Finding the Information in Information Networks

Lise Getoor
University of Maryland, College Park

NASA GSFC Information Science & Technology Colloquium
February 6, 2008
Some Acknowledgements

- **Students:**
  - Indrajit Bhattacharya
  - Mustafa Bilgic
  - Rezarta Islamaj
  - Louis Licamele
  - Qing Lu
  - Galileo Namata
  - Vivek Sehgal
  - Prithvi Sen
  - Elena Zheleva

- **Collaborators:**
  - Chris Diehl
  - Hyunmo Kang
  - Lisa Singh
  - Tina Eliassi-Rad
  - Renee Miller
  - Ben Taskar
  - John Grant
  - Ben Shneiderman
  - Octavian Udrea

- **Funding Sources:**
  - NSF
  - KDD Program
  - Google
  - Microsoft Research
Graphs and Networks everywhere…

- The Web, social networks, communication networks, financial transaction networks, biological networks, etc.

Others available at Mark Newman’s gallery:
http://www-personal.umich.edu/~mejn/networks/
Wealth of Data

- Inundated with data describing networks
- But much of the data is
  - noisy and incomplete
  - at WRONG level of abstraction for analysis

Graph Identification

Graph Alignment
Graph Transformations

Data Graph $\Rightarrow$ Information Graph

1. Entity Resolution: mapping email addresses to people
2. Link Prediction: predicting social relationship based on communication
3. Collective Classification: labeling nodes in the constructed social network

HP Labs, Huberman & Adamic
Vision: Image Parsing

Graph Partitioning + Graph Matching

Bio: Graph Identification

Biological Networks: protein-protein, transcriptional regulation, signaling

Courtesy of Chris Wiggins, Columbia
Bio: Graph Alignment

Kelley, Brian P. et al. PNAS03
Roadmap

- The Problem

- The Components
  - Entity Resolution
  - Collective Classification
  - Link Prediction

- Putting It All Together

- Open Questions
Entity Resolution

- The Problem
- Relational Entity Resolution
- Algorithms
InfoVis Co-Author Network Fragment

before

after
The Entity Resolution Problem

Issues:
1. Identification
2. Disambiguation
Attribute-based Entity Resolution

Pair-wise classification

- “J Smith” vs. “James Smith”
  - Score: 0.8
- “Jim Smith” vs. “James Smith”
  - Score: 0.1
- “J Smith” vs. “James Smith”
  - Score: ?
- “John Smith” vs. “James Smith”
  - Score: 0.7
- “Jon Smith” vs. “James Smith”
  - Score: 0.05
- “Jonthan Smith” vs. “James Smith”
  - Score: ?

1. Choosing threshold: precision/recall tradeoff
2. Inability to disambiguate
3. Perform transitive closure?
Entity Resolution

- The Problem
- Relational Entity Resolution
- Algorithms
Relational Entity Resolution

- References not observed independently
  - Links between references indicate relations between the entities
  - Co-author relations for bibliographic data
  - To, cc: lists for email

- Use relations to improve identification and disambiguation

Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05
Relational Identification

Very similar names. Added evidence from shared co-authors
Relational Disambiguation

Very similar names but no shared collaborators
Collective Entity Resolution

One resolution provides evidence for another => joint resolution
Entity Resolution with Relations

- Naïve Relational Entity Resolution
  - Also compare attributes of related references
  - Two references have co-authors w/ similar names

- Collective Entity Resolution
  - Use *discovered entities* of related references
  - Entities cannot be identified independently
  - Harder problem to solve
Entity Resolution

- The Problem
- Relational Entity Resolution
- Algorithms
  - Relational Clustering (RC-ER)
    - Bhattacharya & Getoor, DMKD’04, Wiley’06, DE Bulletin’06, TKDD’07


Relational Clustering (RC-ER)
Relational Clustering (RC-ER)

P1
C. Walshaw
M. Cross
M. G. Everett
S. Johnson

P2
C. Walshaw
M. Cross
M. Everett
S. Johnson
K. McManus

P4
Alfred V. Aho
Jefferey D. Ullman
Stephen C. Johnson

P5
A. Aho
J. Ullman
S. Johnson
Cut-based Formulation of RC-ER

- Good separation of attributes
  - Many cluster-cluster relationships
    - Aho-Johnson1, Aho-Johnson2, Everett-Johnson1

- Worse in terms of attributes
  - Fewer cluster-cluster relationships
    - Aho-Johnson1, Everett-Johnson2
Objective Function

- **Minimize:**

\[ \sum_{i} \sum_{j} w_A \cdot \text{sim}_A(c_i, c_j) + w_R \cdot \text{sim}_R(c_i, c_j) \]

- **Greedy clustering algorithm:** merge cluster pair with max reduction in objective function

\[ \Delta(c_i, c_j) = w_A \cdot \text{sim}_A(c_i, c_j) + w_R \cdot |N(c_i) \cap N(c_j)| \]

- **Similarity of attributes**
- **Common cluster neighborhood**
Measures for Attribute Similarity

- Use best available measure for each attribute
  - Name Strings: *Soft TF-IDF, Levenstein, Jaro*
  - Textual Attributes: *TF-IDF*

- Aggregate to find similarity between clusters
  - Single link, Average link, Complete link
  - Cluster representative
Comparing Cluster Neighborhoods

- Consider neighborhood as multi-set

- Different measures of set similarity
  - Common Neighbors: Intersection size
  - Jaccard’s Coefficient: Normalize by union size
  - Adar Coefficient: Weighted set similarity
  - Higher order similarity: Consider neighbors of neighbors
Relational Clustering Algorithm

1. Find similar references using ‘blocking’
2. Bootstrap clusters using attributes and relations
3. Compute similarities for cluster pairs and insert into priority queue
4. Repeat until priority queue is empty
   5. Find ‘closest’ cluster pair
   6. Stop if similarity below threshold
   7. Merge to create new cluster
   8. Update similarity for ‘related’ clusters

- O(n k log n) algorithm w/ efficient implementation
Entity Resolution

- The Problem
- Relational Entity Resolution

Algorithms

- Relational Clustering (RC-ER)
- Probabilistic Model (LDA-ER)
  - *SIAM SDM’06, Best Paper Award*
- Experimental Evaluation
Discovering Groups from Relations

Parallel Processing Research Group

P1: C. Walshaw, M. Cross, M. G. Everett, S. Johnson
P2: C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus
P3: C. Walshaw, M. Cross, M. G. Everett

Bell Labs Group

P4: Alfred V. Aho, Stephen C. Johnson, Jeffrey D. Ullman
P5: A. Aho, S. Johnson, J. Ullman
P6: A. Aho, R. Sethi, J. Ullman
**Latent Dirichlet Allocation ER**

- Entity label \( a \) and group label \( z \) for each reference \( r \)
- \( \Theta \): ‘mixture’ of groups for each co-occurrence
- \( \Phi_z \): multinomial for choosing entity \( a \) for each group \( z \)
- \( V_a \): multinomial for choosing reference \( r \) from entity \( a \)
- Dirichlet priors with \( \alpha \) and \( \beta \)
Entity Resolution

- The Problem
- Relational Entity Resolution

**Algorithms**
- Relational Clustering (RC-ER)
- Probabilistic Model (LDA-ER)
- Experimental Evaluation
Evaluation Datasets

- CiteSeer
  - 1,504 citations to machine learning papers (Lawrence et al.)
  - 2,892 references to 1,165 author entities

- arXiv
  - 29,555 publications from High Energy Physics (KDD Cup’03)
  - 58,515 refs to 9,200 authors

- Elsevier BioBase
  - 156,156 Biology papers (IBM KDD Challenge ’05)
  - 831,991 author refs
  - Keywords, topic classifications, language, country and affiliation of corresponding author, etc
Baselines

- **A**: Pair-wise duplicate decisions w/ attributes only
  - **Names**: Soft-TFIDF with Levenstein, Jaro, Jaro-Winkler
  - **Other textual attributes**: TF-IDF

- **A***: Transitive closure over **A**

- **A+N**: Add attribute similarity of co-occurring refs

- **A+N***: Transitive closure over **A+N**

- Evaluate pair-wise decisions over references
- F1-measure (harmonic mean of precision and recall)
ER over Entire Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CiteSeer</th>
<th>arXiv</th>
<th>BioBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.980</td>
<td>0.976</td>
<td>0.568</td>
</tr>
<tr>
<td>A*</td>
<td>0.990</td>
<td>0.971</td>
<td>0.559</td>
</tr>
<tr>
<td>A+N</td>
<td>0.973</td>
<td>0.938</td>
<td>0.710</td>
</tr>
<tr>
<td>A+N*</td>
<td>0.984</td>
<td>0.934</td>
<td>0.753</td>
</tr>
<tr>
<td>RC-ER</td>
<td>0.995</td>
<td>0.985</td>
<td>0.818</td>
</tr>
<tr>
<td>LDA-ER</td>
<td>0.993</td>
<td>0.981</td>
<td>0.645</td>
</tr>
</tbody>
</table>

- RC-ER & LDA-ER outperform baselines in all datasets
- Collective resolution better than naïve relational resolution
- RC-ER and baselines require threshold as parameter
  - Best achievable performance over all thresholds
- Best RC-ER performance better than LDA-ER
- LDA-ER does not require similarity threshold

*Collective Entity Resolution In Relational Data*, Indrajit Bhattacharya and Lise Getoor, *ACM Transactions on Knowledge Discovery and Datamining*, 2007
### ER over Entire Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CiteSeer</th>
<th>arXiv</th>
<th>BioBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.980</td>
<td>0.976</td>
<td>0.568</td>
</tr>
<tr>
<td>A*</td>
<td>0.990</td>
<td>0.971</td>
<td>0.559</td>
</tr>
<tr>
<td>A+N</td>
<td>0.973</td>
<td>0.938</td>
<td>0.710</td>
</tr>
<tr>
<td>A+N*</td>
<td>0.984</td>
<td>0.934</td>
<td>0.753</td>
</tr>
<tr>
<td>RC-ER</td>
<td>0.995</td>
<td>0.985</td>
<td>0.818</td>
</tr>
<tr>
<td>LDA-ER</td>
<td>0.993</td>
<td>0.981</td>
<td>0.645</td>
</tr>
</tbody>
</table>

- **CiteSeer**: Near perfect resolution; 22% error reduction
- **arXiv**: 6,500 additional correct resolutions; 20% error reduction
- **BioBase**: Biggest improvement over baselines
Roadmap

- The Problem
- The Components
  - Entity Resolution
  - Collective Classification
  - Link Prediction
- Putting It All Together
- Open Questions
Collective Classification

- The Problem
- Collective Relational Classification
- Algorithms
Traditional Classification

Trainig Data

Test Data

Predict $Y$ based on attributes $X_i$
Correlations among linked instances
autocorrelation: labels are likely to be the same
homophily: similar nodes are more likely to be linked
Relational Classification (2)

Training Data

Test Data

Irregular graph structure
Relational Classification (3)

Links between training set & test set learning with partial labels or within network classification
The Problem

- Relational Classification: predicting the category of an object based on its attributes and its links and attributes of linked objects

- Collective Classification: jointly predicting the categories for a collection of connected, unlabelled objects

Neville & Jensen 00, Taskar, Abbeel & Koller 02, Lu & Getoor 03, Neville, Jensen & Galliger 04, Sen & Getoor TR07, Macskassy & Provost 07, Gupta, Diwam & Sarawagi 07, Macskassy 07, McDowell, Gupta & Aha 07
Feature Construction

- Objects are linked to a set of objects. To construct features from this set of objects, we need feature aggregation methods.

Perlich & Provost 03, 04, 05, Popescul & Ungar 03, 05, 06, Lu & Getoor 03, Gupta, Diwam & Sarawagi 07
Objects are linked to a set of objects. To construct features from this set of objects, we need feature aggregation methods.

Instances vs. generics

- Features may refer
  - explicitly to individuals
  - classes or generic categories of individuals

- On one hand, want to model that a particular individual may be highly predictive
- On the other hand, want models to generalize to new situations, with different individuals
Formulation

- Directed Model
  - Collection of Local Conditional Models
  - Inference Algorithms:
    - Iterative Classification Algorithm (ICA)
    - Gibbs Sampling (Gibbs)

- Undirected Model
  - (Pairwise) Markov Random Fields
  - Inference Algorithms:
    - Loopy Belief Propagation (LBP)
    - Gibbs Sampling
    - Mean Field Relaxation Labeling (MF)
Experimental Evaluation

- Comparison of Collective Classification Algorithms
  - Mean Field Relaxation Labeling (MF)
  - Iterative Classification Algorithm (ICA)
  - Loopy Belief Propagation (LBP)
  - Baseline: Content Only

- Datasets
  - Real Data
    - Bibliographic Data (Cora & Citeseer), WebKB, etc.
  - Synthetic Data
    - Data generator which can vary the class label correlations (homophily), attribute noise, and link density
## Results on Real Data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Cora</th>
<th>CiteSeer</th>
<th>WebKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Only</td>
<td>66.51</td>
<td>59.77</td>
<td>62.49</td>
</tr>
<tr>
<td>ICA</td>
<td>74.99</td>
<td>62.46</td>
<td>65.99</td>
</tr>
<tr>
<td>Gibbs</td>
<td>74.64</td>
<td>62.52</td>
<td>65.64</td>
</tr>
<tr>
<td>MF</td>
<td>79.70</td>
<td>62.91</td>
<td>65.65</td>
</tr>
<tr>
<td>LBP</td>
<td>82.48</td>
<td>62.64</td>
<td>65.13</td>
</tr>
</tbody>
</table>

Sen and Getoor, TR 07
Results clearly indicate that algorithms’ performance depends (in non-trivial ways) on structure.
Roadmap

- The Problem
- The Components
  - Entity Resolution
  - Collective Classification
  - Link Prediction
- Putting It All Together
- Open Questions
Link Prediction

- The Problem
- Predicting Relations
- Algorithms
  - Link Labeling
  - Link Ranking
  - Link Existence
Links in Data Graph

Communications Graph
Nodes: Network References
Edges: Communications Events

Node 1

chris@enron.com
chris37
555-450-0981

Email
IM
TXT

555-901-8812

Node 2

liz@enron.com
lizs22

555-901-8812
⇒ Links in Information Graph

Network Graph
Nodes: Entities
Edges: Social Relationships

Node 1

Manager

Chris

Family

Steve

Elizabeth

Tim
Predicting Relations

- Link Labeling
  - Can use similar approaches to collective classification

- Link Ranking
  - Many variations
    - Diehl, Namata, Getoor, *Relationship Identification for Social Network Discovery*, AAAI07
  - ‘Leak detection’
    - Carvalho & Cohen, SDM07

- Link Existence
  - HARD!
  - Huge class skew problem
  - Variations: Link completion, find missing link
Roadmap

- The Problem
- The Components
- Putting It All Together
- Open Questions
Putting Everything together....

Collaborative Social Network Discovery
Entity Resolution
Relationship Identification

Communications Graph
Nodes: Network References
Edges: Communications Events

Network Graph
Nodes: Entities
Edges: Social Relationships
Learning and Inference Hard

- Full Joint Probabilistic Representations
  - Directed vs. Undirected
  - Require sophisticated approximate inference algorithms
  - Tradeoff: hard inference vs. hard learning

- Combinations of Local Classifiers
  - Local classifiers choices
  - Require sophisticated updating and truth maintenance or global optimization via LP
  - Tradeoff: granularity vs. complexity

Many interesting and challenging research problems!!
Roadmap

- The Problem
- The Components
- Putting It All Together
- Open Questions
1. Query-time GI

- Instead of viewing as an off-line knowledge reformulation process

- Consider as real-time data gathering with
  - Varying resource constraints
  - Ability to reason about value of information
  - E.g., what attributes are most useful to acquire? which relationships? which will lead to the greatest reduction in ambiguity?

2. Visual Analytics for GI

- Combining rich statistical inference models with visual interfaces that support knowledge discovery and understanding.

- Because the statistical confidence we may have in any of our inferences may be low, it is important to be able to have a human in the loop, to understand and validate results, and to provide feedback.

- Especially for graph and network data, a well-chosen visual representation, suited to the inference task at hand, can improve the accuracy and confidence of user input.
D-Dupe: An Interactive Tool for Entity Resolution

http://www.cs.umd.edu/projects/linqs/ddupe
C-Group: A Visual Analytic Tool for Pairwise Analysis of Dynamic Group Membership

http://www.cs.umd.edu/projects/linqs/cgroup
HOMER: Tool for Ontology Alignment

http://www.cs.umd.edu/projects/linqs/iliads
SplicePort: Motif Explorer

Islamaj Dogan, Getoor, Wilbur, Mount, Nucleic Acids Research, 2007

http://www.cs.umd.edu/projects/spliceport
3. GI & Privacy

- Obvious privacy concerns that need to be taken into account!!!

- A better theoretical understanding of when graph identification is feasible will also help us understand what must be done to maintain privacy of graph data.

- … Graph Re-Identification: study of anonymization strategies such that the information graph cannot be inferred from released data graph.
Link Re-Identification

Disease data

has hypertension

Communication data

father-of

Robert Lady

call

Search data

Query 1: “how to tell if your wife is cheating on you”

Query 2: “myrtle beach golf course job listings”

Social network data

friends

Zheleva and Getoor, Preserving the Privacy of Sensitive Relationships in Graph Data, PINKDD 2007
Graph Identification can be seen as a process of knowledge reformulation.

In the context where we have some statistical information to help us learn which reformulations are more promising than others.

Inference is the process of transferring the learned knowledge to new situations.
Statistical Relational Learning (SRL)

○ Methods that combine expressive knowledge representation formalisms such as relational and first-order logic with principled probabilistic and statistical approaches to inference and learning

○ Hendrik Blockeel, Mark Craven, James Cussens, Bruce D’Ambrosio, Luc De Raedt, Tom Dietterich, Pedro Domingos, Saso Dzeroski, Peter Flach, Rob Holte, Manfred Jaeger, David Jensen, Kristian Kersting, Heikki Mannila, Andrew McCallum, Tom Mitchell, Ray Mooney, Stephen Muggleton, Kevin Murphy, Jen Neville, David Page, Avi Pfeffer, Claudia Perlich, David Poole, Foster Provost, Dan Roth, Stuart Russell, Taisuke Sato, Jude Shavlik, Ben Taskar, Lyle Ungar and many others

Dagstuhl April 2007
Conclusion

- Relationships matter!
- Structure matters!

- Killer Apps:
  - Biology: Biological Network Analysis
  - Computer Vision: Human Activity Recognition
  - Information Extraction: Entity Extraction & Role labeling
  - Semantic Web: Ontology Alignment and Integration
  - Personal Information Management: Intelligent Desktop

- While there are important pitfalls to take into account (confidence and privacy), there are many potential benefits and payoffs!
Thanks!

http://www.cs.umd.edu/linqs

Work sponsored by the National Science Foundation, KDD program, National Geospatial Agency, Google and Microsoft