Tutorial on Mining Heterogeneous Information Networks

Rokia Missaoui
LARIM
Université du Québec en Outaouais, Canada
http://w3.uqo.ca/missaoui

Acknowledgement

I am grateful to Professor Jiawei Han who kindly allowed me to use his slides on mining heterogeneous information networks. This presentation is a slight adaptation of his presentation material.
(see http://www.cs.uiuc.edu/~hanj/ and the following key reference for more details)

Yizhou Sun, Jiawei Han: Mining Heterogeneous Information Networks: Principles and Methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers 2012
Outline

- Introduction
  - Network representation and kinds of networks
  - Why mining heterogeneous information networks (HINs)?
- Research work of Han’s team on mining HINs
  - Combining ranking with clustering
  - Combining ranking with classification
  - Meta-path based exploration of information networks
  - Role discovery and evolution analysis
- Other contributions
- Our current research on HINs
- Conclusion
- References
Introduction

- **Social networks**
  - A social structure of nodes (e.g., individuals or organizations) that are related to each other by various ties such as friendship, affinity, collaboration, ...

- **Typical social networks**
  - Social bookmarking (Del.icio.us)
  - Friendship networks (Facebook, Myspace)
  - Professional networks (LinkedIn)
  - Media Sharing (Flickr, Youtube)
  - Folksonomy: collaborative tagging using three entities: users, resources and tags

Information network analysis (Sun & Han, 2012)

- Database as an information network: entities and relationships
- Focus on heterogeneous information networks since they contain rich and inter-related semantics
- Data mining (DM) techniques: clustering, classification, ranking, similarity search, link prediction, trends and evolution analysis
- Construction of semantically rich networks by exploring links among node types through DM techniques
- A lot of topics that still need to be explored

Main reference

Yizhou Sun, Jiawei Han: Mining Heterogeneous Information Networks: Principles and Methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers 2012
Network Representations

- A network/graph: \( G = (V, E) \), where \( V \): vertices/nodes, \( E \): edges/links

- Adjacency matrix:
  \[
  A = \begin{pmatrix}
  0 & 1 & 0 & 0 & 1 & 0 \\
  1 & 0 & 1 & 1 & 0 & 0 \\
  0 & 1 & 0 & 1 & 1 & 1 \\
  0 & 1 & 1 & 0 & 0 & 0 \\
  1 & 0 & 1 & 0 & 0 & 0 \\
  0 & 0 & 1 & 0 & 0 & 0 
  \end{pmatrix}
  \]

- Weighted graph:
  - Edges having weight (strength), usually a real number

- Directed network (directed graph): if each edge has a direction

- Labeled graph:
  - Edges have a label (e.g., creation date)

Information Network (IN)

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

- Homogeneous networks
  - Single object type and single link type (one-mode data)
  - Web: a collection of linked Web pages

- Heterogeneous or multi-typed networks
  - Multiple object and link types
  - Medical network: patients, doctors, diseases, treatments
  - Bibliographic network: publications, authors, venues
Information Network

- Information network (Sun & Han, 2012)

Definition 1.1  (Information network) An information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\tau : V \rightarrow A$ and a link type mapping function $\phi : E \rightarrow R$, where each object $v \in V$ belongs to one particular object type $\tau(v) \in A$, each link $e \in E$ belongs to a particular relation $\phi(e) \in R$, and if two links belong to the same relation type, the two links share the same starting object type as well as the ending object type.

- Heterogeneous information network
  - When the number of object types or link types is more than 1

Homogeneous vs. Heterogeneous Networks

Co-author Network  Conference-Author Network
Mining Information Networks

- **Homogeneous** networks can often be derived from their original **heterogeneous** networks
  - E.g., coauthor networks can be derived from author-paper-conference networks by *projection* on authors only
  - Paper citation networks can be derived from a complete bibliographic network with papers and citations projected
- Heterogeneous INs (HINs) carry **richer information** than their corresponding projected homogeneous networks
- Typed HINs vs. non-typed HINs (i.e., not distinguishing different types of nodes)
  - Typed nodes and links imply a more structured IN, and thus often lead to more **informative discovery**

Why Mining INs?

- Information networks are everywhere!
  - Biological networks
  - Bibliographic networks: DBLP, ArXiv, PubMed, ...
  - Social networks: Facebook >100 million active users
  - World Wide Web (WWW): > 3 billion nodes, > 50 billion edges
  - Cyber-physical networks
What Can be Mined from Heterogeneous Networks?

- DBLP: A Computer Science bibliographic database
  - >1.8 M papers, >0.7 M authors, >10K venues, >70K terms (appearing more than once)

A sample publication record in DBLP

<table>
<thead>
<tr>
<th>DBLP Network</th>
<th>Mining Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>How are CS research areas structured?</td>
<td>Clustering</td>
</tr>
<tr>
<td>Who are the leading researchers on Web search?</td>
<td>Ranking</td>
</tr>
<tr>
<td>Who are the peer researchers of Jure Leskovec?</td>
<td>Similarity Search</td>
</tr>
<tr>
<td>Whom will Christos Faloutsos collaborate with in the future?</td>
<td>Relationship Prediction</td>
</tr>
<tr>
<td>Whether will an author publish a paper in KDD, and when?</td>
<td>Relationship Prediction with Time</td>
</tr>
<tr>
<td>Which types of relationships are most influential for an author to decide her topics?</td>
<td>Relation Strength Learning</td>
</tr>
</tbody>
</table>

Bipartite Graphs

- $G = (V_1 \cup V_2, E \subseteq V_1 \times V_2)$
- Incidence matrix where cell $A_{ij} = 1$ if there exists a link between $i$ and $j$, 0 otherwise
- One can convert a two-mode network data into two one-mode data, but with information loss
  - $A \times A^T$ gives the number of nodes in $V_2$ co-linked by both the row and the column in $V_1$. E.g., two authors have papers in both AAAI and ICML
  - $A^T \times A$ gives the number of nodes in $V_1$ which are linked to both the row and the column in $V_2$. E.g., Jack and Tracy have papers in one conference (SDM)
Clustering and Ranking: Two Critical Functions

- Clustering

- Ranking

Comparing apples and oranges?

A better solution: Integrating clustering with ranking

RankClus: Integrating Clustering with Ranking

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

Simple solution: Project the bi-typed network into homogeneous conference network?

Information-loss projection!
A New Methodology: RankClus

- 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!

Simple Ranking vs. Authority Ranking

- Simple Ranking
  - Proportional to # of publications of an author / a conference
  - Considers only immediate neighborhood in the network
  - What about an author publishing 100 papers in very weak conferences?
- Authority Ranking:
  - More sophisticated “rank rules” are needed
  - Propagate the ranking scores in the network over different types
Rules for Authority Ranking

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

- **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
  \[ \tilde{r}_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i, j) \tilde{r}_X(j) + (1 - \alpha) \sum_{j=1}^{n} W_{XY}(i, j) \tilde{r}_Y(j) \]

---

RankClus: Algorithm Framework

- **Initialization**
  - Randomly partition

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

- Until stable
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

RankClus: Clustering & Ranking CS Conferences

<table>
<thead>
<tr>
<th>DB</th>
<th>Network</th>
<th>AT</th>
<th>Theory</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>INFCOM</td>
<td>AAMAS</td>
<td>SODA</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
<td>SIGMETRICS</td>
<td>LICAI</td>
<td>STOC</td>
</tr>
<tr>
<td>3</td>
<td>SIGMOD</td>
<td>ICNP</td>
<td>AAAI</td>
<td>FOCSC</td>
</tr>
<tr>
<td>4</td>
<td>KDD</td>
<td>SIGCOMM</td>
<td>Agents</td>
<td>ICALP</td>
</tr>
<tr>
<td>5</td>
<td>ICDM</td>
<td>MObICOM</td>
<td>AAAI/AAAI</td>
<td>CCC</td>
</tr>
<tr>
<td>6</td>
<td>EDFT</td>
<td>ICDCS</td>
<td>ECAI</td>
<td>SPAA</td>
</tr>
<tr>
<td>7</td>
<td>DASFAA</td>
<td>NETWORKING</td>
<td>RoboCup</td>
<td>PODC</td>
</tr>
<tr>
<td>8</td>
<td>PODS</td>
<td>MobiHoc</td>
<td>IAT</td>
<td>CRYPTO</td>
</tr>
<tr>
<td>9</td>
<td>SDBM</td>
<td>ECC</td>
<td>ICMAS</td>
<td>APPROX-GRAYCOM</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td>CP</td>
<td>EUROCRYPT</td>
</tr>
</tbody>
</table>

Top-10 conferences in 5 clusters using RankClus in DBLP (when k = 15)

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

Tutorial - Mining Heterogeneous Information Networks  EGC’2013 - Toulouse
Time Complexity: Linear to # of Links

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

NetClus [KDD’09]: Beyond Bi-Typed Networks

- Beyond bi-typed information network
  - A Star Network Schema [richer information]
- Split a network into different layers
  - Each represented 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

<table>
<thead>
<tr>
<th>Conference</th>
<th>Rank Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGMOD</td>
<td>0.315</td>
</tr>
<tr>
<td>VLDB</td>
<td>0.306</td>
</tr>
<tr>
<td>ICDE</td>
<td>0.194</td>
</tr>
<tr>
<td>PODS</td>
<td>0.109</td>
</tr>
<tr>
<td>EDBT</td>
<td>0.046</td>
</tr>
<tr>
<td>CIKM</td>
<td>0.019</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author</th>
<th>Rank Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Stonebraker</td>
<td>0.0063</td>
</tr>
<tr>
<td>Surajit Chaudhuri</td>
<td>0.0057</td>
</tr>
<tr>
<td>C. Mohan</td>
<td>0.0053</td>
</tr>
<tr>
<td>Michael J. Carey</td>
<td>0.0052</td>
</tr>
<tr>
<td>David J. DeWitt</td>
<td>0.0051</td>
</tr>
<tr>
<td>H. V. Jagadish</td>
<td>0.0043</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Rank Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>database</td>
<td>0.0529</td>
</tr>
<tr>
<td>system</td>
<td>0.0322</td>
</tr>
<tr>
<td>query</td>
<td>0.0313</td>
</tr>
<tr>
<td>data</td>
<td>0.0251</td>
</tr>
<tr>
<td>object</td>
<td>0.0138</td>
</tr>
<tr>
<td>management</td>
<td>0.0113</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

NetClus: Database System Cluster

database 0.099551
databases 0.070881
system 0.0678563
data 0.0214893
query 0.0133316
systems 0.0110413
queries 0.0090603
management 0.00850744
object 0.00837766
relational 0.0081175
processing 0.00745875
based 0.00736599
distributed 0.0068367
xml 0.00664958
oriented 0.00595597
model 0.00595597
web 0.00595597
information 0.0050518
efficient 0.00465707

<table>
<thead>
<tr>
<th>Author</th>
<th>Rank Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serge Abiteboul</td>
<td>0.0472111</td>
</tr>
<tr>
<td>Victor Vianu</td>
<td>0.0348510</td>
</tr>
<tr>
<td>Jerome Simeon</td>
<td>0.0324529</td>
</tr>
<tr>
<td>Michael J. Carey</td>
<td>0.0288872</td>
</tr>
<tr>
<td>Sophie Cheet</td>
<td>0.0282911</td>
</tr>
<tr>
<td>Daniela Foressu</td>
<td>0.0241411</td>
</tr>
<tr>
<td>Silvia Amer-Yehia</td>
<td>0.0240869</td>
</tr>
<tr>
<td>Donald Kosmann</td>
<td>0.0232118</td>
</tr>
<tr>
<td>Wenfai Fan</td>
<td>0.0225235</td>
</tr>
<tr>
<td>Tora Milo</td>
<td>0.0202201</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Ranking authors in XML
Interesting Results from Other Domains

RankCompete: Organize your photo album automatically!

<table>
<thead>
<tr>
<th>Rank</th>
<th>Treatment</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zidovudine/therapeutic use</td>
<td>0.1079</td>
</tr>
<tr>
<td>2</td>
<td>Anti-HIV Agents/therapeutic use</td>
<td>0.1340</td>
</tr>
<tr>
<td>3</td>
<td>Anti-retroviral Therapy, Highly Active</td>
<td>0.0977</td>
</tr>
<tr>
<td>4</td>
<td>Antiviral Agents/therapeutic use</td>
<td>0.0718</td>
</tr>
<tr>
<td>5</td>
<td>Anti-Retroviral Agents/therapeutic use</td>
<td>0.0536</td>
</tr>
<tr>
<td>6</td>
<td>Interferon Type I/therapeutic use</td>
<td>0.0417</td>
</tr>
<tr>
<td>7</td>
<td>Didanosine/therapeutic use</td>
<td>0.0132</td>
</tr>
<tr>
<td>8</td>
<td>Ganciclovir/therapeutic use</td>
<td>0.0114</td>
</tr>
<tr>
<td>9</td>
<td>HIV Protease Inhibitors/therapeutic use</td>
<td>0.0095</td>
</tr>
<tr>
<td>10</td>
<td>Antineoplastic Combined Chemotherapy</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Rank treatments for AIDS from MEDLINE

From RankClus to RankClass

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

- **RankClass [KDD’11]: Integration of ranking and classification**
  - Ranking: informative understanding & summary of each class
  - Class membership is critical information when ranking objects
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class
Classification Generates Good Ranking Results

- 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
- List objects with the highest confidence measure belonging to conf. & terms

<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>VLDB</td>
<td>KDD</td>
<td>IJCAI</td>
<td>SIGIR</td>
<td></td>
</tr>
<tr>
<td>SIGMOD</td>
<td>SDM</td>
<td>AAAI</td>
<td>ECIR</td>
<td></td>
</tr>
<tr>
<td>ICDE</td>
<td>ICDM</td>
<td>ICML</td>
<td>CIKM</td>
<td></td>
</tr>
<tr>
<td>PODS</td>
<td>PKDD</td>
<td>CVPR</td>
<td>WWW</td>
<td></td>
</tr>
<tr>
<td>EDBT</td>
<td>PAKDD</td>
<td>ECML</td>
<td>WSDM</td>
<td></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>database</td>
<td>data</td>
<td>knowledge</td>
<td>information</td>
<td></td>
</tr>
<tr>
<td>query</td>
<td>clustering</td>
<td>reasoning</td>
<td>web</td>
<td></td>
</tr>
<tr>
<td>system</td>
<td>classification</td>
<td>logic</td>
<td>search</td>
<td></td>
</tr>
<tr>
<td>xml</td>
<td>frequent</td>
<td>cognition</td>
<td>text</td>
<td></td>
</tr>
</tbody>
</table>

Similarity Search: Find Similar Objects in Networks

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

**How to systematically answer these questions?**

**Study similarity search in heterogeneous networks**

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

- Network schema
  - Meta-level description of a network

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

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

**Path instances**

**Meta-path**

**Co-authorship**

**Relation:** Describes the type of relationships

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>1.00</td>
</tr>
<tr>
<td>2</td>
<td>Spiros Papadimitriou</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>Jimeng Sun</td>
<td>0.73</td>
</tr>
<tr>
<td>4</td>
<td>Jia-Yu Pan</td>
<td>0.50</td>
</tr>
<tr>
<td>5</td>
<td>Agus J. M. Traina</td>
<td>0.40</td>
</tr>
<tr>
<td>6</td>
<td>Jure Leskovec</td>
<td>0.38</td>
</tr>
<tr>
<td>7</td>
<td>Csatoni Traina Jr.</td>
<td>0.37</td>
</tr>
<tr>
<td>8</td>
<td>Hanghang Tong</td>
<td>0.37</td>
</tr>
<tr>
<td>9</td>
<td>Deepayan Chakrabarti</td>
<td>0.36</td>
</tr>
<tr>
<td>10</td>
<td>Flip Korn</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Christos’s students or close collaborators

**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>1.00</td>
</tr>
<tr>
<td>2</td>
<td>Jia-Yu Pan</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>Jimeng Sun</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>Jure Leskovec</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>Agus J. M. Traina</td>
<td>0.11</td>
</tr>
<tr>
<td>6</td>
<td>Csatoni Traina Jr.</td>
<td>0.11</td>
</tr>
<tr>
<td>7</td>
<td>Hanghang Tong</td>
<td>0.11</td>
</tr>
<tr>
<td>8</td>
<td>Deepayan Chakrabarti</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Work on similar topics and have similar reputation
One Meta-Path Is “Better” Than Others

- Which pictures are most similar to?

![Diagram showing meta-paths: Image-Tag-Image and Image-Tag-Image-Group-Image-Tag-Image.]

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-6

PathSim: Similarity in Terms of “Peers”

- Why peers?
  - Strongly connected, while similar visibility

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

![Diagram showing different e-readers: B&N Nook, Amazon Kindle, Sony Reader, Kobo eReader.]

Tutorial - Mining Heterogeneous Information Networks  EGC’2013 - Toulouse
Existing Similarity Measures

- Random walk
  - The probability of random walk starting from \( x \) and ending with \( y \) following meta-path \( \mathcal{P} \): 
    \[
    s(x, y) = \sum_{p \in \mathcal{P}} \text{Prob}(p)
    \]
  - Used in Personalized PageRank (P-PageRank) [Jeh and Widom, 2003]

- Pairwise random walk
  - The probability of pairwise random walk starting from \( (x, y) \) and ending with a common object following meta-path \( \mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2) \): 
    \[
    s(x, y) = \sum_{(p_1, p_2) \in (\mathcal{P}_1, \mathcal{P}_2)} \text{Prob}(p_1)\text{Prob}(p_2^{-1})
    \]
  - Used in SimRank [Jeh and Widom, 2002]

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

Only PathSim Can Find Peers

- Limitations of Existing Measures
  - Random walk (RW): Favor highly visible objects
    - objects with large degrees
  - Pairwise random walk (PRW): Favor “pure” objects
    - objects with highly skewed distribution in their in-links or out-links

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

\[
\begin{align*}
\text{PathSim} & = \frac{2 \times |\{p_{x\rightarrow y} : p_{x\rightarrow y} \in \mathcal{P}\}|}{|\{p_{x\rightarrow x} : p_{x\rightarrow x} \in \mathcal{P}\}| + |\{p_{y\rightarrow y} : p_{y\rightarrow y} \in \mathcal{P}\}|}
\end{align*}
\]
Comparing Similarity Measures in DBLP Data

### Which venues are most similar to DASFAA?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DASFAA</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
</tr>
<tr>
<td>3</td>
<td>VLDB</td>
</tr>
<tr>
<td>4</td>
<td>SIGMOD Conference</td>
</tr>
<tr>
<td>5</td>
<td>DBX</td>
</tr>
<tr>
<td>6</td>
<td>TKDE</td>
</tr>
<tr>
<td>7</td>
<td>CIKM</td>
</tr>
<tr>
<td>8</td>
<td>Data Knowl. Eng.</td>
</tr>
<tr>
<td>9</td>
<td>SIGIR</td>
</tr>
<tr>
<td>10</td>
<td>SIGMOD Record</td>
</tr>
</tbody>
</table>

**Favor highly visible objects**

### Which venues are most similar to SIGMOD?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SIGMOD Conf.</td>
</tr>
<tr>
<td>2</td>
<td>ACM SIGMOD D. S. C.</td>
</tr>
<tr>
<td>3</td>
<td>DB for Inter. Des.</td>
</tr>
<tr>
<td>4</td>
<td>IJCAI</td>
</tr>
<tr>
<td>5</td>
<td>CIKM</td>
</tr>
<tr>
<td>6</td>
<td>AFIPS NCC</td>
</tr>
<tr>
<td>7</td>
<td>XQuery Impl. Parad.</td>
</tr>
<tr>
<td>8</td>
<td>CloudDB</td>
</tr>
</tbody>
</table>

**Are these tiny forums most similar to SIGMOD?**

### 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>
PathPredict: Meta-Path Based Relationship Prediction

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

Relationship Prediction vs. Link Prediction

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

- **Relationship prediction** in heterogeneous networks
  - 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?

- Why is homogeneous networks not sufficient?
  - In reality, objects belong to different types are linked together with different types of relations — Need to extend from link prediction to relationship prediction
  - Use heterogeneous topological features instead of homogeneous ones
- Schema-Guided Relationship Prediction
  - Semantic relationships among similar typed links share similar semantics and are comparable and inferable
  - Relationship across different typed links are not directly comparable but their collective behavior will help predicting particular relationships
- Using topological features also encoded by meta paths:
  - E.g., citation relations between authors: \( A \leftarrow P \rightarrow P \rightarrow A \)
  - Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks", ASONAM’11, July 2011.

Meta-Path Based Co-authorship Prediction in DBLP

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

<table>
<thead>
<tr>
<th>Meta-Path</th>
<th>Semantic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A - P \rightarrow P - A )</td>
<td>( a_i ) cites ( a_j )</td>
</tr>
<tr>
<td>( A - P \leftarrow 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 \rightarrow P \rightarrow P - A )</td>
<td>( a_i ) cites papers that cite ( a_j )</td>
</tr>
<tr>
<td>( A - P \leftarrow P \leftarrow P - A )</td>
<td>( a_i ) is cited by papers that are cited by ( a_j )</td>
</tr>
<tr>
<td>( A - P - P \leftarrow P - A )</td>
<td>( a_i ) and ( a_j ) cite the same papers</td>
</tr>
<tr>
<td>( A - P \rightarrow P - P - A )</td>
<td>( a_i ) and ( a_j ) are cited by the same papers</td>
</tr>
</tbody>
</table>

Meta-paths between authors under length 4
Case Study: Predicting Concrete Co-Authors

- High quality predictive power for such a difficult task

<table>
<thead>
<tr>
<th>Query author</th>
<th># Candidates</th>
<th># True relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jiawei Han</td>
<td>11934</td>
<td>36</td>
</tr>
<tr>
<td>Christos Faloutsos</td>
<td>12945</td>
<td>45</td>
</tr>
<tr>
<td>Charn Aggarwal</td>
<td>5166</td>
<td>12</td>
</tr>
<tr>
<td>Jian Pei</td>
<td>4809</td>
<td>42</td>
</tr>
<tr>
<td>Xifeng Yan</td>
<td>1617</td>
<td>8</td>
</tr>
</tbody>
</table>

<p>| Top-5 predicted co-authors for Jian Pei in 2003-2009 |</p>
<table>
<thead>
<tr>
<th>Rank</th>
<th>Hybrid features</th>
<th># Shared authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Philip S. Yu</td>
<td>Philip S. Yu</td>
</tr>
<tr>
<td>2</td>
<td>Raymond T. Ng</td>
<td>Ming-Yi Chen</td>
</tr>
<tr>
<td>3</td>
<td>Omer R. Zaier</td>
<td>Divesh Srivastava</td>
</tr>
<tr>
<td>4</td>
<td>Ling Feng</td>
<td>Kottakrishnan Ramakrishnan</td>
</tr>
<tr>
<td>5</td>
<td>David Wai-Lok Cheung</td>
<td>Jeffrey Xu Yu</td>
</tr>
</tbody>
</table>

- Predict new coauthor relationship in T2 = [2003; 2009]

When Will It Happen?

- From “whether” to “when”
  - “Whether”: Will Jim rent the movie “Avatar” in Netflix?
    - Within 1 month? 3 months? 1 year? Need to build different models!
  - “When”: When will Jim rent the movie “Avatar”?
    - What is the probability Jim will rent “Avatar” within 2 months?
    - By when Jim will rent “Avatar” with 90% probability?
    - What is the expected time it will take for Jim to rent “Avatar”?

- Y. Sun, J. Han, C. C. Aggarwal, and N. Chawla, “When Will It Happen? Relationship Prediction in Heterogeneous Information Networks”, WSDM’12, Feb. 2012
Role Discovery in Networks: Why Does It Matter?

- Objective: Extract semantic meaning from plain links to finely model and better organize information networks
- Challenges
  - Latent semantic knowledge
  - Interdependency
  - Scalability
- Opportunity
  - Human intuition
  - Realistic constraint
  - Crosscheck with collective intelligence
- Methodology: propagate simple intuitive rules and constraints over the whole network
Discovery of Advisor-Advisee Relationships in DBLP Network

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

Mining Evolution and Dynamics in HINs

- 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
- EvoNetClus: Modeling evolution of dynamic heterogeneous networks
  - Co-evolution within a community
    - heterogeneous multi-typed object/links
  - Discovery of evolution structures among different communities
- Y. Sun, et al., "Studying Co-Evolution of Multi-Typed Objects in Dynamic Heterogeneous Information Networks", MLG’10
Evolution: Idea Illustration

- From network sequences to evolutionary communities

Related Work

- Backstrom & Leskovec (2011)
  - Supervised random walks for link prediction and recommendation in HINs
- Dong et al, 2012
  - A ranking factor graph model (RFG) for predicting links in HINs
- Tang et al. (2008, 2012)
  - Community evolution in HINs (called multi-mode networks) using a clustering method on evolving networks
- Davis et al. (2011)
  - Link prediction in HINs using an extension to Adamic/Adar measure and exploiting classification
Our Current Work

- **Goal**
  - Mining heterogeneous information networks

- **Approach**
  - Exploit the potential and theoretical basis of formal concept analysis (FCA) and two of its extensions to manage multidimensionality and heterogeneity in networks
  - Use and adapt a set of findings on concept pruning, core/peripheral node identification, network partitioning (e.g., biclustering and triclustering), taxonomy-based mining to better analyze large networks and extract rich patterns such as groups and association rules

Our Current Work

- **Networks under study**
  - Analysis of a HIN with an interaction network together with an affiliation one for link prediction and recommendation
  - Generalize the approach to an arbitrary HIN
  - Analysis of a network with tridimensional and even n-dimensional data using triadic (and later on polyadic) concept analysis
  - Exploration of our previous work on formal concept analysis (e.g., implications with negation, attribute/object generalization, operations on lattices) to detect richer and user-oriented patterns in HINs
Our Current Work

- Link prediction and recommendation
  - Add a new node $N_i$ and at least a link in a HIN with two types of nodes and links
  - Use formal concept analysis together with concept (cluster) pruning and weighting to suggest a set of links to be added between the new node and existing nodes

![Diagram showing link prediction and recommendation]

Our Current Work

- Networks with tridimensional data
  - Objects with attributes under conditions
  - E.g. events (1..5), researchers (P, N, R, K, S) and roles (a, b, c, d)
    - a: speaker (at a given event), b: organizer
    - c: author, d: PC member
  - E.g., Researcher $K$ attends Event 2 with two roles: author and PC member

![Matrix showing network data]

![Matrix showing network data]
Our Current Work

- Triadic concept analysis (Lehmann & Wille, 1995)

A triadic context $\mathcal{K} := (K_1, K_2, K_3, Y)$ where $Y \subseteq K_1 \times K_2 \times K_3$. The elements of $K_1$, $K_2$ and $K_3$ are called (formal) objects, attributes and conditions, respectively. A triple $(a_1, a_2, a_3)$ in $Y$ means that object $a_1$ has attribute $a_2$ under condition $a_3$.


- Example: The triadic concept $(345, RK, ab)$ means that Events 3, 4 & 5 attract Researchers R & K with roles a and b.
Our Current Work

- Triadic association rules
  - E.g., any role (e.g., speaker) played by $S$ is also played by $P$
  - $N \rightarrow P$: whenever researcher $N$ attends events as a speaker and PC member, $P$ does so


Conclusion (Sun & Han, 2012)

- Rich knowledge can be mined from information networks
- What is the magic?
  - *Heterogeneous, structured information networks!*
- Clustering, ranking and classification: Integrated clustering, ranking and classification: RankClus, RankClass, ...
- Meta-Path based similarity search and relationship prediction
- Role discovery and evolutionary analysis
- Knowledge is power, but knowledge is hidden in *massive links!*
- *Mining heterogeneous information networks*: Much more to be explored!!
Future Research (Sun & Han, 2012)

- 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

- S. Chakrabarti. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2003
### 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

- L. 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
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.
References: Social Network Analysis

- 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: Link and Relationship Prediction

- 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 and Summarization

- 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
- 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, and Jiawei Han, “Graph Cube: On Warehousing and OLAP Multidimensional Networks”, SIGMOD'11
- Peixiang Zhao and Jiawei Han, “On Graph Query Optimization in Large Networks”, Proc. 2010 Int. Conf. on Very Large Data Bases (VLDB’10), Singapore, Sept. 2010

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