On the Power of Mining Heterogeneous Information Networks

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

†University of Illinois at Urbana-Champaign
§University of California at Santa Barbara
‡University of Illinois at Chicago

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

August 27, 2012
Motivation: Why Mining Information Networks?

Part I: Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks

Part II: Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks

Part III: Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering

Part IV: Advanced Topics on Information Network Analysis

Conclusions
What Are Information Networks?

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

They are all treated as Homogeneous Networks!

Social Networking Websites

Biological Network: Protein Interaction

Research Collaboration Network

Product Recommendation Network via Emails
Homogeneous Information Networks

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

  - Ranking web pages [Brin and Page, 1998]
  - Clustering books about politics [Newman, 2006]
  - Link Prediction [Kleinberg, 2003]
Heterogeneous Information Networks

- Multiple object types and/or multiple link types

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

Directly Mining information richer heterogeneous networks
Heterogeneous Networks Are Ubiquitous

- Healthcare
  - Doctor, patient, disease, treatment

- Content sharing websites
  - Video, image, user, comment

- E-Commerce
  - Seller, buyer, product, review

- News
  - Person, organization, location, text
### What Can be Mined from Heterogeneous Networks?

- **DBLP**: A Computer Science bibliographic database

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

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

- **Principle 1**: Use Holistic Network Information
  - Study information propagation across different types of objects and links

- **Principle 2**: Explore Network Meta Structure
  - Meta-path-based similarity search and mining

- **Principle 3**: User-Guided Exploration
  - Relation strength-aware mining with user guidance
Motivation: Why Mining Information Networks?

Part I: Clustering, Ranking and Classification
- Clustering and Ranking in Information Networks
- Classification of Information Networks

Part II: Meta-Path-Based Exploration of Information Networks
- Similarity Search in Information Networks
- Relationship Prediction in Information Networks

Part III: Relation Strength-Aware Mining
- Relation Strength-Aware Clustering of Networks with Incomplete Attributes
- Integrating Meta-Path Selection with User-Guided Clustering

Part IV: Advanced Topics on Information Network Analysis

Conclusions
Ranking and Clustering: Two Critical Functions

- Ranking
  - SIGMOD
  - ICDE
  - ASPLOS
  - DAC
  - CASES
  - ISC
  - DASFAA
  - ADBIS

- Clustering
  - SIGMOD
  - ICDE
  - ASPLOS
  - DAC
  - CASES
  - ISC
  - DASFAA
  - ADBIS

Comparing apples and oranges?

A better solution: Integrating clustering with ranking

Not distinguishing objects in each cluster?

- Database Conferences: SIGMOD, ASPLOS, ICDE, DAC, CASES, ISC, DASFAA, ADBIS
- Hardware and Architecture Conferences: SIGMOD, ASPLOS, ICDE, DAC, CASES, ISC, DASFAA, ADBIS
RankClus: Integrating Clustering with Ranking

[Sun et al., EDBT’09]

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

- Better clustering => Conditional ranking distributions are more distinguishing from each other
  - Conditional ranking distribution serves as the feature of each cluster
    - $P(\cdot | \text{area} = \text{"database"})$ vs. $P(\cdot | \text{area} = \text{"hardware"})$

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

$\left( P(\text{area} = \text{"database"}|\text{SIGMOD}), P(\text{area} = \text{"hardware"}|\text{SIGMOD}) \right)$
**Simple Ranking vs. Authority Ranking**

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

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

What about an author publishing 100 papers in low reputation conferences?
### Rules for Authority Ranking

- **Rule 1:** Highly ranked authors publish *many* papers in highly ranked conferences
  
  \[ \tilde{r}_Y(j) = \sum_{i=1}^{m} W_{YX}(j, i) \tilde{r}_X(i) \]

- **Rule 2:** Highly ranked conferences attract *many* papers from *many* highly ranked authors
  
  \[ \tilde{r}_X(i) = \sum_{j=1}^{n} W_{XY}(i, j) \tilde{r}_Y(j) \]

- **Rule 3:** The rank of an author is enhanced if he or she co-authors with *many* highly ranked authors
  
  \[ \tilde{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i, j) \tilde{r}_X(j) + (1 - \alpha) \sum_{j=1}^{n} W_{YY}(i, j) \tilde{r}_Y(j) \]
Generating New Measure Space

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

- **Step 0: Initialization**
  - Randomly partition

- **Step 1: Ranking**
  - Ranking objects in each sub-network induced from each cluster

- **Step 2: Generating new measure space**
  - Estimate *mixture model coefficients* for each target object

- **Step 3: Adjusting cluster**

- **Step 4: Repeating Steps 1-3 until stable**
Step-by-Step Running Case Illustration

Initially, ranking distributions are mixed together

Improved a little

Improved significantly

Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable
Clustering and Ranking CS Conferences by RankClus

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

Top-10 conferences in 5 clusters using RankClus in DBLP

RankClus outperforms spectral clustering [Shi and Malik, 2000] algorithms on projected homogeneous networks
Beyond bi-typed information network
- A Star Network Schema [richer information]
- Split a network into different layers
  - Each representing by a network cluster
Multi-Typed Networks Lead to Better Results

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

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

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

...
Interesting Results from Other Domains

RankCompete: Organize images automatically!

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

Rank treatments for AIDS from MEDLINE
Outline

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

M. Ji, M. Danilevski, et al., “Graph Regularized Transductive Classification on Heterogeneous Information Networks”, ECMLPKDD'10
Minimize the objective function

$$J(f_1^{(k)}, \ldots, f_m^{(k)})$$

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

$$+ \sum_{i=1}^{m} \alpha_i (f_i^{(k)} - y_i^{(k)})^T (f_i^{(k)} - y_i^{(k)})$$

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

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

Confidence estimation on labeled data and their pre-given labels should be similar
From RankClus to GNetMine & RankClass

- **RankClus [EDBT’09]:** Clustering and ranking working together
  - No training, no available class labels, no expert knowledge

- **GNetMine [PKDD’10]:** Incorp. prior knowledge in networks
  - Classification in heterog. networks, but objects treated equally

- **RankClass [KDD’11]:** Integration of ranking and classification in heterogeneous network analysis
  - Ranking: informative understanding & summary of each class
  - Class membership is critical information when ranking objects
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class
Experiments on DBLP

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>((a%, p%)) of authors and papers labeled</th>
<th>nLB (A-C-P-T)</th>
<th>wvRN (A-C-P-T)</th>
<th>LLGC (A-C-P-T)</th>
<th>GNetMine (A-C-P-T)</th>
<th>RankClass (A-C-P-T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.1%), (0.1%)</td>
<td>25.5</td>
<td>43.5</td>
<td>79.0</td>
<td>81.0</td>
<td>84.5</td>
</tr>
<tr>
<td>(0.2%), (0.2%)</td>
<td>22.5</td>
<td>56.0</td>
<td>83.5</td>
<td>85.0</td>
<td>85.5</td>
</tr>
<tr>
<td>(0.3%), (0.3%)</td>
<td>25.0</td>
<td>59.0</td>
<td>87.0</td>
<td>87.0</td>
<td>87.0</td>
</tr>
<tr>
<td>(0.4%), (0.4%)</td>
<td>25.0</td>
<td>57.0</td>
<td>86.5</td>
<td>89.5</td>
<td>90.5</td>
</tr>
<tr>
<td>(0.5%), (0.5%)</td>
<td>25.0</td>
<td>68.0</td>
<td>90.0</td>
<td>94.0</td>
<td>95.0</td>
</tr>
<tr>
<td>average</td>
<td>24.6</td>
<td>56.7</td>
<td>85.2</td>
<td>87.3</td>
<td>88.5</td>
</tr>
</tbody>
</table>
Experiments with Very Small Training Set

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

<table>
<thead>
<tr>
<th>Top-5 ranked conferences</th>
<th>Database</th>
<th>Data Mining</th>
<th>AI</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VLDB</td>
<td>KDD</td>
<td>IJCAI</td>
<td>SIGIR</td>
</tr>
<tr>
<td></td>
<td>SIGMOD</td>
<td>SDM</td>
<td>AAAI</td>
<td>ECIR</td>
</tr>
<tr>
<td></td>
<td>ICDE</td>
<td>ICDM</td>
<td>ICML</td>
<td>CIKM</td>
</tr>
<tr>
<td></td>
<td>PODS</td>
<td>PKDD</td>
<td>CVPR</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>EDBT</td>
<td>PAKDD</td>
<td>ECML</td>
<td>WSDM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top-5 ranked terms</th>
<th>data</th>
<th>mining</th>
<th>learning</th>
<th>retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>database</td>
<td>data</td>
<td>knowledge</td>
<td>information</td>
</tr>
<tr>
<td></td>
<td>query</td>
<td>clustering</td>
<td>reasoning</td>
<td>web</td>
</tr>
<tr>
<td></td>
<td>system</td>
<td>classification</td>
<td>logic</td>
<td>search</td>
</tr>
<tr>
<td></td>
<td>xml</td>
<td>frequent</td>
<td>cognition</td>
<td>text</td>
</tr>
</tbody>
</table>
Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**
Similarity Search: Find Similar Objects in Networks [Sun et al., VLDB’11]

- DBLP
  - Who are the most similar to “Christos Faloutsos”?

- IMDB
  - Which movies are the most similar to “Little Miss Sunshine”?

- E-Commerce
  - Which products are the most similar to “Kindle”?

How to systematically answer these questions in heterogeneous information networks?
Existing Link-based Similarity Functions

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

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

Objects are connected together via different types of relationships!

“Jim-P1-Ann”
“Mike-P2-Ann”
“Mike-P3-Bob”

Author-Paper-Author

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

Author-Paper-Venue-Paper-Author

- Network schema
  - Meta-level description of a network

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

- Who are most similar to Christos Faloutsos?

Meta-Path: *Author-Paper-Author*

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

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

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

Christos’s students or close collaborators
Work on similar topics and have similar reputation
Some Meta-Path Is “Better” Than Others

- Which pictures are most similar to?

Evaluate the similarity between images according to their linked tags

Evaluate the similarity between images according to tags and groups

Meta-Path: *Image-Tag-Image*

Meta-Path: *Image-Tag-Image-Group-Image-Tag-Image*
**PathSim: Similarity in Terms of “Peers”**

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

- **In addition to meta-path**
  - Need to consider **similarity measures**
Limitations of Existing Similarity Measures

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

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

- PathSim
  - Normalized path count between x and y following meta-path \( \mathcal{P} \)
    \[
    s(x, y) = \frac{2 \times |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightarrow x} : p_{x \rightarrow x} \in \mathcal{P}\}| + |\{p_{y \rightarrow y} : p_{y \rightarrow y} \in \mathcal{P}\}|}
    \]
  - Favor “peers”:
    - objects with strong connectivity and similar visibility under the given meta-path
  - Calculation
    - For \( \mathcal{P}: A_1 - A_2 - \cdots - A_l - A_{l-1} - \cdots - A_1 \)
      - \( M = W_{A_1A_2} W_{A_2A_3} \cdots W_{A_{l-1}A_l} W_{A_lA_{l-1}} \cdots W_{A_3A_2} W_{A_2A_1} \)
      - \( s(x, y) = \frac{2M_{xy}}{M_{xx} + M_{yy}} \)
    - A co-clustering based pruning algorithm is provided
      - 18.23% - 68.04% efficiency improvement over the baseline
Properties of PathSim

- Symmetric
  - \[ s(x, y) = s(y, x) \]

- Self-Maximum
  - \[ s(x, y) \in [0,1], \text{and} \ s(x, x) = 1 \]

- Balance of visibility
  - \[ s(x, y) \leq \frac{2}{\sqrt{M_{xx}/M_{yy}} + \sqrt{M_{yy}/M_{xx}}} \]
    - \( M_{xx} \) is the number of path instances starting from \( x \) and ending with \( x \) following the given meta path

- Limiting behavior
  - If repeating a pattern of meta path infinite times, PathSim degenerates to authority ranking comparison

---

Long meta-path without introducing new relationships is not that helpful!
Find Academic Peers by PathSim

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

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

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

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

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

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

41
Meta-Path: A Key Concept for Mining Heterogeneous Networks

- **Search and Query System**
  - PathSim [Sun et al., VLDB’11]
  - User-guided similarity search [Yu et al., CIKM’12]

- **Relationship Prediction**
  - PathPredict [Sun et al., ASONAM’11]
    - Co-authorship prediction using meta-path-based similarity
  - PathPredict_when [Sun et al., WSDM’12]
    - When a relationship will happen
  - Citation prediction [Yu et al., SDM’12]
    - Meta-path + topic

- **User-Guided Clustering**
  - PathSelClus [Sun et al., KDD’12]
    - Meta-path selection + clustering

- **Recommendation System**
  - Ongoing work
Outline

- **Motivation:** Why Mining Information Networks?

- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks

- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks

- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering

- **Part IV:** Advanced Topics on Information Network Analysis

- Conclusions
Meta-Path-Based Relationship Prediction

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

[Image of a bookshelf with a question mark]
Relationship Prediction vs. Link Prediction

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

- Relationship prediction in heterogeneous networks
  - **Target**: Different types of relationships need different prediction models
  - **Features**: Different connection paths need to be treated separately!
    - **Meta-path-based approach** to define topological features.
PathPredict: Meta-Path Based Co-authorship Prediction in DBLP [Sun et al., ASONAM’11]

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

<table>
<thead>
<tr>
<th>Meta-Path</th>
<th>Semantic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → P → P − A</td>
<td>$a_i$ cites $a_j$</td>
</tr>
<tr>
<td>A → P ← P − A</td>
<td>$a_i$ is cited by $a_j$</td>
</tr>
<tr>
<td>A → P − V − P − A</td>
<td>$a_i$ and $a_j$ publish in the same venues</td>
</tr>
<tr>
<td>A → P − A − P − A</td>
<td>$a_i$ and $a_j$ are co-authors of the same authors</td>
</tr>
<tr>
<td>A → P − T − P − A</td>
<td>$a_i$ and $a_j$ write the same topics</td>
</tr>
<tr>
<td>A → P → P → P − A</td>
<td>$a_i$ cites papers that cite $a_j$</td>
</tr>
<tr>
<td>A → P ← P ← P − A</td>
<td>$a_i$ is cited by papers that are cited by $a_j$</td>
</tr>
<tr>
<td>A → P ← P ← P − A</td>
<td>$a_i$ and $a_j$ cite the same papers</td>
</tr>
<tr>
<td>A → P ← P → P − A</td>
<td>$a_i$ and $a_j$ are cited by the same papers</td>
</tr>
</tbody>
</table>
The Power of PathPredict

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

Social relations play very important role?

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])
From “whether” to “when”

- “Whether”: Will Jim rent the movie “Avatar” in Netflix?
  
- “When”: When will Jim rent the movie “Avatar”?

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”?

- $E(Y)$
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)

**Generalized Linear Model under Weibull Distribution Assumption**

\[
egin{align*}
\log L &= \sum_{i=1}^{n} \left( f_Y(y_i|\theta_i, \lambda)I_{\{y_i<T\}} + P(y_i \geq T|\theta_i, \lambda)I_{\{y_i \geq T\}} \right) \\
LL_W(\beta, \lambda) &= \sum_{i=1}^{n} I_{\{y_i<T\}} \log \left( \frac{\lambda y_i^{\lambda-1}}{e^{-\lambda x_i \beta}} \right) - \sum_{i=1}^{n} \left( \frac{y_i}{e^{-x_i \beta}} \right)^{\lambda}
\end{align*}
\]

**Training Framework**

- **Right Censoring**
- **T**: Right Censoring

49
Author Citation Time Prediction in DBLP

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

\[A \rightarrow P \rightarrow T \rightarrow P \rightarrow A\]
\[A \rightarrow P \leftarrow P \rightarrow P \rightarrow A\]
\[A \rightarrow P \rightarrow A \rightarrow P \rightarrow P \rightarrow A\]
\[A \rightarrow P \rightarrow T \rightarrow P \rightarrow A \rightarrow P \rightarrow P \rightarrow A\]

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

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

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

Under Weibull distribution assumption
Motivation: Why Mining Information Networks?

Part I: Clustering, Ranking and Classification
- Clustering and Ranking in Information Networks
- Classification of Information Networks

Part II: Meta-Path-Based Exploration of Information Networks
- Similarity Search in Information Networks
- Relationship Prediction in Information Networks

Part III: Relation Strength-Aware Mining
- Relation Strength-Aware Clustering of Networks with Incomplete Attributes
- Integrating Meta-Path Selection with User-Guided Clustering

Part IV: Advanced Topics on Information Network Analysis

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

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

<table>
<thead>
<tr>
<th>Age</th>
<th>Salary</th>
<th>Interests</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10K</td>
<td>Sports, Music</td>
<td>Champaign, Boston</td>
</tr>
<tr>
<td>22</td>
<td>50K</td>
<td>Movie, Music, Football</td>
<td>New York</td>
</tr>
<tr>
<td>50</td>
<td>150K</td>
<td>Shopping, Books</td>
<td>Chicago</td>
</tr>
<tr>
<td>52</td>
<td>120K</td>
<td>Painting, Music</td>
<td>Boston</td>
</tr>
<tr>
<td>25</td>
<td>100K</td>
<td>Cooking, Books</td>
<td>Chicago, Seattle</td>
</tr>
</tbody>
</table>

### Customer Segmentation According to Customer Profiles

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>5</td>
</tr>
<tr>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
</tr>
</tbody>
</table>

### Weather Pattern Clustering According to Weather Sensor Records
## Incomplete Attributes

<table>
<thead>
<tr>
<th>Age</th>
<th>Salary</th>
<th>Interests</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10K</td>
<td>Sports, Music</td>
<td>Champaign, Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>50</td>
<td>N/A</td>
<td>Shopping, Books</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>120K</td>
<td>N/A</td>
<td>Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>100K</td>
<td>Cooking, Books</td>
<td>Chicago, Seattle</td>
</tr>
</tbody>
</table>

### Customer Segmentation According to Customer Profiles

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>N/A</td>
<td>15</td>
</tr>
<tr>
<td>N/A</td>
<td>20</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
</tr>
<tr>
<td>85</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Weather Pattern Clustering According to Weather Sensor Records

- **Precip. Sensor Type**
  - N/A
  - 15
  - 20
  - N/A
  - N/A

- **Temp. Sensor Type**
  - N/A
  - 5
  - 15
  - 20
  - N/A
  - N/A
The Links Help!

<table>
<thead>
<tr>
<th>Age</th>
<th>Salary</th>
<th>Interests</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10K</td>
<td>Sports, Music</td>
<td>Champaign, Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>50</td>
<td>N/A</td>
<td>Shopping, Books</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>120K</td>
<td>N/A</td>
<td>Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>100K</td>
<td>Cooking, Books</td>
<td>Chicago, Seattle</td>
</tr>
</tbody>
</table>

Customer Segmentation According to Customer Profiles

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>N/A</td>
<td>15</td>
</tr>
<tr>
<td>N/A</td>
<td>20</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
</tr>
<tr>
<td>85</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Weather Pattern Clustering According to Weather Sensor Records
Example 1: Bibliographic Information Network

Link type:
- Paper-Author, Paper-Venue, (Paper->Paper)

Attribute type:
- Text attribute for Paper type

Goal:
- Clustering authors, venues, papers into different research areas
Example 2: Weather Sensor Information Network

Link type:
- T->P, T->T, P->P, P->T (According to KNN relationships)

Attribute type:
- Temperature attribute for T-typed sensors, Precipitation attribute for P-typed sensors

Goal:
- Clustering both types of sensors into different regional weather patterns
Challenges

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

- Modeling attribute generation and structural consistency in a unified framework

\[
p({\{v[X]\}}_{v \in V_X} \mid x \in X, \Theta \mid G, \gamma, \beta) = \prod_{x \in X} p({\{v[X]\}}_{v \in V_X} \mid \Theta, \beta) p(\Theta \mid G, \gamma)
\]

- Attribute generation as a mixture model

\[
p({\{v[X]\}}_{v \in V_X} \mid \Theta, \beta) = \prod_{v \in V_X} \prod_{x \in v[X]} \sum_{k=1}^{K} \theta_{v,k} p(x \mid \beta_k)
\]

- \(v[X]\): observed values for Attribute \(X\) on Object \(v\)
- \(\Theta\): soft clustering membership matrix
- \(\beta\): parameters associated with each mixture model component

- Structural consistency as a log-linear model

\[
p(\Theta \mid G, \gamma) = \frac{1}{Z(\gamma)} \exp\left\{ \sum_{e=\langle v_i, v_j \rangle \in E} f(\theta_i, \theta_j, e, \gamma) \right\}
\]

- \(\gamma\): relation strength vector
The Objective Function and the Algorithm

Overview

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

\[
g(\Theta, \beta, \gamma) = \log \sum_{X \in \mathcal{X}} p(\{v[X]\} \mid \Theta, \beta) + \log p(\Theta \mid G, \gamma) + \frac{||\gamma||^2}{2\sigma^2}
\]

- Attribute Generation
- Structural Consistency
- Regularization Term
Higher Accuracy and More Stable Clustering Results

Clustering Accuracy Comparisons for AC

Clustering Accuracy Comparisons for Weather Sensor Network
Intuitive relation strength weights

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

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering, Ranking and Classification
  - Clustering and Ranking in Information Networks
  - Classification of Information Networks
- **Part II:** Meta-Path-Based Exploration of Information Networks
  - Similarity Search in Information Networks
  - Relationship Prediction in Information Networks
- **Part III:** Relation Strength-Aware Mining
  - Relation Strength-Aware Clustering of Networks with Incomplete Attributes
  - Integrating Meta-Path Selection with User-Guided Clustering
- **Part IV:** Advanced Topics on Information Network Analysis
- **Conclusions**
Why Meta-Path Selection? [Sun et al., KDD’12]

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

Which meta-path to choose?

{1,2,3,4} {5,6,7,8} {1,3,5,7} {2,4,6,8} {1,3} {2,4} {5,7} {6,8}
The Role of User Guidance

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

Seeds | Meta-path(s) | Clustering Result
---|---|---
{1} {5} |  | {1,2,3,4} {5,6,7,8}
{1} {2} {5} {6} | (a) AOA | {1,3} {2,4} {5,7} {6,8}
{1} {2} {5} {6} | (c) AOA + AVA |
The Problem of User-Guided Clustering with Meta-Path Selection

- **Input:**
  - The target type for clustering: $T$
  - Number of clusters: $K$
    - Seeds in *some* of the clusters: $L_1, L_2, \ldots, L_K$
  - $M$ Candidate meta-paths starting from $T$: $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_M$

- **Output:**
  - The *quality weight* for each candidate meta-path in the clustering process
    - $\alpha_m$
  - The *clustering results* that are consistent with the user guidance
    - $\theta_i$
Existing Link-based User-Guided Clustering Approaches

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

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

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

Objective Function

\[
J = \sum_i \left( \sum_m \log P(\pi_{i,m}|\alpha_m w_{i,m}, \theta_i, B_m) + \sum_k 1_{\{t_i \in  \mathcal{L}_k\}} \lambda \log \theta_{ik} \right)
\]
Part 1: Modeling the Relationship Generation

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

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

E.g., $P(\mathcal{A}_0 | \Theta)$

(a) AOA

(b) AVA
For each soft clustering probability vector $\theta_i$:

- Model it as generated from a Dirichlet prior
  - If $t_i$ is labeled as a seed in Cluster $k^*$
    - $\theta_i \sim Dir(\lambda e_{k^*} + \mathbf{1})$
    - $e_{k^*}$ is an all-zero vector except for item $k^*$, which is 1
    - $\lambda$ is the user confidence for the guidance
  - If $t_i$ is not labeled in any cluster
    - $\theta_i \sim Dir(\mathbf{1})$
    - The prior density is uniform, a special case of Dirichlet distribution

$$p(\theta_i|\lambda) = \begin{cases} 
\prod_k \theta_{ik}^{1_{\{t_i \in \mathcal{L}_k\}} \lambda} = \theta_{ik^*}^\lambda, & \text{if } t_i \text{ is labeled and } t_i \in \mathcal{L}_{k^*}, \\
1, & \text{if } t_i \text{ is not labeled.}
\end{cases}$$
Part 3: Modeling the Quality Weights for Meta-Paths

- Model quality weight $\alpha_m$ as the relative weight for each relationship in $\mathcal{W}_m$
  - Observation of relationships: $\mathcal{W}_m \rightarrow \alpha_m \mathcal{W}_m$
- Further assume relationship generation with Dirichlet Prior: $\pi_{i,m} \sim \text{Dir}(1)$
- The best $\alpha_m$: the most likely to generate current clustering-based parameters
  $$\alpha^*_m = \arg \max_{\alpha_m} \prod_i P(\pi_{i,m} | \alpha_m, w_{i,m}, \theta_i, B_m)$$

  - when $\alpha_m$ is small, $\pi_{i,m}$ is more likely to be a uniform distribution
    - Random generated
  - when $\alpha_m$ is large, $\pi_{i,m}$ is more likely to be $\frac{w_{i,m}}{n_{i,m}}$, what we observed
    - Consistent with the observation
The Learning Algorithm

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

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

- Step 2: Optimize $\alpha$ given $\Theta$
  - In general, the higher likelihood of observing $W_m$ given $\Theta$, the higher $\alpha_m$

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

- Datasets
  - DBLP
    - Object Types: Authors, Venues, Papers, Terms
    - Relation Types: AP, PA, VP, PV, TP, PT
  - Yelp
    - Object Types: Users, Businesses, Reviews, Terms
    - Relation Types: UR, RU, BR, RB, TR, RT
DBLP-T1: Clustering Venues According to Research Areas

- **Task:**
  - Target objects: venues
  - Number of clusters: 4;
  - Candidate meta-paths: $V-P-A-P-V$, $V-P-T-P-V$

- **Output:**
  - **Weights:**
    - $V-P-A-P-V$: 1576 (0.0017 per relationship)
    - $V-P-T-P-V$: 17001 (0.0003 per relationship)
  - **Clustering results:**

<table>
<thead>
<tr>
<th>#S</th>
<th>Measure</th>
<th>PathSelClus</th>
<th>LP</th>
<th>ITC</th>
<th>LP_voting</th>
<th>LP_soft</th>
<th>ITC_voting</th>
<th>ITC_soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accuracy</td>
<td>0.9950</td>
<td>0.650</td>
<td>0.690</td>
<td>0.650</td>
<td>0.6650</td>
<td>0.6450</td>
<td>0.5100</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.9906</td>
<td>0.618</td>
<td>0.698</td>
<td>0.618</td>
<td>0.5801</td>
<td>0.5903</td>
<td>0.5316</td>
</tr>
<tr>
<td>2</td>
<td>Accuracy</td>
<td>1</td>
<td>0.750</td>
<td>0.845</td>
<td>0.750</td>
<td>0.8200</td>
<td>0.8950</td>
<td>0.8700</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>1</td>
<td>0.673</td>
<td>0.775</td>
<td>0.673</td>
<td>0.7492</td>
<td>0.8321</td>
<td>0.7942</td>
</tr>
</tbody>
</table>
Yelp-T2: Clustering Restaurants According to Categories

- **Task:**
  - Target objects: restaurants
  - Number of clusters: 6;
  - Candidate meta-paths: \( B-R-U-R-B, B-R-T-R-B \).

- **Output:**
  - **Weights:**
    - \( B-R-U-R-B : 6000 \) \((0.1716 \text{ per relationship, compared with } 0.5864 \text{ for clustering shopping categories})\)
    - \( B-R-T-R-B: 2.9522\times 10^7 \) \((0.0138 \text{ per relationship})\)
Motivation: Why Mining Information Networks?

Part I: Clustering, Ranking and Classification
- Clustering and Ranking in Information Networks
- Classification of Information Networks

Part II: Meta-Path-Based Exploration of Information Networks
- Similarity Search in Information Networks
- Relationship Prediction in Information Networks

Part III: Relation Strength-Aware Mining
- Relation Strength-Aware Clustering of Networks with Incomplete Attributes
- Integrating Meta-Path Selection with User-Guided Clustering

Part IV: Advanced Topics on Information Network Analysis

Conclusions
1. Role Discovery in Network: Why It Matters?

Army communication network (imaginary)

Automatically infer

Commander

Captain

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

- Input: DBLP research publication network
- Output: Potential advising relationship and its ranking \((r, [\text{st, ed}])\)
- Ref. C. Wang, J. Han, et al., “Mining Advisor-Advisee Relationships from Research Publication Networks”, SIGKDD 2010
2. Graph/Network Summarization: Graph Compression

- Extract common subgraphs and simplify graphs by condensing these subgraphs into nodes.
OLAP on Information Networks [Chen, ICDM’08]

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

- **Group by “Gender”**

<table>
<thead>
<tr>
<th>Gender</th>
<th>COUNT(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
</tr>
</tbody>
</table>

- **Group by “Gender” and “Location”**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Location</th>
<th>COUNT(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>CA</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>CA</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>WA</td>
<td>2</td>
</tr>
<tr>
<td>Male</td>
<td>IL</td>
<td>3</td>
</tr>
<tr>
<td>Male</td>
<td>NY</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>NY</td>
<td>1</td>
</tr>
</tbody>
</table>
OLAP on Graph Cube [Zhao et al., SIGMOD’ 11]

- Cuboid query
  - Return as output the aggregate network corresponding to a specific multidimensional space (cuboid)
    - What is the aggregate network between various genders?
    - What is the aggregate network between various gender and location combinations?
Many networks are with time information
- E.g., according to paper publication year, DBLP networks can form network sequences

Motivation: Model evolution of communities in heterogeneous network
- Automatically detect the best number of communities in each timestamp
- Model the smoothness between communities of adjacent timestamps
- Model the evolution structure explicitly
  - Birth, death, split
Case Study on DBLP

- Tracking database and information system community evolution

<table>
<thead>
<tr>
<th>98-99</th>
<th>00-01</th>
<th>02-03</th>
<th>04-05</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECOOP</td>
<td>H. Garcia-Molina</td>
<td>ICSE</td>
<td>Elisa Bertino</td>
</tr>
<tr>
<td>DEXA</td>
<td>Elisa Bertino</td>
<td>VLDB</td>
<td>Jiawei Han</td>
</tr>
<tr>
<td>ICDE</td>
<td>Robert Fox</td>
<td>ICDE</td>
<td>B. C. Ooi</td>
</tr>
<tr>
<td>ICSE</td>
<td>Philip S. Yu</td>
<td>SIGMOD</td>
<td>Robert Fox</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>H-P Kriegel</td>
<td>DEXA</td>
<td>H-P Kriegel</td>
</tr>
<tr>
<td>VLDB</td>
<td>Matthias Jarke</td>
<td>HICSS</td>
<td>Kian-Lee Tan</td>
</tr>
<tr>
<td>Info. &amp; soft</td>
<td>system software object data oriented distributed design database web information</td>
<td>Info. &amp; Soft</td>
<td>software system object oriented verification design distributed</td>
</tr>
</tbody>
</table>

- 1998-2001:
  - Learning knowledge information reasoning data retrieval decision logic
  - Didier Dubois
  - J. Y. Halpern
  - Henri Prade
  - TREC
  - ICML
  - TREC
  - SIGIR
  - COLING
  - ACL
  - UAI
  - UCAI
  - KDD
  - Daphne Koller

- 2002-2003:
  - ICSE
  - S. L. P. Jones
  - COMPSAC
  - T. A. Henzinger
  - M. Piattini
  - ICSM
  - E. M. Clarke
  - SEKE
  - Mark Harman
  - Software system oriented verification checking engineering

- 2004-2005:
  - SigSoft
  - M. Piattini
  - ICSE
  - Mark Harman
  - ICSE
  - Baowen Xu
  - APSEC
  - M. C. Rinard
  - ASE
  - S. Ducasse
  - SEKE
  - T. A. Henzinger
  - time design object uml

- 2006-2007:
  - VLDB
  - Hongjun Lu
  - SIGIR
  - Philip S. Yu
  - TREC
  - Jiawei Han
  - ICDE
  - Divesh Srivastava
  - ICDM
  - Kian-Lee Tan
  - CIKM
  - Didier Dubois
  - COLING
  - Agrawal
  - TKDE
  - data web mining information retrieval learning xml query clustering model system database semantic search
  - SIGMOD
  - R. Ramakrishnan
  - CIKM
  - KDD
  - data web mining information retrieval learning xml query clustering model database search classification

- 2008-2009:
  - SIGMOD
  - R. Ramakrishnan
  - TREC
  - Barry Smyth
  - ICDE
  - H-P Kriegel
  - ICDM
  - C. Faloutsos
  - TKDE
  - R. Ramakrishnan
  - KDD
  - data web mining information retrieval learning xml query clustering model database search classification
Case Study on Delicious.com

### Tags
- Security
- Terrorism
- Politics
- Travel
- USA
- Airport
- Israel
- Obama
- CIA
- Afghanistan

### Website

### Users

### Delicious Schema

<table>
<thead>
<tr>
<th>C1:</th>
<th>C2:</th>
<th>C3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>Google</td>
<td>Security</td>
</tr>
<tr>
<td>Terrorism</td>
<td>China</td>
<td>China</td>
</tr>
<tr>
<td>Politics</td>
<td>Internet</td>
<td>Internet</td>
</tr>
<tr>
<td>Travel</td>
<td>Privacy</td>
<td>Microsoft</td>
</tr>
<tr>
<td>USA</td>
<td>Politics</td>
<td>Privacy</td>
</tr>
<tr>
<td>Airport</td>
<td>Censorship</td>
<td>Censorship</td>
</tr>
<tr>
<td>Israel</td>
<td>Facebook</td>
<td>Politics</td>
</tr>
<tr>
<td>Obama</td>
<td>Business</td>
<td>Browser</td>
</tr>
<tr>
<td>CIA</td>
<td>Terrorism</td>
<td>USA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C1:</th>
<th>C2:</th>
<th>C3:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mac</td>
<td>Iphone</td>
<td>Iphone</td>
</tr>
<tr>
<td>Apple</td>
<td>Apple</td>
<td>Apple</td>
</tr>
<tr>
<td>Iphone</td>
<td>Twitter</td>
<td>Mac</td>
</tr>
<tr>
<td>Windows</td>
<td>Mobile</td>
<td>Mobile</td>
</tr>
<tr>
<td>Tablet</td>
<td>Twitter</td>
<td>Software</td>
</tr>
<tr>
<td>Ipad</td>
<td>Apps</td>
<td>Apps</td>
</tr>
<tr>
<td>Apple</td>
<td>Blog</td>
<td>Business</td>
</tr>
<tr>
<td>Iphone</td>
<td>Newspapers</td>
<td>Osx</td>
</tr>
<tr>
<td>Macbook</td>
<td>Technology</td>
<td>Radio</td>
</tr>
<tr>
<td>Tutorial</td>
<td>Technology</td>
<td>Media</td>
</tr>
<tr>
<td>Drm</td>
<td>Technology</td>
<td>Media</td>
</tr>
</tbody>
</table>

### Event Count Graph

- **C1**: Tags
- **C2**: Websites
- **C3**: Users

### Event Count

<table>
<thead>
<tr>
<th>Week</th>
<th>Event Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Additional Categories
- Health
- Depression
- Sleep
- Teenagers
- Dubai
- Tallest
- BBC
- Building
- Architecture
- Mentalhealth
- Weather
- UK
- Photography
- Photo
- Haiti
- Photos
- 2010
- BBC
- Snow
- Earthquake
- Haiti
- Photography
- BBC
- Earthquake
- Photos
- UK
- 2010
- Disaster
- Travel
- Wildlife
Outline

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

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

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

References: Some Overview Papers

- L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. ICML'01
References: Some Influential Papers

- S. Brin and L. Page. The anatomy of a large-scale hyper-textual web search engine. WWW'98.
- G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. KDD'02
- D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD'03
- J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. COCOON'99
- J. M. Kleinberg. Small world phenomena and the dynamics of information. NIPS'01
- R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. FOCS'00
References: Clustering and Ranking (1)

- E. Airoldi, D. Blei, S. Fienberg and E. Xing, “Mixed Membership Stochastic Blockmodels”, JMLR’08
- Liangliang Cao, Andrey Del Pozo, Xin Jin, Jiebo Luo, Jiawei Han, and Thomas S. Huang, “RankCompete: Simultaneous Ranking and Clustering of Web Photos”, WWW’10
- G. Jeh and J. Widom, “SimRank: a measure of structural-context similarity”, KDD'02
- Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han, “Community Outliers and their Efficient Detection in Information Networks”, KDD'10
- J. Shi and J. Malik, “Normalized cuts and image Segmentation”, CVPR'97
- Yizhou Sun, Yintao Yu, and Jiawei Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema", KDD’09
- Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", EDBT’09
References: Clustering and Ranking (2)

- Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu, "iTopicModel: Information Network-Integrated Topic Modeling", ICDM’09
- Yizhou Sun, Charu C. Aggarwal, and Jiawei Han, "Relation Strength-Aware Clustering of Heterogeneous Information Networks with Incomplete Attributes", PVLDB 5(5), 2002
- A. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. KDD'04
- Xiaoxin Yin, Jiawei Han, Philip S. Yu. "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", VLDB'06.
- Yintao Yu, Cindy X. Lin, Yizhou Sun, Chen Chen, Jiawei Han, Binbin Liao, Tianyi Wu, ChengXiang Zhai, Duo Zhang, and Bo Zhao, "iNextCube: Information Network-Enhanced Text Cube", VLDB'09 (demo)
- X. Yin, J. Han, and P. S. Yu. Cross-relational clustering with user's guidance. KDD'05
References: Network Classification (1)

- Jing Gao, Feng Liang, Wei Fan, Yizhou Sun, and Jiawei Han, "Bipartite Graph-based Consensus Maximization among Supervised and Unsupervised Models ", NIPS'09
- M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, “Graph-based classification on heterogeneous information networks”, ECMLPKDD’10.
- M. Ji, J. Jan, and M. Danilevsky, “Ranking-based Classification of Heterogeneous Information Networks”, KDD’11.
- Q. Lu and L. Getoor, “Link-based classification”, ICML'03
- D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks”, CIKM'03
References: Network Classification (2)

- J. Neville, D. Jensen, L. Friedland, and M. Hay. Learning relational probability trees. KDD’03
- Jennifer Neville, David Jensen, “Relational Dependency Networks”, JMLR’07
- M. J. Rattigan, M. Maier, and D. Jensen. Graph clustering with network structure indices. ICML’07
- B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. IJCAI’01
- X. Yin, J. Han, J. Yang, and P. S. Yu, “CrossMine: Efficient Classification across Multiple Database Relations”, ICDE'04.

R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. WWW'03

P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. WWW'04

D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. PKDD'05


P. Domingos and M. Richardson. Mining the network value of customers. KDD'01


J. Kubica, A. Moore, and J. Schneider. Tractable group detection on large link data sets. ICDM'03
References: Data Quality & Search in Networks

- I. Bhattacharya and L. Getoor, “Iterative record linkage for cleaning and integration”, Proc. SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'04)
- Y. Sun, J. Han, T. Wu, X. Yan, and Philip S. Yu, “PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks”, VLDB’11.
- X. Yin, J. Han, and P. S. Yu, “Object Distinction: Distinguishing Objects with Identical Names by Link Analysis”, ICDE'07.
- X. Yin, J. Han, and P. S. Yu, “Truth Discovery with Multiple Conflicting Information Providers on the Web”, IEEE TKDE, 20(6):796-808, 2008
- P. Zhao and J. Han, “On Graph Query Optimization in Large Networks”, VLDB’10.
References: Link and Relationship Prediction

- D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks”, CIKM ’03,
- Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal and Jiawei Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks“, ASONAM’11.
- Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla, "When Will It Happen? --- Relationship Prediction in Heterogeneous Information Networks", WSDM’12.
- B. Taskar, M. fai Wong, P. Abbeel, and D. Koller, “Link prediction in relational data”, NIPS ’03.
- Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han, "Citation Prediction in Heterogeneous Bibliographic Networks”, SDM’12.
References: Role Discovery, Summarization and OLAP

- Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu, "Graph OLAP: Towards Online Analytical Processing on Graphs", ICDM 2008
- Chen Chen, Xifeng Yan, Feida Zhu, Jiawei Han, and Philip S. Yu, "Graph OLAP: A Multi-Dimensional Framework for Graph Data Analysis", KAIS 2009.
- Xin Jin, Jiebo Luo, Jie Yu, Gang Wang, Dhiraj Joshi, and Jiawei Han, “iRIN: Image Retrieval in Image Rich Information Networks”, WWW'10 (demo paper)
- Lu Liu, Feida Zhu, Chen Chen, Xifeng Yan, Jiawei Han, Philip Yu, and Shiqiang Yang, “Mining Diversity on Networks", DASFAA'10
- Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. SIGMOD'08
- Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo, “Mining Advisor-Advisee Relationships from Research Publication Networks", KDD'10
- Zhijun Yin, Manish Gupta, Tim Weninger and Jiawei Han, “LINKREC: A Unified Framework for Link Recommendation with User Attributes and Graph Structure ”, WWW’10
- Peixiang Zhao, Xiaolei Li, Dong Xin, Jiawei Han. Graph Cube: On Warehousing and OLAP Multidimensional Networks, SIGMOD’11
References: Network Evolution

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