Mining Heterogeneous Information Networks

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MARCH 6, 2018
Outline

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

- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks

- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Information Networks
  - Relationship Prediction in Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation recommendation in heterogeneous networks

- Summary
Ubiquitous Information Networks

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

World-Wide Web

Yeast protein interaction network

Co-author network

Social network sites
Heterogeneous Information Networks

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

Co-author Network

Conference-Author Network

Tom
Mary
Alice
Bob
Cindy
Tracy
Jack
Mike
Lucy
Jim

SIGMOD
VLDB
EDBT
KDD
ICDM
SDM
AAAI
ICML
Mining Heterogeneous Information Networks

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

- **Bibliographic information networks:** DBLP, ArXive, PubMed
  - Entity types: paper (P), venue (V), author (A), and term (T)
  - Relation type: authors write papers, venues publish papers, papers contain terms

- **Twitter information network, and other social media network**
  - Objects types: user, tweet, hashtag, and term
  - Relation/link types: users follow users, users post tweets, tweets reply tweets, tweets use terms, tweets contain hashtags

- **Flickr information network**
  - Object types: image, user, tag, group, and comment
  - Relation types: users upload images, image contain tags, images belong to groups, users post comments, and comments comment on images

- **Healthcare information network**
  - Object types: doctor, patient, disease, treatment, and device
  - Relation types: treatments used-for diseases, patients have diseases, patients visit doctors
Structures Facilitate Mining Heterogeneous Networks

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

- Venue
- Paper
- Author
- DBLP Bibliographic Network

- KDD
- VLDB
- GenClus
- NetClus
- PathSim

- Movie
- Studio
- The IMDB Movie Network
- Actor
- Movie
- Director

- The Facebook Network

It works well for ego-networks!
What Can be Mined from Heterogeneous Networks?

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

Knowledge hidden in DBLP Network

<table>
<thead>
<tr>
<th>Mining Functions</th>
<th>Knowledge hidden in DBLP Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranking</td>
<td>Who are the <strong>leading</strong> researchers on Web search?</td>
</tr>
<tr>
<td>Similarity Search</td>
<td>Who are the <strong>peer</strong> researchers of Jure Leskovec?</td>
</tr>
<tr>
<td>Relationship Prediction</td>
<td>Whom <strong>will</strong> Christos Faloutsos <strong>collaborate with</strong>?</td>
</tr>
<tr>
<td>Relation Strength Learning</td>
<td>Which <strong>relationships</strong> are most <strong>influential</strong> for an author to decide her topics?</td>
</tr>
<tr>
<td>Network Evolution</td>
<td>How was the field of Data Mining <strong>emerged</strong> or <strong>evolving</strong>?</td>
</tr>
<tr>
<td>Outlier/anomaly detection</td>
<td>Which authors are <strong>rather different</strong> from his/her peers in IR?</td>
</tr>
</tbody>
</table>
Principles of Mining Heterogeneous Information Net

- Information propagation across heterogeneous nodes & links
  - *How to compute* ranking scores, similarity scores, and clusters, and how to make good use of class labels, across heterogeneous nodes and links
  - *Objects in the networks are interdependent and knowledge can only be mined using the holistic information in a network*

- Search and mining by exploring network meta structures
  - Heter. info networks: semi-structured and typed
  - Network schema: a meta structure, guidance of search and mining
  - Meta-path based similarity search and mining

- User-guided exploration of information networks
  - Automatically select the right relation (or meta-path) combinations with appropriate weights for a particular search or mining task
  - User-guided or feedback-based network exploration is a strategy
Outline

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

- Clustering without ranking? Think about no PageRank dark time before Google
- Ranking will make more sense within a particular cluster
  - Einstein in physics vs. Turing in computer science

Why not integrate ranking with clustering & classification?

- High-ranked objects should be more important in a cluster than low-ranked ones
- Why treat every object the same weight in the same cluster?
  - But how to get their weight?

Integrate ranking with clustering/classification in the same process

- Ranking, as the feature, is conditional (i.e., relative) to a specific cluster
- Ranking and clustering may mutually enhance each other
- Ranking-based clustering: RankClus [EDBT’09], NetClus [KDD’09]
A Bi-Typed Network Model and Simple Ranking

- A bi-typed network model
  - Let \( X \) represents type venue
  - \( Y \): Type author
- The DBLP network can be represented as matrix \( W \)
- Our task: Rank-based clustering of heterogeneous network \( W \)
- Simple Ranking
  - Proportional to # of publications of an author and a venue
  - Considers only immediate neighborhood in the network

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

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

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

- Methodology: **Propagate** the ranking scores in the network over different types

- Rule 1: Highly ranked authors publish *many* papers in highly ranked venues
  \[ r_Y(j) = \sum_{i=1}^{m} W_{YX}(j,i) r_X(i) \]

- Rule 2: Highly ranked venues attract *many* papers from *many* highly ranked authors
  \[ r_X(i) = \sum_{j=1}^{n} W_{XY}(i,j) r_Y(j) \]

- Rule 3: The rank of an author is enhanced if he or she co-authors with *many* highly ranked authors
  \[ r_Y(i) = \alpha \sum_{j=1}^{m} W_{YX}(i,j) r_X(j) + (1 - \alpha) \sum_{j=1}^{n} W_{XY}(i,j) r_Y(j) \]

- Other ranking functions are quite possible (e.g., using domain knowledge)
  - Ex. Journals may weight more than conferences in science
A ranking function is not only related to the link property, but also depends on domain knowledge

- Ex: Journals may weight more than conferences in science

Ranking functions can be defined on multi-typed networks

- Ex: PopRank takes into account the impact both from the same type of objects and from the other types of objects, with different impact factors for different types

Use expert knowledge, for example,

- TrustRank semi-automatically separates reputable, good objects from spam ones
- Personalized PageRank uses expert ranking as query and generates rank distributions w.r.t. such knowledge

A research problem that needs systematic study
The k-means algorithm has two steps at each iteration (in the E-M framework):

- **Expectation Step** (E-step): Given the current cluster centers, each object is assigned to the cluster whose center is closest to the object: An object is expected to belong to the closest cluster.

- **Maximization Step** (M-step): Given the cluster assignment, for each cluster, the algorithm adjusts the center so that the sum of distance from the objects assigned to this cluster and the new center is minimized.

**The (EM) algorithm**: A framework to approach maximum likelihood or maximum a posteriori estimates of parameters in statistical models.

- **E-step** assigns objects to clusters according to the current probabilistic clustering or parameters of probabilistic clusters.

- **M-step** finds the new clustering or parameters that minimize the sum of squared error (SSE) or maximize the expected likelihood.
From Conditional Rank Distribution to E-M Framework

- Given a bi-typed bibliographic network, how can we use the conditional rank scores to further improve the clustering results?

- Conditional rank distribution as cluster feature
  - For each cluster $C_k$, the conditional rank scores, $r_X|C_k$ and $r_Y|C_k$, can be viewed as conditional rank distributions of $X$ and $Y$, which are the features for cluster $C_k$

- Cluster membership as object feature
  - From $p(k|o_i) \propto p(o_i|k)p(k)$, the higher its conditional rank in a cluster $(p(o_i|k))$, the higher possibility an object will belong to that cluster $(p(k|o_i))$
  - Highly ranked attribute object has more impact on determining the cluster membership of a target object

- Parameter estimation using the Expectation-Maximization algorithm
  - E-step: Calculate the distribution $p(z = k|y_j, x_i, \Theta)$ based on the current value of $\Theta$
  - M-Step: Update $\Theta$ according to the current distribution
RankClus: Integrating Clustering with Ranking

- An EM styled Algorithm
  - Initialization
    - Randomly partition
  - Repeat
    - Ranking
      - Ranking objects in each sub-network induced from each cluster
    - Generating new measure space
    - Estimate \textit{mixture model coefficients} for each target object
    - Adjusting cluster
  - Until change < threshold

An E-M framework for iterative enhancement

RankClus [EDBT’09]: Ranking and clustering mutually enhancing each other in an E-M framework
Initially, ranking distributions are mixed together

Improved a little

Improved significantly

Stable

Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

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

Normalized Mutual Info. (NMI):

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

D1: med. separated & med. density
D2: med. separated & low density
D3: med. Separated & high density
D4: highly separated & med. density
D5: poorly separated & med. density
## RankClus: Clustering & Ranking CS Venues in DBLP

### Top-10 venues in 5 clusters generated by RankClus in DBLP

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

Top-10 conferences in 5 clusters using RankClus in DBLP (when k = 15)
Time Complexity: Linear to # of Links

- At each iteration, $|E|$: edges in network, $m$: # of target objects, $K$: # of clusters
  - Ranking for sparse network
    - $\sim O(|E|)$
  - Mixture model estimation
    - $\sim O(K|E| + mK)$
  - Cluster adjustment
    - $\sim O(mK^2)$
  - In all, linear to $|E|$
    - $\sim O(K|E|)$
- Note: SimRank will be at least quadratic at each iteration since it evaluates distance between every pair in the network
NetClus: Ranking-Based Clustering with Star Network Schema

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

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

- Generate initial partitions for target objects and induce initial net-clusters from the original network

- Repeat // An E-M Framework
  - Build ranking-based probabilistic generative model for each net-cluster
  - Calculate the posterior probabilities for each target object
  - Adjust their cluster assignment according to the new measure defined by the posterior probabilities to each cluster

- Until the clusters do not change significantly

- Calculate the posterior probabilities for each attribute object in each net-cluster
NetClus on the DBLP Network

- NetClus initialization: $G = (V, E, W)$, weight $w_{x_ix_j}$ linking $x_i$ and $x_j$
  - $V = A \cup C \cup D \cup T$, where $D$ (paper), $A$ (author), $C$ (conf.), $T$ (term)
    
    $w_{x_ix_j} = \begin{cases} 
    1, & \text{if } x_i(x_j) \in A \cup C \text{ and } x_j(x_i) \in D, \\
    & \text{and } x_i \text{ has link to } x_j \\
    c, & \text{if } x_i(x_j) \in T \text{ and } x_j(x_i) \in D \text{ and } x_i(x_j) \\
    & \text{appears } c \text{ times in } x_j(x_i), \\
    0, & \text{otherwise.}
    \end{cases}$

- Simple ranking:

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

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

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

- For DBLP:

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

$$P(A|T_A, G) = W_{AD}D^{-1}_{DC}W_{DC}P(C|T_C, G)$$
Multi-Typed Networks Lead to Better Results

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

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

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

<table>
<thead>
<tr>
<th>Conference</th>
<th>AAAI</th>
<th>CIKM</th>
<th>CVPR</th>
<th>ECIR</th>
<th>ECML</th>
<th>EDBT</th>
<th>ICDE</th>
<th>ICDM</th>
<th>ICML</th>
<th>IJCAI</th>
<th>KDD</th>
<th>PAKDD</th>
<th>PKDD</th>
<th>PODS</th>
<th>SDM</th>
<th>SIGIR</th>
<th>SIGMOD</th>
<th>VLDB</th>
<th>WSDM</th>
<th>WWW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0022667</td>
<td>0.00899168</td>
<td><strong>0.934024</strong></td>
<td>0.0300042</td>
<td>0.0247133</td>
<td>CIKM 0.150053</td>
<td>0.310172</td>
<td>0.00723807</td>
<td>0.444524</td>
<td>0.0880127</td>
<td>CVPR 0.000163812</td>
<td>0.00763072</td>
<td><strong>0.931496</strong></td>
<td>0.0281342</td>
<td>0.032575</td>
<td>ECIR 3.47023e-05</td>
<td>0.00657402</td>
<td><strong>0.978391</strong></td>
<td>0.00787288</td>
<td>EDBT 0.573362</td>
</tr>
</tbody>
</table>
NetClus: Experiment on DBLP: Database System Cluster

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

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

### RankCompete: Organize your photo album 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>

### MedRank: Rank treatments for AIDS from Medline

- Use multi-typed image features to build up heterogeneous networks
- Explore multiple types in a star schema network
iNextCube: Information Network-Enhanced Text Cube

Architecture of iNextCube

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Conf/Journal</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>0.075232</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
<td>0.062318</td>
</tr>
<tr>
<td>3</td>
<td>SIGIR</td>
<td>0.053869</td>
</tr>
<tr>
<td>4</td>
<td>SIGMOD</td>
<td>0.048740</td>
</tr>
<tr>
<td>5</td>
<td>KDD</td>
<td>0.028168</td>
</tr>
<tr>
<td>6</td>
<td>IEEE Trans. Scal. Data Eng.</td>
<td>0.024118</td>
</tr>
<tr>
<td>7</td>
<td>IEEE Data Eng. Mag.</td>
<td>0.022618</td>
</tr>
<tr>
<td>8</td>
<td>ACM Trans. Comput. Syst.</td>
<td>0.020792</td>
</tr>
<tr>
<td>9</td>
<td>ACM Trans. Comput. Syst.</td>
<td>0.020606</td>
</tr>
<tr>
<td>10</td>
<td>FOCS</td>
<td>0.015857</td>
</tr>
<tr>
<td>11</td>
<td>TREC</td>
<td>0.014045</td>
</tr>
<tr>
<td>12</td>
<td>In Proc. Manage.</td>
<td>0.011904</td>
</tr>
<tr>
<td>13</td>
<td>ICDE</td>
<td>0.011839</td>
</tr>
<tr>
<td>14</td>
<td>VLDB</td>
<td>0.011544</td>
</tr>
<tr>
<td>15</td>
<td>SIGIR</td>
<td>0.011441</td>
</tr>
<tr>
<td>16</td>
<td>CLUDE</td>
<td>0.009575</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>David A. Bost</td>
<td>0.009409</td>
</tr>
<tr>
<td>2</td>
<td>Michael Steinberg</td>
<td>0.007072</td>
</tr>
<tr>
<td>3</td>
<td>Richard T. Hodgins</td>
<td>0.006152</td>
</tr>
<tr>
<td>4</td>
<td>David J. Dewitt</td>
<td>0.004660</td>
</tr>
<tr>
<td>5</td>
<td>Suraj Chaudhuri</td>
<td>0.004424</td>
</tr>
<tr>
<td>6</td>
<td>Michael J. Carey</td>
<td>0.004195</td>
</tr>
<tr>
<td>7</td>
<td>Won Kim</td>
<td>0.004167</td>
</tr>
<tr>
<td>8</td>
<td>Hector Quesnel-Molina</td>
<td>0.003793</td>
</tr>
<tr>
<td>9</td>
<td>Michael J. Franklin</td>
<td>0.003775</td>
</tr>
<tr>
<td>10</td>
<td>Marwan Mindell</td>
<td>0.003753</td>
</tr>
<tr>
<td>11</td>
<td>C. Mohan</td>
<td>0.003580</td>
</tr>
<tr>
<td>12</td>
<td>Philip A. Bernstein</td>
<td>0.003459</td>
</tr>
<tr>
<td>13</td>
<td>Arne Seeger</td>
<td>0.003191</td>
</tr>
<tr>
<td>14</td>
<td>Dai Suda</td>
<td>0.003159</td>
</tr>
<tr>
<td>15</td>
<td>Compaq Weisner</td>
<td>0.003043</td>
</tr>
<tr>
<td>16</td>
<td>Jennifer Wilcox</td>
<td>0.003033</td>
</tr>
<tr>
<td>17</td>
<td>H. Y. Jagadish</td>
<td>0.002918</td>
</tr>
<tr>
<td>18</td>
<td>Jeffrey F. Naughton</td>
<td>0.002911</td>
</tr>
<tr>
<td>19</td>
<td>Raghu Ramaswamy</td>
<td>0.002832</td>
</tr>
<tr>
<td>20</td>
<td>Joseph M. Hellerstein</td>
<td>0.002793</td>
</tr>
<tr>
<td>21</td>
<td>Raphael Agarwal</td>
<td>0.002769</td>
</tr>
<tr>
<td>22</td>
<td>Umeshwar Dayal</td>
<td>0.002763</td>
</tr>
<tr>
<td>23</td>
<td>Serge Abiteboul</td>
<td>0.002737</td>
</tr>
</tbody>
</table>

In sub-area Algorithms and Theory of Computation, the top ranked authors are:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Philip S. Yu</td>
<td>0.009881</td>
</tr>
<tr>
<td>2</td>
<td>Jianzong Han</td>
<td>0.007163</td>
</tr>
<tr>
<td>3</td>
<td>Chuan C. Aggarwal</td>
<td>0.005583</td>
</tr>
<tr>
<td>4</td>
<td>Cristian Falbolesi</td>
<td>0.005140</td>
</tr>
<tr>
<td>5</td>
<td>Beng Chin Ooi</td>
<td>0.003431</td>
</tr>
<tr>
<td>6</td>
<td>Minyi Sun Chien</td>
<td>0.003209</td>
</tr>
<tr>
<td>7</td>
<td>Hsin-Chieh Hwang</td>
<td>0.003160</td>
</tr>
<tr>
<td>8</td>
<td>Wei Wang</td>
<td>0.003009</td>
</tr>
<tr>
<td>9</td>
<td>Jiongshi Tan</td>
<td>0.002389</td>
</tr>
<tr>
<td>10</td>
<td>Nick Koudas</td>
<td>0.002260</td>
</tr>
<tr>
<td>11</td>
<td>H. V. Jagadish</td>
<td>0.002196</td>
</tr>
</tbody>
</table>

In area Algorithms and Theory of Computation, the top ranked conferences/journals are:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Conf/Journal</th>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>STOC</td>
<td>1</td>
<td>Andrew Chi-Chih Yao</td>
<td>0.009409</td>
</tr>
<tr>
<td>2</td>
<td>FOCS</td>
<td>2</td>
<td>Christos H. Papadimitriou</td>
<td>0.007072</td>
</tr>
<tr>
<td>3</td>
<td>IEEE Comp. Mag.</td>
<td>3</td>
<td>Robert E. Tarjan</td>
<td>0.006152</td>
</tr>
<tr>
<td>4</td>
<td>SIGMOD</td>
<td>4</td>
<td>David Eppstein</td>
<td>0.004660</td>
</tr>
<tr>
<td>5</td>
<td>J. Comput. Syst.</td>
<td>5</td>
<td>Minu Sinha</td>
<td>0.004424</td>
</tr>
<tr>
<td>6</td>
<td>ACM</td>
<td>6</td>
<td>Avi Wigderson</td>
<td>0.004195</td>
</tr>
<tr>
<td>7</td>
<td>CGIS</td>
<td>7</td>
<td>John H. Reif</td>
<td>0.004167</td>
</tr>
<tr>
<td>8</td>
<td>Theor. Comput. Sys.</td>
<td>8</td>
<td>Bernard Chazelle</td>
<td>0.003793</td>
</tr>
</tbody>
</table>
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
- ...

Outline

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

- Similarity measure/search is the base for cluster analysis
- Who are the most similar to Christos Faloutsos based on the DBLP network?
- Meta-Path: **Meta-level description** of a path between two objects
  - A path on network schema
  - Denote an existing or concatenated relation between two object types
- Different meta-paths tell different semantics

<table>
<thead>
<tr>
<th>Meta-Path: Author-Paper-Author</th>
<th>Meta-Path: Author-Paper-Venue-Paper-Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>Author</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
</tr>
<tr>
<td>2</td>
<td>Spiros Papadimitriou</td>
</tr>
<tr>
<td>3</td>
<td>Jineng Sun</td>
</tr>
<tr>
<td>4</td>
<td>Jia-Yu Pan</td>
</tr>
<tr>
<td>5</td>
<td>Agma J. M. Traina</td>
</tr>
<tr>
<td>6</td>
<td>Jure Leskovec</td>
</tr>
<tr>
<td>7</td>
<td>Caetano Traina Jr.</td>
</tr>
<tr>
<td>8</td>
<td>Hanghang Tong</td>
</tr>
<tr>
<td>9</td>
<td>Deepayan Chakrabarti</td>
</tr>
<tr>
<td>10</td>
<td>Flip Korn</td>
</tr>
</tbody>
</table>

Christos’ students or close collaborators
Work in similar fields with similar reputation
Existing Popular Similarity Measures for Networks

- **Random walk (RW):**
  - The probability of random walk starting at \( x \) and ending at \( y \), with meta-path \( P \)
  
  \[
  s(x, y) = \sum_{p \in P} \text{prob}(p)
  \]
  
  - Used in Personalized PageRank (P-Pagerank) (Jeh and Widom 2003)
  - Favors **highly visible** objects (i.e., objects with large degrees)

- **Pairwise random walk (PRW):**
  - The probability of pairwise random walk starting at \((x, y)\) and ending at a common object (say \( z \)), following a meta-path \((P_1, P_2)\)
  
  \[
  s(x, y) = \sum_{(p_1, p_2) \in (P_1, P_2)} \text{prob}(p_1)\text{prob}(p_2)
  \]
  
  - Used in SimRank (Jeh and Widom 2002)
  - Favors **pure** objects (i.e., objects with highly skewed distribution in their in-links or out-links)

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

- SimRank (Jeh and Widom 2002)
  - Base: objects are maximally similar to themselves, i.e., $s_0(a, b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases}$
  - Induction: Two objects are considered to be similar if they are referenced by similar objects
    $$s(a, b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{l(a)} \sum_{j=1}^{l(b)} s(I_i(a), I_j(b))$$
  - The computation is quite costly: Many efficient computation methods proposed
    Glen Jeh and Jennifer Widom. SimRank: A Measure of Structural-Context Similarity. In KDD'02

- Personalized PageRank (P-Pagerank) (Jeh and Widom 2003)
  - P-PageRank score $x$ is defined as: $x = \alpha Px + (1 - \alpha)b$, where $P$ is a transition matrix of the network $G$, $b$ is a stochastic vector, called personalized vector, and $\alpha \in (0, 1)$ is the teleportation constant
  - Efficient computation methods are also studied (e.g., Maehara, et al., VLDB’14)
    Glen Jeh and Jennifer Widom. Scaling Personalized Web Search, In WWW 2003
Which Similarity Measure Is Better for Finding Peers?

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

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

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

<table>
<thead>
<tr>
<th>Author \ Conf.</th>
<th>SIGMOD</th>
<th>VLDB</th>
<th>ICDM</th>
<th>KDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Jim</td>
<td>50</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mary</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Bob</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ann</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure \ Author</th>
<th>Jim</th>
<th>Mary</th>
<th>Bob</th>
<th>Ann</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-PageRank</td>
<td>0.376</td>
<td>0.013</td>
<td>0.016</td>
<td>0.005</td>
</tr>
<tr>
<td>SimRank</td>
<td>0.716</td>
<td>0.572</td>
<td>0.713</td>
<td>0.184</td>
</tr>
<tr>
<td>Random Walk</td>
<td>0.8983</td>
<td>0.0238</td>
<td>0.0390</td>
<td>0</td>
</tr>
<tr>
<td>Pairwise R.W.</td>
<td>0.5714</td>
<td>0.4440</td>
<td>0.5556</td>
<td>0</td>
</tr>
<tr>
<td>PathSim (APCPA)</td>
<td>0.083</td>
<td>0.8</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Comparison of Multiple Measures: A Toy Example
**Example with DBLP: Find Academic Peers by PathSim**

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

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

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

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

---

**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 Silverstein</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>

---

**Meta-Path:**

- Author
- Paper
- Venue
- Paper
- Author
Some Meta-Path Is “Better” Than Others

Which pictures are most similar to this one?

Evaluate the similarity between images according to their linked tags

Meta-Path: Image-Tag-Image

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

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
Comparing Similarity Measures in DBLP Data

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Conference</th>
<th>Rank</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SIGMOD Conf.</td>
<td>1</td>
<td>SIGMOD Conf.</td>
</tr>
<tr>
<td>2</td>
<td>Found. and Trends in DB</td>
<td>2</td>
<td>VLDB</td>
</tr>
<tr>
<td>3</td>
<td>ACM SIGMOD D. S. C.</td>
<td>3</td>
<td>ICDE</td>
</tr>
<tr>
<td>5</td>
<td>DB for Inter. Des.</td>
<td>5</td>
<td>SIGMOD Rec.</td>
</tr>
<tr>
<td>6</td>
<td>IPSJ</td>
<td>6</td>
<td>ACM Trans. DB Syst.</td>
</tr>
<tr>
<td>7</td>
<td>CIDR</td>
<td>7</td>
<td>TKDE</td>
</tr>
<tr>
<td>8</td>
<td>AFIPS NCC</td>
<td>8</td>
<td>PODS</td>
</tr>
<tr>
<td>9</td>
<td>XQuery Impl. Parad</td>
<td>9</td>
<td>VLDB J.</td>
</tr>
<tr>
<td>10</td>
<td>CleanDB</td>
<td>10</td>
<td>EDBT</td>
</tr>
</tbody>
</table>

Table 6: SimRank vs. PathSim on query: “SIGMOD”
Long Meta-Path May Not Carry the Right Semantics

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

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

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

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

Table 8: Top-10 similar conferences to “SIGMOD” under path schemas with different lengths
Co-Clustering-Based Pruning Algorithm

- Meta-Path based similarity computation can be costly
- The overall cost can be reduced by storing commuting matrices for short path schemas and computing top-$k$ queries on line

Framework

- Generate co-clusters for materialized commuting matrices for feature objects and target objects
- Derive upper bound for similarity between object and target cluster and between object and object
- Safely prune target clusters and objects if the upper bound similarity is lower than current threshold
- Dynamically update top-$k$ threshold
Meta-Path: A Key Concept for Heterogeneous Networks

- Meta-path based mining
  - PathPredict [Sun et al., ASONAM’11]
    - Co-authorship prediction using meta-path based similarity
  - PathPredict_when [Sun et al., WSDM’12]
    - When a relationship will happen
  - Citation prediction [Yu et al., SDM’12]
    - Meta-path + topic
- Meta-path learning
  - User Guided Meta-Path Selection [Sun et al., KDD’12]
  - Meta-path selection + clustering
Outline

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

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

Which meta-path(s) to choose?
User Guidance Determines Clustering Results

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

Seeds | Meta-path(s) | Clustering
---|---|---
{1} {5} | \{1\} \{5\} \{1,2,3,4\} \{5,6,7,8\} | (a) AOA

Seeds | Meta-path(s) | Clustering
---|---|---
{1} {2} {5} {6} | \{1\} \{2\} \{5\} \{6\} \{1,3\} \{2,4\} \{5,7\} \{6,8\} | (c) AOA + AVA

Problem: User-guided clustering with meta-path selection

Input:
- The target type for clustering $T$
- # of clusters $k$
- Seeds in some clusters: $L_1, \ldots, L_k$
- Candidate meta-paths: $P_1, \ldots, P_M$

Output:
- Weight of each meta-path: $w_1, \ldots, w_m$
- Clustering results that are consistent with the user guidance
PathSelClus: A Probabilistic Modeling Approach

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

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

$$P(W_m|\Pi_m, \Theta, B_m) = \prod_i P(w_{i,m}|\pi_{i,m}, \Theta, B_m) = \prod_i \prod_j (\pi_{ij,m})^{w_{ij,m}}$$
Part 2: Modeling the Guidance from Users

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

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

- Model quality weight $\alpha_m$ as the relative weight for each relationship in $W_m$
  - Observation of relationships: $W_m \rightarrow \alpha_m W_m$
  - 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
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

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

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

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

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

<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>0.7435</td>
<td>0.1137</td>
<td>0.1758</td>
<td>0.2112</td>
<td>0.2112</td>
<td>0.2430</td>
<td>0.2022</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6517</td>
<td>0.0323</td>
<td>0.0178</td>
<td>0.0578</td>
<td>0.0578</td>
<td>0.2308</td>
<td>0.2490</td>
</tr>
<tr>
<td>2%</td>
<td>Accuracy</td>
<td>0.8004</td>
<td>0.1264</td>
<td>0.1910</td>
<td>0.2202</td>
<td>0.2202</td>
<td>0.2762</td>
<td>0.2792</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6803</td>
<td>0.0487</td>
<td>0.0150</td>
<td>0.0801</td>
<td>0.0801</td>
<td>0.2099</td>
<td>0.2907</td>
</tr>
<tr>
<td>5%</td>
<td>Accuracy</td>
<td>0.8125</td>
<td>0.2653</td>
<td>0.2200</td>
<td>0.2437</td>
<td>0.2437</td>
<td>0.3049</td>
<td>0.3240</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6894</td>
<td>0.1111</td>
<td>0.0220</td>
<td>0.1212</td>
<td>0.1212</td>
<td>0.2252</td>
<td>0.2692</td>
</tr>
</tbody>
</table>

Users try different kinds of food
Outline

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

- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks

- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks

- **Summary**
Classification: Knowledge Propagation Across Heterogeneous Typed Networks

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

**Classification:** Labeled knowledge propagates through multi-typed objects across heterogeneous networks [KDD’11]
GNetMine: Methodology

- Classification of networked data can be essentially viewed as a process of *knowledge propagation*, where information is propagated from labeled objects to unlabeled ones through links until a stationary state is achieved.

- A novel graph-based regularization framework to address the classification problem on heterogeneous information networks.

- Respect the link type differences by preserving consistency over each relation graph corresponding to each type of links separately.

  - Mathematical intuition: Consistency assumption

  - The confidence \( f \) of two objects \( x_{ip} \) and \( x_{jq} \) belonging to class \( k \) should be similar if \( x_{ip} \leftrightarrow x_{jq} \) \( (R_{ij,pq} > 0) \).

  - \( f \) should be similar to the given ground truth on the labeled data.
GNetMine: Graph-Based Regularization

- Minimize the objective function

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

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

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

Smoothness constraints: objects linked together should share similar estimations of confidence belonging to class \(k\)
Normalization term applied to each type of link separately: reduce the impact of popularity of objects
Confidence estimation on labeled data and their pre-given labels should be similar
RankClass: Ranking-Based Classification

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

- RankClus [EDBT’09]: Clustering and ranking working together
  - No training, no available class labels, no expert knowledge
- GNetMine [PKDD’10]: Incorp. label information in networks
  - Classification in heterog. networks, but objects treated equally
- RankClass [KDD’11]: Integration of ranking and classification in heterogeneous network analysis
  - Ranking: informative understanding & summary of each class
  - Class membership is critical information when ranking objects
  - Let ranking and classification mutually enhance each other!
  - Output: Classification results + ranking list of objects within each class
Why Rank-Based Classification?

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

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

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

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

Class 1
- Ranking distribution: \( P(x|T(x), 1) \)

Class 2
- Ranking distribution: \( P(x|T(x), 2) \)

Class 3
- Ranking distribution: \( P(x|T(x), 3) \)

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

\[ \text{class}(x) = \arg\max_{1 \leq k \leq K} P(k|x, T(x)) \]

The RankClass Framework

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

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

Converge

Calculate the posterior probability, i.e., \( P(k|x, T(x)) \)
Graph-Based Ranking

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

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

The initial ranking score

Weighted average of the neighbors' ranking scores
Ideally, the within-class ranking should be performed within the sub-network corresponding to each class.

Use the ranking distribution to describe each class.

Gradually emphasize the network on highly ranked objects, and weaken the network on lowly ranked objects.
Comparing Classification Accuracy on the DBLP Data

- **Class:** Four research areas: DB, DM, AI, IR
- **Four types of objects**
  - Paper(14376), Conf.(20), Author(14475), Term(8920)
- **Three types of relations**
  - Paper-conf., paper-author, paper-term
- **Algorithms for comparison**
  - LLGC [Zhou et al. NIPS’03]
  - wvRN) [Macskassy et al. JMLR’07]
  - nLB [Lu et al. ICML’03, Macskassy et al. JMLR’07]
Object Ranking in Each Class: Experiment

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

<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></td>
<td>SIGIR</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>SDM</td>
<td>AAAI</td>
<td></td>
<td>ECIR</td>
</tr>
<tr>
<td>ICDE</td>
<td>ICDM</td>
<td>ICML</td>
<td></td>
<td>CIKM</td>
</tr>
<tr>
<td>PODS</td>
<td>PKDD</td>
<td>CVPR</td>
<td></td>
<td>WWW</td>
</tr>
<tr>
<td>EDBT</td>
<td>PAKDD</td>
<td>ECML</td>
<td></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>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>
Outline

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

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

- Relationship prediction in heterogeneous networks
  - Different types of relationships need different prediction models
  - Different connection paths need to be treated separately!
    - **Meta-path-based approach** to define topological features
Rich semantics contained in structured, text-rich heterogeneous networks

Homogeneous networks, such as coauthor networks, miss too much critically important information

Schema-guided relationship prediction

Semantic relationships among similar typed links share similar semantics and are comparable and inferable

Relationships across different typed links are not directly comparable but their collective behavior will help predict particular relationships

Meta-paths: encoding topological features

E.g., citation relations between authors:  A — P → P — A

Y. Sun, R. Barber, M. Gupta, C. Aggarwal and J. Han, "Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks", ASONAM'11
PathPredict: Meta-Path Based Relationship Prediction

- Who will be your new coauthors in the next 5 years?
- Meta path-guided prediction of links and relationships
- Philosophy: Meta path relationships among similar typed links share similar semantics and are comparable and inferable
- Co-author prediction (A→P→A) [Sun et al., ASONAM’11]
- Use topological features encoded by meta paths, e.g., citation relations between authors (A→P→P→A)

<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>
Logistic Regression-Based Prediction Model

- Training and test pair: \(<x_i, y_i> = \text{<history feature list, future relationship label>}

<table>
<thead>
<tr>
<th></th>
<th>A—P—A</th>
<th>A—P—V</th>
<th>A—P—T</th>
<th>A—P→P</th>
<th>A—P</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Mike, Ann&gt;</td>
<td>4</td>
<td>5</td>
<td>100</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>&lt;Mike, Jim&gt;</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

- Logistic Regression Model
  - Model the probability for each relationship as
    \[ p_i = \frac{e^{x_i \beta}}{e^{x_i \beta} + 1} \]
  - \( \beta \) is the coefficients for each feature (including a constant 1)
  - MLE (Maximum Likelihood Estimation)
  - Maximize the likelihood of observing all the relationships in the training data
    \[ L = \prod_i p_i^{y_i} (1 - p_i)^{(1-y_i)} \]
Selection among Competitive Measures

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

- **Path Count**: Number of path instances between authors following $R$
  
  \[ PC_R(a_i, a_j) \]

- **Normalized Path Count**: Normalize path count following $R$ by the “degree” of authors
  
  \[ NPC_R(a_i, a_j) = \frac{PC_R(a_i, a_j) + PC_{R^{-1}}(a_j, a_i)}{PC_R(a_i, \cdot) + PC_R(\cdot, a_j)} \]

- **Random Walk**: Consider one way random walk following $R$
  
  \[ RW_R(a_i, a_j) = \frac{PC_R(a_i, a_j)}{PC_R(a_i, \cdot)} \]

- **Symmetric Random Walk**: Consider random walk in both directions
  
  \[ SRW_R(a_i, a_j) = RW_R(a_i, a_j) + RW_{R^{-1}}(a_j, a_i) \]
Performance Comparison: Homogeneous vs. Heterogeneous Topological Features

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Topological features</th>
<th>Accuracy</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP2hop</td>
<td>common neighbor</td>
<td>0.6053</td>
<td>0.6537</td>
</tr>
<tr>
<td></td>
<td>homogeneous PC</td>
<td>0.6433</td>
<td>0.7098</td>
</tr>
<tr>
<td></td>
<td>heterogeneous PC</td>
<td><strong>0.6545</strong></td>
<td><strong>0.7230</strong></td>
</tr>
<tr>
<td>HP3hop</td>
<td>common neighbor</td>
<td>0.6589</td>
<td>0.7078</td>
</tr>
<tr>
<td></td>
<td>homogeneous PC</td>
<td>0.6990</td>
<td>0.7998</td>
</tr>
<tr>
<td></td>
<td>rooted PageRank</td>
<td>0.6433</td>
<td>0.7098</td>
</tr>
<tr>
<td></td>
<td>heterogeneous PC</td>
<td><strong>0.7173</strong></td>
<td><strong>0.8158</strong></td>
</tr>
<tr>
<td>LP2hop</td>
<td>common neighbor</td>
<td>0.5995</td>
<td>0.6415</td>
</tr>
<tr>
<td></td>
<td>homogeneous PC</td>
<td>0.6154</td>
<td>0.6868</td>
</tr>
<tr>
<td></td>
<td>heterogeneous PC</td>
<td><strong>0.6300</strong></td>
<td><strong>0.6935</strong></td>
</tr>
<tr>
<td>LP3hop</td>
<td>common neighbor</td>
<td>0.6804</td>
<td>0.7195</td>
</tr>
<tr>
<td></td>
<td>homogeneous PC</td>
<td>0.6901</td>
<td>0.7883</td>
</tr>
<tr>
<td></td>
<td>heterogeneous PC</td>
<td><strong>0.7147</strong></td>
<td><strong>0.8046</strong></td>
</tr>
</tbody>
</table>

Notation: HP2hop: highly productive source authors with 2-hops reaching target authors.
Comparison among Different Topological Features

- Hybrid heterogeneous topological feature is the best

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

$PC1$: homogeneous common neighbor
$PCSum$: homogeneous path count
The Power of PathPredict: Experiment on DBLP

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

<table>
<thead>
<tr>
<th>Meta Path</th>
<th>p-value</th>
<th>significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A → P → P → A</td>
<td>0.0377</td>
<td>**</td>
</tr>
<tr>
<td>A ↔ P ↔ P → A</td>
<td>0.0077</td>
<td>***</td>
</tr>
<tr>
<td>A → P → V → P → A</td>
<td>1.2974e-174</td>
<td>****</td>
</tr>
<tr>
<td>A → P → A → P → A</td>
<td>1.1484e-126</td>
<td>****</td>
</tr>
<tr>
<td>A → P → T → P → A</td>
<td>3.4867e-51</td>
<td>****</td>
</tr>
<tr>
<td>A → P → P → P → A</td>
<td>0.7459</td>
<td></td>
</tr>
<tr>
<td>A ↔ P ↔ P ↔ P → A</td>
<td>0.0647</td>
<td>*</td>
</tr>
<tr>
<td>A → P → P ↔ P → A</td>
<td>9.7641e-11</td>
<td>****</td>
</tr>
<tr>
<td>A ↔ P ↔ P → P → A</td>
<td>0.0966</td>
<td>*</td>
</tr>
</tbody>
</table>

1. *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$; ****: $p < 0.001$

Social relations play more important role?

Evaluation of the prediction power of different meta-paths


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

- From “whether” to “when”
  - “Whether”: Will Jim rent the movie “Avatar” in Netflix?
  - “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”? $E(Y)$

Output: A distribution of time!

May provide useful information to supply chain management
The Relationship Building Time Prediction Model

- Solution
- Directly model relationship building time: \( P(Y=t) \)
- Geometric distribution, Exponential distribution, Weibull distribution
- Use generalized linear model
- Deal with censoring (relationship builds beyond the observed time interval)

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

Training Framework

T: Right Censoring

Generalized Linear Model under Weibull Distribution Assumption

\[
LL_W(\beta, \lambda) = \sum_{i=1}^{n} I_{y_i<T} \log \frac{\lambda y_i^{\lambda-1}}{e^{-\lambda x_i \beta}} - \sum_{i=1}^{n} \left( \frac{y_i}{e^{-x_i \beta}} \right)^\lambda
\]
Author Citation Time Prediction in DBLP

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

\[
\begin{align*}
A - P - T - P - A \\
A - P \leftrightarrow P \rightarrow P - A \\
A - P - A - P \rightarrow P - A \\
A - P - T - P - A - P \rightarrow P - A
\end{align*}
\]

- 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
Outline

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

- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks

- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - **ClusCite:** Citation Recommendation in Heterogeneous Information Networks

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

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

Collaborative filtering methods suffer from the data sparsity issue

A small set of users & items have a large number of ratings
Most users & items have a small number of ratings

Personalized recommendation with heterog. Networks [WSDM’14]
Different users may have different behaviors or preferences

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

Different users may be interested in the same movie for different reasons
Preference Propagation-Based Latent Features

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

Propagate user implicit feedback along each meta-path

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

Observation 1: Different meta-paths may have different importance

Global Recommendation Model

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

Observation 2: Different users may require different models

Personalized Recommendation Model

\[
\hat{r}_p(u_i, e_j) = \sum_{k=1}^{c} \text{sim}(C_k, u_i) \sum_{q=1}^{L} \theta_q^{\{k\}} \cdot \hat{U}_i(q) \hat{V}_j(q)^T
\]
Parameter Estimation

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

\[
\min_\Theta \quad - \sum_{u_i \in U} \sum_{(e_a > e_b) \in R_i} \ln \sigma(\hat{r}(u_i, e_a) - \hat{r}(u_i, e_b)) + \frac{\lambda}{2} ||\Theta||^2_2
\]

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

\[
\sigma(x) = \frac{1}{1+e^{-x}}.
\]

Soft cluster users with NMF + k-means

For each user cluster, learn one model with Eq. (3)

Generate personalized model for each user on the fly with Eq. (2)

Learning Personalized Recommendation Model
Experiments: Heterogeneous Network Modeling Enhances the Quality of Recommendation

- Datasets

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

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

HeteRec personalized recommendation (HeteRec-p) leads to the best recommendation.
Citation Recommendation: Given a manuscript (title, abstract and/or content) and its attributes (authors, target venues), suggest a small set of high quality references.

- X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, “ClusCite: Effective Citation Recommendation by Information Network-Based Clustering”, KDD’14

- Paper-specific recommendation model: heterogeneous network modeling
  - Captures paper-paper relevance of different semantics
  - Enables authority propagation between different types of objects

Are Google Scholar and PubMed satisfactory?
Each group follow distinct behavioral patterns and adopt different criteria in deciding relevance and authority of a candidate paper.
We explore the principle that: citations tend to be softly clustered into different interest groups, based on the heterogeneous network structures.

**Derive group membership for query manuscript**

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

**Paper-specific recommendation:** by integrating learned models of query’s related interest groups

**Phase I: Joint Learning (offline)**

**Phase II: Recommendation (online)**
## Performance Comparison on DBLP and PubMed

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

<table>
<thead>
<tr>
<th>Method</th>
<th>DBLP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>PubMed</th>
<th></th>
<th></th>
<th></th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@10</td>
<td>P@20</td>
<td>R@20</td>
<td>R@50</td>
<td>MRR</td>
<td>P@10</td>
<td>P@20</td>
<td>R@20</td>
<td>R@50</td>
<td>MRR</td>
</tr>
<tr>
<td>BM25</td>
<td>0.1260</td>
<td>0.0902</td>
<td>0.1431</td>
<td>0.2146</td>
<td>0.4107</td>
<td>0.1847</td>
<td>0.1349</td>
<td>0.1754</td>
<td>0.2470</td>
<td>0.4971</td>
</tr>
<tr>
<td>PopRank</td>
<td>0.0112</td>
<td>0.0098</td>
<td>0.0155</td>
<td>0.0308</td>
<td>0.0451</td>
<td>0.0438</td>
<td>0.0314</td>
<td>0.0402</td>
<td>0.0814</td>
<td>0.2012</td>
</tr>
<tr>
<td>TopicSim</td>
<td>0.0328</td>
<td>0.0273</td>
<td>0.0432</td>
<td>0.0825</td>
<td>0.1161</td>
<td>0.0761</td>
<td>0.0685</td>
<td>0.0855</td>
<td>0.1516</td>
<td>0.3254</td>
</tr>
<tr>
<td>Link-PLSA-LDA</td>
<td>0.1023</td>
<td>0.0893</td>
<td>0.1295</td>
<td>0.1823</td>
<td>0.3748</td>
<td>0.1439</td>
<td>0.1002</td>
<td>0.1589</td>
<td>0.2015</td>
<td>0.4079</td>
</tr>
<tr>
<td>L2-LR</td>
<td>0.2274</td>
<td>0.1677</td>
<td>0.2471</td>
<td>0.3547</td>
<td>0.4866</td>
<td>0.2527</td>
<td>0.1959</td>
<td>0.2504</td>
<td>0.3981</td>
<td>0.5308</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.2372</td>
<td>0.1799</td>
<td>0.2733</td>
<td>0.3621</td>
<td>0.4989</td>
<td>0.2534</td>
<td>0.1954</td>
<td>0.2499</td>
<td>0.382</td>
<td>0.5187</td>
</tr>
<tr>
<td>MixFea</td>
<td>0.2261</td>
<td>0.1689</td>
<td>0.2473</td>
<td>0.3636</td>
<td>0.5002</td>
<td>0.2699</td>
<td>0.2025</td>
<td>0.2519</td>
<td>0.4021</td>
<td>0.5041</td>
</tr>
<tr>
<td>ClusCite-Rel</td>
<td>0.2402</td>
<td>0.1872</td>
<td>0.2856</td>
<td>0.4015</td>
<td>0.5156</td>
<td>0.2766</td>
<td>0.2221</td>
<td>0.2753</td>
<td>0.4305</td>
<td>0.5524</td>
</tr>
<tr>
<td>ClusCite</td>
<td>0.2429</td>
<td>0.1958</td>
<td>0.2993</td>
<td>0.4279</td>
<td>0.5481</td>
<td>0.3019</td>
<td>0.2434</td>
<td>0.3129</td>
<td>0.4587</td>
<td>0.5787</td>
</tr>
</tbody>
</table>

- **20.19% improvement in Recall@20; 14.79% in MRR**, over the best performing compared method, on PubMed
Outline

- **Motivation:** Why Mining Information Networks?
- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks
- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - Task-Guided and Path-Augmented Heterogeneous Network Embedding
- **Summary**
T. Chen and Y. Sun, Task-guided and Path-augmented Heterogeneous Network Embedding for Author Identification, WSDM’17

Given an anonymized paper (often: double-blind review), with
- Venue (e.g., WSDM)
- Year (e.g., 2017)
- Keywords (e.g., “heterogeneous network embedding”)
- References (e.g., [Chen et al., IJCAI’16] )

Can we predict its authors?

Previous work on author identification: Feature engineering

New approach: Heterogeneous Network Embedding
- Embedding: automatically represent nodes into lower dimensional feature vectors
- Heterogeneous network embedding: Key challenge—select the best type of info due to the heterogeneity of the network
Task-Guided and Path-Augmented Embedding

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

- Consider the ego-network of $p$:
  - $X_p = (X^1_p, X^2_p, ..., X^T_p)$,
  - $T$: # types of nodes associated with paper type
  - $X^t_p$: the set of nodes with type $t$ associated with paper $p$
- $u_a$: embedding of author $a$
- $u_n$: embedding of node $n$
- $V_p$: embedding of paper $p$
- Weighted average of all the neighbors
- The score function between $p$ and $a$ is:
- Ranking-based objective: maximize the difference between authors $b$ and $a$: 
  $\max \left( 0, f(p, b) - f(p, a) + \xi \right)$
  $f(p, a) = u_a^T V_p = u_a^T \left( \sum_t w_t V_p^{(t)} \right)$
  $= u_a^T \left( \sum_t w_t \sum_{n \in X_p^{(t)}} u_n / |X_p^{(t)}| \right)$
Path-Augmented Embedding

- Why not just task-guided embedding?
  - Supervised labels expensive to obtain
  - The rich structured information of heterogeneous info-net is not fully explored

- Path-Augmented Embedding
  - Prepare meta-paths that are potentially related to the task
  - Apply general purpose embedding
  - For each meta-path-based relation
    - The probability of reaching node $j$ from node $i$ via meta-path $r$ via their embeddings
      \[
      P(j|i; r) = \frac{\exp(u_i^T u_j)}{\sum_{j' \in \text{DST}(r)} \exp(u_i^T u_{j'})}
      \]
  - Use negative sampling to approximate the distribution
  - Goal: maximize the likelihood to observing all the paths under each meta-path
The Joint Model and How to select meta-paths?

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

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

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

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

<table>
<thead>
<tr>
<th>Table 1: Node statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper</td>
</tr>
<tr>
<td>Train</td>
</tr>
<tr>
<td>Test</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Length-2 link statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>17M</td>
</tr>
</tbody>
</table>
Which Meta-Paths Are Selected?

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

Paths are sorted according to their performance

Only paths that can help improve the author identification task are shown

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

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

The performance of the combined model when meta-paths are added gradually
Author Identification: Performance Comparison

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

Table 5: Author identification performance comparison.

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

- **Performance over different groups of authors**
The Real Game and Case Study

Treat all the authors as candidates

Top ranked authors for queried paper
Outline

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

- **Part I:** Clustering and Ranking in Heterogeneous Information Networks
  - Clustering and Ranking in Information Networks
  - Similarity Search in Information Networks
  - User-Guided Meta-Path based Clustering in Heterogeneous Networks

- **Part II:** Classification and Prediction in Heterogeneous Information Networks
  - Classification of Heterogeneous Information Networks
  - Relationship Prediction in Heterogeneous Information Networks
  - Recommendation with Heterogeneous Information Networks
  - ClusCite: Citation Recommendation in Heterogeneous Information Networks

- Summary
Heterogeneous information networks are ubiquitous

Most datasets can be “organized” or “transformed” into “structured” multi-typed heterogeneous info. networks

Examples: DBLP, IMDB, Flickr, Google News, Wikipedia, ...

Surprisingly rich knowledge can be mined from structured heterogeneous info. networks

Clustering, ranking, classification, path prediction, ......

Knowledge is power, but knowledge is hidden in massive, but “relatively structured” nodes and links!

Key issue: Construction of trusted, semi-structured heterogeneous networks from unstructured data

From data to knowledge: Much more to be explored but heterogeneous network mining has shown high promise!
References

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