Mining Heterogeneous Information Networks: The Next Frontier

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

Collaborated with many, especially Yizhou Sun, Ming Ji, Chi Wang, Tim Weninger, Xiaoxin Yin, Bo Zhao

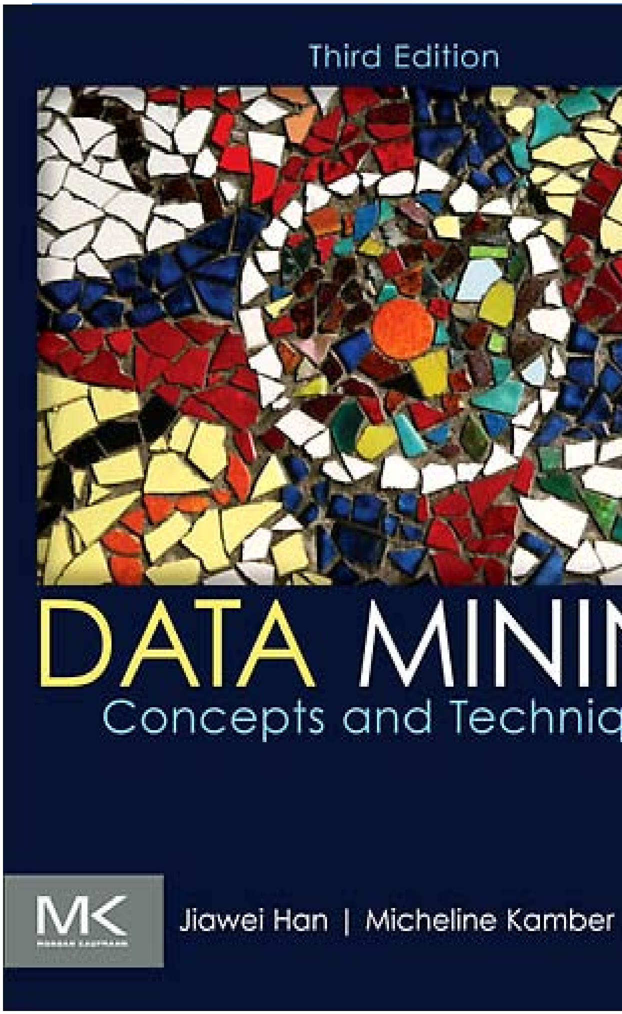
Acknowledgements: NSF, ARL, NASA, AFOSR (MURI), Microsoft, IBM, Yahoo!, Google, HP Lab & Boeing

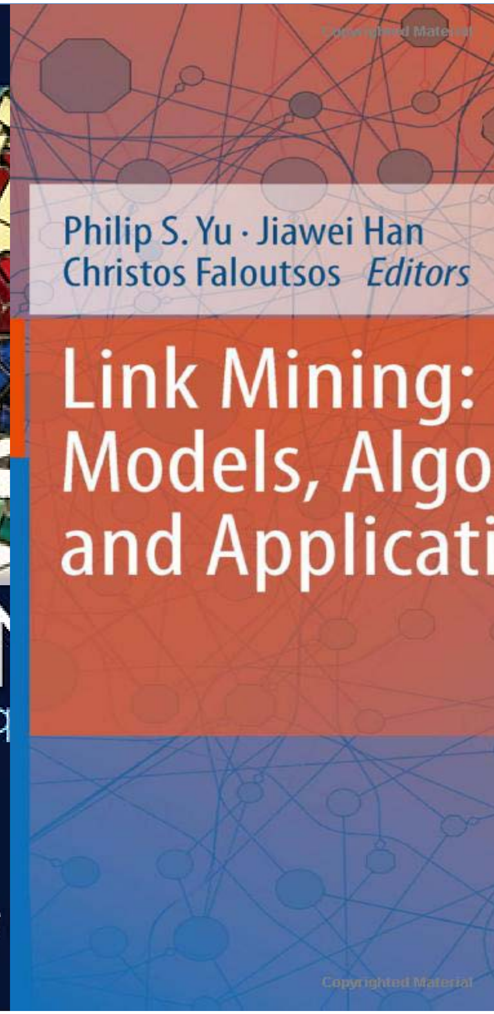
August 20, 2012

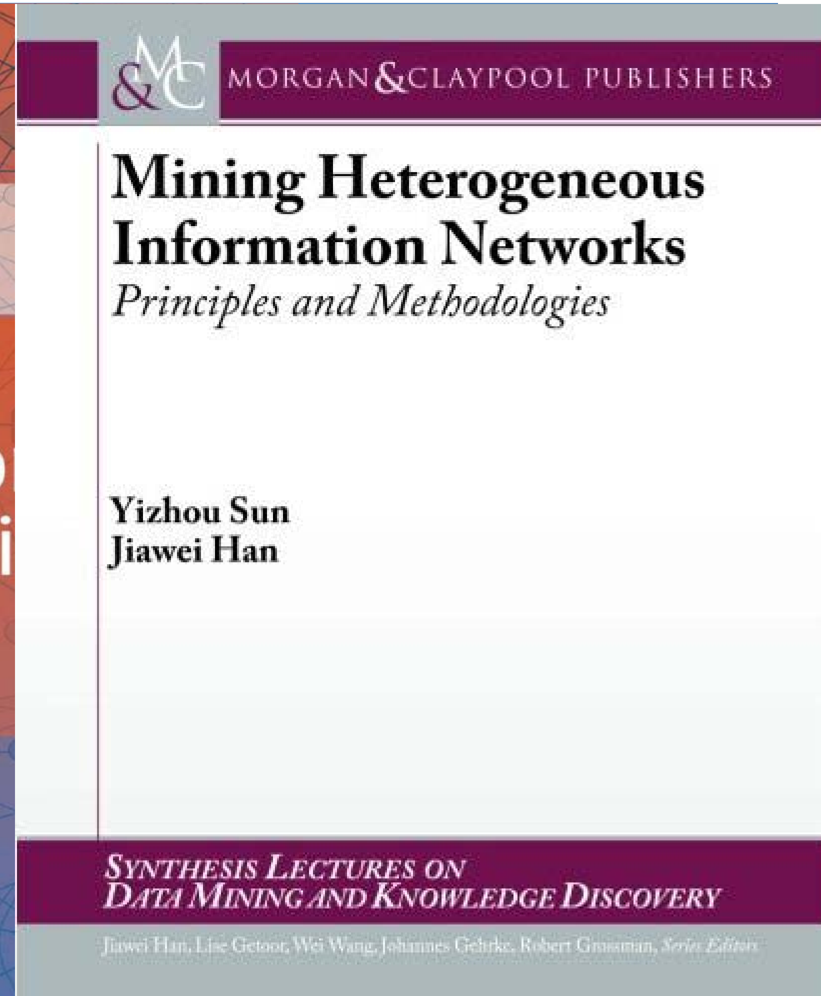
Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

From Data Mining to Mining Info. Networks







Han, Kamber and Pei,
Data Mining, 3rd ed. 2011

Yu, Han and Faloutsos (eds.),
Link Mining, 2010

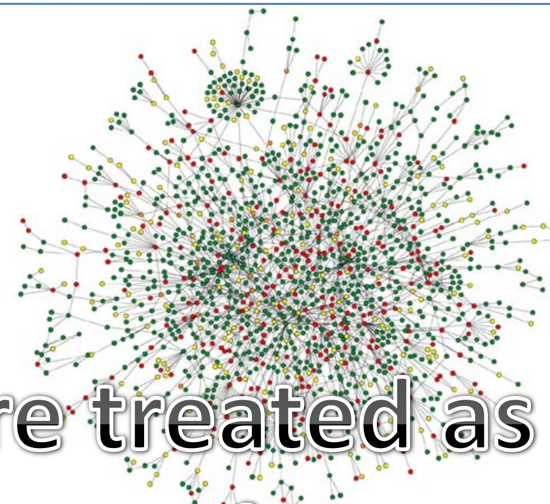
Sun and Han, Mining Heterogeneous
Information Networks, 2012

Why Is Mining Het. Info Net. the Next Frontier?

- **Data Mining Research: An Evolutionary path**
  - Mining simple data ⇒ mining complex data (structures, sequences, graphs/networks, heterogeneous info. networks)
  - *Heterogeneous* information networks vs. *homogeneous* information networks
  - Modeling the world as heterogeneous information networks
    - Captures the nature & rich info. of interconnected data
- **Mining heterogeneous information networks is**
  - **Necessary**: Reflecting the real nature of interconnected data
  - **Challenging**: Complexity, diversity, scalability, …
  - **Rewarding**: doable, exciting, efficient, as shown here

Where There Is Information, There Are Networks!

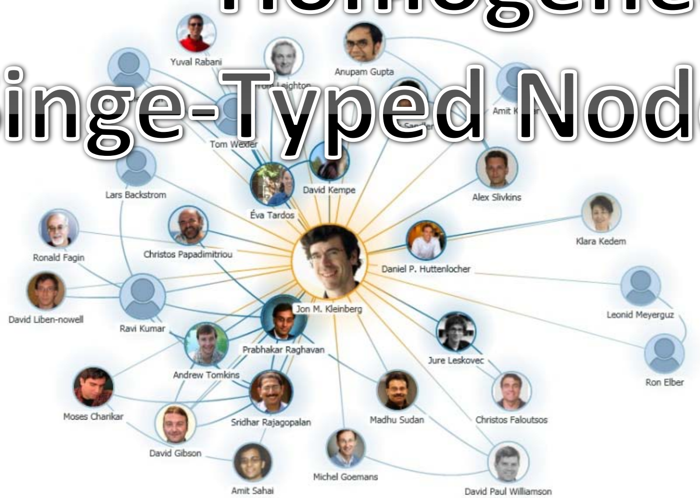
Social Networking Websites



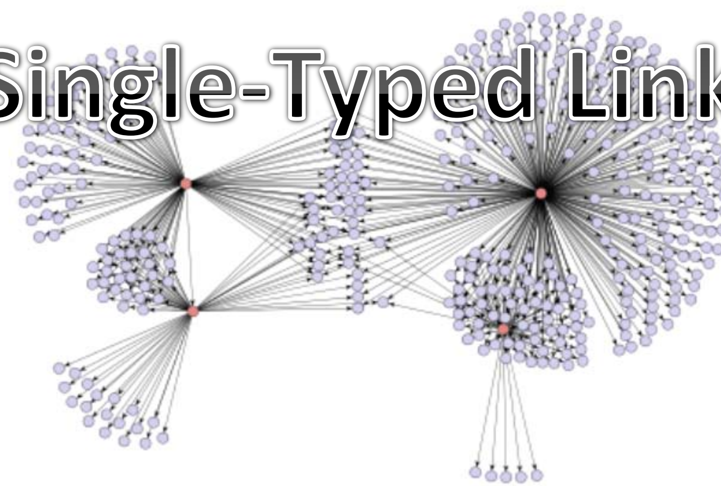
Biological Network: Protein Interaction

But most of them are treated as Homogeneous Networks —

Singe-Typed Nodes, Single-Typed Links!



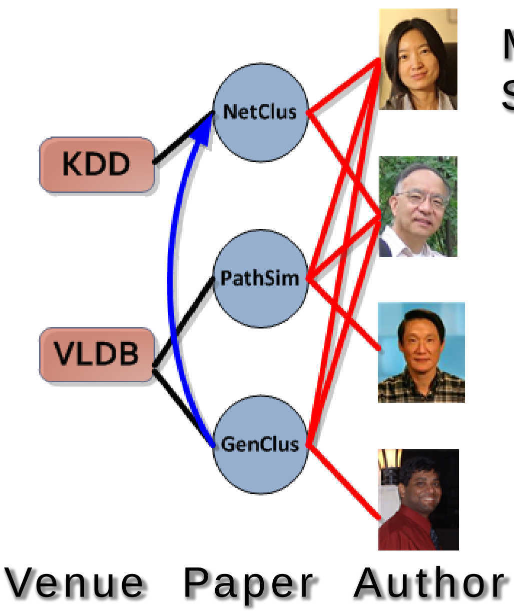
Research Collaboration Network



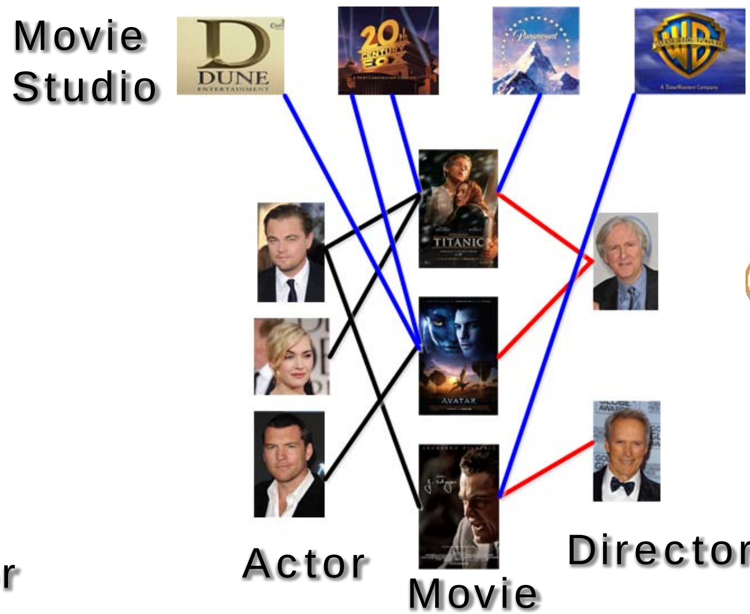
Product Recommendation Network via Emails

The Real World: Heterogeneous Networks

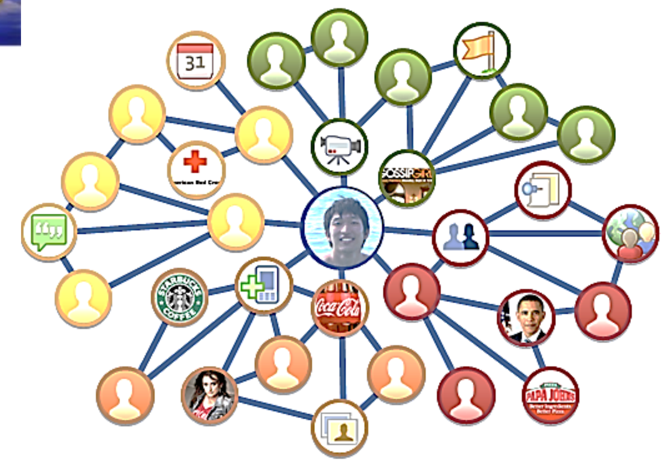
- Multiple object types and/or multiple link types



DBLP Bibliographic Network



The IMDB Movie Network



The Facebook Network

Homogeneous networks are *information loss* projection of heterogeneous networks!

Directly mining information-richer heterogeneous networks

Heterogeneous Networks Are Ubiquitous

- Healthcare
  - Doctor, patient, disease, treatment

- Online source code repository
  - Project, developer, programming language, project category, code, comments, …
- E-Commerce
  - Seller, buyer, product, review
- News
  - Person, organization, location, text

What Can be Mined from Heterogeneous Networks?

- DBLP: A Computer Science bibliographic database

26 Yizhou Sun, Jiawei Han, Charu C. Aggarwal, Nitesh V. Chawla: When will it happen?: relationship prediction in heterogeneous information networks. WSDM 2012: 663-672

A sample publication record in DBLP (>1.8 M papers, >0.7 M authors, >10 K venues), …

Knowledge hidden in DBLP Network	Mining Functions
How are CS research areas **structured?**	Clustering
Who are the **leading** researchers on Web search?	Ranking
What are the most essential **terms, venues, authors** in **AI**?	Classification + Ranking
Who are the **peer** researchers of Jure Leskovec?	Similarity Search
Whom **will** Christos Faloutsos **collaborate with**?	Relationship Prediction
Which types of **relationships** are most **influential** for an author to decide her topics?	Relation Strength Learning
How was the field of Data Mining **emerged** or **evolving**?	Network Evolution
Which authors are **rather different** from his/her peers in IR?	Outlier/anomaly detection

Outline

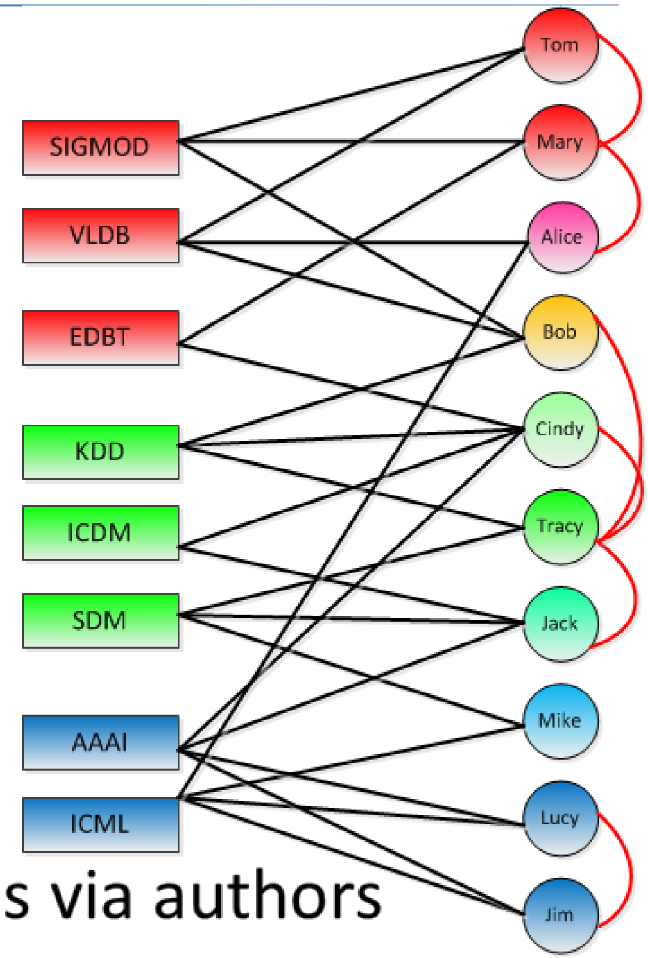
- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

RankClus: Integrated Clustering and Ranking in Heterogeneous Networks

- Clustering authors in one huge cluster without distinction?
  - Thinking about the power of PageRank!
- Ranking globally without considering clusters
  - Rank apples and bananas together?
- Integrated clustering with ranking
  - Ranking, as the feature of the cluster, is conditional (i.e., relative) to a specific cluster
  - E.g., VLDB's rank in Theory vs. its rank in the DB area
- RankClus: Clustering and ranking are mutually enhanced
  - Philosophy: Not all objects are equal in clustering!
- Y. Sun, et al., "*RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis*", EDBT'09

RankClus: Integrating Clustering with Ranking

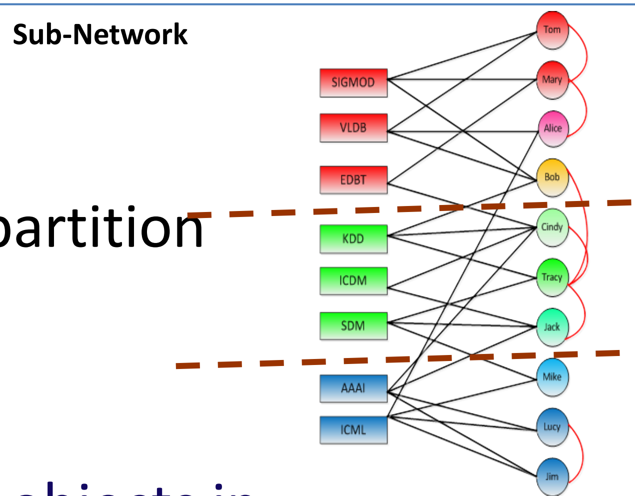
- A case study on bi-typed DBLP network
- Links exist between
  - Conference (X) and author (Y)
  - Author (Y) and author (Y)
- A matrix denoting the weighted links
  - $W = \begin{bmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{bmatrix}$
- Goal:
  - Clustering and ranking conferences via authors

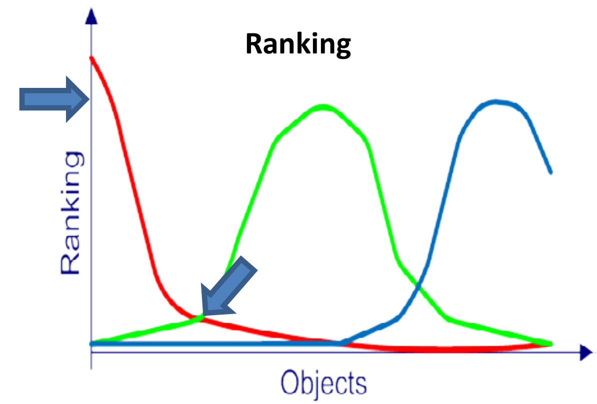


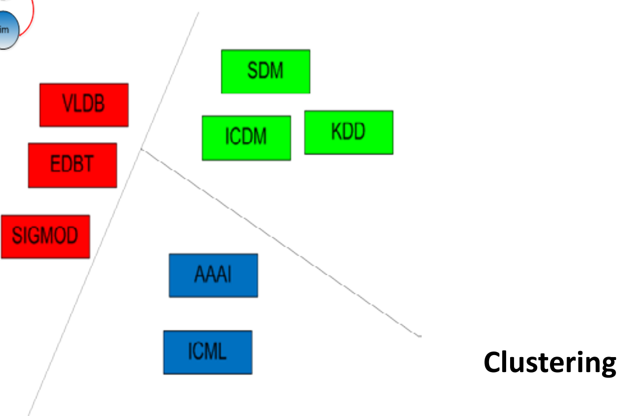
Project the bi-typed network into homogeneous conference network?
→ Information-loss projection!

RankClus: Algorithm Framework

- Initialization
  - Randomly partition
- Repeat
  - Ranking
    - Ranking objects in each sub-network induced from each cluster
  - Generating new measure space
    - Estimate **mixture model coefficients** for each target object
  - Adjusting cluster
- Until stable







Simple Ranking vs. Authority Ranking

- Simple Ranking
  - Proportional to # of publications of an author or a venue
  - Considers only **immediate neighborhood** in the network

What about an author publishing many papers in bogus conferences?

- Authority Ranking:
  - More sophisticated "rank rules" are needed
  - **Propagate** the ranking scores in the network over different types

Rules for 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* 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

Clustering and ranking two fields: DB/DM & HW/CA

Initially, ranking distributions are mixed together

Improved a little

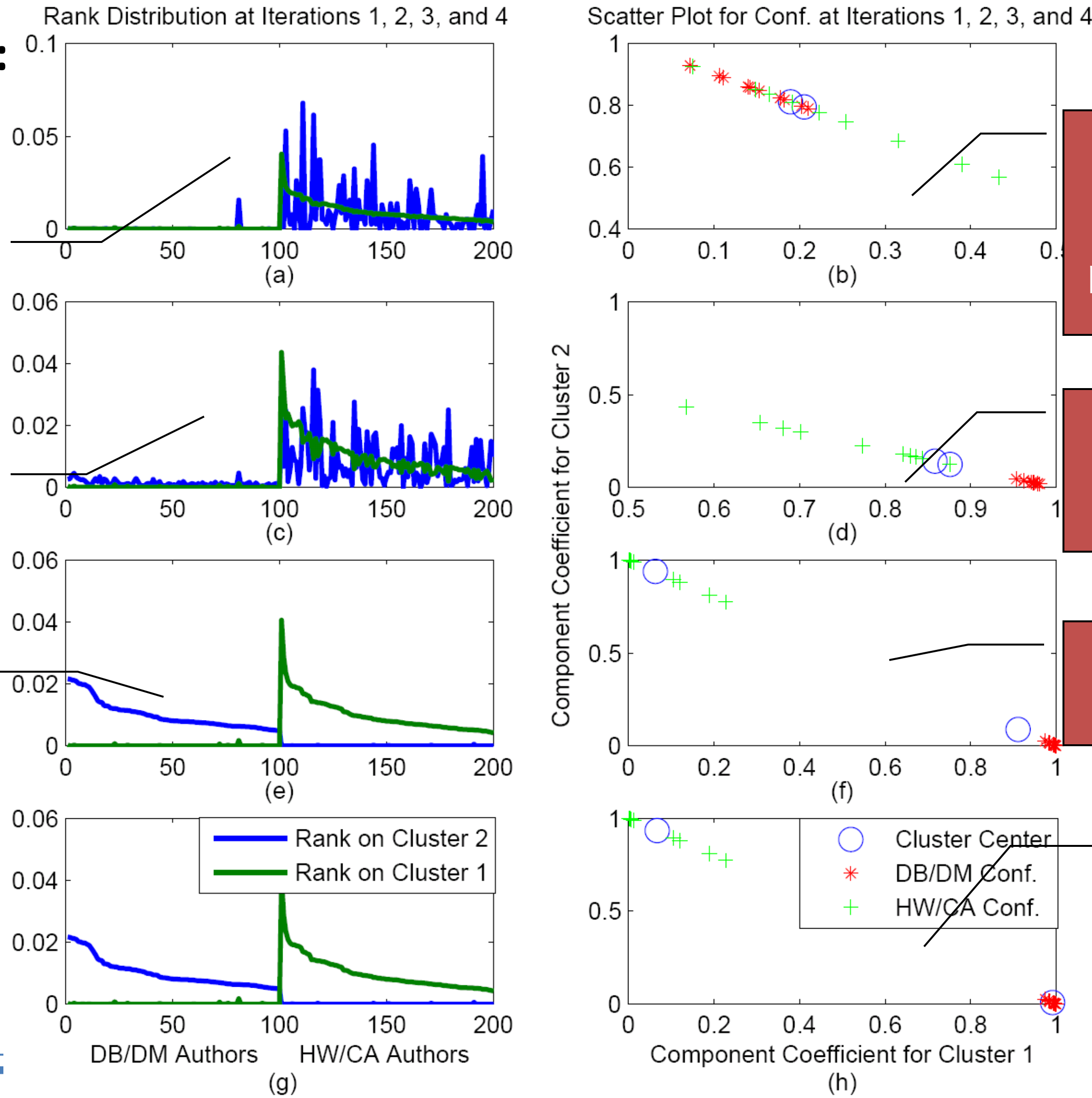
Improved significantly

Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable



Experiment on Dataset: DBLP

- 2676 conferences and 20,000 authors with publications from 1998 to 2007
- Both conf.-author and co-author relationships are used
- K=15 (select only 5 clusters here)

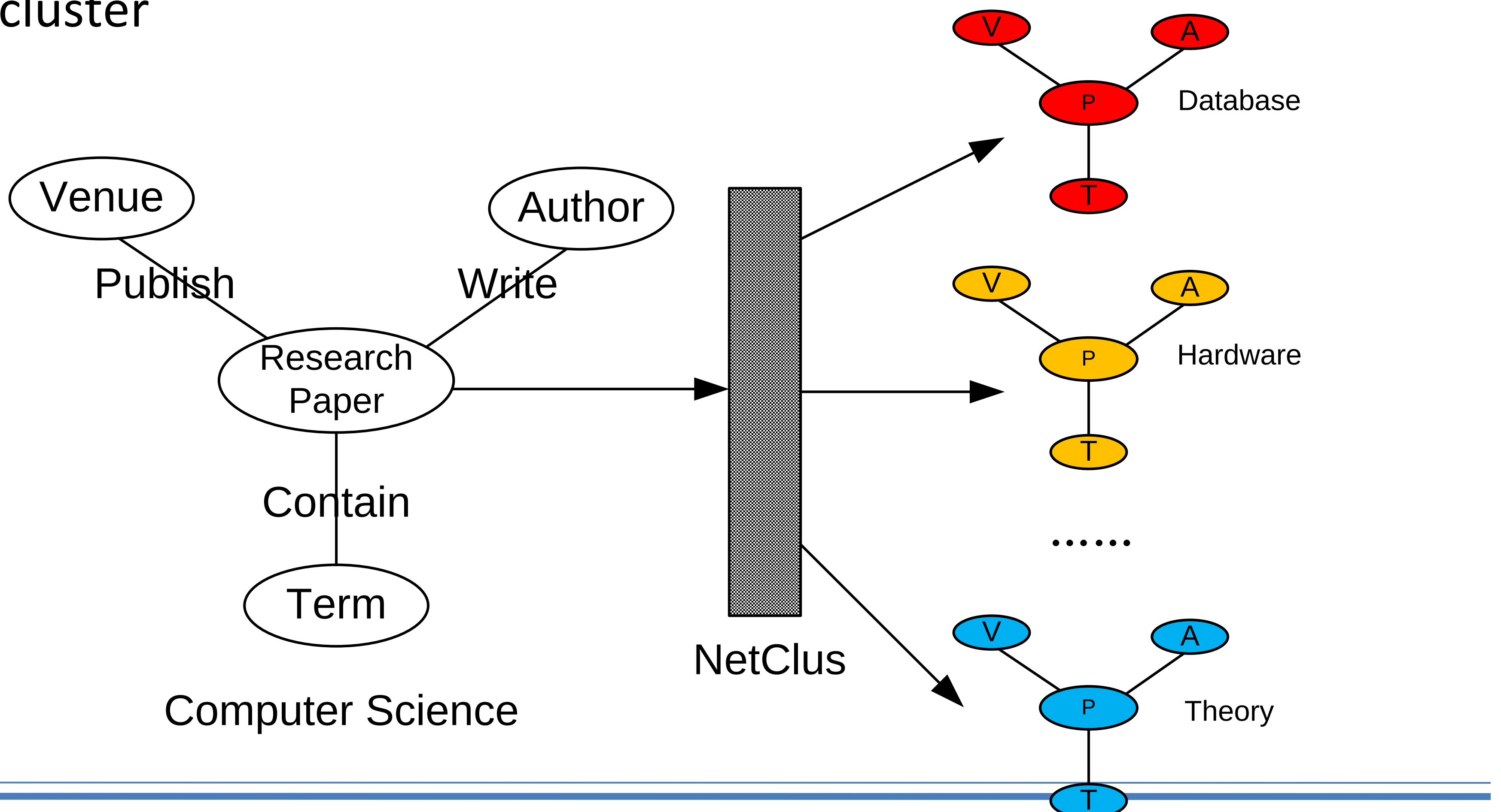
Table 5: Top-10 Conferences in 5 Clusters Using RankClus

	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: ~O(K|E|), where K is the number of clusters

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



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

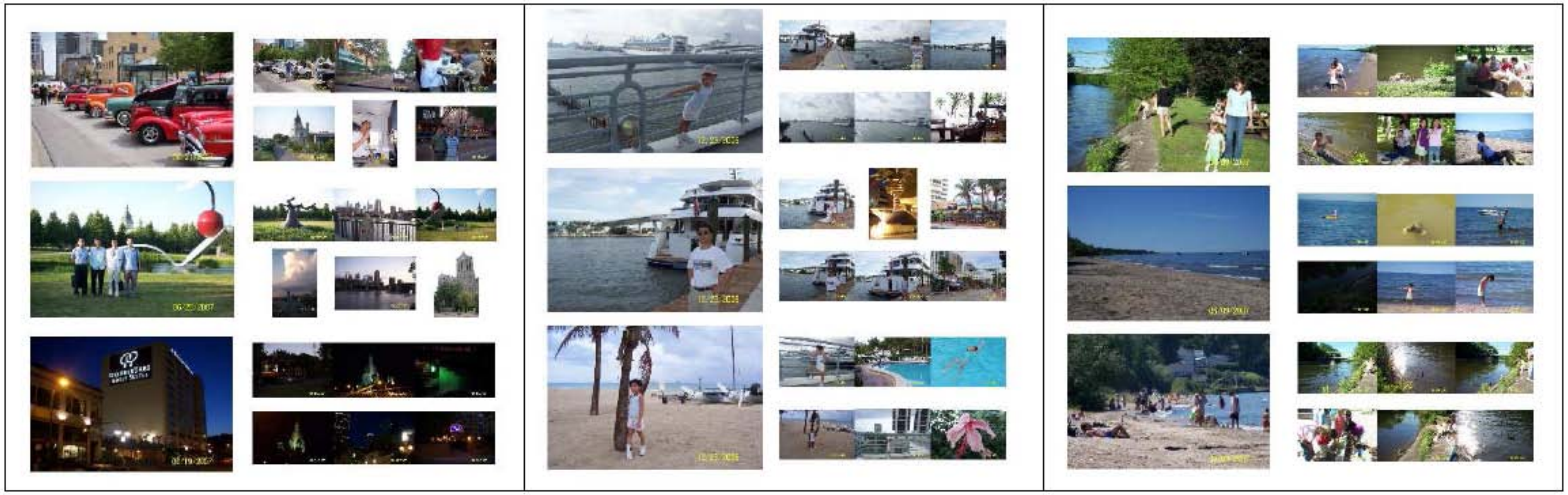
VLDB 0.318495
SIGMOD Conf. 0.313903
ICDE 0.188746
PODS 0.107943
EDBT 0.0436849

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

Rank-Based Clustering for Others



RankCompete: Organize your photo album automatically!

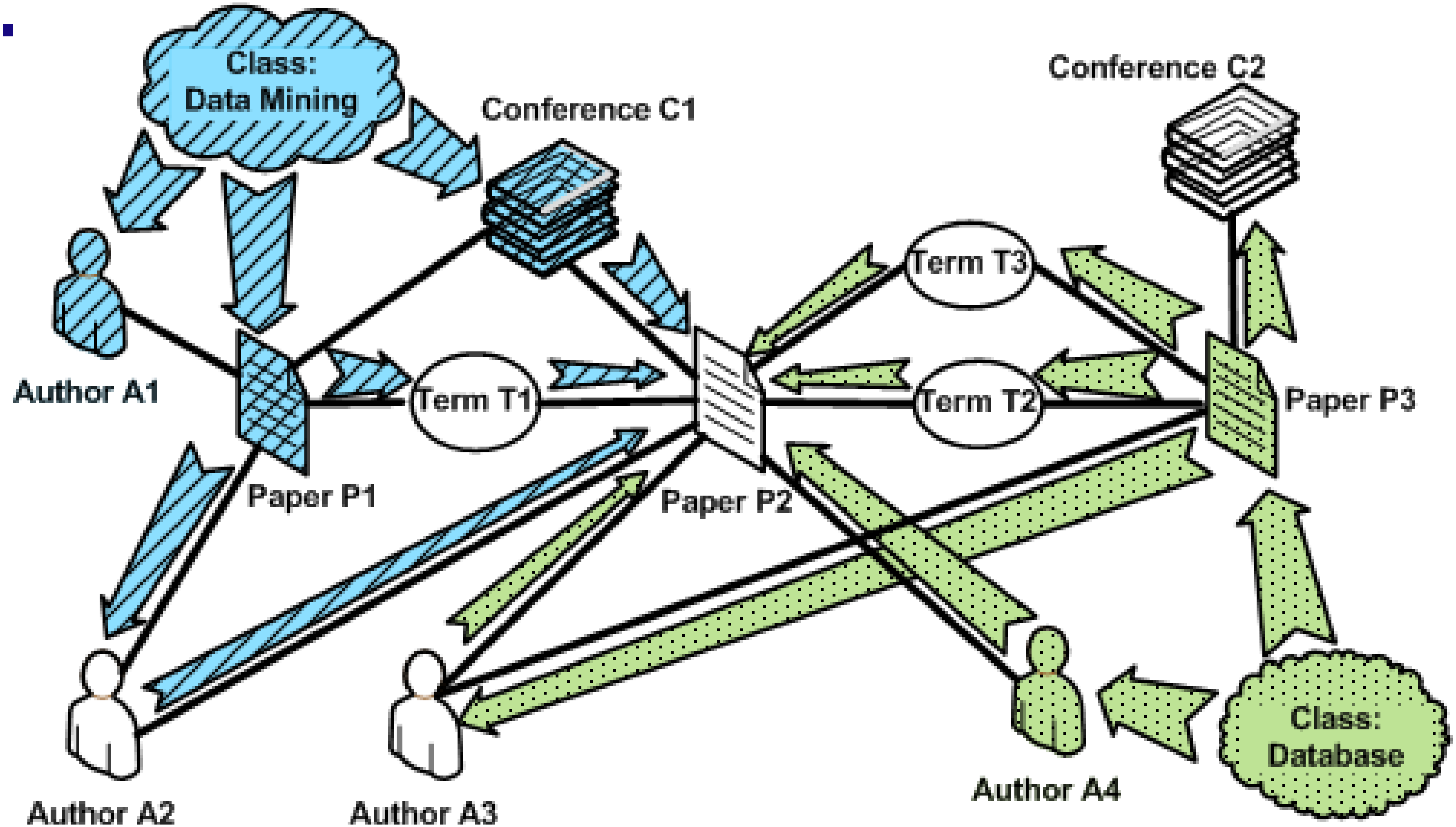
	Top 10 Treatments	Ranking
1	Zidovudine/therapeutic use	0.1679
2	Anti-HIV Agents/therapeutic use	0.1340
3	Antiretroviral Therapy, Highly Active	0.0977
4	Antiviral Agents/therapeutic use	0.0718
5	Anti-Retroviral Agents/therapeutic use	0.0236
6	Interferon Type I/therapeutic use	0.0147
7	Didanosine/therapeutic use	0.0132
8	Ganciclovir/therapeutic use	0.0114
9	HIV Protease Inhibitors/therapeutic use	0.0105
10	Antineoplastic Combined Chemotherapy	0.0103

Rank treatments for AIDS from MEDLINE

Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

Classification: Knowledge Propagation



M. Ji, M. Danilevski, et al., "Graph Regularized Transductive Classification on Heterogeneous Information Networks", ECMLPKDD'10

GNetMine: Graph-Based Regularization

- Minimize the objective function

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

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

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

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

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

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 [M. Ji et al., KDD'11]: 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

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

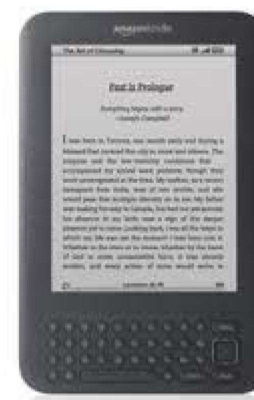
	Database	Data Mining	AI	IR
Top-5 ranked conferences	VLDB	KDD	IJCAI	SIGIR
	SIGMOD	SDM	AAAI	ECIR
	ICDE	ICDM	ICML	CIKM
	PODS	PKDD	CVPR	WWW
	EDBT	PAKDD	ECML	WSDM
Top-5 ranked terms	data	mining	learning	retrieval
	database	data	knowledge	information
	query	clustering	reasoning	web
	system	classification	logic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

Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

Similarity Search: Find Similar Objects in Networks

- DBLP
  - Who are the most similar to “Christos Faloutsos”?
- IMDB
  - Which movies are the most similar to “Little Miss Sunshine”?
- E-Commerce
  - Which products are the most similar to “Kindle”?



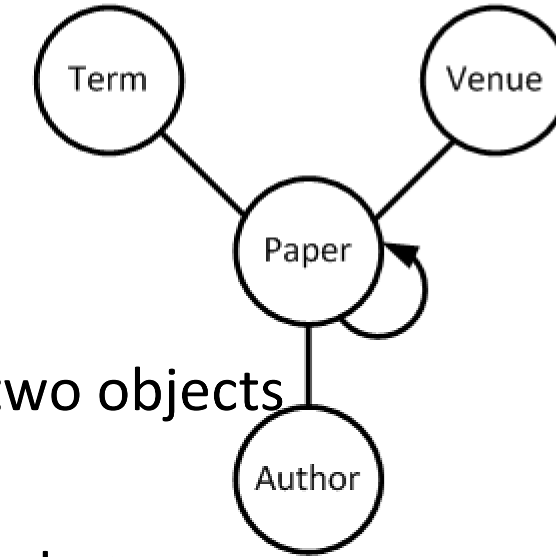
How to systematically answer these questions ?

Study similarity search in heterogeneous networks

- Y. Sun, J. Han, X. Yan, P. S. Yu, and Tianyi Wu, “**PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks**”, VLDB'11

Network Schema and Meta-Path

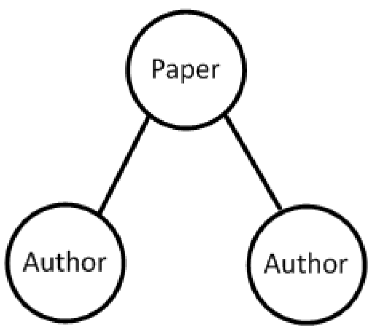
- Network schema
  - Meta-level description of a network



- Meta-Path
  - **Meta-level description** of a path between two objects
  - **A path** on network schema
  - Denote an existing or concatenated **relation** between two object types

"Jim-P1-Ann"
"Mike-P2-Ann"
"Mike-P3-Bob"
…

Path Instances



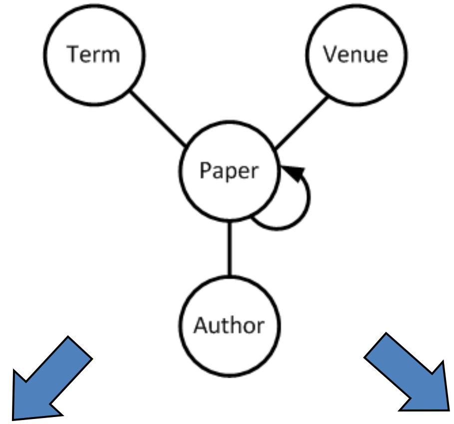
Meta-Path

Co-authorship

Relation: Describe the Type of Relationships

Different Meta-Paths Tell Different Semantics

- Who are most similar to Christos Faloutsos?



Meta-Path: ***Author-Paper-Author***

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

Christos's students or close collaborators

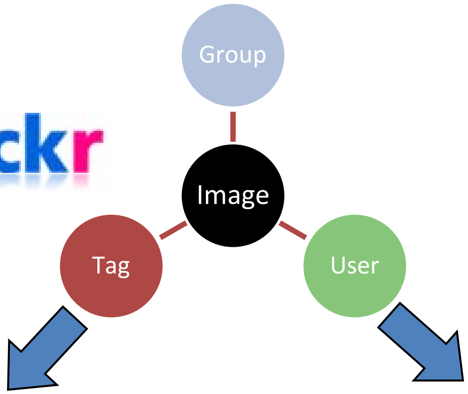
Meta-Path: ***Author-Paper-Venue-Paper-Author***

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

Work on similar topics and have similar reputation

Some Meta-Path Is "Better" Than Others

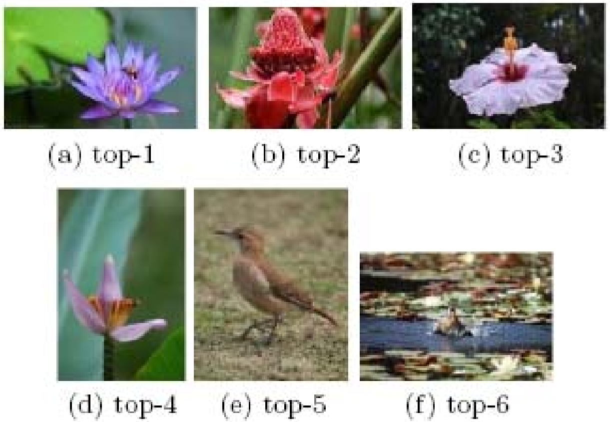
- Which pictures are most similar to  ?



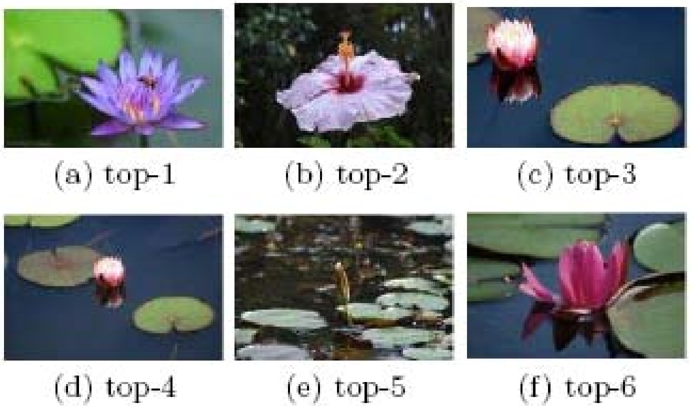
Evaluate the similarity between images according to their linked tags

Evaluate the similarity between images according to tags and groups

Meta-Path: ***Image-Tag-Image***



Meta-Path: ***Image-Tag-Image-Group-Image-Tag-Image***



Some Similarity Measure Is "Better" Than Others

- Anhai Doan
  - CS, Wisconsin
  - Database area
  - PhD: 2002

- Jignesh Patel
  - CS, Wisconsin
  - Database area
  - PhD: 1998

Meta-Path: ***Author-Paper-Venue-Paper-Author***

Rank	P-PageRank	SimRank	PathSim
1	AnHai Doan	AnHai Doan	AnHai Doan
2	Philip S. Yu	Douglas W. Cornell	Jignesh M. Patel
3	Jiawei Han	Adam Silberstein	Amol Deshpande
4	Hector Garcia-Molina	Samuel DeFazio	Jun Yang
5	Gerhard Weikum	Curt Ellmann	Renée J. Miller

- Amol Deshpande
  - CS, Maryland
  - Database area
  - PhD: 2004

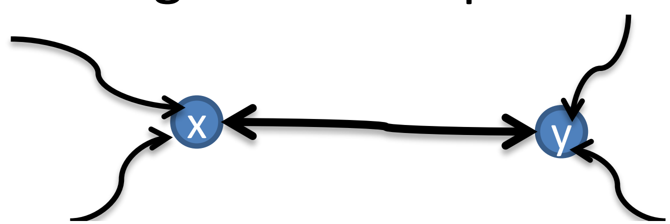
- Jun Yang
  - CS, Duke
  - Database area
  - PhD: 2001

PathSim vs. Some Popular Measures

- Popular object similarity measures in networks
  - Random walk (RW) or Personalized PageRank: Favors **highly visible** objects (i.e., objects with large degrees)
  - Pairwise random walk (PRW) (or SimRank): Favors **"pure"** objects (i.e., objects with highly skewed distribution in their in-links or out-links)

Note: P-PageRank and SimRank do not distinguish object type and relationship type

- PathSim
  - Favor **"peers"**: objects with strong connectivity and similar visibility under the given meta-path



$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

Comparison with Other Measures: A Toy Example

Who is most similar to Mike?

(a) Adjacency matrix W_{AC}.

	SIGMOD	VLDB	ICDE	KDD
Mike	2	1	0	0
Jim	50	20	0	0
Mary	2	0	1	0
Bob	2	1	0	0
Ann	0	0	1	1

(b) Similarity between Mike and other authors.

	Jim	Mary	Bob	Ann
P-PageRank	**0.3761**	0.0133	**0.0162**	0.0046
SimRank	**0.7156**	0.5724	**0.7125**	0.1844
RW	**0.8983**	0.0238	**0.0390**	0
PRW	**0.5714**	0.4444	**0.5556**	0
PathSim	0.0826	**0.8**	**1**	0

Comparing Similarity Measures in DBLP Data

Which venues are most similar to DASFAA?

Favor highly visible objects

(a) P-PageRank: $CPAPC$

Rank	Conference
1	DASFAA
2	ICDE
3	VLDB
4	SIGMOD Conference
5	DEXA
6	TKDE
7	CIKM
8	Data Knowl. Eng.
9	SIGIR
10	SIGMOD Record

(b) PathSim: $CPAPC$

Rank	Conference
1	DASFAA
2	DEXA
3	WAIM
4	APWeb
5	CIKM
6	WISE
7	ICDE
8	Data Knowl. Eng.
9	PAKDD
10	EDBT

Table 5: P-PageRank vs. PathSim on query: "DASFAA"

Which venues are most similar to SIGMOD?

These tiny forums most similar to SIGMOD?

(a) SimRank: $CPAPC$

Rank	Conference
1	SIGMOD Conf.
2	Found. and Trends in DB
3	ACM SIGMOD D. S. C.
4	HPTS
5	DB for Inter. Des.
6	IPSJ
7	CIDR
8	AFIPS NCC
9	XQuery Impl. Parad
10	CleanDB

(b) PathSim: $CPAPC$

Rank	Conference
1	SIGMOD Conf.
2	VLDB
3	ICDE
4	IEEE Data Eng. Bull.
5	SIGMOD Rec.
6	ACM Trans. DB Syst.
7	TKDE
8	PODS
9	VLDB J.
10	EDBT

Table 6: SimRank vs. PathSim on query: "SIGMOD"

Long Meta-Path May Not Carry the Right Semantics

- Repeat the meta-path 2, 4, and infinite times for conference similarity query

(a) Path: $(CPAPC)^2$

Rank	Term	Score
1	SIGMOD Conference	1
2	VLDB	0.981
3	ICDE	0.949
4	TKDE	0.650
5	SIGMOD Record	0.630
6	IEEE Data Eng. Bull.	0.530
7	PODS	0.467
8	ACM Trans. Database Syst.	0.429
9	EDBT	0.420
10	CIKM	0.410

(b) Path: $(CPAPC)^4$

Rank	Term	Score
1	SIGMOD Conference	1
2	VLDB	0.997
3	ICDE	0.996
4	TKDE	0.787
5	SIGMOD Record	0.686
6	PODS	0.586
7	KDD	0.553
8	CIKM	0.540
9	IEEE Data Eng. Bull.	0.532
10	J. Comput. Syst. Sci	0.463

(c) Path: $(CPAPC)^\infty$

Rank	Term	Score
1	SIGMOD Conference	1
2	AAAI	0.9999
3	ESA	0.9999
4	IEEE Trans. on Commun.	0.9999
5	STACS	0.9997
6	PODC	0.9996
7	NIPS	0.9993
8	Comput. Geom.	0.9992
9	ICC	0.9991
10	ICDE	0.9984

Table 8: Top-10 similar conferences to "SIGMOD" under path schemas with different lengths

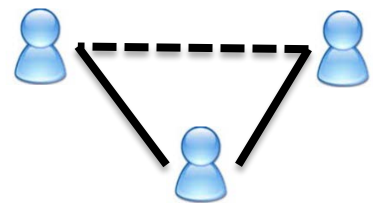
- Efficient support of top-k similarity queries
  - Co-clustering based pre-computation (i.e., materialization) of meta-path matrices

Outline

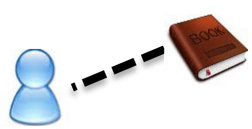
- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

PathPredict: Meta-Path Based Relationship Prediction

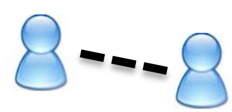
- Previous work: Link prediction in homogeneous networks [Liben-Nowell and Kleinberg, 2003, Hasan et al., 2006]
  - E.g., friendship prediction



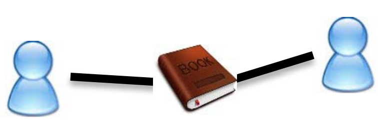
- Relationship prediction in heterogeneous networks [ASONAM'11]
  - Predict what to write, where to submit, whom to coauthor, …
  - Different types of relationships need different prediction models



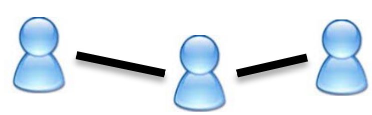
vs.



  - Different connection paths need to be treated separately!
    - **Use meta-paths** to define topological features

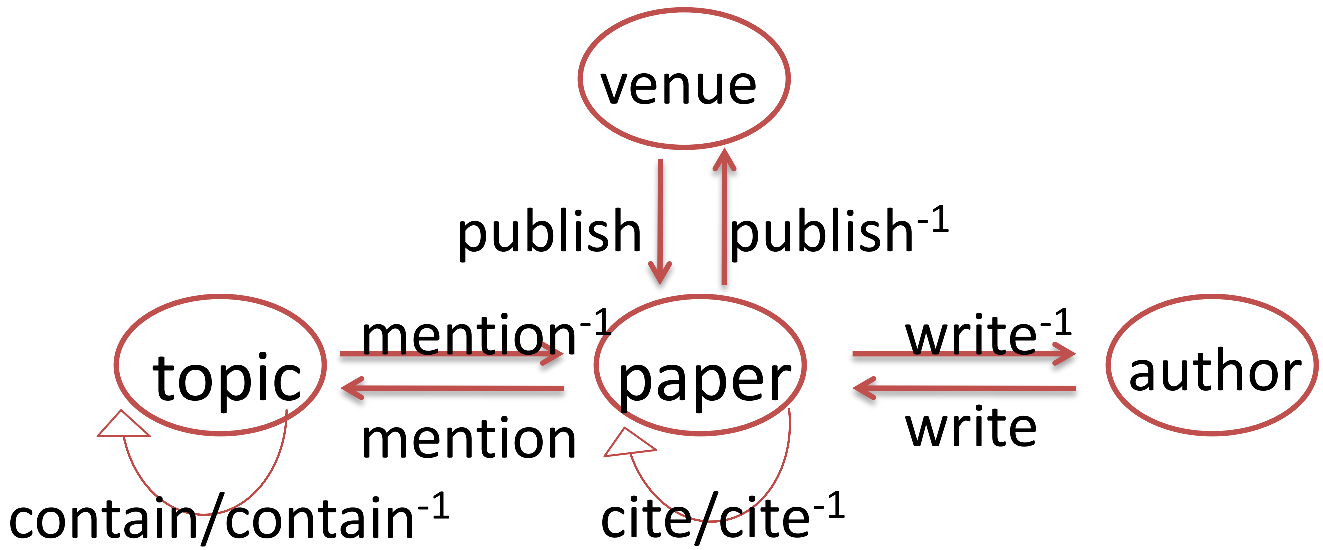


vs.



Guidance: Meta Path in Bibliographic Network

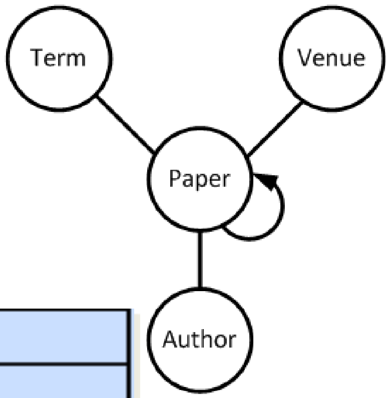
- Relationship prediction: meta path-guided prediction
- Meta path relationships among similar typed links share similar semantics and are comparable and inferable



- Co-author prediction (A—P—A) using topological features also encoded by meta paths, e.g., citation relations between authors (A—P→P—A)

Meta-Path Based Co-authorship Prediction in DBLP

- Co-authorship prediction problem
  - Whether two authors are going to collaborate for the first time
- Co-authorship encoded in meta-path
  - Author-Paper-Author
- Topological features encoded in meta-paths



Meta-Path	Semantic Meaning
$A - P \rightarrow P - A$	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

Meta-paths between authors under length 4

The Power of PathPredict

- Explain the prediction power of each meta-path
  - Wald Test for logistic regression
- Higher prediction accuracy than using projected homogeneous network
  - **11%** higher in prediction accuracy

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

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 Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

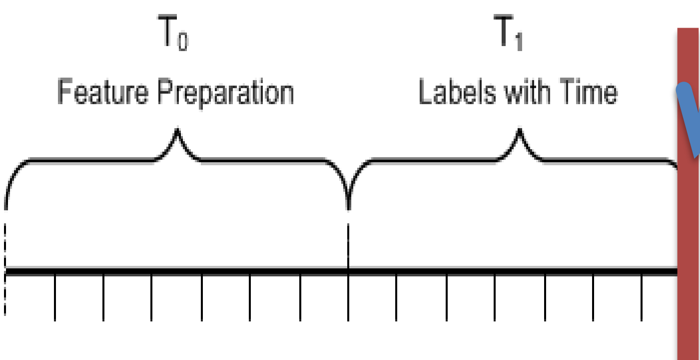
Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors!

(Feature collected in [1996, 2002]; Test period in [2003,2009])

When Will It Happen?—When Will You Cite Him?

- The Relationship Building Time Prediction Model [WSDM'12]
  - Directly **model relationship building time**: P(Y=t)
    - Geometric distribution, Exponential distribution, Weibull distribution
  - Use **generalized linear model**
    - Deal with censoring (relationship builds beyond the observed time interval)

***T*: Right Censoring**



Training Framework

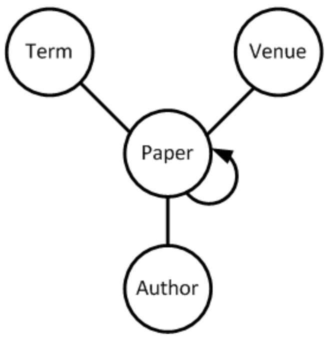
$$\log L = \sum_{i=1}^{n} \left(f_Y(y_i|\theta_i,\lambda) I_{\{y_i<T\}} + P(y_i \geq T|\theta_i,\lambda) I_{\{y_i \geq T\}} \right)$$

Generalized Linear Model
under Weibull Distribution Assumption

$$LL_W(\boldsymbol{\beta},\lambda) = \sum_{i=1}^{n} I_{\{y_i<T\}} \log \frac{\lambda y_i^{\lambda-1}}{e^{-\lambda \mathbf{X}_i \beta}} - \sum_{i=1}^{n} \left(\frac{y_i}{e^{-\mathbf{X}_i \beta}}\right)^{\lambda}$$

Author Citation Time Prediction in DBLP

- Top-4 meta-paths for author citation time prediction



$$A - P - T - P - A$$ Study the same topic

$$A - P \leftarrow P \rightarrow P - A$$ Co-cited by the same paper

$$A - P - A - P \rightarrow P - A$$ Follow co-authors' citation

$$A - P - T - P - A - P \rightarrow P - A$$ Follow the citations of authors who study the same topic

Social relations are less important in author citation prediction than in co-author prediction.

- Predict when Philip S. Yu will cite a new author

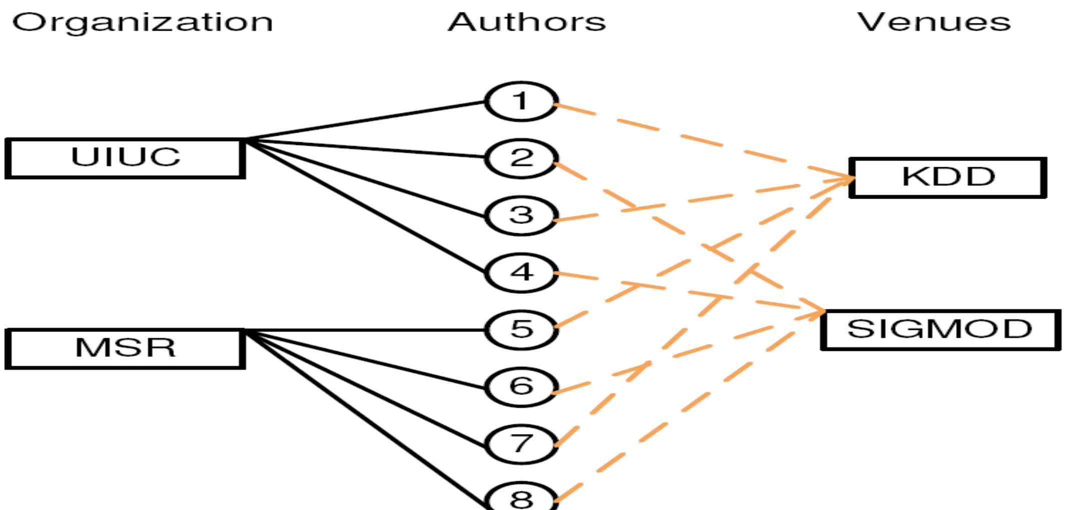
a_i	a_j	Ground Truth	Median	Mean	25% quantile	75% quantile
Philip S. Yu	Ling Liu	1	2.2386	3.4511	0.8549	4.7370
Philip S. Yu	Christian S. Jensen	3	2.7840	4.2919	1.0757	5.8911
Philip S. Yu	C. Lee Giles	0	8.3985	12.9474	3.2450	17.7717
Philip S. Yu	Stefano Ceri	0	0.5729	0.8833	0.2214	1.2124
Philip S. Yu	David Maier	9+	2.5675	3.9581	0.9920	5.4329
Philip S. Yu	Tong Zhang	9+	9.5371	14.7028	3.6849	20.1811
Philip S. Yu	Rudi Studer	9+	9.7752	15.0698	3.7769	20.6849

Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

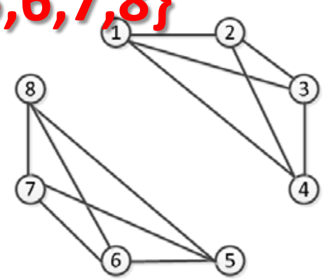
Why User Guidance in Clustering?

- Different users may like to get different clusters
  - Clustering authors based on their connections in the network



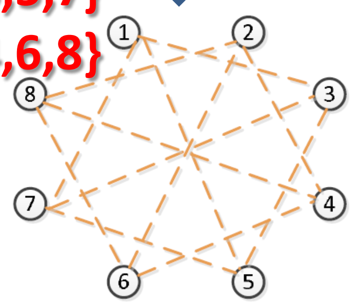
Which meta-path to choose?

{1,2,3,4}
{5,6,7,8}



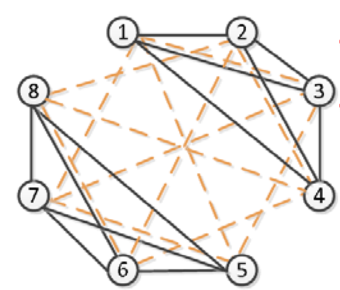
(a) AOA

{1,3,5,7}
{2,4,6,8}



(b) AVA

{1,3}
{2,4}
{5,7}
{6,8}

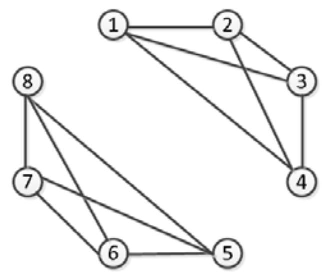


(c) AOA + AVA

User Guidance Determines Clustering Results

- Different user preferences (e.g., by seeding desired clusters) lead to the choice of different met-paths

{1}
{5}



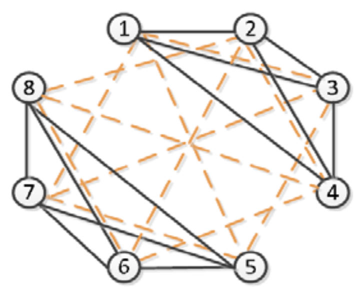
(a) AOA

+

{1,2,3,4}
{5,6,7,8}

Seeds **Meta-path(s)** **Clustering**

{1}
{2}
{5}
{6}



(c) AOA + AVA

+

{1,3}
{2,4}
{5,7}
{6,8}

Seeds **Meta-path(s)** **Clustering**

Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

Challenge I: Automated Construction of Heterogeneous Info. Networks

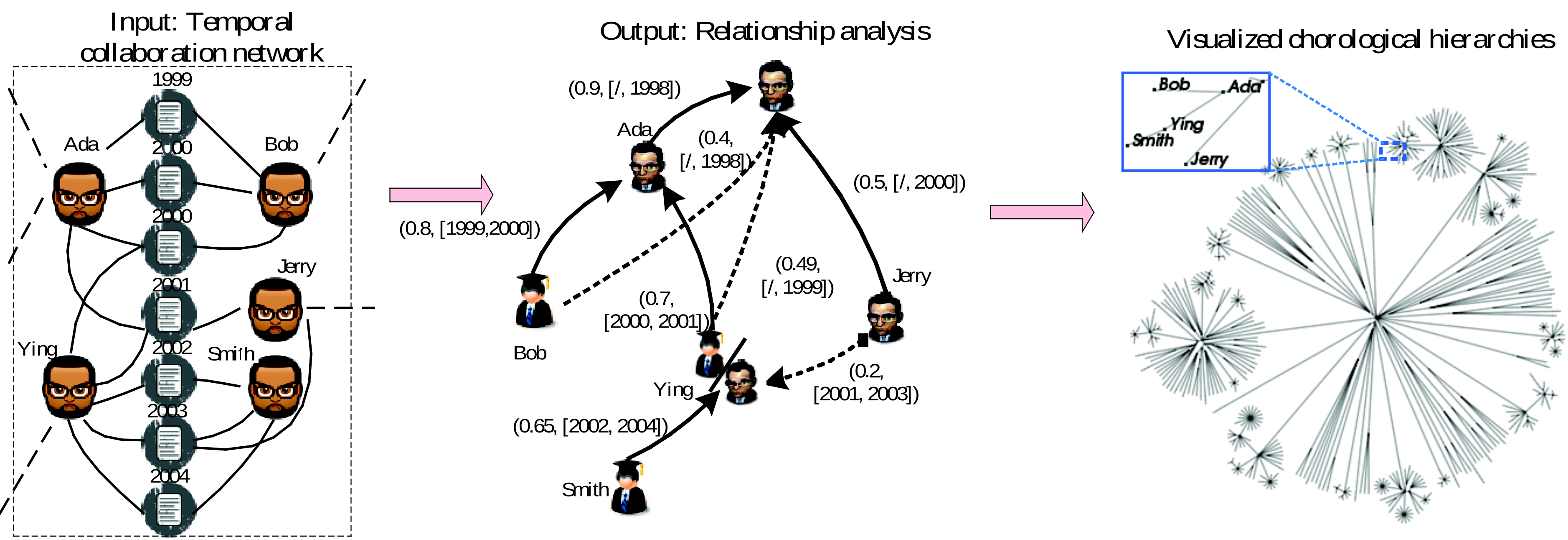
- Much of the real world data is unstructured or partially structured
  - News, Wikipedia, blogs, multimedia data, …

- **Challenge: Generation of structured heterogeneous info. networks from unstructured data**

- Entity/type/information extraction: NLP, ML, DB, Web, ……
- Role and hidden structure discovery (KDD'10, SDM'12)
- Web structure discovery by parallel path growth (WWW'11)
- Integration of structure and unstructured information networks
- Progressive refinement and self-boosting
  - Boosting information network construction and refinement by information network mining

Role Discovery: Mining Advisor-Advisee Relationships in DBLP Network

- Propagation of simple, commonly accepted constraints in Time-Constrained Probabilistic Factor Graph (TPFG)
  - *"Advisor has more publications and longer history than advisee at the time of advising"*
  - *"Once an advisee becomes advisor, s/he will not become advisee again"*



Role Discovery: Performance & Case Study

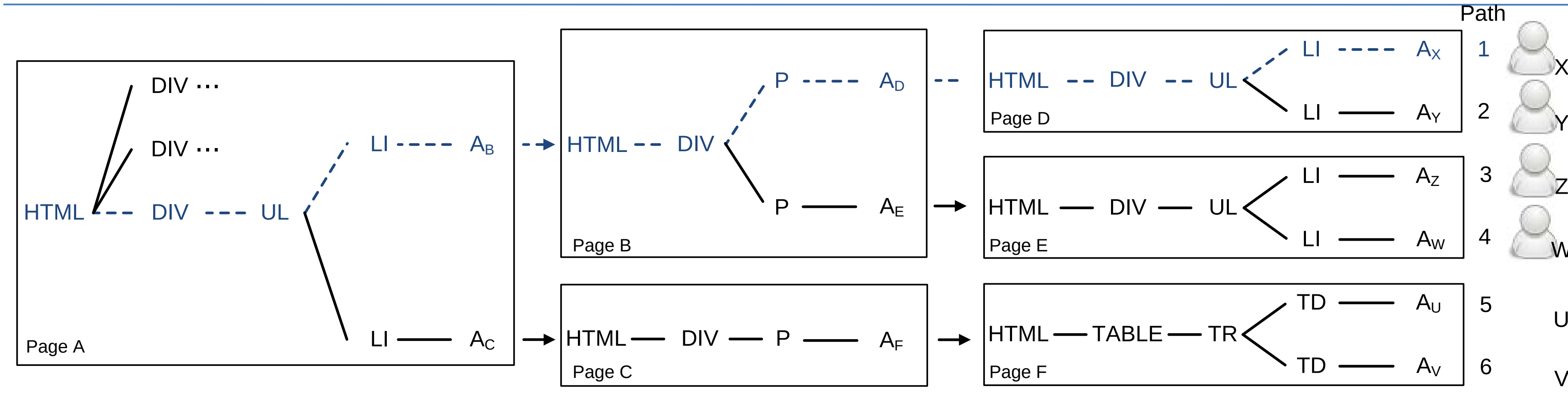
- DBLP data: 654, 628 authors, 1076,946 publications, years provided
- Labeled data: MathGealogy Project; AI Gealogy Project; Homepage

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

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”

Web Structure Discovery by Growing Parallel Paths



Finding home pages of CS professors at UIUC

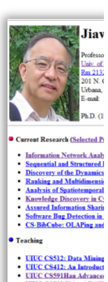
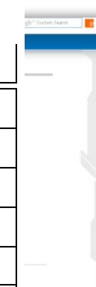


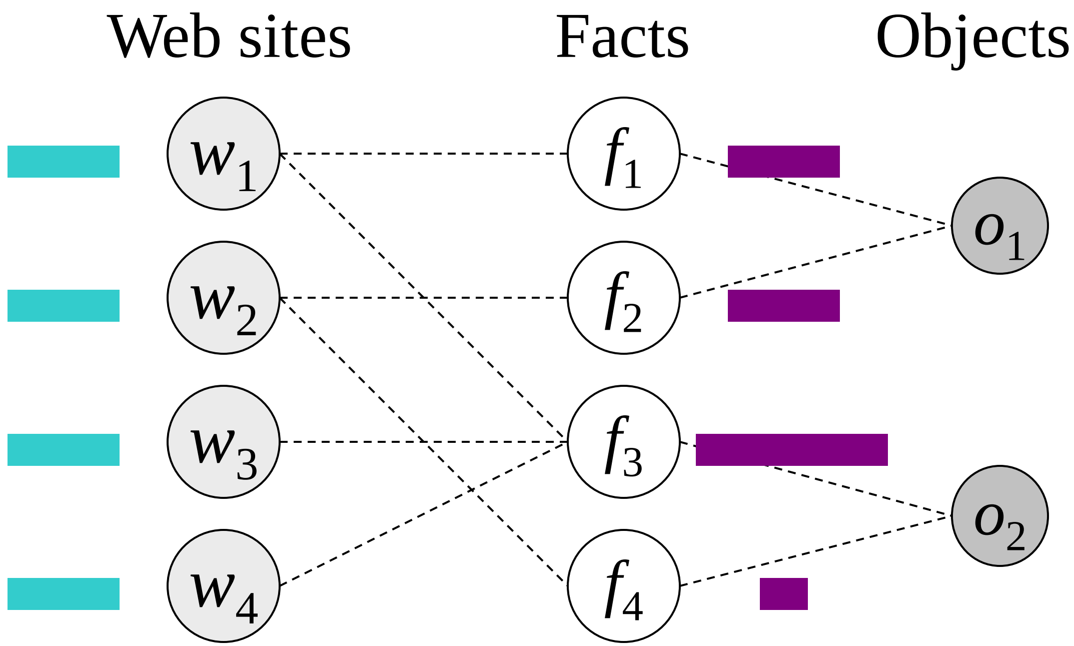
Table 1: Entity-page discovery results

Domain	Reference Page	Example Entity	Count	k-Shortest Paths Precision	k-Shortest Paths Recall	Path Removal Precision	Path Removal Recall
CS Faculty	`cs.*.edu`	Various	1,410	75.4	57.6	90.3	87.4
UIUC CS Courses	`cs.illinois.edu`	CS410	84	96.7	100	100	100
UIUC CS Groups	`cs.illinois.edu`	DAIS	36	100	100	100	100
Representatives	`house.gov`	Tim Johnson	441	100	100	100	100
Senators	`senate.gov`	Dick Durbin	100	55.3	100	100	100
Senate Committees	`senate.gov`	Finance	40	100	100	100	100
House Committees	`house.gov`	Ways and Means	45	100	100	100	100
Football Teams	`espn.go.com`	Illinois Fighting Illini	238	100	100	100	100
Football Players	`espn.go.com`	Nathan Scheelhaase	10,154	100	100	100	100



Challenge II: Enhancing the Quality of Heterogeneous Info. Networks

- Info. networks could be untrustworthy, error-prone, missing, …
- TruthFinder [KDD'07]: Inference on trustworthiness by constructing heterogeneous info. networks

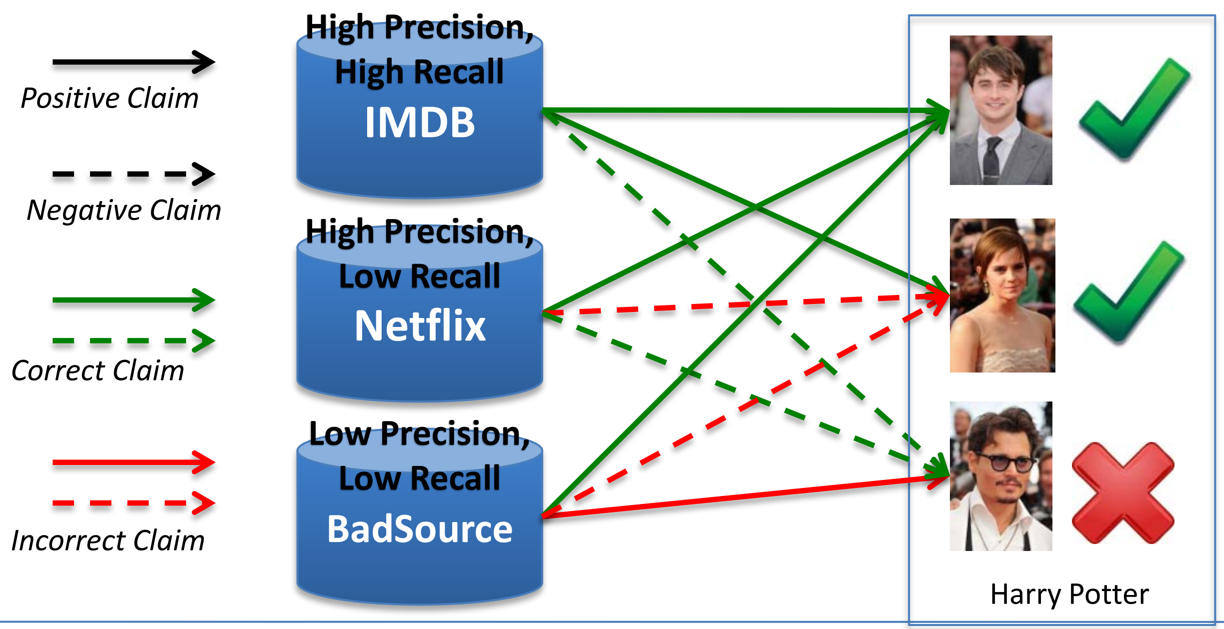


- True facts and trustable websites mutually enhance each other and will become apparent after some iterations

Truth Discovery: Multiple Truth Value and Handling False Negatives

- Voting may not always work well: Some sources tend to miss true values (False Negatives), while some others tend to produce false claims (False Positives)
- Why Latent Truth Model (LTM)? Modeling two-sided quality to support multiple true values per entity for truth finding [VLDB'12]

Generating Implicit Negative Claims:



Truth Discovery: Effectiveness of Latent Truth Model

Experimental datasets: Large and real

- **Book Authors from abebooks.com** (1263 books, 879 sources, 48153 claims, 2420 book-author, 100 labeled)
- **Movie Directors from Bing** (15073 movies, 12 sources, 108873 claims, 33526 movie-director, 100 labeled)

Effectiveness of Latent Truth Model:

	Results on book data					*Results on movie data*				
	One-sided error			*Two-sided error*		*One-sided error*			*Two-sided error*	
	Precision	*Recall*	*FPR*	*Accuracy*	*F1*	*Precision*	*Recall*	*FPR*	*Accuracy*	*F1*
LTMinc	**1.000**	0.995	**0.000**	**0.995**	**0.997**	0.943	0.914	0.150	**0.897**	**0.928**
LTM	**1.000**	0.995	**0.000**	**0.995**	**0.997**	0.943	0.908	0.150	0.892	0.925
3-Estimates	**1.000**	0.863	**0.000**	0.880	0.927	0.945	0.847	0.133	0.852	0.893
Voting	**1.000**	0.863	**0.000**	0.880	0.927	0.855	0.908	0.417	0.821	0.881
TruthFinder	0.880	**1.000**	1.000	0.880	0.936	0.731	**1.000**	1.000	0.731	0.845
Investment	0.880	**1.000**	1.000	0.880	0.936	0.731	**1.000**	1.000	0.731	0.845
HubAuthority	**1.000**	0.322	**0.000**	0.404	0.488	**1.000**	0.620	**0.000**	0.722	0.765
AvgLog	**1.000**	0.169	**0.000**	0.270	0.290	**1.000**	0.025	**0.000**	0.287	0.048
LTMpos	0.880	**1.000**	1.000	0.880	0.936	0.731	**1.000**	1.000	0.731	0.845
PooledInvestment	**1.000**	0.142	**0.000**	0.245	0.249	**1.000**	0.025	**0.000**	0.287	0.048

- Model source quality in other data integration tasks, e.g. entity resolution.
- Trustworthiness in multi-genre networks (text-rich networks, social networks, etc.)

Challenge III: Extending the Horizon of the Study

- Going deep: **Meta (schema) level analysis ⇒ object level analysis**
  - Integration of statistical analysis with rich network topology
- Going broad: **Broaden the scope at meta-level**
  - Star schema ⇒ Entity-relationship schema
- **OLAP mining on multi-dimensional information networks**
  - E.g., authors ⇒ institutions; conferences ⇒ research subareas
- **Mining mission-based or user-relevant hidden networks**
  - Only a portion of multi-networks relevant to a task/query
- Information harvesting: **Discovery-driven similarity queries**
- **Mining cyber-physical networks** (networks with spatiotemporal, text, sensor, image/video/multimedia data and streams)

Outline

- Why Mining Heterogeneous Information Networks?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - RankClus: Ranking-Based Clustering in InfoNet
  - RankClass: Ranking-Based Classification in InfoNet
- Meta Path: A Key to Mining Heterogeneous Information Networks
  - PathSim: A New Metric for Finding Similar Objects in Heterogeneous Networks
  - PathPredict: Relationship Prediction in Info. Networks
  - Path-Selection: A User-Guided Learning Approach
- Challenges in Mining Heterogeneous Info. Networks
- Conclusions

Conclusions

- **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, …
- **Surprisingly rich knowledge can be mined from such structured heterogeneous info. networks**
  - Clustering, ranking, classification, data cleaning, trust analysis, role discovery, similarity search, relationship prediction, ……
  - Meta path holds a key to effective mining and exploration!
- **Knowledge is power, but knowledge is hidden in massive, but "relatively structured" nodes and links!**
- **Much more to be explored in information network mining!**

References of the Talk

- Y. Sun and J. Han, ***Mining Heterogeneous Information Networks: Principles and Methodologies***, Morgan & Claypool Publishers, 2012
- J. Han, Y. Sun, X. Yan, and P. S. Yu, "***Mining Heterogeneous Information Networks***", SIGMOD'10, KDD'10, ICDE'12 tutorials.
- M. Ji, J. Han, and M. Danilevsky, "***Ranking-Based Classification of Heterogeneous Information Networks***", KDD'11.
- 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.
- Tim Weninger, Marina Danilevsky, et al., "**WinaCS: Construction and Analysis of Web-Based Computer Science Information Networks**", ACM SIGMOD'11 (system demo)
- Y. Sun, J. Han, C. C. Aggarwal, N. Chawla, "**When Will It Happen? Relationship Prediction in Heterogeneous Information Networks**", WSDM'12