiawei Han - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/h/Han:Jiawei.html
Most Visited
Loading...
dblp.uni-trier.de
Computer Science Bibliography
Universität Trier
Jiawei Han
Vis
List of publications from the DBLP Bibliography Server - FAQ
Facets and more with CompleteSearch
Coauthor Index - Ask others: ACM DL/Guide - CiteSeer - CSB - MetaPress - Google - MSN - Yahoo
author:jiaweihan:
Home Page
2009
361 EE Jiawei Han, Xifeng Yan, Philip S. Yu: Scalable OLAP and mining of information networks. EDBT 2009: 1159
360 EE Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, Tianyi Wu: RankClus: integrating clustering with ranking for heterogeneous information network analysis. EDBT 2009: 565-576
359 EE Bhavani M. Thuraisingham, Latifur Khan, Murat Kantarcioglu, Sonia Chib, Jiawei Han, Sang Son: Real-Time Knowledge Discovery and Dissemination for Intelligence Analysis. HICSS 2009: 1-12
358 EE Bolin Ding, David Lo, Jiawei Han, Siau-Cheng Khoo: Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database. ICDE 2009: 1024-1035
357 EE Xiaolei Li, Zhenhui Li, Jiawei Han, Jae-Gil Lee: Temporal Outlier Detection in Vehicle Traffic Data. ICDE 2009: 1319-1322
356 EE Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Multi-partition Multi-chunk Ensemble Technique to Classify Concept-Drifting Data Streams. PAKDD 2009: 363-375
2008
Refine by VENUE
International Journal of Software Engineering and Knowledge Engineering (IJSEKE) (1)
[all 1]
Refine by YEAR
1992 (1)
[all 1]
hide facet boxes
Done
Venue
Author
Publish
Write
Research Paper
Contain
Term

StarNet of Delicious.com



StartNet for IMDB





NetClus: Distinguishing Conferences

- AAAI 0.0022667 0.00899168 0.934024 0.0300042 0.0247133
- CIKM 0.150053 0.310172 0.00723807 0.444524 0.0880127
- CVPR 0.000163812 0.00763072 0.931496 0.0281342 0.032575
- ECIR 3.47023e-05 0.00712695 0.00657402 0.978391 0.00787288
- ECML 0.00077477 0.110922 0.814362 0.0579426 0.015999
- EDBT 0.573362 0.316033 0.00101442 0.0245591 0.0850319
- ICDE 0.529522 0.376542 0.00239152 0.0151113 0.0764334
- ICDM 0.000455028 0.778452 0.0566457 0.113184 0.0512633
- ICML 0.000309624 0.050078 0.878757 0.0622335 0.00862134
- IJCAI 0.00329816 0.0046758 0.94288 0.0303745 0.0187718
- KDD 0.00574223 0.797633 0.0617351 0.067681 0.0672086
- PAKDD 0.00111246 0.813473 0.0403105 0.0574755 0.0876289
- PKDD 5.39434e-05 0.760374 0.119608 0.052926 0.0670379
- PODS 0.78935 0.113751 0.013939 0.00277417 0.0801858
- SDM 0.000172953 0.841087 0.058316 0.0527081 0.0477156
- SIGIR 0.00600399 0.00280013 0.00275237 0.977783 0.0106604
- SIGMOD 0.689348 0.223122 0.0017703 0.00825455 0.0775055
- VLDB 0.701899 0.207428 0.00100012 0.0116966 0.0779764
- WSDM 0.00751654 0.269259 0.0260291 0.683646 0.0135497
- WWW 0.0771186 0.270635 0.029307 0.451857 0.171082

NetClus: Database System Cluster

database 0.0995511
databases 0.0708818
system 0.0678563
data 0.0214893
query 0.0133316
systems 0.0110413
queries 0.0090603
management 0.00850744
object 0.00837766
relational 0.0081175
processing 0.00745875
based 0.00736599
distributed 0.0068367
xml 0.00664958
oriented 0.00589557
design 0.00527672
web 0.00509167
information 0.0050518
model 0.00499396
efficient 0.00465707

VLDB 0.318495
SIGMOD Conf. 0.313903
ICDE 0.188746
PODS 0.107943
EDBT 0.0436849

author	rank score
Serge Abiteboul	0.0472111
Victor Vianu	0.0348510
Jerome Simeon	0.0324529
Michael J. Carey	0.0288872
Sophie Cluet	0.0282911
Daniela Florescu	0.0241411
Sihem Amer-Yahia	0.0240869
Donald Kossmann	0.0232118
Wenfei Fan	0.0225235
Tova Milo	0.0202201
...	...

Ranking authors in XML

Surajit Chaudhuri 0.00678065
Michael Stonebraker 0.00616469
Michael J. Carey 0.00545769
C. Mohan 0.00528346
David J. DeWitt 0.00491615
Hector Garcia-Molina 0.00453497
H. V. Jagadish 0.00434289
David B. Lomet 0.00397865
Raghu Ramakrishnan 0.0039278
Philip A. Bernstein 0.00376314
Joseph M. Hellerstein 0.00372064
Jeffrey F. Naughton 0.00363698
Yannis E. Ioannidis 0.00359853
Jennifer Widom 0.00351929
Per-Ake Larson 0.00334911
Rakesh Agrawal 0.00328274
Dan Suciu 0.00309047
Michael J. Franklin 0.00304099
Umeshwar Dayal 0.00290143
Abraham Silberschatz 0.00278185

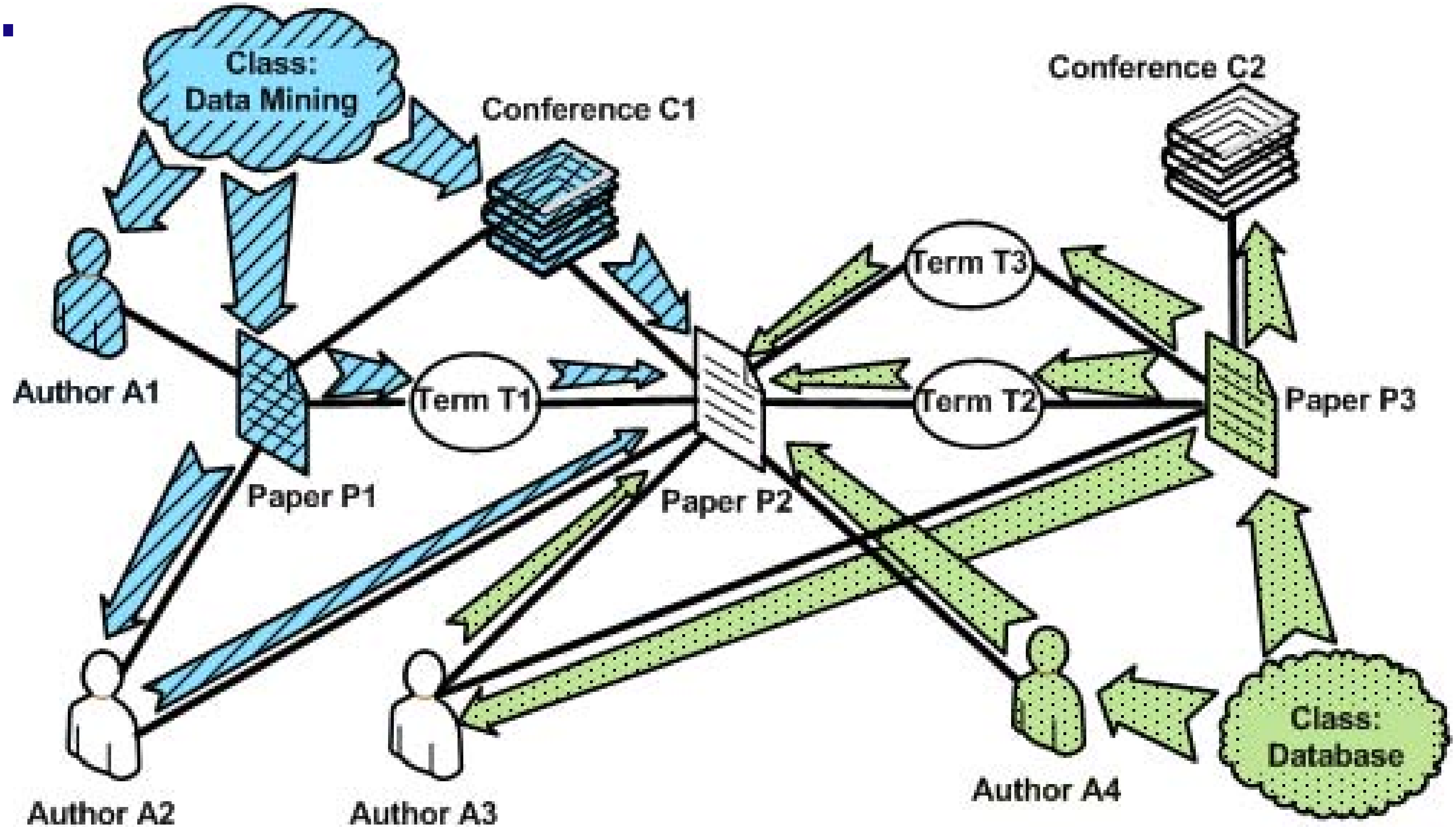
Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

From RankClus to GNetMine & RankClass

- **RankClus [EDBT'09]: Clustering and ranking working together**
  - No training, no available class labels, no expert knowledge
- **GNetMine [PKDD'10]: Incorp. prior knowledge in networks**
  - Classification in heterog. networks, but objects treated equally
- **RankClass [KDD'11 sub]: Integration of ranking and classification in heterogeneous network analysis**
  - Ranking: informative understanding & summary of each class
  - Class membership is critical information when ranking objects
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class

Classification: Knowledge Propagation

- 

GNetMine: Methodology

- M. Ji, et al., "Graph Regularized Transductive Classification on Heterogeneous Information Networks", ECMLPKDD'10
- Classifying networked data: a ***knowledge propagation*** process
- Information is propagated from labeled objects to unlabeled ones through links until a stationary state is achieved
- A novel **graph-based regularization framework** to address the classification problem on heterogeneous information networks
- Respect the link type differences by preserving consistency over each relation graph corresponding to each type of links separately
  - Mathematical intuition: Consistency assumption
  - The confidence (f)of two objects (x_{ip} and x_{jq}) belonging to class k should be similar if $x_{ip} \leftrightarrow x_{jq}$ ($R_{ij,pq} > 0$)
  - f should be similar to the given ground truth

GNetMine: Graph-Based Regularization

- Minimize the objective function

$$J(\boldsymbol{f}_1^{(k)}, \ldots, \boldsymbol{f}_m^{(k)})$$

User preference: how much do you value this relationship / ground truth?

$$= \sum_{i,j=1}^{m} \lambda_{ij} \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} R_{ij,pq} \left(\frac{1}{\sqrt{D_{ij,pp}}} f_{ip}^{(k)} - \frac{1}{\sqrt{D_{ji,qq}}} f_{jq}^{(k)} \right)^2$$

$$+ \sum_{i=1}^{m} \alpha_i (\boldsymbol{f}_i^{(k)} - \mathbf{y}_i^{(k)})^T (\boldsymbol{f}_i^{(k)} - \mathbf{y}_i^{(k)})$$

Smoothness constraints: objects linked together should share *similar* estimations of confidence belonging to class *k*

Normalization term applied to each type of link separately: reduce the impact of popularity of nodes

Confidence estimation on labeled data and their pre-given labels should be similar

Experiments on DBLP

- Class: Four research areas (communities)
  - Database, data mining, AI, information retrieval
- Four types of objects
  - Paper (14376), Conf. (20), Author (14475), Term (8920)
- Three types of relations
  - Paper-conf., paper-author, paper-term
- Algorithms for comparison
  - Learning with Local and Global Consistency (LLGC) [Zhou et al. NIPS 2003] – also the homogeneous version of our method
  - Weighted-vote Relational Neighbor classifier (wvRN) [Macskassy et al. JMLR 2007]
  - Network-only Link-based Classification (nLB) [Lu et al. ICML 2003, Macskassy et al. JMLR 2007]

Performance Study on the DBLP Data Set

Table 3: Comparison of classification accuracy on authors (%)

$(a\%, p\%)$ of authors and papers labeled	nLB (A-A)	nLB (A-C-P-T)	wvRN (A-A)	wvRN (A-C-P-T)	LLGC (A-A)	LLGC (A-C-P-T)	GNetMine (A-C-P-T)	RankClass (A-C-P-T)
(0.1%, 0.1%)	25.4	26.0	40.8	34.1	41.4	61.3	82.9	**83.9**
(0.2%, 0.2%)	28.3	26.0	46.0	41.2	44.7	62.2	83.4	**85.6**
(0.3%, 0.3%)	28.4	27.4	48.6	42.5	48.8	65.7	86.7	**88.3**
(0.4%, 0.4%)	30.7	26.7	46.3	45.6	48.7	66.0	87.2	**88.8**
(0.5%, 0.5%)	29.8	27.3	49.0	51.4	50.6	68.9	87.5	**89.2**
average	28.5	26.7	46.3	43.0	46.8	64.8	85.5	**87.2**

Table 4: Comparison of classification accuracy on papers (%)

$(a\%, p\%)$ of authors and papers labeled	nLB (P-P)	nLB (A-C-P-T)	wvRN (P-P)	wvRN (A-C-P-T)	LLGC (P-P)	LLGC (A-C-P-T)	GNetMine (A-C-P-T)	RankClass (A-C-P-T)
(0.1%, 0.1%)	49.8	31.5	62.0	42.0	67.2	62.7	**79.2**	77.7
(0.2%, 0.2%)	73.1	40.3	71.7	49.7	72.8	65.5	**83.5**	83.0
(0.3%, 0.3%)	77.9	35.4	77.9	54.3	76.8	66.6	83.2	**83.6**
(0.4%, 0.4%)	79.1	38.6	78.1	54.4	77.9	70.5	83.7	**84.7**
(0.5%, 0.5%)	80.7	39.3	77.9	53.5	79.0	73.5	84.1	**84.8**
average	72.1	37.0	73.5	50.8	74.7	67.8	82.7	**82.8**

Table 5: Comparison of classification accuracy on conferences (%)

$(a\%, p\%)$ of authors and papers labeled	nLB (A-C-P-T)	wvRN (A-C-P-T)	LLGC (A-C-P-T)	GNetMine (A-C-P-T)	RankClass (A-C-P-T)
(0.1%, 0.1%)	25.5	43.5	79.0	81.0	**84.5**
(0.2%, 0.2%)	22.5	56.0	83.5	85.0	**85.5**
(0.3%, 0.3%)	25.0	59.0	**87.0**	**87.0**	**87.0**
(0.4%, 0.4%)	25.0	57.0	86.5	89.5	**90.5**
(0.5%, 0.5%)	25.0	68.0	90.0	94.0	**95.0**
average	24.6	56.7	85.2	87.3	**88.5**

Experiments with Very Small Training Set

- DBLP: 4-fields data set (DB, DM, AI, IR) forming a heterog. info. network
- Rank objects within each class (with extremely limited label information)
- Obtain High classification accuracy and excellent rankings within each class

	Database	Data Mining	AI	IR
Top-5 ranked conferences	VLDB	KDD	IJCAI	SIGIR
	SIGMOD	SDM	AAAI	ECIR
	ICDE	ICDM	ICML	CIKM
	PODS	PKDD	CVPR	WWW
	EDBT	PAKDD	ECML	WSDM
Top-5 ranked terms	data	mining	learning	retrieval
	database	data	knowledge	information
	query	clustering	reasoning	web
	system	classification	logic	search
	xml	frequent	cognition	text

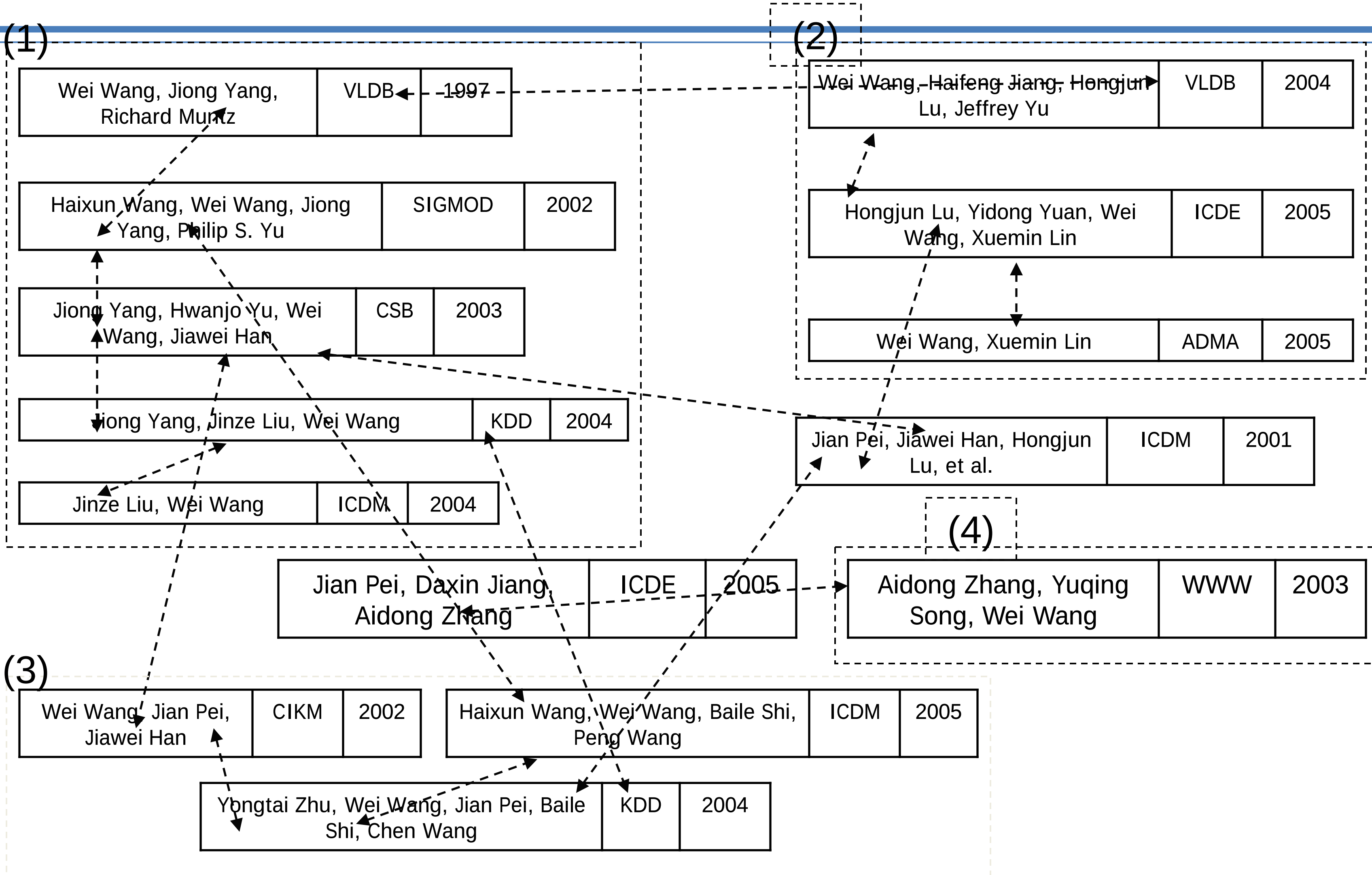
Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Data Cleaning by Link Analysis

- Object reconciliation vs. object distinction as data cleaning tasks
- Link analysis may take advantages of redundancy and make facilitate entity cross-checking and validation
- Object distinction: Different people/objects do share names
  - In AllMusic.com, 72 songs and 3 albums named “Forgotten” or “The Forgotten”
  - In DBLP, 141 papers are written by at least 14 “Wei Wang”
- New challenges of object distinction:
  - Textual similarity cannot be used
- Distinct: Object distinction by information network analysis
  - X. Yin, J. Han, and P. S. Yu, “Object Distinction: Distinguishing Objects with Identical Names by Link Analysis”, ICDE'07

Entity Distinction: The "Wei Wang" Challenge in DBLP



(1) Wei Wang at UNC
(2) Wei Wang at UNSW, Australia
(3) Wei Wang at Fudan Univ., China
(4) Wei Wang at SUNY Buffalo

DISTINCT: Distinguish Objects w. Identical Names

- Measure similarity between references
  - Link-based similarity: Linkages between references
    - References to the same object are more likely to be connected (Using random walk probability)
  - Neighborhood similarity
    - Neighbor tuples of each reference can indicate similarity between their contexts
- **Self-boosting: Training using the "same" bulky data set**
- Reference-based clustering
  - Group references according to their similarities
  - *Use average neighborhood similarity* and *collective random walk probability*

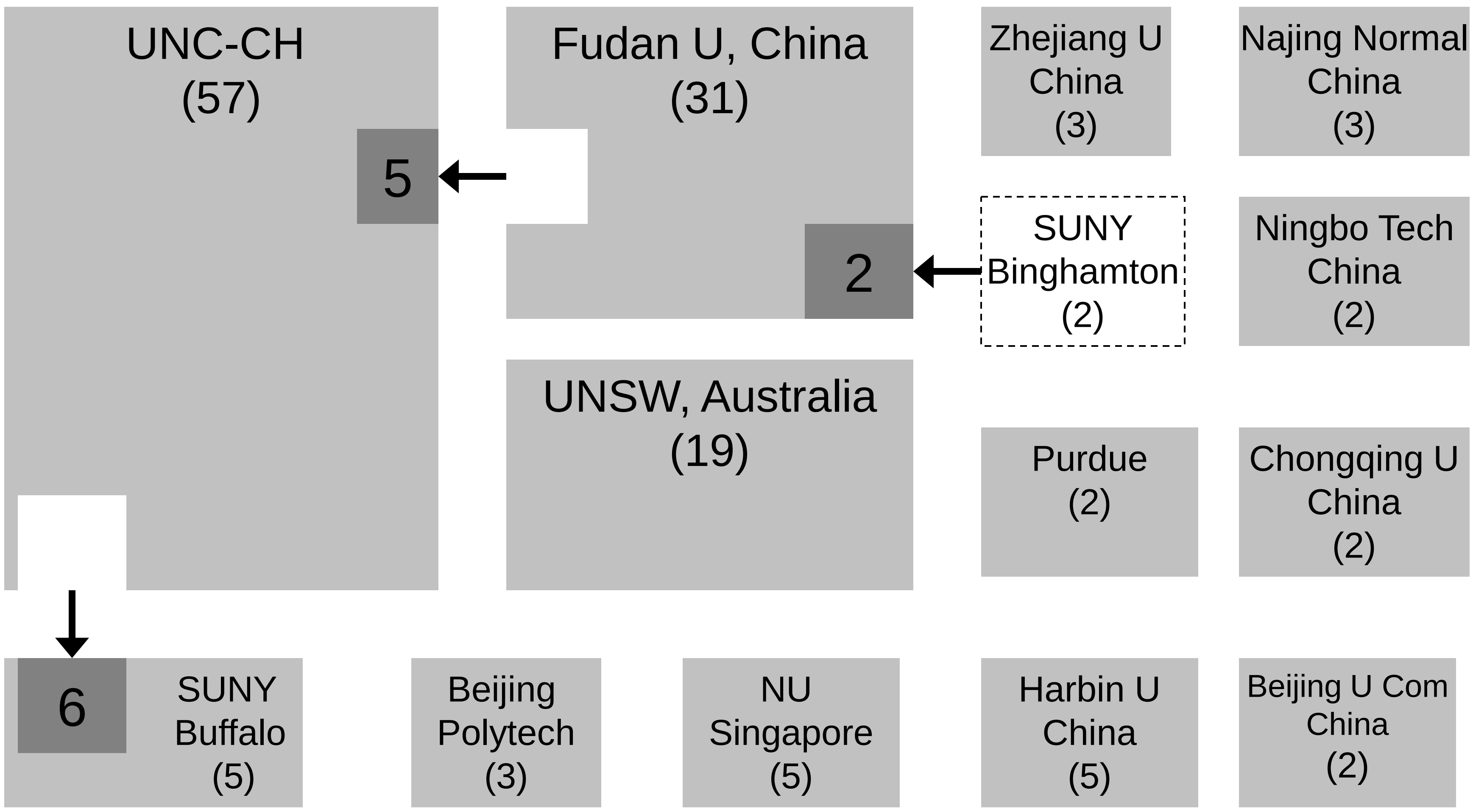
Training with the "Same" Data Set

- Build a training set automatically
  - Select distinct names, e.g., Johannes Gehrke
  - The collaboration behavior within the same community share some similarity
  - Training parameters using a typical and large set of "unambiguous" examples
- Use SVM to learn a model for combining different join paths
  - Each join path is used as two attributes (with link-based similarity and neighborhood similarity)
  - The model is a weighted sum of all attributes

Real Cases: DBLP Popular Names

Name	*Num_authors*	*Num_refs*	*accuracy*	*precision*	*recall*	*f-measure*
Hui Fang	3	9	1.0	1.0	1.0	1.0
Ajay Gupta	4	16	1.0	1.0	1.0	1.0
Joseph Hellerstein	2	151	0.81	1.0	0.81	0.895
Rakesh Kumar	2	36	1.0	1.0	1.0	1.0
Michael Wagner	5	29	0.395	1.0	0.395	0.566
Bing Liu	6	89	0.825	1.0	0.825	0.904
Jim Smith	3	19	0.829	0.888	0.926	0.906
Lei Wang	13	55	0.863	0.92	0.932	0.926
Wei Wang	14	141	0.716	0.855	0.814	0.834
Bin Yu	5	44	0.658	1.0	0.658	0.794
Average			0.81	0.966	0.836	0.883

Distinguishing Different "Wei Wang"s



Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Truth Validation by Info. Network Analysis

- The trustworthiness problem of the web (according to a survey):
  - 54% of Internet users trust news web sites most of time
  - 26% for web sites that sell products
  - 12% for blogs
- TruthFinder: Truth discovery on the Web by link analysis
  - Among multiple conflict results, can we automatically identify which one is likely the true fact?
- Veracity (conformity to truth):
  - Given conflicting information provided by multiple web sites, how to discover the true fact about each object?
- X. Yin, J. Han, P. S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web", TKDE'08

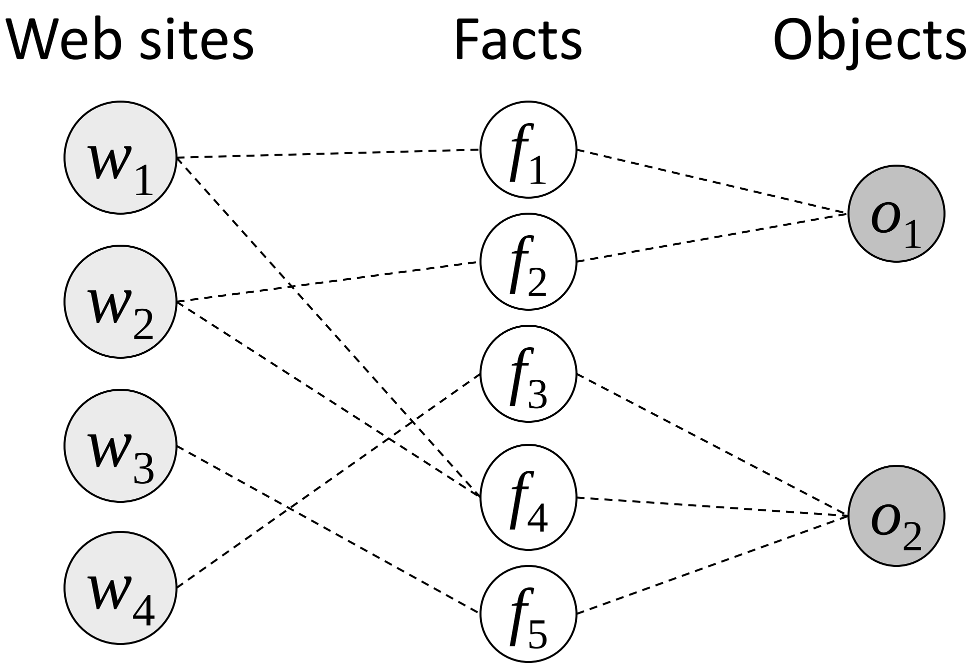
Conflicting Information on the Web

- Different websites often provide conflicting info. on a subject, e.g., Authors of *"Rapid Contextual Design"*

Online Store	***Authors***
Powell's books	Holtzblatt, Karen
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Cornwall books	Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood
Mellon's books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley

Our Setting: Info. Network Analysis

- Each object has a set of ***conflictive*** facts
  - E.g., different author names for a book
- And each web site provides some facts
- How to find the true fact for each object?

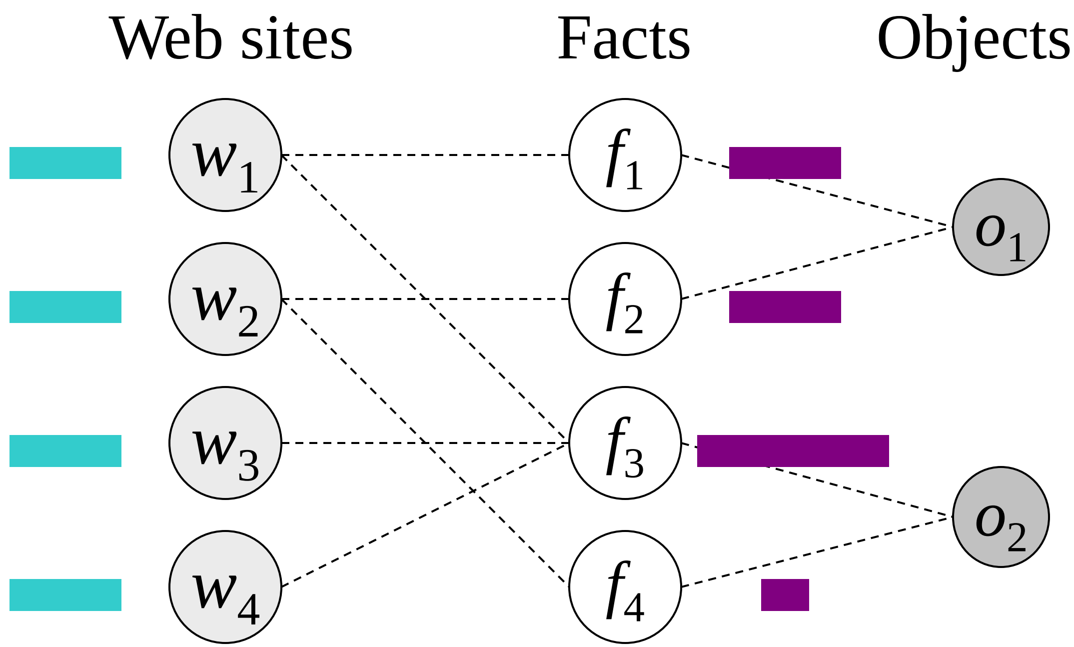


Basic Heuristics for Problem Solving

1. There is usually **only one true fact** for a property of an object
2. This true fact **appears to be the same or similar** on different web sites
   - E.g., “Jennifer Widom” vs. “J. Widom”
3. **The false facts on different web sites are less likely to be the same or similar**
   - False facts are often introduced by random factors
4. **A web site that provides mostly true facts for many objects will likely provide true facts for other objects**

Inference on Trustworthness

- Inference of web site trustworthiness & fact confidence



- True facts and trustable web sites will become apparent after some iterations

TruthFinder: Iterative Mutual Enhancement

- ***Confidence of facts*** ↔ ***Trustworthiness of web info providers***
  - A fact has *high confidence* if it is provided by (many) trustworthy web sites
  - A web info provider is *trustworthy* if it provides many facts with high confidence
- TruthFinder mechanism:
  - Initially, each web site is equally trustworthy
  - Based on the above four heuristics, infer fact confidence from web site trustworthiness, and then backwards
  - Repeat until achieving stable state

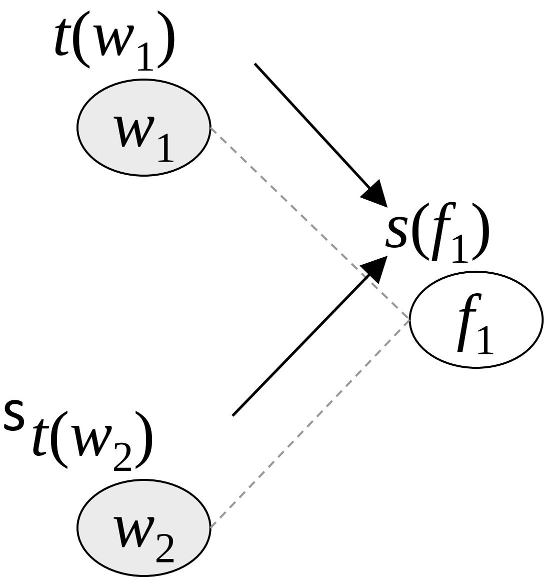
Computational Model: t(w) and s(f)

- **The trustworthiness of a web site *w*: *t*(*w*)**
  - Average confidence of facts it provides

$$t(w) = \frac{\sum_{f \in F(w)} s(f)}{|F(w)|}$$

Sum of fact confidence

Set of facts provided by w



- **The confidence of a fact *f*: *s*(*f*)**
  - One minus the probability that all web sites providing *f* are wrong

Probability that w is wrong

$$s(f) = 1 - \prod_{w \in W(f)} (1 - t(w))$$

Set of websites providing f

Experiments: Finding Truth of Facts

- Determining authors of books
  - Dataset contains 1265 books listed on abebooks.com
  - We analyze 100 random books (using book images)

Case	***Voting***	***TruthFinder***	***Barnes & Noble***
Correct	71	85	64
Miss author(s)	12	2	4
Incomplete names	18	5	6
Wrong first/middle names	1	1	3
Has redundant names	0	2	23
Add incorrect names	1	5	5
No information	0	0	2

Experiments: Trustable Info Providers

- Finding trustworthy information sources
  - Most trustworthy bookstores found by TruthFinder vs. Top ranked bookstores by Google (query "bookstore")

TruthFinder

Bookstore	***trustworthiness***	***#book***	***Accuracy***
TheSaintBookstore	0.971	28	0.959
MildredsBooks	0.969	10	1.0
Alphacraze.com	0.968	13	0.947

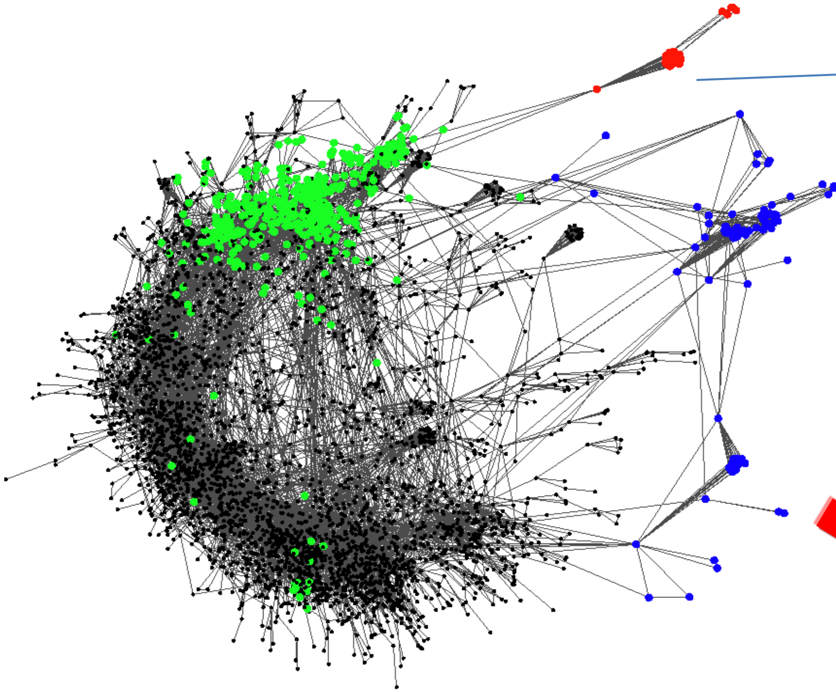
Google

Bookstore	***Google rank***	***#book***	***Accuracy***
Barnes & Noble	1	97	0.865
Powell's books	3	42	0.654

Outline

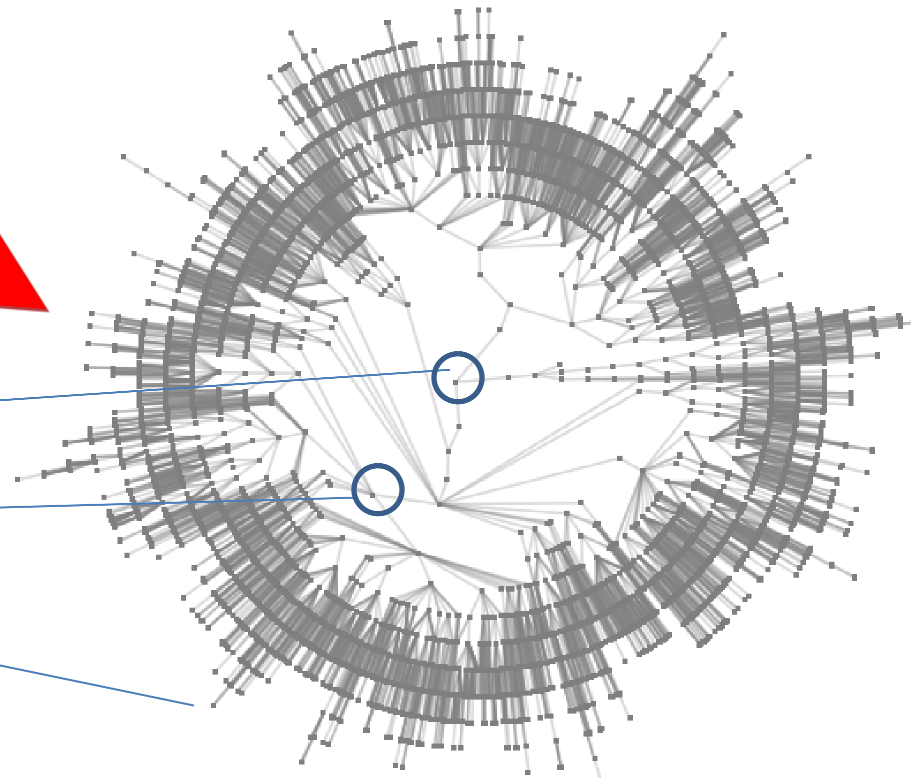
- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Role Discovery in Network: Why Does It Matter?



A “dirty” Information Network
(imaginary)

Cleaned/Inferred
Adversarial Network

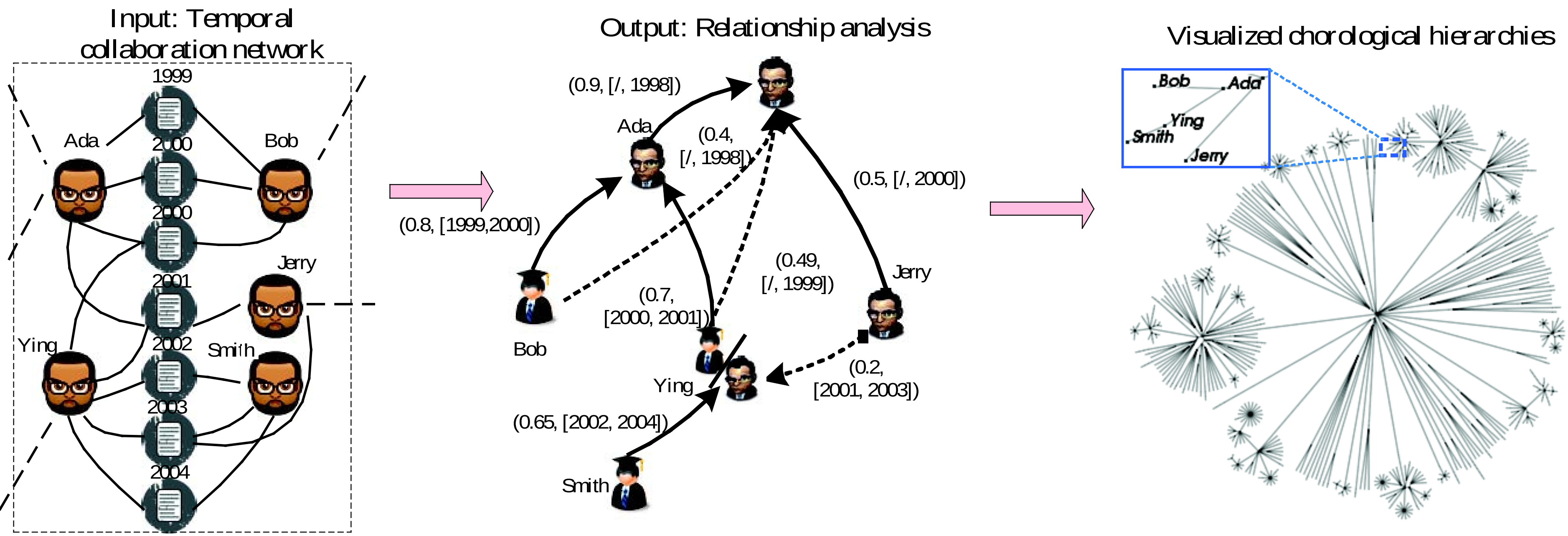


Automatically infer

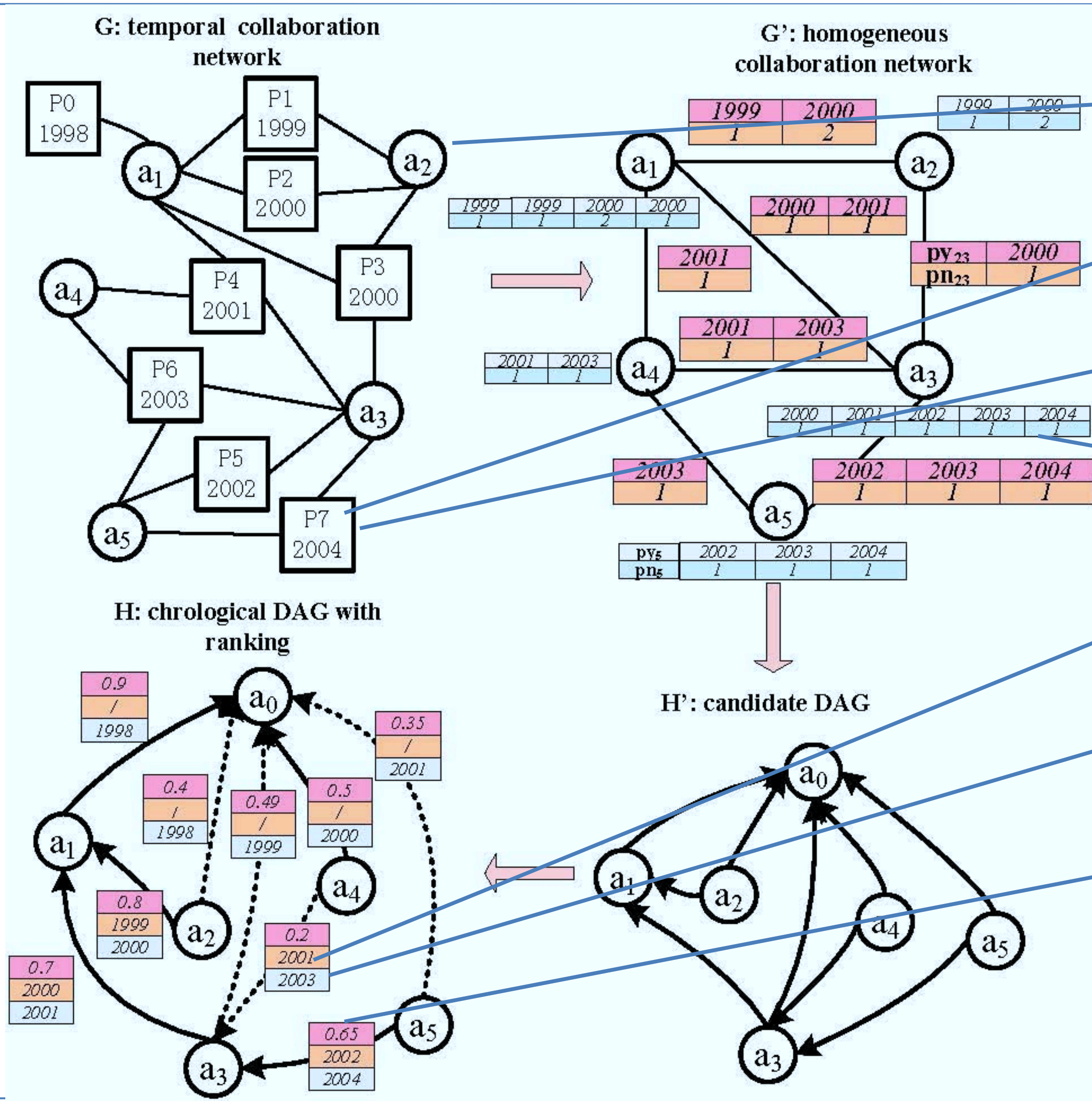
- Chief
- Cell Lead
- Insurgent

Discovery of Advisor-Advisee Relationships in DBLP Network

- Input: DBLP research publication network
- Output: Potential advising relationship and its ranking (r, [st, ed])
- C. Wang, J. Han, et al., *"Mining Advisor-Advisee Relationships from Research Publication Networks"*, KDD 2010

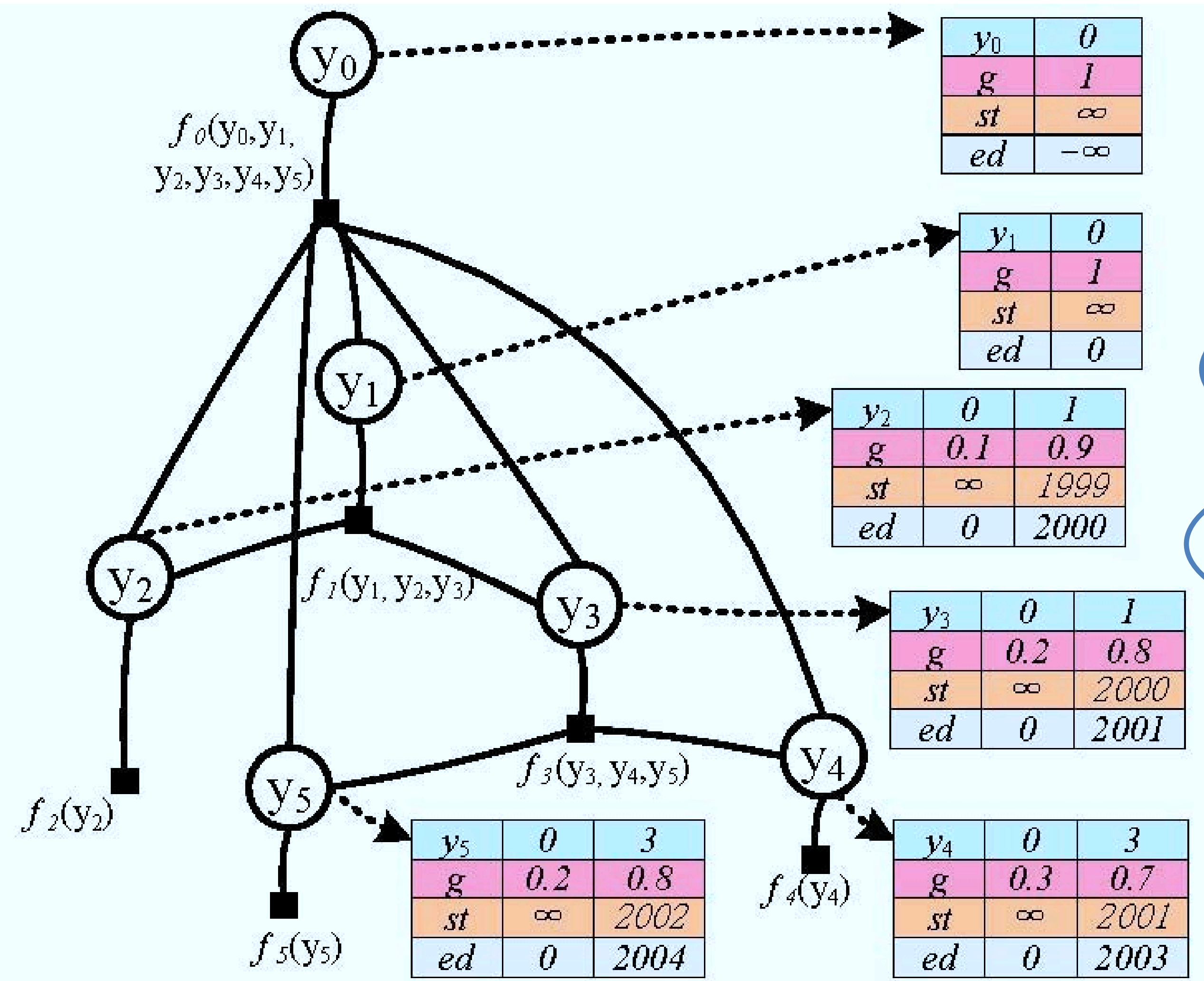


Overall Framework



- a_i: author i
- p_j: paper j
- py: paper year
- pn: paper#
- st_{i,y_i}: starting time
- ed_{i,y_i}: ending time
- r_{i,y_i}: ranking score

Time-Constrained Probabilistic Factor Graph (TPFG)



- y_x: a_x's advisor
- $st_{x,yx}$: starting time
  $ed_{x,yx}$: ending time
- $g(y_x, st_x, ed_x)$ is predefined 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

Experiment Results

- DBLP data: 654, 628 authors, 1076,946 publications, years provided
- Labeled data: MathGealogy Project; AI Gealogy Project; 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

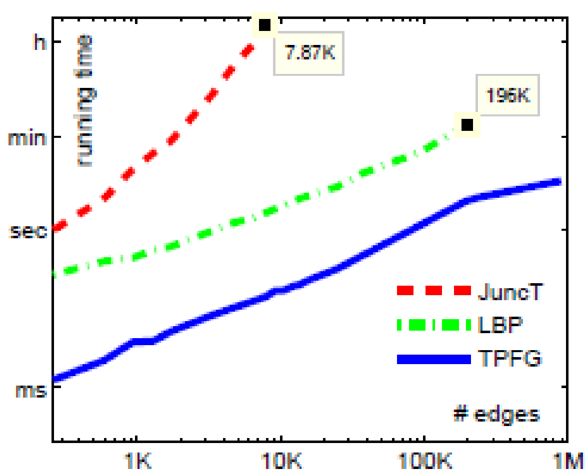
Supervised learning

Empirical parameter

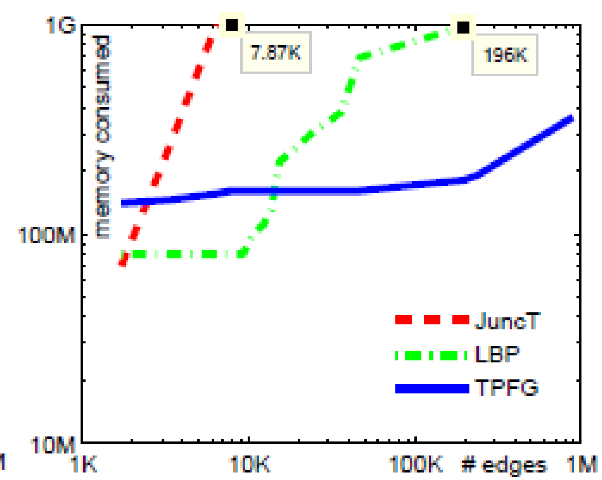
optimized parameter

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

Outline

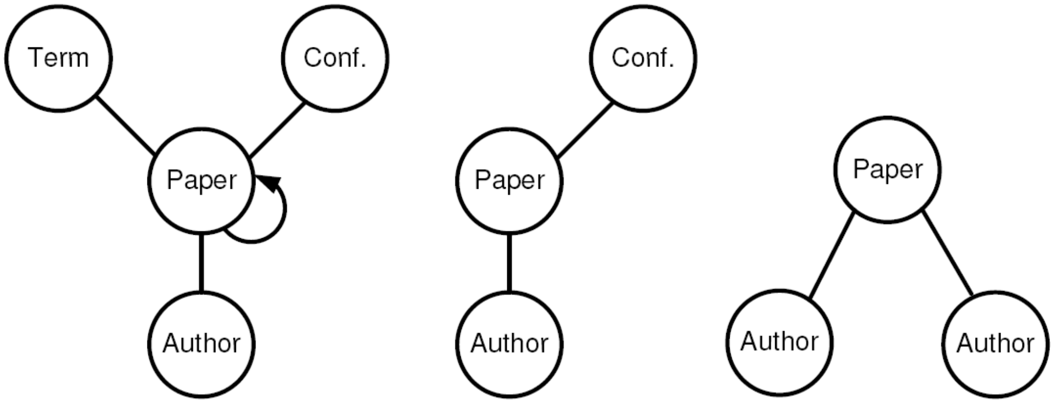
- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Finding Similar Objects in Networks

- Y. Sun et al, "**PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks**", **VLDB'11**
- Search top-k similar objects of the same type in a network
  - Find researchers most similar with "Christos Faloutsos"?
- Feature space
  - Traditional data: attributes denoted as numerical (or categorical) value set or vector
  - Network data: A relation sequence called "**meta path**"
- Measure defined on the feature space
  - Cosine, Euclidean distance, Jaccard coefficient, etc.
  - **PathSim**: $s(i, j) = 2M_P(i, j)/(M_P(i, i) + M_P(j, j))$
    - $M_P(i, j)$: Matrix corresp. to a meta-path from object i to j

Meta-Path for DBLP Queries

- Meta-Path: A path of InfoNet attributes, e.g., APC, APA
- Who are most similar to Christos Faloutsos?



(a) InfoNet Schema (b) Path Schema: APC/CPA (c) Path Schema: APA

(a) Path: APA

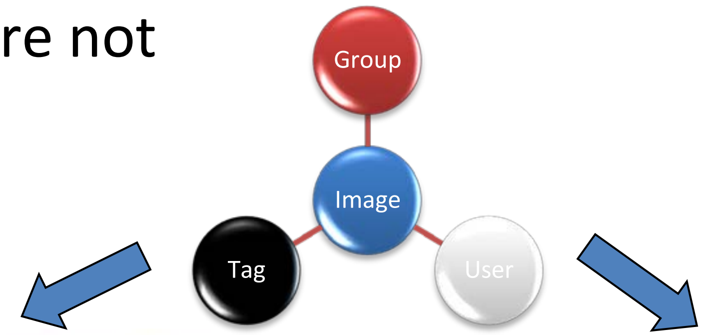
Rank	Author	Score
1	Christos Faloutsos	1
2	Spiros Papadimitriou	0.127
3	Jimeng Sun	0.12
4	Jia-Yu Pan	0.114
5	Agma J. M. Traina	0.110
6	Jure Leskovec	0.096
7	Caetano Traina Jr.	0.096
8	Hanghang Tong	0.091
9	Deepayan Chakrabarti	0.083
10	Flip Korn	0.053

(b) Path: $APCPA$

Rank	Author	Score
1	Christos Faloutsos	1
2	Jiawei Han	0.842
3	Rakesh Agrawal	0.838
4	Jian Pei	0.8
5	Charu C. Aggarwal	0.739
6	H. V. Jagadish	0.705
7	Raghu Ramakrishnan	0.697
8	Nick Koudas	0.689
9	Surajit Chaudhuri	0.677
10	Divesh Srivastava	0.661

Flickr: Which Pictures Are Most Similar?

- Some path schema leads to similarity closer to human intuition
- But some others are not



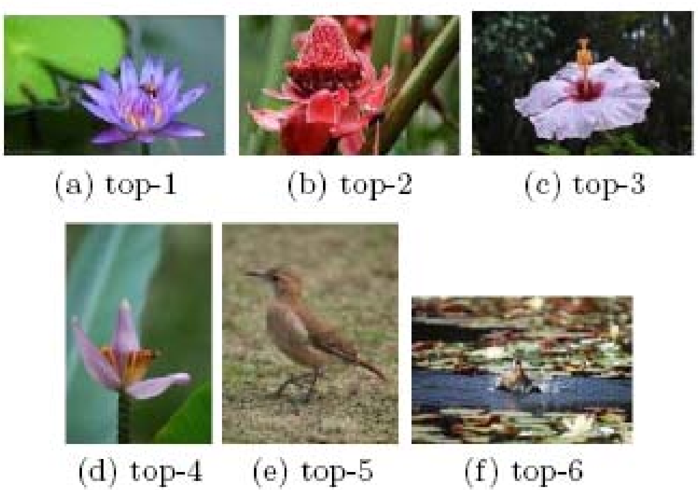


Figure 5: Top-6 images in Flickr network under path schema *ITI*

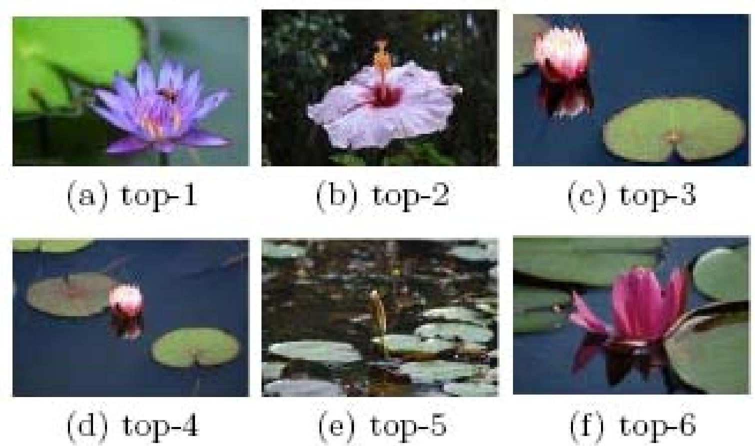


Figure 6: Top-6 images in Flickr network under path schema *ITIGITI*

Outline

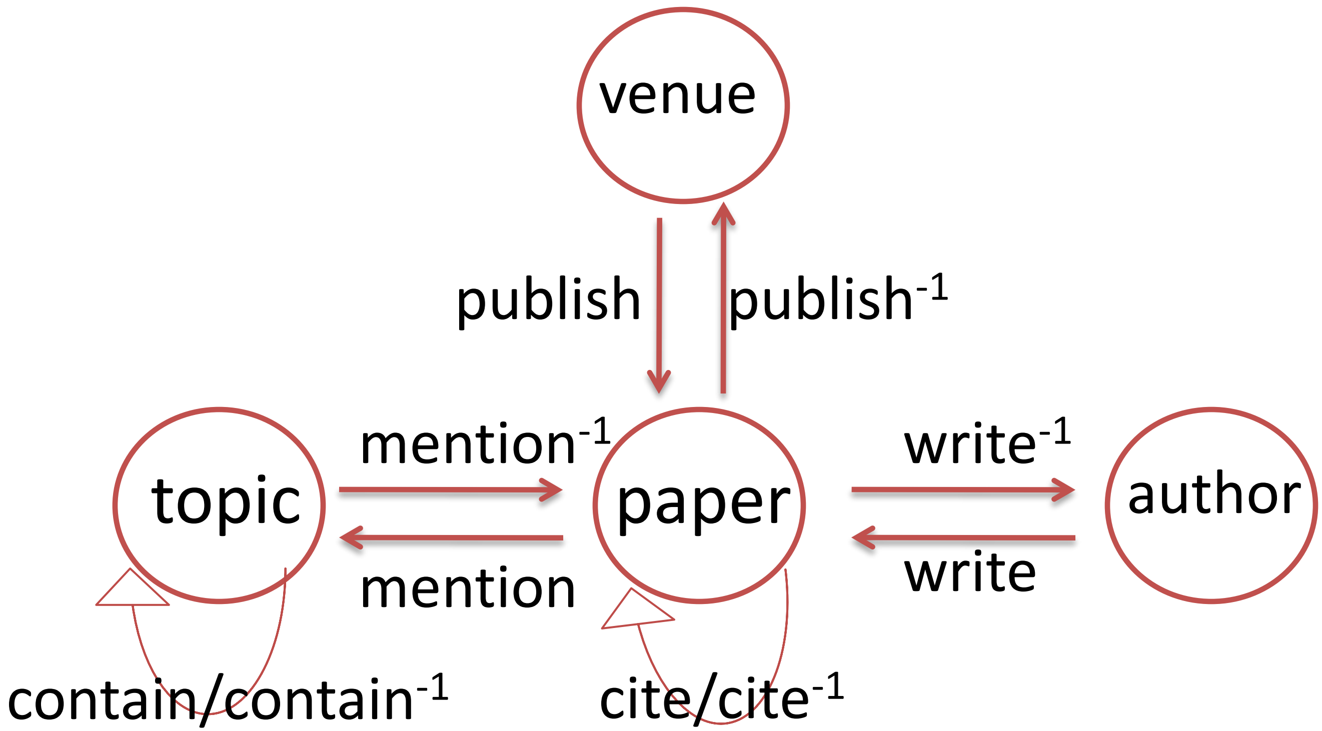
- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Relationship Prediction in Heterogeneous Info Networks

- Why Prediction of Co-Author Relationship in DBLP?
  - Prediction of relationships between different types of nodes in heterogeneous networks, e.g., what papers should he writes?
- Traditional link prediction
  - Studies on homogeneous networks
  - E.g., co-author networks in DBLP, friendship networks (e.g., facebook)
- Relationship prediction
  - Study the roles of topological features in heterogeneous networks in predicting the co-author relationship building
- Y. Sun, et al., "Co-Author Relationship Prediction in Heterog. Bibliographic Networks", Int. Conf. on Advances in Social Network Analysis and Mining (ASONAM'11), July 2011

Guidance: Meta Path in Bibliographic Network

- Schema of object type relationships in a bibliographic Networks
- Underneath structure: A directed graph
- Relationship prediction: meta path-guided prediction



Meta Path-Guided Relationship Prediction

- Meta path relationships among similar typed links share similar semantics and are comparable and inferable
- Relationship across different typed links are not directly comparable but their collective behavior will help predicting particular relationships
- Example: Co-author prediction: Predict whether two existing authors will build a relationship in the future following the relation encoded by a meta path:

$$A \xrightarrow{write} P \xrightarrow{write^{-1}} A$$

  - Using topological features also encoded by meta paths:
    - E.g., citation relations between authors

$$A \xrightarrow{write} P \xrightarrow{cite} P \xrightarrow{write^{-1}} A.$$

Meta-Paths & Their Prediction Power

- List all the meta-paths in bibliographic network up to length 4

Meta Path	Semantic Meaning of the Relation
$A - P - A$	a_i and a_j are coauthors (the target relation)
$A - P \rightarrow P - A$	a_i cites a_j
$A - P \leftarrow P - A$	a_i is cited by a_j
$A - P - V - P - A$	a_i and a_j publish in the same venues
$A - P - A - P - A$	a_i and a_j are co-authors of the same authors
$A - P - T - P - A$	a_i and a_j write the same topics
$A - P \rightarrow P \rightarrow P - A$	a_i cites papers that cite a_j
$A - P \leftarrow P \leftarrow P - A$	a_i is cited by papers that are cited by a_j
$A - P \rightarrow P \leftarrow P - A$	a_i and a_j cite the same papers
$A - P \leftarrow P \rightarrow P - A$	a_i and a_j are cited by the same papers

- Investigate their respective power for coauthor relationship prediction
  - Which meta-path has more prediction power?
  - How to combine them to achieve the best quality of prediction

Selection among Competitive Measures

4 measures that defines a relationship *R* encoded by a meta path

- Path Count: #path instances between authors following *R*

$$PC_R(a_i, a_j)$$

- Normalized Path Count: Normalize path count following *R* by the "degree" of authors

$$NPC_R(a_i, a_j) = \frac{PC_R(a_i, a_j) + PC_{R^{-1}}(a_j, a_i)}{PC_R(a_i, \cdot) + PC_R(\cdot, a_j)}$$

- Random Walk: Consider one way random walk following *R*

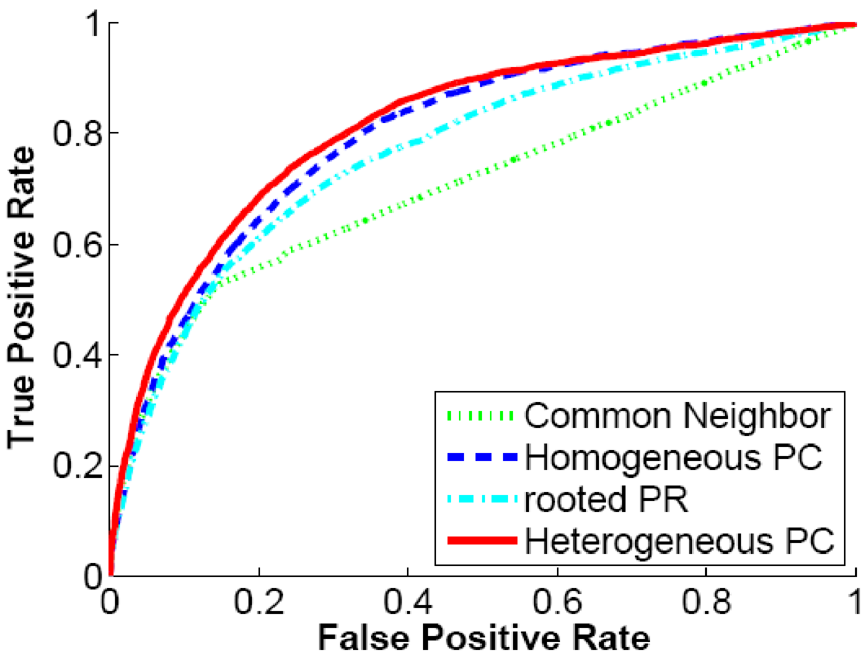
$$RW_R(a_i, a_j) = \frac{PC_R(a_i, a_j)}{PC_R(a_i, \cdot)}$$

- Symmetric Random Walk: Consider random walk in both directions

$$SRW_R(a_i, a_j) = RW_R(a_i, a_j) + RW_{R^{-1}}(a_j, a_i)$$

Performance Comparison: Homogeneous vs. Heterogeneous Topological Features

- Homogeneous features
  - Only consider co-author sub-network (common neighbor; rooted PageRank)
  - Mix all types together (homogeneous path count)
- Heterogeneous feature
  - Heterogeneous path count



Dataset	Topological features	Accuracy	AUC
HP2hop	common neighbor	0.6053	0.6537
	homogeneous PC	0.6433	0.7098
	heterogeneous PC	**0.6545**	**0.7230**
HP3hop	common neighbor	0.6589	0.7078
	homogeneous PC	0.6990	0.7998
	rooted PageRank	0.6433	0.7098
	heterogeneous PC	**0.7173**	**0.8158**
LP2hop	common neighbor	0.5995	0.6415
	homogeneous PC	0.6154	0.6868
	heterogeneous PC	**0.6300**	**0.6935**
LP3hop	common neighbor	0.6804	0.7195
	homogeneous PC	0.6901	0.7883
	heterogeneous PC	**0.7147**	**0.8046**

Notation: *HP2hop*: highly productive source authors with 2-hops reaching target authors

Case Study in CS Bibliographic Network

- The learned significance for each meta path under measure "normalized path count" for HP-3hop dataset

Meta Path	p-value	significance level[1]
$A - P \rightarrow P - A$	0.0378	**
$A - P \leftarrow P - A$	0.0077	***
$A - P - V - P - A$	1.2974e-174	****
$A - P - A - P - A$	1.1484e-126	****
$A - P - T - P - A$	3.4867e-51	****
$A - P \rightarrow P \rightarrow P - A$	0.7459	
$A - P \leftarrow P \leftarrow P - A$	0.0647	*
$A - P \rightarrow P \leftarrow P - A$	9.7641e-11	****
$A - P \leftarrow P \rightarrow P - A$	0.0966	*

[1] *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

Case Study: Predicting Concrete Co-Authors

- High quality predictive power for such a difficult task

QUERY AUTHOR SUMMARIZATION

Query author	# Candidates	# True relationships
Jiawei Han	11934	36
Christos Faloutsos	12945	45
Charu Aggarwal	5166	12
Jian Pei	4809	42
Xifeng Yan	1617	8

TOP-5 PREDICTED CO-AUTHORS FOR JIAN PEI IN 2003-2009

Rank	Hybrid heterogeneous features	# Shared authors
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanara
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

* Authors in bold format are the true new co-authors of Jian in the t period 2003-2009.

- Using data in $T0$ =[1989; 1995] and $T1$ = [1996; 2002]
- Predict new coauthor relationship in $T2$ = [2003; 2009]

TOP-10 PREDICTED CO-AUTHORS FOR JIAWEI HAN

Rank	Hybrid features	# Shared authors
1	**Hans-Peter Kriegel**	Elisa Bertino
2	Christos Faloutsos	Sushil Jajodia
3	Divesh Srivastava	Hector Garcia-Molina
4	H. V. Jagadish	**Hans-Peter Kriegel**
5	Bing Liu[1]	Christos Faloutsos
6	Johannes Gehrke	Divyakant Agrawal
7	George Karypis	Elke A. Rundensteiner
8	**Charu C. Aggarwal**	Amr El Abbadi
9	Mohammed Javeed Zaki	Krithi Ramamritham
10	Wynne Hsu	Stefano Ceri

[1] Although not included in the time interval T_2, Bing Liu co-authored with Jiawei in Year 2010.

Recall@50 COMPARISON

Query author	Hybrid Features	Random	# Shared authors
Jiawei Han	0.1111	0.0042	0.0833
Christos Faloutsos	0.0889	0.0039	0.1111
Charu Aggarwal	0.4167	0.0097	0.3333
Jian Pei	0.2619	0.0104	0.2619
Xifeng Yan	0.875	0.0309	0.5
Avg.	**0.3507**	0.0118	0.2579

Outline

- Why Data Mining with Heterogeneous Info. Networks?
- RankClus: Integrated Clustering and Ranking in InfoNet
- RankClass: Classification with Heterog. Info. Networks
- Distinct: Object Distinction by InfoNet Analysis
- TruthFinder: Trust Analysis and Data Validation
- Role Discovery in Heterogeneous Info. Networks
- PathSim: Finding Similar Objects in Networks
- PathPredict: Relationship Prediction in Info. Networks
- Conclusions: Where Does the Power Come from?

Conclusions: Where Does the Power Come from?

- Heterogeneous information networks are ubiquitous
  - Most datasets can be "organized" or "transformed" into "*structured*" multi-typed heterogeneous info. networks
  - Examples: DBLP, IMDB, Flickr, Google News, Wikipedia, …
  - Structures can be progressively mined from less organized data sets by info. network analysis
  - Surprisingly rich knowledge can be mine from such structured heterogeneous info. networks
  - Clustering, ranking, classification, data cleaning, trust analysis, role discovery, similarity search, relationship prediction, ……
- Data mining by exploring the power of heterog. info. networks
  - Much more to be explored!!!

References for the Talk

- J. Han, Y. Sun, X. Yan, and . S. Yu, "***Mining Heterogeneous Information Networks***" (tutorial), KDD'10.
- M. Ji, et al., "***Graph Regularized Transductive Classification on Heterogeneous Information Networks***", ECMLPKDD'10.
- Y. Sun, J. Han, et al., "***RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis***", EDBT'09
- Y. Sun, Y. Yu, and J. Han, "***Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema***", KDD'09
- Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "***PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks***", VLDB'11
- Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "***Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks***", ASONAM'11
- C. Wang, J. Han, et al.,, , "***Mining Advisor-Advisee Relationships from Research Publication Networks***", KDD'10.
- X. Yin, J. Han, and P. S. Yu, "***Object Distinction: Distinguishing Objects with Identical Names by Link Analysis***", ICDE'07
- X. Yin, J. Han, and P. S. Yu, "***Truth Discovery with Multiple Conflicting Information Providers on the Web***", IEEE TKDE, 20(6), 2008