Collective Graph Identification

Lise Getoor
University of Maryland, College Park

Joint work with Galileo Namata

DIMACS/CCICADA Workshop on Data Quality Metrics
Feb 3, 2011
Motivation: Network Analysis

Who are the “central” individuals?

What are the communities?

What are the common interaction patterns/motifs?
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 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

Communications Graph

Nodes: Network References
Edges: Communications Events

Network Graph

Nodes: Entities
Edges: Social Relationships

HP Labs, Huberman & Adamic
Overview: Graph Identification

- Many real world datasets are relational in nature
  - Social Networks – people related by relationships like friendship, family, enemy, boss_of, etc.
  - Biological Networks – proteins are related to each other based on if they physically interact
  - Communication Networks – email addresses related by who emailed whom
  - Citation Networks – papers linked by which other papers they cite, as well as who the authors are
- However, the observations describing the data are noisy and incomplete
- **graph identification problem** is to infer the appropriate information graph from the data graph
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
The Entity Resolution Problem

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

Pair-wise classification

“J Smith”  “James Smith”  ?
“Jim Smith”  “James Smith”  0.8
“J Smith”  “James Smith”  ?
“John Smith”  “James Smith”  0.1
“Jon Smith”  “James Smith”  0.7
“Jonthan Smith”  “James Smith”  0.05

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 with 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


P1: “JOSTLE: Partitioning of Unstructured Meshes for Massively Parallel Machines”, C. Walshaw, M. Cross, M. G. Everett, **S. Johnson**


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
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:**

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

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

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

- Weight for attributes
- Similarity of attributes
- Weight for relations
- Similarity based on relational edges between \(c_i\) and \(c_j\)
- 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
5. 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

- Stephen P. Johnson
- Chris Walshaw
- Kevin McManus
- Mark Cross
- Martin Everett

Bell Labs Group

- Stephen C. Johnson
- Alfred V. Aho
- Ravi Sethi
- Jeffrey D. Ullman

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
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

- 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

**Algorithm**

<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><strong>0.995</strong></td>
<td><strong>0.985</strong></td>
<td><strong>0.818</strong></td>
</tr>
<tr>
<td>LDA-ER</td>
<td>0.993</td>
<td>0.981</td>
<td>0.645</td>
</tr>
</tbody>
</table>

*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><strong>0.995</strong></td>
<td><strong>0.985</strong></td>
<td><strong>0.818</strong></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

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)

Training Data

Test Data

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
Example: Linked Bibliographic Data

Objects:
- Papers
- Authors
- Institutions

Links:
- Citation
- Co-Citation
- Author-of
- Author-affiliation

Labels:
- 
- 
- 

Objects: Papers
Authors
Institutions

Links:
Citation
Co-Citation
Author-of
Author-affiliation
Feature Construction

- Objects are linked to a set of objects. To construct features from this set of objects, we need feature aggregation methods:
  
  Kramer, Lavrac & Flach 01, Perlich & Provost 03, 04, 05, Popescul & Ungar 03, 05, 06, Lu & Getoor 03, Gupta, Diwam & Sarawagi 07
Formulation

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

- Global Models
  - (Pairwise) Markov Random Fields
  - Inference Algorithms:
    - Loopy Belief Propagation (LBP)
    - Mean Field Relaxation Labeling (MF)
Learn model from fully labeled training set
ICA: Inference (1)

Step 1: Bootstrap using object attributes only
ICA: Inference (2)

Step 2: Iteratively update the category of each object, based on linked object’s categories
Experimental Evaluation

- Comparison of Collective Classification Algorithms
  - Mean Field Relaxation Labeling (MF)
  - Iterative Classification Algorithm (ICA)
  - Gibbs Sampling (Gibbs)
  - 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 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><strong>62.91</strong></td>
<td>65.65</td>
</tr>
<tr>
<td>LBP</td>
<td><strong>82.48</strong></td>
<td>62.64</td>
<td>65.13</td>
</tr>
</tbody>
</table>
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

Node 1

Email
chris@enron.com
chris37
555-450-0981

IM
lizs22

TXT
555-901-8812

Node 2

Email
liz@enron.com
⇒ Links in Information Graph

Network Graph

Nodes: Entities
Edges: Social Relationships

Manager

Family

Chris

Steve

Elizabeth

Tim
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
Graph Identification

- **Goal:**
  - Given an **input graph** infer a **complete** and **clean output graph**

- Three major components:
  - **Entity Resolution (ER):** Infer the set of nodes
  - **Collective Classification (CC):** Infer the node labels
  - **Link Prediction (LP):** Infer the set of edges

- **Problem:** The components are intra and inter-dependent
Dependencies

- Intra-dependent
  - Two nodes more likely to be co-referent if their neighbors are co-referent
  - Two nodes are more likely to be linked if they link to common nodes
  - Label of a node depends on the labels of related nodes

- Inter-dependent
  - Two nodes are more likely to be co-referent if they have the same *inferred* label
  - Two nodes are more likely to be linked depending on their *inferred* labels
  - Label of a node depends on *inferred* linked nodes
Classifiers

- **Base Classifiers**
  - Can use any conditional model as base classifier (i.e., logistic regression, decision trees, SVMs, naïve Bayes, etc.)
  - Local Classifiers – use only local attribute info for a node or edge
  - Relational Classifiers – can use info from relational neighborhood
Classifiers

- **Base Classifiers**
  - Can use any conditional model as base classifier (i.e., logistic regression, decision trees, SVMs, naïve Bayes, etc.)
  - Local Classifiers – use only local attribute info for a node or edge
  - Relational Classifiers – can use info from relational neighborhood

- **Collective classifiers**
  - Use local classifiers to bootstrap classification process
  - Iteratively apply relational classifiers
Classifiers

- Base Classifiers
  - Can use any conditional model as base classifier (i.e., logistic regression, decision trees, SVMs, naïve Bayes, etc.)
  - Local Classifiers – use only local attribute info for a node or edge
  - Relational Classifiers – can use info from relational neighborhood

- Collective classifiers
  - Use local classifiers to bootstrap classification process
  - Iteratively apply relational classifiers

- Coupled Classifiers
  - Apply the collective classifiers in order such that collective classifiers can use the predictions of earlier classifiers when computing relational features
    - Pipeline – Apply the components one at a time, in a particular sequence
    - Coupled Collective Classifiers – Apply components iteratively
Coupled Collective Classification (C³) Algorithm

- Focus is on coupling the inference of the three components using conditional models
- Conditional models applied in two phases
  - Phase 1: Local models using only local features
    - Bootstraps the process
  - Phase 2: Relational models using intra- and inter-relational features
    - Infer assignments using local and intra- and inter-relational information
- Cyclic dependencies handled by iteratively apply relational models
C³ Variants

- Capture more dependencies can also mean introducing more channels for error propagation

- Variant 1: Confidence-Based Inference
  - Some predictions are more confident than others
  - Commit more confident predictions earlier

- Variant 2: Stacked Learning (Kou & Cohen 07)
  - Instead of using the true assignments for relational features during training, use inferred assignments
Datasets:

- Citation Networks
  - Citeseer – 3312 paper nodes, 4732 citation edges, 6 possible labels
  - Cora – 2708 paper nodes, 5428 citation edges, 7 possible labels
- Partitioned to three disjoint networks and created noisy versions of each; varied amount of noise (Low, Medium, High)
- Given noisy network, infer the original network

Conditional models: linear SVM

Evaluate average F1 performance over ER, LP, CC
Algorithms

- **Baselines:**
  - LOCAL: apply only the local models
  - INTRA: apply relational classifiers using only intra-relational features

- **PIPELINE:** apply collective classifiers for each component in the pipeline

- **C³ Variants:**
  - C³: the basic algorithm
  - C³+C: C³ using confidence based inference
  - C³+S: C³ using stacking
  - C³+SC: C³ using stacking and confidence based inference

- **Gibbs:** apply pseudo-Gibbs sampling over the conditional models
General Trends: Citeseer

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.999</td>
<td>0.853</td>
<td>0.656</td>
<td>0.836</td>
<td>0.993</td>
<td>0.707</td>
<td>0.633</td>
<td>0.778</td>
<td>0.954</td>
<td>0.650</td>
<td>0.602</td>
<td>0.735</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRA</td>
<td>0.999</td>
<td>0.852</td>
<td>0.660</td>
<td>0.837</td>
<td>0.995</td>
<td>0.706</td>
<td>0.639</td>
<td>0.780</td>
<td>0.956</td>
<td>0.647</td>
<td>0.621</td>
<td>0.741</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELN</td>
<td>0.999</td>
<td>0.906</td>
<td>0.684</td>
<td>0.863</td>
<td>0.995</td>
<td>0.851</td>
<td>0.675</td>
<td>0.840</td>
<td>0.956</td>
<td>0.780</td>
<td>0.634</td>
<td>0.790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENL</td>
<td>0.999</td>
<td>0.916</td>
<td>0.679</td>
<td>0.865</td>
<td>0.995</td>
<td>0.872</td>
<td>0.665</td>
<td>0.844</td>
<td>0.956</td>
<td>0.808</td>
<td>0.633</td>
<td>0.799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>0.999</td>
<td>0.852</td>
<td>0.678</td>
<td>0.843</td>
<td>0.994</td>
<td>0.706</td>
<td>0.666</td>
<td>0.789</td>
<td>0.953</td>
<td>0.647</td>
<td>0.625</td>
<td>0.742</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNE</td>
<td>0.999</td>
<td>0.852</td>
<td>0.663</td>
<td>0.838</td>
<td>0.994</td>
<td>0.706</td>
<td>0.643</td>
<td>0.781</td>
<td>0.953</td>
<td>0.647</td>
<td>0.608</td>
<td>0.736</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEL</td>
<td>0.999</td>
<td>0.916</td>
<td>0.660</td>
<td>0.858</td>
<td>0.993</td>
<td>0.872</td>
<td>0.639</td>
<td>0.835</td>
<td>0.959</td>
<td>0.812</td>
<td>0.621</td>
<td>0.797</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLE</td>
<td>0.999</td>
<td>0.863</td>
<td>0.660</td>
<td>0.840</td>
<td>0.993</td>
<td>0.754</td>
<td>0.639</td>
<td>0.795</td>
<td>0.955</td>
<td>0.694</td>
<td>0.621</td>
<td>0.757</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gibbs</td>
<td>0.999</td>
<td>0.924</td>
<td>0.676</td>
<td>0.866</td>
<td>0.942</td>
<td>0.891</td>
<td>0.666</td>
<td>0.833</td>
<td>0.613</td>
<td>0.840</td>
<td>0.621</td>
<td>0.691</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³</td>
<td>0.999</td>
<td>0.917</td>
<td>0.683</td>
<td>0.866</td>
<td>0.995</td>
<td>0.870</td>
<td>0.670</td>
<td>0.845</td>
<td>0.959</td>
<td>0.809</td>
<td>0.638</td>
<td>0.802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+C</td>
<td>0.999</td>
<td>0.917</td>
<td>0.684</td>
<td>0.867</td>
<td>0.995</td>
<td>0.872</td>
<td>0.667</td>
<td>0.845</td>
<td>0.957</td>
<td>0.810</td>
<td>0.634</td>
<td>0.800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+S</td>
<td>0.999</td>
<td>0.917</td>
<td>0.700</td>
<td>0.872</td>
<td>0.996</td>
<td>0.868</td>
<td>0.684</td>
<td>0.849</td>
<td>0.965</td>
<td>0.775</td>
<td>0.651</td>
<td>0.797</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+SC</td>
<td>0.999</td>
<td>0.918</td>
<td>0.701</td>
<td>0.873</td>
<td>0.995</td>
<td>0.869</td>
<td>0.681</td>
<td>0.848</td>
<td>0.962</td>
<td>0.773</td>
<td>0.654</td>
<td>0.797</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Capturing more dependencies result in improved performance
- C³ algorithm generally the best performing for each task and overall
### General Trends: Cora

<table>
<thead>
<tr>
<th></th>
<th>Low Noise</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>High Noise</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td>ER (f1)</td>
<td>LP (f1)</td>
<td>NL (f1)</td>
<td>Avg.</td>
<td>ER (f1)</td>
<td>LP (f1)</td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.983</td>
<td>0.816</td>
<td>0.719</td>
<td>0.839</td>
<td>0.950</td>
<td>0.702</td>
<td>0.682</td>
<td>0.778</td>
<td>0.910</td>
<td>0.483</td>
<td>0.613</td>
<td>0.669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRA</td>
<td>0.975</td>
<td>0.812</td>
<td>0.735</td>
<td>0.841</td>
<td>0.938</td>
<td>0.694</td>
<td>0.694</td>
<td>0.775</td>
<td>0.886</td>
<td>0.470</td>
<td>0.657</td>
<td>0.671</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELN</td>
<td>0.975</td>
<td>0.906</td>
<td>0.774</td>
<td>0.885</td>
<td>0.938</td>
<td>0.867</td>
<td>0.722</td>
<td>0.842</td>
<td>0.886</td>
<td>0.762</td>
<td>0.657</td>
<td>0.768</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENL</td>
<td>0.975</td>
<td>0.918</td>
<td>0.765</td>
<td>0.886</td>
<td>0.938</td>
<td>0.882</td>
<td>0.728</td>
<td>0.849</td>
<td>0.886</td>
<td>0.774</td>
<td>0.663</td>
<td>0.774</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>0.972</td>
<td>0.812</td>
<td>0.764</td>
<td>0.849</td>
<td>0.932</td>
<td>0.694</td>
<td>0.711</td>
<td>0.779</td>
<td>0.892</td>
<td>0.470</td>
<td>0.632</td>
<td>0.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LNE</td>
<td>0.974</td>
<td>0.812</td>
<td>0.739</td>
<td>0.842</td>
<td>0.937</td>
<td>0.694</td>
<td>0.674</td>
<td>0.768</td>
<td>0.895</td>
<td>0.470</td>
<td>0.610</td>
<td>0.659</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEL</td>
<td>0.977</td>
<td>0.916</td>
<td>0.735</td>
<td>0.876</td>
<td>0.943</td>
<td>0.881</td>
<td>0.694</td>
<td>0.839</td>
<td>0.897</td>
<td>0.806</td>
<td>0.657</td>
<td>0.787</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLE</td>
<td>0.975</td>
<td>0.837</td>
<td>0.735</td>
<td>0.849</td>
<td>0.942</td>
<td>0.769</td>
<td>0.694</td>
<td>0.802</td>
<td>0.894</td>
<td>0.628</td>
<td>0.657</td>
<td>0.726</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gibbs</td>
<td>0.943</td>
<td>0.932</td>
<td>0.772</td>
<td>0.882</td>
<td>0.742</td>
<td><strong>0.895</strong></td>
<td>0.690</td>
<td>0.776</td>
<td>0.365</td>
<td><strong>0.835</strong></td>
<td>0.620</td>
<td>0.607</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³</td>
<td>0.977</td>
<td><strong>0.919</strong></td>
<td>0.767</td>
<td>0.888</td>
<td>0.943</td>
<td>0.880</td>
<td>0.724</td>
<td>0.849</td>
<td>0.892</td>
<td>0.792</td>
<td>0.663</td>
<td>0.782</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+C</td>
<td>0.976</td>
<td>0.918</td>
<td>0.772</td>
<td>0.899</td>
<td>0.943</td>
<td>0.882</td>
<td>0.716</td>
<td>0.847</td>
<td>0.894</td>
<td>0.797</td>
<td>0.660</td>
<td>0.784</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+S</td>
<td><strong>0.984</strong></td>
<td>0.915</td>
<td><strong>0.790</strong></td>
<td><strong>0.896</strong></td>
<td><strong>0.961</strong></td>
<td><strong>0.882</strong></td>
<td><strong>0.767</strong></td>
<td><strong>0.870</strong></td>
<td><strong>0.921</strong></td>
<td>0.809</td>
<td><strong>0.684</strong></td>
<td><strong>0.804</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C³+SC</td>
<td>0.983</td>
<td>0.916</td>
<td>0.786</td>
<td>0.895</td>
<td>0.962</td>
<td>0.880</td>
<td>0.759</td>
<td>0.867</td>
<td>0.919</td>
<td>0.802</td>
<td>0.682</td>
<td>0.801</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Capturing more dependencies result in improved performance
- C³ algorithm generally the best performing for each task and overall
Improvements are Significant

**Citeseer**

<table>
<thead>
<tr>
<th></th>
<th>LOCA</th>
<th>INTRA</th>
<th>ELN</th>
<th>ENL</th>
<th>LEN</th>
<th>LNE</th>
<th>NEL</th>
<th>NLE</th>
<th>Gibbs</th>
<th>C³</th>
<th>C³+C</th>
<th>C³+S</th>
<th>C³+SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTRA</td>
<td>4</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ELN</td>
<td>8</td>
<td>5</td>
<td>--</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENL</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>--</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LEN</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>--</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LNE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NEL</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>--</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NLE</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gibbs</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C³</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>--</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C³+C</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>--</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C³+S</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>--</td>
<td>0</td>
</tr>
<tr>
<td>C³+SC</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>--</td>
</tr>
</tbody>
</table>

- Performed paired t-test (> 95%) between all algorithms pairs
- $C³$ significantly outperforms other models in most cases
## Improvements are Significant

### Cora

<table>
<thead>
<tr>
<th>LOCA</th>
<th>INTRA</th>
<th>ELN</th>
<th>ENL</th>
<th>LEN</th>
<th>LNE</th>
<th>NEL</th>
<th>NLE</th>
<th>Gibbs</th>
<th>C3</th>
<th>C3+C</th>
<th>C3+S</th>
<th>C3+SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL</td>
<td>--</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTRA</td>
<td>1</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ELN</td>
<td>5</td>
<td>4</td>
<td>--</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENL</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>--</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LEN</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LNE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>--</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NEL</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>--</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NLE</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>--</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gibbs</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>--</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Performed paired t-test (> 95%) between all algorithms pairs
- C^3 significantly outperforms other models in most cases
Summary so far…

- Graph identification is a general framework for dealing with noisy structured data.
- Here, we saw a preliminary approach based on collections of local classifiers.
- Many open issues….
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?

- Query-time Entity Resolution, Bhattacharya & Getoor, Journal of Artificial Intelligence Research, 2007

- Active Learning for Networked Data, Bilgic, Mihalkova & Getoor, International Conference on Machine Learning, 2010
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.
Three Tools

D-Dupe

C-Group

G-View
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
Some relevant work

Preserving the Privacy of Sensitive Relationships in Graph Data, Zheleva and Getoor, PINKDD 07


To Join or Not to Join: the Illusion of Privacy in Online Social Networks, Zheleva and Getoor, WWW 2009
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.
Conclusion

- Graph Identification
  - can be seen as a process of data cleaning and knowledge reformulation
  - In the context where we have some relational information that tells us about the structure of the graph that helps us to define features and statistical information to help us learn which reformulations are more promising than others

- 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, Microsoft and Yahoo!

KDD Program