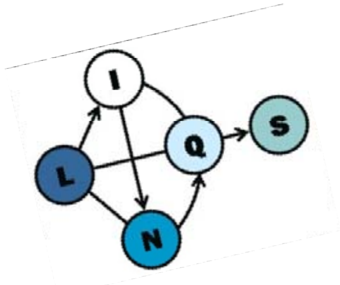
Graph Identification

Lise Getoor

University of Maryland, College Park

Visiting Researcher, MSR Search Labs

SIAM Annual Mtg
July 11, 2008



Some Acknowledgements

- Students:

  Indrajit Bhattacharya, Louis Licamele, Vivek Sehgal
  Mustafa Bilgic, Qing Lu, Prithvi Sen
  Rezarta Islamaj, Galileo Namata, Elena Zheleva

- Collaborators:

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

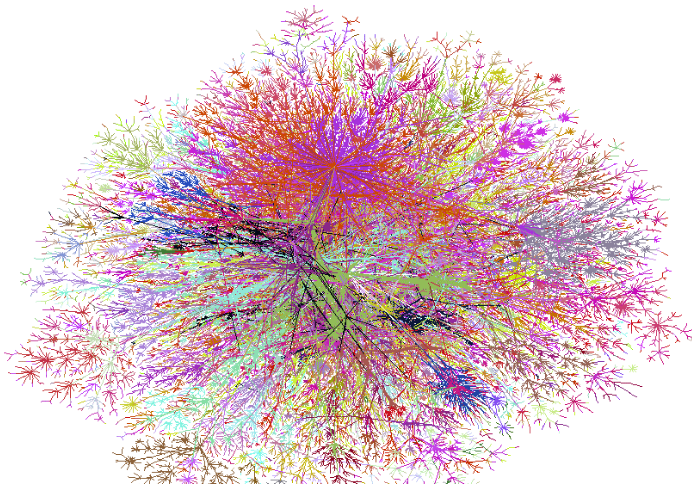
- Funding Sources:

  NSF, National Geospatial-Intelligence Agency, Google, Microsoft Research

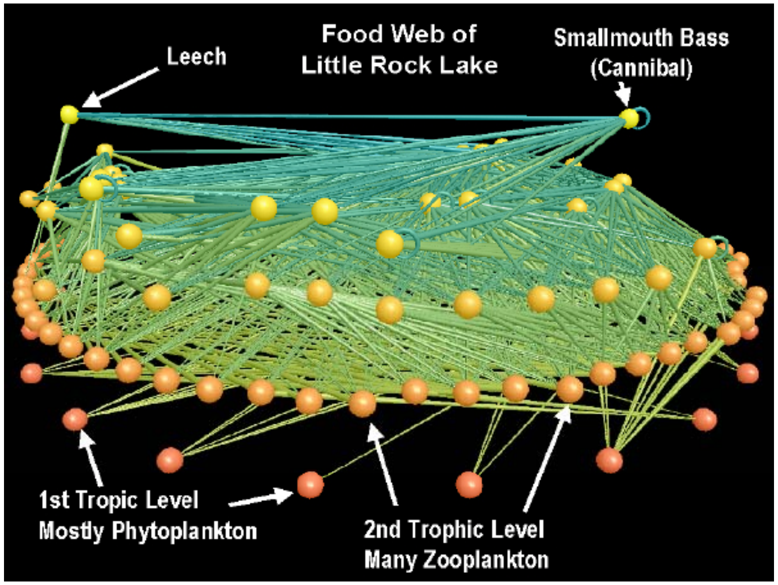
  KDD Program

Graphs and Networks *everywhere*…

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



Internet Map, Burch and Cheswick



Food Web, Martinez

Others available at Mark Newman's gallery:
http://www-personal.umich.edu/~mejn/networks/

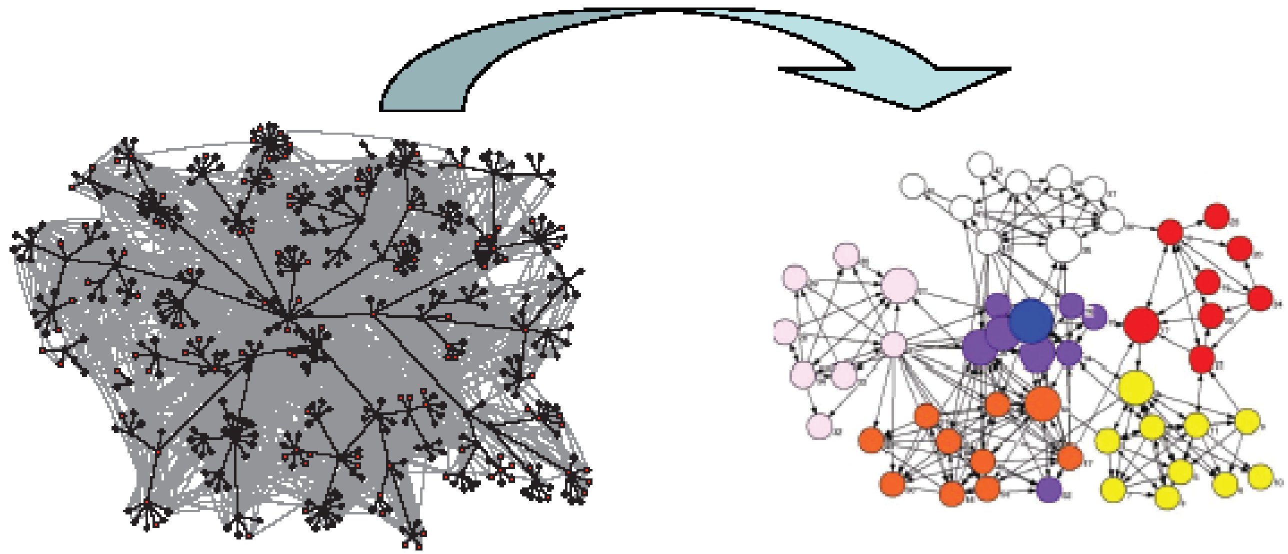
Wealth of Data

- 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



Communications Graph
Nodes: Network References
Edges: Communications Events

Network Graph
Nodes: Entities
Edges: Social Relationships

Data Graph ⇒ 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

Other Applications…

- Natural Language Processing: co-reference resolution, role labeling, sentiment analysis, …
- Computer Vision: correspondence analysis, scene understanding, activity recognition, …
- Computational Biology: protein-protein interaction networks, transcriptional regulation, signaling, …
- Databases: data cleaning, schema and ontology alignment, personal information management, …
- Web Search: extracting useful information from web pages, query logs, click logs, and more…

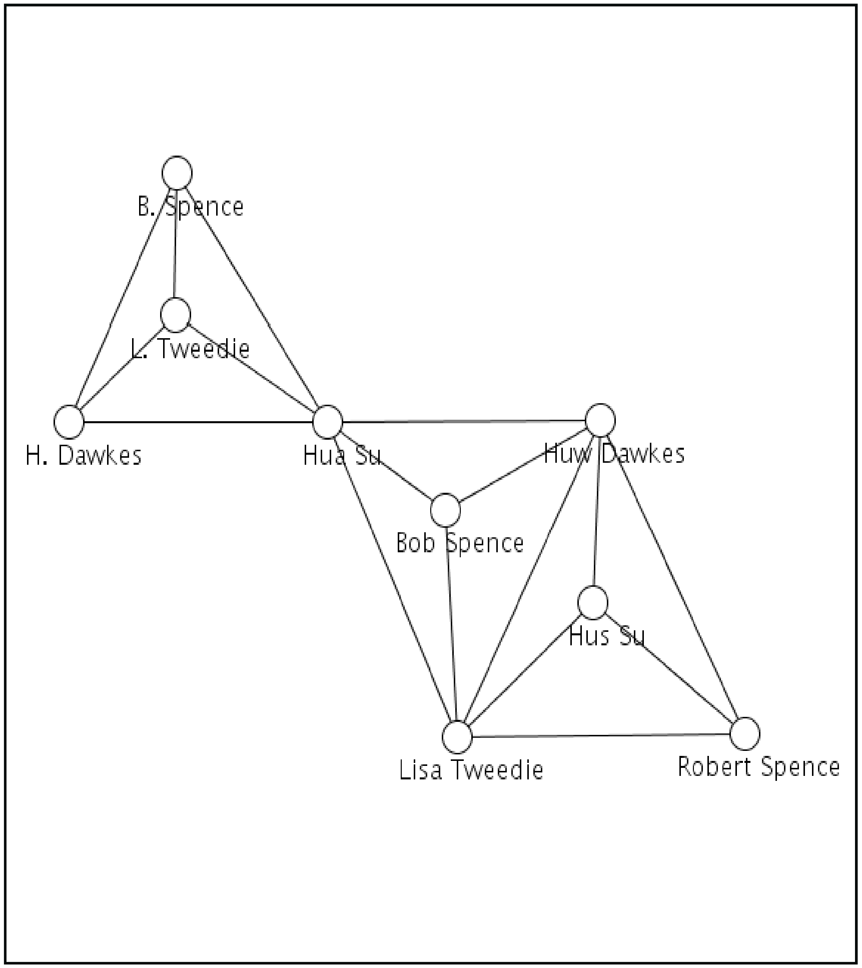
Roadmap

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

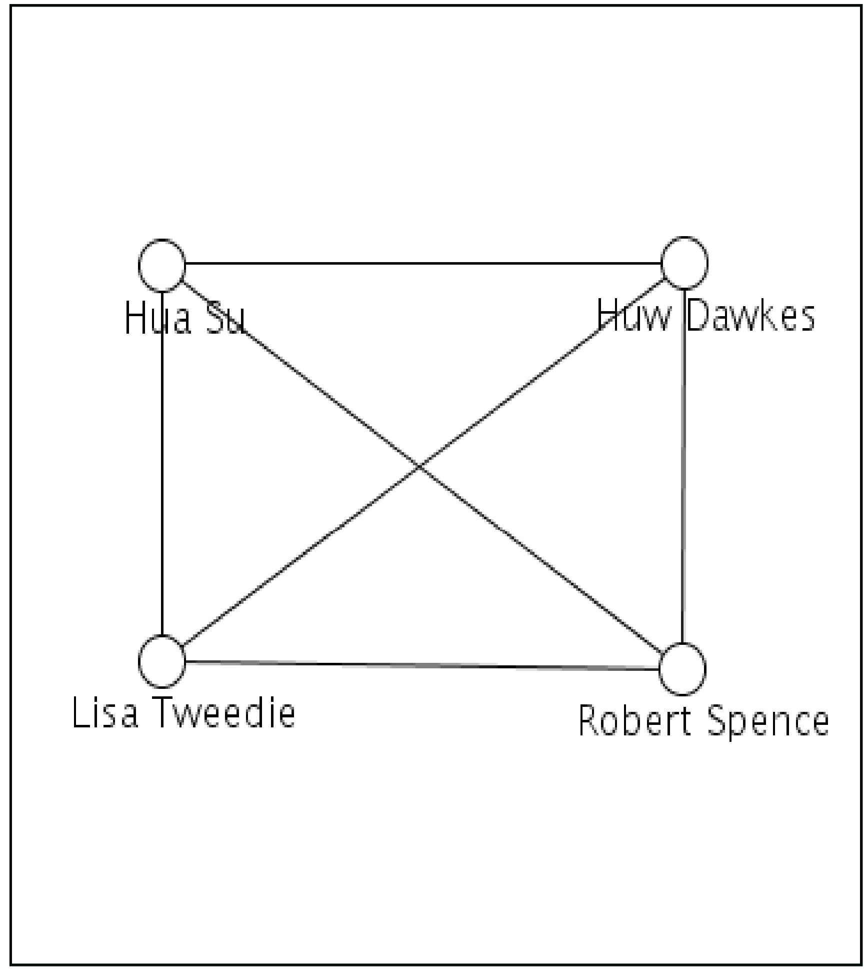
Entity Resolution

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

InfoVis Co-Author Network Fragment

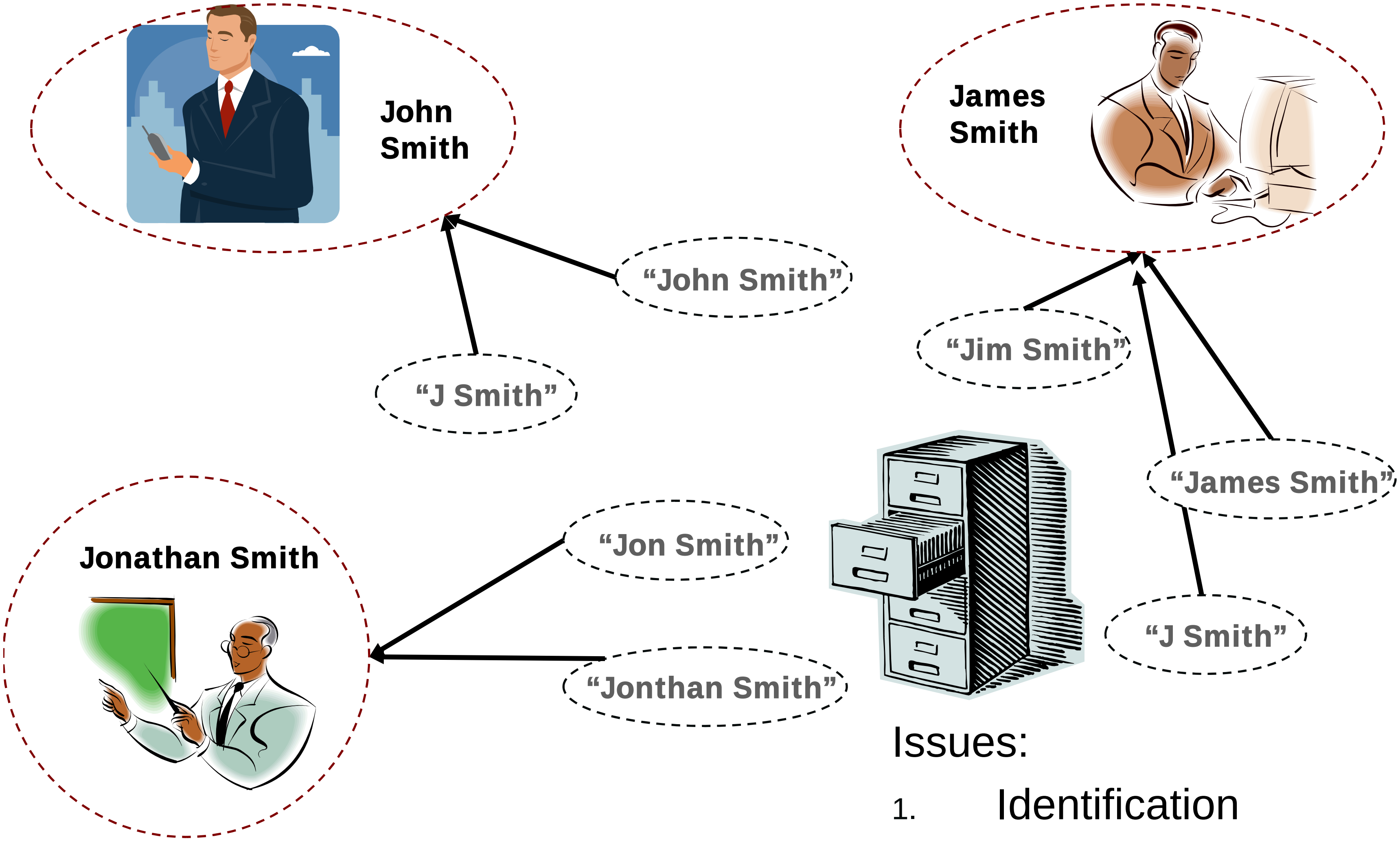


before



after

The Entity Resolution Problem



Issues:

1. Identification
2. Disambiguation

Attribute-based Entity Resolution

Pair-wise classification

"J Smith"	"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

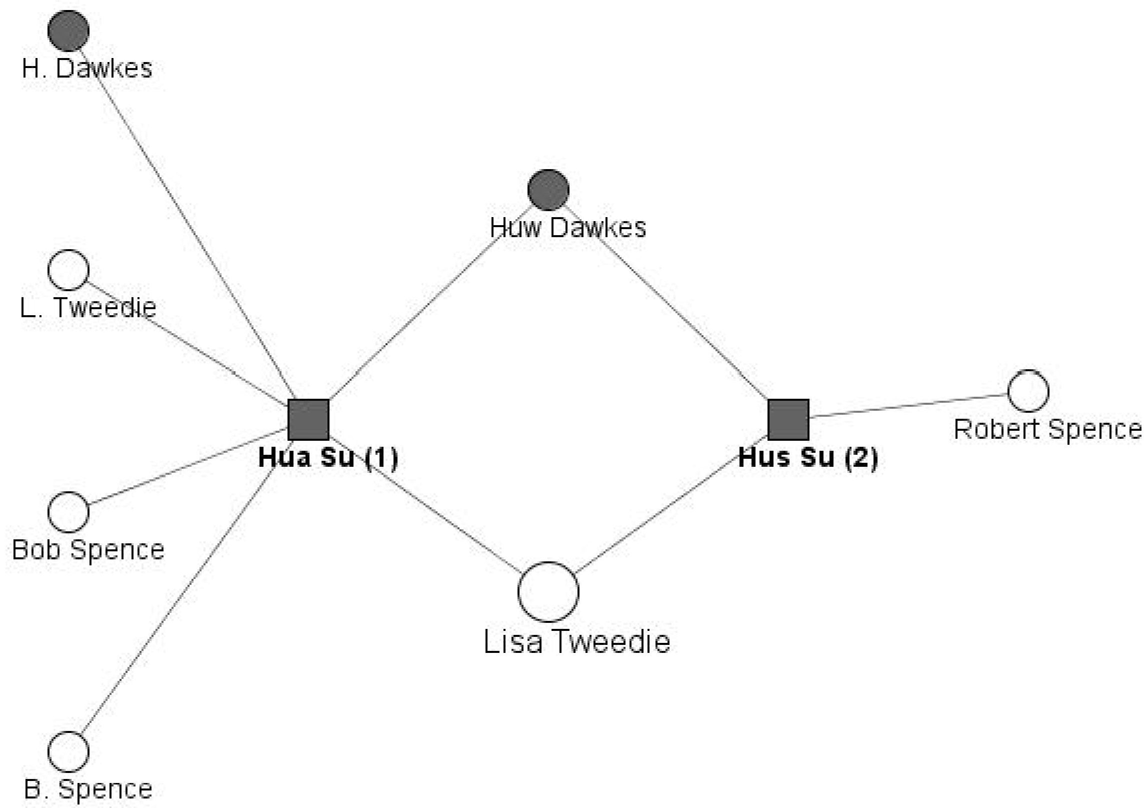
- The Problem
- **Relational Entity Resolution**
- Algorithms

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

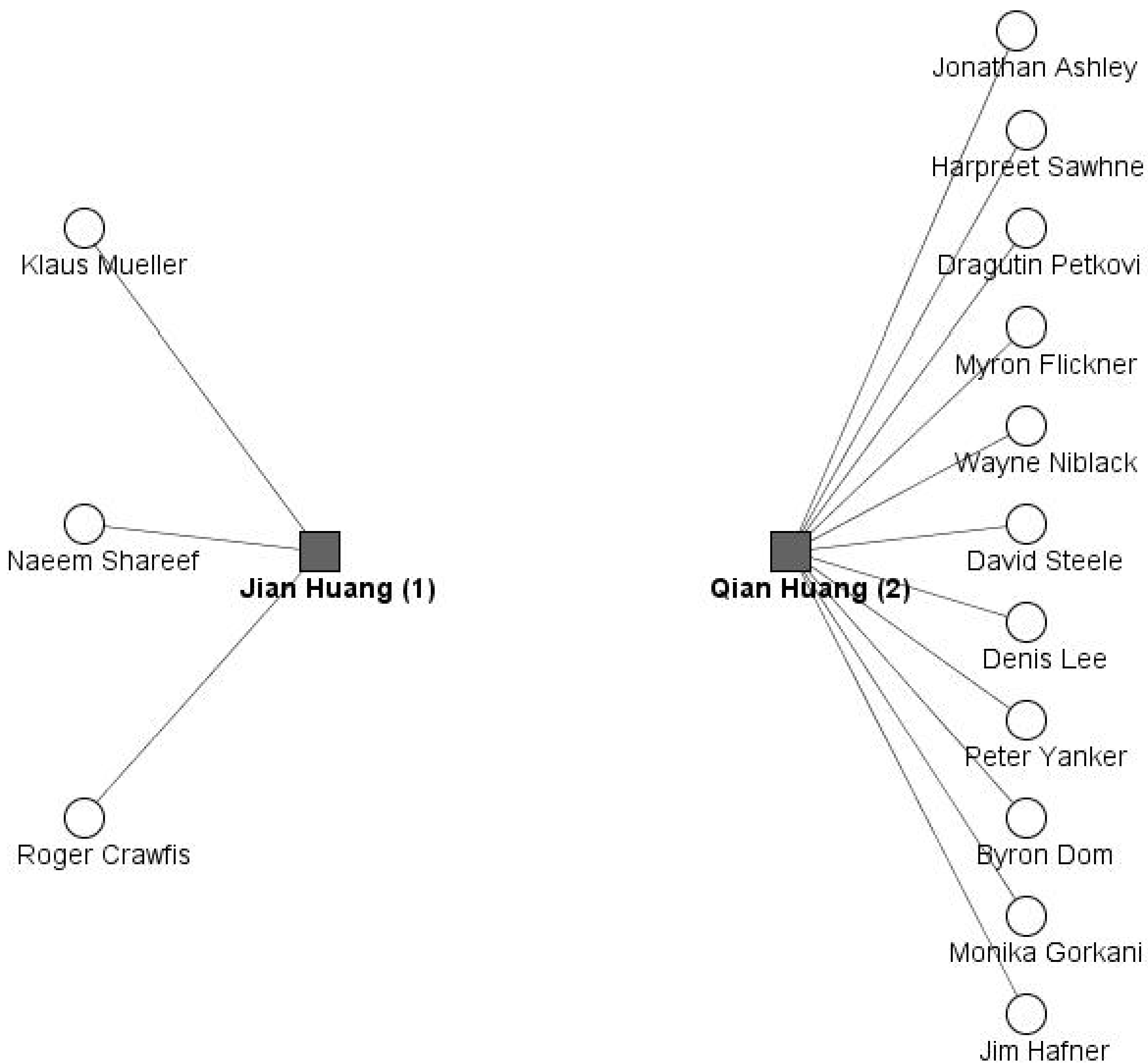
Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05

Relational Identification



Very similar names.
Added evidence from shared co-authors

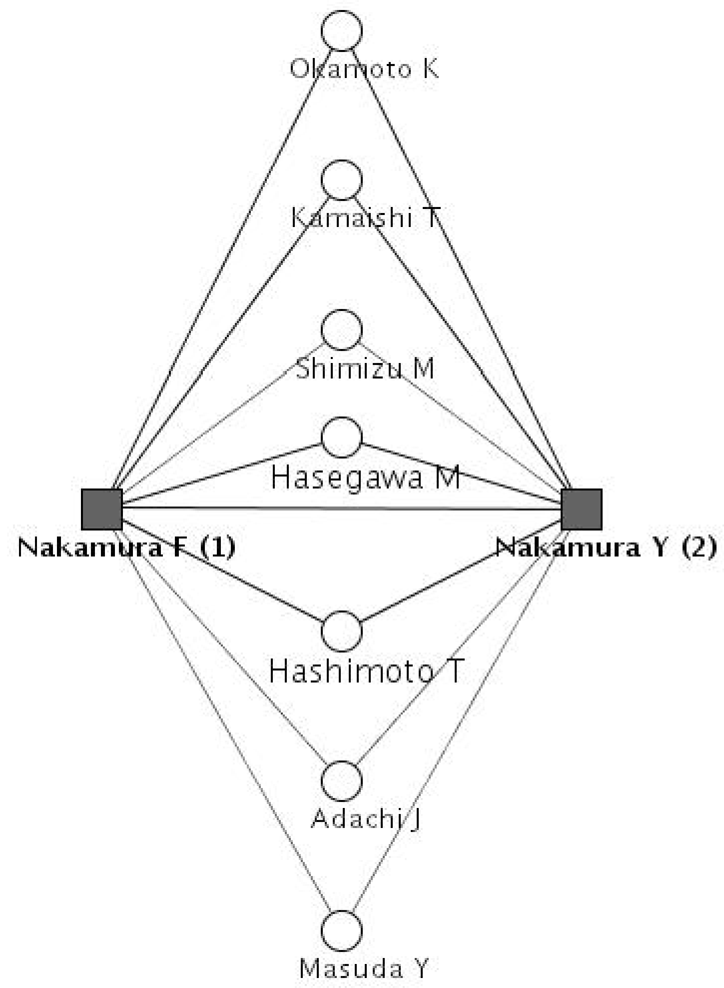
Relational Disambiguation



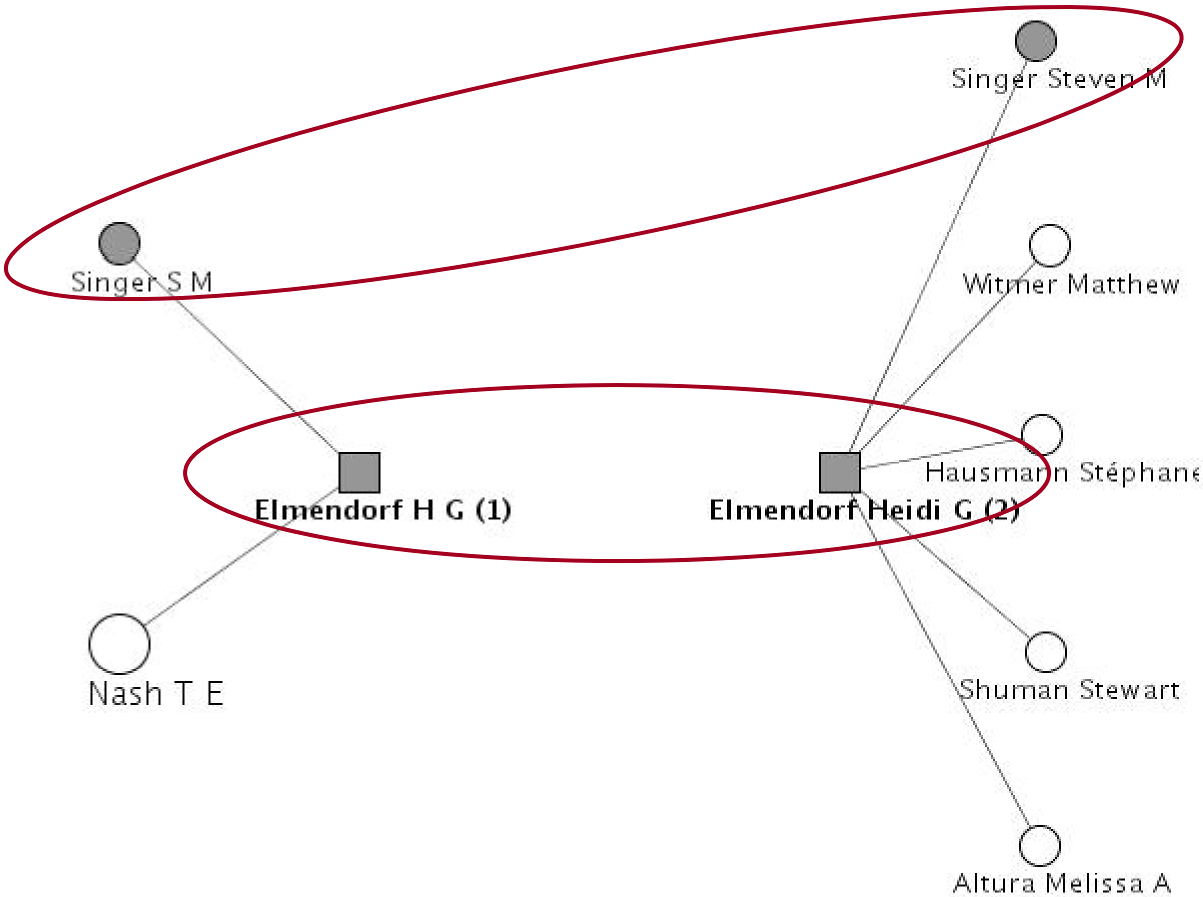
Very similar names but no shared collaborators

Relational Constraints

Co-authors are typically distinct



Collective Entity Resolution



One resolution provides evidence for another => joint resolution

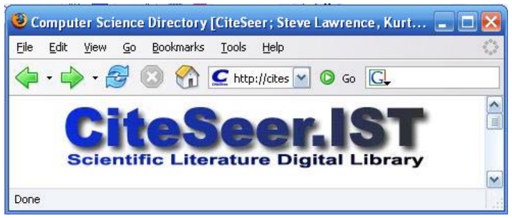
Entity Resolution with Relations

- 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*



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

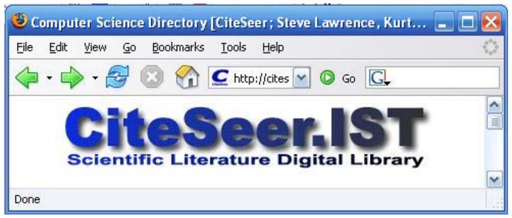
P2: **“***Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies”*, C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus

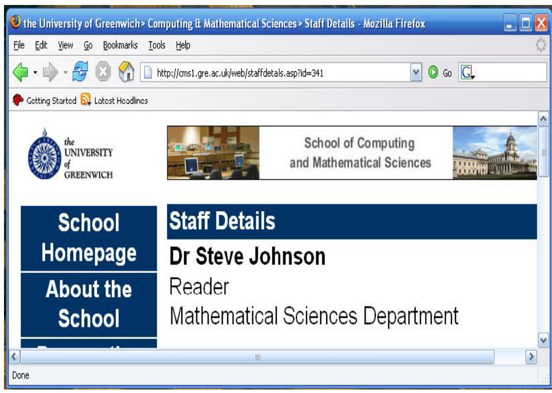
P3: **“***Dynamic Mesh Partitioning: A Unied Optimisation and Load-Balancing Algorithm”*, C. Walshaw, M. Cross, M. G. Everett

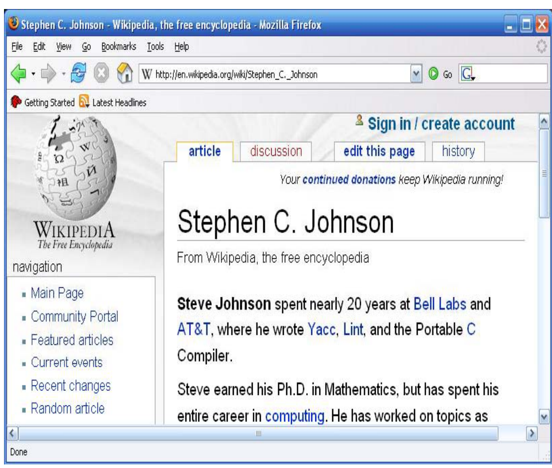
P4: **“***Code Generation for Machines with Multiregister Operations”*, Alfred V. Aho, Stephen C. Johnson, Jefferey D. Ullman

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

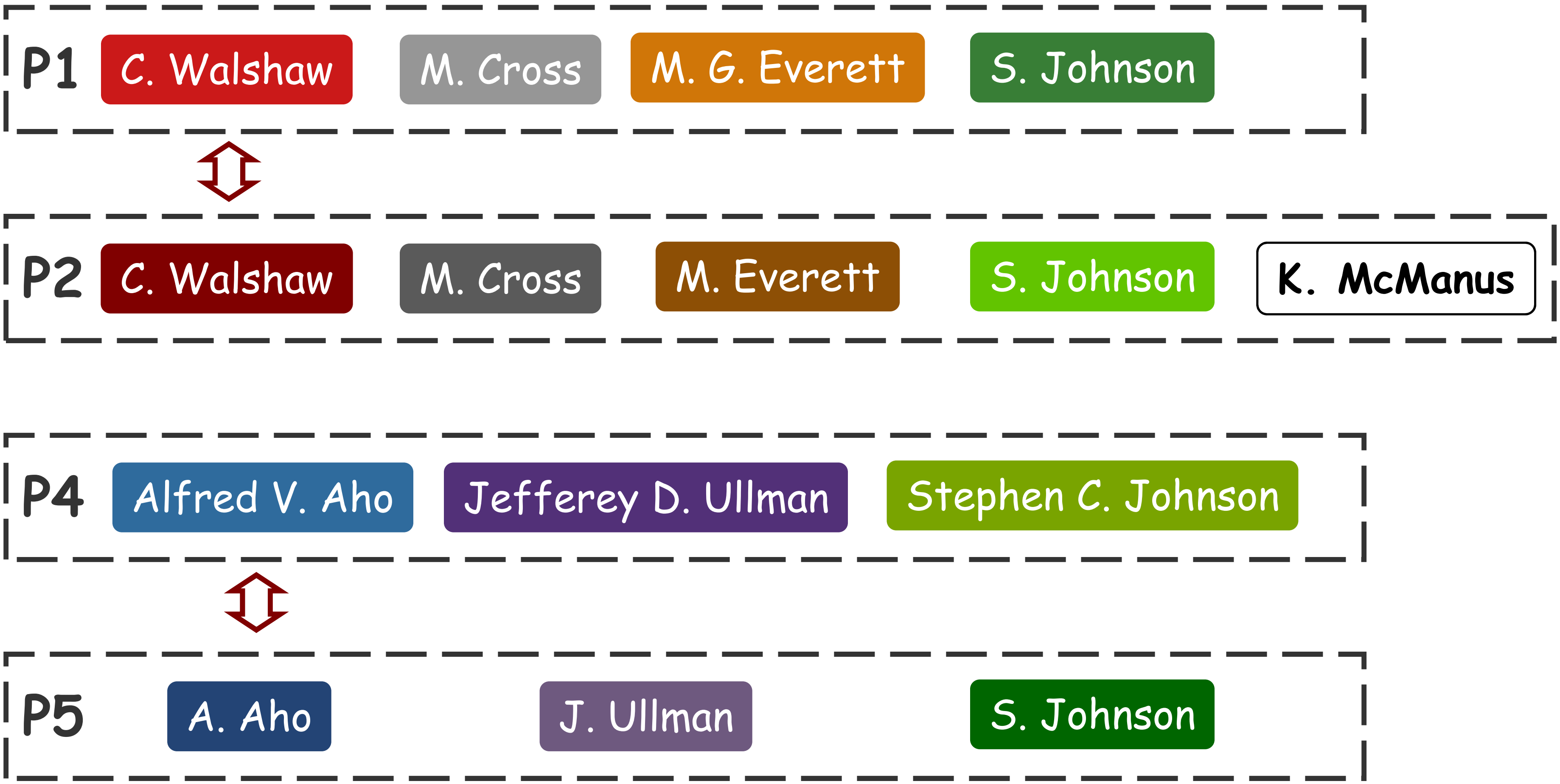
P3: "*Dynamic Mesh Partitioning: A Unied Optimisation and Load-Balancing Algorithm*", C. Walshaw, M. Cross, M. G. Everett

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

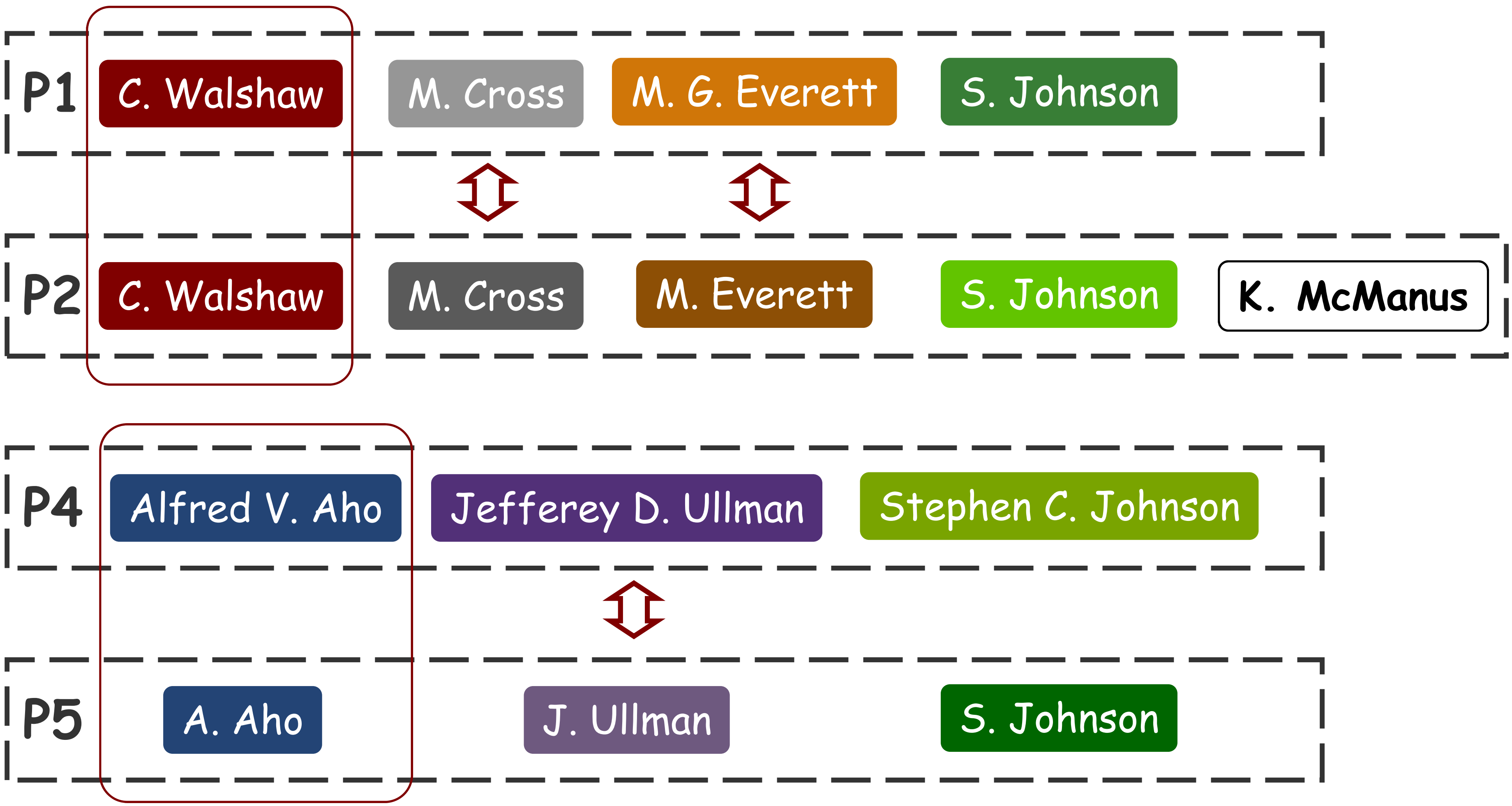
P5: "*Deterministic Parsing of Ambiguous Grammars*", A. Aho, **S. Johnson**, J. Ullman

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

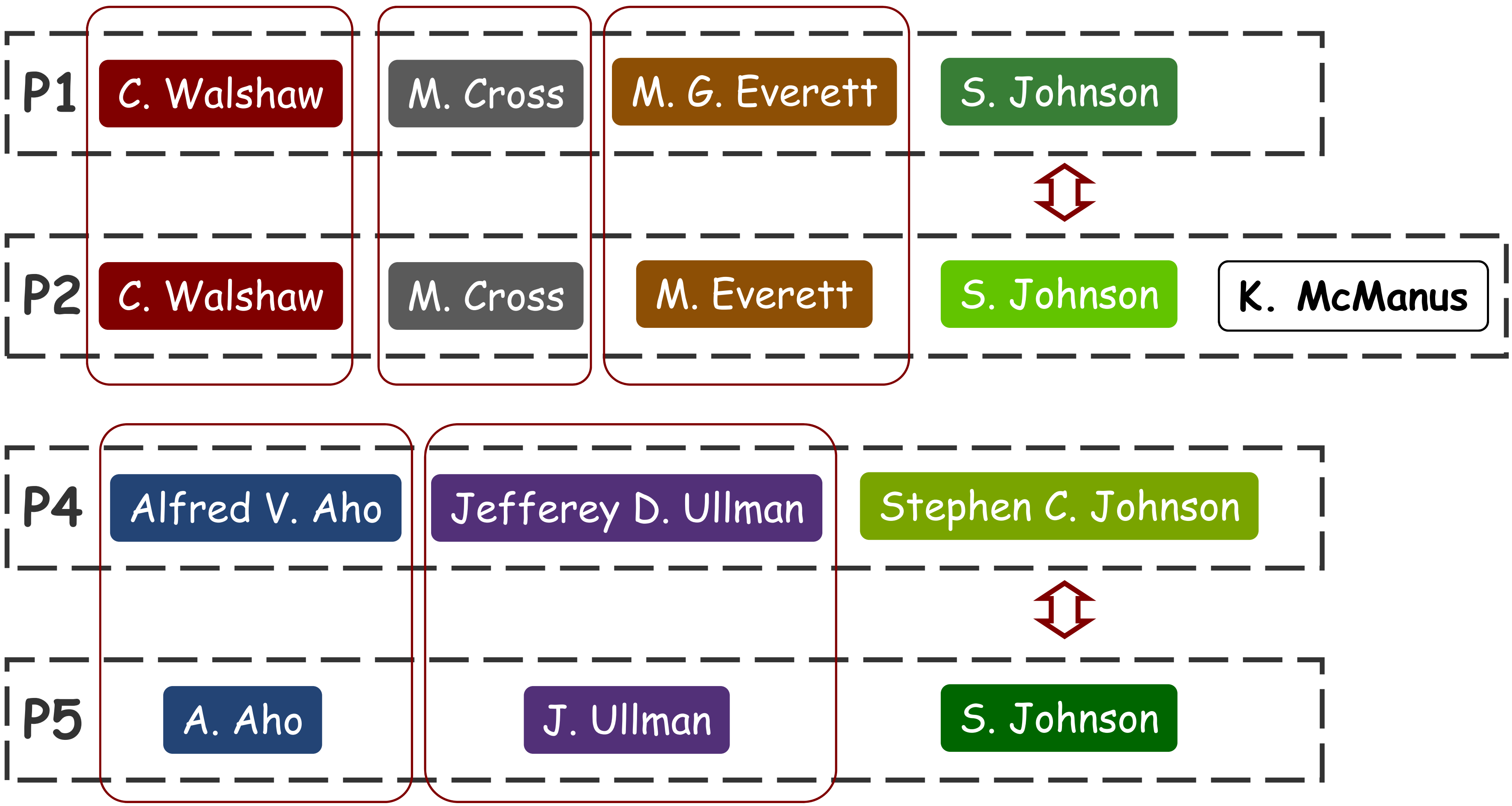
Relational Clustering (RC-ER)



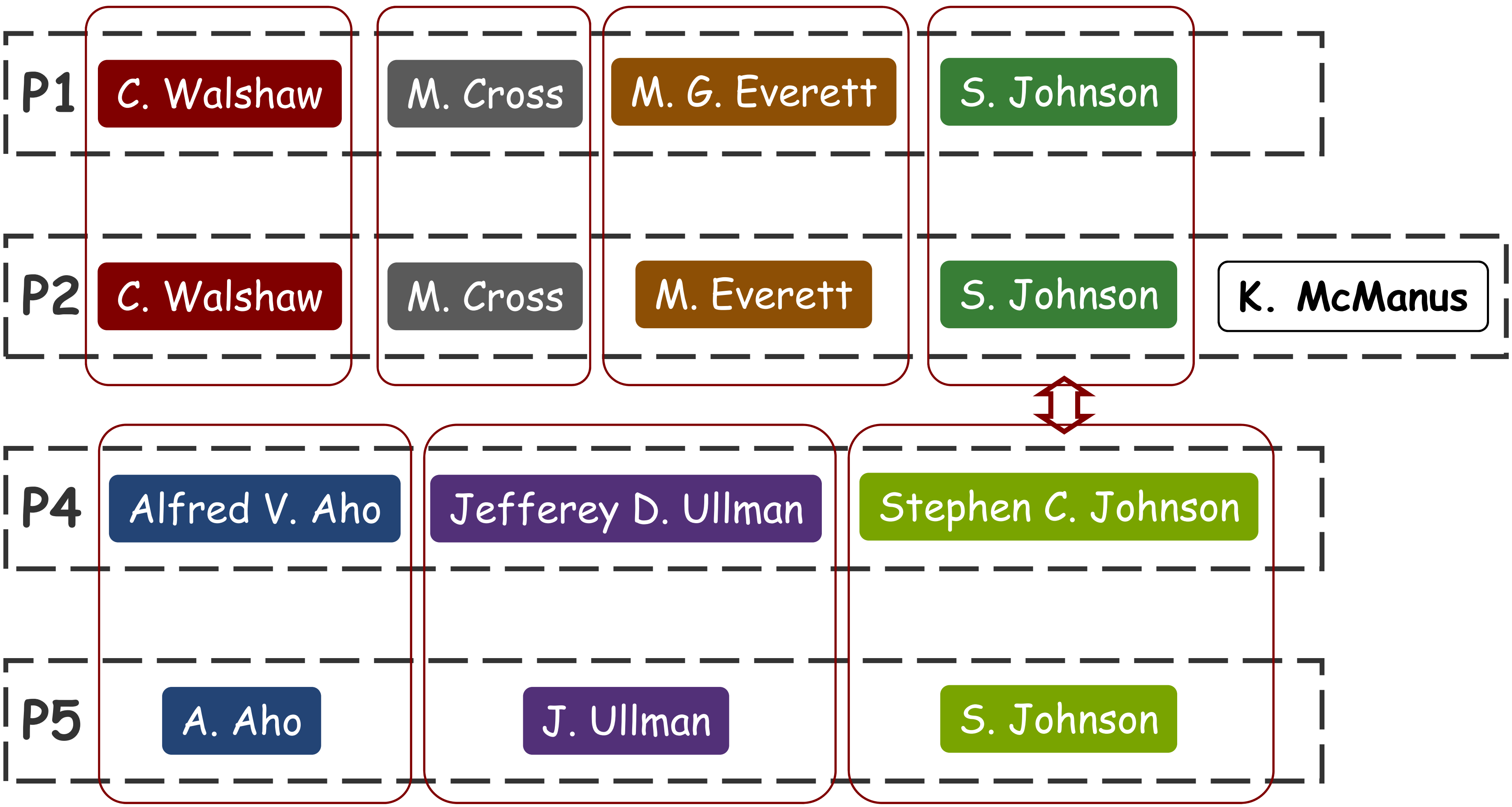
Relational Clustering (RC-ER)



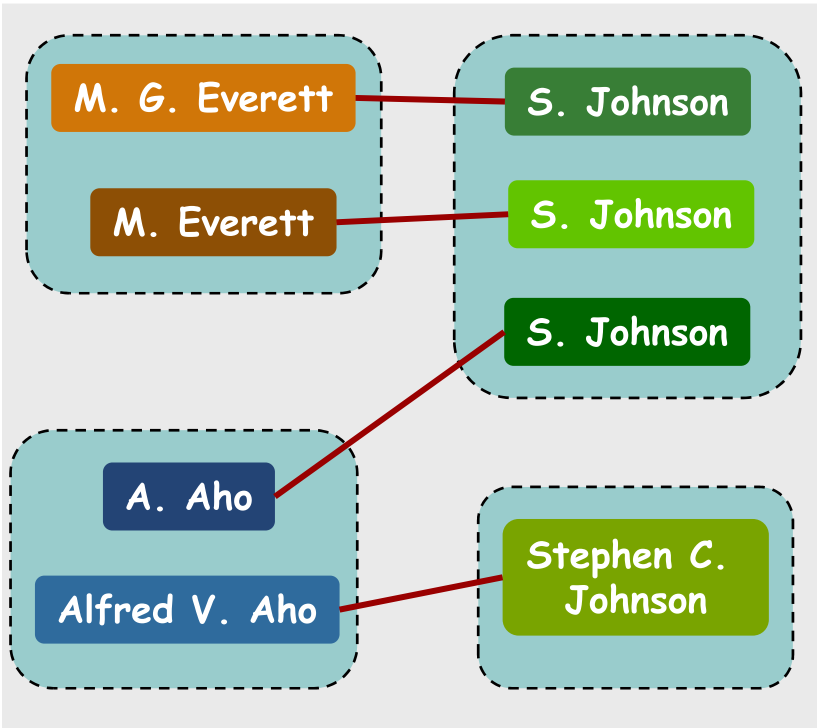
Relational Clustering (RC-ER)

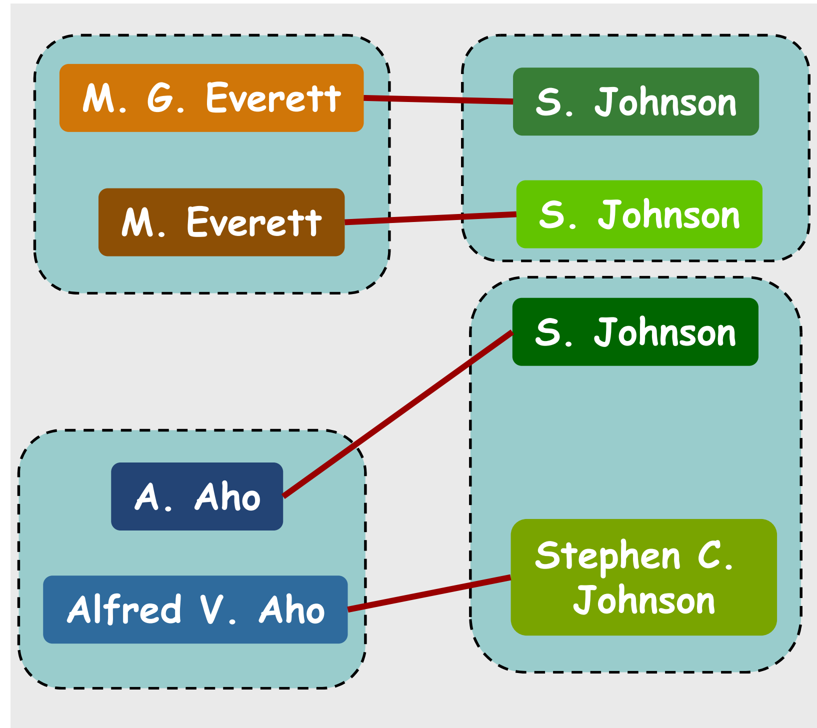


Relational Clustering (RC-ER)



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:**

$$\sum_i \sum_j \underbrace{w_A}_{\text{weight for attributes}} \underbrace{sim_A(c_i, c_j)}_{\text{similarity of attributes}} + \underbrace{w_R}_{\text{weight for relations}} \underbrace{sim_R(c_i, c_j)}_{\text{Similarity based on relational edges between } c_i \text{ and } c_j}$$

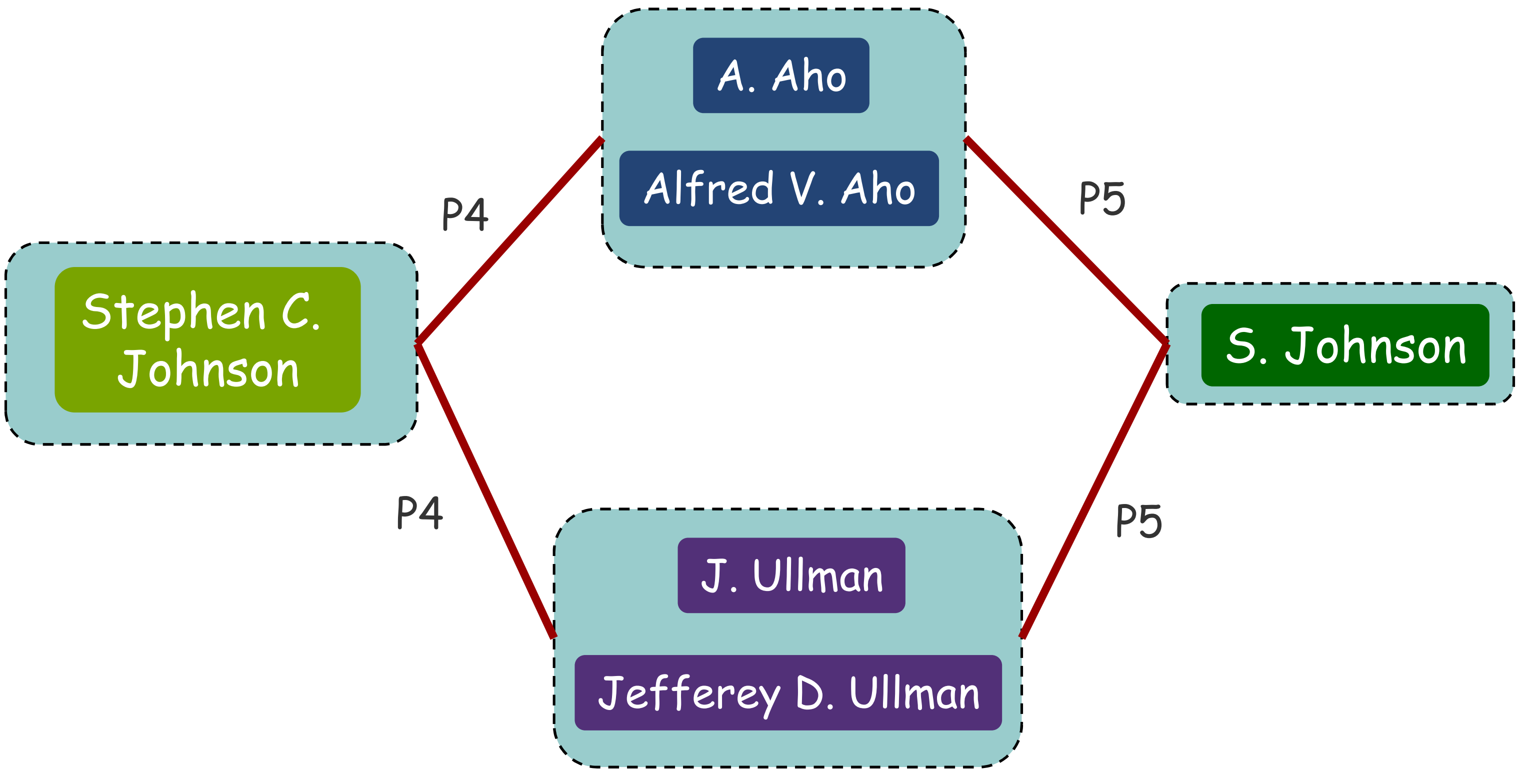
- **Greedy clustering algorithm:** merge cluster pair with max reduction in objective function

Measures for Attribute Similarity

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

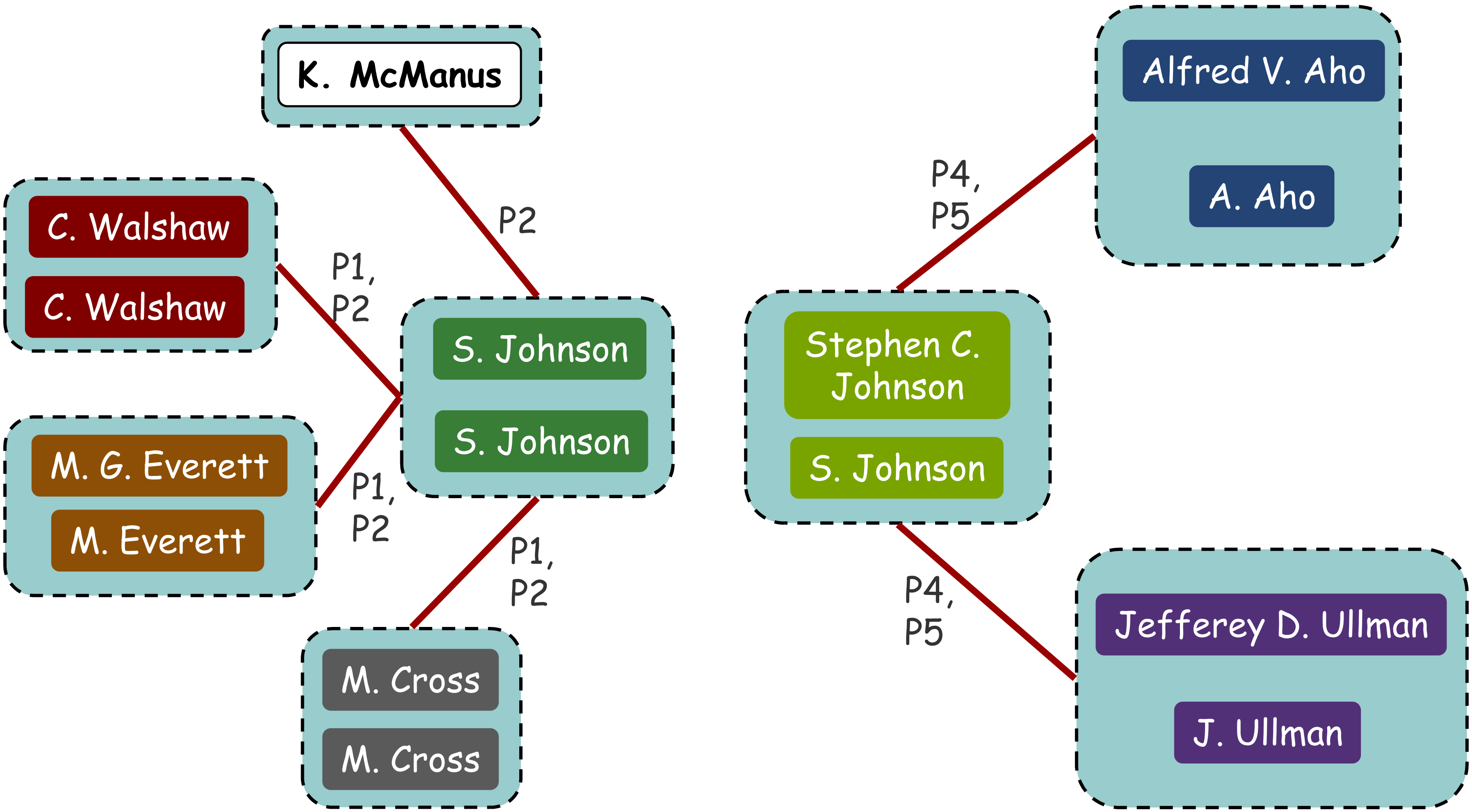
- Aggregate to find similarity between clusters
  - Single link, Average link, Complete link
  - Cluster representative

Relational Similarity: Example 1



All neighborhood clusters are shared: high relational similarity

Relational Similarity: Example 2



No neighborhood cluster is shared: no relational similarity

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)
  - **Experimental Evaluation**

Probabilistic Generative Model for Collective Entity Resolution

- Model how references co-occur in data

1. Generation of references from entities

2. Relationships between underlying entities
   - Groups of entities instead of pair-wise relations

Discovering Groups from Relations

Stephen P Johnson

Chris Walshaw

Kevin McManus

Mark Cross

Martin Everett

Parallel Processing Research Group

Stephen C Johnson

Alfred V Aho

Ravi Sethi

Jeffrey D Ullman

Bell Labs Group

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

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

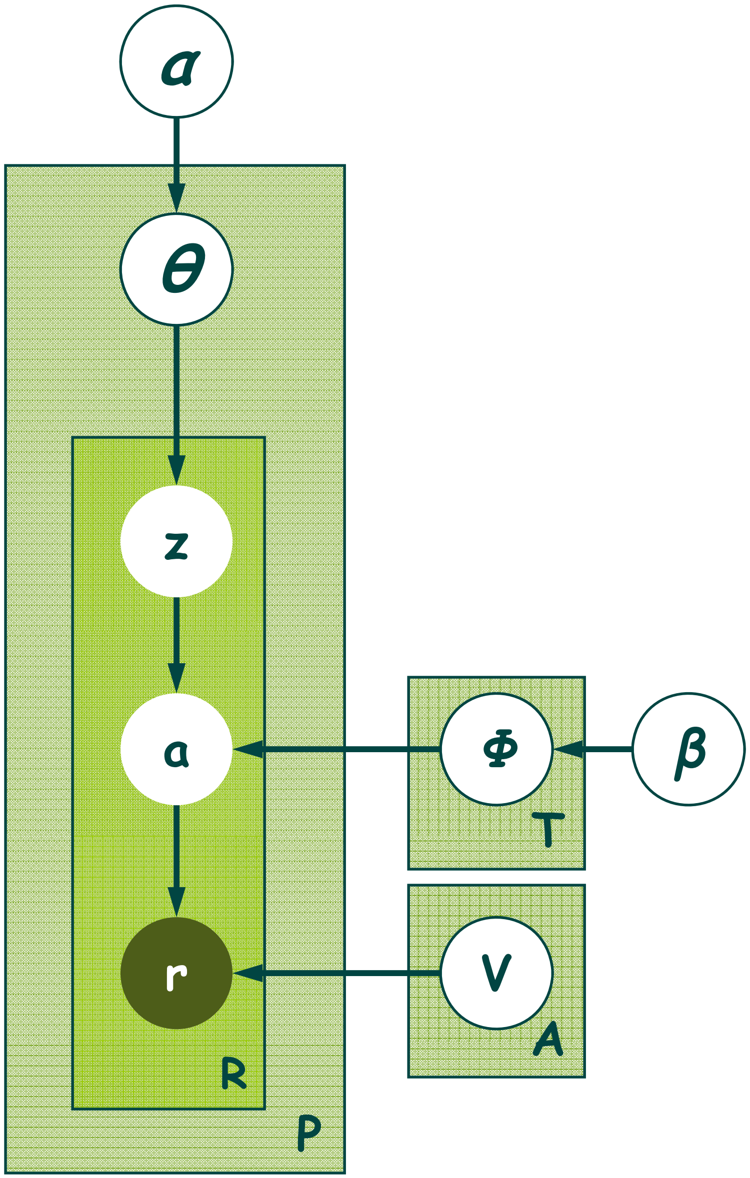
P3: C. Walshaw, M. Cross, M. G. Everett

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

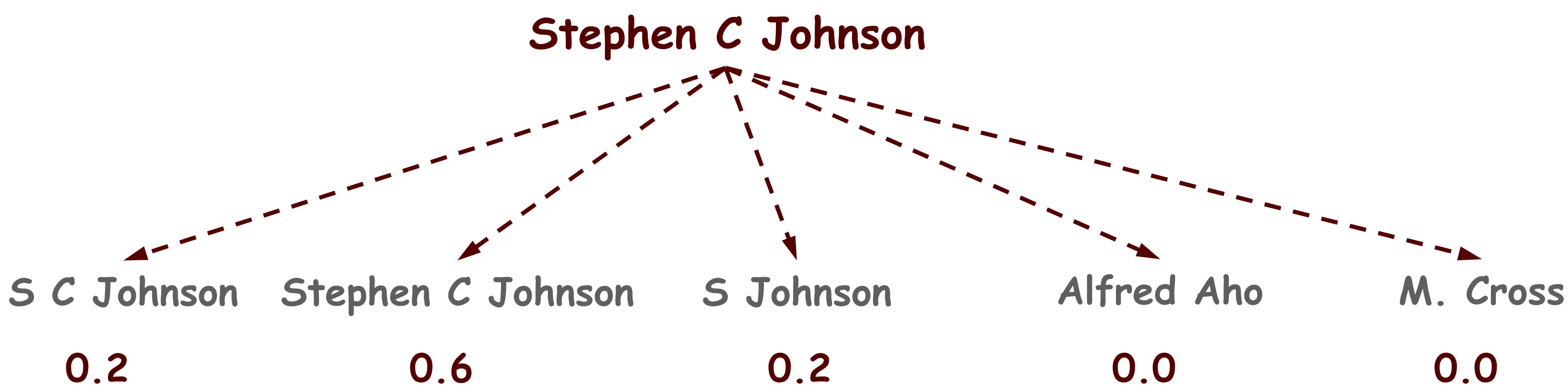


- Entity label ***a*** and group label ***z*** for each reference ***r***
- ***Θ*: 'mixture' of groups for each co-occurrence**
- ***Φ*** $\boldsymbol{\Phi_z}$**:** multinomial for choosing entity ***a*** for each group ***z***
- $\boldsymbol{V_a}$**:** multinomial for choosing reference ***r*** from entity ***a***
- Dirichlet priors with ***α*** and ***β***

Generating References from Entities

- Entities are not directly observed
    1. Hidden attribute for each entity
    2. Similarity measure for pairs of attributes

- A distribution over attributes for each entity



Approx. Inference Using Gibbs Sampling

- Conditional distribution over labels for each ref.
- Sample next labels from conditional distribution
- Repeat over all references until convergence

$$P(z_i = t \mid \mathbf{z}_{-i}, \mathbf{a}, \mathbf{r}) \propto \frac{n^{DT}_{d_i t} + \alpha / T}{n^{DT}_{d_i *} + \alpha} \times \frac{n^{AT}_{a_i t} + \beta / A}{n^{AT}_{* t} + \beta}$$

$$P(a_i = a \mid \mathbf{z}, \mathbf{a}_{-i}, \mathbf{r}) \propto \frac{n^{AT}_{a_i t} + \beta / A}{n^{AT}_{* t} + \beta} \times Sim(r_i, v_a)$$

- Converges to most likely number of entities

Faster Inference: Split-Merge Sampling

- Naïve strategy reassigns references individually
- Alternative: allow entities to merge or split
- For entity a_i, find conditional distribution for
  1. Merging with existing entity a_j
  2. Splitting back to last merged entities
  3. Remaining unchanged
- Sample next state for a_i from distribution
- $O(n\ g + e)$ time per iteration compared to $O(n\ g + n\ e)$

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**

- RC-ER outperform attribute-only baselines in all datasets
- RC-ER better than naïve relational resolution in all datasets
- RC-ER and baselines require threshold as parameter
  - Reporting best achievable performance over all thresholds

***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**

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

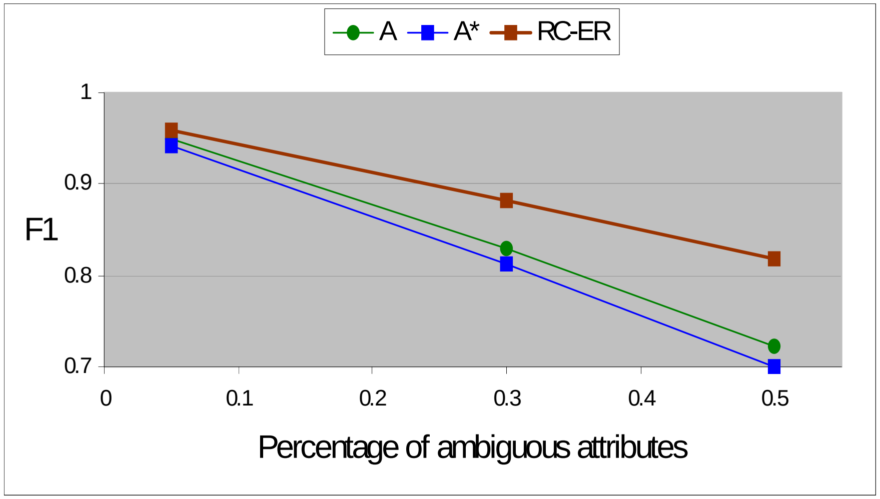
Performance for Specific Names

Name	Best F1 for ATTR/ATTR*	F1 for LDA-ER
cho_h	0.80	1.00
davis_a	0.67	0.89
kim_s	0.93	0.99
kim_y	0.93	0.99
lee_h	0.88	0.99
lee_j	0.98	1.00
liu_j	0.95	0.97
sarkar_s	0.67	1.00
sato_h	0.82	0.97
sato_t	0.85	1.00
shin_h	0.69	1.00
veselov_a	0.78	1.00
yamamoto_k	0.29	1.00
yang_z	0.77	0.97
zhang_r	0.83	1.00
zhu_z	0.57	1.00

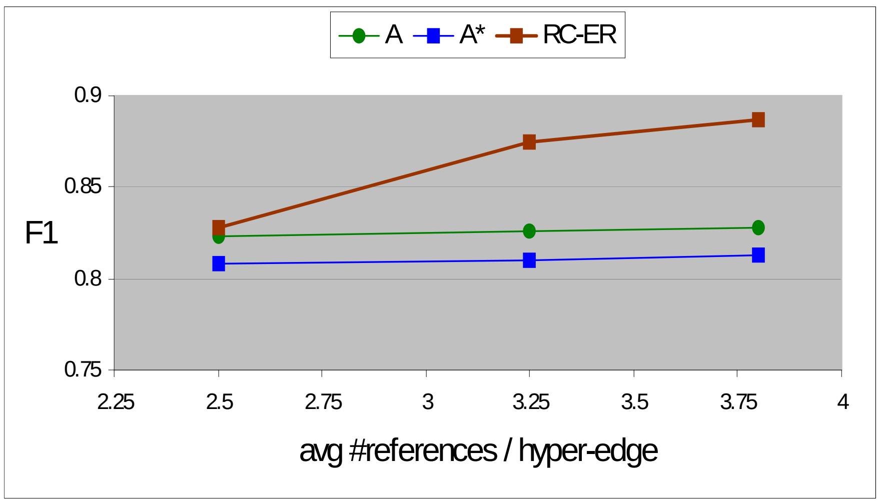
arXiv

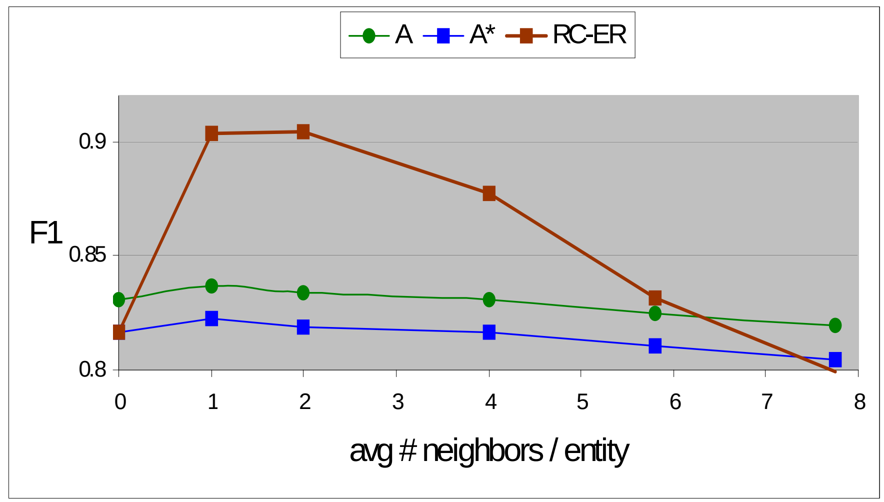
Significantly larger improvements for ‘ambiguous names’

Trends in Synthetic Data



Bigger improvement with
- bigger % of ambiguous refs
- more refs per co-occurrence
- more neighbors per entity





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

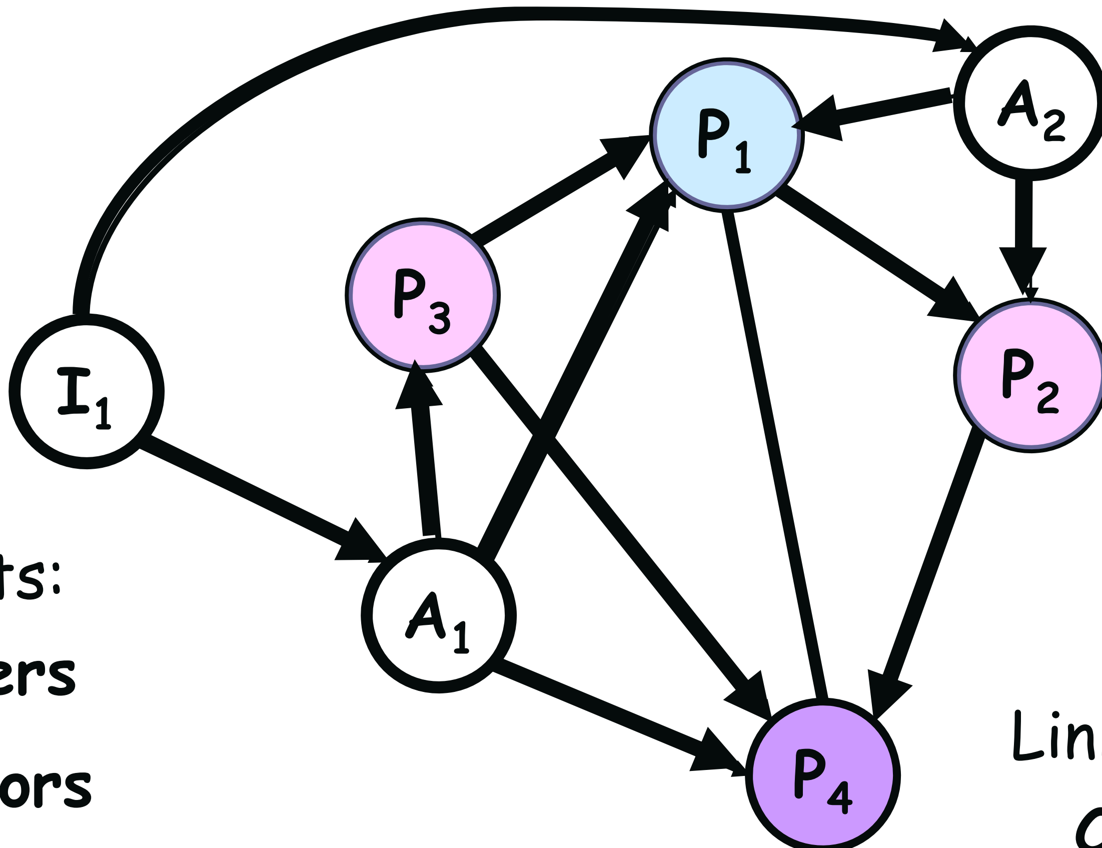
The Problem

- 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:

Feature Construction

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

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

Perlich & Provost 03, 04, 05, Popescul & Ungar 03, 05, 06, Lu & Getoor 03, Gupta, Diwam & Sarawagi 07

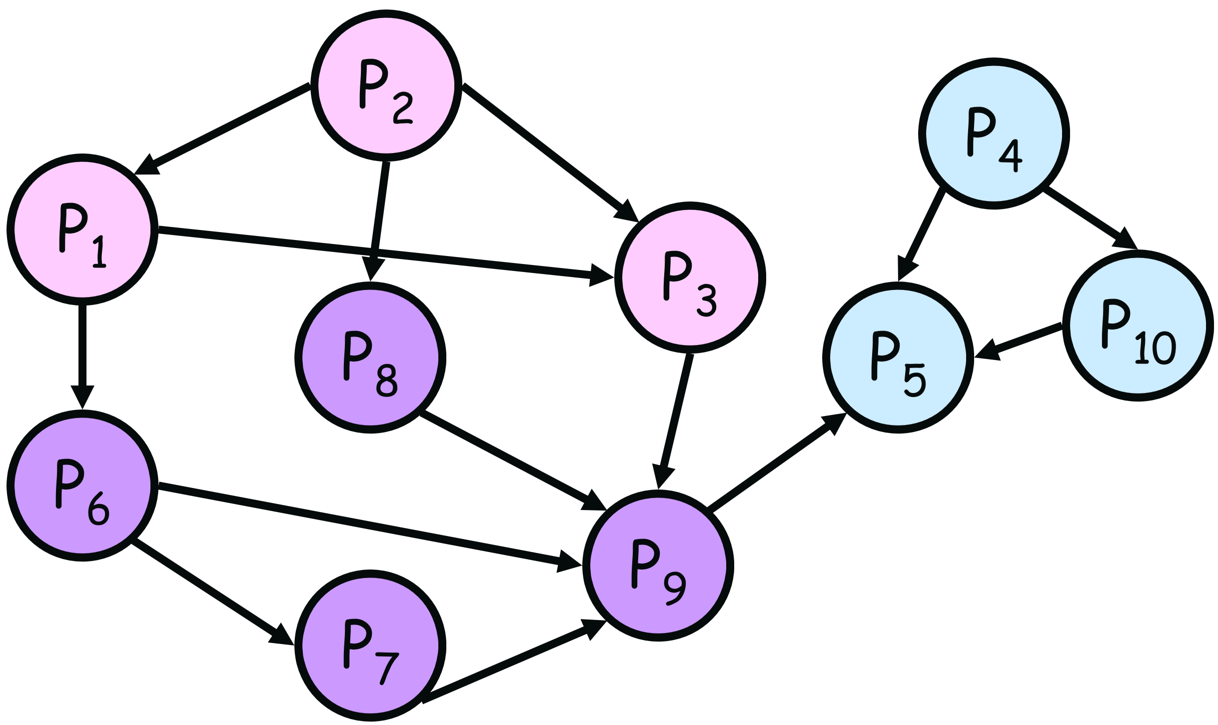
Formulation

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

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

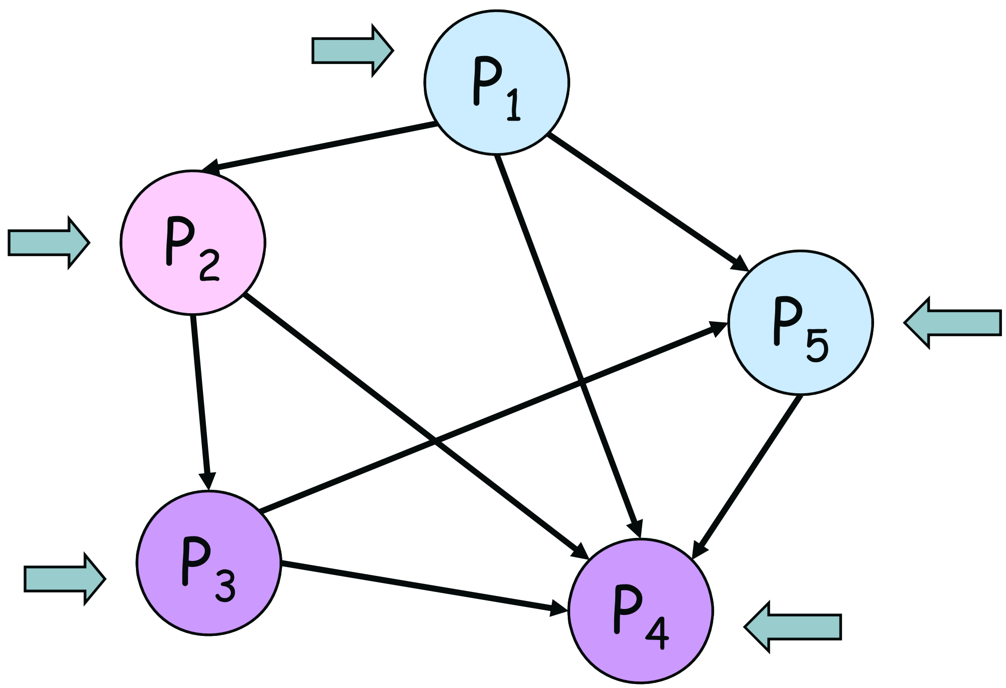
ICA: Learning

- label set:



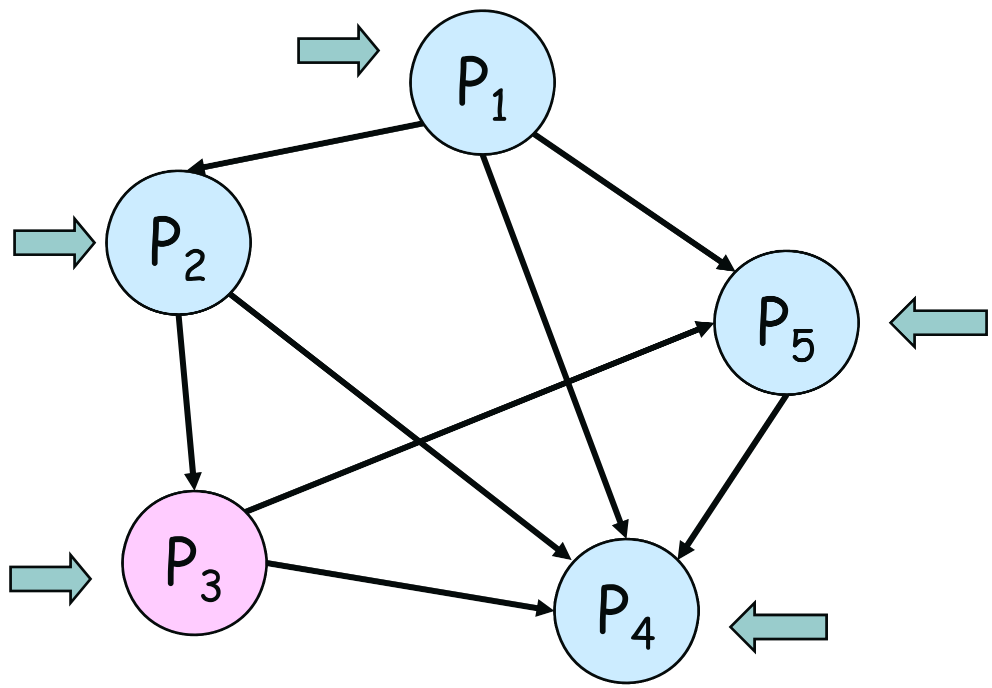
Learn model from fully labeled training set

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)
  - Loopy Belief Propagation (LBP)
  - Baseline: Content Only

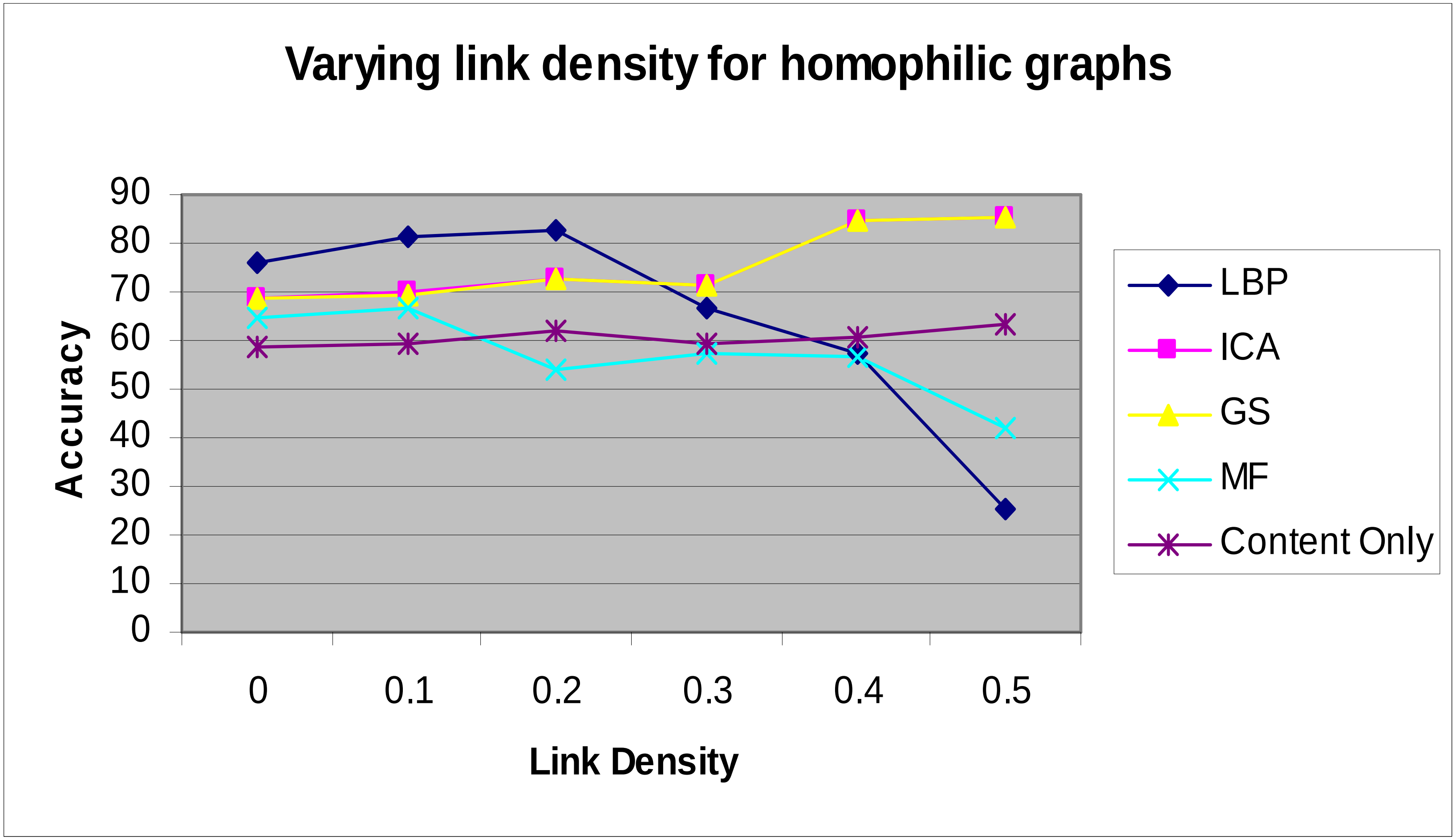
- Datasets
  - Real Data
    - Bibliographic Data (Cora & Citeseer), WebKB, etc.
  - Synthetic Data
    - Data generator which can vary the class label correlations (homophily), attribute noise, and link density

Results on Real Data

Algorithm	Cora	CiteSeer	WebKB
Content Only	66.51	59.77	62.49
ICA	74.99	62.46	**65.99**
Gibbs	74.64	62.52	65.64
MF	79.70	**62.91**	65.65
LBP	**82.48**	62.64	65.13

Sen and Getoor, TR 07

Effect of Structure



Results clearly indicate that algorithms' performance depends (in non-trivial ways) on structure

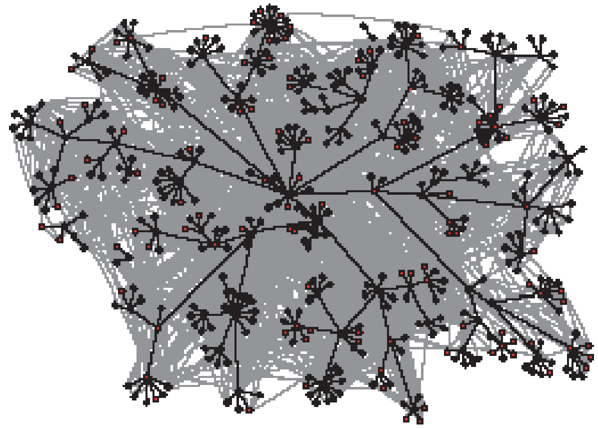
Roadmap

- 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

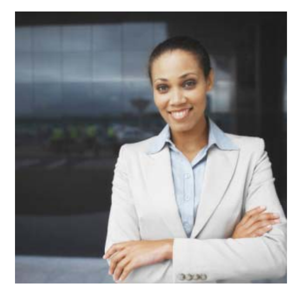
Links in Data Graph



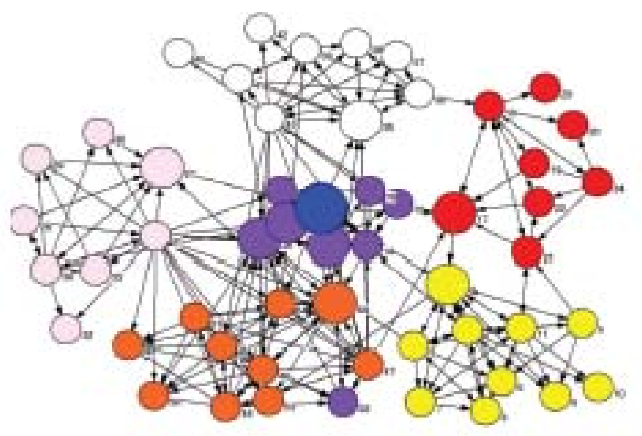
Communications Graph
Nodes: Network References
Edges: Communications Events

Node 1 ⟷ Node 2

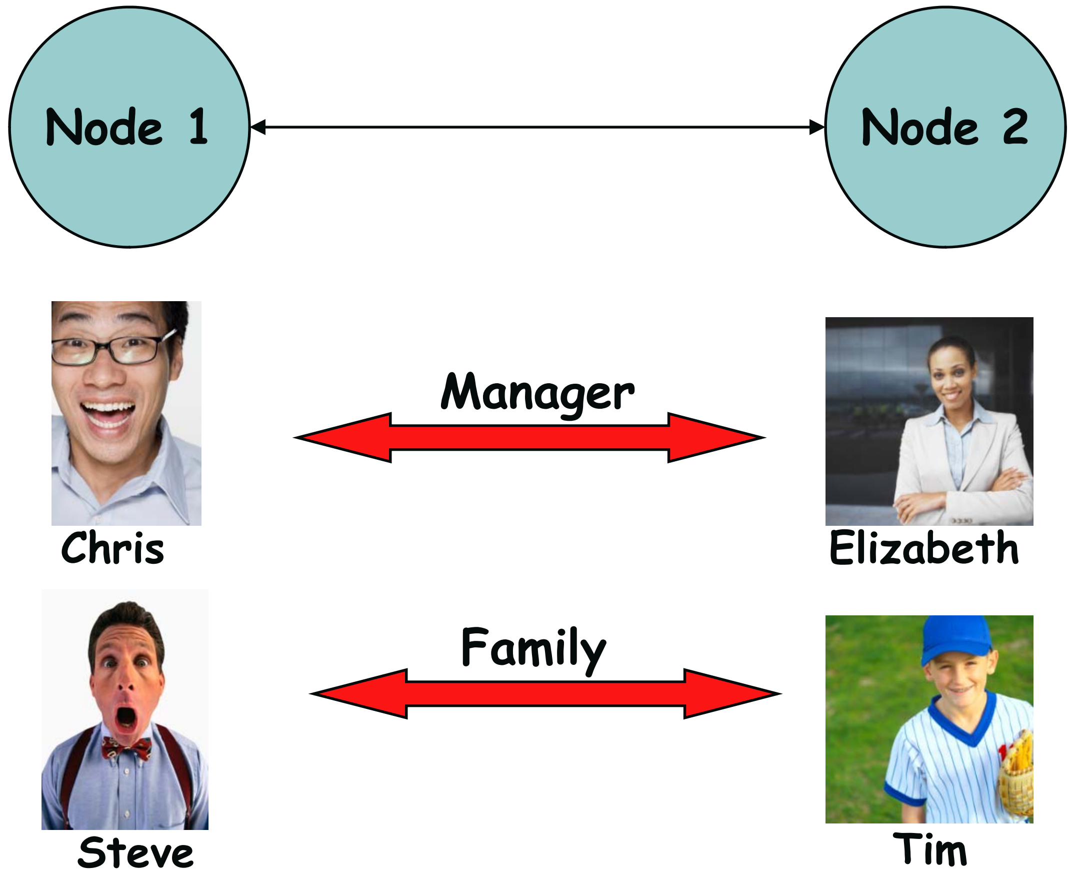
chris@enron.com	Email ⟷	liz@enron.com
chris37	IM ⟷	lizs22
555-450-0981	TXT ⟷	555-901-8812



⇒ Links in Information Graph



Network Graph
Nodes: Entities
Edges: Social Relationships



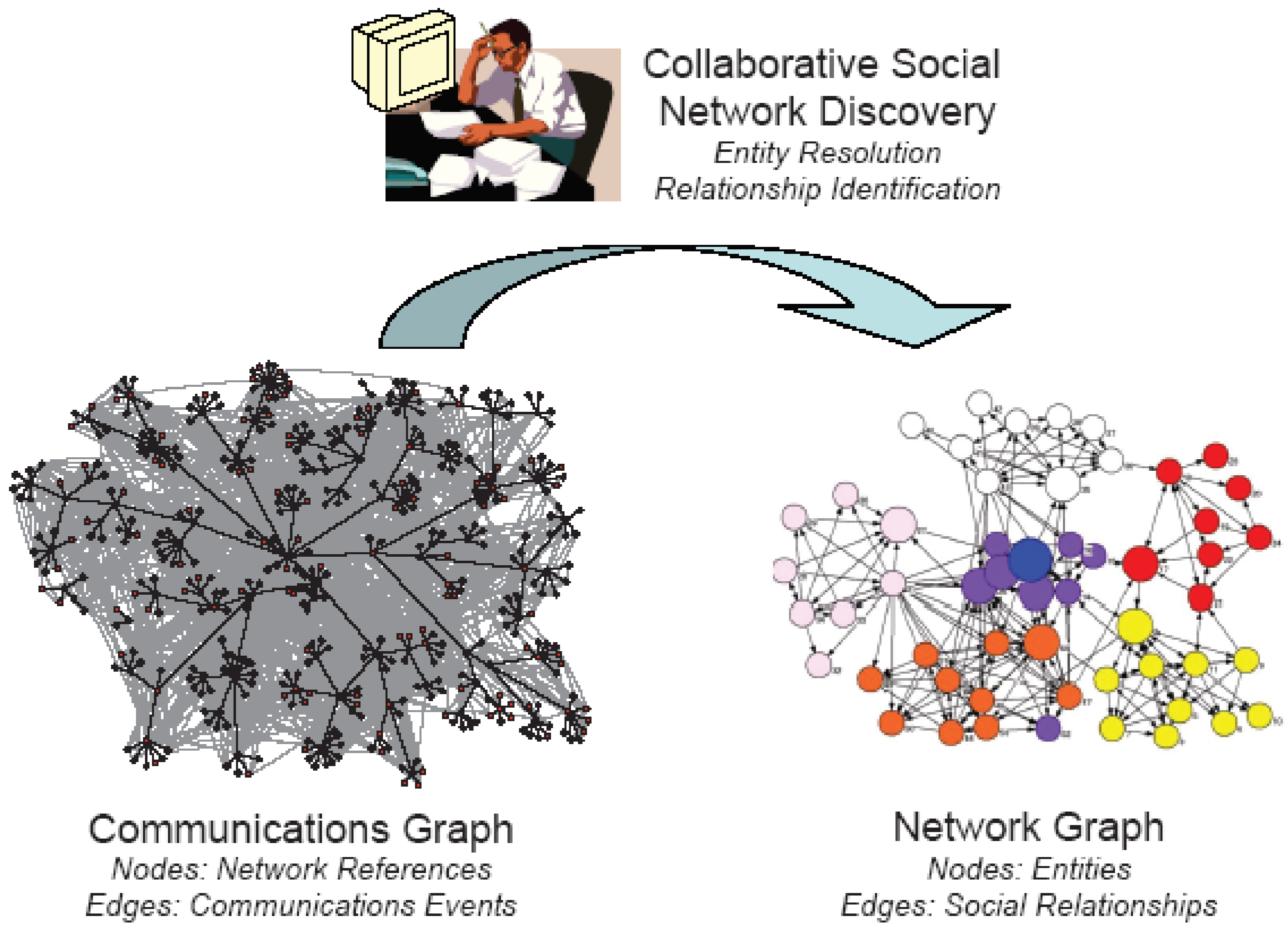
Predicting Relations

- 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….

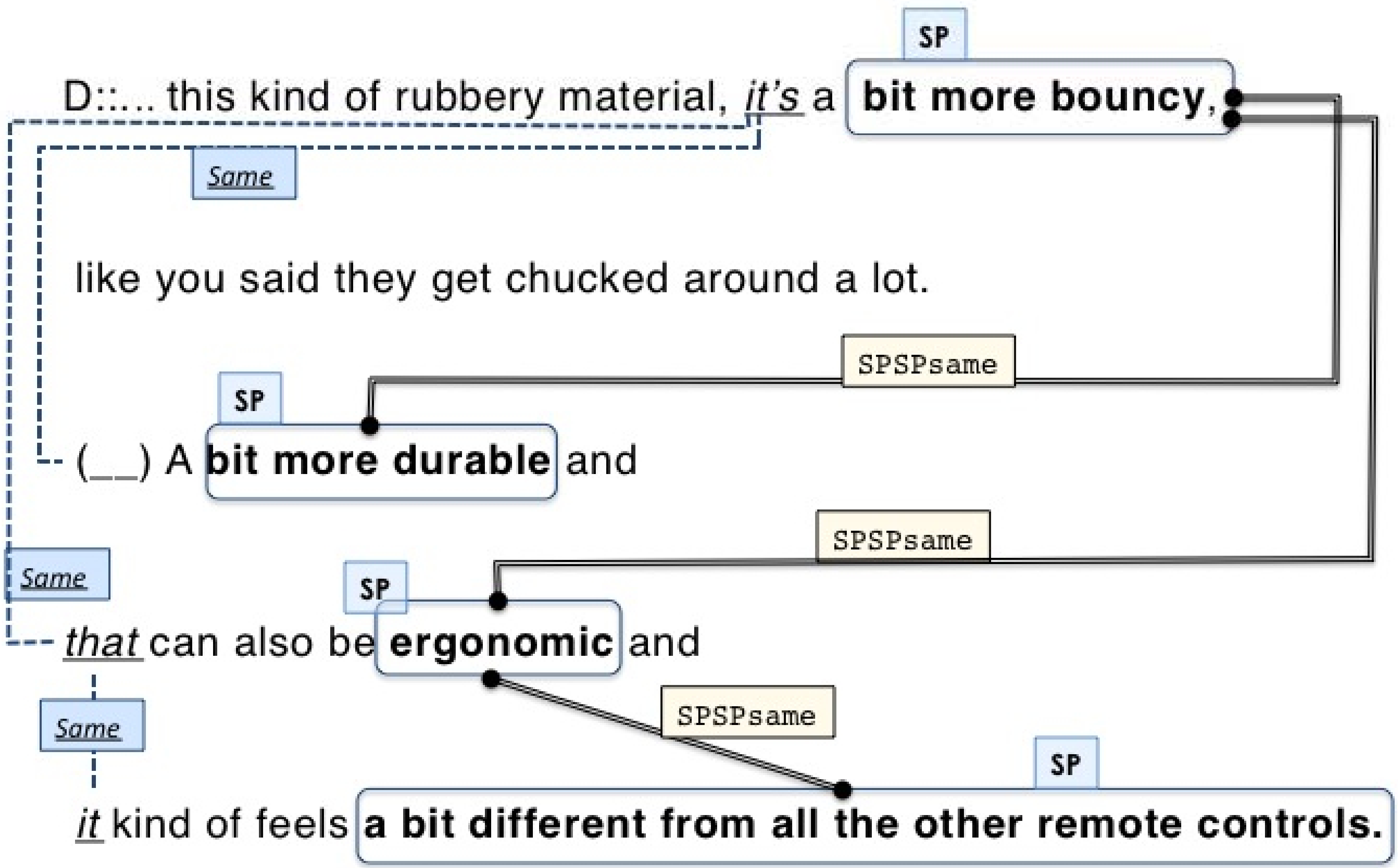


Learning and Inference **Hard**

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

Opinion Analysis



Roadmap

- The Problem
- The Components
- Putting It All Together
- **Open Questions**

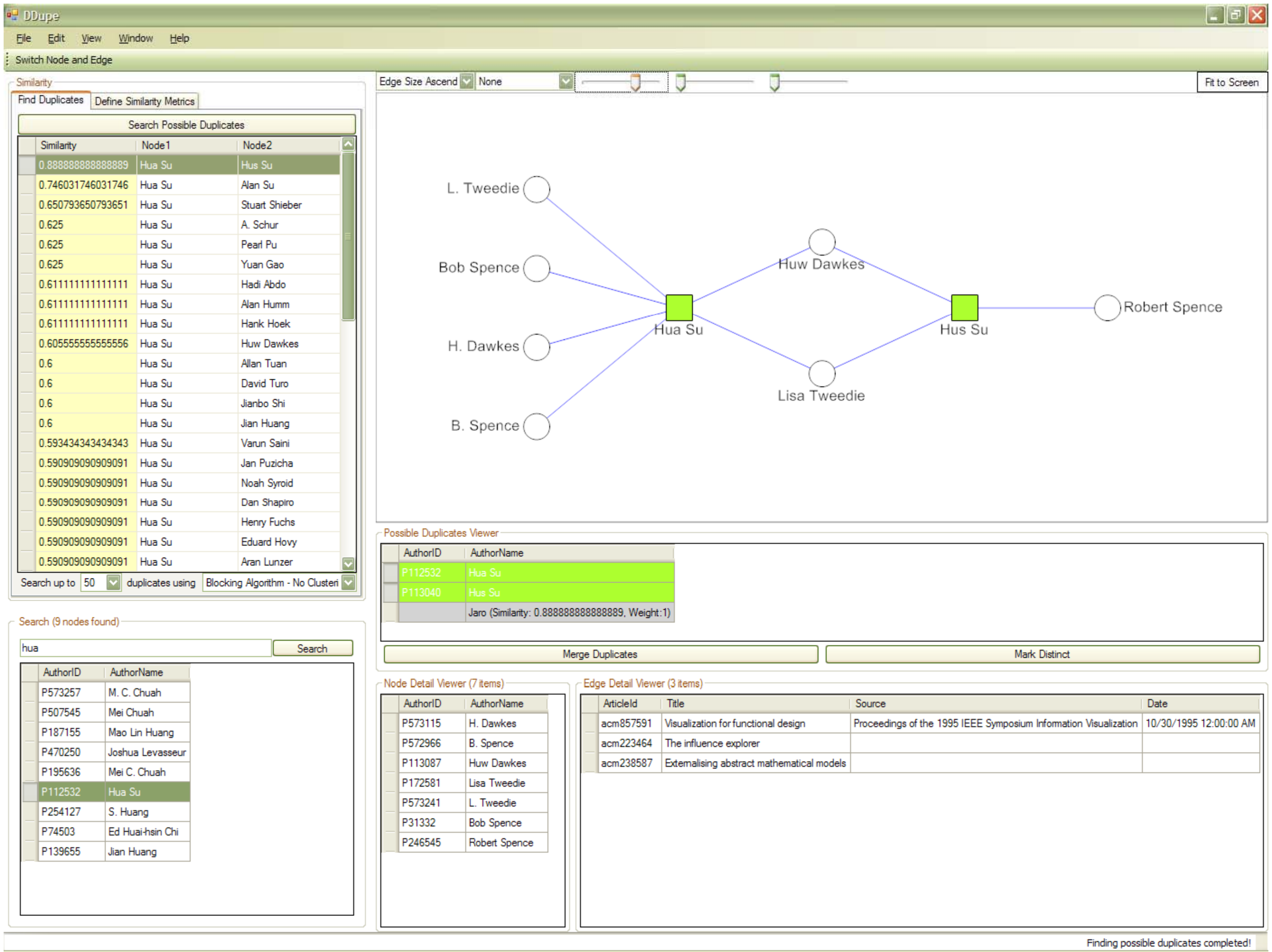
1. Query-time GI

- 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?
- Bhattacharya & Getoor, *Query-time Entity Resolution,* JAIR 2007.

2. Visual Analytics for GI

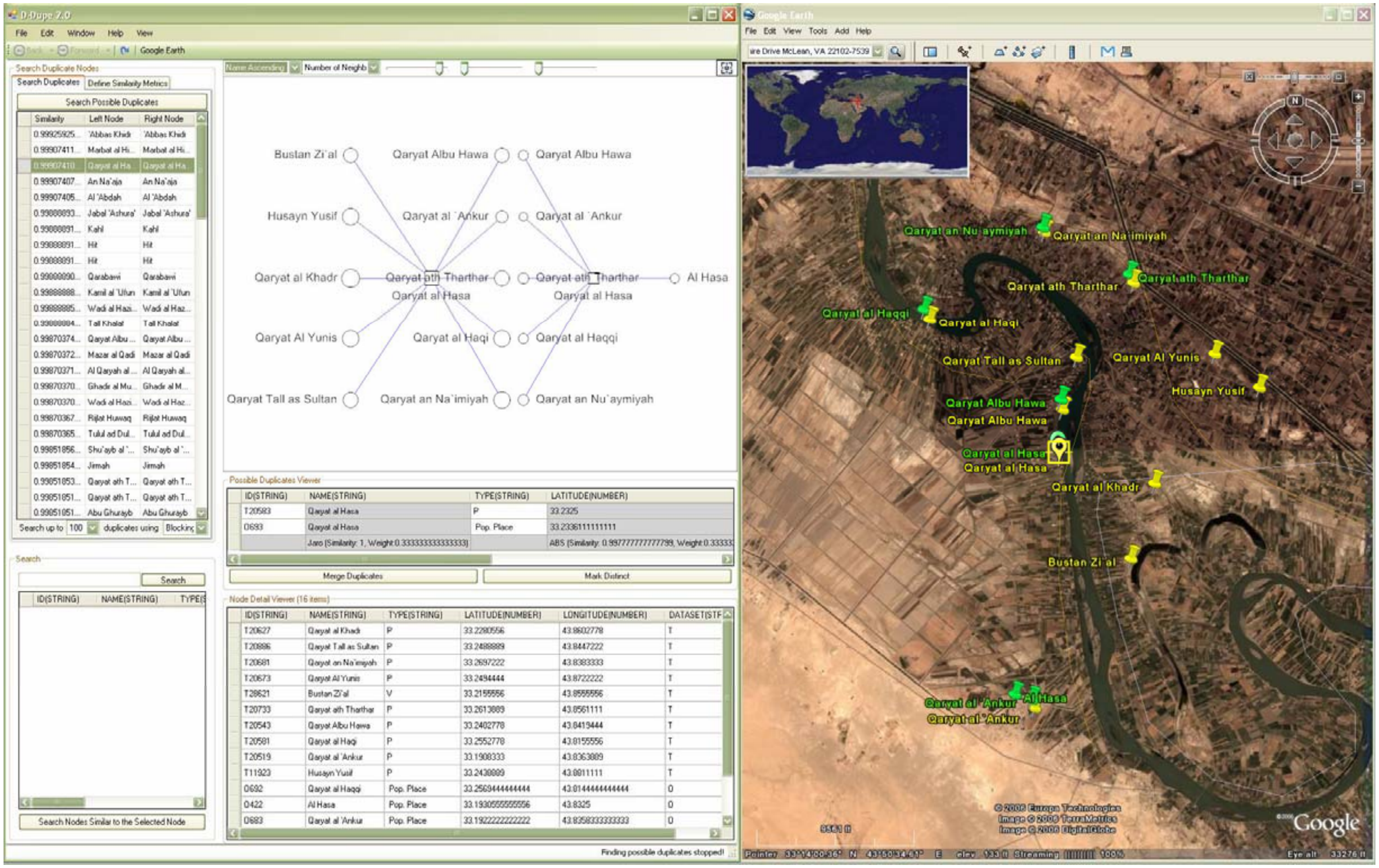
- Combining rich statistical inference models with visual interfaces that support knowledge discovery and understanding
- Because the statistical confidence we may have in any of our inferences may be low, it is important to be able to have a human in the loop, to understand and validate results, and 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



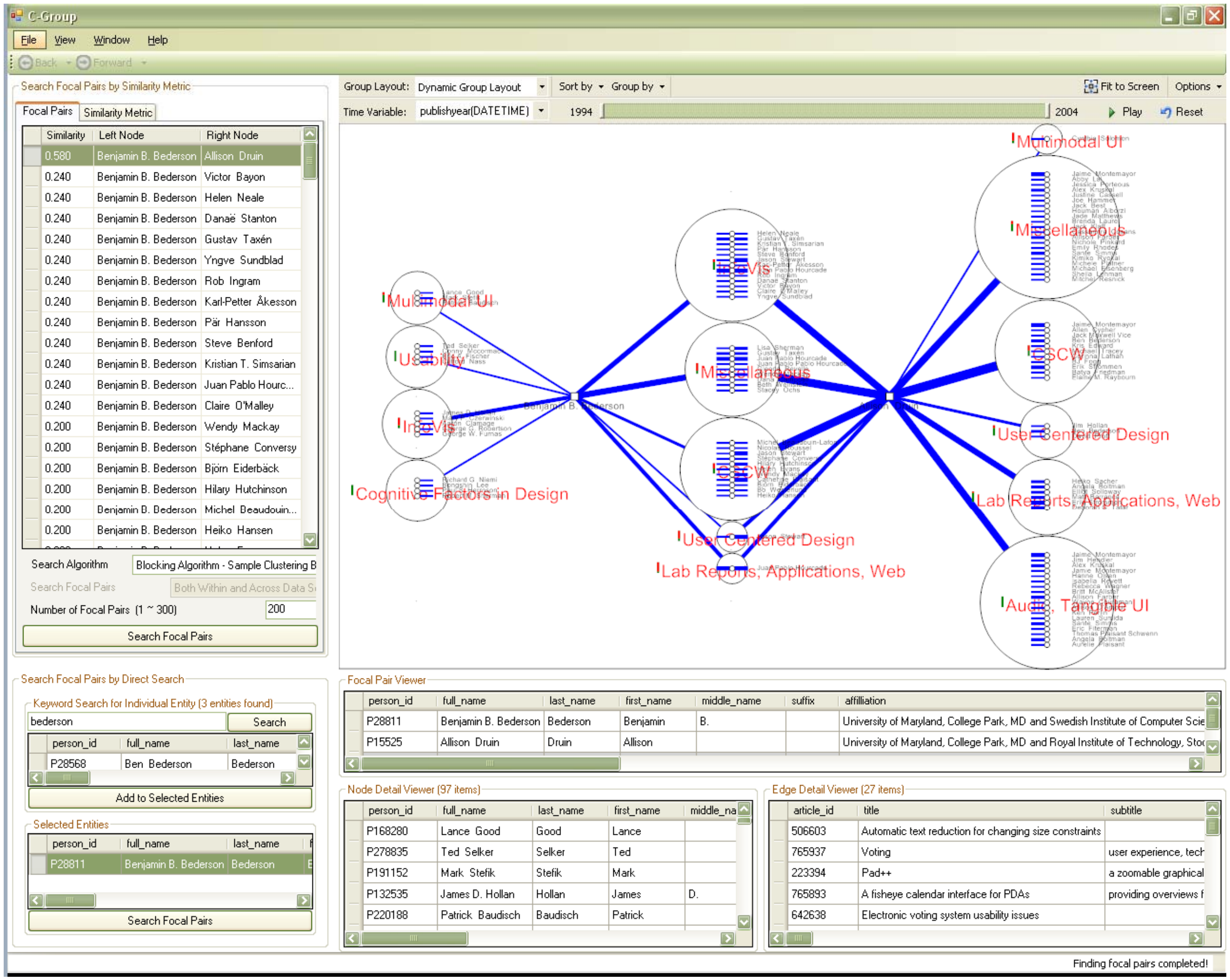
http://www.cs.umd.edu/projects/linqs/ddupe

GeoDDupe: Tool for Interactive Entity Resolution in Geospatial Data



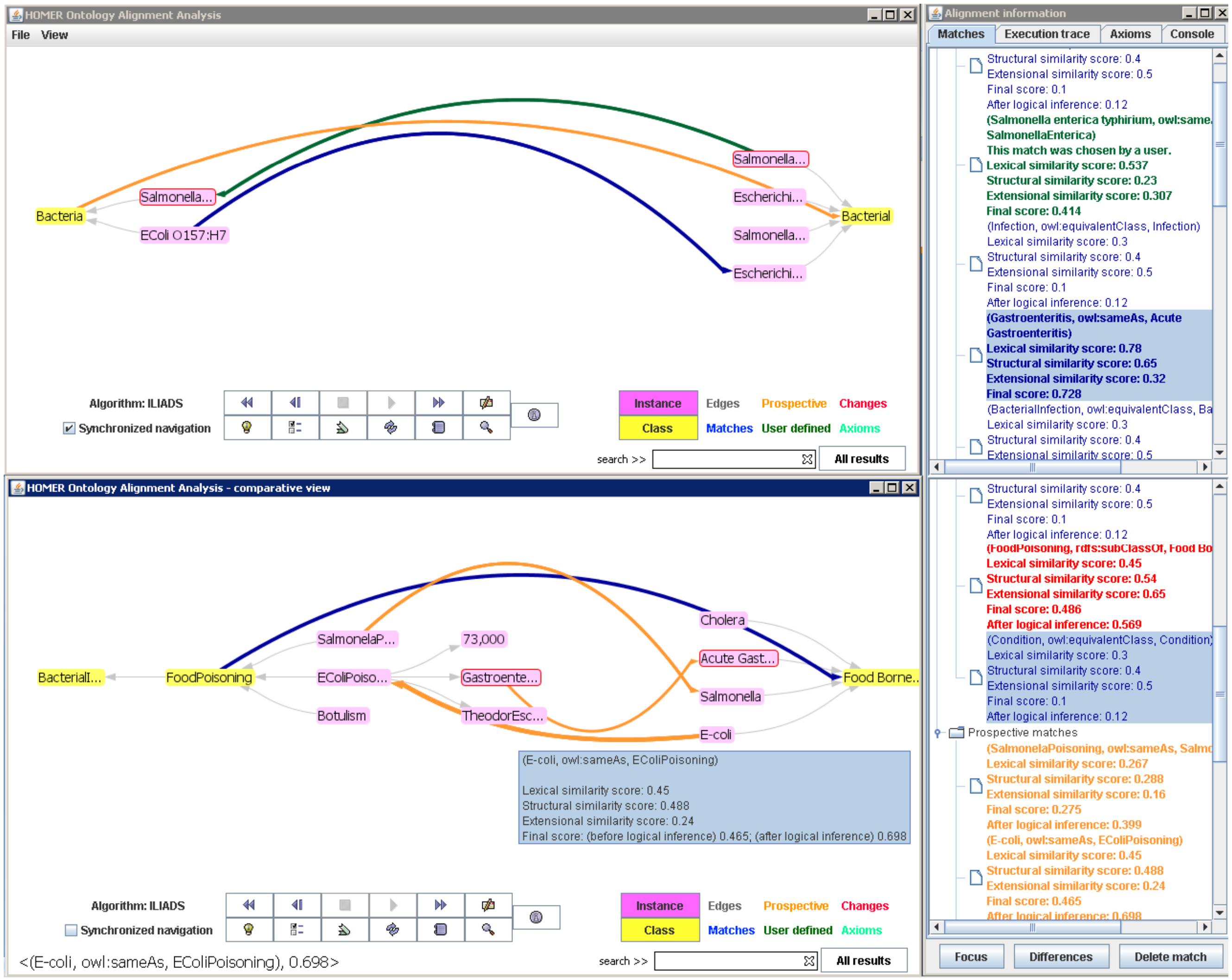
Kang, Sehgal, Getoor, IV 07
http://www.cs.umd.edu/projects/linqs/geoddupe

C-Group: A Visual Analytic Tool for Pairwise Analysis of Dynamic Group Membership



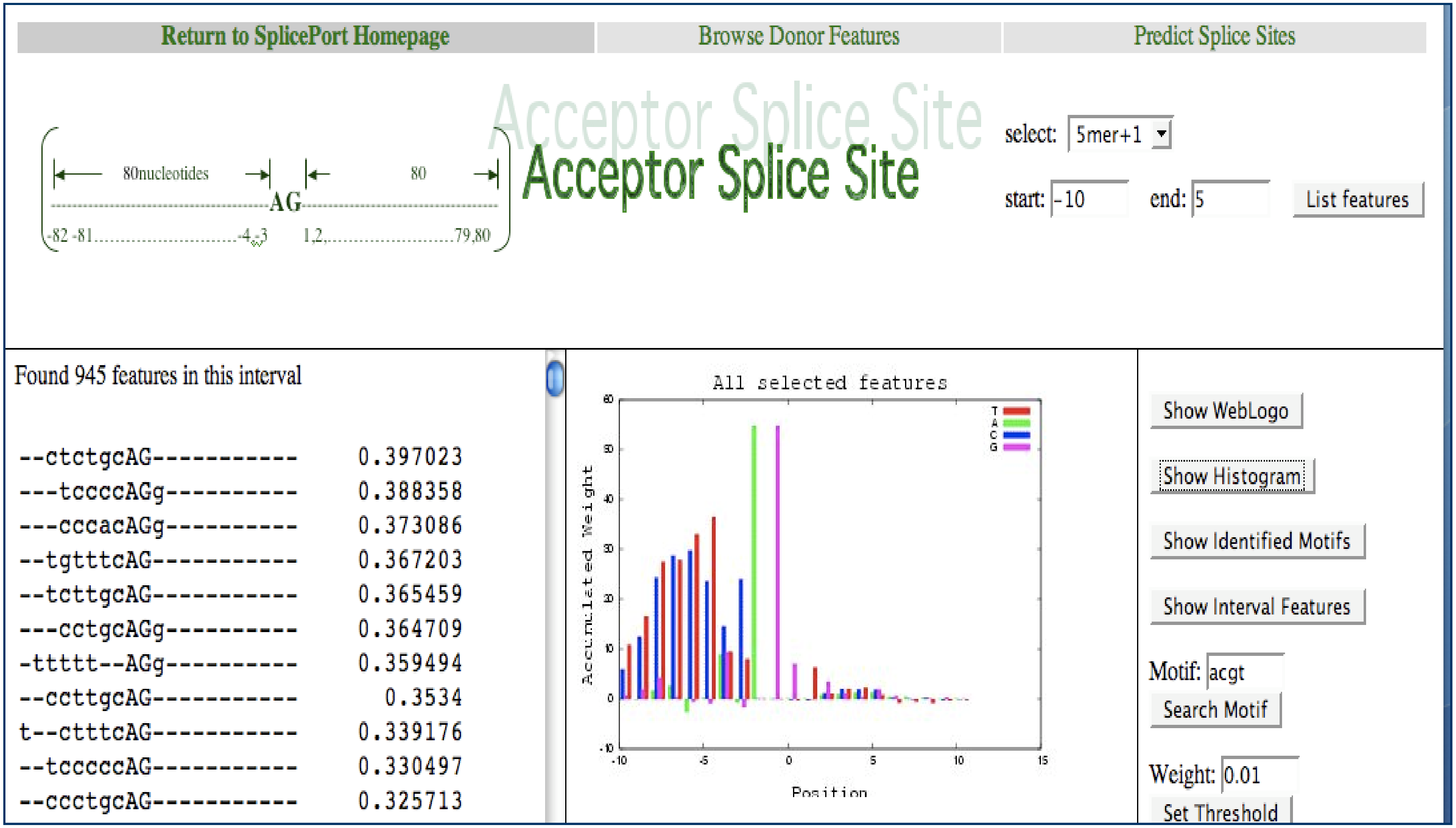
http://www.cs.umd.edu/projects/linqs/cgroup

HOMER: Tool for Ontology Alignment



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

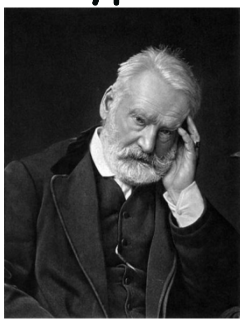
- A better theoretical understanding of when graph identification is feasible will also help us understand what must be done to maintain privacy of graph data

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

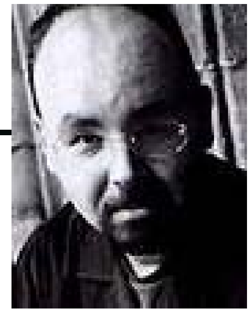
Link Re-Identification

Disease data

has hypertension



father-of



Communication data

?

call

Robert Lady

Search data

Query 1:

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

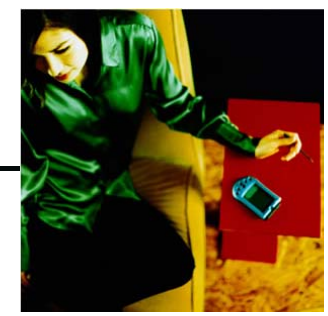
same-user

Query 2:

“myrtle beach golf course job listings”

Social network data

friends



Zheleva and Getoor, Preserving the Privacy of Sensitive Relationshops in Graph Data, PINKDD 2007

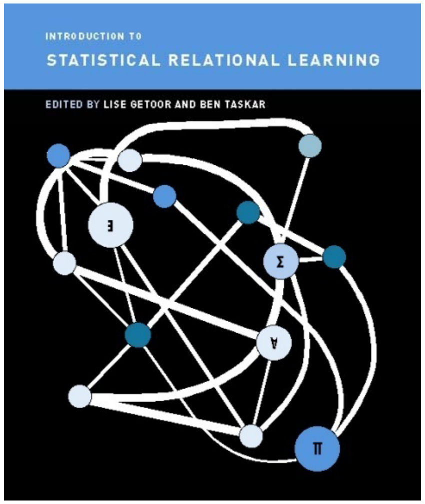
Summary: GIA & AI

- 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

Dagstuhl April 2007



- 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

- 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