Mining Heterogeneous Information Networks

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

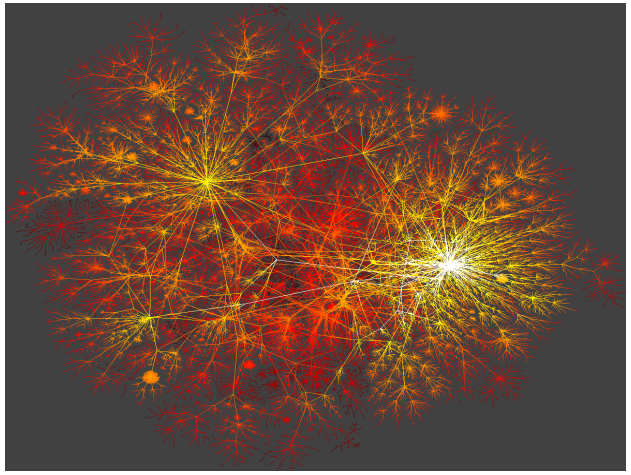
FEBRUARY 4, 2017

Outline

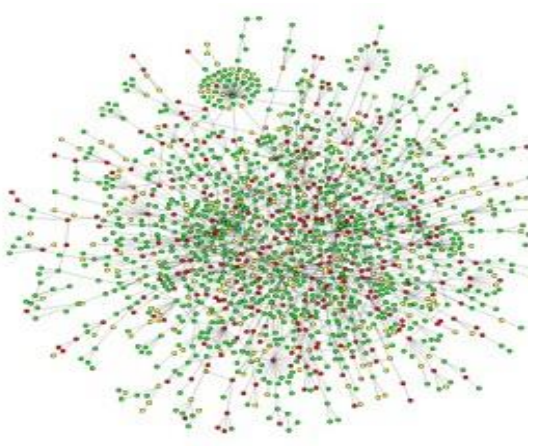
- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Information Networks
  - Relationship Prediction in Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation recommendation in heterogeneous networks
- Summary

Ubiquitous Information Networks

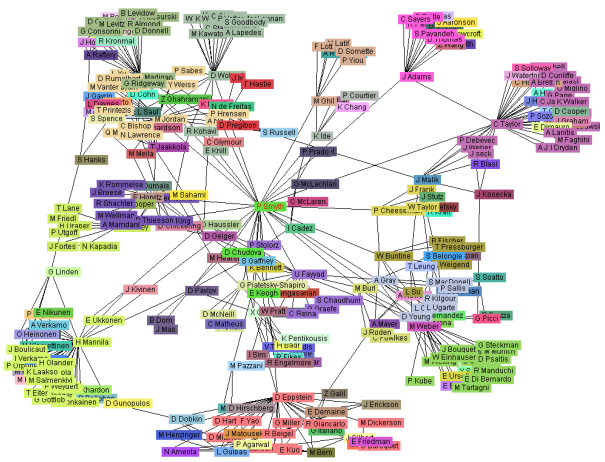
- Graphs and substructures: Chemical compounds, visual objects, circuits, XML
- Biological networks
- Bibliographic networks: DBLP, ArXiv, PubMed, …
- Social networks: Facebook >100 million active users
- World Wide Web (WWW): > 3 billion nodes, > 50 billion arcs
- Cyber-physical networks



World-Wide Web



Yeast protein interaction network

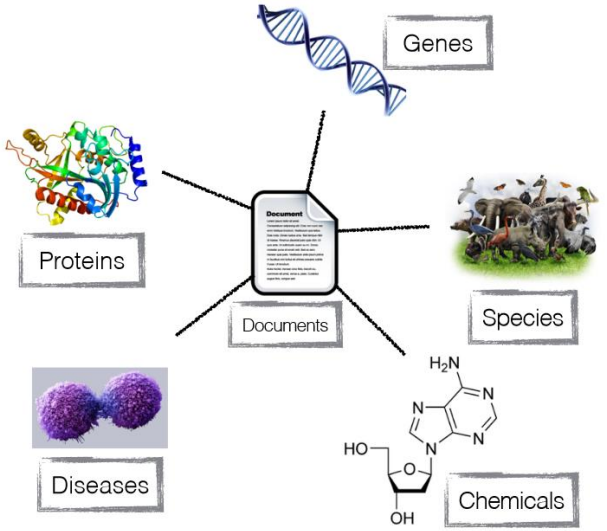


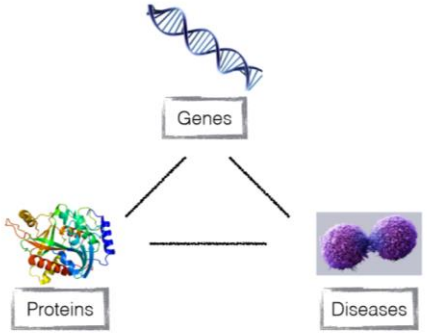
Co-author network

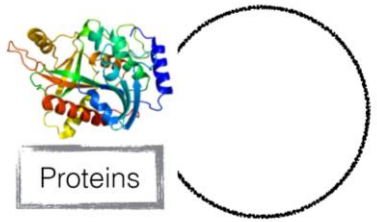
Social network sites

Heterogeneous Information Networks

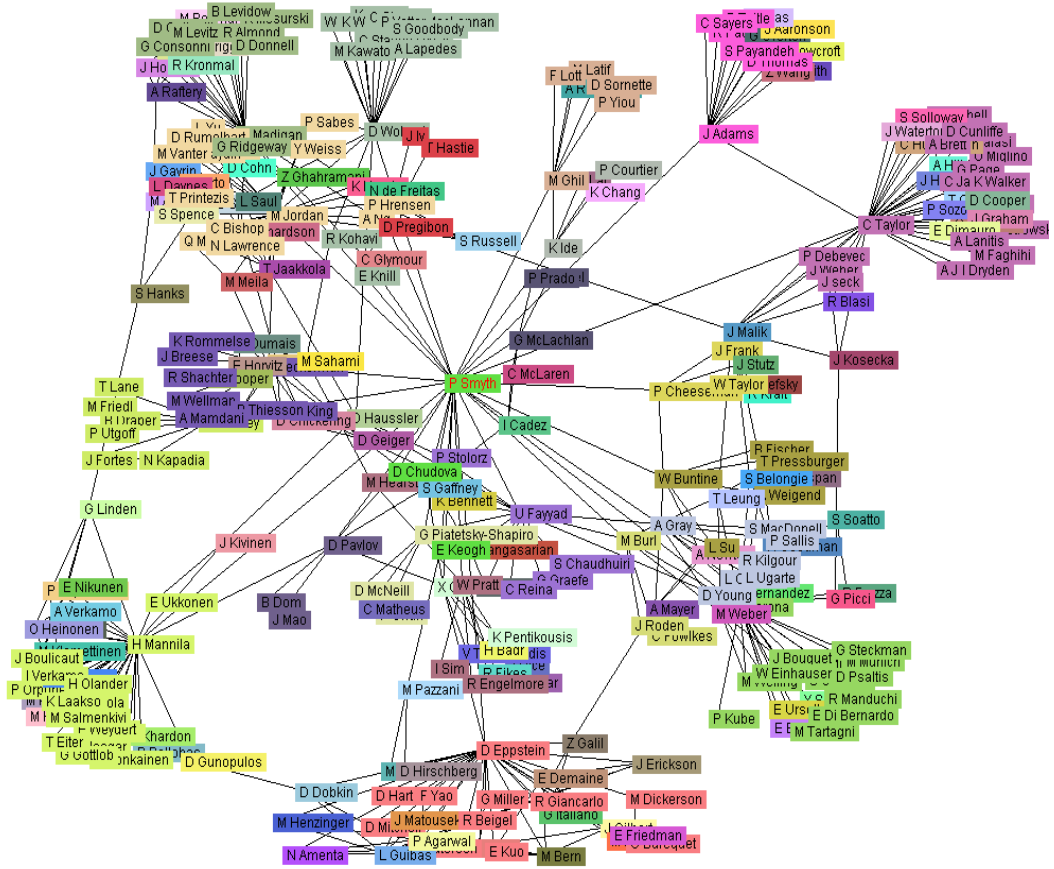
- Homogeneous vs. heterogeneous networks
  - Homogeneous networks: Single object type and single link type
    - Single model social networks (e.g., friends)
    - WWW: A collection of linked Web pages
  - Heterogeneous networks: Multiple object and link types
    - Medical network: Patients, doctors, diseases, contacts, treatments
    - Bibliographic network: Publications, authors, venues (e.g., DBLP > 2 million papers)
- Homogeneous networks are often resulted from projection of heterogeneous networks
  - E.g., coauthor network from its original DBLP network

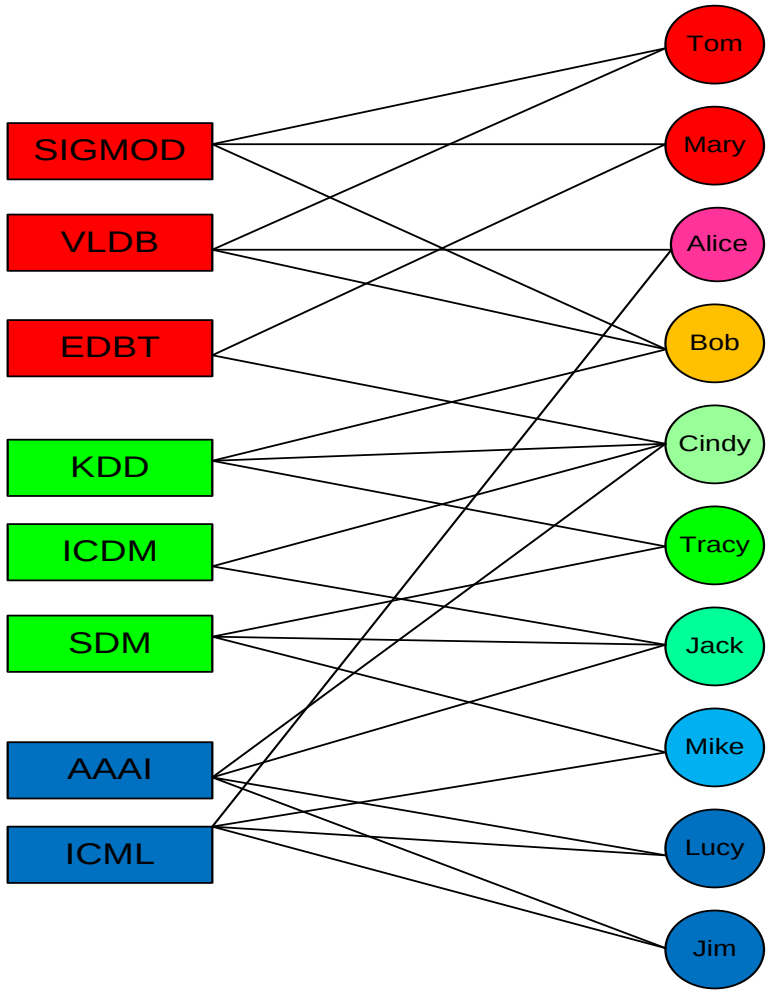






Homogeneous vs. Heterogeneous Networks



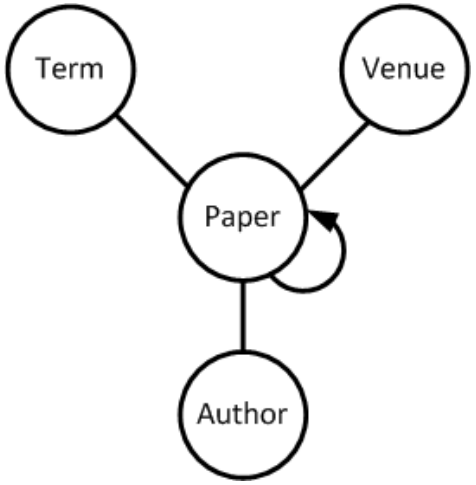


Co-author Network

Conference-Author Network

Mining Heterogeneous Information Networks

- Homogeneous networks can often be derived from their original heterogeneous networks
  - Ex. Coauthor networks can be derived from author-paper-conference networks by projection on authors
  - Paper citation networks can be derived from a complete bibliographic network with papers and citations projected
- Heterogeneous networks carry richer information than their corresponding projected homogeneous networks
- Typed heterogeneous network vs. non-typed heterogeneous network (i.e., not distinguishing different types of nodes)
  - Typed nodes and links imply more structures, leading to richer discovery
- Mining *semi-structured* heterogeneous information networks
  - Clustering, ranking, classification, prediction, similarity search, etc.

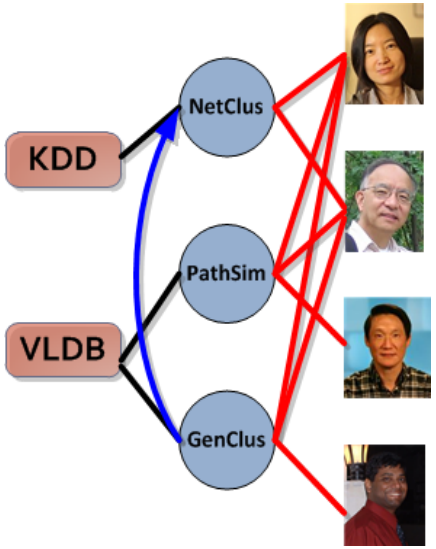


Examples of Heterogeneous Information Networks

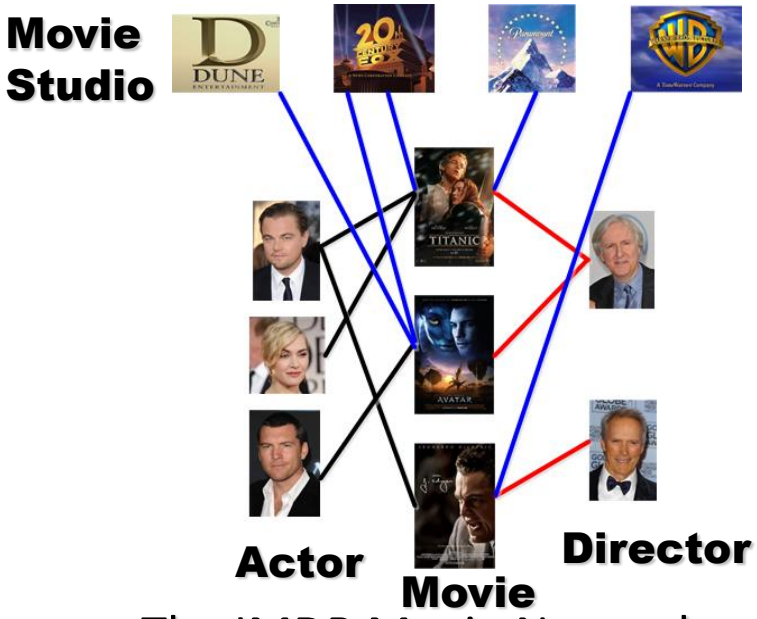
- **Bibliographic information networks:** DBLP, ArXive, PubMed
  - Entity types: *paper (P), venue* (V), *author (A), and term (T)*
  - *Relation type: authors write papers, venues publish papers, papers contain terms*
- **Twitter information network, and other social media network**
  - Objects types: *user, tweet, hashtag, and term*
  - *Relation/*link types: users *follow users, users post tweets, tweets reply tweets, tweets use terms, tweets contain hashtags*
- **Flickr information network**
  - Object types: *image, user, tag, group, and comment*
  - Relation types: users *upload* images, image *contain* tags, images *belong to* groups, users *post* comments, and comments *comment* on images
- **Healthcare information network**
  - Object types: *doctor, patient, disease, treatment,* and *device*
  - *Relation types: treatments used-for diseases,* patients *have* diseases, *patients visit doctors*

Structures Facilitate Mining Heterogeneous Networks

- Network construction: generates structured networks from unstructured text data
  - Each node: an entity; each link: a relationship between entities
  - Each node/link may have attributes, labels, and weights
  - Heterogeneous, multi-typed networks: e.g., Medical network: Patients, doctors, diseases, contacts, treatments



DBLP Bibliographic Network



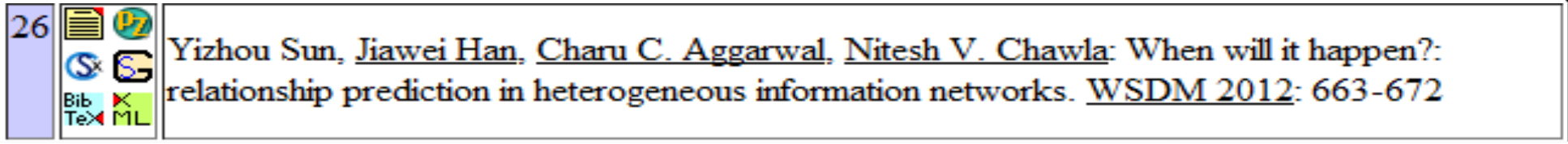
The IMDB Movie Network

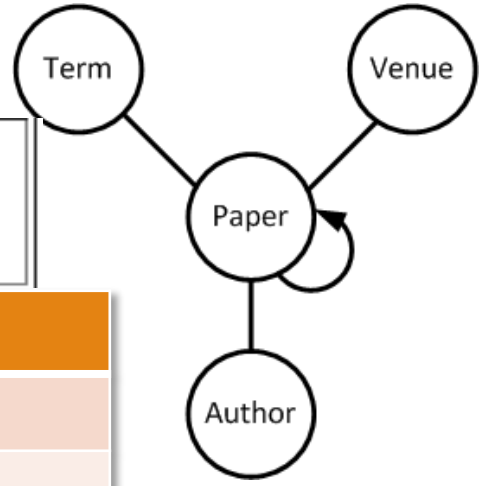

The Facebook Network

It works well for ego-networks!

What Can be Mined from Heterogeneous Networks?

- A homogeneous network can be derived from its "parent" heterogeneous network
  - Ex. Coauthor networks from the original author-paper-conference networks
- Heterogeneous networks carry richer info. than the projected homogeneous networks
- Typed nodes & links imply more structures, leading to richer discovery
- Ex.: DBLP: A Computer Science bibliographic database (network)





Knowledge hidden in DBLP Network	Mining Functions
Who are the **leading** researchers on Web search?	Ranking
Who are the **peer** researchers of Jure Leskovec?	Similarity Search
Whom **will** Christos Faloutsos **collaborate with**?	Relationship Prediction
Which **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

Principles of Mining Heterogeneous Information Net

- **Information propagation across heterogeneous nodes & links**
  - *How to compute* ranking scores, similarity scores, and clusters, and how to make good use of class labels, across heterogeneous nodes and links
  - *Objects in the networks are interdependent and knowledge can only be mined using the holistic information in a network*
- **Search and mining by exploring network meta structures**
  - Heter. info networks: semi-structured and typed
  - Network schema: a meta structure, guidance of search and mining
  - Meta-path based similarity search and mining
- **User-guided exploration of information networks**
  - Automatically select the right relation (or meta-path) combinations with appropriate weights for a particular search or mining task
  - User-guided or feedback-based network exploration is a strategy

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Information Networks
  - Relationship Prediction in Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation recommendation in heterogeneous networks
- Summary

Ranking-Based Clustering in Heterogeneous Networks

- Clustering and ranking: Two critical functions in data mining
  - Clustering without ranking? Think about no PageRank dark time before Google
  - Ranking will make more sense within a particular cluster
    - Einstein in physics vs. Turing in computer science
- Why not integrate ranking with clustering & classification?
  - High-ranked objects should be more important in a cluster than low-ranked ones
  - Why treat every object the same weight in the same cluster?
    - But how to get their weight?
- Integrate ranking with clustering/classification in the same process
  - Ranking, as the feature, is conditional (i.e., relative) to a specific cluster
  - Ranking and clustering may mutually enhance each other
  - Ranking-based clustering: RankClus [EDBT'09], NetClus [KDD'09]

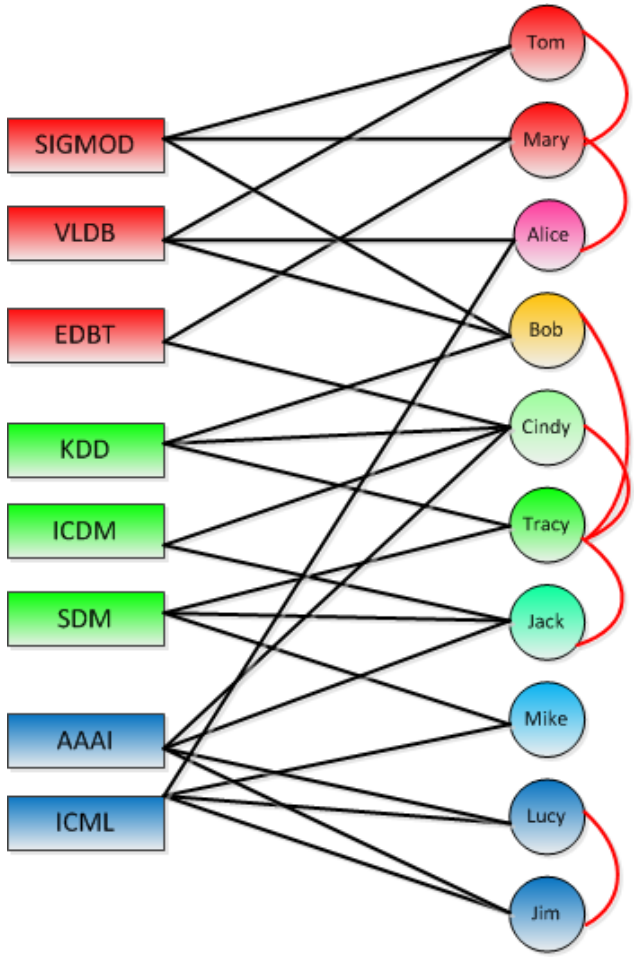
A Bi-Typed Network Model and Simple Ranking

- A bi-typed network model
  - Let X represents type *venue*
    - Y: Type *author*

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

- The DBLP network can be represented as matrix W
- Our task: Rank-based clustering of heterogeneous network W
- Simple Ranking
  - Proportional to # of publications of an author and a venue
  - Considers only **immediate neighborhood** in the network

$$\begin{cases} \vec{r}_X(x) = \dfrac{\sum_{j=1}^{n} W_{XY}(x, j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i, j)} \\ \vec{r}_Y(y) = \dfrac{\sum_{i=1}^{n} W_{XY}(i, y)}{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{XY}(i, j)} \end{cases}$$



A bi-typed network

But what about an author publishing many papers only in very weak venues?

The RankClus Methodology

- Ranking as the feature of the cluster
  - Ranking is conditional on a specific cluster
    - E.g., VLDB's rank in Theory vs. its rank in the DB area
  - The distributions of ranking scores over objects are different in each cluster
- Clustering and ranking are mutually enhanced
  - Better clustering: Rank distributions for clusters are more distinguishing from each other
  - Better ranking: Better metric for objects is learned from the ranking
- Not every object should be treated equally in clustering!
- Y. Sun, et al., "*RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis*", EDBT'09

Authority Ranking

- Methodology: **Propagate** the ranking scores in the network over different types
- Rule 1: Highly ranked authors publish *many* papers in highly ranked venues

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

- Rule 2: Highly ranked venues 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

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

- Other ranking functions are quite possible (e.g., using domain knowledge)
  - Ex. Journals may weight more than conferences in science

Alternative Ranking Functions

- A ranking function is not only related to the link property, but also depends on domain knowledge
  - Ex: Journals may weight more than conferences in science
- Ranking functions can be defined on multi-typed networks
  - Ex: PopRank takes into account the impact both from the same type of objects and from the other types of objects, with different impact factors for different types
- Use expert knowledge, for example,
  - TrustRank semi-automatically separates reputable, good objects from spam ones
  - Personalized PageRank uses expert ranking as query and generates rank distributions w.r.t. such knowledge
- A research problem that needs systematic study

The EM (Expectation Maximization) Algorithm

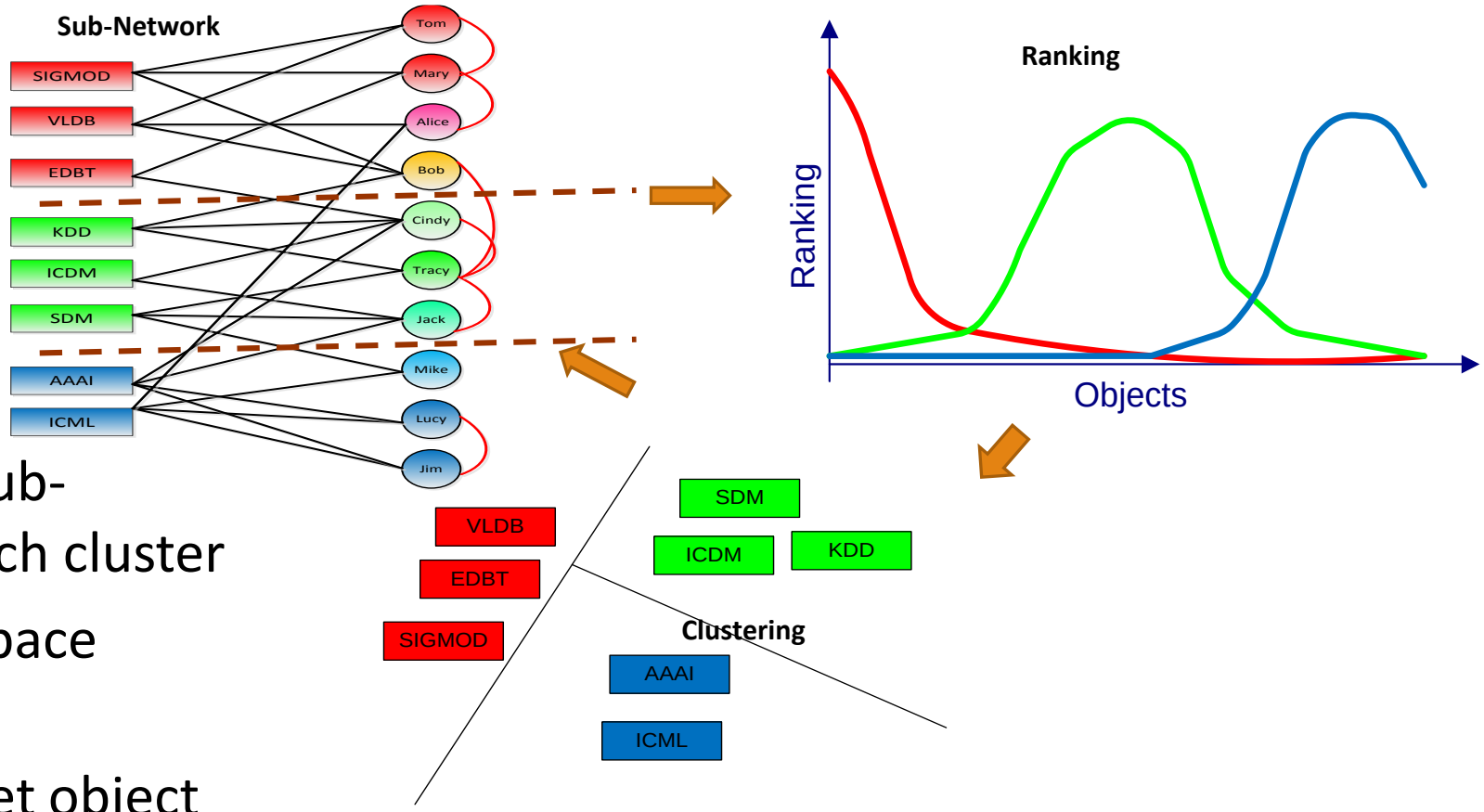
- The k-means algorithm has two steps at each iteration (in the E-M framework):
  - **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is *expected to belong to the closest cluster*
  - **Maximization Step** (M-step): Given the cluster assignment, for each cluster, the algorithm *adjusts the center* so that *the sum of distance* from the objects assigned to this cluster and the new center is minimized
- **The (EM) algorithm:** A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.
  - **E-step** assigns objects to clusters according to the current probabilistic clustering or parameters of probabilistic clusters
  - **M-step** finds the new clustering or parameters that minimize the sum of squared error (SSE) or maximize the expected likelihood

From Conditional Rank Distribution to E-M Framework

- Given a bi-typed bibliographic network, how can we use the conditional rank scores to further improve the clustering results?
- Conditional rank distribution as cluster feature
  - For each cluster C_k, *the conditional rank scores,* $r_X|C_K$ *and* $r_Y|C_K$, *can be viewed* as conditional rank distributions of *X and Y, which are the features for cluster* C_k
- Cluster membership as object feature
  - From $p(k|o_i) \propto p(o_i|k)p(k)$, *the higher its conditional* rank in a cluster ($p(o_i|k)$), *the higher possibility an object will belong to that cluster* ($p(k|o_i)$)
  - Highly ranked attribute object has more impact on determining the cluster membership of a target object
- Parameter estimation using the Expectation-Maximization algorithm
  - E-step: Calculate the distribution $p(z = k|y_j, x_i, \Theta)$ *based on* the current value of Θ
  - M-Step: Update Θ according to the current *distribution*

RankClus: Integrating Clustering with Ranking

- **An EM styled Algorithm**
  - **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** change < threshold

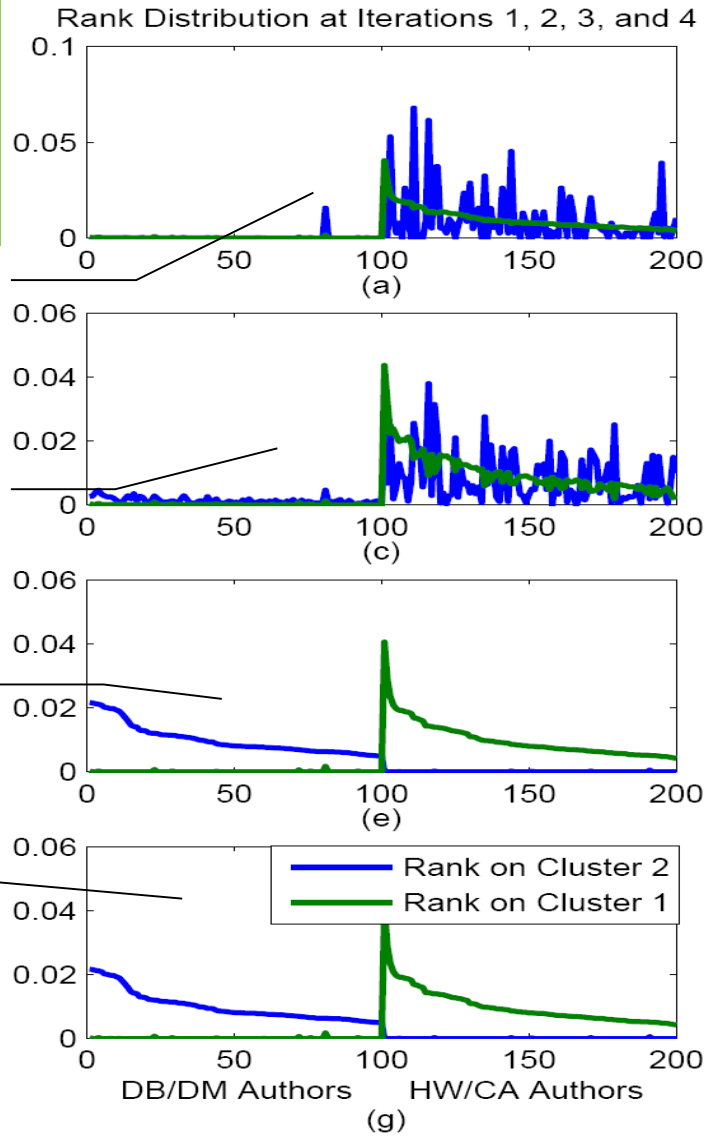


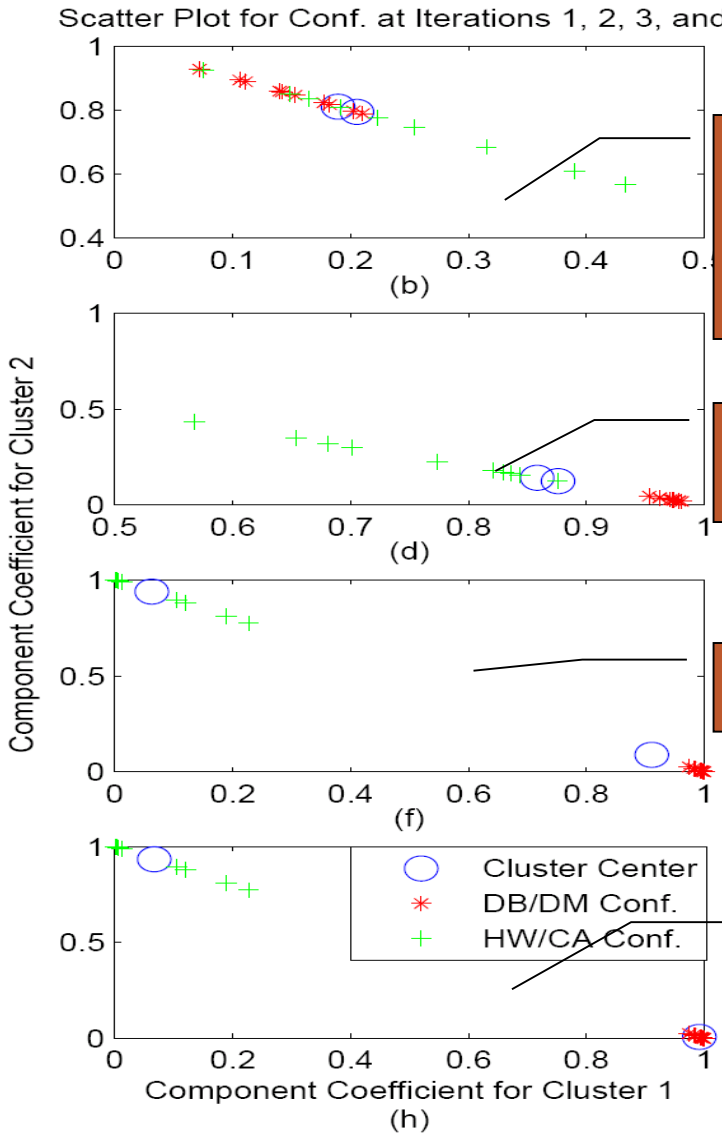
An E-M framework for iterative enhancement

RankClus [EDBT'09]: Ranking and clustering mutually enhancing each other in an E-M framework

Step-by-Step Running of RankClus

Clustering and ranking of two fields: DB/DM vs. HW/CA (hardware/ computer architecture)





Initially, ranking distributions are mixed together

Improved a little

Improved significantly

Stable

Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

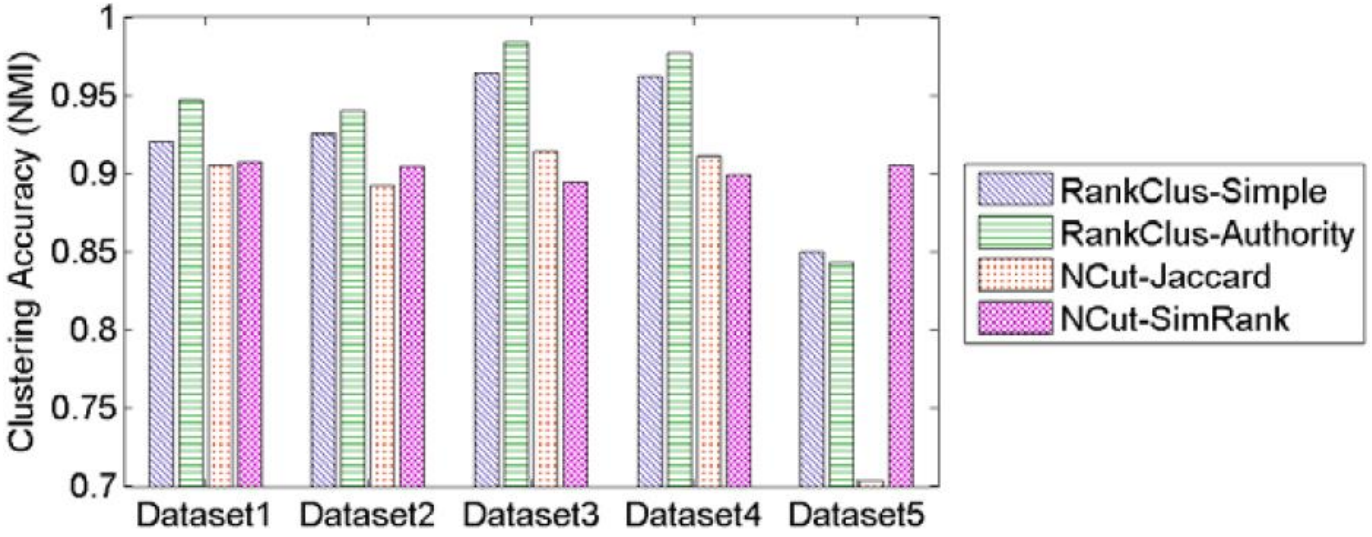
Stable

Clustering Performance (NMI)

- Compare the clustering accuracy: For *N objects, K clusters, and two clustering results, let n(i, j), be # objects with cluster label i in the 1st clustering* result (say generated by the new alg.) and that w. cluster label *j in the 2nd clustering result (say the* ground truth)
- Normalized Mutual Info. (NMI):

$$\frac{\sum_{i=1}^{K}\sum_{j=1}^{K} p(i,j)\log\left(\frac{p(i,j)}{p_1(j)p_2(i)}\right)}{\sqrt{\sum_{j=1}^{K} p_1(j)\log p_1(j)\sum_{i=1}^{K} p_2(i)\log p_2(i)}}$$

  - joint distribution: $p(i, j) = n(i,j)/N$
  - *row distr.* $p_1(j) = \sum_{i=1}^{K} p(i, j)$ *column distr.* $p_2(i) = \sum_{j=1}^{K} p(i, j)$



D1: med. separated & med. density
D2: med. separated & low density
D3: med. Separated & high density
D4: highly separated & med. density
D5: poorly separated & med. density

RankClus: Clustering & Ranking CS Venues in DBLP

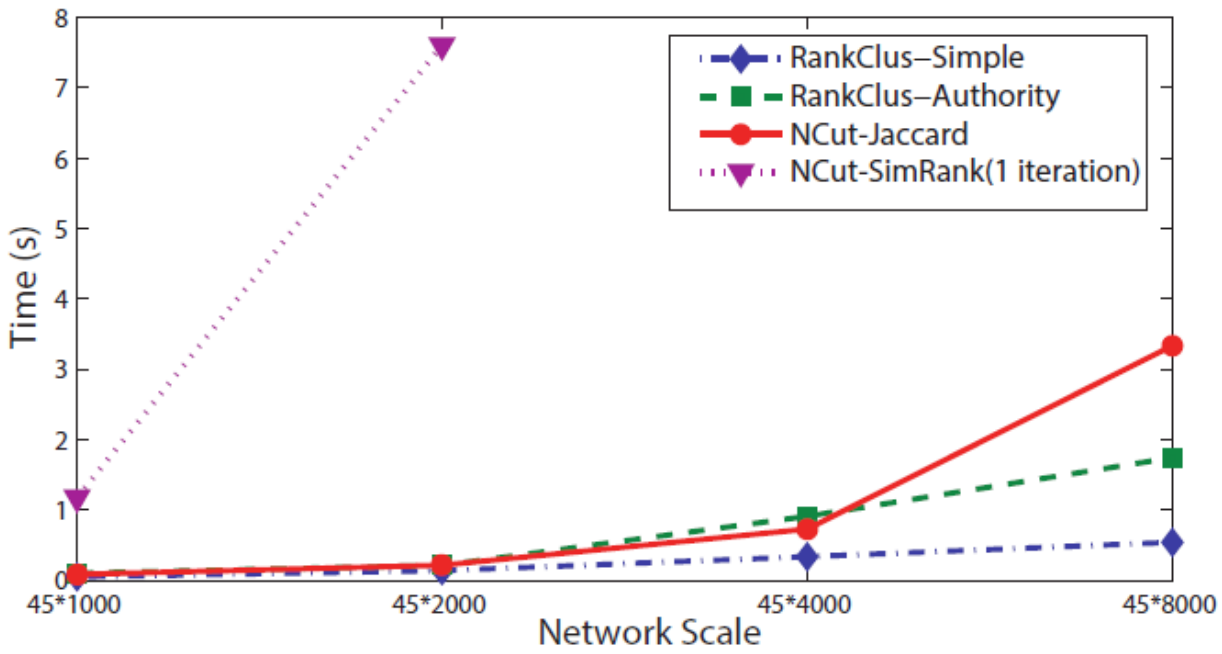
Table 2.4: Top-10 venues in 5 clusters generated by RankClus in DBLP

Rank	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 (when k = 15)

Time Complexity: Linear to # of Links

- At each iteration, |E|: edges in network, m: # of target objects, K: # of clusters
  - Ranking for sparse network
    - ~O(|E|)
  - Mixture model estimation
    - ~O(K|E|+mK)
  - Cluster adjustment
    - ~O(mK^2)
- In all, linear to |E|
  - ~O(K|E|)



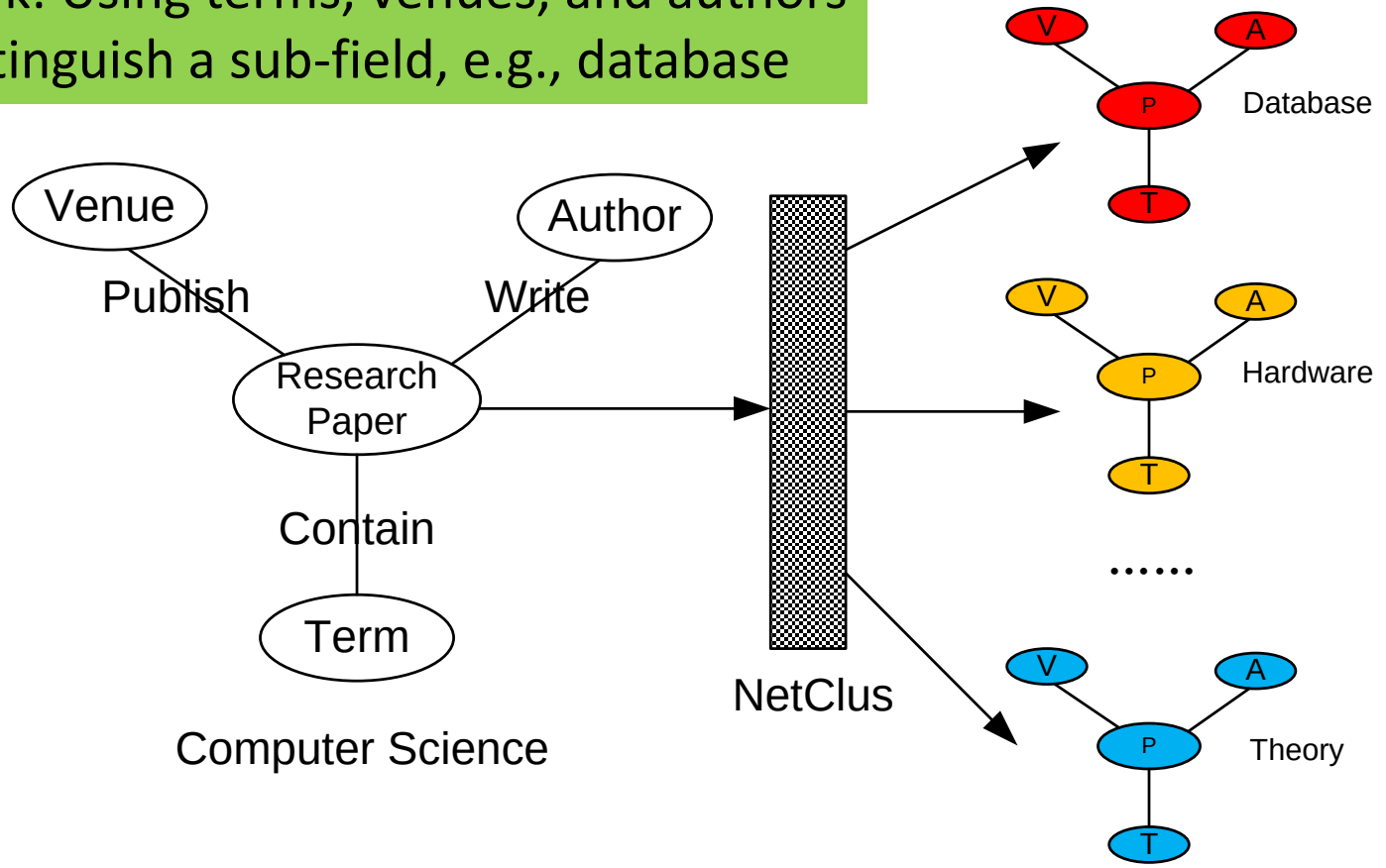
Comparison of algorithms on execution time

- Note: SimRank will be at least quadratic at each iteration since it evaluates distance between every pair in the network

NetClus: Ranking-Based Clustering with Star Network Schema

- Beyond bi-typed network: Capture more semantics with multiple types
- Split a network into multi-subnetworks, each a (multi-typed) net-cluster [KDD'09]

DBLP network: Using terms, venues, and authors to jointly distinguish a sub-field, e.g., database



The NetClus Algorithm

- Generate initial partitions for target objects and induce initial net-clusters from the original network
- Repeat // An E-M Framework
  - Build ranking-based probabilistic generative model for each net-cluster
  - Calculate the posterior probabilities for each target object
  - Adjust their cluster assignment according to the new measure defined by the posterior probabilities to each cluster
- Until the clusters do not change significantly
- Calculate the posterior probabilities for each attribute object in each net-cluster

NetClus on the DBLP Network

- NetClus initialization: G = (V, E, W), weight w_{xixj} linking x_i and x_j
  - V = A U C U D U T, where D (paper), A (author), C (conf.), T(term)

$$w_{x_i x_j} = \begin{cases} 1, \text{ if } x_i(x_j) \in A \cup C \text{ and } x_j(x_i) \in D, \\ \quad \text{and } x_i \text{ has link to } x_j \\ c, \text{ if } x_i(x_j) \in T \text{ and } x_j(x_i) \in D \text{ and } x_i(x_j) \\ \quad \text{appears c times in } x_j(x_i), \\ 0, \text{ otherwise.} \end{cases}$$

- Simple ranking:

$$p(x|T_x, G) = \frac{\sum_{y \in N_G(x)} W_{xy}}{\sum_{x' \in T_x} \sum_{y \in N_G(x')} W_{x'y}}$$

- Authority ranking for type Y based on type X, through the center type Z:

$$P(Y|T_Y, G) = W_{YZ} W_{ZX} P(X|T_X, G)$$

- For DBLP:

$$P(C|T_C, G) = W_{CD} D_{DA}^{-1} W_{DA} P(A|T_A G)$$
$$P(A|T_A, G) = W_{AD} D_{DC}^{-1} W_{DC} P(C|T_C, G)$$

Multi-Typed Networks Lead to Better Results

- The network cluster for database area: Conferences, Authors, and Terms
  - NetClus leads to better clustering and ranking than RankClus

Conference	Rank Score
SIGMOD	0.315
VLDB	0.306
ICDE	0.194
PODS	0.109
EDBT	0.046
CIKM	0.019
...	...

Author	Rank Score
Michael Stonebraker	0.0063
Surajit Chaudhuri	0.0057
C. Mohan	0.0053
Michael J. Carey	0.0052
David J. DeWitt	0.0051
H. V. Jagadish	0.0043
...	...

Term	Rank Score
database	0.0529
system	0.0322
query	0.0313
data	0.0251
object	0.0138
management	0.0113
...	...

  - NetClus vs. RankClus: **16%** higher accuracy on conference clustering

NetClus: Distinguishing Conferences

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

NetClus: Experiment on DBLP: Database System Cluster

Term	Venue	Author
database 0.0995511	VLDB 0.318495	Surajit Chaudhuri 0.00678065
databases 0.0708818	SIGMOD Conf. 0.313903	Michael Stonebraker 0.00616469
system 0.0678563	ICDE 0.188746	Michael J. Carey 0.00545769
data 0.0214893	PODS 0.107943	C. Mohan 0.00528346
query 0.0133316	EDBT 0.0436849	David J. DeWitt 0.00491615
systems 0.0110413		Hector Garcia-Molina 0.00453497
queries 0.0090603		H. V. Jagadish 0.00434289
management 0.00850744		David B. Lomet 0.00397865
object 0.00837766		Raghu Ramakrishnan 0.0039278
relational 0.0081175		Philip A. Bernstein 0.00376314
processing 0.00745875		Joseph M. Hellerstein 0.00372064
based 0.00736599		Jeffrey F. Naughton 0.00363698
distributed 0.0068367		Yannis E. Ioannidis 0.00359853
xml 0.00664958		Jennifer Widom 0.00351929
oriented 0.00589557		Per-Ake Larson 0.00334911
design 0.00527672		Rakesh Agrawal 0.00328274
web 0.00509167		Dan Suciu 0.00309047
information 0.0050518		Michael J. Franklin 0.00304099
model 0.00499396		Umeshwar Dayal 0.00290143
efficient 0.00465707		Abraham Silberschatz 0.00278185

- NetClus generates high quality clusters for all of the three participating types in the DBLP network
  - Quality can be easily judged by our commonsense knowledge
- Highly-ranked objects: Objects centered in the cluster

Rank-Based Clustering: Works in Multiple Domains

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

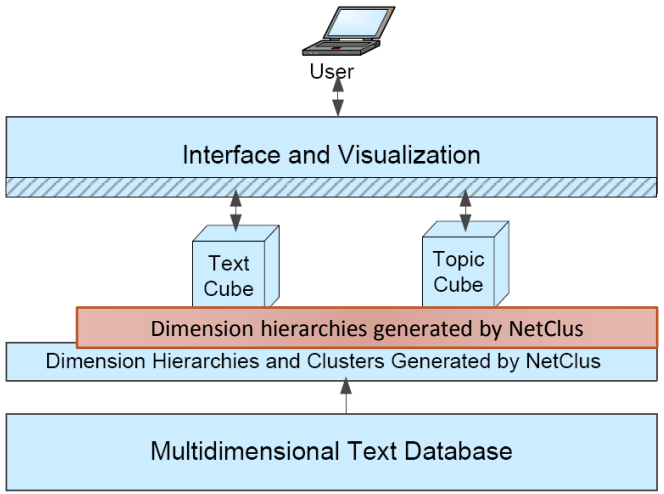
MedRank: Rank treatments for AIDS from Medline

Use multi-typed image features to build up heterogeneous networks

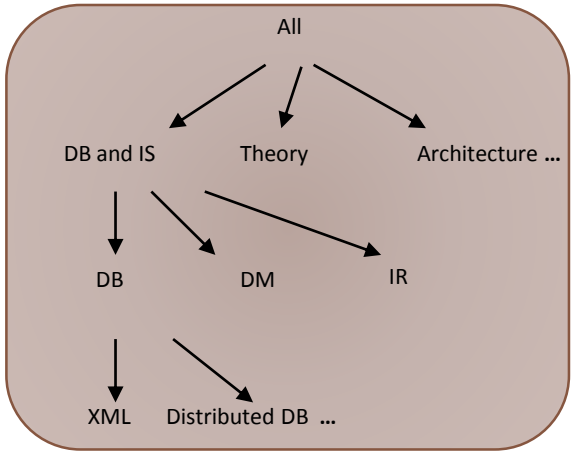
Explore multiple types in a star schema network

iNextCube: Information Network-Enhanced Text Cube

Demo: iNextCube.cs.uiuc.edu



Architecture of iNextCube



Net-cluster Hierarchy

Database and Information System | Search

In area **Database and Information System**, the top ranked conferences/journals are:

Rank	Conf/Journal	Score
1	SIGMOD Conference	0.075232
2	VLDB	0.062318
3	ICDE	0.053869
4	SIGIR	0.048740
5	KDD	0.028168
6	IEEE Trans. Knowl. Data Eng.	0.024118
7	SIGMOD Record	0.022818
8	IEEE Data Eng. Bull.	0.020792
9	CIKM	0.020606
10	ACM Trans. Database Syst.	0.015887
11	PODS	0.015577
12	TREC	0.014045
13	Inf. Process. Manage.	0.011904
14	ICDM	0.011839
15	VLDB J.	0.011544
16	EDBT	0.011441
17	SIGIR Forum	0.009575

Data Mining | Search

In sub-area **Data Mining**, the top ranked authors are:

Rank	Author	Score
1	Philip S. Yu	0.009881
2	Jiawei Han	0.007163
3	Charu C. Aggarwal	0.005638
4	Christos Faloutsos	0.005140
5	Beng Chin Ooi	0.003431
6	Ming-Syan Chen	0.003309
7	Hans-Peter Kriegel	0.003289
8	Wei Wang	0.003160
9	Kian-Lee Tan	0.003009
10	Nick Koudas	0.002989
11	H. V. Jagadish	0.002960

Database | Search

In sub-area **Database**, the top ranked authors are:

Rank	Author	Score
1	David B. Lomet	0.009049
2	Michael Stonebraker	0.007072
3	Richard T. Snodgrass	0.006152
4	David J. DeWitt	0.004660
5	Surajit Chaudhuri	0.004424
6	Michael J. Carey	0.004195
7	Won Kim	0.004167
8	Hector Garcia-Molina	0.003793
9	Michael J. Franklin	0.003773
10	Marianne Winslett	0.003753
11	C. Mohan	0.003580
12	Philip A. Bernstein	0.003459
13	Arie Segev	0.003191
14	Dan Suciu	0.003189
15	Gerhard Weikum	0.003043
16	Jennifer Widom	0.003033
17	H. V. Jagadish	0.002918
18	Jeffrey F. Naughton	0.002911
19	Raghu Ramakrishnan	0.002832
20	Joseph M. Hellerstein	0.002793
21	Rakesh Agrawal	0.002769
22	Umeshwar Dayal	0.002763
23	Serge Abiteboul	0.002737

Author/conference/ term ranking for each research area. The research areas can be at *different levels*.

Algorithms and Theory of Computation | Algorithms and Theory of Computation | Search

In area **Algorithms and Theory of Comp** conferences/journals are:

Rank	Conf/Journal
1	STOC
2	FOCS
3	SIAM J. Comput.
4	SODA
5	J. Comput. Syst. Sci.
6	J. ACM
7	CoRR
8	Theor. Comput. Sci.

In area **Algorithms and Theory of Computation**, the t authors are:

Rank	Author
1	Andrew Chi-Chih Yao
2	Christos H. Papadimitriou
3	Robert Endre Tarjan
4	David Eppstein
5	Micha Sharir
6	Avi Wigderson
7	John H. Reif
8	Bernard Chazelle

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
- …

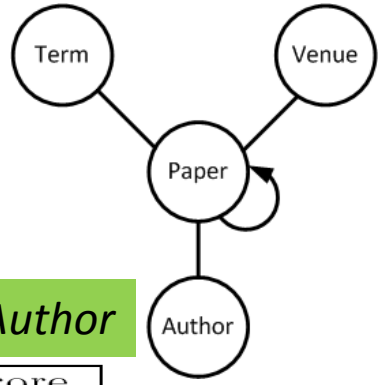

LIFE IS

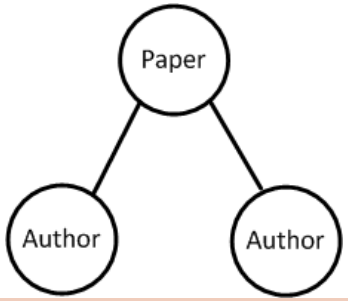
Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Information Networks
  - Relationship Prediction in Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation recommendation in heterogeneous networks
- Summary

Similarity Search in Heterogeneous Networks

- Similarity measure/search is the base for cluster analysis
- Who are the most similar to *Christos Faloutsos* based on the DBLP 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





Co-authorship
Meta-path: A-P-A

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' 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 in similar fields with similar reputation

Existing Popular Similarity Measures for Networks

- Random walk (RW):
  - The probability of random walk starting at *x* and ending at *y*, with meta-path *P*

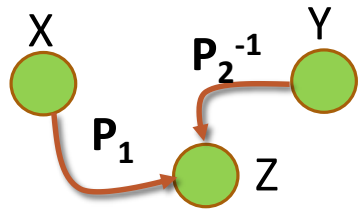
$$s(x, y) = \sum_{p \in P} prob(p)$$

  - Used in Personalized PageRank (P-Pagerank) (Jeh and Widom 2003)
  - Favors **highly visible** objects (i.e., objects with large degrees)

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

- Pairwise random walk (PRW):
  - The probability of pairwise random walk starting at (*x, y*) and ending at a common object (say *z*), following a meta-path (P_1, P_2)

$$s(x, y) = \sum_{(p_1, p_2) \in (P_1, P_2)} prob(p_1) prob(p_2)$$



    - Used in SimRank (Jeh and Widom 2002)
    - Favors ***pure*** objects (i.e., objects with highly skewed distribution in their in-links or out-links)

SimRank and Personalized PageRank

- SimRank (Jeh and Widom 2002)
  - Base: objects are maximally similar to themselves, i.e., $s_0(a,b) = \begin{cases} 1\ , & \text{if } a = b\ , \\ 0\ , & \text{if } a \neq b\ . \end{cases}$
  - Induction: Two objects are considered to be similar if they are referenced by similar objects $s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$
  - The computation is quite costly: Many efficient computation methods proposed

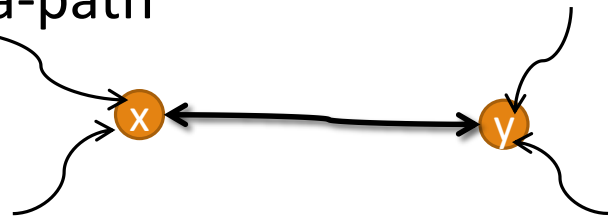
Glen Jeh and Jennifer Widom. SimRank: A Measure of Structural-Context Similarity. In KDD'02

- Personalized PageRank (P-Pagerank) (Jeh and Widom 2003)
  - P-PageRank score x is defined as: $x = \alpha P x + (1 - \alpha) b$, where P is a transition matrix of the network G, b is a stochastic vector, called *personalized vector*, and $\alpha \in (0, 1)$ is the *teleportation constant*
  - Efficient computation methods are also studied (e.g., Maehara, et al., VLDB'14)

Glen Jeh and Jennifer Widom. Scaling Personalized Web Search, In *WWW* 2003

Which Similarity Measure Is Better for Finding Peers?

- PathSim: Favors peers
  - **Peers**: Objects with strong connectivity and similar visibility with a given meta-path



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

  - Meta-path: APCPA
  - Mike publishes similar # of papers as Bob and Mary
  - Other measures find Mike is closer to Jim

Author\Conf.	SIGMOD	VLDB	ICDM	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

Measure\Author	Jim	Mary	Bob	Ann
P-PageRank	**0.376**	0.013	0.016	0.005
SimRank	**0.716**	0.572	0.713	0.184
Random Walk	**0.8983**	0.0238	0.0390	0
Pairwise R.W.	**0.5714**	0.4440	0.5556	0
PathSim (APCPA)	0.083	**0.8**	**1**	0

Comparison of Multiple Measures: A Toy Example

Example with DBLP: 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***

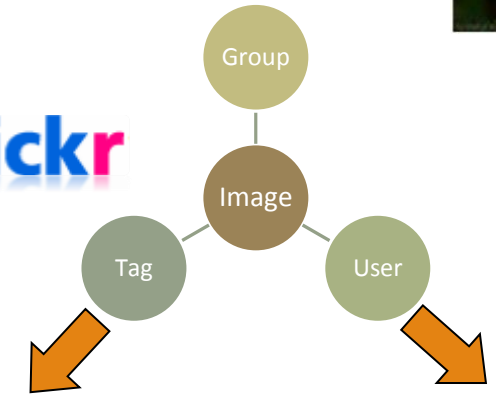
Rank	P-PageRank	SimRank	PathSim
1	AnHai Doan	AnHai Doan	AnHai Doan
2	Philip S. Yu	Douglas W. Cornell	Jignesh M. Patel
3	Jiawei Han	Adam Silberstein	Amol Deshpande
4	Hector Garcia-Molina	Samuel DeFazio	Jun Yang
5	Gerhard Weikum	Curt Ellmann	Renée J. Miller

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

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

Some Meta-Path Is "Better" Than Others

- Which pictures are most similar to this one?



Evaluate the similarity between images according to their linked tags

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

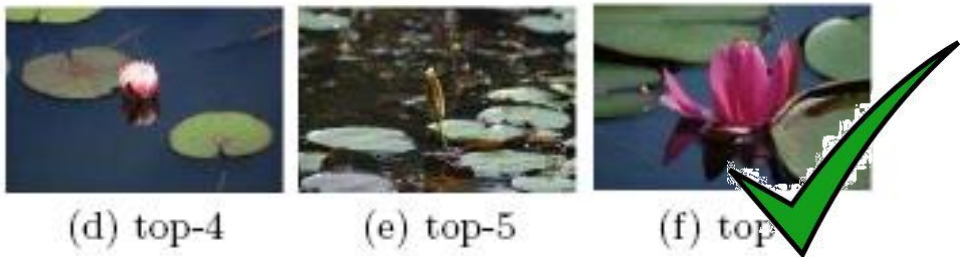
(a) top-1 (b) top-2 (c) top-3

(d) top-4 (e) top-5 (f) top-6

Evaluate the similarity between images according to tags and groups

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

(a) top-1 (b) top-2 (c) top-3



(d) top-4 (e) top-5 (f) top

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?

Are 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

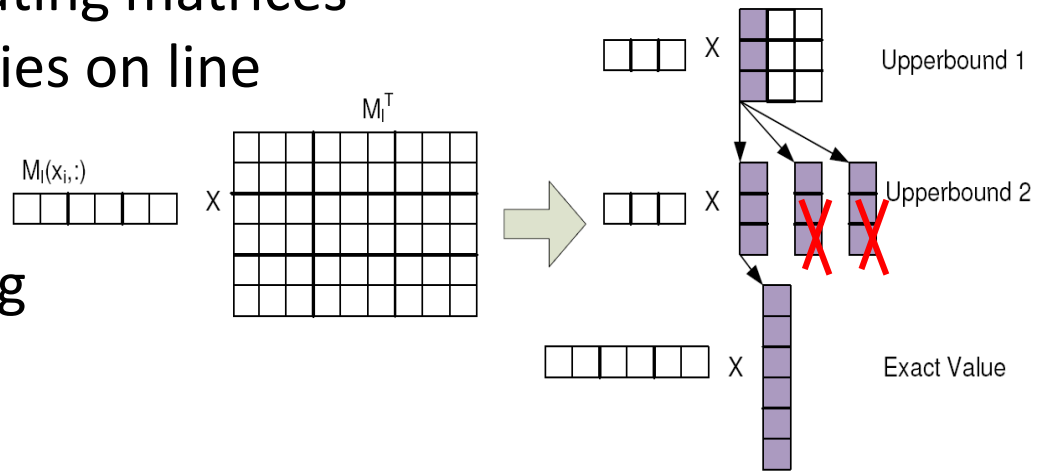
(c) Path: $(CPAPC)^\infty$

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

Co-Clustering-Based Pruning Algorithm

- Meta-Path based similarity computation can be costly
- The overall cost can be reduced by storing commuting matrices for short path schemas and computing top-*k* queries on line
- Framework
  - Generate co-clusters for materialized commuting matrices for feature objects and target objects
  - Derive upper bound for similarity between object and target cluster and between object and object
    - Safely prune target clusters and objects if the upper bound similarity is lower than current threshold
  - Dynamically update top-*k* threshold



Meta-Path: A Key Concept for Heterogeneous Networks

- Meta-path based mining
  - 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
- Meta-path learning
  - User Guided Meta-Path Selection [Sun et al., KDD'12]
    - Meta-path selection + clustering

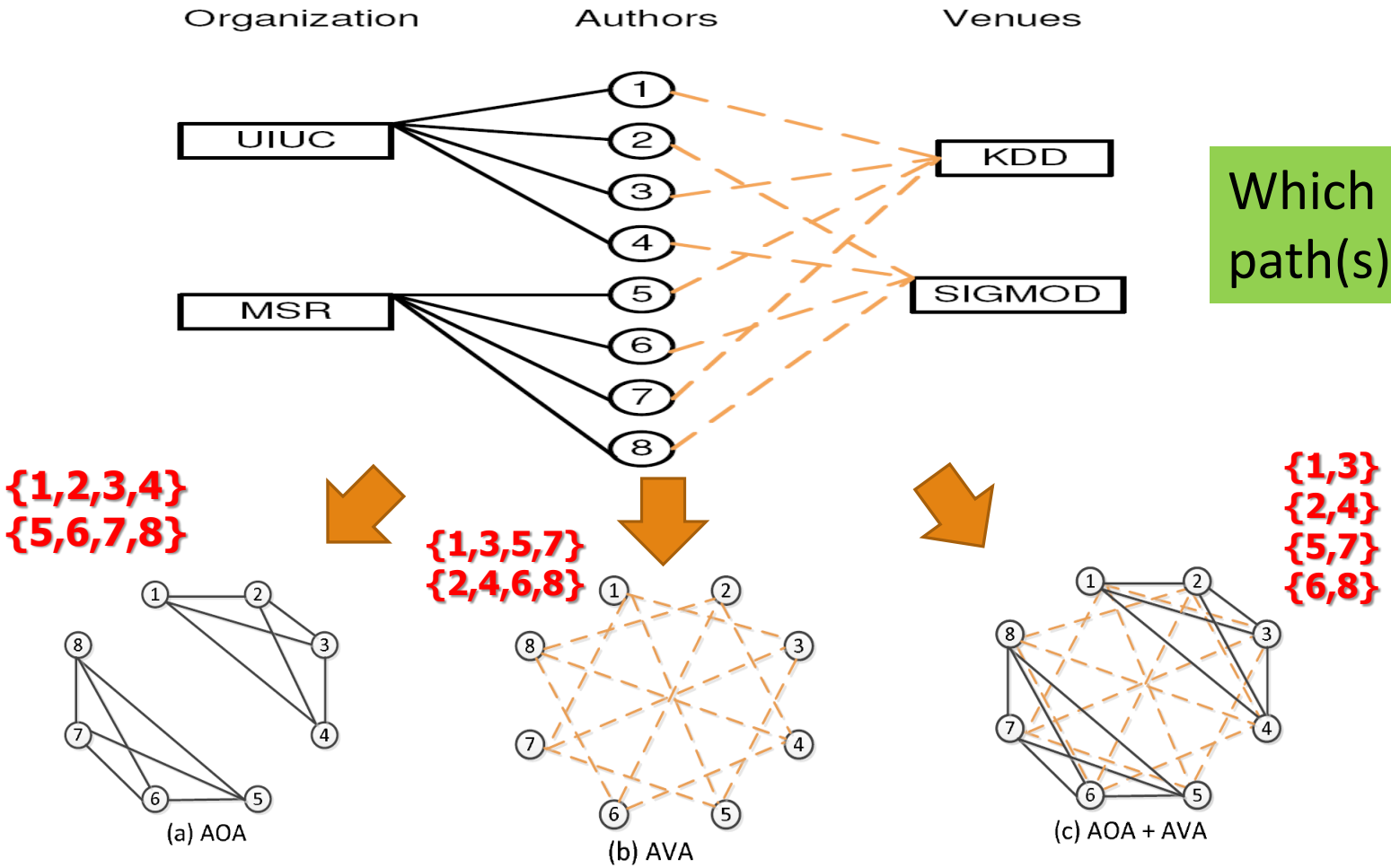

NOBELS FREDSSENTER • NOBEL PEACE CENTER
epea
EUROPEISKE IDENTITETER
epea
EUROPEAN IDENTITIES
7 March–15 September
VELKOMMEN
NOBELS HAGE
NOBEL FIELD
VISIT PEACE

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks 
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks 
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Information Networks
  - Relationship Prediction in Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation recommendation in heterogeneous networks
- Summary

Why User Guidance in Clustering?

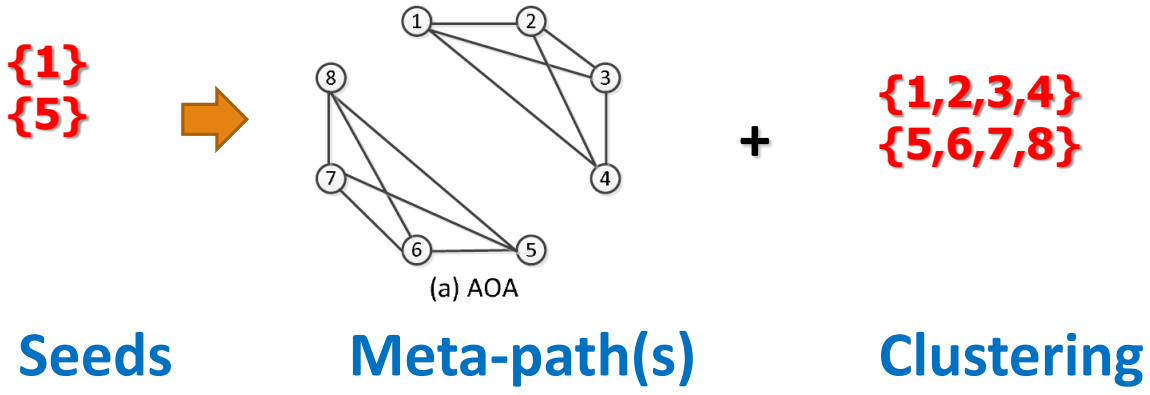
- Different users may like to get different clusters for different clustering goals
  - Ex. Clustering authors based on their connections in the network



Which meta-path(s) to choose?

User Guidance Determines Clustering Results

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



- **Problem: User-guided clustering with meta-path selection**
- Input:
  - The target type for clustering T
  - # of clusters k
  - Seeds in some clusters: $L_1, \ldots, L_k$
  - Candidate meta-paths: $P_1, \ldots, P_M$
- Output:
  - Weight of each meta-path: $w_1, \ldots, w_m$
  - Clustering results that are consistent with the user guidance

PathSelClus: A Probabilistic Modeling Approach

- Part 1: Modeling the Relationship Generation
  - A good clustering result should lead to high likelihood in observing existing relationships
    - 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

Part 1: Modeling the Relationship Generation

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

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

52

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^*
      - The prior density is a K-d Dirichlet distribution with parameter vector $\lambda \boldsymbol{e}_{k^*} + \mathbf{1}$
        - » $\boldsymbol{e}_{k^*}$ 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
      - The prior density is uniform, a special case of Dirichlet distribution, with parameter vector $\mathbf{1}$

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

53

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: $W_m \rightarrow \alpha_m W_m$
- The best α_m: the most likely to generate current clustering-based parameters
  - $$\alpha_m^* = \arg\max_{\alpha_m} \prod_i P(\boldsymbol{\pi}_{i,m} | \alpha_m \mathbf{w}_{i,m}, \boldsymbol{\theta}_i, B_m)$$

Dirichlet Distribution

    - 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

54

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

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

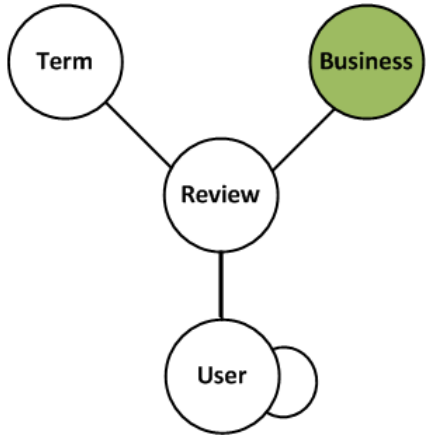
  - Step 2: Optimize $\boldsymbol{\alpha}$ given Θ
    - In general, the higher likelihood of observing W_m given Θ, the higher α_m

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

55

Effectiveness of Meta-Path Selection

- Experiments on Yelp data
  - Object Types: Users, Businesses, Reviews, Terms
  - Relation Types: UR, RU, BR, RB, TR, RT
- Task: Candidate meta-paths: *B-R-U-R-B, B-R-T-R-B*
  - Target objects: restaurants
  - # of clusters: 6
- Output:
  - PathSelClus vs. others
    - High accuracy
  - Restaurant vs. shopping
    - For restraunts, meta-path B-R-U-R-B weighs only 0.1716
    - For clustering shopping, B-R-U-R-B weighs 0.5864



%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

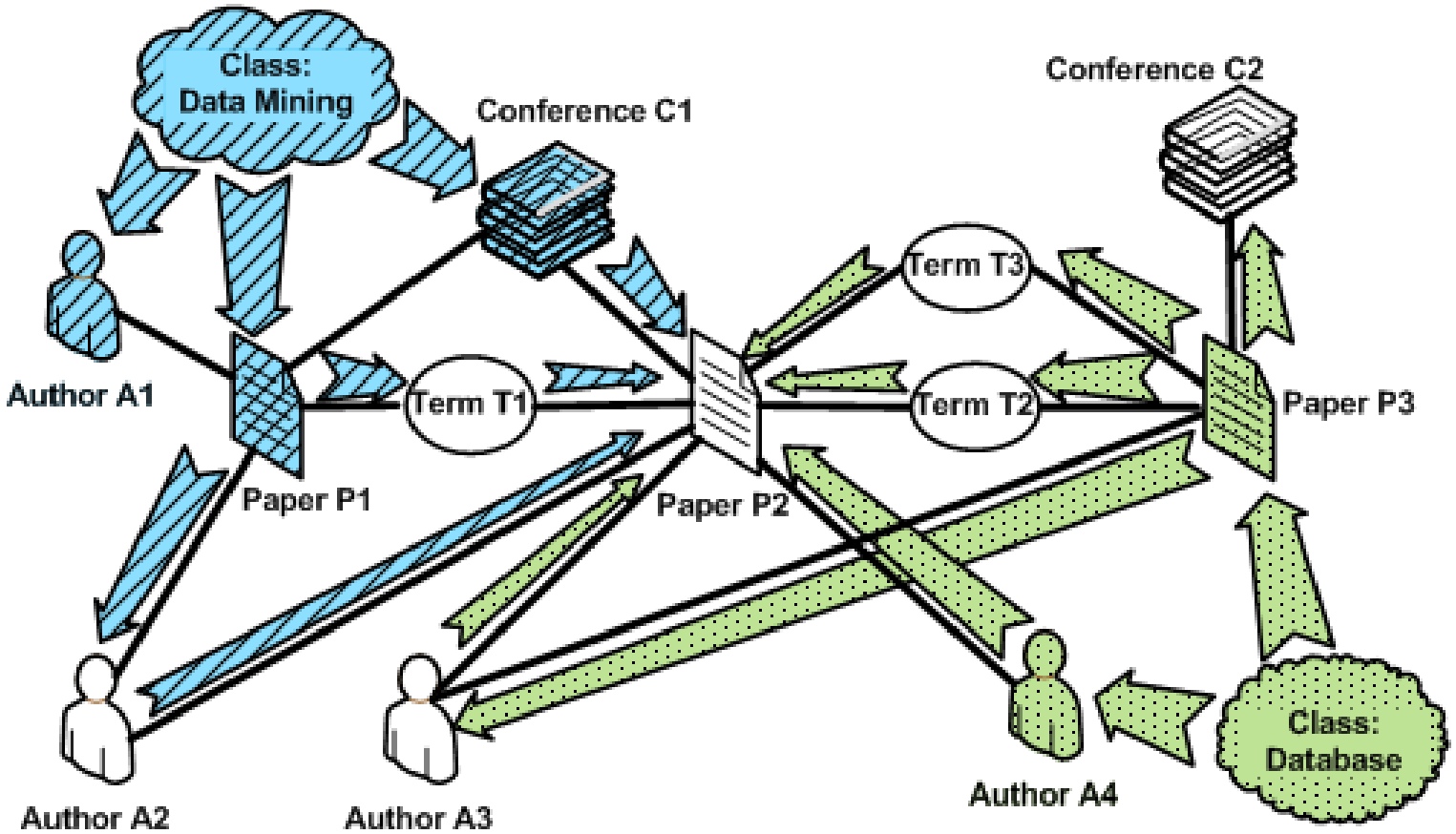
Users try different kinds of food

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks
- Summary

Classification: Knowledge Propagation Across Heterogeneous Typed Networks

- **RankClass [Ji et al., KDD'11]:** Ranking-based classification
  - Highly ranked objects will play more role in classification
  - Class membership and ranking are statistical distributions
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class



Classification: Labeled knowledge propagates through multi-typed objects across heterogeneous networks [KDD'11]

GNetMine: Methodology

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

GNetMine: Graph-Based Regularization

- Minimize the objective function

$$J(\boldsymbol{f}_1^{(k)},\ldots,\boldsymbol{f}_m^{(k)}) = \sum_{i,j=1}^{m} \lambda_{ij} \sum_{p=1}^{n_i}\sum_{q=1}^{n_j} R_{ij,pq}\left(\frac{1}{\sqrt{D_{ii,pp}}} f_{ip}^{(k)} - \frac{1}{\sqrt{D_{ji,qq}}} f_{jq}^{(k)}\right)^2 + \sum_{i=1}^{m} \alpha_i (\boldsymbol{f}_i^{(k)} - \mathbf{y}_i^{(k)})^T (\boldsymbol{f}_i^{(k)} - \mathbf{y}_i^{(k)})$$

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

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 objects

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

RankClass: Ranking-Based Classification

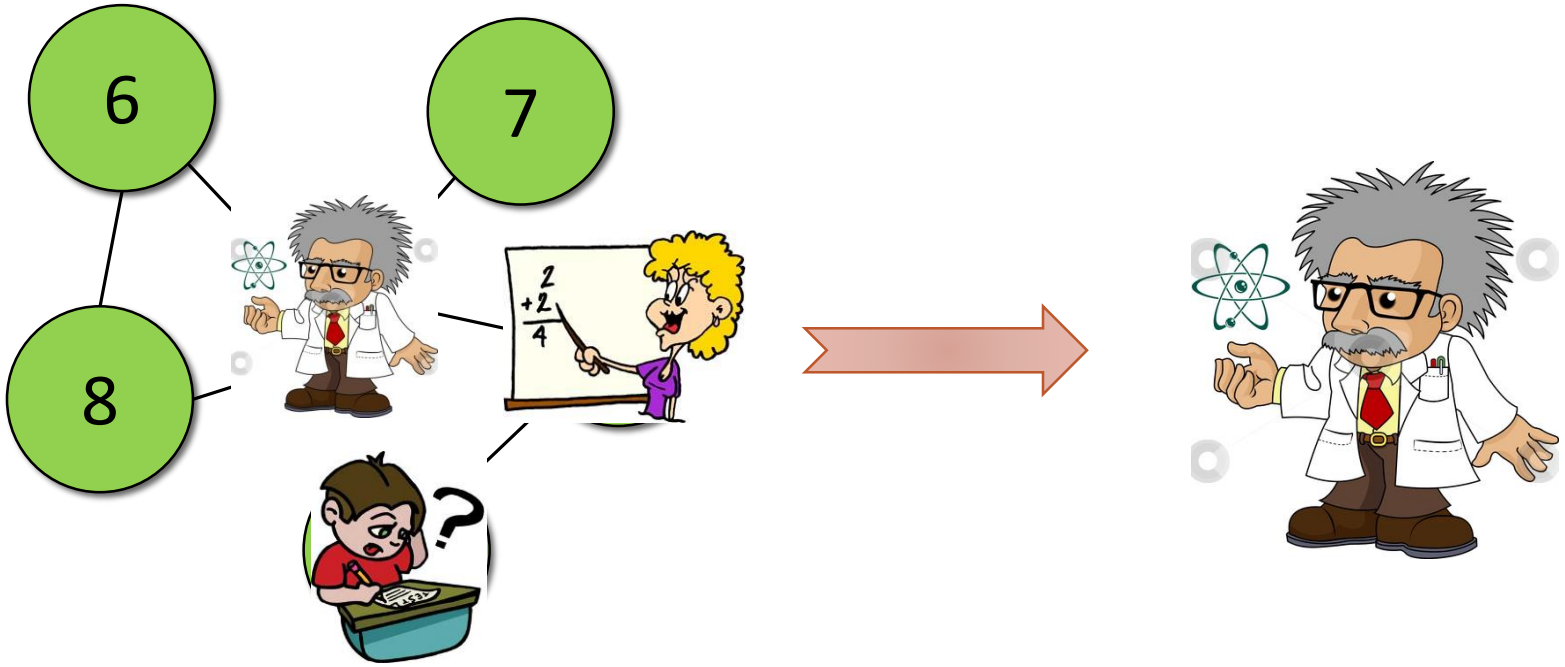
- **Classification in heterogeneous networks**
  - **Knowledge propagation:** Class label knowledge propagated across multi-typed objects through a heterogeneous network
- **GNetMine [Ji et al., PKDD'10]:** Objects are treated equally
- **RankClass [Ji et al., KDD'11]:** Ranking-based classification
  - Highly ranked objects will play more role in classification
  - An object can only be ranked high in some focused classes
  - Class membership and ranking are stat. distributions
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class

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. label information 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

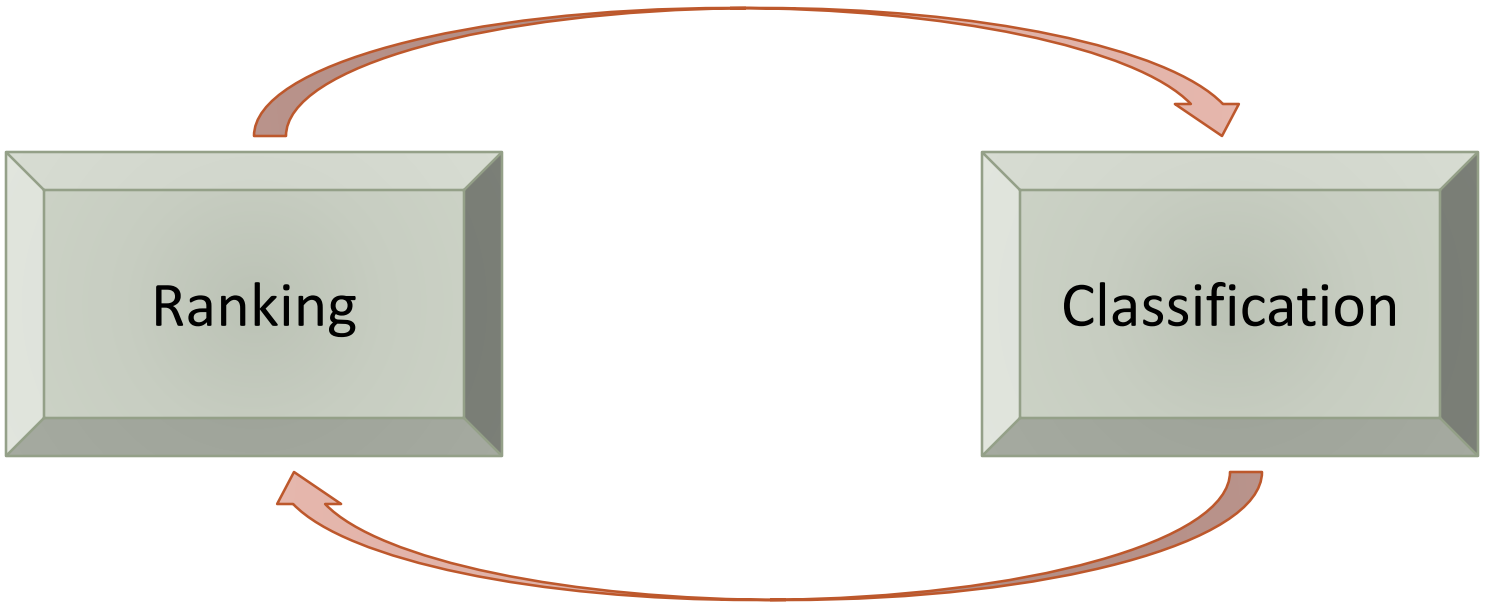
Why Rank-Based Classification?

- Different objects within one class have different importance/visibility!
- The ranking of objects within one class serves as an informative understanding and summary of the class



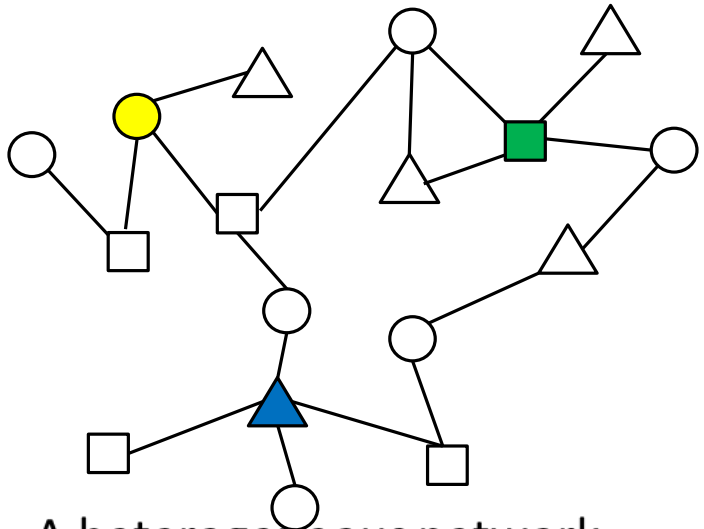
Motivation

- Why not do ranking after classification, or vice versa?
  - Because they mutually enhance each other, not unidirectional.
- RankClass: iteratively let ranking and classification mutually enhance each other



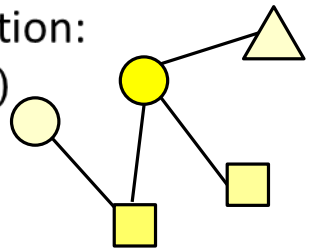
RankClass: Ranking-Based Classification

Class: a group of multi-typed nodes sharing the same topic + within-class ranking

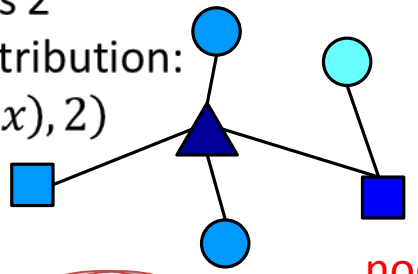


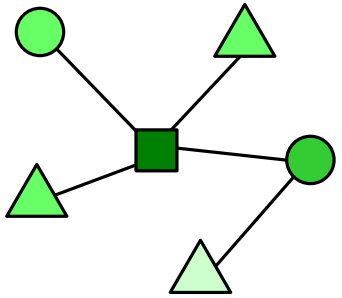
A heterogeneous network contains $m(=3)$ types of nodes

Class 1
ranking distribution:
$P(x|T(x),1)$



Class 2
ranking distribution:
$P(x|T(x),2)$





Class 3
ranking distribution:
$P(x|T(x),3)$

node
the type of x
class

Initialize $t=1$, $P(x|T(x),k)^0$

Ranking: update $P(x|T(x),k)^t$

Classification: use $P(x|T(x),k)^t$ to describe each class, softly partition the network

$t=t+1$

Converge

Calculate the posterior probability, i.e., $P(k|x,T(x))$

$$\text{class}(x) = \underset{1\le k\le K}{\text{argmax}}\, P(k|x,T(x))$$

The RankClass Framework

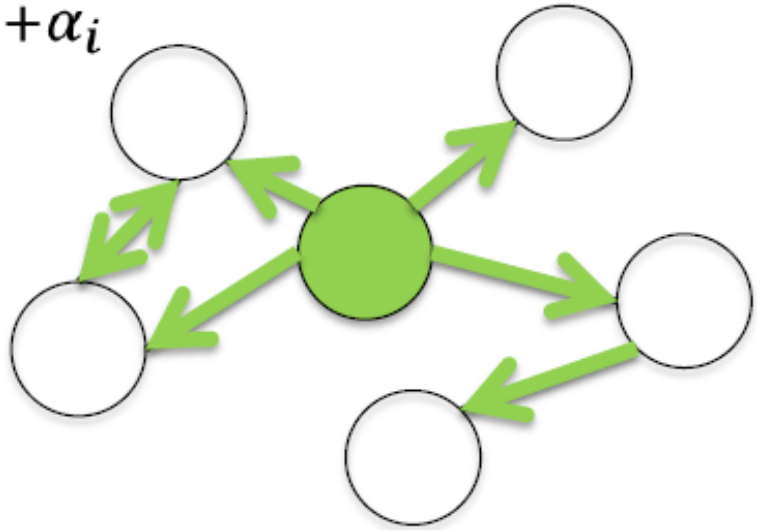
Graph-Based Ranking

- Intuitive idea: authority propagation
  - Objects linked together are likely to share similar ranking scores within class k
- Update the ranking score of each object by looking t the ranking of its neighbors

The initial ranking score

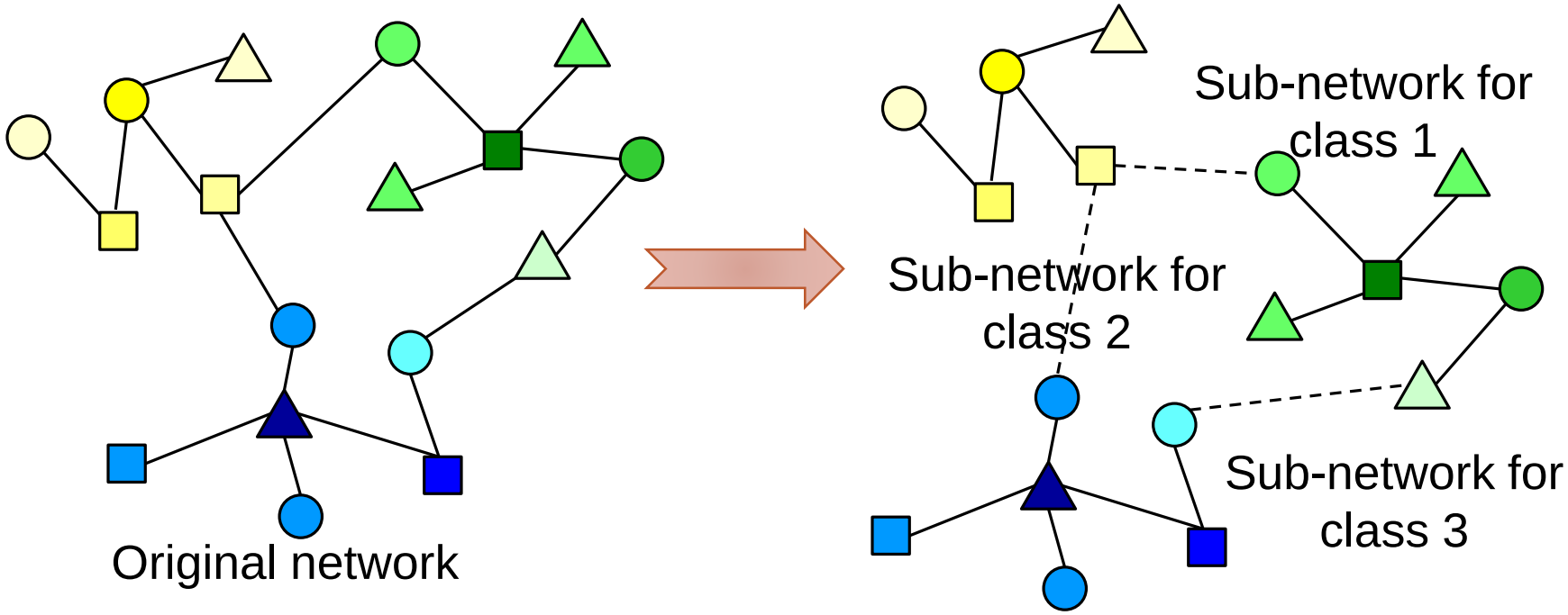
$$P(x_{ip}|\chi_i,k)^{t+1} \propto \frac{\sum_{j=1}^{m}\sum_{q=1}^{n_j}\lambda_{ij}S_{ij,pq}P(x_{jq}|\chi_j,k)^{t}+\alpha_i P(x_{ip}|\chi_i,k)^{0}}{\sum_{j=1}^{m}\lambda_{ij}+\alpha_i}$$

Weighted average of the neighbors' ranking scores



Partition the Network Softly

- Ideally, the within-class ranking should be performed within the sub-network corresponding to each class
  - Use the ranking distribution to describe each class
  - Gradually emphasize the network on highly ranked objects, and weaken the network on lowly ranked objects



Comparing Classification Accuracy on the DBLP Data

- Class: Four research areas: DB, DM, AI, IR
- 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
  - LLGC [Zhou et al. NIPS'03]
  - wvRN) [Macskassy et al. JMLR'07]
  - nLB [Lu et al. ICML'03, Macskassy et al. JMLR'07]

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**

Object Ranking in Each Class: Experiment

- 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 ranking within each class

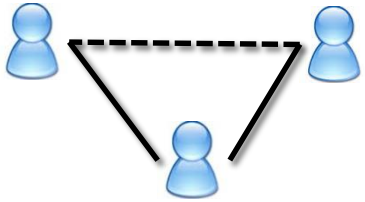
	Database	Data Mining	AI	IR
Top-5 ranked conferences	VLDB	KDD	IJCAI	SIGIR
	SIGMOD	SDM	AAAI	ECIR
	ICDE	ICDM	ICML	CIKM
	PODS	PKDD	CVPR	WWW
	EDBT	PAKDD	ECML	WSDM
Top-5 ranked terms	data	mining	learning	retrieval
	database	data	knowledge	information
	query	clustering	reasoning	web
	system	classification	logic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

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks
- Summary

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
  - Different types of relationships need different prediction models

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

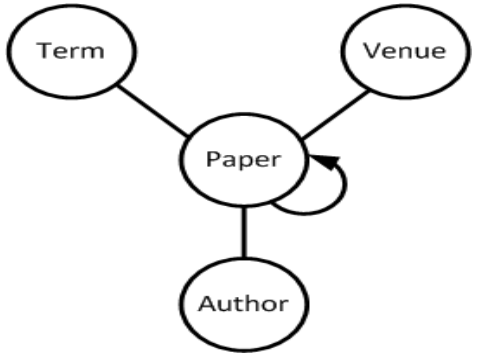
Why Prediction Using Heterogeneous Info Networks?

- Rich semantics contained in structured, text-rich heterogeneous networks
  - Homogeneous networks, such as coauthor networks, miss too much criticially important information
- Schema-guided relationship prediction
  - Semantic relationships among similar typed links share similar semantics and are comparable and inferable
  - Relationships across different typed links are not directly comparable but their collective behavior will help predict particular relationships
- Meta-paths: encoding topological features write cite write
  - E.g., citation relations between authors: A — P → P — A
- Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks", ASONAM'11

PathPredict: Meta-Path Based Relationship Prediction

- Who will be your new coauthors in the next 5 years?
  - Meta path-guided prediction of links and relationships

- Philosophy: Meta path relationships among similar typed links share similar semantics and are comparable and inferable
- Co-author prediction (A—P—A) [Sun et al., ASONAM'11]
  - Use topological features encoded by meta paths, e.g., citation relations between authors (A—P→P—A)



Meta-Path

Semantic Meaning

Meta-paths between authors of length ≤ 4

$A-P\rightarrow P-A$	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

Logistic Regression-Based Prediction Model

- Training and test pair: <$\mathbf{x}_i$, y_i> = <history feature list, future relationship label>

	A—P—A—P—A	A—P—V—P—A	A—P—T—P—A	A—P→P—A	A—P—A
<Mike, Ann>	4	5	100	3	**Yes = 1**
<Mike, Jim>	0	1	20	2	**No = 0**

- Logistic Regression Model
  - Model the probability for each relationship as

$$p_i = \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{e^{\mathbf{x}_i \boldsymbol{\beta}} + 1}$$

    - β is the coefficients for each feature (including a constant 1)
  - MLE (Maximum Likelihood Estimation)
    - Maximize the likelihood of observing all the relationships in the training data

$$L = \prod_i p_i^{y_i} (1 - p_i)^{(1-y_i)}$$

Selection among Competitive Measures

We study four measures that defines a relationship *R* encoded by a meta path

- Path Count: Number of path instances between authors following *R*

$$PC_R(a_i, a_j)$$

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

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

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

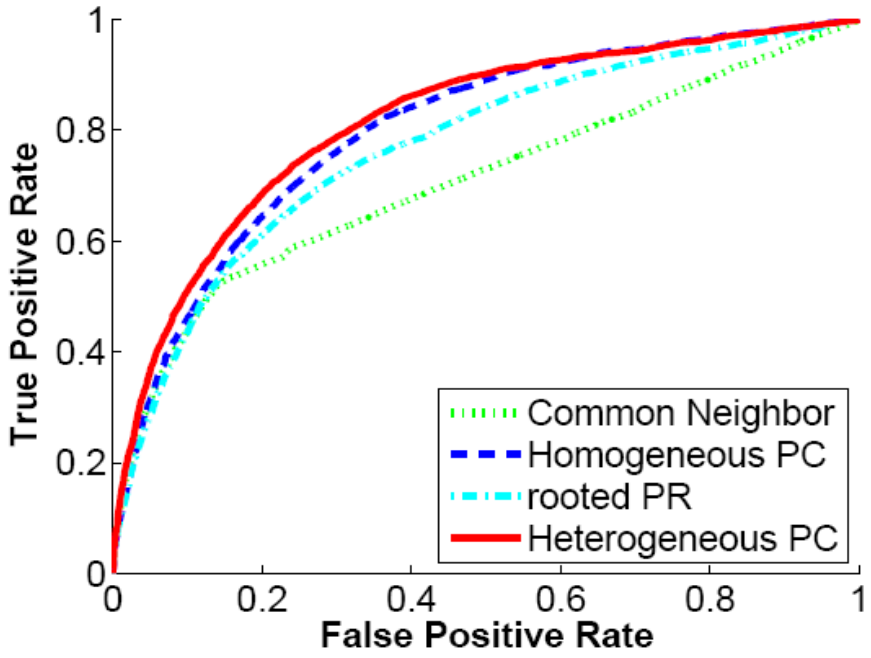
$$RW_R(a_i, a_j) = \frac{PC_R(a_i,a_j)}{PC_R(a_i,\cdot)}$$

- Symmetric Random Walk: Consider random walk in both directions

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

Performance Comparison: Homogeneous vs. Heterogeneous Topological Features

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

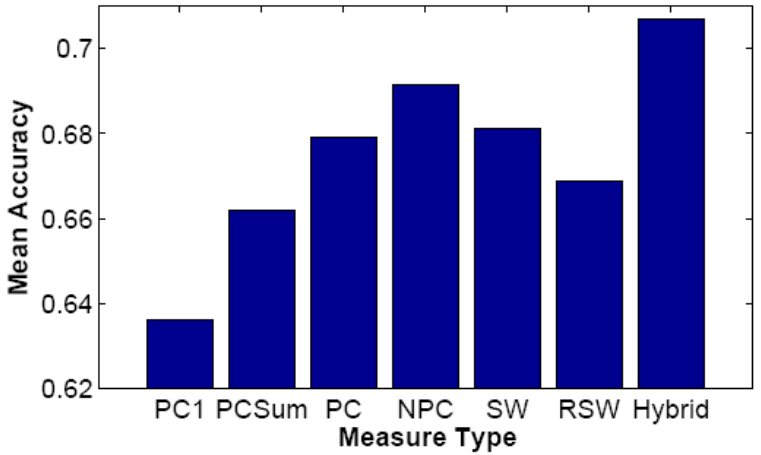


Dataset	Topological features	Accuracy	AUC
$HP2hop$	common neighbor	0.6053	0.6537
	homogeneous PC	0.6433	0.7098
	heterogeneous PC	**0.6545**	**0.7230**
$HP3hop$	common neighbor	0.6589	0.7078
	homogeneous PC	0.6990	0.7998
	rooted PageRank	0.6433	0.7098
	heterogeneous PC	**0.7173**	**0.8158**
$LP2hop$	common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

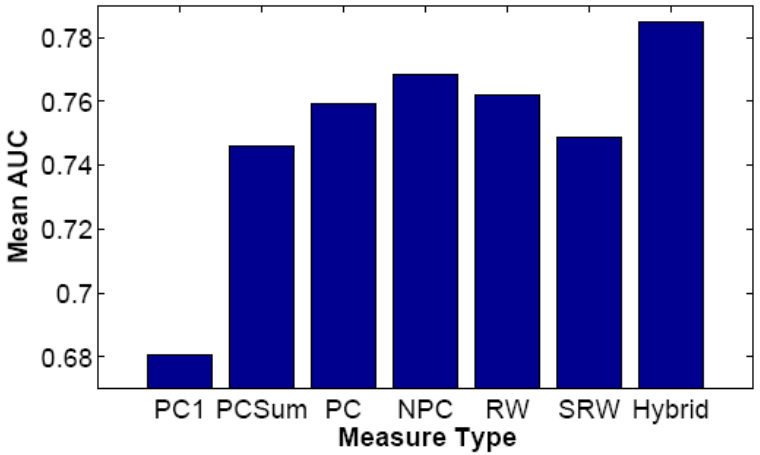
Notation: *HP2hop*: highly productive source authors with 2-hops reaching target authors

Comparison among Different Topological Features

- Hybrid heterogeneous topological feature is the best



(a) Mean accuracy



(b) Mean AUC

Notations
(1) the path count (*PC*)
(2) the normalized path count (*NPC*)
(3) the random walk (*RW*)
(4) the symmetric random walk (*SRW*)

*PC*1: homogeneous common neighbor
PCSum: homogeneous path count

The Power of PathPredict: Experiment on DBLP

- Explain the prediction power of each meta-path
  - Wald Test for logistic regression

Evaluation of the prediction power of different meta-paths

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

Social relations play more important role?

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

Prediction of new coauthors of Jian Pei in [2003-2009]

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

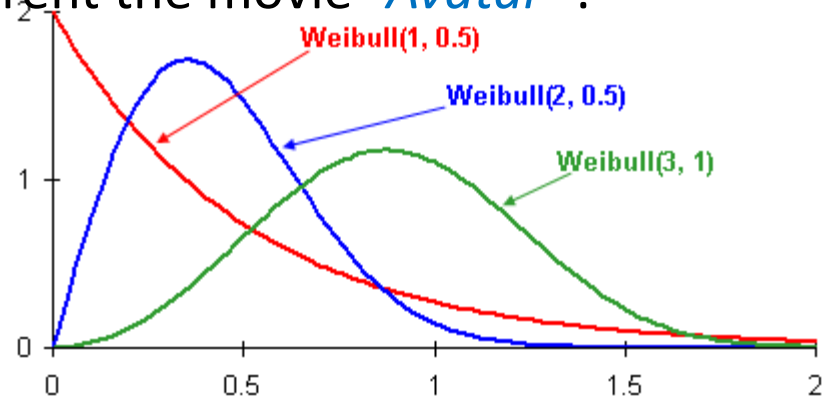
Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors!
(Feature collected in [1996, 2002]; Test period in [2003,2009])

When Will It Happen?

- 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"*?



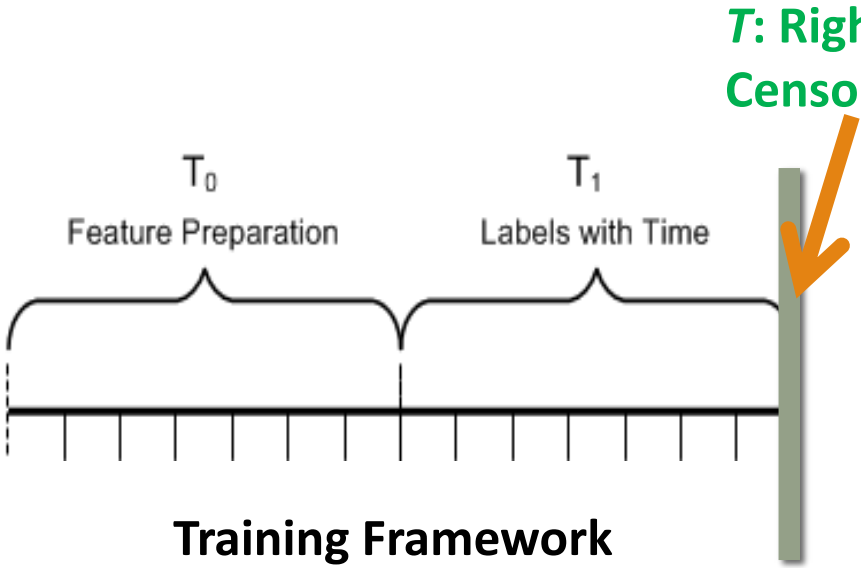
Output: A distribution of time!

  - 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)



Training Framework

***T*: Right Censoring**

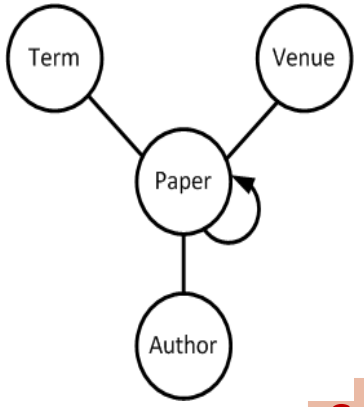
$$\log L = \sum_{i=1}^{n} (f_Y(y_i|\theta_i, \lambda) I_{\{y_i<T\}} + P(y_i \geq T|\theta_i, \lambda) I_{\{y_i \geq T\}})$$

Generalized Linear Model under Weibull Distribution Assumption

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

Author Citation Time Prediction in DBLP

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



$$A - P - T - P - A$$
$$A - P \leftarrow P \rightarrow P - A$$
$$A - P - A - P \rightarrow P - A$$
$$A - P - T - P - A - P \rightarrow P - A$$

Study the same topic

Co-cited by the same paper

Follow co-authors' citation

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

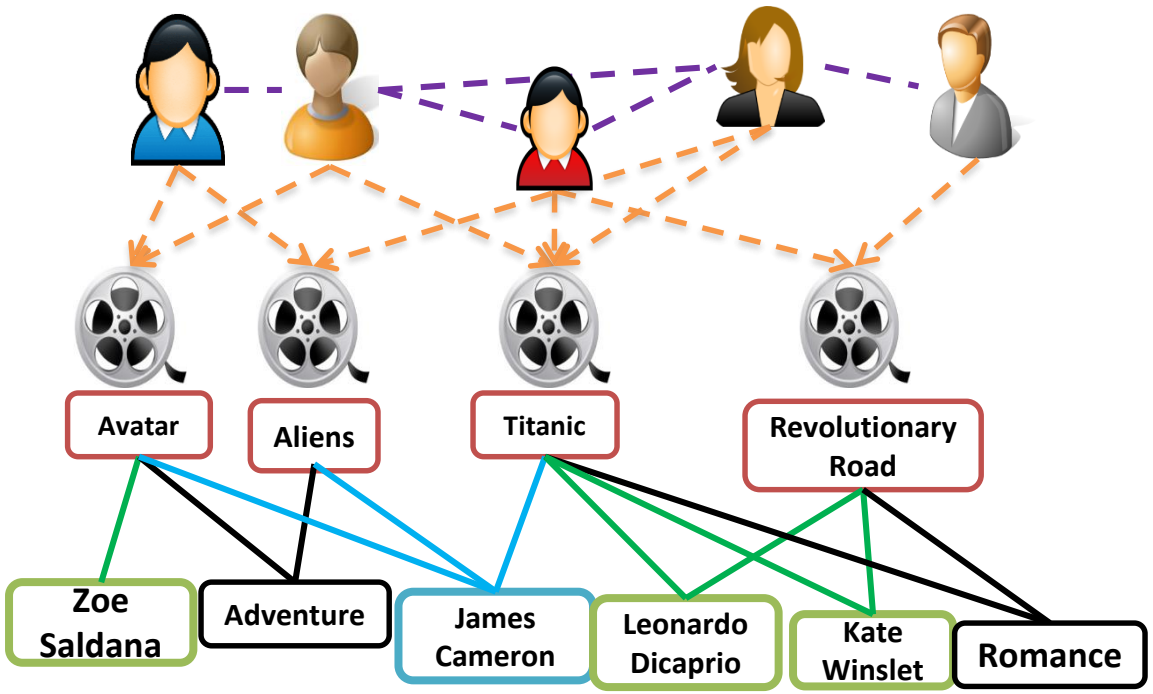
Under Weibull distribution assumption

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

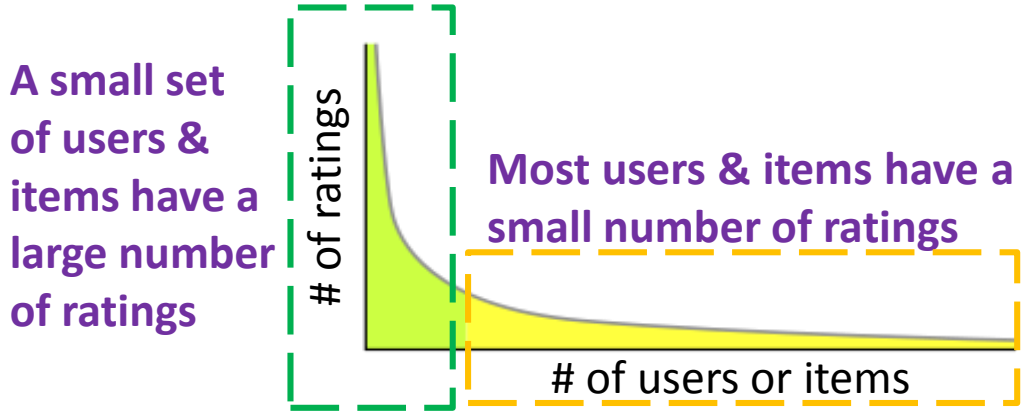
Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks
- Summary

Enhancing the Power of Recommender Systems by Heterog. Info. Network Analysis



Collaborative filtering methods suffer from the data sparsity issue

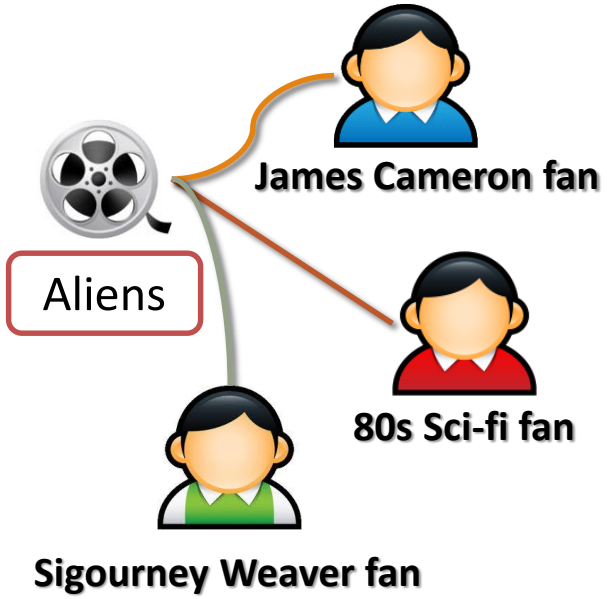


Personalized recommendation with heterog. Networks [WSDM'14]

- Heterogeneous relationships complement each other
- Users and items with limited feedback can be connected to the network by **different types of paths**
  - Connect new users or items in the information network
- Different users may require different models: Relationship heterogeneity makes **personalized recommendation** models easier to define

Relationship Heterogeneity Based Personalized Recommendation Models

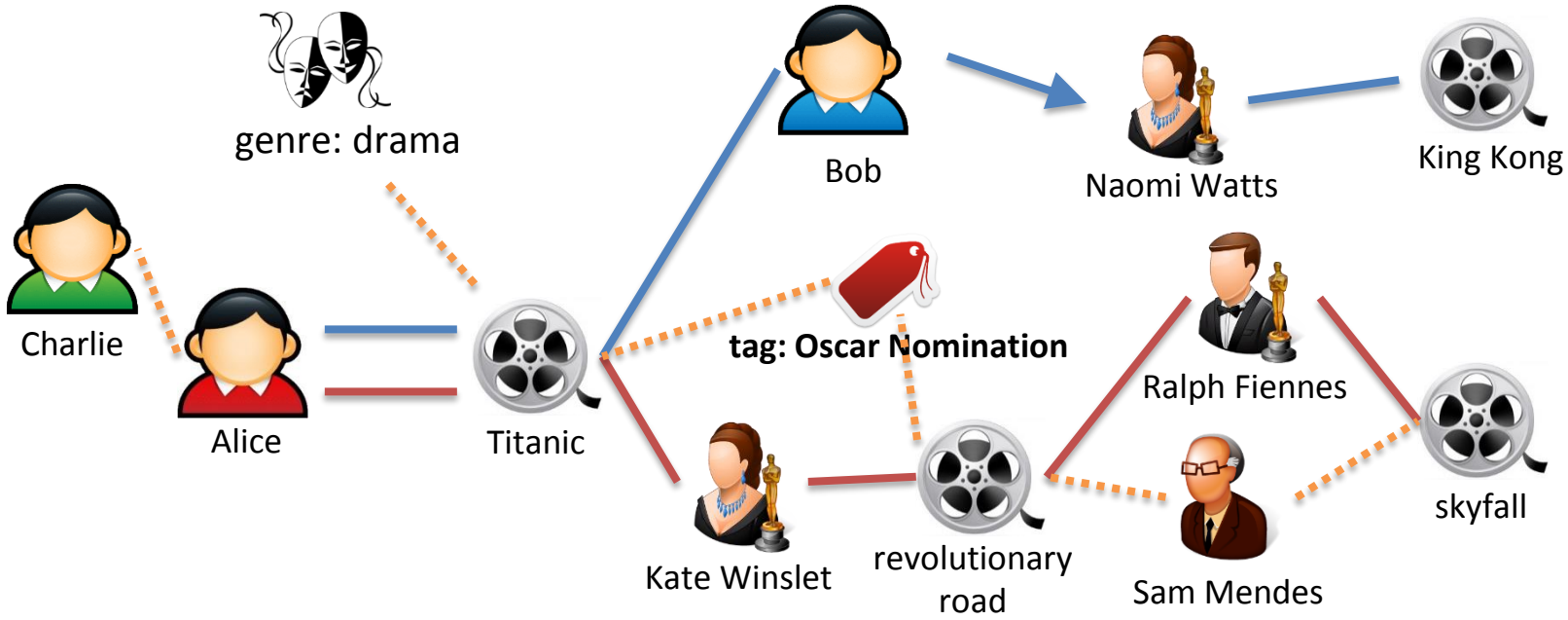
Different users may have different behaviors or preferences



Different users may be interested in the same movie for different reasons

- Two levels of personalization
  - Data level
    - Most recommendation methods use one model for all users and rely on personal feedback to achieve personalization
  - Model level
    - With different entity relationships, we can learn personalized models for different users to further distinguish their differences

Preference Propagation-Based Latent Features



Generate *L* different **meta-path** (path types) connecting users and items

Propagate user implicit feedback along each meta-path

Calculate latent-features for users and items for each meta-path with **NMF** related method

Recommendation Models

Observation 1: Different meta-paths may have different importance

Global Recommendation Model

ranking score

features for user *i* and item *j*

$$\hat{r}(u_i, e_j) = \sum_{q=1}^{L} \theta_q \cdot \hat{U}_i^{(q)} \hat{V}_j^{(q)T} \quad (1)$$

the q-th meta-path

Observation 2: Different users may require different models

Personalized Recommendation Model

user-cluster similarity

$$\hat{r}_p(u_i, e_j) = \sum_{k=1}^{c} sim(C_k, u_i) \sum_{q=1}^{L} \theta_q^{\{k\}} \cdot \hat{U}_i^{(q)} \hat{V}_j^{(q)T} \quad (2)$$

c total soft user clusters

Parameter Estimation

- Bayesian personalized ranking (Rendle UAI'09)
- Objective function

sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$.

$$\min_{\Theta} \; -\sum_{u_i \in \mathcal{U}} \sum_{(e_a > e_b) \in R_i} \ln \sigma(\hat{r}(u_i, e_a) - \hat{r}(u_i, e_b)) + \frac{\lambda}{2}\|\Theta\|_2^2 \qquad (3)$$

for each correctly ranked item pair
i.e., u_i gave feedback to e_a but not e_b

Soft cluster users with NMF + k-means → **For each user cluster, learn one model with Eq. (3)** → **Generate personalized model for each user on the fly with Eq. (2)**

Learning Personalized Recommendation Model

Experiments: Heterogeneous Network Modeling Enhances the Quality of Recommendation

- Datasets

Name	#Items	#Users	#Ratings	#Entities	#Links
IM100K	943	1360	89,626	60,905	146,013
Yelp	11,537	43,873	229,907	285,317	570,634

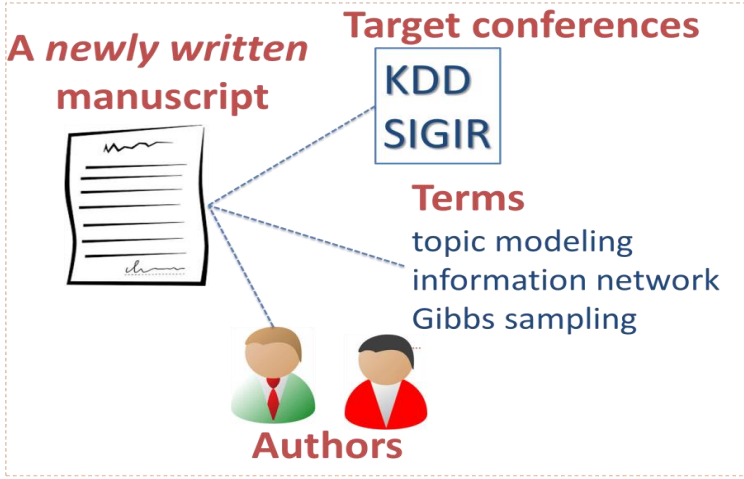
- Comparison methods
  - Popularity: recommend the most popular items to users
  - Co-click: conditional probabilities between items
  - NMF: non-negative matrix factorization on user feedback
  - Hybrid-SVM: use Rank-SVM with plain features (utilize both user feedback and information network)

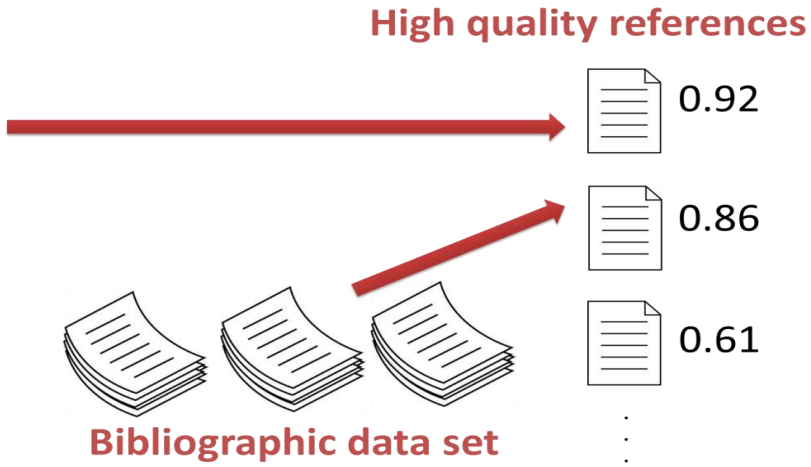
Method	IM100K				Yelp			
	Prec1	Prec5	Prec10	MRR	Prec1	Prec5	Prec10	MRR
Popularity	0.0731	0.0513	0.0489	0.1923	0.00747	0.00825	0.00780	0.0228
Co-Click	0.0668	0.0558	0.0538	0.2041	0.0147	0.0126	0.01132	0.0371
NMF	0.2064	0.1661	0.1491	0.4938	0.0162	0.0131	0.0110	0.0382
Hybrid-SVM	0.2087	0.1441	0.1241	0.4493	0.0122	0.0121	0.0110	0.0337
HeteRec-g	0.2094	0.1791	0.1614	0.5249	0.0165	0.0144	0.0129	0.0422
HeteRec-l	**0.2121**	**0.1932**	**0.1681**	**0.5530**	**0.0213**	**0.0171**	**0.0150**	**0.0513**

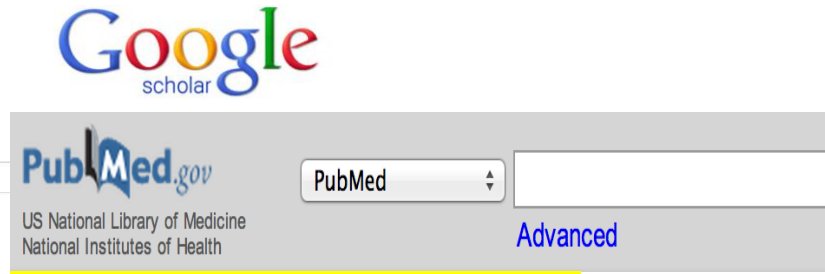
HeteRec personalized recommendation (HeteRec-p) leads to the best recommendation

ClusCite: Citation Recommendation by Info. Net-Based Clustering

Citation Recommendation: Given a manuscript (title, abstract and/or content) and its attributes (authors, target venues), suggest a small set of high quality references

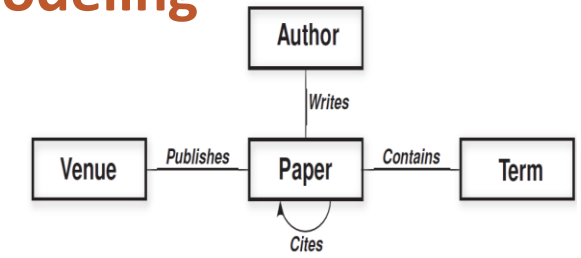






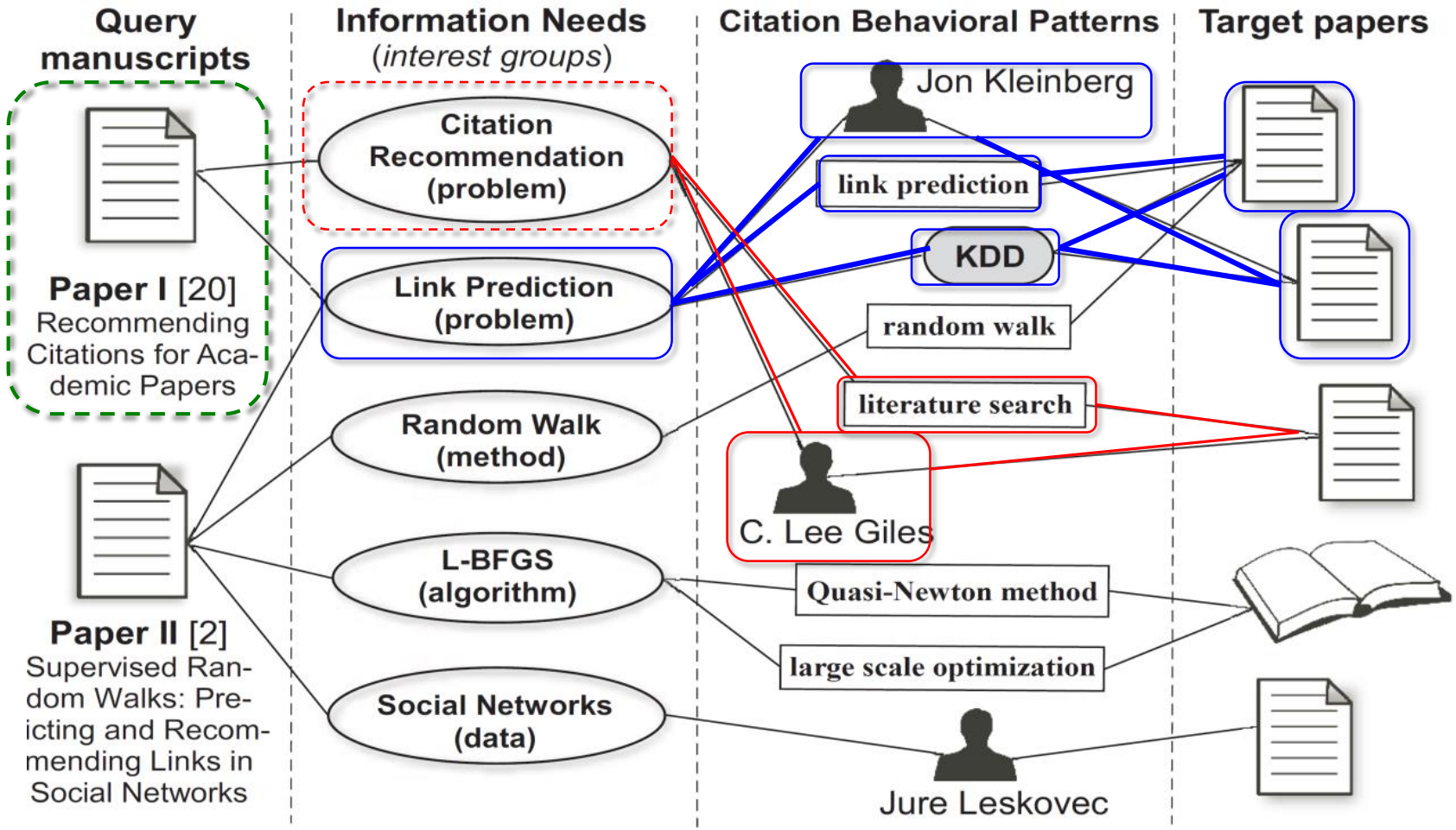
Are Google Scholar and PubMed satisfactory?

- X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, "ClusCite: Effective Citation Recommendation by Information Network-Based Clustering", KDD'14
- **Paper-specific recommendation model: heterogeneous network modeling**
  - Captures paper-paper relevance of different semantics
  - Enables authority propagation between different types of objects



Observation: Distinct Citation Behavioral Patterns

Each group follow distinct behavioral patterns and adopt different criteria in deciding relevance and authority of a candidate paper



The ClusCite Framework

We explore the **principle** that: citations tend to be ***softly*** clustered into different ***interest groups***, based on the heterogeneous network structures

Derive group membership for query manuscript

For *different* interest groups, learn *distinct* models on *finding* relevant papers and *judging* authority of papers

Phase I: Joint Learning (offline)

Paper-specific recommendation:

by integrating learned models of query's related interest groups

Phase II: Recommendation (online)

Performance Comparison on DBLP and PubMed

- 17.68% improvement in Recall@50; 9.57% in MRR, over the best performing compared method, on DBLP
- 20.19% improvement in Recall@20; 14.79% in MRR, over the best performing compared method, on PubMed

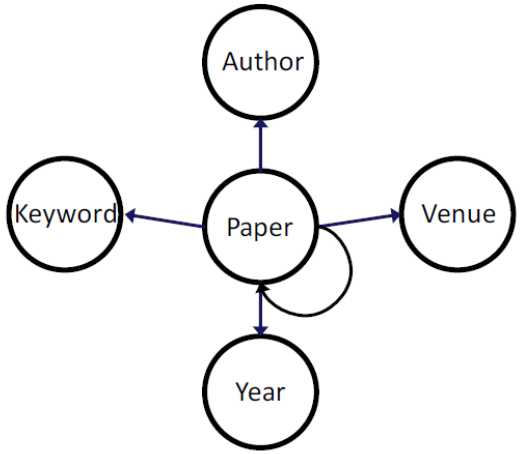
Method	DBLP					PubMed				
	P@10	P@20	R@20	R@50	MRR	P@10	P@20	R@20	R@50	MRR
BM25	0.1260	0.0902	0.1431	0.2146	0.4107	0.1847	0.1349	0.1754	0.2470	0.4971
PopRank	0.0112	0.0098	0.0155	0.0308	0.0451	0.0438	0.0314	0.0402	0.0814	0.2012
TopicSim	0.0328	0.0273	0.0432	0.0825	0.1161	0.0761	0.0685	0.0855	0.1516	0.3254
Link-PLSA-LDA	0.1023	0.0893	0.1295	0.1823	0.3748	0.1439	0.1002	0.1589	0.2015	0.4079
L2-LR	0.2274	0.1677	0.2471	0.3547	0.4866	0.2527	0.1959	0.2504	0.3981	0.5308
RankSVM	0.2372	0.1799	0.2733	0.3621	0.4989	0.2534	0.1954	0.2499	0.382	0.5187
MixFea	0.2261	0.1689	0.2473	0.3636	0.5002	0.2699	0.2025	0.2519	0.4021	0.5041
ClusCite-Rel	0.2402	0.1872	0.2856	0.4015	0.5156	0.2786	0.2221	0.2753	0.4305	0.5524
ClusCite	**0.2429**	**0.1958**	**0.2993**	**0.4279**	**0.5481**	0.3019	0.2434	0.3129	0.4587	0.5787

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - Task-Guided and Path-Augmented Heterogeneous Network Embedding
- Summary

Task-Guided and Path-Augmented Heterogeneous Network Embedding

- T. Chen and Y. Sun, Task-guided and Path-augmented Heterogeneous Network Embedding for Author Identification, WSDM'17
- Given an anonymized paper (often: double-blind review), with
  - Venue (e.g., WSDM)
  - Year (e.g., 2017)
  - Keywords (e.g., "heterogeneous network embedding")
  - References (e.g., [Chen et al., IJCAI'16])
- Can we predict its authors?
- Previous work on author identification: Feature engineering
- New approach: Heterogeneous Network Embedding
  - Embedding: automatically represent nodes into lower dimensional feature vectors
  - Heterogeneous network embedding: Key challenge—select the best type of info due to the heterogeneity of the network



Task-Guided and Path-Augmented Embedding

- Task-guided and path-augmented embedding: A Semi-Supervised framework
  - **Task guided embedding** vs. general network embedding: Task-guided embedding takes care of supervised labels
    - E.g., “Yizhou Sun” should be close to Keyword “Heterogeneous information network”
  - **Meta-path-based augmentation**: Path-augmented embedding takes care of the global structure of networks
    - E.g., Keyword “Heterogeneous network embedding” should be close to Keyword “node representation”
- The Combined Model
  - Joint training of two types of embedding
  - Path selection is performed to pick most informative meta-paths for network embedding

Task-Guided Embedding

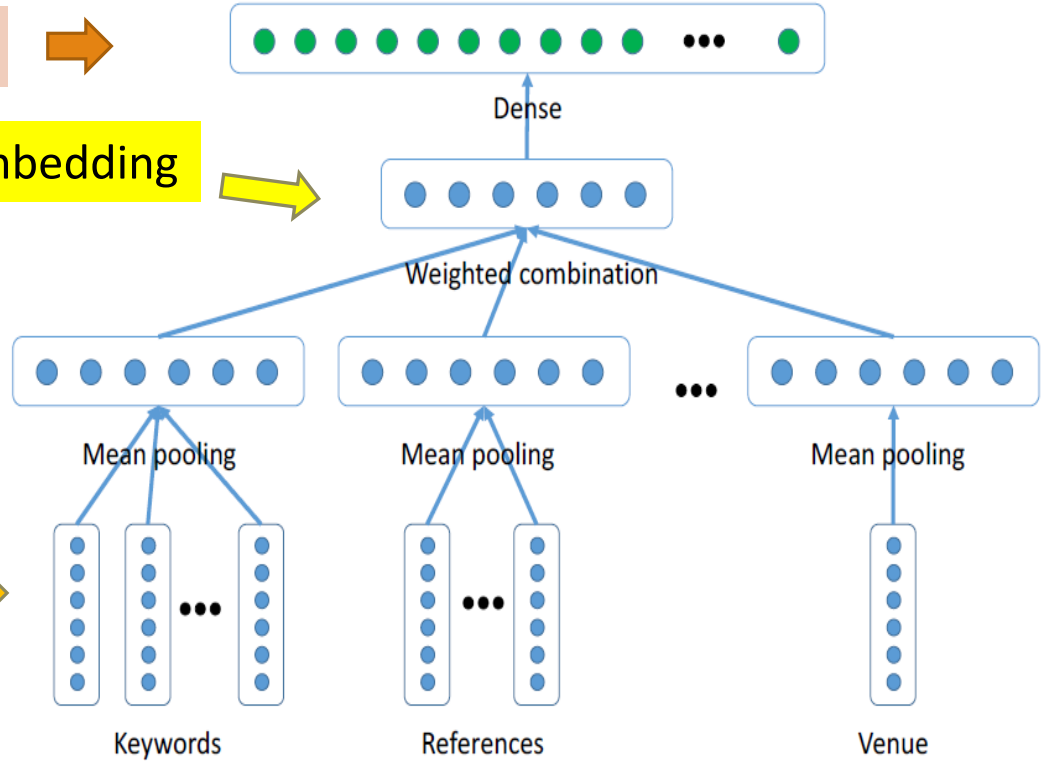
- Consider the ego-network of p:
  - $X_p = (X_p^1, X_p^2, \ldots, X_p^T)$,
    - T: # types of nodes associated with paper type
    - X_p^t: the set of nodes with type *t* associated with paper p
- u_a: embedding of author *a*
- u_n: embedding of node *n*
- V_p: embedding of paper *p*
  - *Weighted average of all the neighbors*
- The score function between *p* and *a* is:
- Ranking-based objective: maximize the difference between authors b and a:

Author score

Paper embedding

Node type embedding

Node embedding



The embedding architecture for author identification

$$f(p,a) = u_a^T V_p = u_a^T \left(\sum_t w_t V_p^{(t)} \right)$$

$$= u_a^T \left(\sum_t w_t \sum_{n \in X_p^{(t)}} u_n / |X_p^{(t)}| \right)$$

Soft hinge loss

$$\max\left(0, f(p,b) - f(p,a) + \xi\right)$$

Path-Augmented Embedding

- Why not just task-guided embedding?
  - Supervised labels expensive to obtain
  - The rich structured information of heterogeneous info-net is not fully explored
- Path-Augmented Embedding
  - Prepare meta-paths that are potentially related to the task
  - Apply general purpose embedding
- For each meta-path-based relation
  - The probability of reaching node *j*

    from node *i* via meta-path *r* via their embeddings

  - Use negative sampling to approximate the distribution

$$P(j|i;r) = \frac{\exp(u_i^T u_j)}{\sum_{j' \in DST(r)} \exp(u_i^T u_{j'})}$$

- Goal: maximize the likelihood to observing all the paths under each meta-path

The Joint Model and How to select meta-paths?

- The joint model
  - Objective function

$$\mathcal{L} = (1-\omega)\mathcal{L}_{task-specific} + \omega\mathcal{L}_{network-general} + \Omega(\mathcal{M})$$
$$= (1-\omega)\mathbb{E}_{(p,a,b)}\left[\max\left(0, f(p,b) - f(p,a) + \xi\right)\right]$$
$$+ \omega\mathbb{E}_{(r,i,j)}\left[-\log \hat{P}(j|i;r)\right] + \lambda\sum_i \|u_i\|_2^2$$

- How to select meta-paths?
  - A greedy strategy (so many ways to weigh meta-paths but may not be effective)
    - Step 1: Rank single meta-path according to their performance
    - Step 2: Greedily add the current best meta-path into current pool, stop until the performance deteriorates
  - Different meta-paths will be selected for different tasks

Identification of Anonymous Authors: Experiments

- Dataset:
  - AMiner Citation data set
  - Papers before 2012 are used in training, and papers on and after 2012 are used as test

Table 1 : Node statistics

	Paper	Author	keyword	Venue	Year
Train	1.6M	1M	4M	7K	60
Test	34K	62K	42K	1K	2

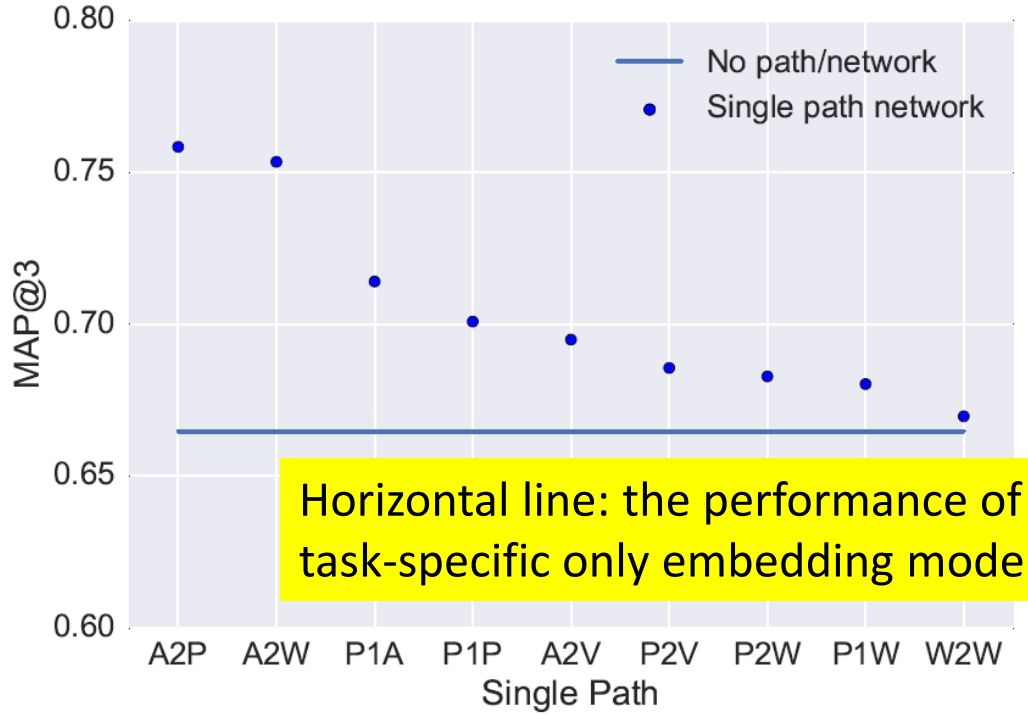
Table 3 : Length-2 link statistics

A-P-A	A-P-P	A-P-V	A-P-W	A-P-Y	P-P-V	P-P-W	V-P-W	W-P-W	Y-P-W
17M	18M	4M	38M	4M	3M	27M	12M	118M	12M

- Baselines
  - Supervised feature-based baselines (i.e. LR, SVM, RF, LambdaMart)
    - Manually crafted features
  - Task-specific embedding
  - Network-general embedding
  - Pre-training + Task-specific embedding
    - Take general embedding as initialization of task-specific embedding

Which Meta-Paths Are Selected?

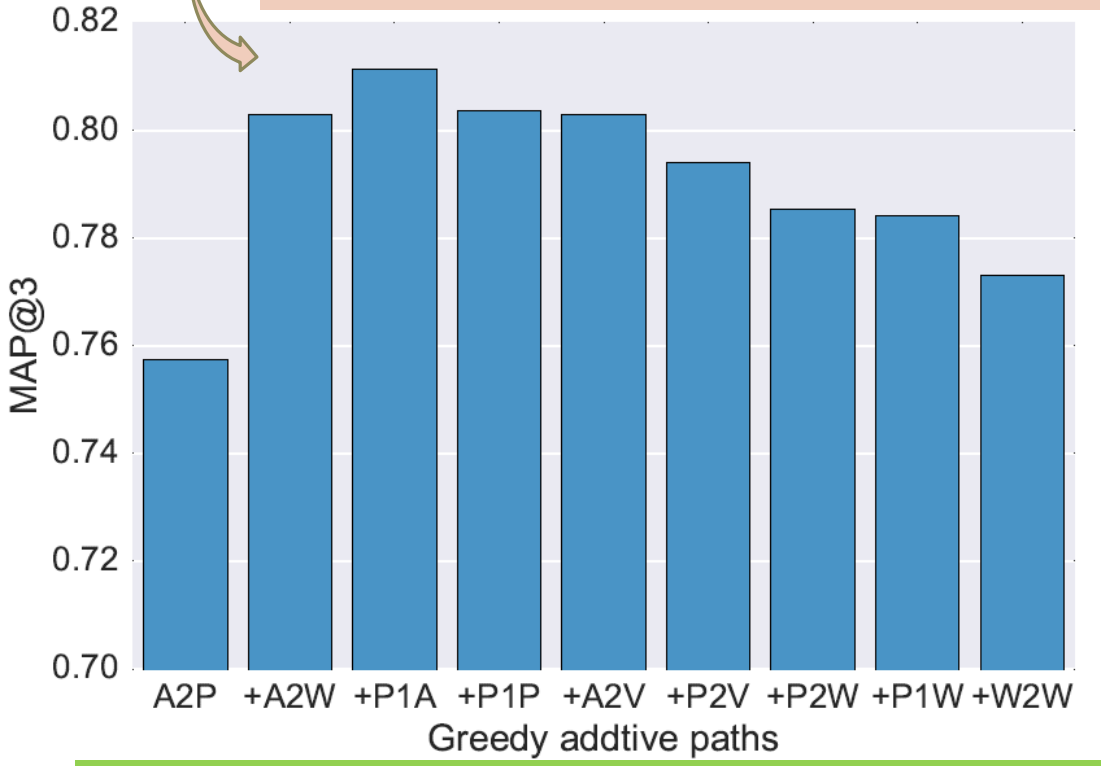
- A-P-P: author *write* paper *cite* paper
- A-P-W: author *write* paper *contain* keyword
- P-A: paper *written-by* author



Horizontal line: the performance of task-specific only embedding model

- Paths are sorted according to their performance
- Only paths that can help improve the author identification task are shown

- The first several paths are most relevant and helpful
- Latter ones can be harmful to use in network-general embedding



The performance of the combined model when meta-paths are added gradually

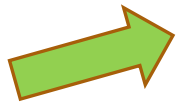
Author Identification: Performance Comparison

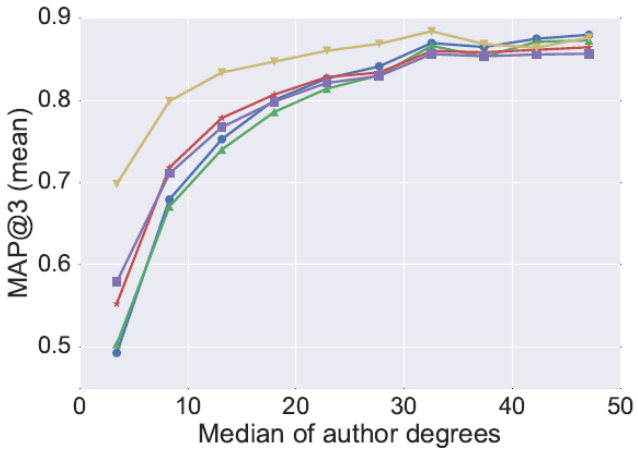
- Accuracy: choose author candidate as true authors + negative authors

Table 5 : Author identification performance comparison.

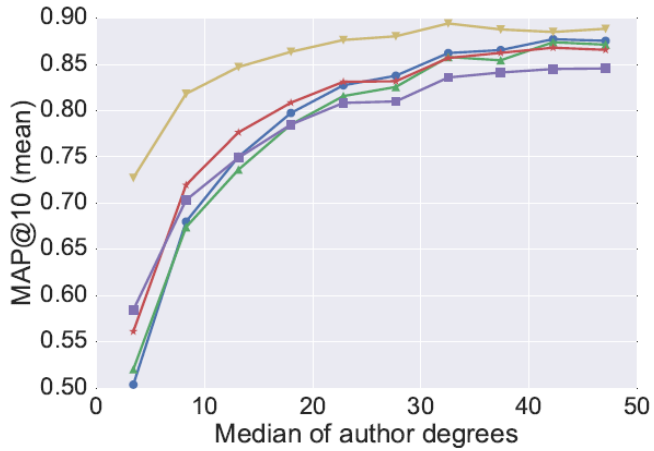
Models	MAP@3	MAP@10	Recall@3	Recall@10
LR	0.7289	0.7321	0.6721	0.8209
SVM	0.7332	0.7365	0.6748	0.8267
RF	0.7509	0.7543	0.6921	0.8381
LambdaMart	0.7511	0.7420	0.6869	0.8026
Task-specific	0.6876	0.7088	0.6523	0.8298
Pre-train+Task.	0.7722	0.7962	0.7234	0.9014
Network-general	0.7563	0.7817	0.7105	0.8903
Combined	**0.8113**	**0.8309**	**0.7548**	**0.9215**

- Performance over different groups of authors

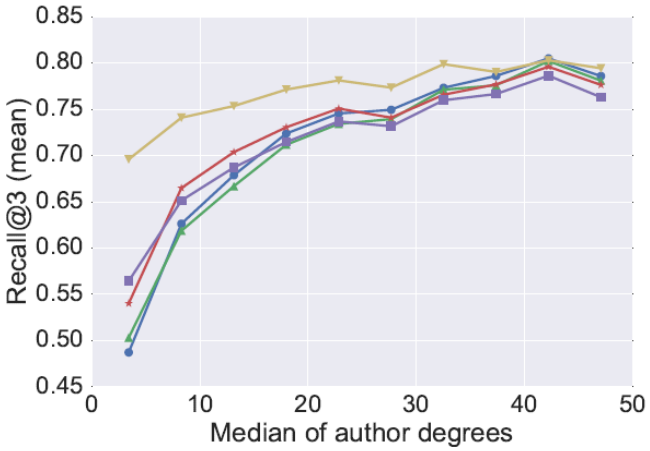




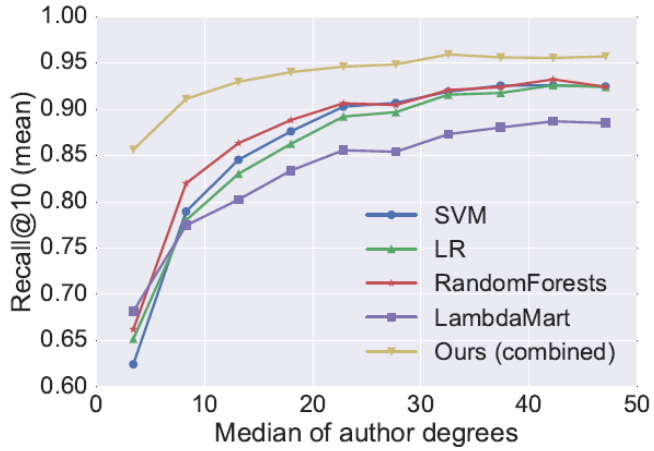
(c) MAP@3



(d) MAP@10

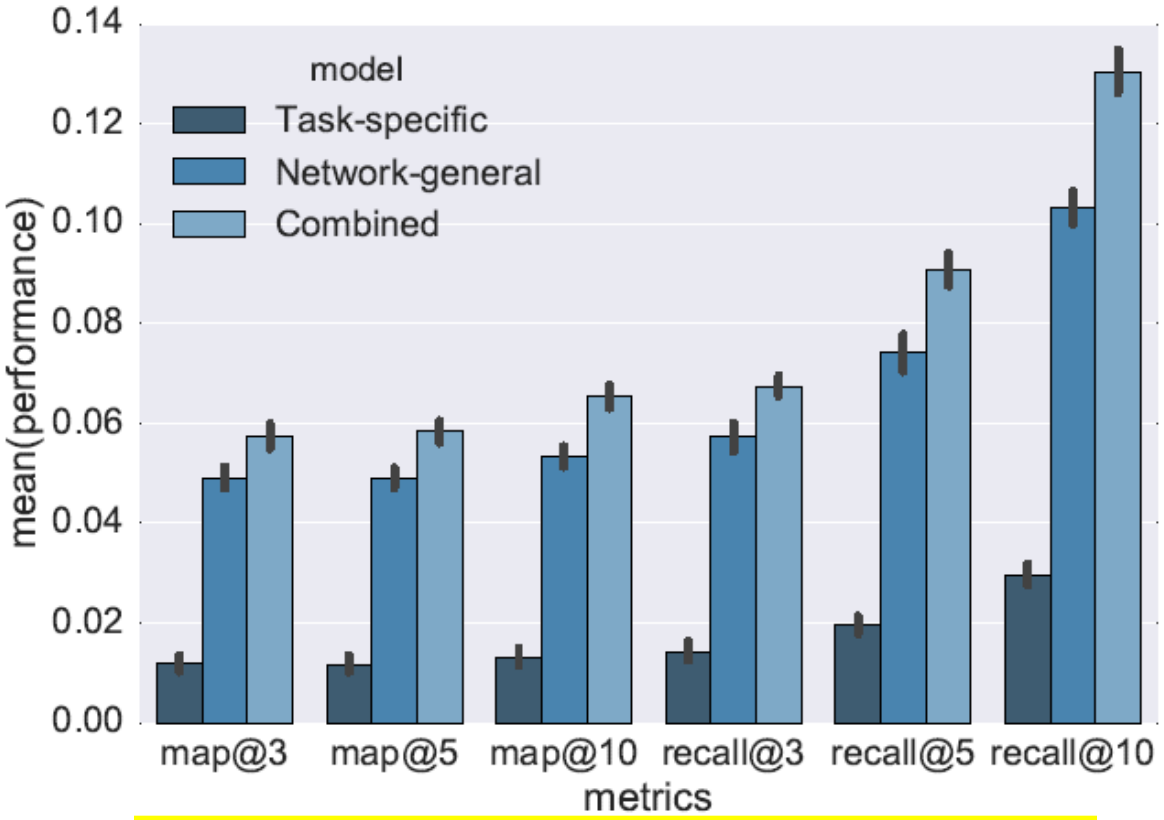


(e) Recall@3



(f) Recall@10

The Real Game and Case Study



Treat all the authors as candidates

Top ranked authors for queried paper

(a) "Active learning for networked data based on non-progressive diffusion model"

Ground-truth	Task-specific	Network-general	Combined
Z. Yang	L. Yu	J. Leskovec	**J. Tang**
J. Tang	Y. Gao	A. Ahmed	H. Liu
B. Xu	J. Wang	L. Getoor	Y. Guo
C. Xing	H. Liu	S.-D. Lin	X. Shi
	Y. Gao	D. Chakrabarti	W. Fan
	Z. Wang	P. Melville	B. Zhang
	Z. Zhang	T. Eliassi-Rad	S.-D. Lin
	J. Zhu	G. Lebanon	H. Zha
	Y. Ye	Y. Sun	L. H. Ungar
	R. Pan	L. H. Ungar	C. Wang

(b) "CatchSync: catching sync. behavior in large directed graphs"

Ground-truth	Task-specific	Network-general	Combined
M. Jiang	H. Wang	L. Akoglu	**C. Faloutsos**
P. Cui	H. Tong	T. Eliassi-Rad	A. Gionis
A. Beutel	**C. Faloutsos**	U. Kang	L. Akoglu
C. Faloutsos	D. Chakrabarti	H. Tong	J. Kleinberg
S. Yang	H. Yang	D. Chakrabarti	J. Leskovec
	G. Konidaris	A. Gionis	D. Chakrabarti
	I. Stanton	X. Yan	A. X. Zheng
	C. Wang	**C. Faloutsos**	T. Eliassi-Rad
	Y. Yang	J. Leskovec	U. Kang
	S. Kale	C. Tsourakakis	H. Tong

Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks
- Summary

Summary

- **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 structured heterogeneous info. networks**
  - Clustering, ranking, classification, path prediction, ……
- **Knowledge is power, but knowledge is hidden in massive, but "relatively structured" nodes and links!**
- **Key issue: Construction of trusted, semi-structured heterogeneous networks from unstructured data**
- **From data to knowledge: Much more to be explored but heterogeneous network mining has shown high promise!**

References

- M. Ji, M. Danilevsky, and J. Han, "Ranking-Based Classification of Heterogeneous Information Networks", KDD'11
- X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, "ClusCite: Effective Citation Recommendation by Information Network-Based Clustering", KDD'14
- Y. Sun and J. Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers, 2012
- 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
- Y. Sun, J. Han, C. C. Aggarwal, N. Chawla, "When Will It Happen? Relationship Prediction in Heterogeneous Information Networks", WSDM'12
- Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, T. Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", EDBT'09
- T. Chen and Y. Sun, Task-guided and Path-augmented Heterogeneous Network Embedding for Author Identification, WSDM'17
- X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han, "Personalized Entity Recommendation: A Heterogeneous Information Network Approach", WSDM'14