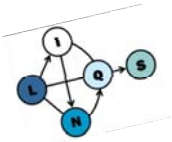
Collective Entity Resolution

Lise Getoor
University of Maryland, College Park

Joint work with Indrajit Bhattacharya

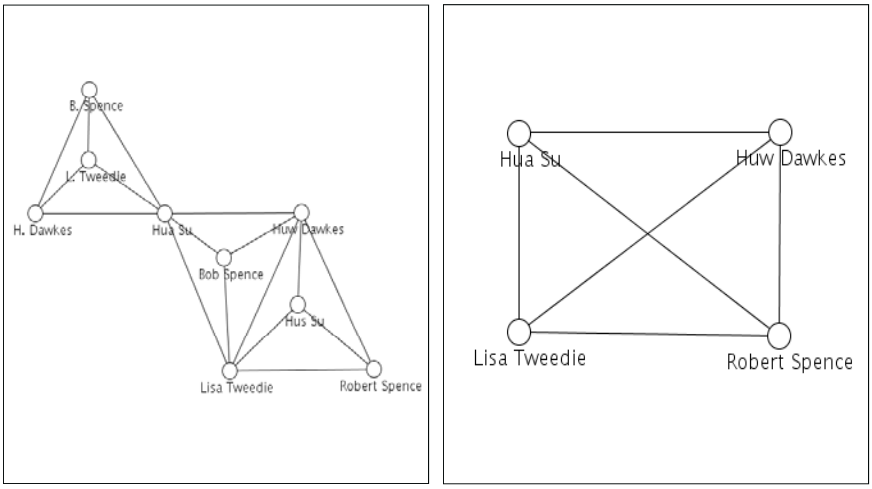
US Census Bureau Conference on
Person Validation and Entity Resolution
May 23, 2011



Entity Resolution

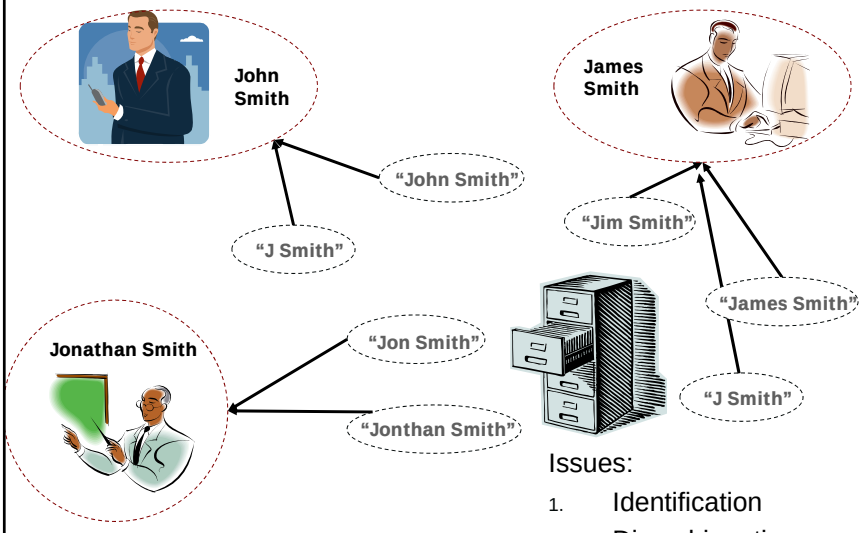
- **The Problem**
- Relational Entity Resolution
- Algorithms
- Open Issues

InfoVis Co-Author Network Fragment



before after

The Entity Resolution Problem

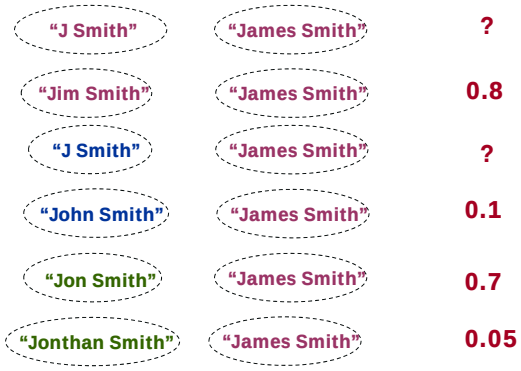


Issues:

1. Identification
2. Disambiguation

Attribute-based Entity Resolution

Pair-wise classification



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

Entity Resolution

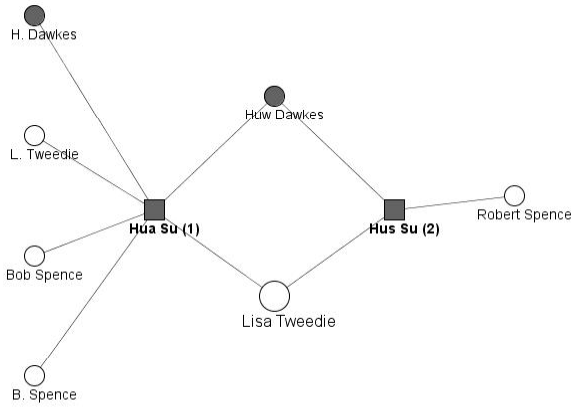
- The Problem
- **Relational Entity Resolution**
- Algorithms
- Open Issues

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

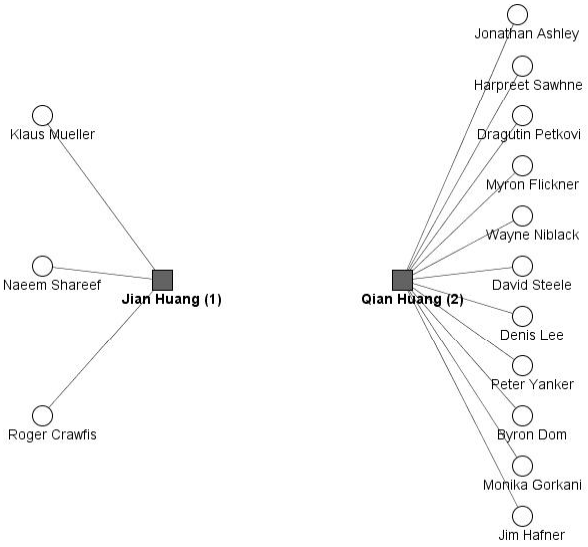
Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05, many others….

Relational Identification



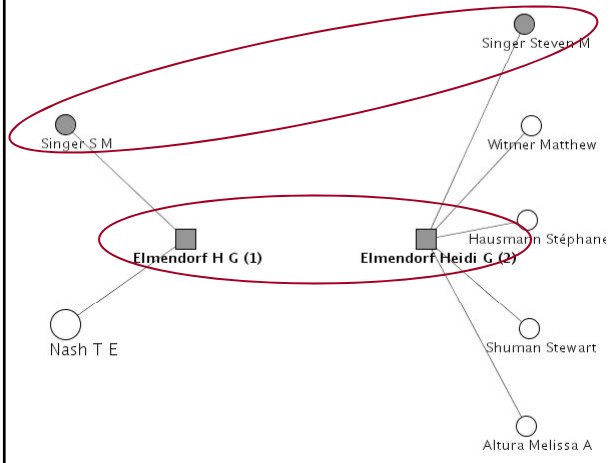
Very similar names.
Added evidence from shared co-authors

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*
- Open Issues

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

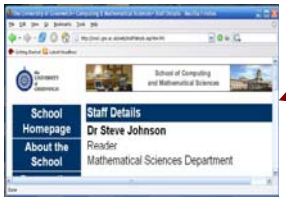
P2: "*Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies*", C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus

P3: "*Dynamic Mesh Partitioning: A Unied Optimisation and Load-Balancing Algorithm*", C. Walshaw, M. Cross, M. G. Everett

P4: "*Code Generation for Machines with Multiregister Operations*", Alfred V. Aho, Stephen C. Johnson, Jefferey D. Ullman

P5: "*Deterministic Parsing of Ambiguous Grammars*", A. Aho, S. Johnson, J. Ullman

P6: "*Compilers: Principles, Techniques, and Tools*", A. Aho, R. Sethi, J. Ullman



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

P2: "*Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies*", C. Walshaw, M. Cross, M. G. Everett, **S. Johnson**, K. McManus

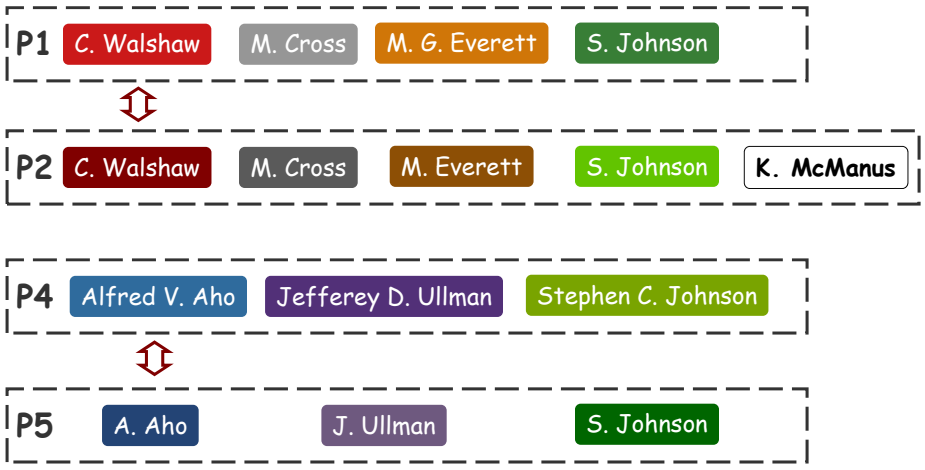
P3: "*Dynamic Mesh Partitioning: A Unied Optimisation and Load-Balancing Algorithm*", C. Walshaw, M. Cross, M. G. Everett

P4: "*Code Generation for Machines with Multiregister Operations*", Alfred V. Aho, **Stephen C. Johnson**, Jefferey D. Ullman

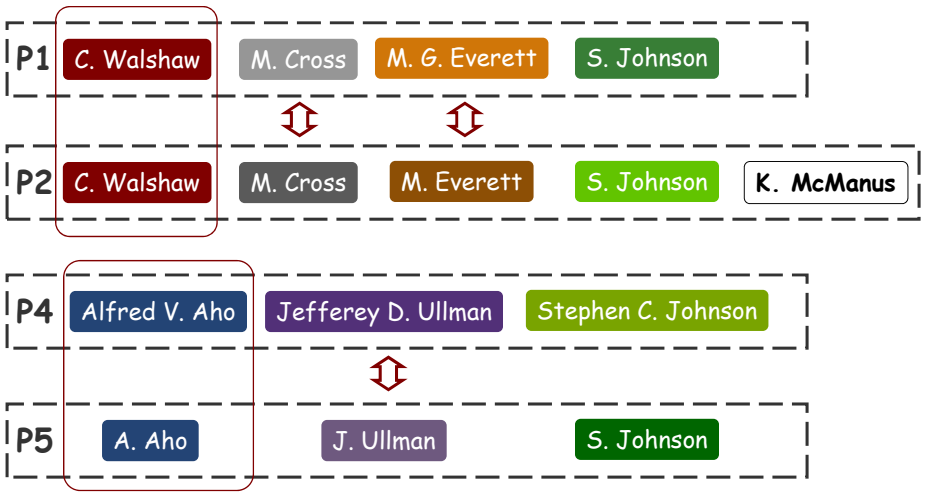
P5: "*Deterministic Parsing of Ambiguous Grammars*", A. Aho, **S. Johnson**, J. Ullman

P6: "*Compilers: Principles, Techniques, and Tools*", A. Aho, R. Sethi, J. Ullman

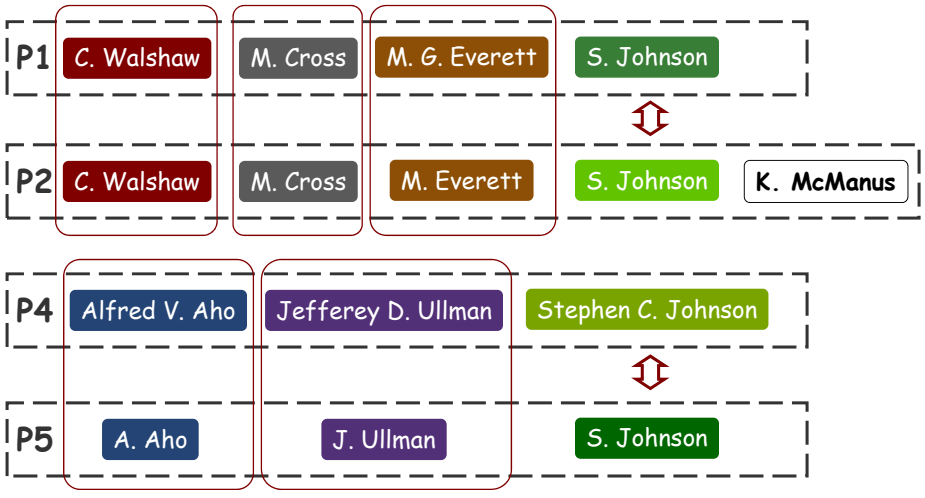
Relational Clustering (RC-ER)



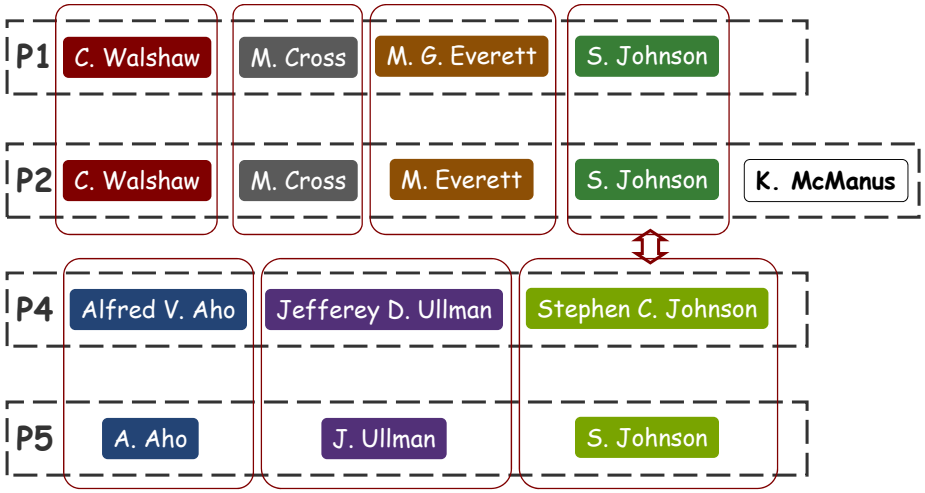
Relational Clustering (RC-ER)



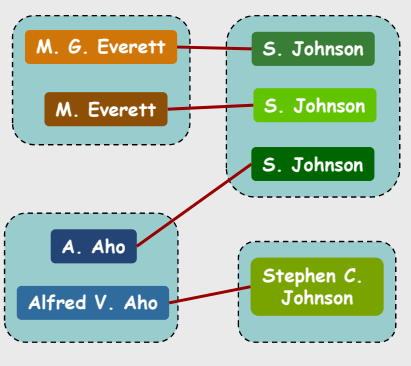
Relational Clustering (RC-ER)



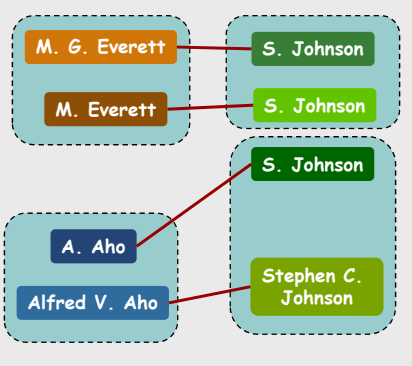
Relational Clustering (RC-ER)



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:**

$$\sum_i \sum_j w_A sim_A(c_i, c_j) + w_R sim_R(c_i, c_j)$$

weight for attributes | similarity of attributes | weight for relations | Similarity based on relational edges between c_i and c_j

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

$$\Delta(c_i, c_j) = w_A sim_A(c_i, c_j) + w_R(|N(c_i) \cap N(c_j)|)$$

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

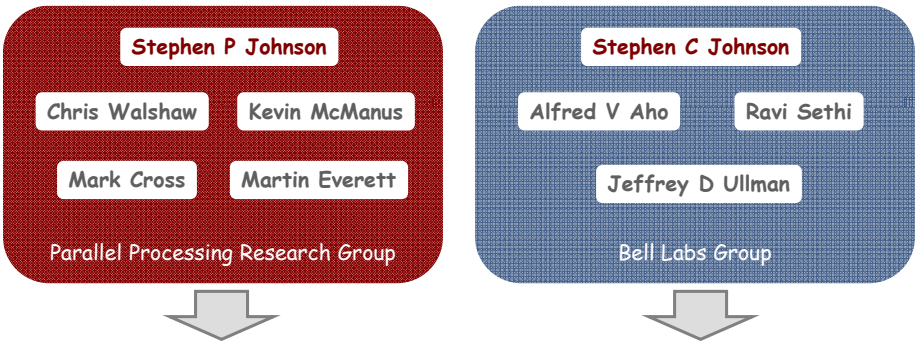
4. Repeat until priority queue is empty
5. Find ‘closest’ cluster pair
6. Stop if similarity below threshold
7. Merge to create new cluster
8. Update similarity for ‘related’ clusters

- 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
- Open Issues

Discovering Groups from Relations



P1: C. Walshaw, M. Cross, M. G. Everett, **S. Johnson**

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

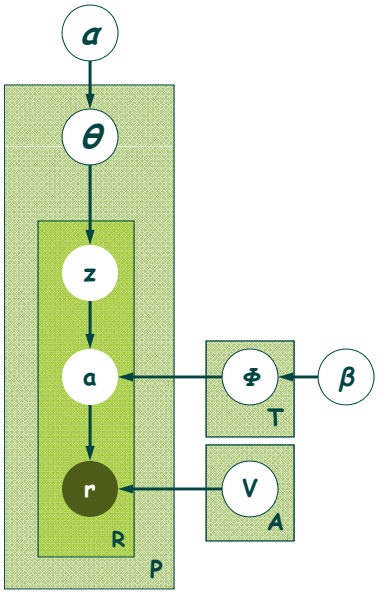
P3: C. Walshaw, M. Cross, M. G. Everett

P4: Alfred V. Aho, **Stephen C. Johnson**, Jefferey 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***
- ***Θ*:** 'mixture' of groups for each co-occurrence
- Φ_z**:** multinomial for choosing entity ***a*** for each group ***z***
- V_a**:** multinomial for choosing reference ***r*** from entity ***a***
- Dirichlet priors with α and β

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

Algorithm	CiteSeer	arXiv	BioBase
A	0.980	0.976	0.568
A*	0.990	0.971	0.559
A+N	0.973	0.938	0.710
A+N*	0.984	0.934	0.753
RC-ER	**0.995**	**0.985**	**0.818**
LDA-ER	0.993	0.981	0.645

- 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

Algorithm	CiteSeer	arXiv	BioBase
A	0.980	0.976	0.568
A*	0.990	0.971	0.559
A+N	0.973	0.938	0.710
A+N*	0.984	0.934	0.753
RC-ER	**0.995**	**0.985**	**0.818**
LDA-ER	0.993	0.981	0.645

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

Entity Resolution

- The Problem
- Relational Entity Resolution
- Algorithms
- **Open Issues**

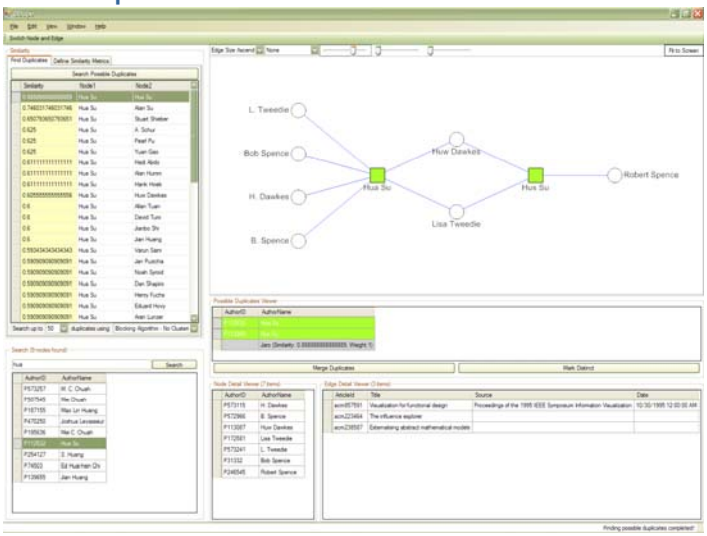
1. Query-time ER

- Instead of viewing as an off-line data cleaning 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 ER

- 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

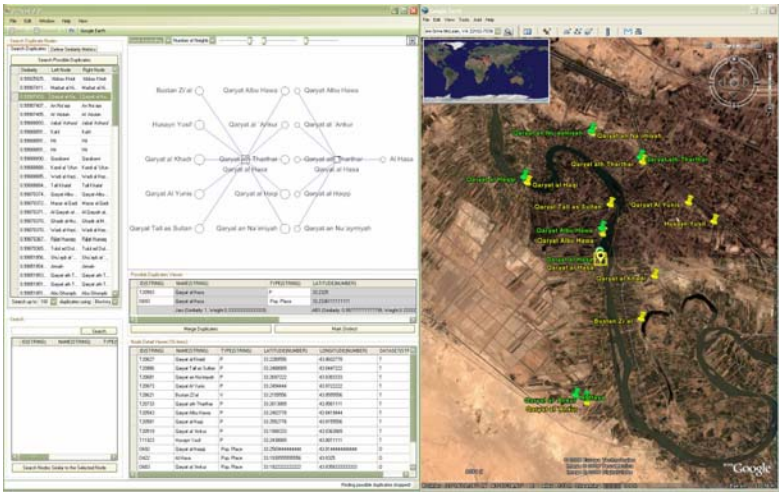
D-Dupe: An Interactive Tool for ER



Kang, Getoor, Shneiderman, Bilgic, Licamele, TCGV 08

http://www.cs.umd.edu/projects/lings/ddupe

GeoDDupe: Tool for Interactive Entity Resolution in Geospatial Data

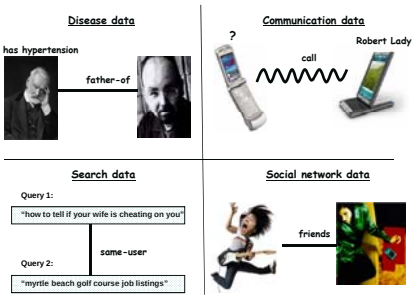


Kang, Sehgal, Getoor, IV 07
http://www.cs.umd.edu/projects/lings/geoddupe

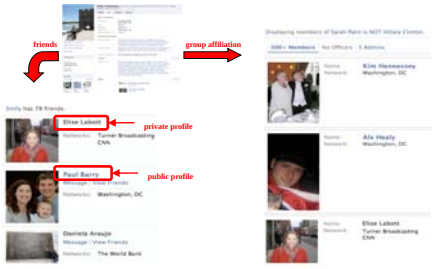
3. ER & Privacy

- Obvious privacy concerns that need to be taken into account!!!
- A better theoretical understanding of when person identification is feasible will also help us understand what must be done to maintain privacy of graph data
- … Person Re-Identification: study of anonymization strategies such that the identities **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**

***Privacy in Social Networks: A Survey*, Zheleva and Getoor, book chapter in Social Network Data Analytics 2010.**

Conclusion

- Collective Entity Resolution
  - Process of **data cleaning** and **knowledge reformulation** where we have relational information that tells us about the structure of the domain
  - Structure helps us define features and propagate information, and allows us to improve the overall accuracy
- 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