IEEE/ACM Intl. Conf. on ASONAM, Istanbul, Aug., 2012

On the Power of Mining Heterogeneous Information Networks

Yizhou Sun[†] Jiawei Han[†] Xifeng Yan [§] Philip S. Yu[‡]

[†]University of Illinois at Urbana-Champaign

[§] University of California at Santa Barbara

[‡]University of Illinois at Chicago

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

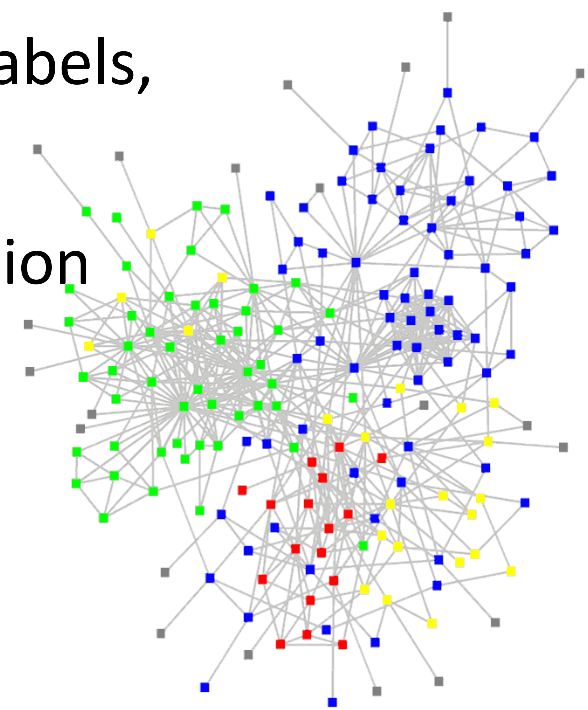
August 27, 2012

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

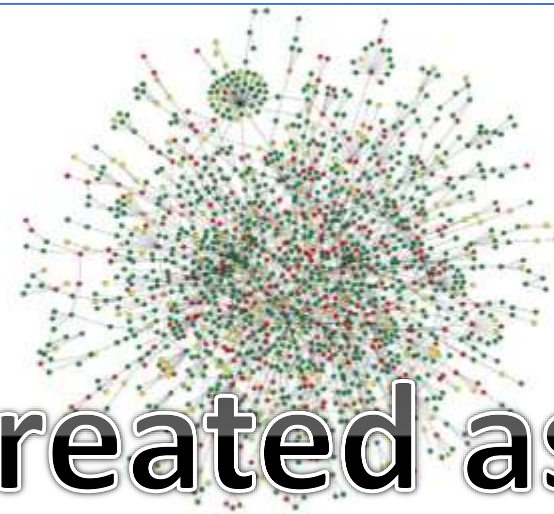
What Are Information Networks?

- Information network: A network where each node represents an entity (e.g., actor in a social network) and each link (e.g., tie) a relationship between entities
    - Each node/link may have attributes, labels, and weights
    - Link may carry rich semantic information



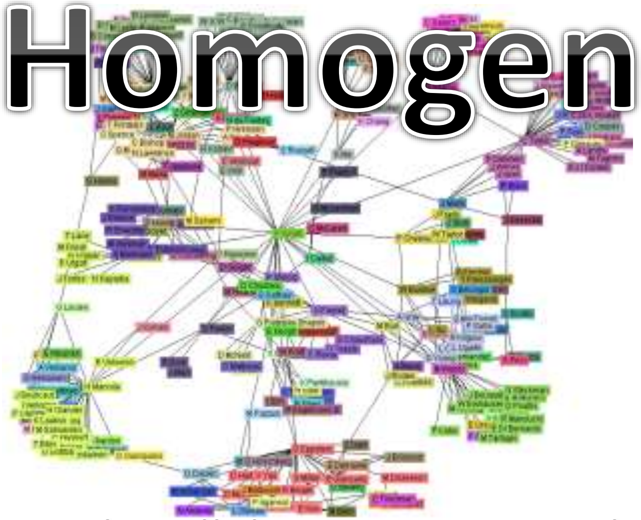
Information Networks Are Everywhere

Social Networking Websites

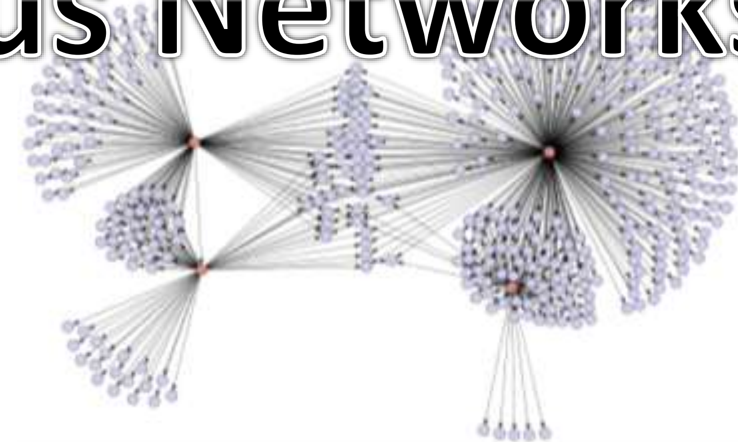


Biological Network: Protein Interaction

They are all treated as Homogeneous Networks!



Research Collaboration Network

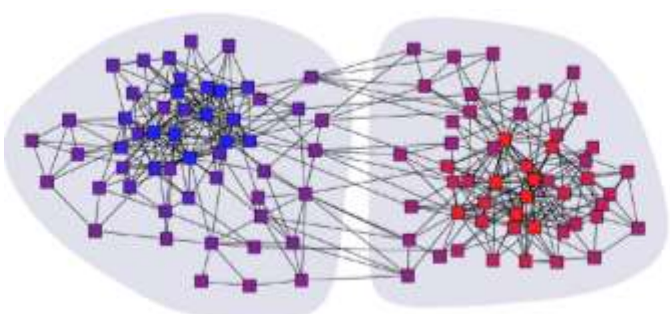


Product Recommendation Network via Emails

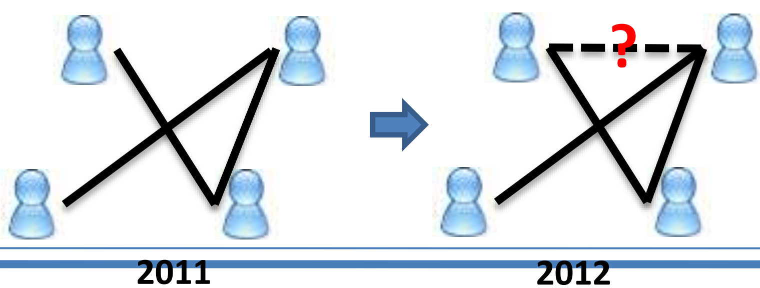
Homogeneous Information Networks

- Single object type and single link type
  - Link analysis based applications

Ranking web pages [Brin and Page, 1998]



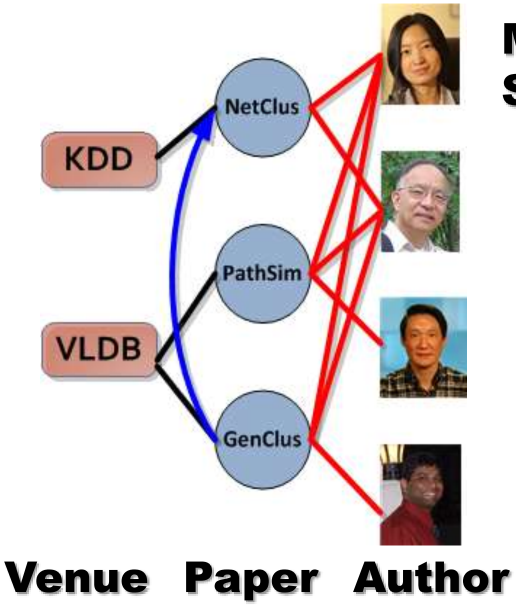
Clustering books about politics [Newman, 2006]



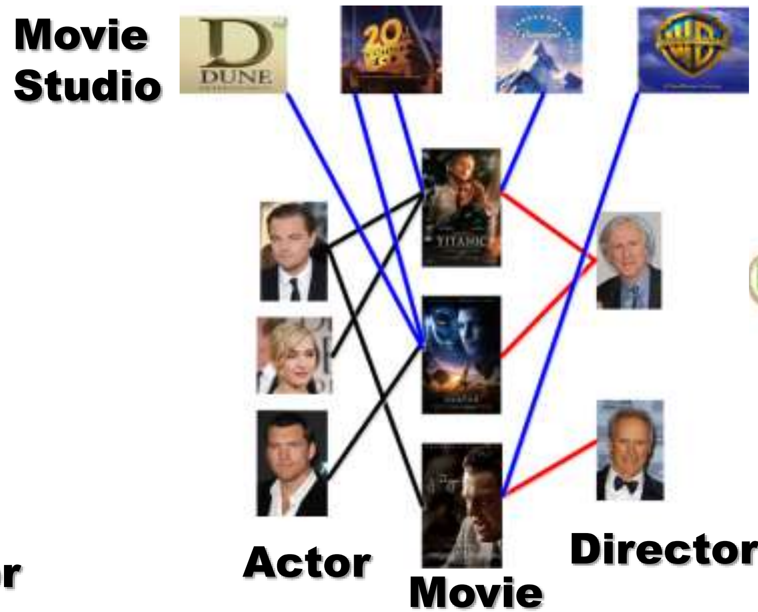
Link Prediction [Kleinberg, 2003]

Heterogeneous Information Networks

- Multiple object types and/or multiple link types



DBLP Bibliographic Network



The IMDB Movie Network

The Facebook Network

1. Homogeneous networks are ***Information loss*** projection of heterogeneous networks!
2. ***New problems*** are emerging in heterogeneous networks!

Directly Mining information richer heterogeneous networks

Heterogeneous Networks Are Ubiquitous

- Healthcare
  - Doctor, patient, disease, treatment

- Content sharing websites
  - Video, image, user, comment
- E-Commerce
  - Seller, buyer, product, review
- News
  - Person, organization, location, text

What Can be Mined from Heterogeneous Networks?

- DBLP: A Computer Science bibliographic database

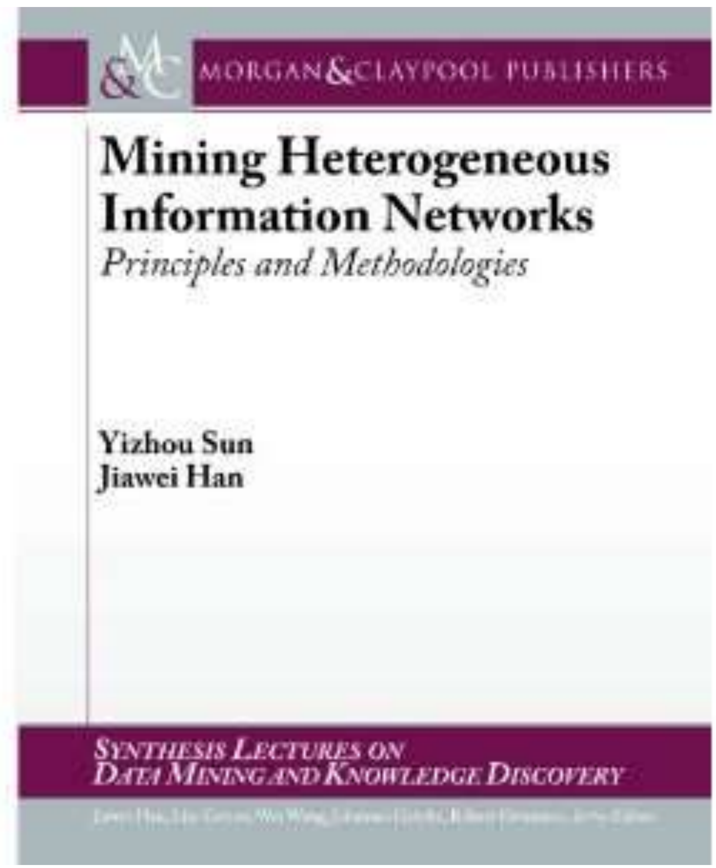
26	Yizhou Sun, Jiawei Han, Charu C. Aggarwal, Nitesh V. Chawla: When will it happen?: relationship prediction in heterogeneous information networks. WSDM 2012: 663-672

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

Knowledge hidden in DBLP Network	Mining Functions	Publications
How are CS research areas **structured**?	Clustering	EDBT'09, KDD'09, ICDM'09
Who are the **leading** researchers on Web search?	Ranking	EDBT'09, KDD'09,
Who are the **peer** researchers of Jure Lescovec?	Similarity Search	VLDB'11
Whom **will** Christos Faloutsos **collaborate with** in the future?	Relationship Prediction	ASONAM'11
Whether **will** an author **publish** a paper in KDD, and **when**?	Relationship Prediction with Time	WSDM'12
Which types of **relationships** are most **influential** for an author to decide her topics?	Relation Strength Learning	VLDB'12, KDD'12

Principles of Mining Heterogeneous Information Networks

- **Principle 1**: Use Holistic Network Information
  - Study information propagation across different types of objects and links
- **Principle 2**: Explore Network Meta Structure
  - Meta-path-based similarity search and mining
- **Principle 3**: User-Guided Exploration
  - Relation strength-aware mining with user guidance

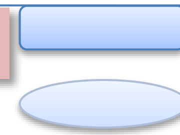


Outline

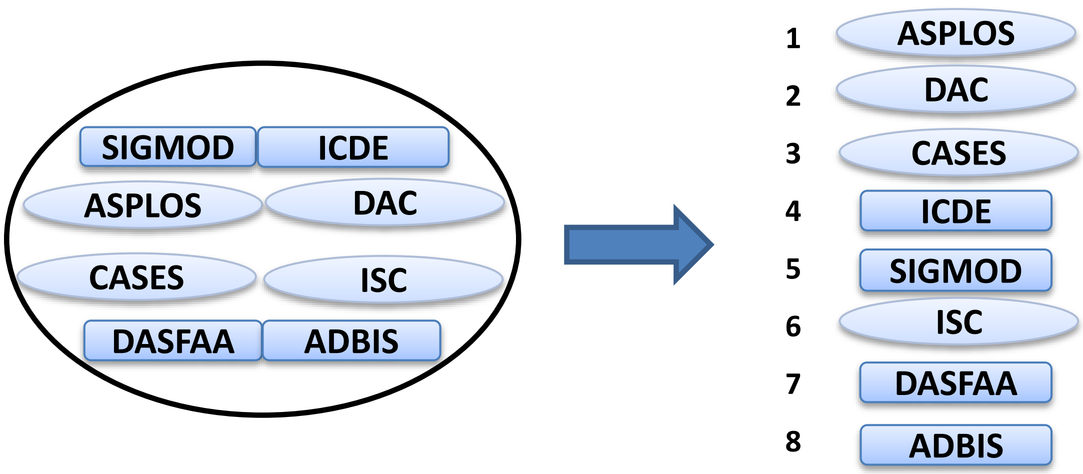
- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

Ranking and Clustering: Two Critical Functions

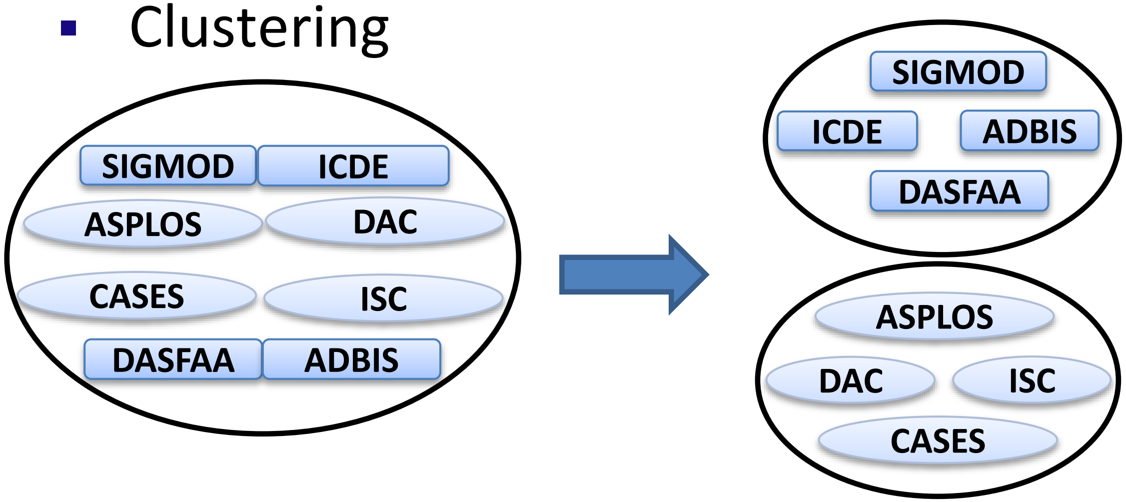
Comparing apples and oranges?

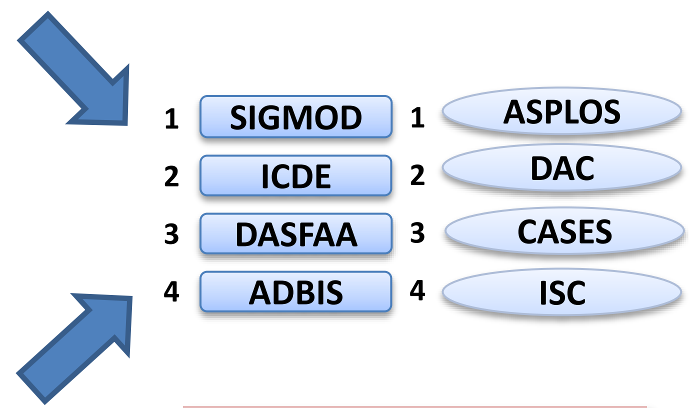


- Ranking



- Clustering



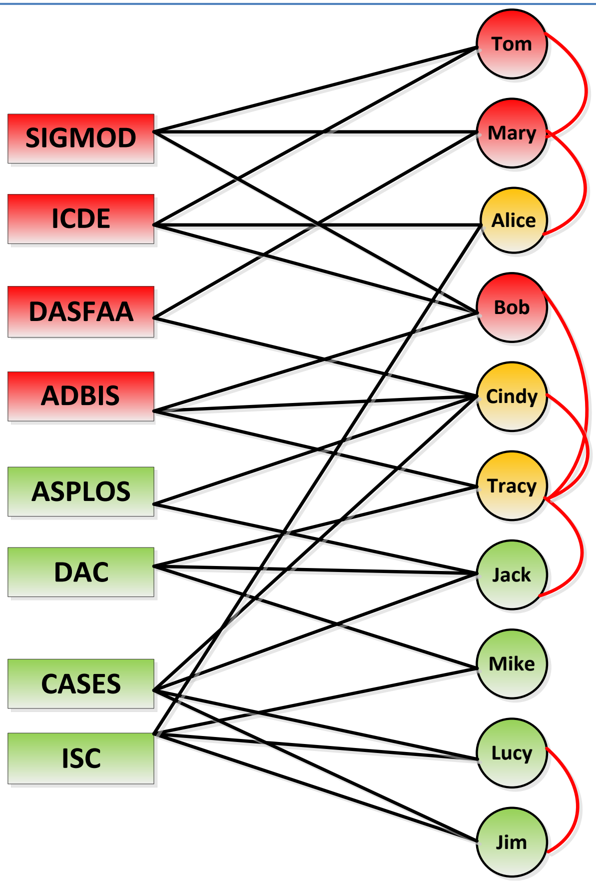


A better solution: Integrating clustering with ranking

Not distinguishing objects in each cluster?

RankClus: Integrating Clustering with Ranking [Sun et al., EDBT'09]

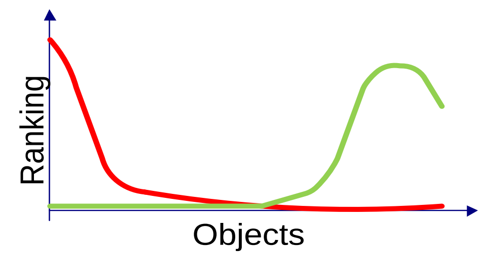
- A case study on bi-typed DBLP network
  - Links exist between
    - Conference (X) and author (Y)
    - Author (Y) and author (Y)
  - A matrix denoting the weighted links
    - $W = \begin{bmatrix} W_{XX} & W_{XY} \\ W_{YX} & W_{YY} \end{bmatrix}$
  - Goal:
    - Clustering and ranking conferences via authors
      - **Simple solution: Project the bi-typed network into homogeneous conference network + spectral clustering [Shi & Malik, 2000]**



Idea: Ranking and Clustering Mutually Enhance Each Other

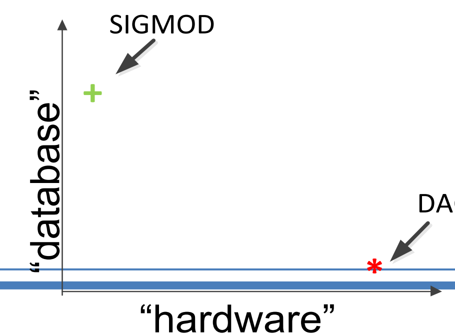
- Better clustering => Conditional ranking distributions are more distinguishing from each other
  - Conditional ranking distribution serves as the feature of each cluster

P(•|area = "database") vs. P(•|area = "hardware")



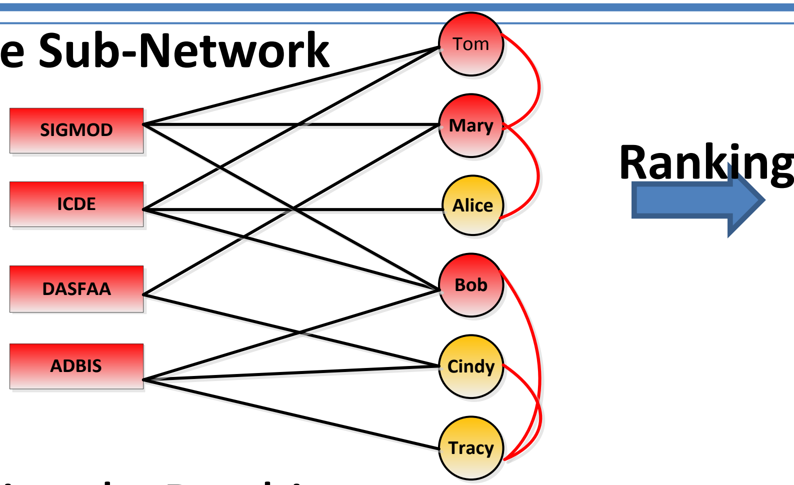
- Better ranking => Better metric for objects can be learned from the ranking for better clustering
  - Posterior probabilities for each object in each cluster serves as the new metric for each object

(P(area = "database"|SIGMOD), P(area = "hardware"|SIGMOD))



Simple Ranking vs. Authority Ranking

Database Sub-Network



P(SIGMOD|area = “database”)?
P(Tom|area = “database”)?

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

What about an author publishing 100 papers in low reputation 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

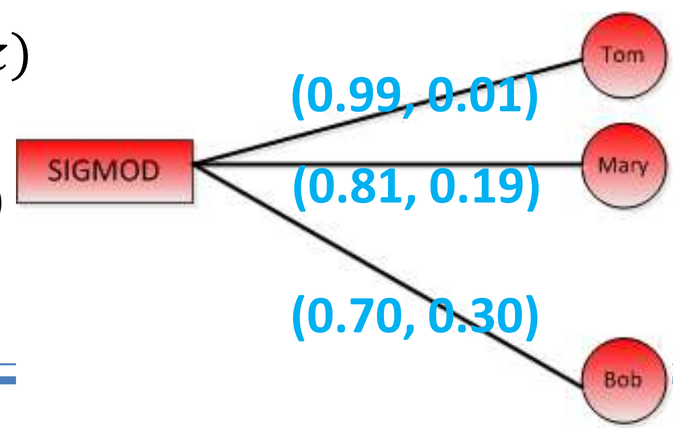
$$\vec{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i,j)\vec{r}_Y(j)$$

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

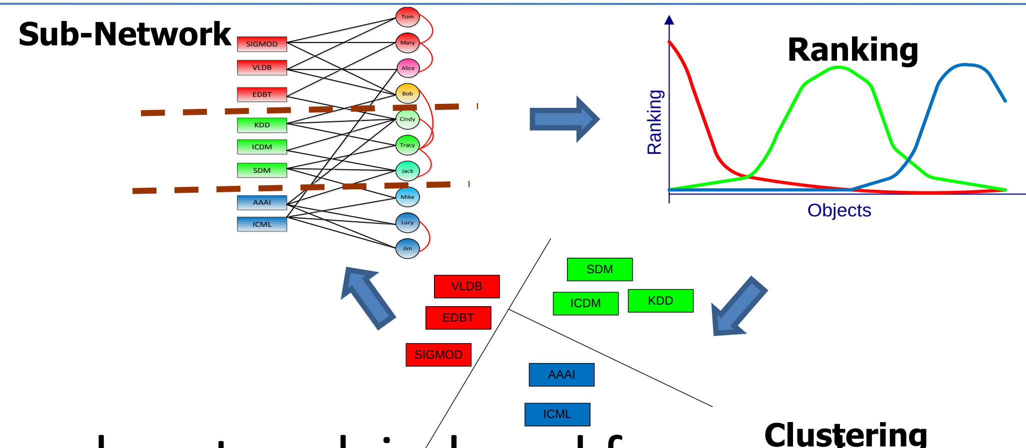
$$\vec{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j)\vec{r}_X(j) + (1-\alpha) \sum_{j=1}^{n} W_{YY}(i,j)\vec{r}_Y(j)$$

Generating New Measure Space

- Input: Conditional ranking distributions for each cluster
  - $P_X(i|k)$: $e.g.,\ P_X(SIGMOD|area = \text{"database"})$
- Output: Each conference *i* is mapped into a new measure space
  - $i: \left(\pi_{i,1}, \ldots, \pi_{i,K}\right), where\ \pi_{i,k} = P_X(k|i)$
    - E.g., SIGMOD: $(P(\text{"}database\text{"}|SIGMOD), P(\text{"}hardware\text{"}|SIGMOD))$
- Solution
  - $P_X(k|i) \propto P(k) \times P_X(i|k)$
  - Calculate cluster size $P(k)$
    - Maximize the log-likelihood of generating all the links
      - $P(i,j) = \sum_k P(k) \times P_X(i|k) \times P_Y(j|k)$
    - EM algorithm
      - $P(k|i,j) \propto P(k) \times P_X(i|k) \times P_Y(j|k)$
      - $P(k) \propto \sum_{ij} W_{XY}(i,j) P(k|i,j)$

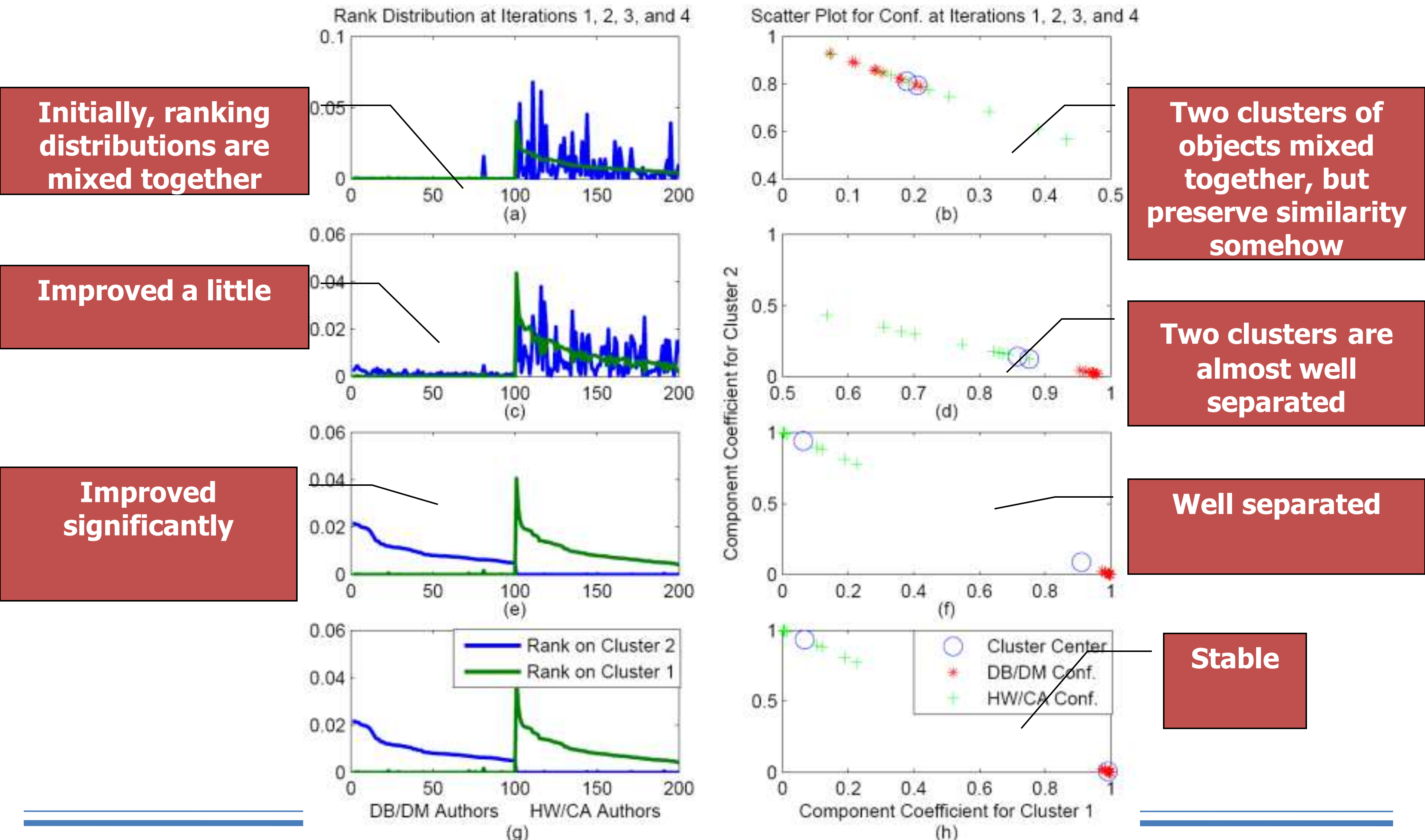


The Algorithm Framework



- Step 0: Initialization
  - Randomly partition
- Step 1: Ranking
  - Ranking objects in each sub-network induced from 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

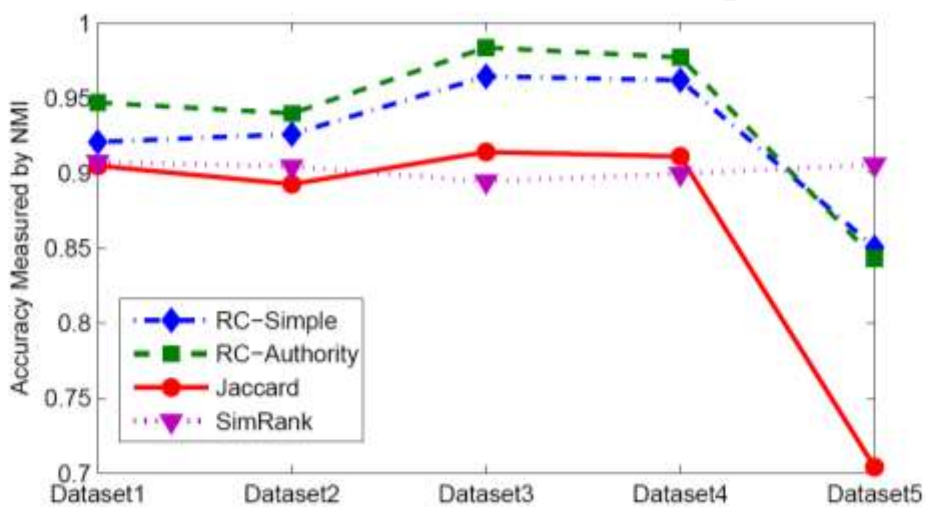
Step-by-Step Running Case Illustration



Clustering and Ranking CS Conferences by 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

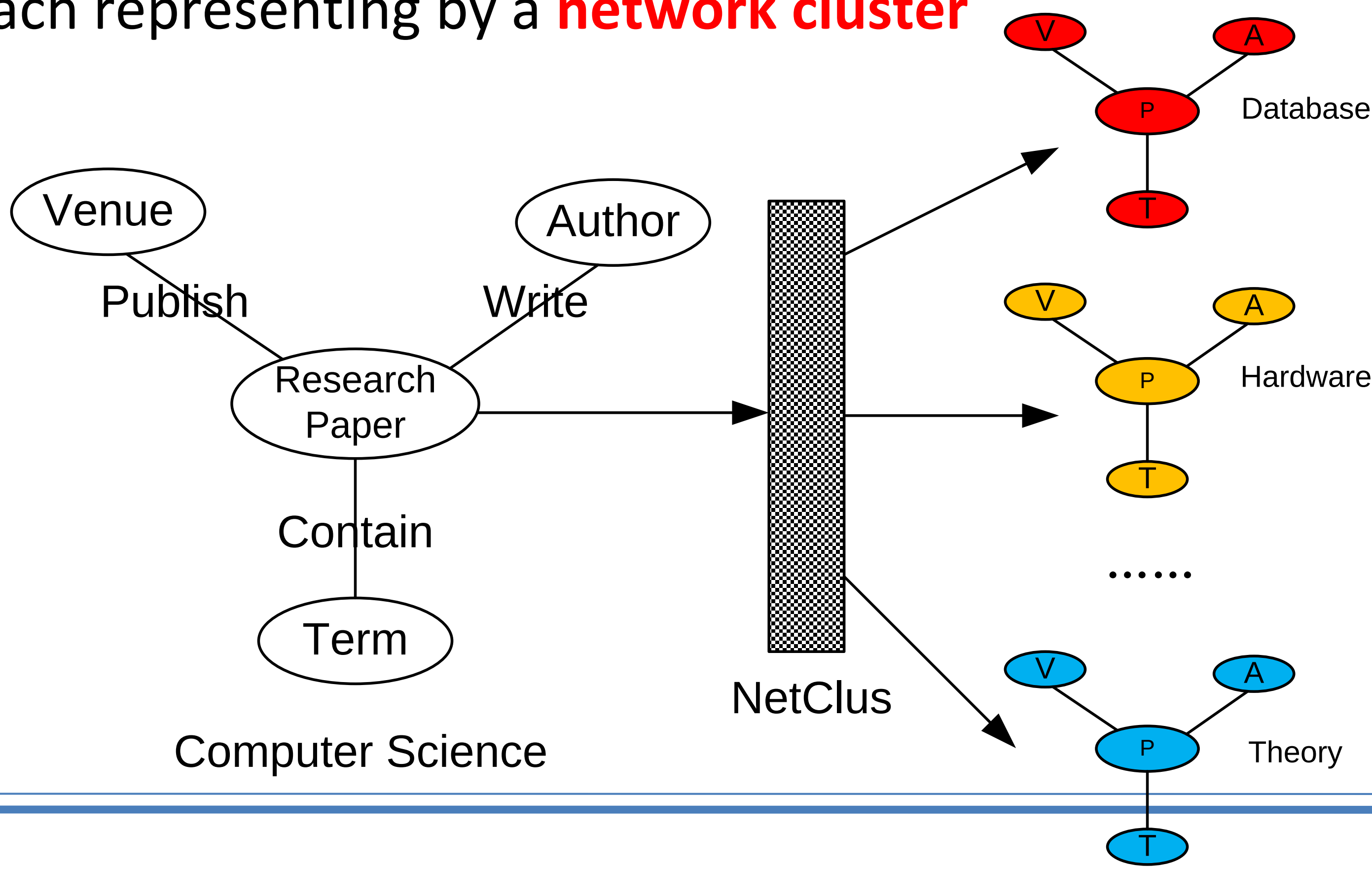
Top-10 conferences in 5 clusters using RankClus in DBLP



RankClus outperforms *spectral clustering [Shi and Malik, 2000]* algorithms on projected homogeneous networks

NetClus [Sun et al., KDD'09]: Beyond Bi-Typed Networks

- Beyond bi-typed information network
  - A Star Network Schema [**richer information**]
- Split a network into different layers
  - Each representing by a **network cluster**



Multi-Typed Networks Lead to Better Results

- The network cluster for database area: Conferences, Authors, and Terms
  - Better clustering and ranking than RankClus

Conference	Rank Score	Author	Rank Score	Term	Rank Score
SIGMOD	0.315	Michael Stonebraker	0.0063	database	0.0529
VLDB	0.306	Surajit Chaudhuri	0.0057	system	0.0322
ICDE	0.194	C. Mohan	0.0053	query	0.0313
PODS	0.109	Michael J. Carey	0.0052	data	0.0251
EDBT	0.046	David J. DeWitt	0.0051	object	0.0138
CIKM	0.019	H. V. Jagadish	0.0043	management	0.0113
...	...	...	...	...	...

  - NetClus vs. RankClus: **16%** higher accuracy on conference clustering in terms of Normalized Mutual Information

Impact of RankClus Methodology

- RankCompete [Cao et al., WWW'10]
  - Extend to the domain of web images
- RankClus in Medical Literature [Li et al., Working paper]
  - Ranking treatments for diseases
- RankClass [Ji et al., KDD'11]
  - Integrate classification with ranking
- Trustworthy Analysis [Gupta et al., WWW'11] [Khac Le et al., IPSN'11]
  - Integrate clustering with trustworthiness score
- Topic Modeling in Heterogeneous Networks [Deng et al., KDD'11]
  - Propagate topic information among different types of objects
- …

Interesting Results from Other Domains

RankCompete: Organize images 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

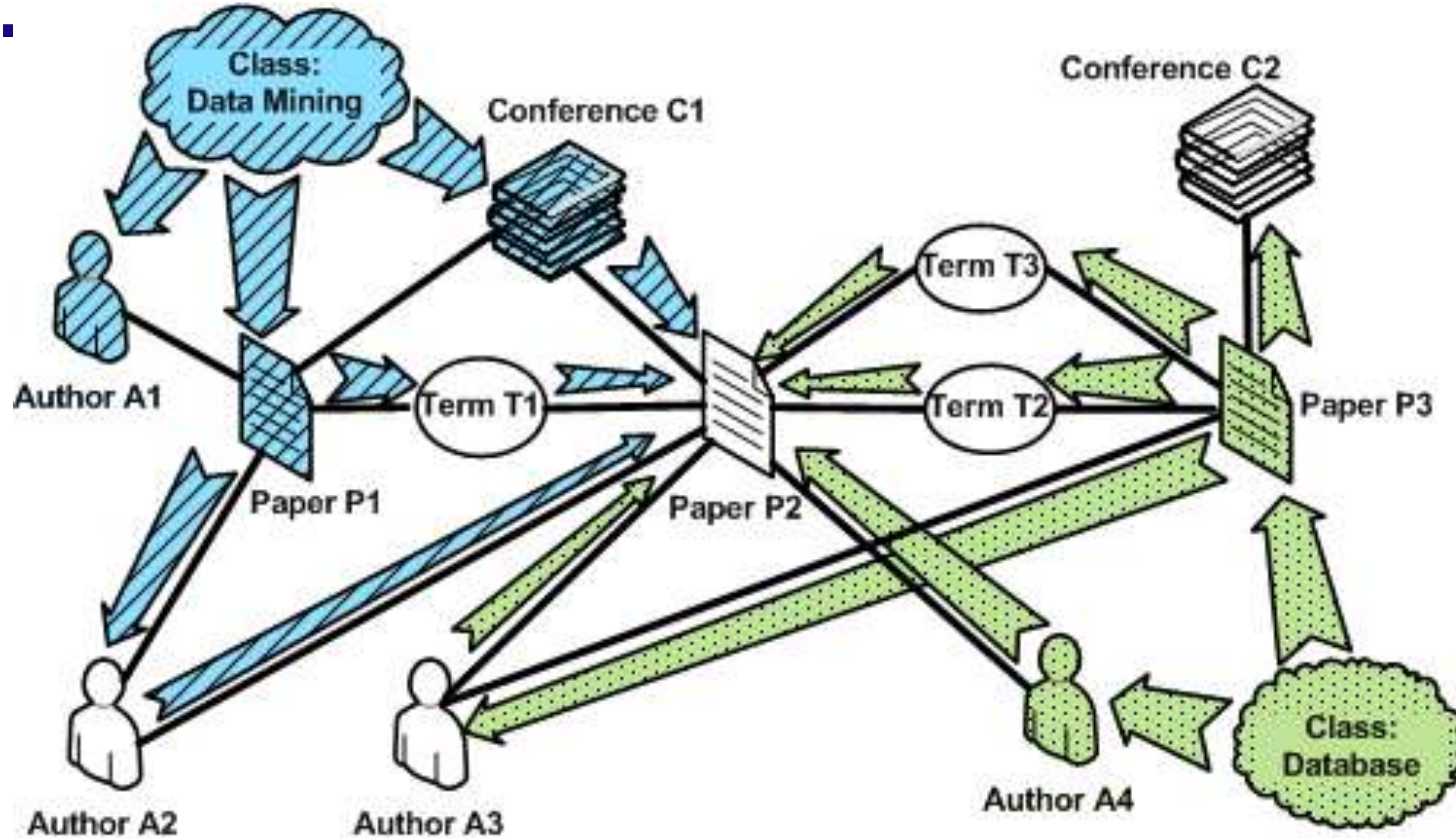
Rank treatments for AIDS from MEDLINE

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

Classification: Knowledge Propagation

▪



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

GNetMine: Graph-Based Regularization [Ji, PKDD'10]

- 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 [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

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

Similarity Search: Find Similar Objects in Networks [Sun et al., VLDB'11]

- 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 in heterogeneous information networks?

Existing Link-based Similarity Functions

- Existing similarity functions in networks
  - Personalized PageRank (P-PageRank) [Jeh and Widom, 2003]
  - SimRank [Jeh and Widom, 2002]

- Drawbacks
  - Do not distinguish object type and link type
  - Limitations on the similarity measures
    - To return highly visible objects or pure objects in the network

Network Schema and Meta-Path

Objects are connected together via different types of relationships!

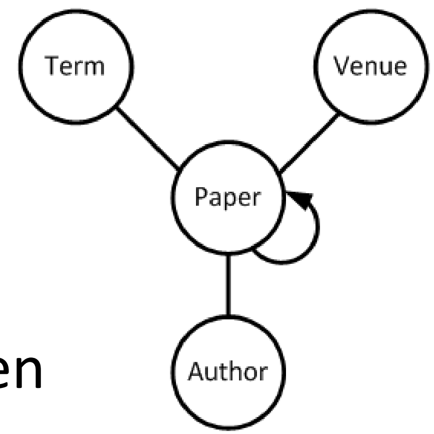
"Jim-P1-Ann"
"Mike-P2-Ann"
"Mike-P3-Bob"

Author-Paper-Author

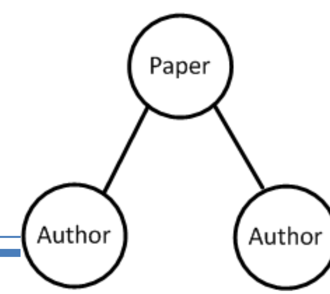
"Jim-P1-SIGMOD-P2-Ann"
"Mike-P3-SIGMOD-P2-Ann"
"Mike-P4-KDD-P5-Bob"

Author-Paper-Venue-Paper-Author

- Network schema
  - Meta-level description of a network

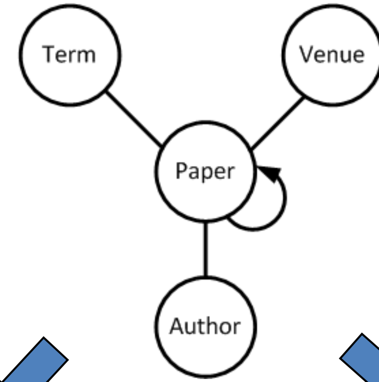


- Meta-Path
  - **Meta-level description** of a path between two objects
  - **A path** on network schema
  - Denote an existing or concatenated **relation** between two object types



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

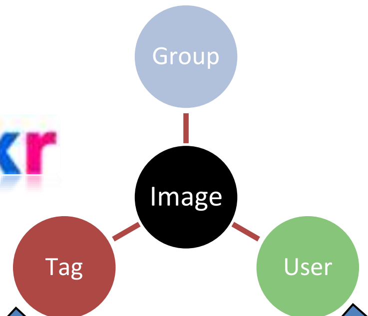
Meta-Path: *Author-Paper-Venue-Paper-Author*

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  ?

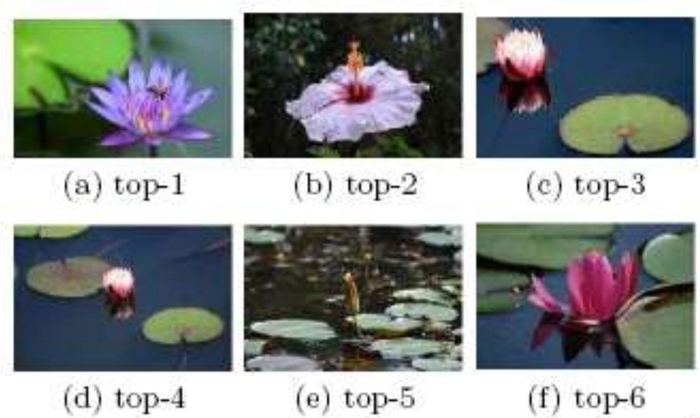


Evaluate the similarity between images according to their linked tags

Evaluate the similarity between images according to tags and groups

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

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



PathSim: Similarity in Terms of "Peers"

- Why peers?
  - Strongly connected, while **similar visibility**

Amazon Kindle

B&N Nook

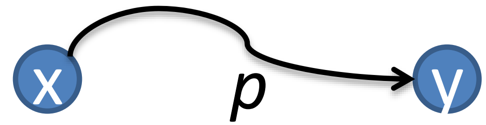
Sony Reader

Kobo eReader

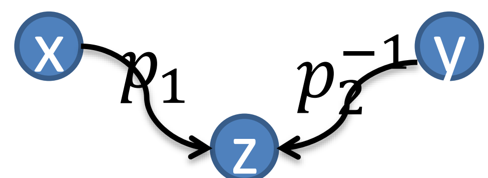
- In addition to meta-path
  - Need to consider **similarity measures**

Limitations of Existing Similarity Measures

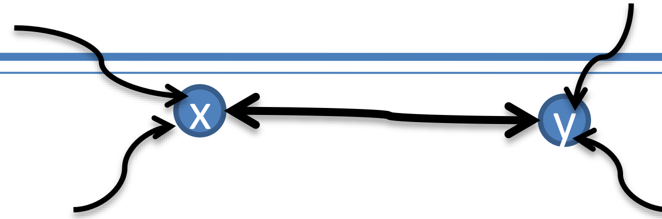
- Random walk (RW)
  - $s(x,y) = \sum_{p \in \mathcal{P}} Prob(p)$
  - Used in **Personalized PageRank (P-PageRank)**
  - Favor **highly visible** objects
    - objects with large degrees



- Pairwise random walk (PRW)
  - $s(x,y) = \sum_{(p_1,p_2) \in (\mathcal{P}_1,\mathcal{P}_2)} Prob(p_1)Prob(p_2^{-1})$
  - Used in **SimRank**
  - Favor **“pure”** objects
    - objects with highly skewed distribution in their in-links or out-links



Only PathSim Can Find Peers



- PathSim
  - Normalized path count between x and y following meta-path $\mathcal{P}$

$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{\underbrace{\boxed{|\{p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in \mathcal{P}\}|}}_{\text{Visibility of x}} + \underbrace{\boxed{|\{p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in \mathcal{P}\}|}}_{\text{Visibility of y}}}$$

  - Favor **"peers"**:
    - objects with strong connectivity and similar visibility under the given meta-path
  - Calculation
    - For $\mathcal{P}: A_1 - A_2 - \cdots - A_l - A_{l-1} - \cdots - A_1$
      - $M = W_{A_1A_2} W_{A_2A_3} \ldots W_{A_{l-1}A_l} W_{A_lA_{l-1}} \ldots W_{A_3A_2} W_{A_2A_1}$
      - $s(x,y) = \frac{2M_{xy}}{M_{xx}+M_{yy}}$
      - A co-clustering based pruning algorithm is provided
        - » 18.23% - 68.04% efficiency improvement over the baseline

Properties of PathSim

- Symmetric
  - $s(x, y) = s(y, x)$
- Self-Maximum
  - $s(x, y) \in [0,1], and\ s(x, x) = 1$
- Balance of visibility
  - $s(x, y) \leq \frac{2}{\sqrt{M_{xx}/M_{yy}} + \sqrt{M_{yy}/M_{xx}}}$
    - M_{xx} is the number of path instances starting from *x* and ending with *x* following the given meta path
- Limiting behavior
  - If repeating a pattern of meta path infinite times, PathSim degenerates to authority ranking comparison

Long meta-path without introducing new relationships is not that helpful!

Find Academic Peers by PathSim

- 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

Meta-Path: A Key Concept for Mining Heterogeneous Networks

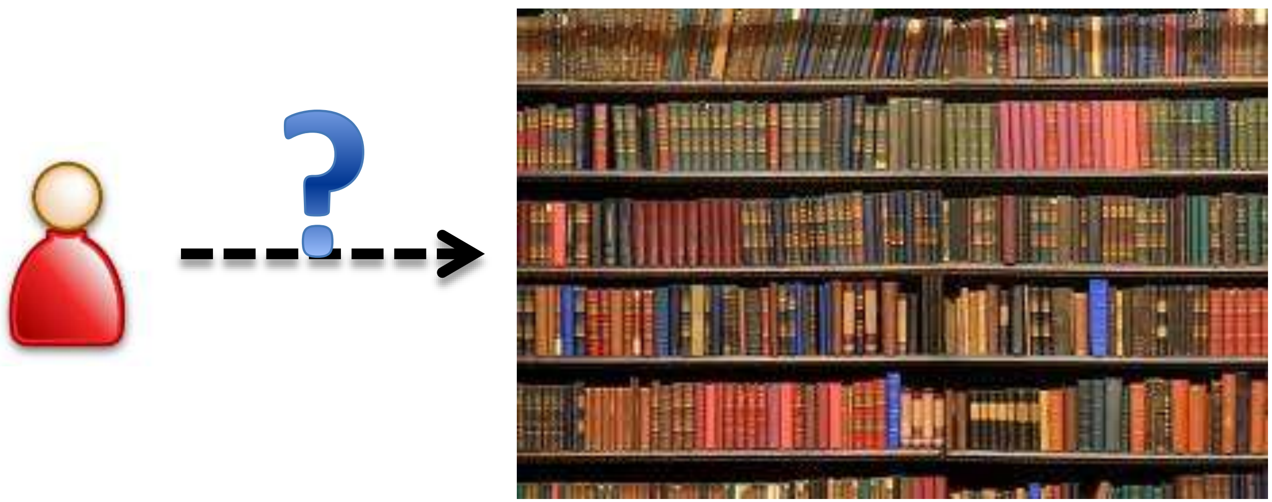
- Search and Query System
  - PathSim [Sun et al., VLDB'11]
  - User-guided similarity search [Yu et al., CIKM'12]
- Relationship Prediction
  - PathPredict [Sun et al., ASONAM'11]
    - Co-authorship prediction using meta-path-based similarity
  - PathPredict_when [Sun et al., WSDM'12]
    - When a relationship will happen
  - Citation prediction [Yu et al., SDM'12]
    - Meta-path + topic
- User-Guided Clustering
  - PathSelClus [Sun et al., KDD'12]
    - Meta-path selection + clustering
- Recommendation System
  - Ongoing work

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

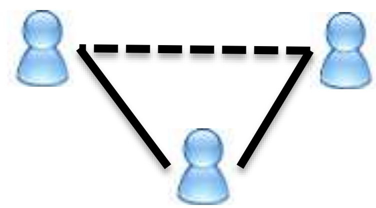
Meta-Path-Based Relationship Prediction

- Wide applications
  - Whom should I **collaborate** with?
  - Which paper should I **cite** for this topic?
  - Whom else should I **follow** on Twitter?
  - Whether Ann will **buy** the book “Steve Jobs”?
  - Whether Bob will **click** the ad on hotel?
  - …



Relationship Prediction vs. Link Prediction

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

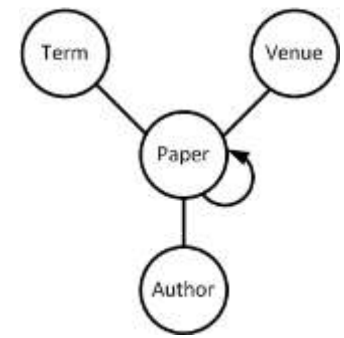


- Relationship prediction in heterogeneous networks
  - Target: Different types of relationships need different prediction models

  - Features: Different connection paths need to be treated separately!
    - **Meta-path-based approach** to define topological features.

PathPredict: Meta-Path Based Co-authorship Prediction in DBLP [Sun et al., ASONAM'11]

- 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

Social relations play very important role?

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

- Higher prediction accuracy than using projected homogeneous network
  - **11%** higher in prediction accuracy

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**

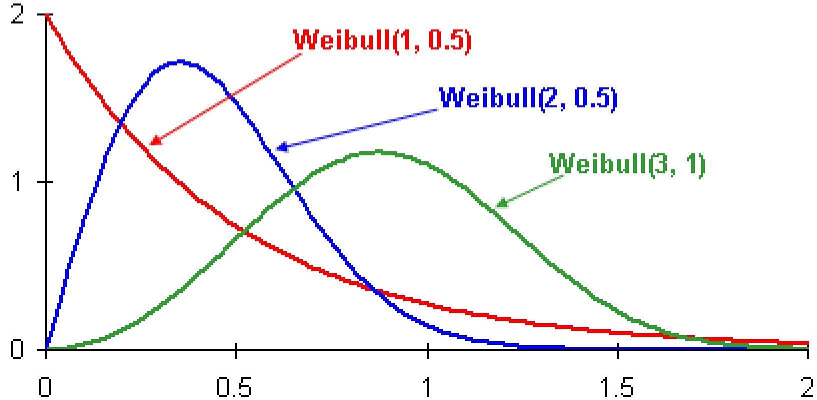
Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors!
(Feature collected in [1996, 2002]; Test period in [2003,2009])

When Will It Happen? [Sun et al., WSDM'12]

- From "whether" to "when"
  - "Whether": Will *Jim* rent the movie *"Avatar"* in Netflix?

Output: P(X=1)=?

  - "When": When will *Jim* rent the movie *"Avatar"*?



Output: A distribution of time!

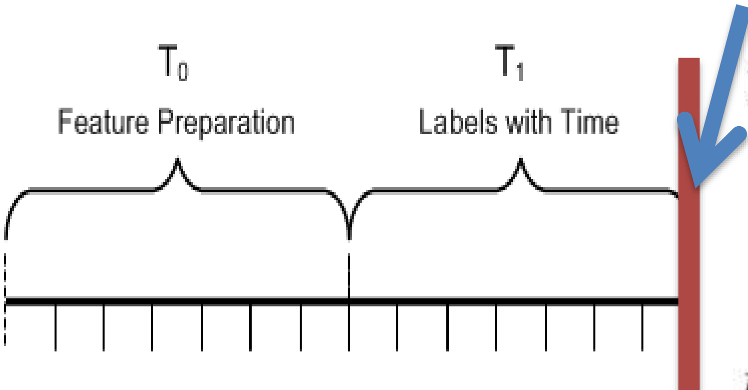
- What is the probability Jim will rent "Avatar" ***within 2 months***?
  - $P(Y \leq 2)$
- ***By when*** Jim will rent "Avatar" with 90% probability?
  - $t: P(Y \leq t) = 0.9$
- What is the ***expected time*** it will take for Jim to rent "Avatar"?
  - $E(Y)$

May provide useful information to supply chain management

The Relationship Building Time Prediction Model

- Solution
  - 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

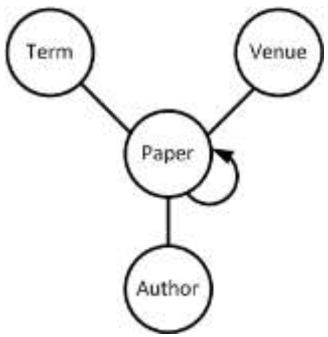
$$\log L = \sum_{i=1}^{n} \left(f_Y(y_i|\theta_i, \lambda) I_{\{y_i < T\}} + P(y_i \geq T|\theta_i, \lambda) I_{\{y_i \geq T\}} \right)$$

Generalized Linear Model
under Weibull Distribution Assumption

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

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

Under Weibull distribution assumption

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

Relation Strength-Aware Clustering of Heterogeneous InfoNet with Incomplete Attributes [Sun et al., VLDB'12]

- Content-Rich Heterogeneous information networks become increasingly popular
  - Heterogeneous links + (incomplete) attributes
  - Examples
    - Social media
    - E-Commerce
    - Cyber-physical system
- Soft clustering objects using both link information and attribute information
  - E-Commerce: customers, products, comments, …
  - Social websites: people, groups, books, posts, …
- Understanding the strengths for different relations in determining object's cluster

The Attribute-Based Clustering Problem

Age	Salary	Interests	Locations
20	10K	Sports, Music	Champaign, Boston
22	50K	Movie, Music, Football	New York
50	150K	Shopping, Books	Chicago
52	120K	Painting, Music	Boston
25	100K	Cooking, Books	Chicago, Seattle

Customer Segmentation According to Customer Profiles

Temperature (F)	Precipitation (mm)
60	5
70	15
56	0
80	12
85	15

Weather Pattern Clustering According to Weather Sensor Records

Incomplete Attributes

Object level: Missing data obs.

Age	Salary	Interests	Locations
20	10K	Sports, Music	Champaign, Boston
N/A	N/A	N/A	N/A
50	N/A	Shopping, Books	N/A
52	120K	N/A	Boston
N/A	100K	Cooking, Books	Chicago, Seattle

Customer Segmentation According to Customer Profiles

Schema level: Some type of objects only contains partial attribute types

Temperature (F)	Precipitation (mm)
N/A	5
N/A	15
N/A	20
80	N/A
85	N/A

P

Precip. Sensor Type

T

Temp. Sensor Type

Weather Pattern Clustering According to Weather Sensor Records

The Links Help!

Age	Salary	Interests	Locations
20	10K	Sports, Music	Champaign, Boston
N/A	N/A	N/A	N/A
50	N/A	Shopping, Books	N/A
52	120K	N/A	Boston
N/A	100K	Cooking, Books	Chicago, Seattle

Friendship
Family relationship
Schoolmate relationship
Colleague relationship
……

Customer Segmentation According to Customer Profiles

Temperature (F)	Precipitation (mm)
N/A	5
N/A	15
N/A	20
80	N/A
85	N/A

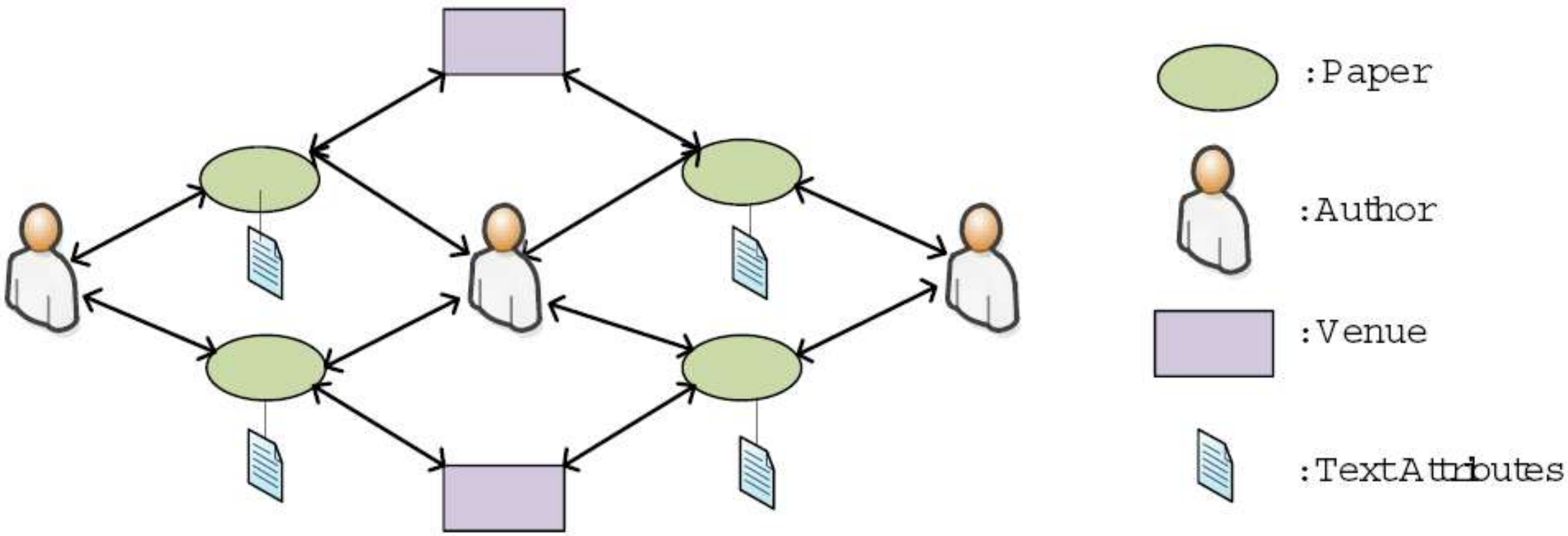
P
Precip. Sensor Type

T
Temp. Sensor Type

KNN relationship

Weather Pattern Clustering According to Weather Sensor Records

Example 1: Bibliographic Information Network



Link type:

- **Paper-Author, Paper-Venue, (Paper->Paper)**

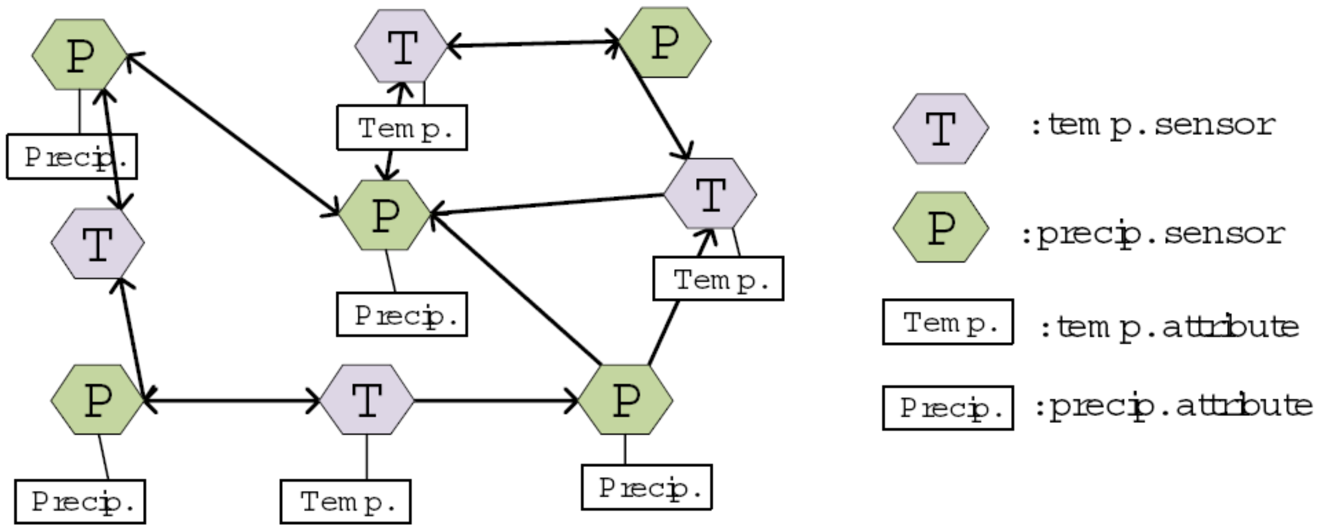
Attribute type:

- **Text attribute for Paper type**

Goal:

- **Clustering authors, venues, papers into different research areas**

Example 2: Weather Sensor Information Network



Link type:

- **T->P, T->T, P->P, P->T (According to KNN relationships)**

Attribute type:

- **Temperature attribute for T-typed sensors, Precipitation attribute for P-typed sensors**

Goal:

- **Clustering both types of sensors into different regional weather patterns**

Challenges

- Attributes are **incomplete** for objects
  - Not every type of objects contained the user specified attributes
    - E.g., Temperature typed sensors are only associated with temperature attributes
  - Missing value
    - E.g., some sensor may contain no observations due to malfunctioning
- Links are **heterogeneous**
  - Different types of links carry different importance in enhancing the quality of attribute-based clustering results
    - E.g., which type of links are more trustable to determine a person's political interest: friendship or person-like-book relationship?

Solution Overview

- Modeling attribute generation and structural consistency in a unified framework

$$p(\{\{v[X]\}_{v\in V_X}\}_{X\in\mathcal{X}}, \Theta|G,\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{X\in\mathcal{X}} p(\{v[X]\}_{v\in V_X}|\Theta,\boldsymbol{\beta})p(\Theta|G,\boldsymbol{\gamma})$$

  - Attribute generation as a mixture model
    - $p(\{v[X]\}_{v\in V_X}|\Theta,\boldsymbol{\beta}) = \prod_{v\in V_X}\prod_{x\in v[X]}\sum_{k=1}^{K}\theta_{v,k}p(x|\boldsymbol{\beta}_k)$
    - *v[X]: observed values for Attribute X on Object v*
    - *Θ: soft clustering membership matrix*
    - *$\boldsymbol{\beta}$: parameters associated with each mixture model component*
  - Structural consistency as a log-linear model
    - $p(\Theta|G,\boldsymbol{\gamma}) = \frac{1}{Z(\boldsymbol{\gamma})}\exp\{\sum_{e=\langle v_i,v_j\rangle\in E} f(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,e,\boldsymbol{\gamma})\}$
    - *$\boldsymbol{\gamma}$: relation strength vector*

The Objective Function and the Algorithm Overview

$$g(\Theta, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \boxed{\log \sum_{X \in \mathcal{X}} p(\{v[X]\}_{v \in V_X} | \Theta, \boldsymbol{\beta})} + \boxed{\log p(\Theta | G, \boldsymbol{\gamma})} - \boxed{\frac{||\boldsymbol{\gamma}||^2}{2\sigma^2}}$$

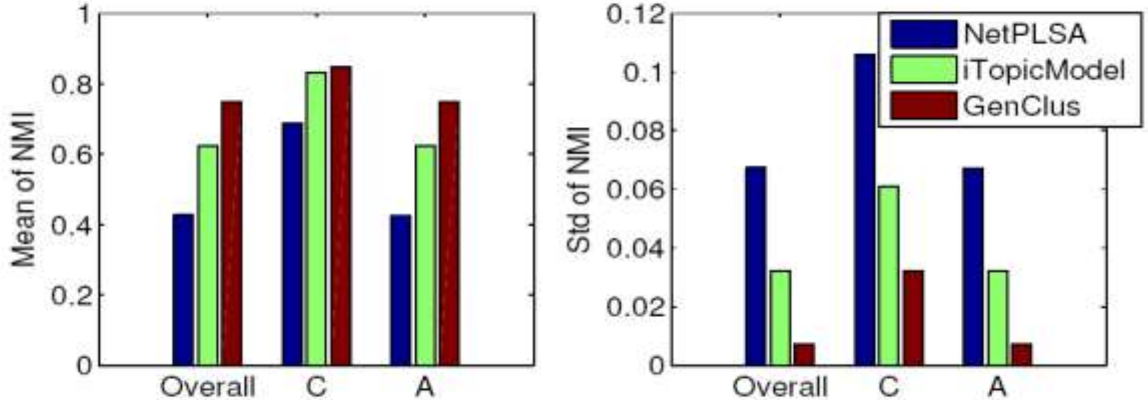
Attribute Generation

Structural Consistency

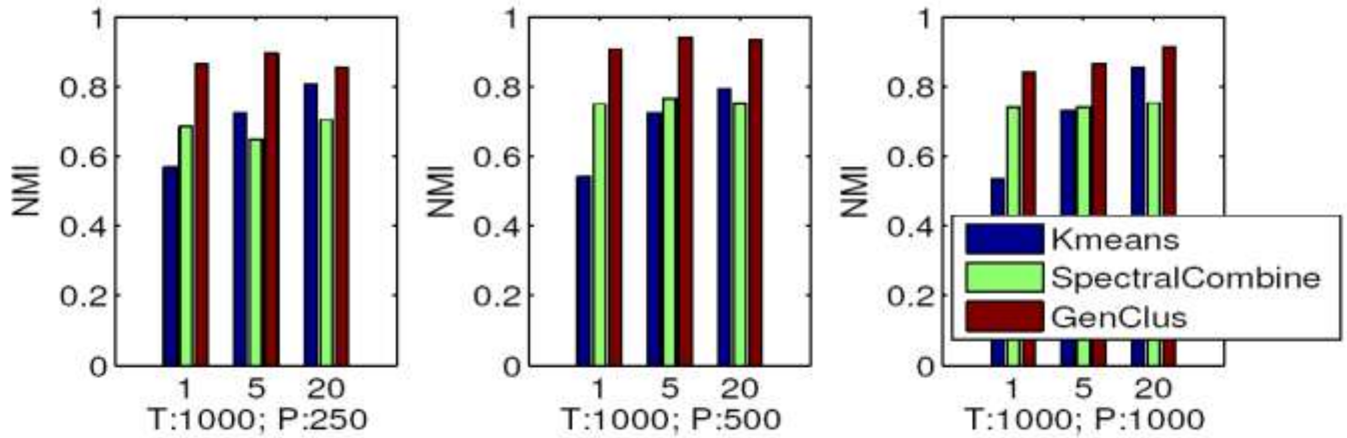
Regularization Term

- The clustering algorithm
  - Iterative algorithm
    - Step 1: Fix the relation strength and optimize the clustering result
      - Cluster optimization
    - Step 2: Fix the clustering result and optimize the relation strength
      - Relation strength learning

Higher Accuracy and More Stable Clustering Results

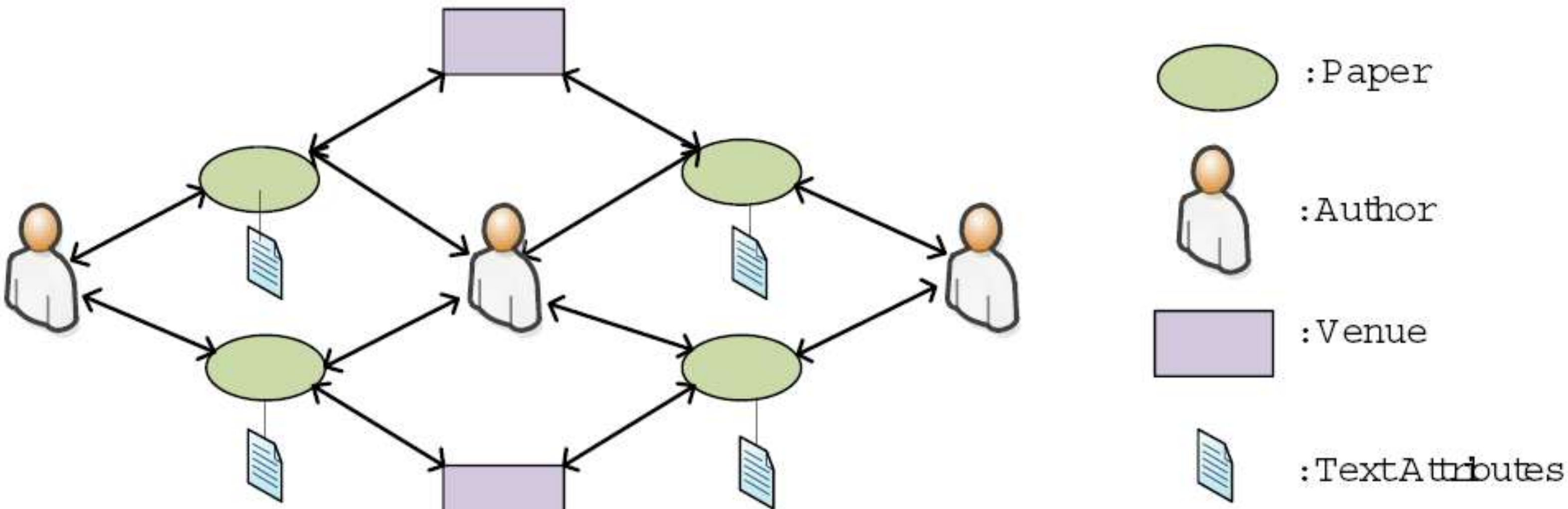


Clustering Accuracy Comparisons for AC

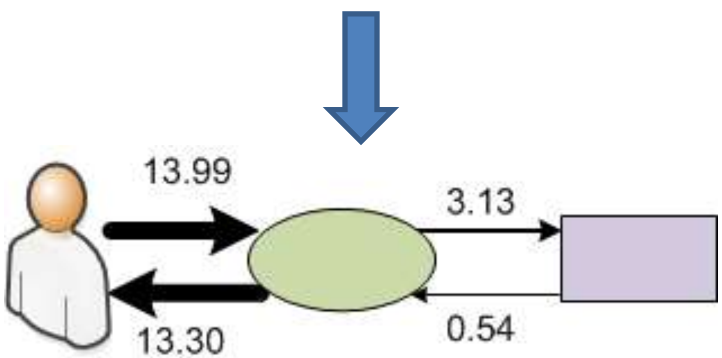


Clustering Accuracy Comparisons for Weather Sensor Network

Intuitive relation strength weights



DBLP Bibliographic Network



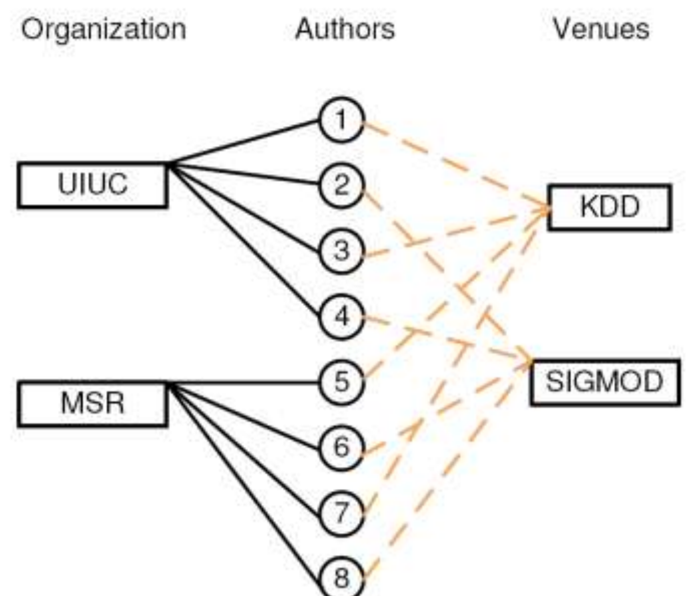
A paper's research area is more determined by its authors than its venue (13.30 vs. 3.13)

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

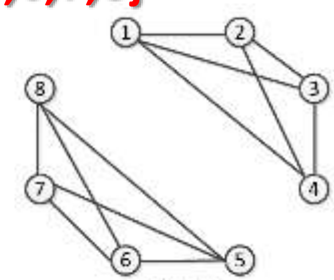
Why Meta-Path Selection? [Sun et al., KDD'12]

- Goal: Clustering authors based on their connection in the network



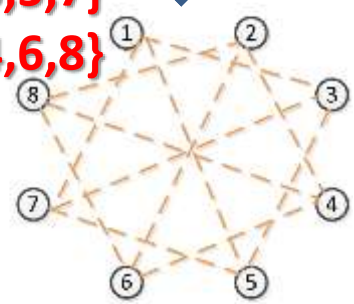
Which meta-path to choose?

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



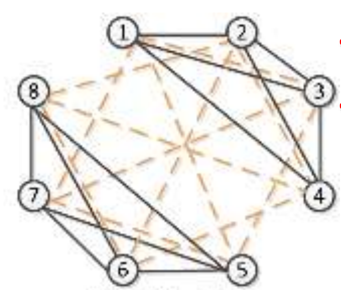
(a) AOA

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



(b) AVA

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

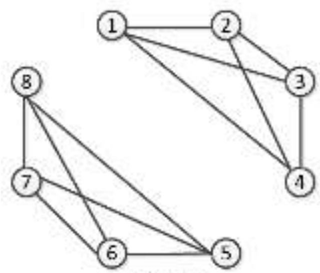


(c) AOA + AVA

The Role of User Guidance

- It is users' responsibility to specify their clustering purpose
  - Say, by giving seeds in each cluster

{1}
{5}



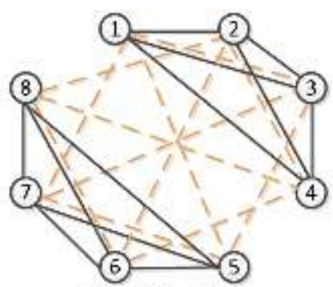
(a) AOA

+

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

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

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



(c) AOA + AVA

+

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

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

The Problem of User-Guided Clustering with Meta-Path Selection

- Input:
  - The target type for clustering: T
  - Number of clusters: K
    - **Seeds in *some* of the clusters: $\boldsymbol{L_1, L_2, \dots, L_K}$**
  - M Candidate meta-paths starting from T: $\boldsymbol{\mathcal{P}}_1, \boldsymbol{\mathcal{P}}_2, \dots, \boldsymbol{\mathcal{P}}_M$
- Output:
  - The quality weight for each candidate meta-path in the clustering process
    - α_m
  - The clustering results that are consistent with the user guidance
    - $\boldsymbol{\theta}_i$

Existing Link-based User-Guided Clustering Approaches

- Link-based clustering algorithms on homogeneous networks
  - Treat all types of links equally important (Zhu et al., 2003)

- Distinguish different relations in HIN, but use *ALL* the relations in the network
  - Do not distinguish different clustering tasks with different semantic meanings (Long et al., 2007)

The Probabilistic Model

- Part 1: Modeling the Relationship Generation
  - A good clustering result should lead to high likelihood in observing existing relationships
    - Keep in mind: higher quality relations should count more in the total likelihood
- Part 2: Modeling the Guidance from Users
  - The more consistent with the guidance, the higher probability of the clustering result
- Part 3: Modeling the Quality Weights for Meta-Paths
  - The more consistent with the clustering result, the higher quality weight

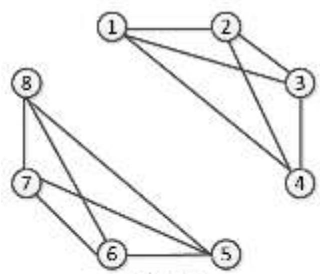
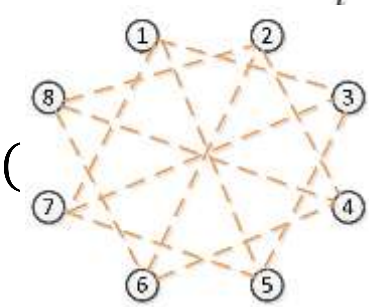
Objective Function

$$J = \sum_i \left(\sum_m \log P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m) + \sum_k \mathbf{1}_{\{t_i \in \mathcal{L}_k\}} \lambda \log \theta_{ik} \right)$$

Part 1: Modeling the Relationship Generation

- For each meta path $\boldsymbol{\mathcal{P}}_m$, let the relation matrix be W_m:
  - The relationship $\langle t_i, f_{j,m} \rangle$ is generated under a mixture of multinomial distributions
    - $\pi_{ij,m} = P(j|i,m) = \sum_k P(k|i)P(j|k,m) = \sum_k \theta_{ik}\beta_{kj,m}$
    - θ_{ik}: the probability that t_i belongs to Cluster *k*
    - $\beta_{kj,m}$: the probability that feature object $f_{j,m}$ appearing in Cluster *k*
  - The probability to observing all the relationships in $\boldsymbol{\mathcal{P}}_m$

$$P(W_m|\Pi_m, \Theta, B_m) = \prod_i P(\mathbf{w}_{i,m}|\boldsymbol{\pi}_{i,m}, \Theta, B_m) = \prod_i \prod_j (\pi_{ij,m})^{w_{ij,m}}$$

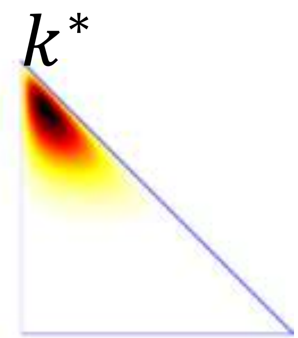
E.g., $P($  $|\Theta)$ $P($  $|\Theta)$

(a) AOA

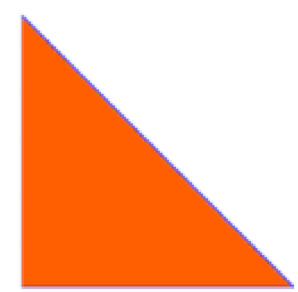
(b) AVA

Part 2: Modeling the Guidance from Users

- For each soft clustering probability vector θ_i:
  - Model it as generated from a Dirichlet prior
    - If t_i is labeled as a seed in Cluster k^*
      - $\theta_i \sim Dir(\lambda \boldsymbol{e}_{k^*} + \mathbf{1})$
        - » $\boldsymbol{e}_{k^*}$ is an all-zero vector except for item k^*, which is 1
        - » λ is the user confidence for the guidance



    - If t_i is not labeled in any cluster
      - $\theta_i \sim Dir(\mathbf{1})$
        - » The prior density is uniform, a special case of Dirichlet distribution



$$p(\boldsymbol{\theta}_i|\lambda) = \begin{cases} \prod_k \theta_{ik}^{\mathbf{1}_{\{t_i \in \mathcal{L}_k\}}\lambda} = \theta_{ik^*}^{\lambda}, & \text{if } t_i \text{ is labeled and } t_i \in \mathcal{L}_{k^*}, \\ 1, & \text{if } t_i \text{ is not labeled.} \end{cases}$$

Part 3: Modeling the Quality Weights for Meta-Paths

- Model quality weight α_m as the relative weight for each relationship in W_m
  - Observation of relationships: $W_m \rightarrow \alpha_m W_m$
- Further assume relationship generation with Dirichlet Prior: $\boldsymbol{\pi}_{i,m} \sim \mathrm{Dir}(\mathbf{1})$
- The best α_m: the most likely to generate current clustering-based parameters

Dirichlet Distribution

  - $$\alpha_m^* = \arg\max_{\alpha_m} \prod_i P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m)$$
    - when α_m is small, $\pi_{i,m}$ is more likely to be a uniform distribution
      - Random generated
    - when α_m is large, $\pi_{i,m}$ is more likely to be $\frac{w_{i,m}}{n_{i,m}}$, what we observed
      - Consistent with the observation

The Learning Algorithm

- An *Iterative algorithm* that the clustering result Θ and quality weight vector $\boldsymbol{\alpha}$ mutually enhance each other
  - Step 1: Optimize Θ given $\boldsymbol{\alpha}$
    - θ_i is determined by all the relation matrices with different weights α_m, as well as the labeled seeds

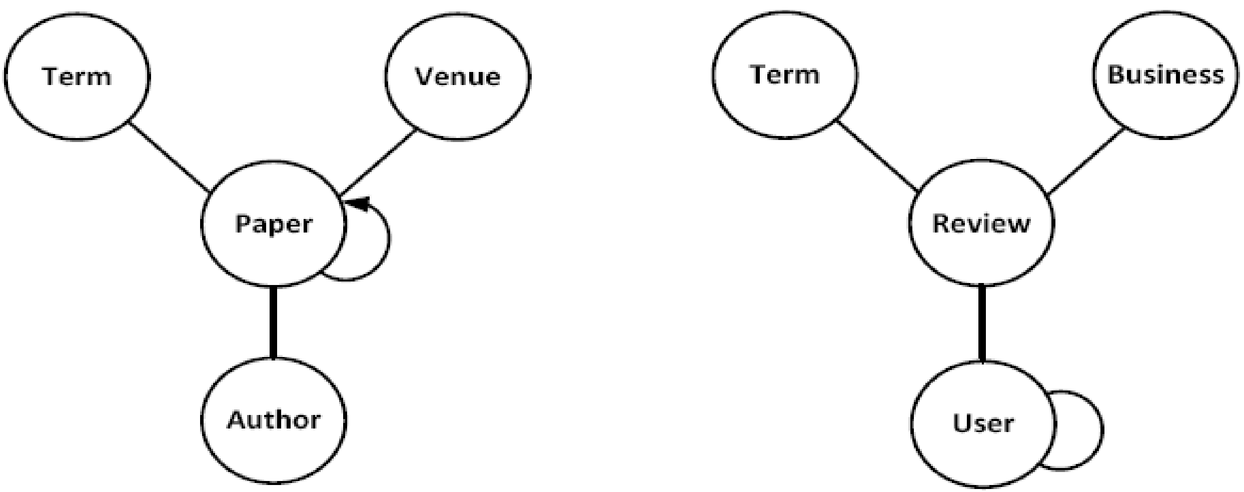
$$\theta_{ik}^t \propto \sum \alpha_m \sum w_{ij,m} p(z_{ij,m} = k|\Theta^{t-1}, B^{t-1}) + \mathbf{1}_{\{t_i \in \mathcal{L}_k\}}\lambda$$

  - Step 2: Optimize $\boldsymbol{\alpha}$ given Θ
    - In general, the higher likelihood of observing W_m given Θ, the higher α_m

$$\alpha_m^t = \alpha_m^{t-1} \frac{\sum_i \left(\psi(\alpha_m^{t-1} n_{im} + |F_m|) n_{i,m} - \sum_j \psi(\alpha_m^{t-1} w_{ij,m} + 1) w_{ij,m}\right)}{-\boxed{\sum_i \sum_j w_{ij,m} \log \pi_{ij,m}}}$$

Experiments

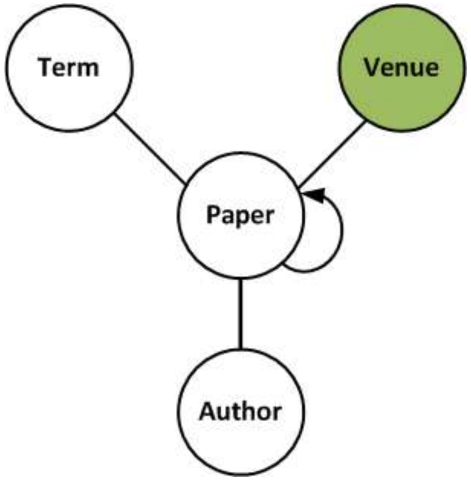
- Datasets
  - DBLP
    - Object Types: Authors, Venues, Papers, Terms
    - Relation Types: AP, PA, VP, PV, TP, PT
  - Yelp
    - Object Types: Users, Businesses, Reviews, Terms
    - Relation Types: UR, RU, BR, RB, TR, RT



(a) DBLP (b) Yelp

DBLP-T1: Clustering Venues According to Research Areas

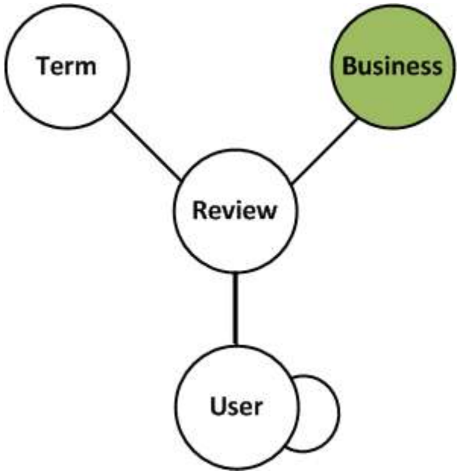
- Task:
  - Target objects: venues
  - Number of clusters: 4;
  - Candidate meta-paths: *V-P-A-P-V, V-P-T-P-V*
- Output:
  - Weights:
    - *V-P-A-P-V: 1576 (0.0017 per relationship)*
    - *V-P-T-P-V: 17001 (0.0003 per relationship)*
  - Clustering results:



#S	Measure	PathSelClus	LP	ITC	LP_voting	LP_soft	ITC_voting	ITC_soft
1	Accuracy	**0.9950**	0.6500	0.6900	0.6500	0.6650	0.6450	0.5100
	NMI	**0.9906**	0.6181	0.6986	0.6181	0.5801	0.5903	0.5316
2	Accuracy	1	0.7500	0.8450	0.7500	0.8200	0.8950	0.8700
	NMI	1	0.6734	0.7752	0.6734	0.7492	0.8321	0.7942

Yelp-T2: Clustering Restaurants According to Categories

- Task:
  - Target objects: restaurants
  - Number of clusters: 6;
  - Candidate meta-paths: *B-R-U-R-B, B-R-T-R-B.*
- Output:
  - Weights:
    - *B-R-U-R-B : 6000 (0.1716 per relationship, compared with 0.5864 for clustering shopping categories)*
    - *B-R-T-R-B: 2.9522*$\times 10^7$ *(0.0138 per relationship)*

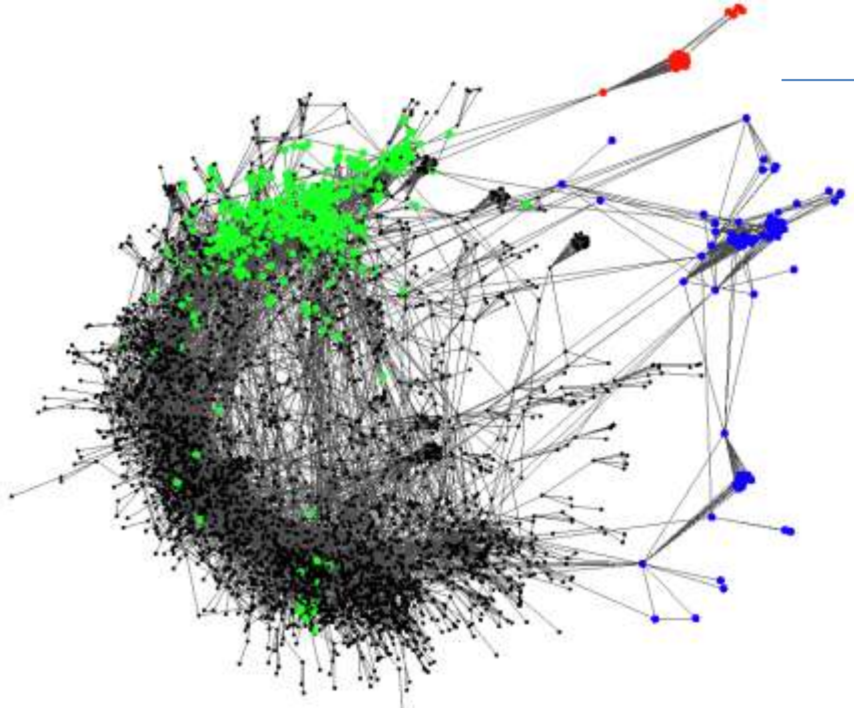


%S	Measure	PathSelClus	LP	ITC	LP_voting	LP_soft	ITC_voting	ITC_soft
1%	Accuracy	**0.7435**	0.1137	0.1758	0.2112	0.2112	0.2430	0.2022
	NMI	**0.6517**	0.0323	0.0178	0.0578	0.0578	0.2308	0.2490
2%	Accuracy	**0.8004**	0.1264	0.1910	0.2202	0.2202	0.2762	0.2792
	NMI	**0.6803**	0.0487	0.0150	0.0801	0.0801	0.2099	0.2907
5%	Accuracy	**0.8125**	0.2653	0.2200	0.2437	0.2437	0.3049	0.3240
	NMI	**0.6894**	0.1111	0.0220	0.1212	0.1212	0.2252	0.2692

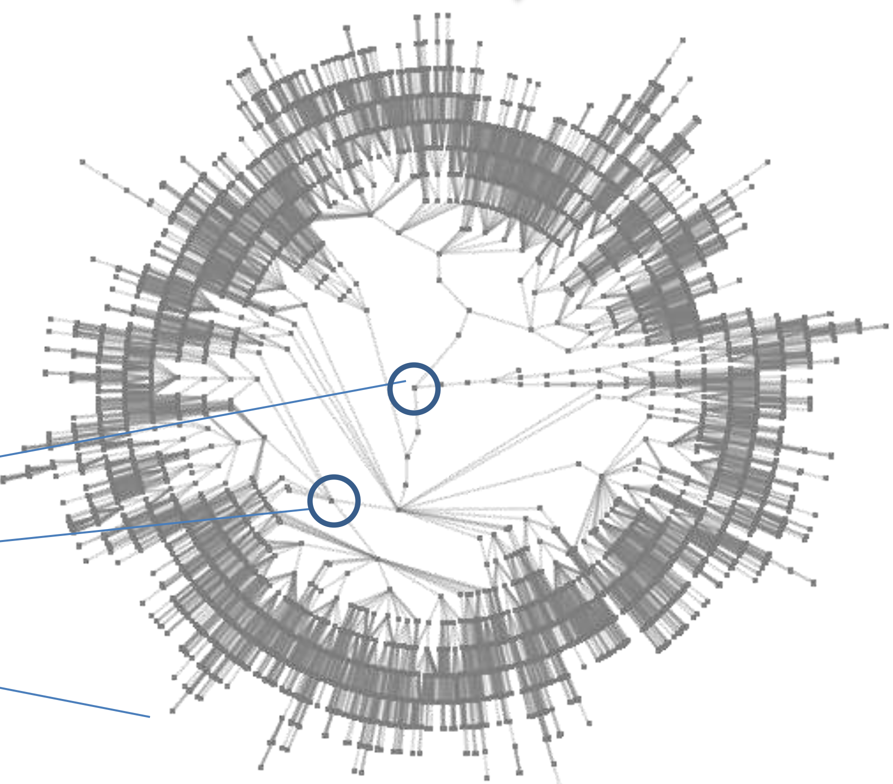
Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

1. Role Discovery in Network: Why It Matters?



Army communication network (imaginary)



Automatically infer

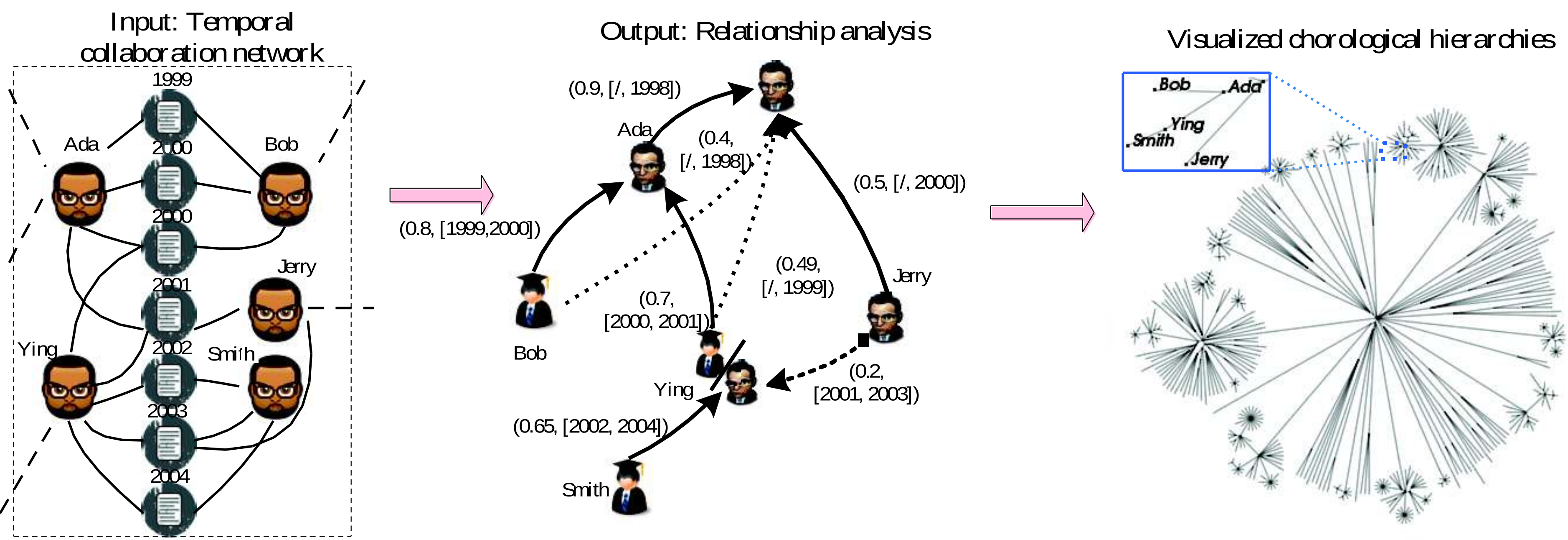
Commander

Captain

Solider

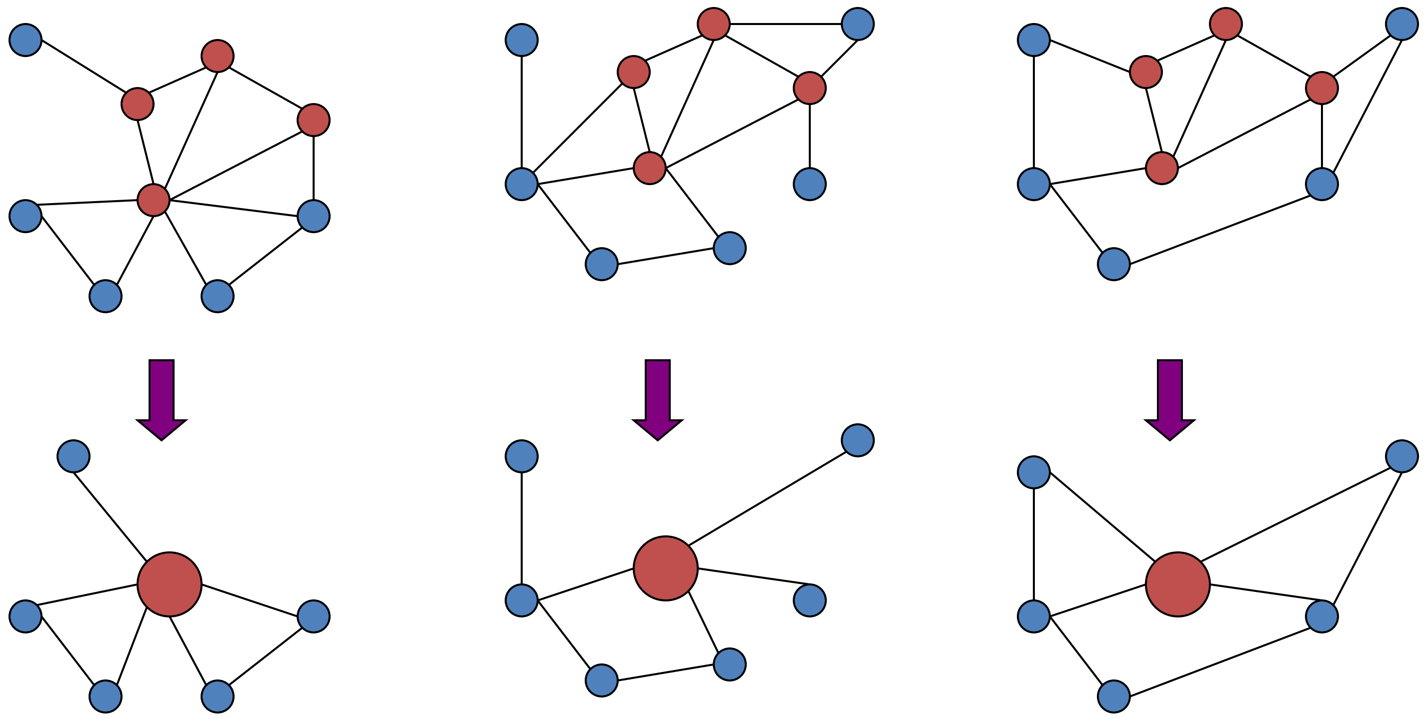
Discovery of Advisor-Advisee Relationships in DBLP Network [Wang, KDD'10]

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



2. Graph/Network Summarization: Graph Compression

- Extract common subgraphs and simplify graphs by condensing these subgraphs into nodes

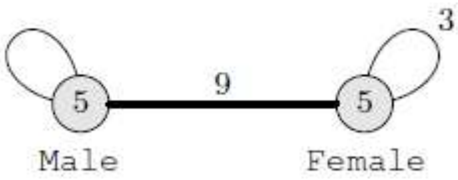


OLAP on Information Networks [Chen, ICDM’08]

- Why OLAP information networks?
- Advantages of OLAP: Interactive exploration of multi-dimensional and multi-level space in a data cube Infonet
  - Multi-dimensional: Different perspectives
  - Multi-level: Different granularities
- InfoNet OLAP: Roll-up/drill-down and slice/dice on information network data
  - Traditional OLAP cannot handle this, because they ignore links among data objects
- Handling two kinds of InfoNet OLAP
  - Informational OLAP
  - Topological OLAP

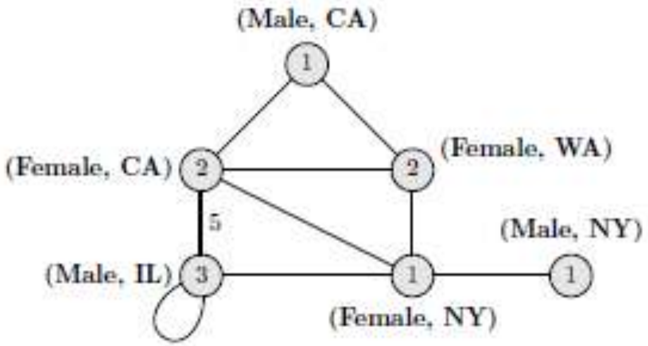
Conventional Group-by v.s. Network Summarization

Gender	COUNT(*)
Male	5
Female	5



Group by "Gender"

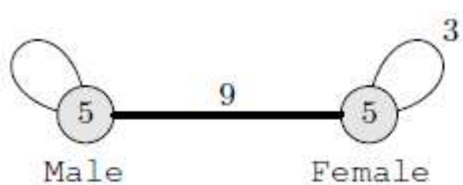
Gender	Location	COUNT(*)
Male	CA	1
Female	CA	2
Female	WA	2
Male	IL	3
Male	NY	1
Female	NY	1

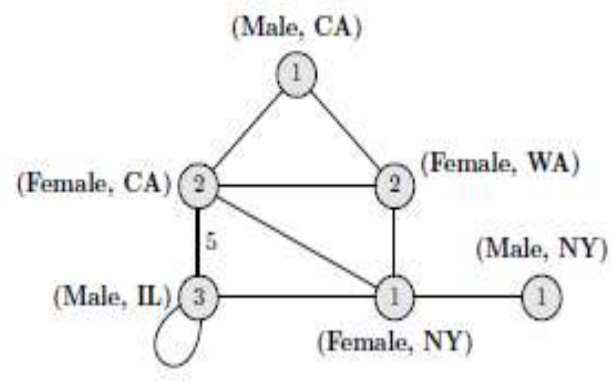


Group by "Gender" and "Location"

OLAP on Graph Cube [Zhao et al., SIGMOD' 11]

- Cuboid query
  - Return as output the aggregate network corresponding to a specific multidimensional space (**cuboid**)
    - *What is the aggregate network between various genders?*
    - *What is the aggregate network between various gender and location combinations?*



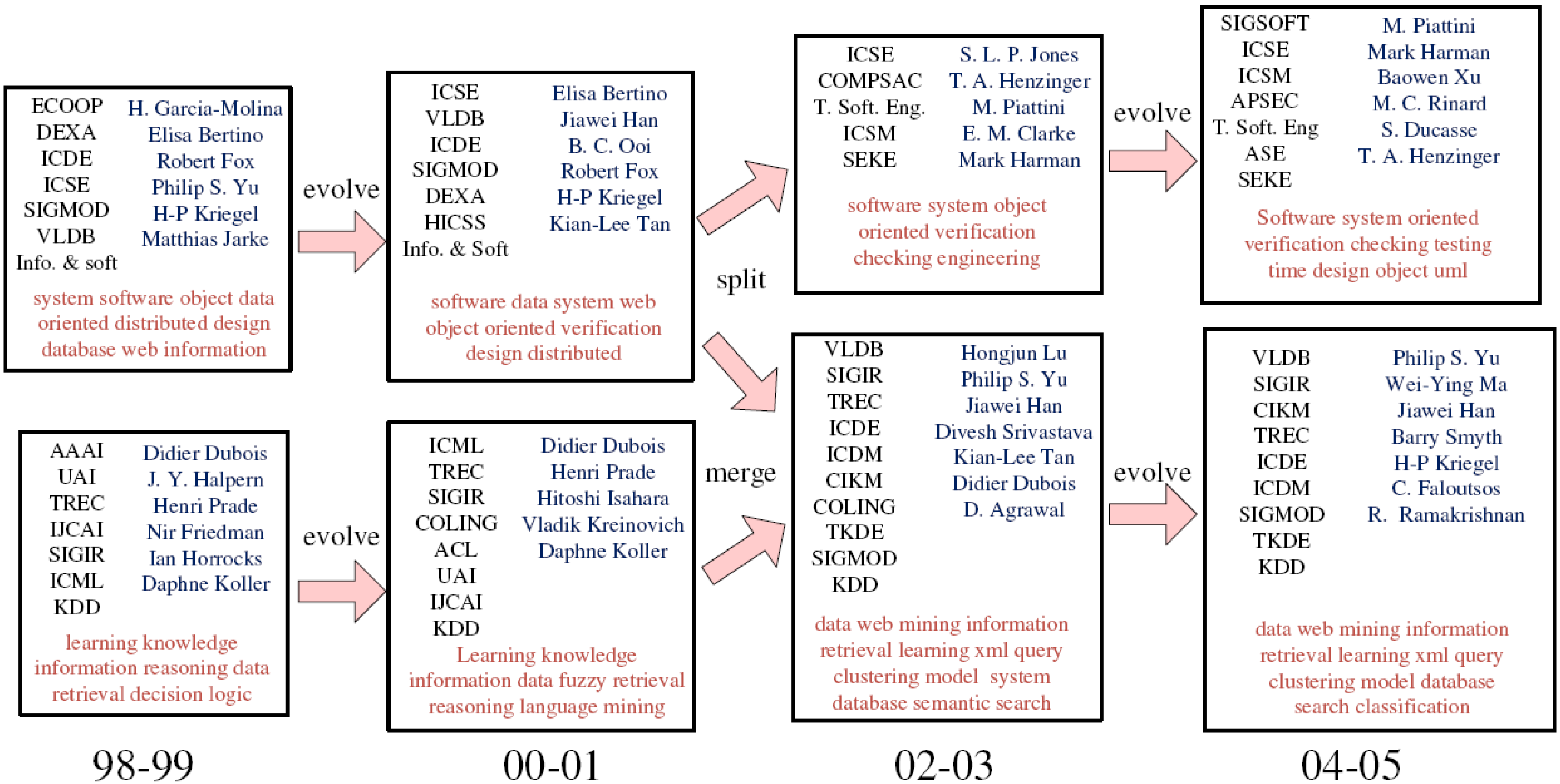


3. Mining Evolution and Dynamics of InfoNet [Sun et al., MLG'10]

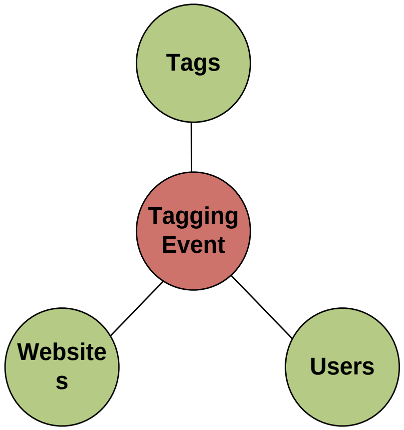
- Many networks are with time information
  - E.g., according to paper publication year, DBLP networks can form network sequences
- Motivation: Model evolution of communities in heterogeneous network
  - Automatically detect the best number of communities in each timestamp
  - Model the smoothness between communities of adjacent timestamps
  - Model the evolution structure explicitly
    - Birth, death, split

Case Study on DBLP

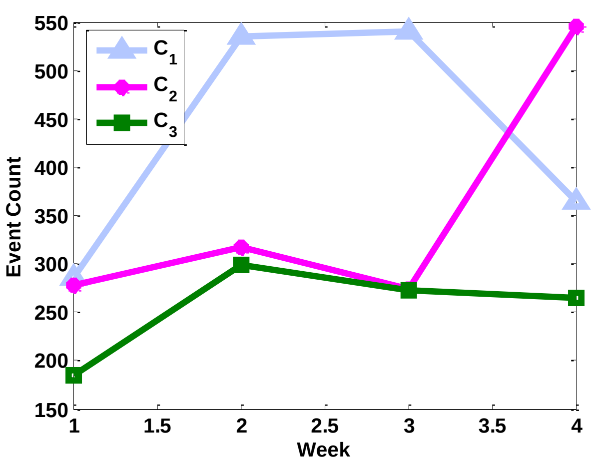
- Tracking database and information system community evolution



Case Study on Delicious.com



Delicious Schema



	Jan. 1 - Jan. 7	Jan. 8 – Jan. 14	Jan. 15 – Jan. 21	Jan. 22 – Jan. 28
C_1:	Security	Google	Security	Google
	Terrorism	China	Google	Security
	Politics	Security	China	China
	Travel	Internet	Internet	Internet
	Usa	Privacy	Microsoft	Privacy
	Airport	Politics	Privacy	Digg
	Israel	Censorship	Censorship	Politics
	Obama	Facebook	Politics	Datenschutz
	CIA	Business	Browser	Facebook
	Afghanistan	Terrorism	USA	USA
C_2:	Mac	Iphone	Iphone	Ipad
	Apple	Apple	Apple	Apple
	Iphone	Twitter	Mac	Iphone
	Windows	Mac	Mobile	Technology
	Tablet	Mobile	Twitter	Tablet
	Ipod	Apps	Software	Mac
	Tips	Ratio	Apps	Mobile
	Macbook	Blog	Business	Newspapers
	Tutorial	Newspapers	Osx	Kindle
	Drm	Technology	Radio	Media
C_3:	Health	Weather	Haiti	Haiti
	Depression	UK	Photography	BBC
	Sleep	Photography	BBC	Photography
	Teenagers	Photo	Earthquake	Animals
	Dubai	Haiti	Photos	Earthquake
	Tallest	Photos	UK	2010
	BBC	2010	2010	Photos
	Building	BBC	Disaster	Nature
	Architecture	Snow	Travel	Funny
	Mentalhealth	Earthquake	Wildlife	Theonion

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**

Conclusions

- Rich knowledge can be mined from information networks
- What is the magic?
  - ***Heterogeneous*, semi-structured *information networks***!
- Clustering, ranking and classification: Integrated clustering, ranking and classification: RankClus, NetClus, GNetMine, …
- Meta-Path-based similarity search and relationship prediction
- User-guided relation strength-aware mining
- Knowledge is power, but knowledge is hidden in massive links!
- ***Mining heterogeneous information networks***: Much more to be explored!!

Future Research

- Discovering ontology and structure in information networks
- Discovering and mining hidden information networks
- Mining information networks formed by structured data linking with unstructured data (text, multimedia and Web)
- Mining cyber-physical networks (networks formed by dynamic sensors, image/video cameras, with information networks)
- Enhancing the power of knowledge discovery by transforming massive unstructured data: Incremental information extraction, role discovery, … ⇒ multi-dimensional structured info-net
- Mining noisy, uncertain, un-trustable massive datasets by information network analysis approach
- Turning Wikipedia and/or Web into structured or semi-structured databases by heterogeneous information network analysis

References: Books on Network Analysis

- A.-L. Barabasi. Linked: How Everything Is Connected to Everything Else and What It Means. Plume, 2003.
- M. Buchanan. Nexus: Small Worlds and the Groundbreaking Theory of Networks. W. W. Norton & Company, 2003.
- P. J. Carrington, J. Scott, and S. Wasserman. Models and Methods in Social Network Analysis. Cambridge University Press, 2005.
- S. Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003.
- D. J. Cook and L. B. Holder. Mining Graph Data. John Wiley & Sons, 2007.
- J. Davies, D. Fensel, and F. van Harmelen. Towards the Semantic Web: Ontology-Driven Knowledge Management. John Wiley & Sons, 2003.
- A. Degenne and M. Forse. Introducing Social Networks. Sage Publications, 1999.
- M. O. Jackson. Social and Economic Networks. Princeton University Press, 2010.
- D. Easley and J. Kleinberg. Networks, Crowds, and Markets. Cambridge University Press, 2010.
- D. Fensel, W. Wahlster, H. Lieberman, and J. Hendler. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.
- L. Getoor and B. Taskar (eds.). Introduction to statistical learning. In MIT Press, 2007.
- B. Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.
- M. E. J. Newman. Networks: An Introduction. Oxford University Press, 2010
- J. P. Scott. Social Network Analysis: A Handbook. Sage Publications, 2005.
- J. Watts. Six Degrees: The Science of a Connected Age. W. W. Norton & Company, 2003.
- D. J.Watts. Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, 2003.
- S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

References: Some Overview Papers

- T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. Scientific American, May 2001.
- C. Cooper and A Frieze. A general model of web graphs. Algorithms, 22, 2003.
- S. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. ACM Comput. Surv., 38, 2006.
- T. Dietterich, P. Domingos, L. Getoor, S. Muggleton, and P. Tadepalli. Structured machine learning: The next ten years. Machine Learning, 73, 2008
- S. Dumais and H. Chen. Hierarchical classification of web content. SIGIR'00.
- S. Dzeroski. Multirelational data mining: An introduction. ACM SIGKDD Explorations, July 2003.
- L. Getoor. Link mining: a new data mining challenge. SIGKDD Explorations, 5:84{89, 2003.
- L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. ICML'01
- D. Jensen and J. Neville. Data mining in networks. In Papers of the Symp. Dynamic Social Network Modeling and Analysis, National Academy Press, 2002.
- T. Washio and H. Motoda. State of the art of graph-based data mining. SIGKDD Explorations, 5, 2003.

References: Some Influential Papers

- A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. Computer Networks, 33, 2000.
- S. Brin and L. Page. The anatomy of a large-scale hyper-textual web search engine. WWW'98.
- S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. M. Kleinberg. Mining the web's link structure. COMPUTER, 32, 1999.
- M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. ACM SIGCOMM'99
- M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In Proc. Natl. Acad. Sci. USA 99, 2002.
- B. A. Huberman and L. A. Adamic. Growth dynamics of world-wide web. Nature, 399:131, 1999.
- G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. KDD'02
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD'03
- J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. COCOON'99
- J. M. Kleinberg. Small world phenomena and the dynamics of information. NIPS'01
- R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. FOCS'00
- M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45, 2003.

References: Clustering and Ranking (1)

- E. Airoldi, D. Blei, S. Fienberg and E. Xing, “Mixed Membership Stochastic Blockmodels”, JMLR’08
- Liangliang Cao, Andrey Del Pozo, Xin Jin, Jiebo Luo, Jiawei Han, and Thomas S. Huang, “RankCompete: Simultaneous Ranking and Clustering of Web Photos”, WWW’10
- G. Jeh and J. Widom, “SimRank: a measure of structural-context similarity”, KDD'02
- Jing Gao, Feng Liang,Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han, “Community Outliers and their Efficient Detection in Information Networks", KDD'10
- M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks”, Physical Review E, 2004
- M. E. J. Newman and M. Girvan, “Fast algorithm for detecting community structure in networks”, Physical Review E, 2004
- J. Shi and J. Malik, “Normalized cuts and image Segmentation”, *CVPR'97*
- Yizhou Sun, Yintao Yu, and Jiawei Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema", KDD’09
- Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", EDBT’09

References: Clustering and Ranking (2)

- Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu, "iTopicModel: Information Network-Integrated Topic Modeling", ICDM'09
- Yizhou Sun, Charu C. Aggarwal, and Jiawei Han, "*Relation Strength-Aware Clustering of Heterogeneous Information Networks with Incomplete Attributes*", PVLDB 5(5), 2002
- A. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. KDD'04
- Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation", IEEE Trans. Pattern Anal. Mach. Intell., 1993.
- X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: A structural clustering algorithm for networks. KDD'07
- Xiaoxin Yin, Jiawei Han, Philip S. Yu. "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", VLDB'06.
- Yintao Yu, Cindy X. Lin, Yizhou Sun, Chen Chen, Jiawei Han, Binbin Liao, Tianyi Wu, ChengXiang Zhai, Duo Zhang, and Bo Zhao, "iNextCube: Information Network-Enhanced Text Cube", VLDB'09 (demo)
- X. Yin, J. Han, and P. S. Yu. Cross-relational clustering with user's guidance. KDD'05

References: Network Classification (1)

- A. Appice, M. Ceci, and D. Malerba. Mining model trees: A multi-relational approach. ILP'03
- Jing Gao, Feng Liang, Wei Fan, Yizhou Sun, and Jiawei Han, "Bipartite Graph-based Consensus Maximization among Supervised and Unsupervised Models ", NIPS'09
- L. Getoor, N. Friedman, D. Koller and B. Taskar, "Learning Probabilistic Models of Link Structure", JMLR'02.
- L. Getoor, E. Segal, B. Taskar and D. Koller, "Probabilistic Models of Text and Link Structure for Hypertext Classification", IJCAI WS 'Text Learning: Beyond Classification', 2001.
- L. Getoor, N. Friedman, D. Koller, and A. Pfeffer, "Learning Probabilistic Relational Models", chapter in Relation Data Mining, eds. S. Dzeroski and N. Lavrac, 2001.
- M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, "Graph-based classification on heterogeneous information networks", ECMLPKDD'10.
- M. Ji, J. Jan, and M. Danilevsky, "Ranking-based Classification of Heterogeneous Information Networks", KDD'11.
- Q. Lu and L. Getoor, "Link-based classification", ICML'03
- D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks", CIKM'03

References: Network Classification (2)

- J. Neville, B. Gallaher, and T. Eliassi-Rad. Evaluating statistical tests for within-network classifiers of relational data. ICDM'09.
- J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. KDD'03
- Jennifer Neville, David Jensen, "Relational Dependency Networks", JMLR'07
- M. Szummer and T. Jaakkola, "Partially labeled classication with markov random walks", In NIPS, volume 14, 2001.
- M. J. Rattigan, M. Maier, and D. Jensen. Graph clustering with network structure indices. ICML'07
- P. Sen, G. M. Namata, M. Galileo, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. AI Magazine, 29, 2008.
- B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. IJCAI'01
- B. Taskar, P. Abbeel, M.F. Wong, and D. Koller, "Relational Markov Networks", chapter in L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning, 2007
- X. Yin, J. Han, J. Yang, and P. S. Yu, "CrossMine: Efficient Classification across Multiple Database Relations", ICDE'04.
- D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency", In *NIPS 16*, Vancouver, Canada, 2004.
- X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation", Technical Report, 2002.

References: Social Network Analysis

- B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. P. Sheth, I. B. Arpinar, A. Joshi, and T. Finin. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. WWW'06
- R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. WWW'03
- P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. WWW'04
- D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. PKDD'05
- P. Domingos. Mining social networks for viral marketing. IEEE Intelligent Systems, 20, 2005.
- P. Domingos and M. Richardson. Mining the network value of customers. KDD'01
- P. DeRose, W. Shen, F. Chen, A. Doan, and R. Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. VLDB'07
- G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. IEEE Computer, 35, 2002.
- J. Kubica, A. Moore, and J. Schneider. Tractable group detection on large link data sets. ICDM'03

References: Data Quality & Search in Networks

- I. Bhattacharya and L. Getoor, “Iterative record linkage for cleaning and integration”, Proc. SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'04)
- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, “Integrating conflicting data: The role of source dependence”, PVLDB, 2(1):550–561, 2009.
- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava, “Truth discovery and copying detection in a dynamic world”, PVLDB, 2(1):562–573, 2009.
- H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, “Two supervised learning approaches for name disambiguation in author citations”, ICDL'04.
- Y. Sun, J. Han, T. Wu, X. Yan, and Philip S. Yu, “PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks”, VLDB’11.
- 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):796-808, 2008
- P. Zhao and J. Han, “On Graph Query Optimization in Large Networks”, VLDB’10.

References: Link and Relationship Prediction

- V. Leroy, B. B. Cambazoglu, and F. Bonchi, “Cold start link prediction”, *KDD ’10*.
- D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks”, *CIKM ’03*,
- R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, “New perspectives and methods in link prediction”, *KDD’10*.
- Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal and Jiawei Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks“, ASONAM’11.
- Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla, "When Will It Happen? --- Relationship Prediction in Heterogeneous Information Networks", WSDM’12.
- B. Taskar, M. fai Wong, P. Abbeel, and D. Koller, “Link prediction in relational data”, NIPS ’03.
- Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han, "Citation Prediction in Heterogeneous Bibliographic Networks”, SDM’12.

References: Role Discovery, Summarization and OLAP

- D. Archambault, T. Munzner, and D. Auber. Topolayout: Multilevel graph layout by topological features. IEEE Trans. Vis. Comput. Graph, 2007.
- Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu, "Graph OLAP: Towards Online Analytical Processing on Graphs", ICDM 2008
- Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu, "Graph OLAP: A Multi-Dimensional Framework for Graph Data Analysis", KAIS 2009.
- Xin Jin, Jiebo Luo, Jie Yu, Gang Wang, Dhiraj Joshi, and Jiawei Han, "*iRIN: Image Retrieval in Image Rich Information Networks*", WWW'10 (demo paper)
- Lu Liu, Feida Zhu, Chen Chen, Xifeng Yan, Jiawei Han, Philip Yu, and Shiqiang Yang, "Mining Diversity on Networks", DASFAA'10
- Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. SIGMOD'08
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo, "Mining Advisor-Advisee Relationships from Research Publication Networks ", KDD'10
- Zhijun Yin, Manish Gupta, Tim Weninger and Jiawei Han, "LINKREC: A Unified Framework for Link Recommendation with User Attributes and Graph Structure ", WWW'10
- Peixiang Zhao, Xiaolei Li, Dong Xin, Jiawei Han. Graph Cube: On Warehousing and OLAP Multidimensional Networks, SIGMOD'11

References: Network Evolution

- L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. KDD'06
- M.-S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. VLDB'09
- J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. KDD'05
- Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta, Bo Zhao, “Community Evolution Detection in Dynamic Heterogeneous Information Networks”, KDD-MLG’10