



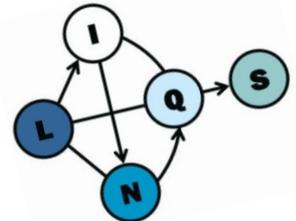
Finding the Information in Information Networks

Lise Getoor

University of Maryland, College Park



NASA GSFC Information Science & Technology Colloquium
February 6, 2008



Some Acknowledgements

o Students:

Indrajit Bhattacharya

Mustafa Bilgic

Rezarta Islamaj

Louis Licamele

Qing Lu

Galileo Namata

Vivek Sehgal

Prithvi Sen

Elena Zheleva

o Collaborators:

Chris Diehl

Tina Eliassi-Rad

John Grant

Hyunmo Kang

Renee Miller

Ben Shneiderman

Lisa Singh

Ben Taskar

Octavian Udrea

o Funding Sources:



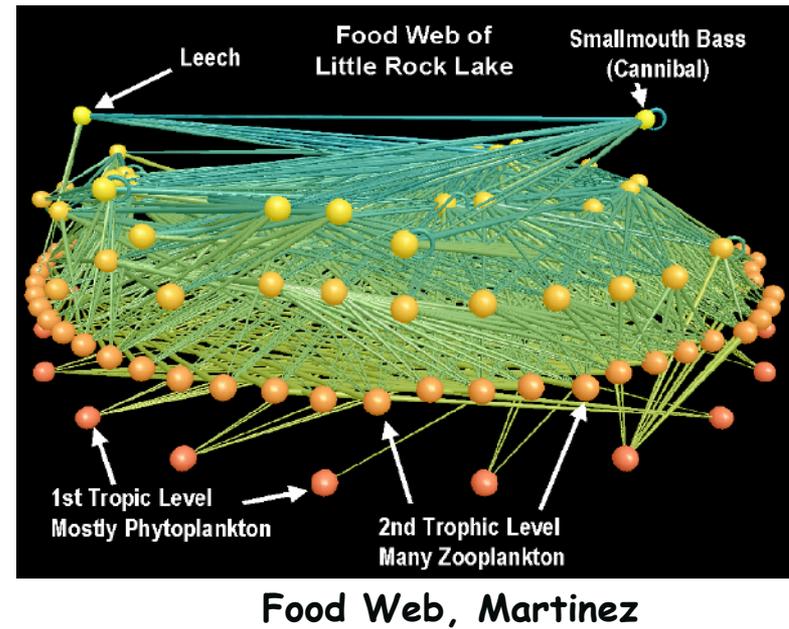
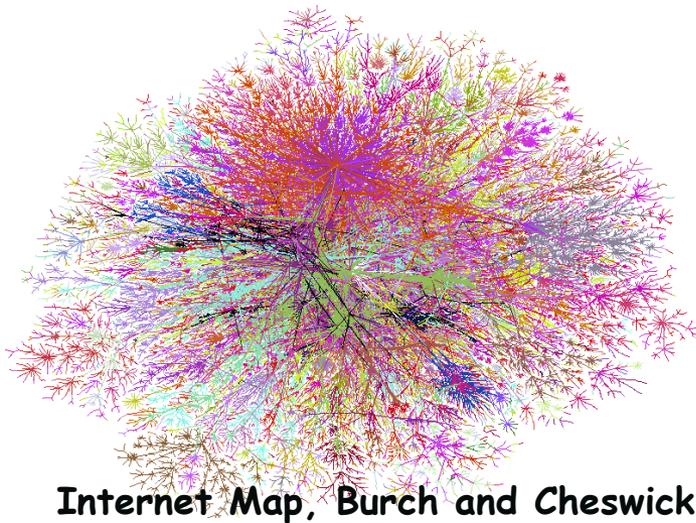
Google

Microsoft®
Research

KDD Program

● ● ● 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/>

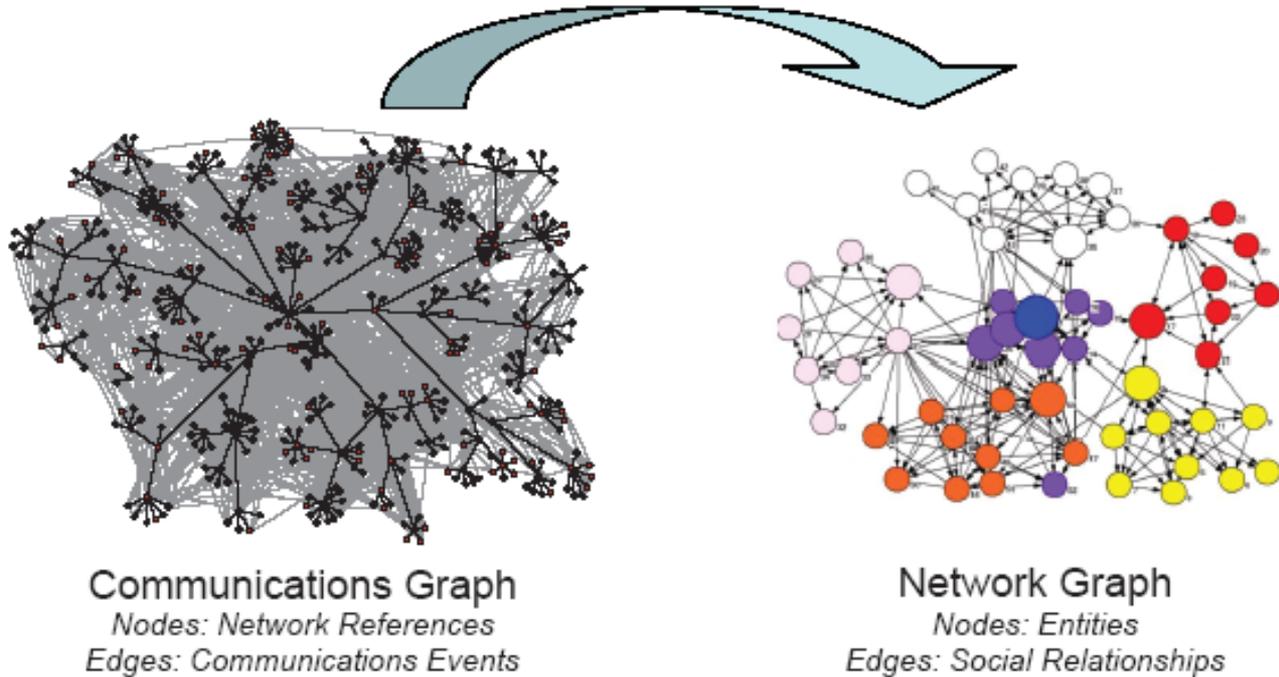
● ● ● Wealth of Data

- 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

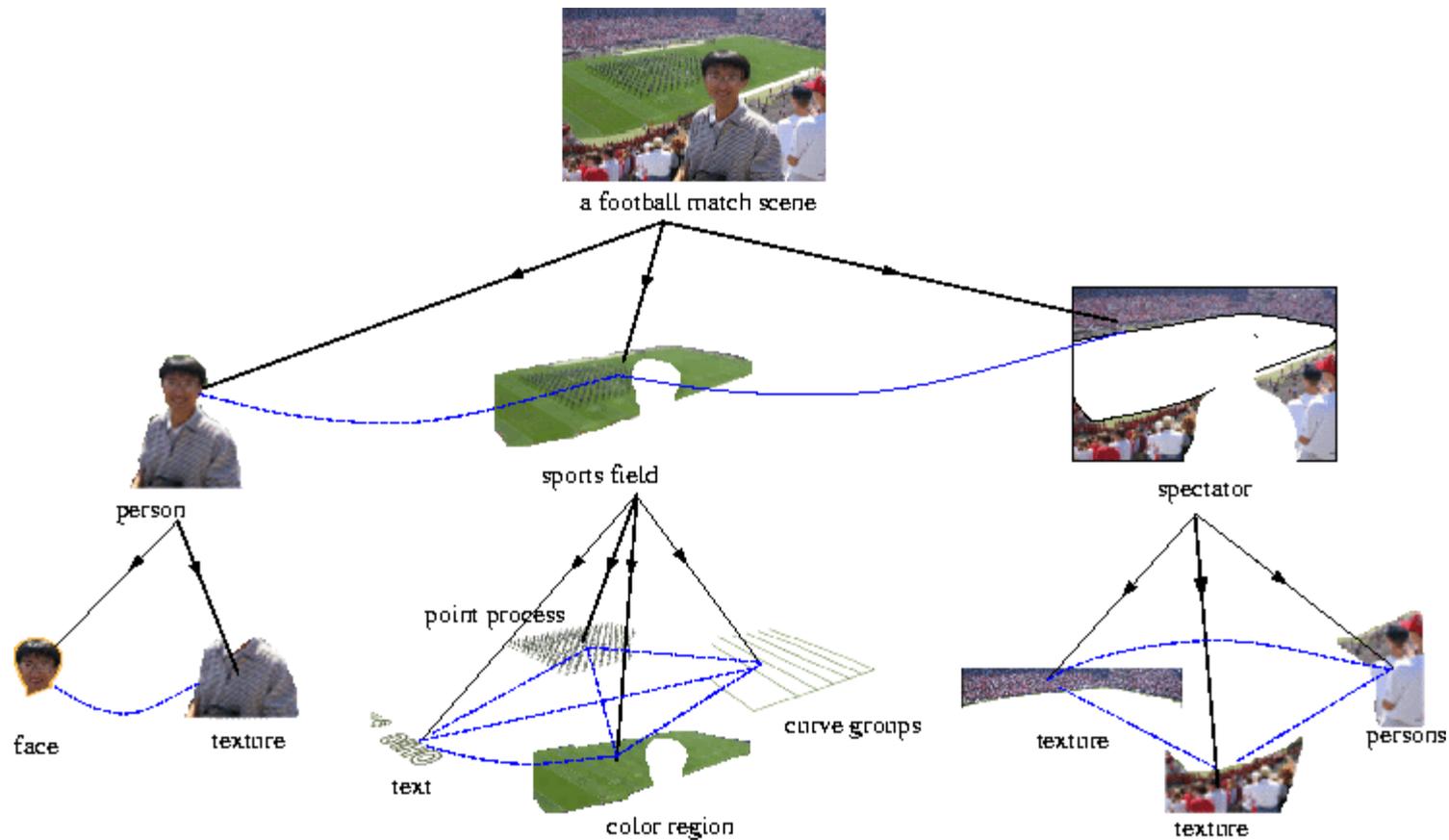
Graph Transformations



Data Graph \Rightarrow Information Graph

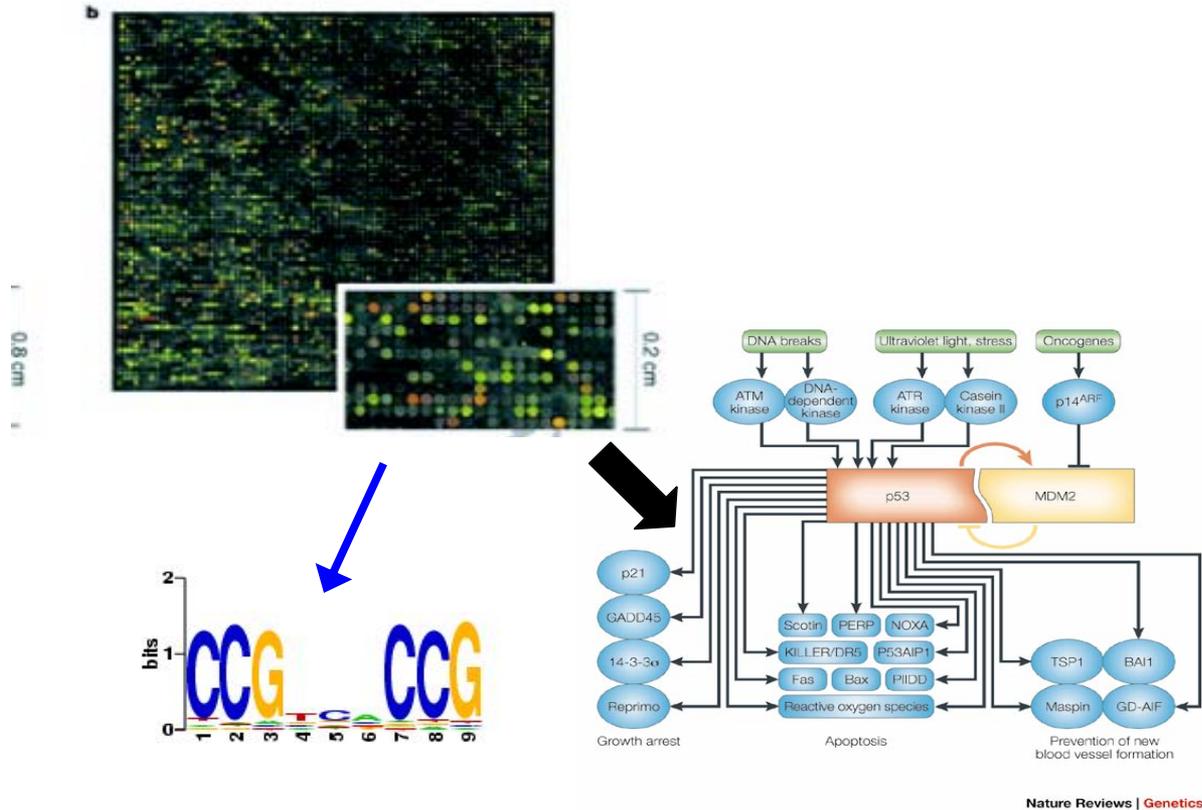
1. **Entity Resolution:** mapping email addresses to people
2. **Link Prediction:** predicting social relationship based on communication
3. **Collective Classification:** labeling nodes in the constructed social network

● ● ● Vision: Image Parsing



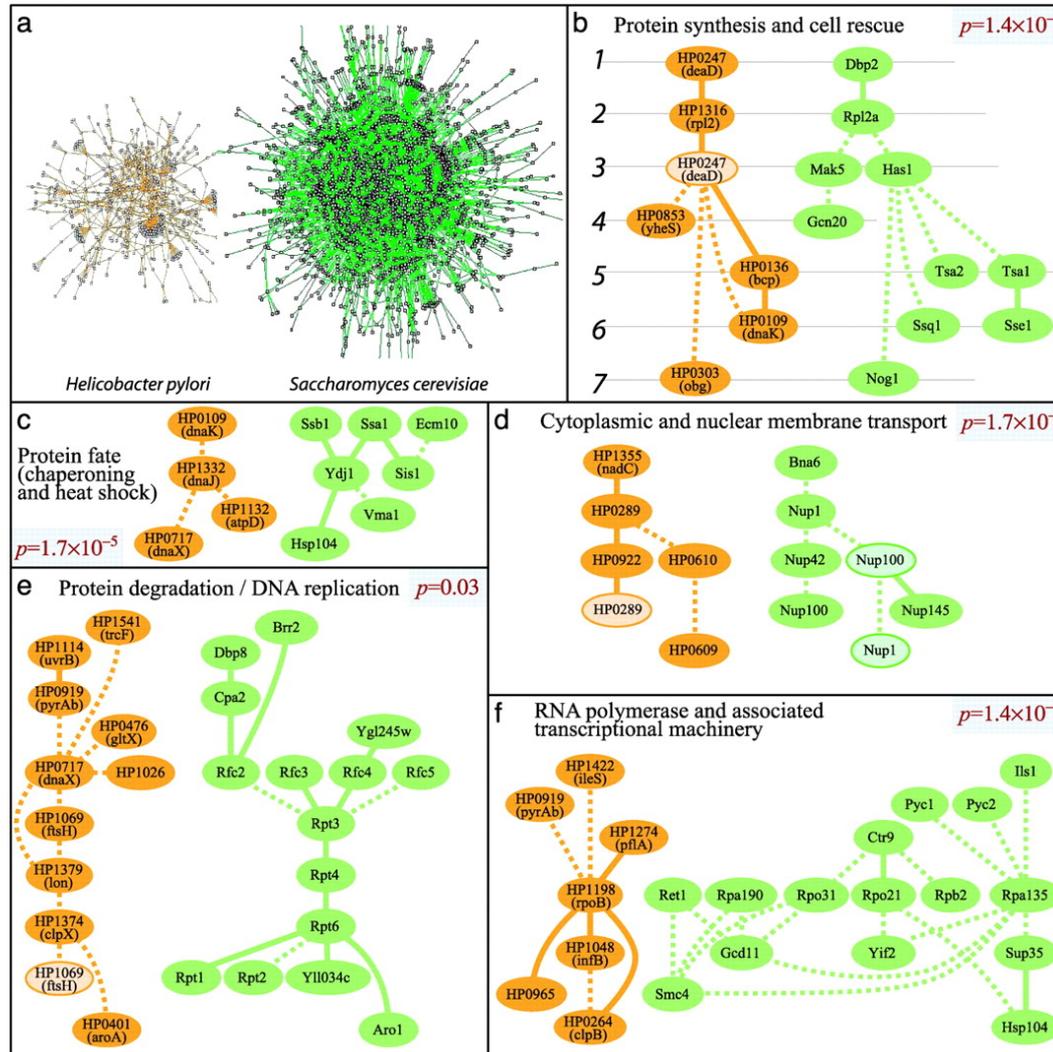
Graph Partitioning + Graph Matching

Bio: Graph Identification



Biological Networks: protein-protein, transcriptional regulation, signaling

Bio: Graph Alignment



● ● ● Roadmap

- The Problem

- **The Components**

- Entity Resolution
- Collective Classification
- Link Prediction

- Putting It All Together

- Open Questions

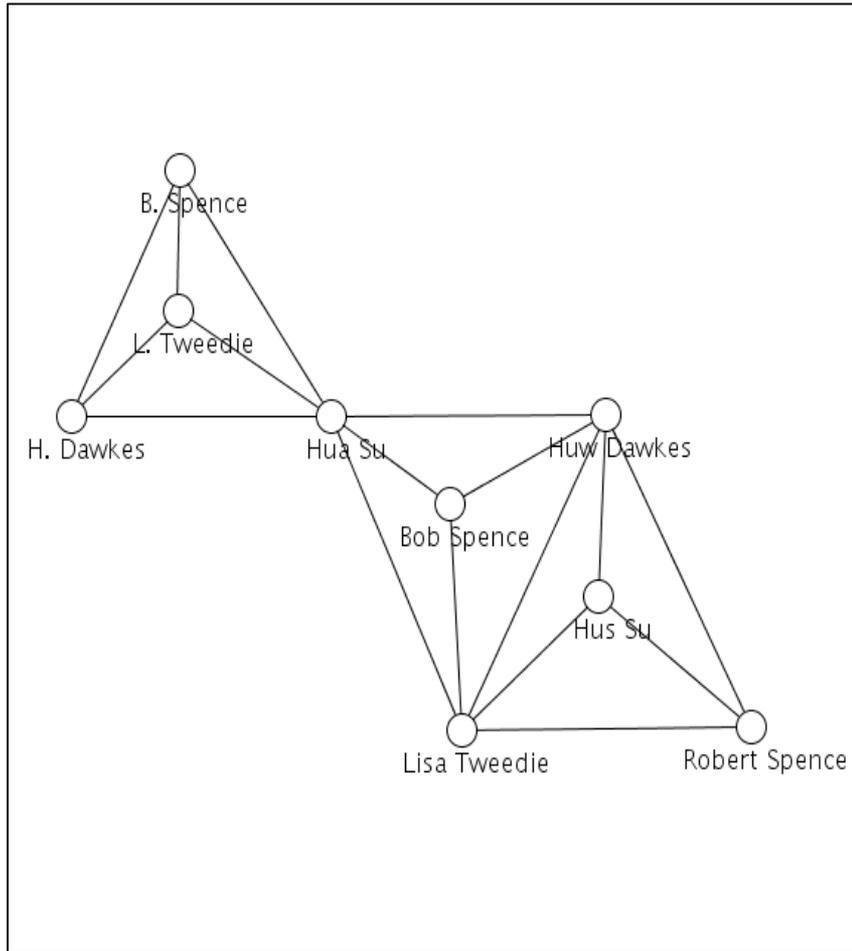
- ● ● Entity Resolution

- **The Problem**

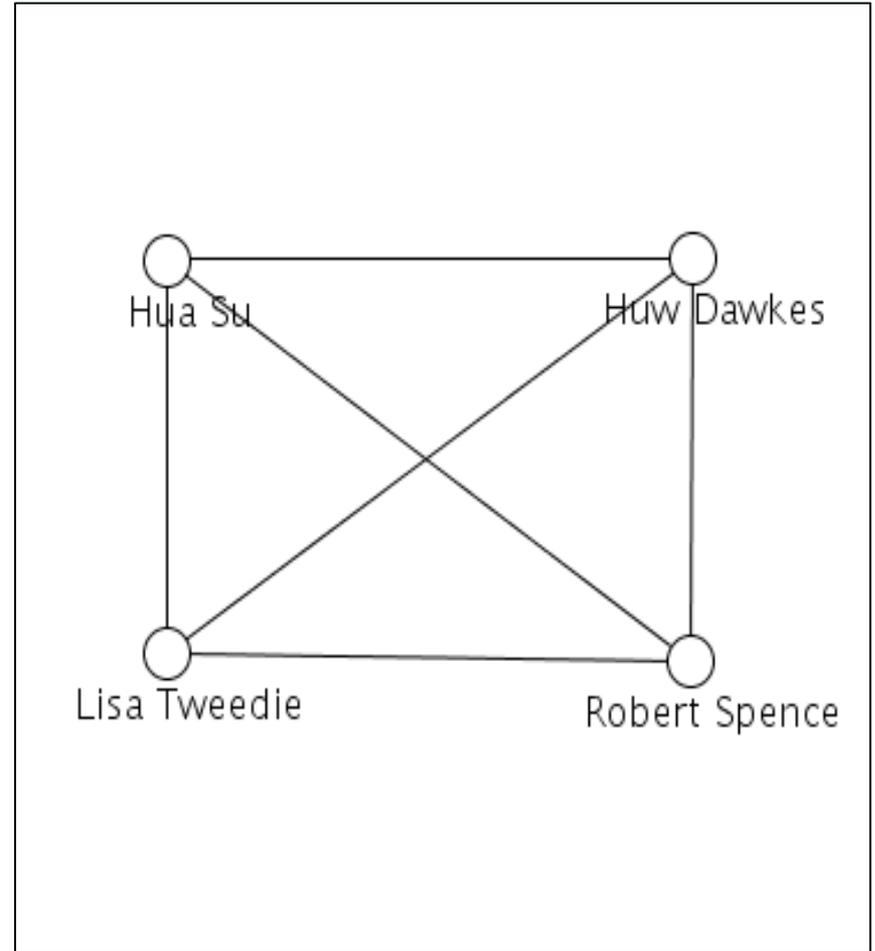
- Relational Entity Resolution

- Algorithms

● ● ● InfoVis Co-Author Network Fragment

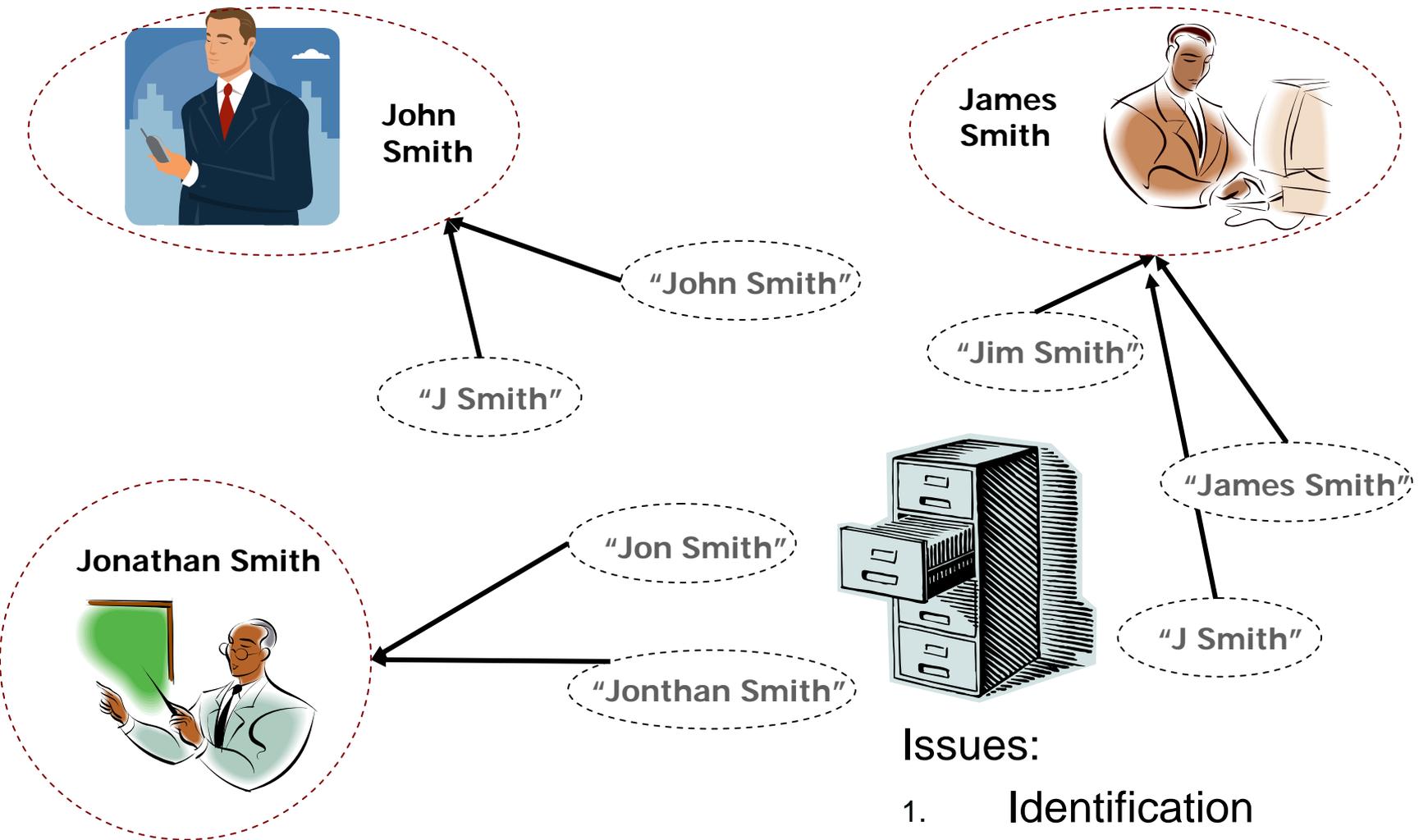


before



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**

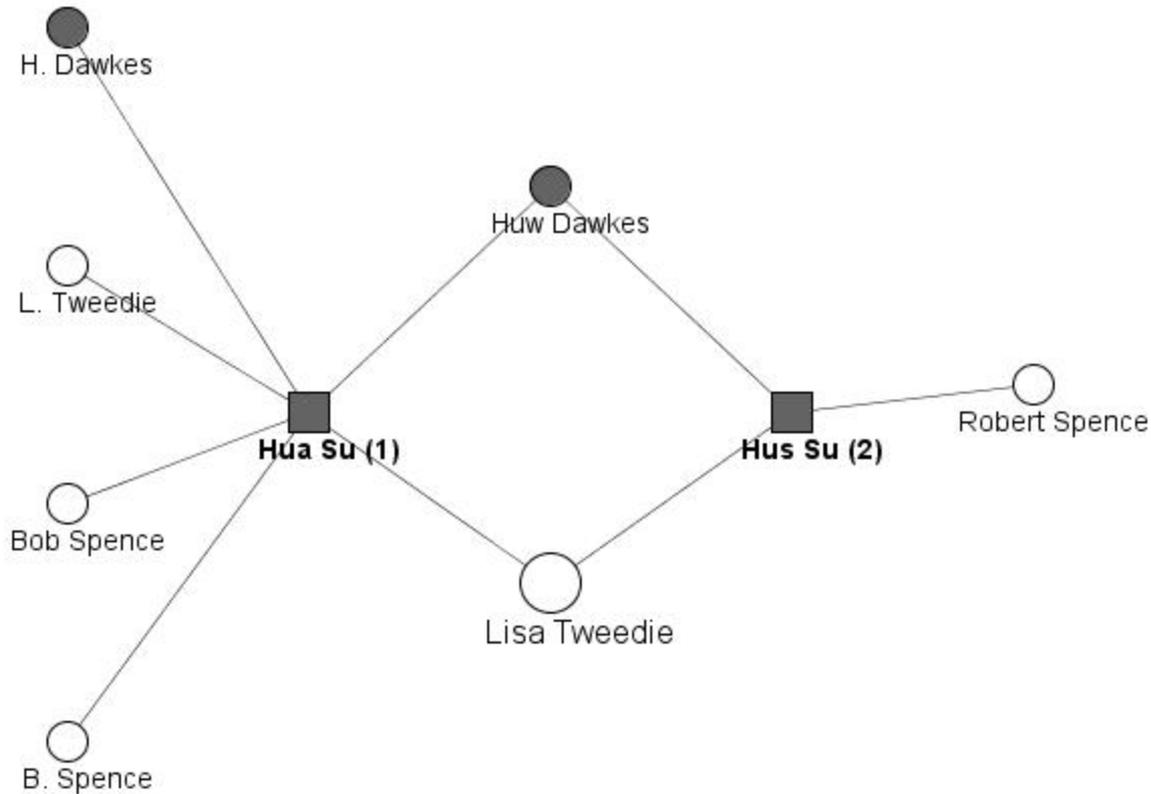
- Algorithms

● ● ● 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

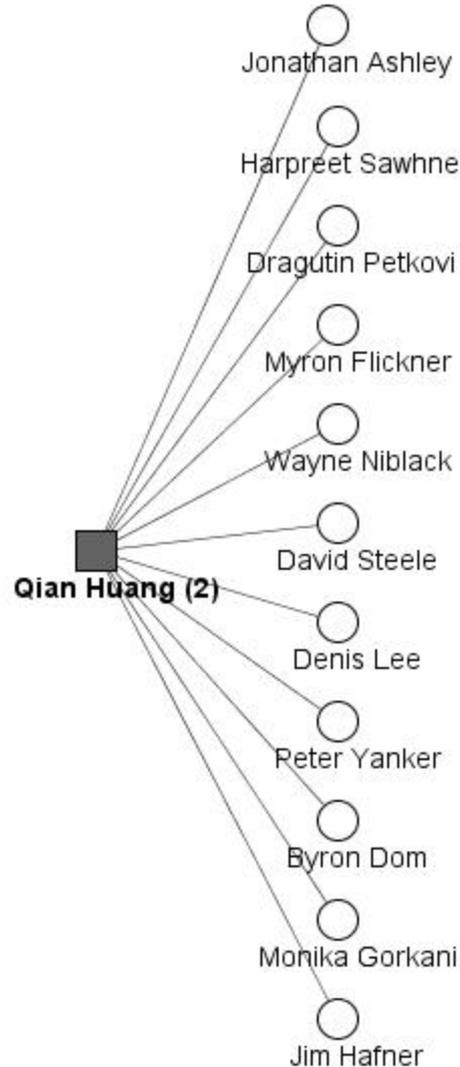
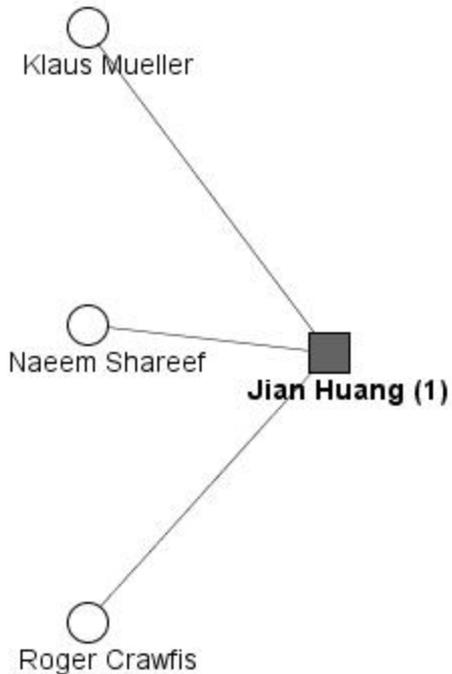
Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05

● ● ● Relational Identification



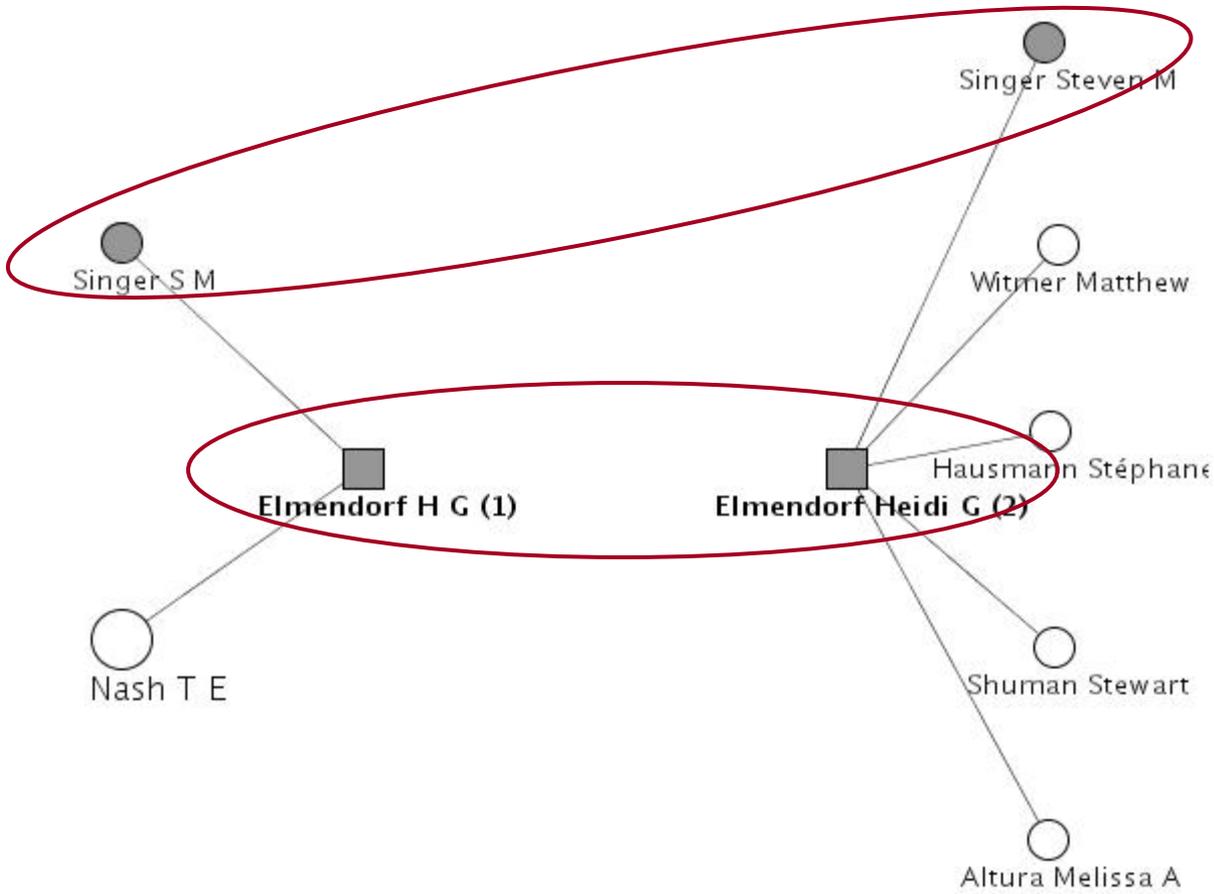
Very similar names.
Added evidence from
shared co-authors

● ● ● Relational Disambiguation



Very similar names
but no shared
collaborators

● ● ● Collective Entity Resolution



One resolution provides evidence for another => joint resolution

● ● ● Entity Resolution with Relations

○ Naïve Relational Entity Resolution

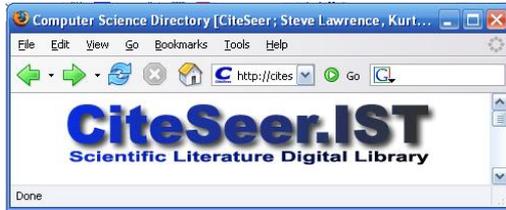
- Also compare attributes of related references
- Two references have co-authors w/ similar names

○ **Collective Entity Resolution**

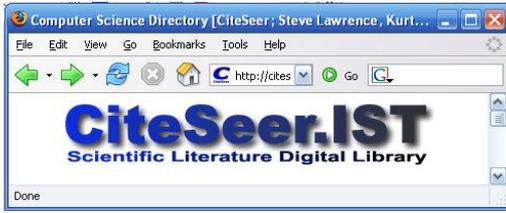
- Use **discovered entities** of related references
- Entities cannot be identified independently
- Harder problem to solve

● ● ● Entity Resolution

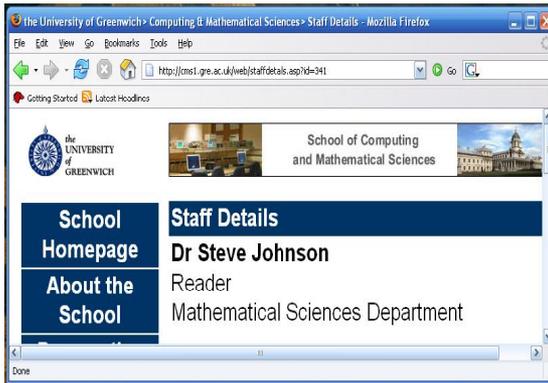
- The Problem
- Relational Entity Resolution
- **Algorithms**
 - **Relational Clustering (RC-ER)**
 - *Bhattacharya & Getoor, DMKD'04, Wiley'06, DE Bulletin'06, TKDD'07*



- 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 Unified 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



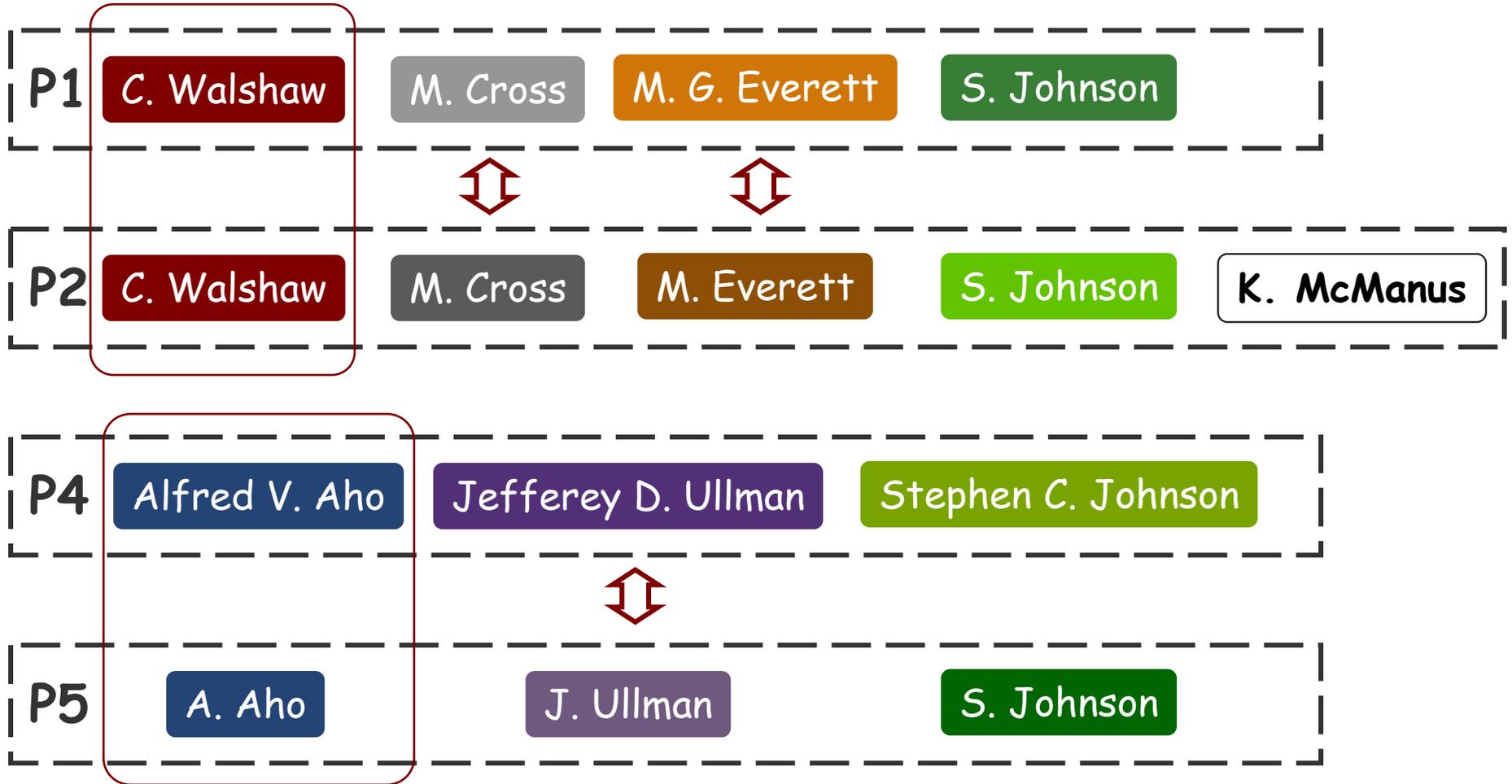
P5: “*Deterministic Parsing of Ambiguous Grammars*”, A. Aho, **S. Johnson**, J. Ullman

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

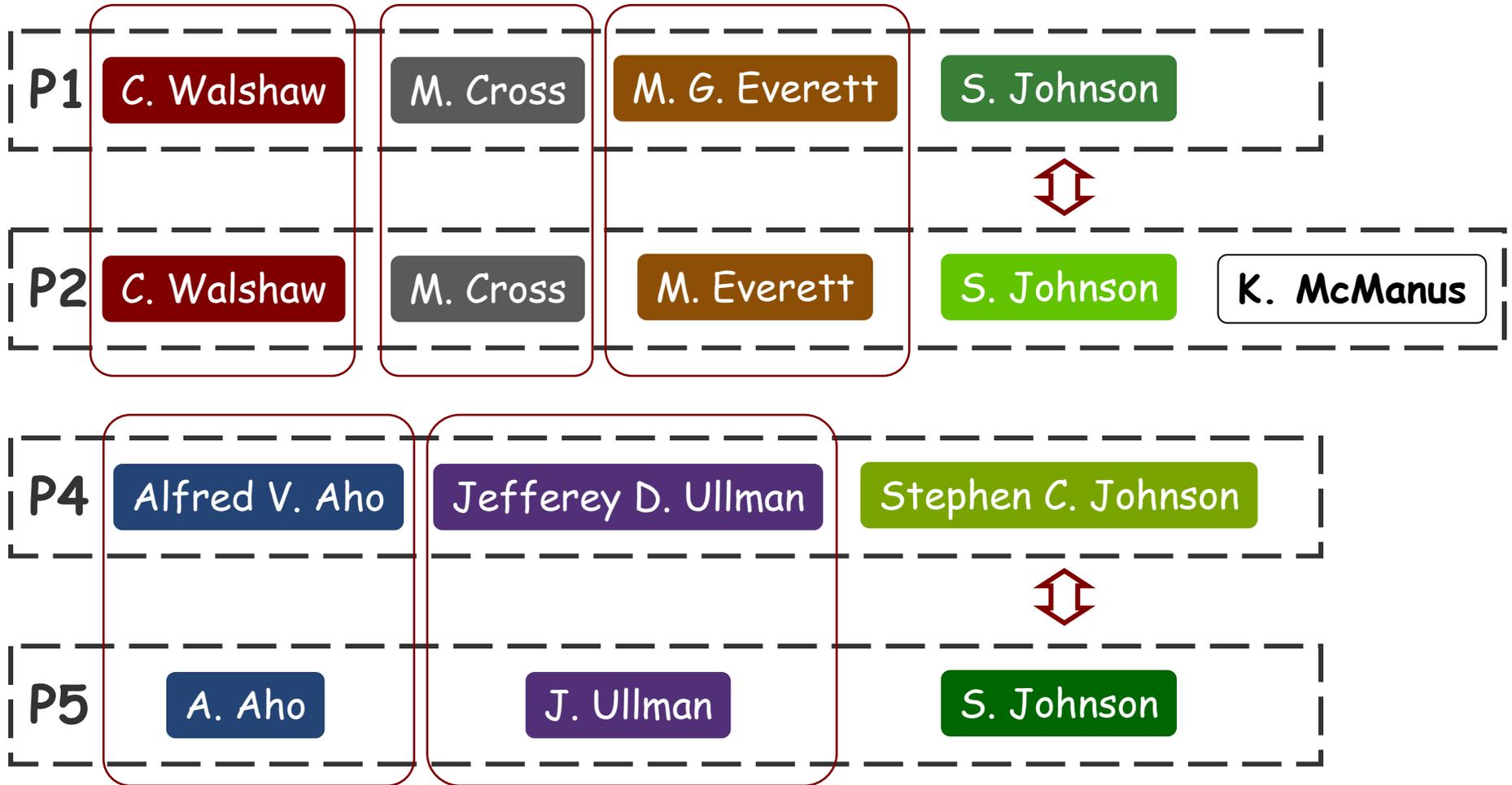
● ● ● Relational Clustering (RC-ER)



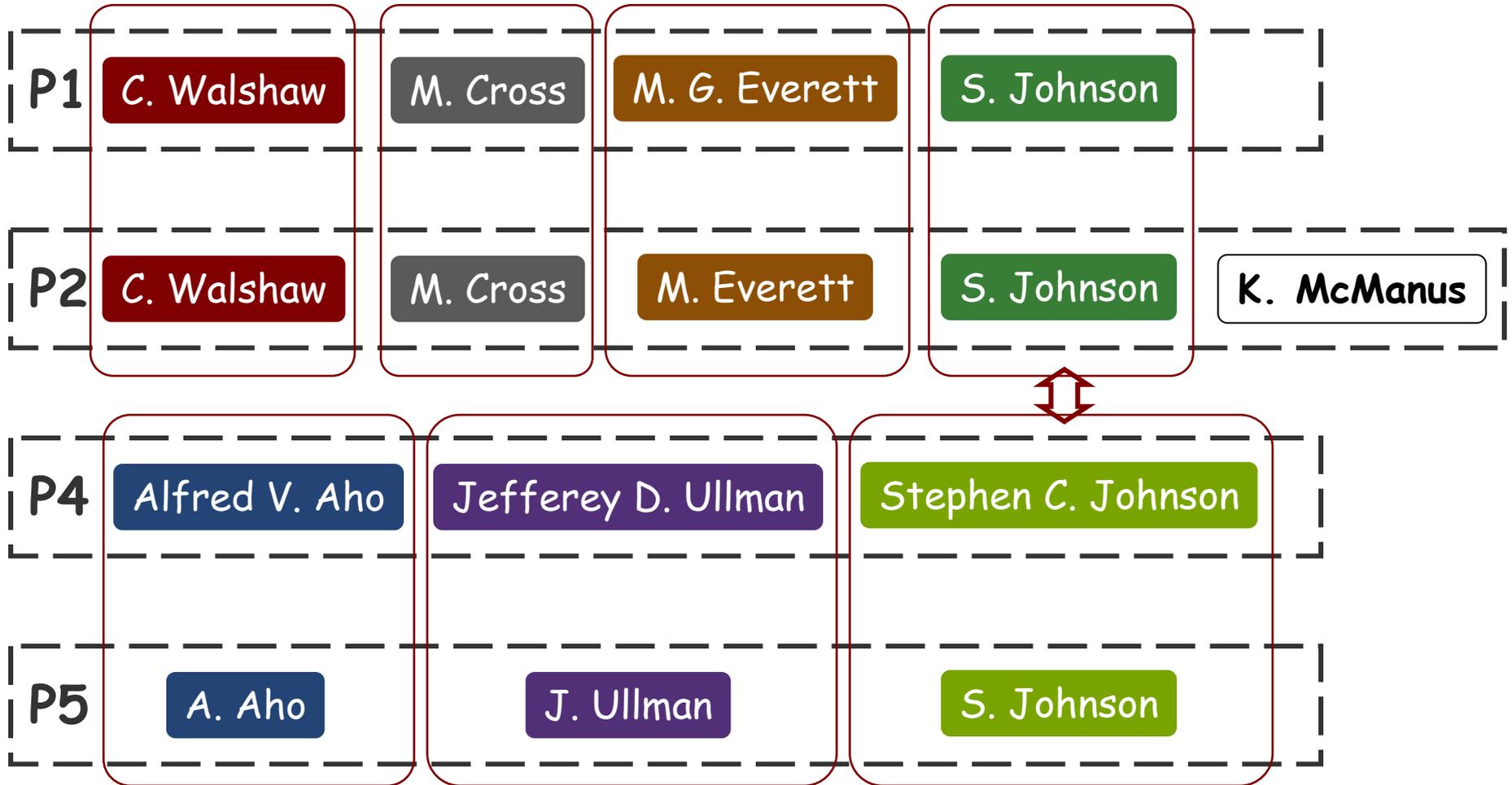
Relational Clustering (RC-ER)



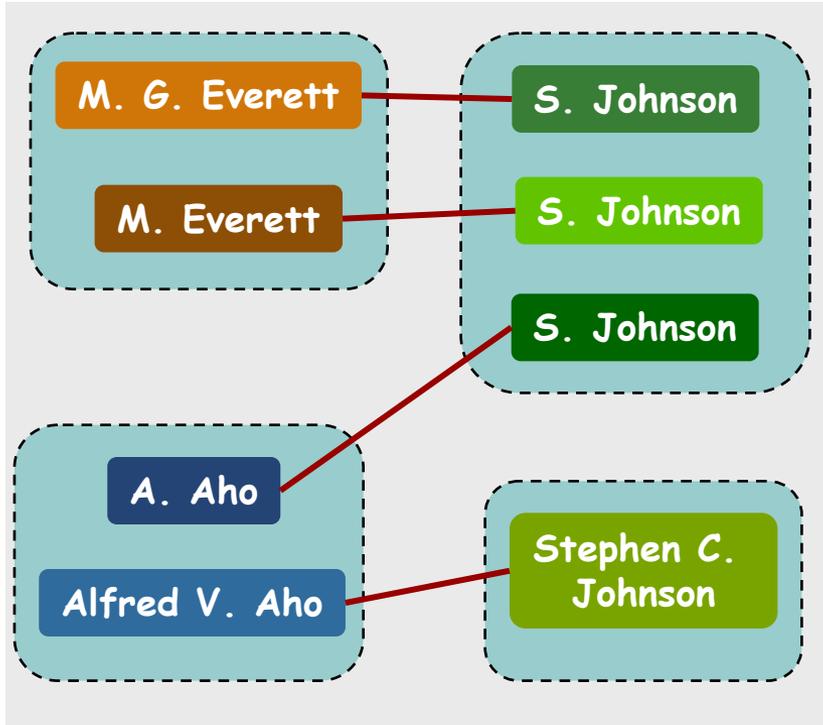
● ● ● Relational Clustering (RC-ER)



● ● ● Relational Clustering (RC-ER)

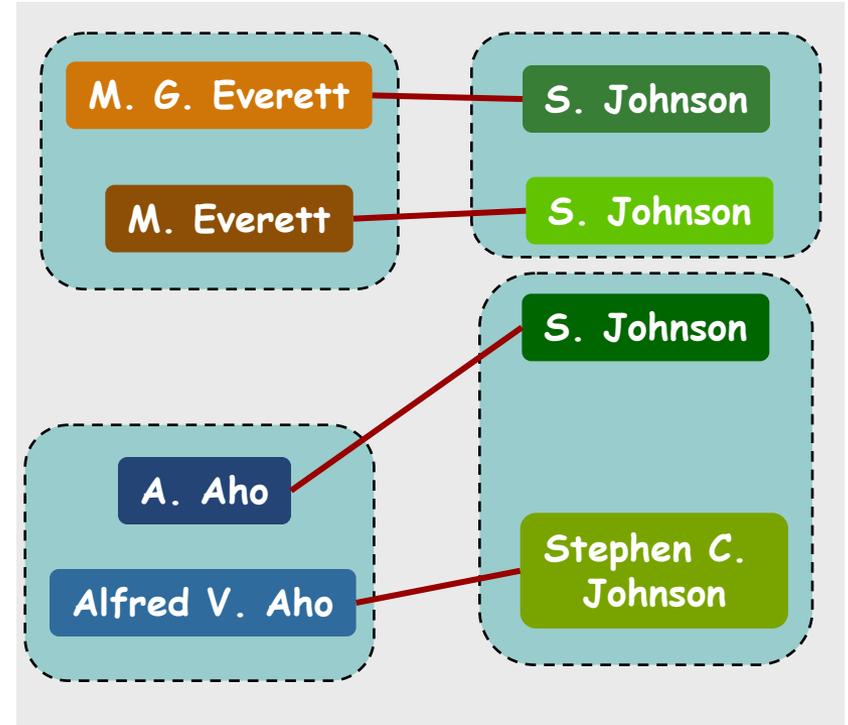


● ● ● Cut-based Formulation of RC-ER



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

Discovering Groups from Relations

Stephen P Johnson

Chris Walshaw

Kevin McManus

Mark Cross

Martin Everett

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

Stephen C Johnson

Alfred V Aho

Ravi Sethi

Jeffrey D Ullman

Bell Labs Group

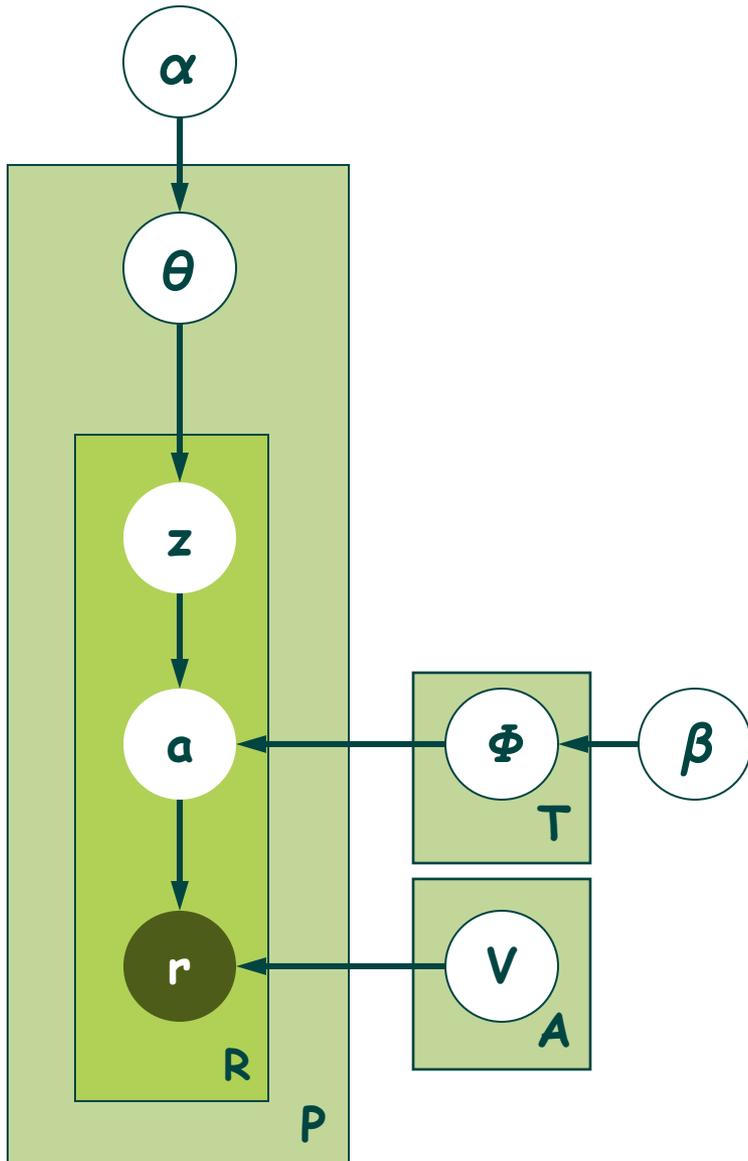


P4: Alfred V. Aho, **Stephen C. Johnson**,
Jefferey D. Ullman

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

● ● ● Roadmap

- The Problem

- **The Components**

 - Entity Resolution

 - **Collective Classification**

 - Link Prediction

 - Putting It All Together

- Open Questions

● ● ● Collective Classification

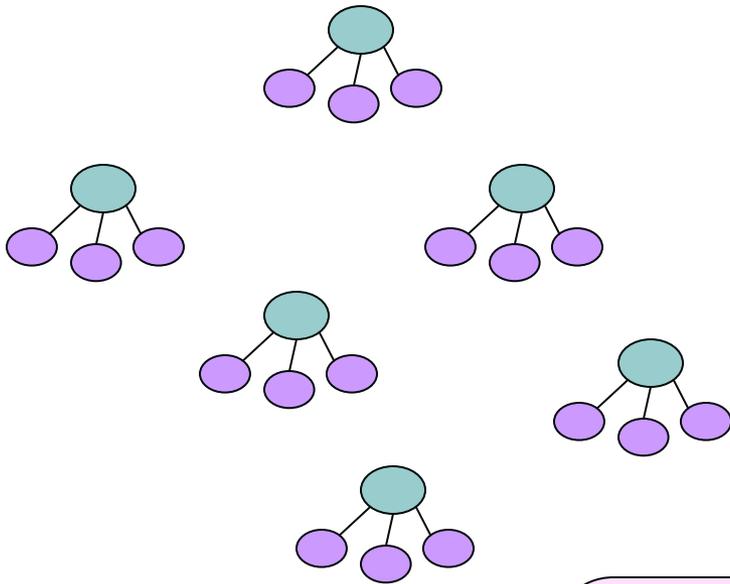
- **The Problem**

- Collective Relational Classification

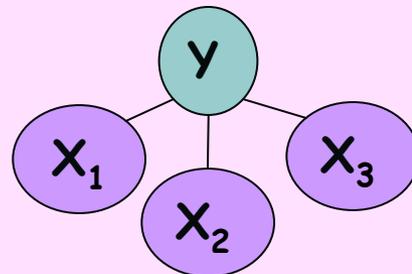
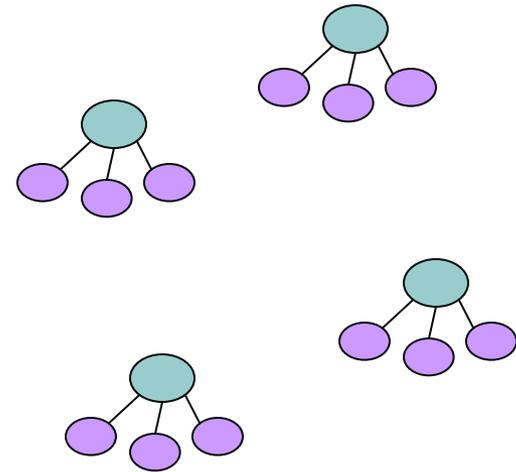
- Algorithms

● ● ● Traditional Classification

Training Data



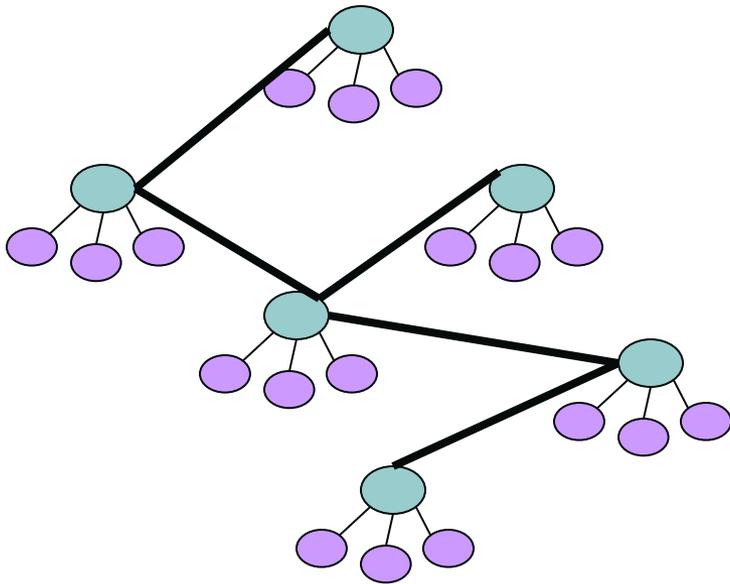
Test Data



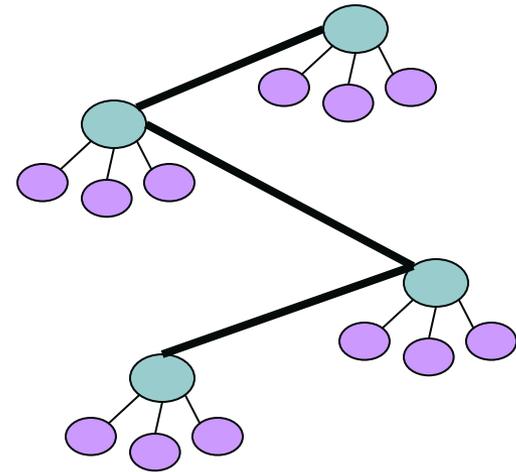
**Predict Y based on
attributes X_i**

● ● ● Relational Classification (1)

Training Data



Test Data



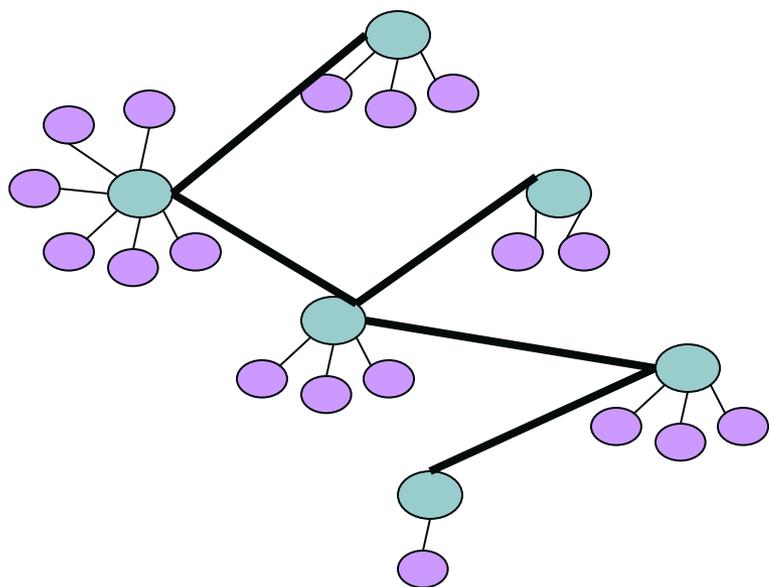
Correlations among linked instances

autocorrelation: labels are likely to be the same

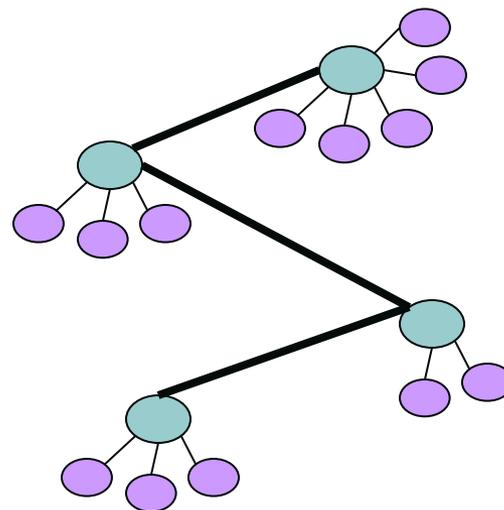
homophily: similar nodes are more likely to be linked

● ● ● Relational Classification (2)

Training Data



Test Data

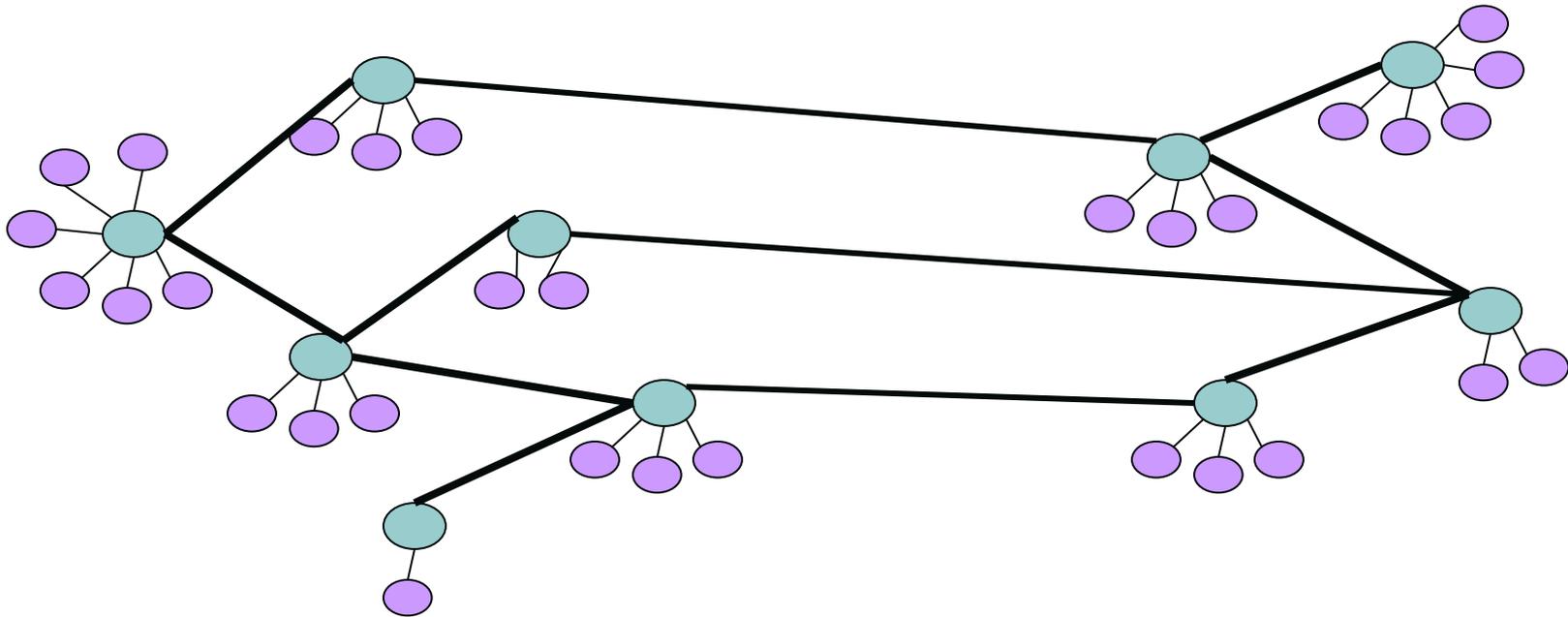


Irregular graph structure

● ● ● 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

● ● ● 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

● ● ● Feature Construction

- 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)

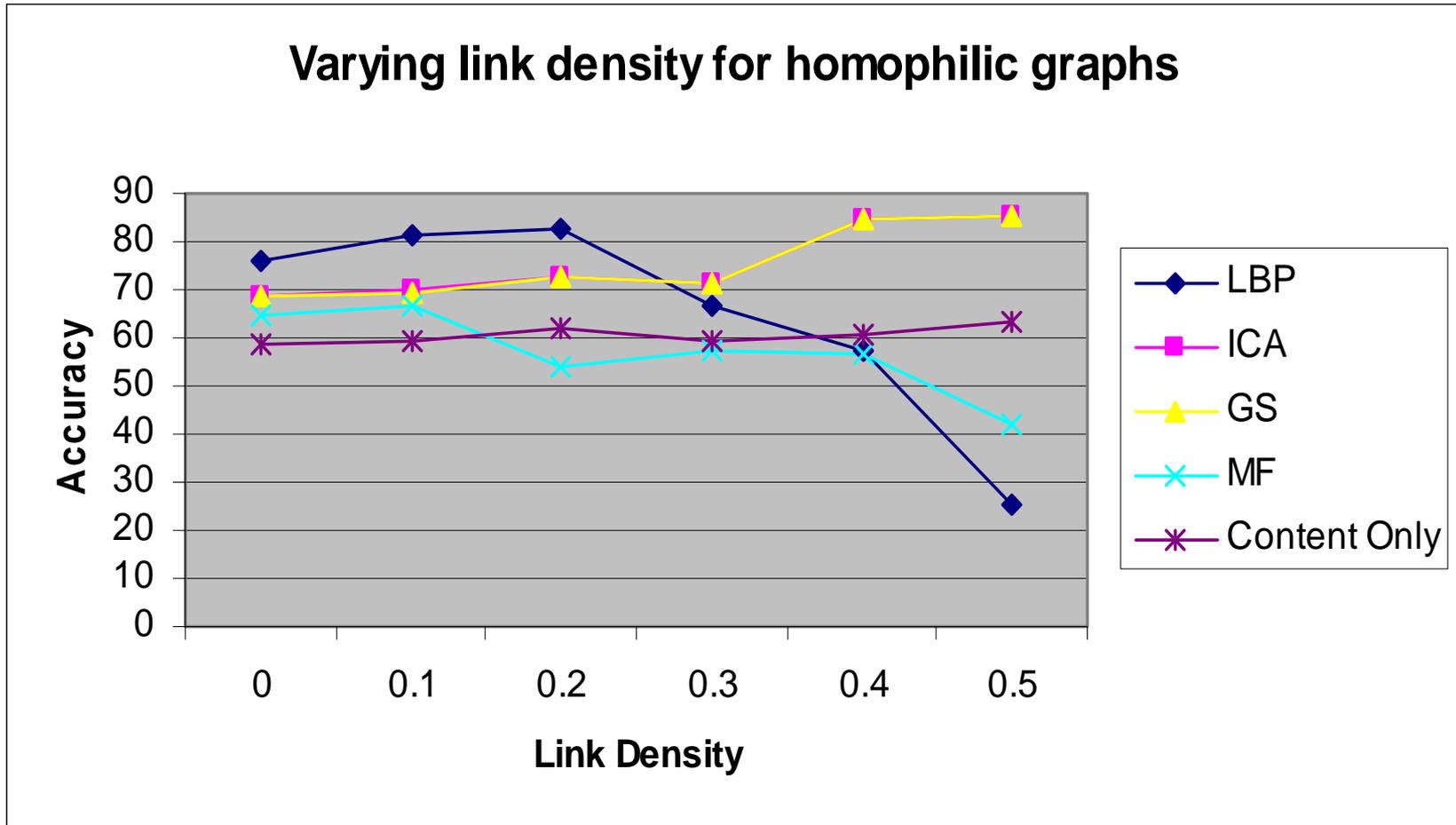
● ● ● Experimental Evaluation

- 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

| 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 |

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**

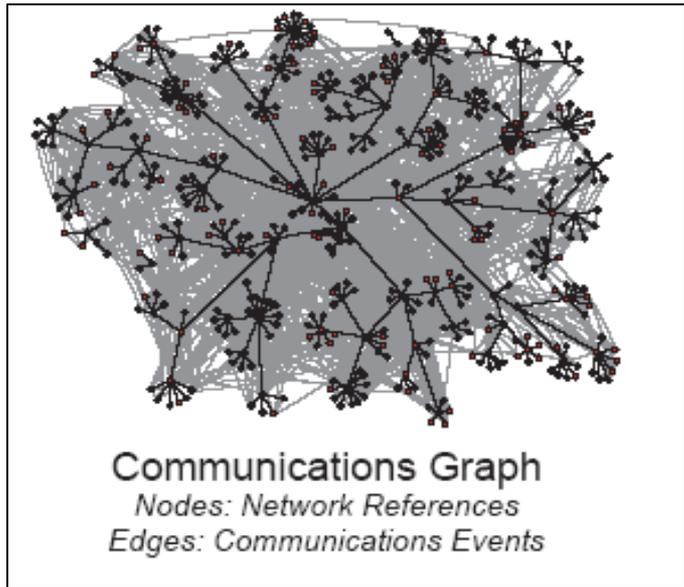
- Putting It All Together

- Open Questions

● ● ● Link Prediction

- **The Problem**
- Predicting Relations
- Algorithms
 - Link Labeling
 - Link Ranking
 - Link Existence

Links in Data Graph



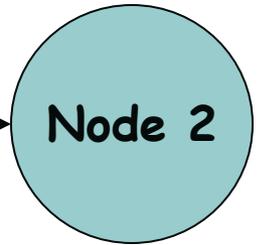
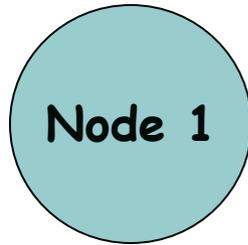
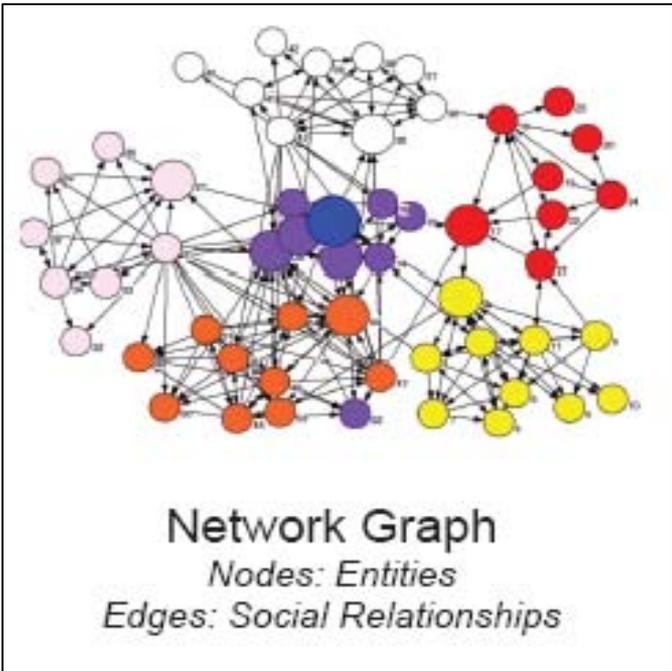
chris@enron.com ← Email → liz@enron.com

chris37 ← IM → lizs22

555-450-0981 ← TXT → 555-901-8812



• • • ⇒ Links in Information Graph



Chris



Elizabeth



Steve



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

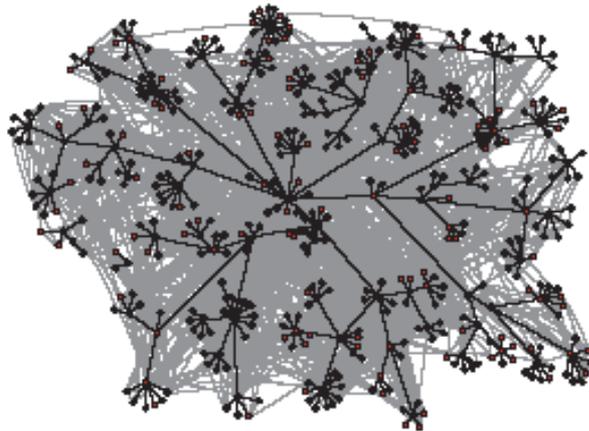
● ● ● Roadmap

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

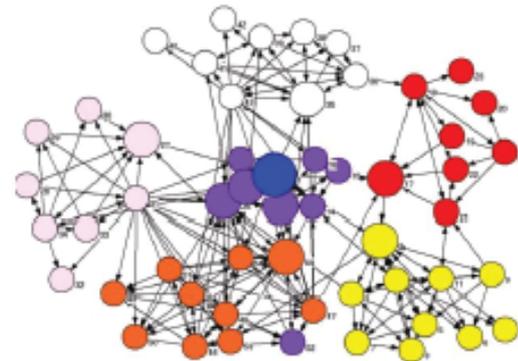
Putting Everything together....



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?
- 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

The screenshot displays the D-Dupe application window with the following components:

- Similarity Table:** A table for finding possible duplicates.

| Similarity | Node1 | Node2 |
|-------------------|--------|----------------|
| 0.888888888888889 | Hua Su | Hus Su |
| 0.746031746031746 | Hua Su | Alan Su |
| 0.650793650793651 | Hua Su | Stuart Shieber |
| 0.625 | Hua Su | A. Schur |
| 0.625 | Hua Su | Pearl Pu |
| 0.625 | Hua Su | Yuan Gao |
| 0.611111111111111 | Hua Su | Hadi Abdo |
| 0.611111111111111 | Hua Su | Alan Humm |
| 0.611111111111111 | Hua Su | Hank Hoek |
| 0.605555555555556 | Hua Su | Huw Dawkes |
| 0.6 | Hua Su | Allan Tuan |
| 0.6 | Hua Su | David Turo |
| 0.6 | Hua Su | Jianbo Shi |
| 0.6 | Hua Su | Jian Huang |
| 0.593434343434343 | Hua Su | Varun Saini |
| 0.590909090909091 | Hua Su | Jan Puzicha |
| 0.590909090909091 | Hua Su | Noah Syroid |
| 0.590909090909091 | Hua Su | Dan Shapiro |
| 0.590909090909091 | Hua Su | Henry Fuchs |
| 0.590909090909091 | Hua Su | Eduard Hovy |
| 0.590909090909091 | Hua Su | Aran Lunzer |
- Network Graph:** A graph showing nodes and edges. Nodes include L. Tweedie, Bob Spence, H. Dawkes, B. Spence, Hua Su, Hus Su, Lisa Tweedie, Huw Dawkes, and Robert Spence. Hua Su and Hus Su are highlighted in green.
- Possible Duplicates Viewer:** A table showing identified duplicates.

| AuthorID | AuthorName |
|---|------------|
| P112532 | Hua Su |
| P113040 | Hus Su |
| Jaro (Similarity: 0.888888888888889, Weight: 1) | |
- Search Results:** A search for 'hua' returned 9 nodes.

| AuthorID | AuthorName |
|----------|------------------|
| P573257 | M. C. Chuah |
| P507545 | Mei Chuah |
| P187155 | Mao Lin Huang |
| P470250 | Joshua Levasseur |
| P195636 | Mei C. Chuah |
| P112532 | Hua Su |
| P254127 | S. Huang |
| P74503 | Ed Huai-hsin Chi |
| P139655 | Jian Huang |
- Node Detail Viewer (7 items):**

| AuthorID | AuthorName |
|----------|---------------|
| P573115 | H. Dawkes |
| P572966 | B. Spence |
| P113087 | Huw Dawkes |
| P172581 | Lisa Tweedie |
| P573241 | L. Tweedie |
| P31332 | Bob Spence |
| P246545 | Robert Spence |
- Edge Detail Viewer (3 items):**

| Articleid | Title | Source | Date |
|-----------|--|--|------------------------|
| acm857591 | Visualization for functional design | Proceedings of the 1995 IEEE Symposium Information Visualization | 10/30/1995 12:00:00 AM |
| acm223464 | The influence explorer | | |
| acm238587 | Externalising abstract mathematical models | | |

Buttons at the bottom include 'Merge Duplicates' and 'Mark Distinct'. A status bar at the bottom right reads 'Finding possible duplicates completed!'.

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

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

The screenshot displays the C-Group software interface, which is used for pairwise analysis of dynamic group membership. The main window shows a network graph with nodes representing research topics and edges representing relationships between them. The nodes are labeled with research topics such as 'Multimodal UI', 'Multitouch', 'User-Centered Design', 'Lab Reports, Applications, Web', 'Augmented, Tangible UI', 'Cognitive Factors in Design', 'In-Vis', 'Ir-Vis', 'Usability', and 'Miscellaneous'. The edges are labeled with researcher names, indicating their involvement in the research topics.

The interface includes several panels and controls:

- Search Focal Pairs by Similarity Metric:** A table showing similarity scores between nodes. The top row shows a similarity of 0.580 between Benjamin B. Bederson and Allison Druin.
- Search Focal Pairs by Direct Search:** A panel for keyword search, showing results for 'bederson'.
- Focal Pair Viewer:** A table showing details for selected focal pairs, including person_id, full_name, last_name, first_name, middle_name, suffix, and affiliation.
- Node Detail Viewer:** A table showing details for selected nodes, including person_id, full_name, last_name, first_name, middle_name, and subtitle.
- Edge Detail Viewer:** A table showing details for selected edges, including article_id, title, and subtitle.

The main graph area shows a central node 'Benjamin B. Bederson' connected to various other nodes. The nodes are represented by circles containing a bar chart and text labels. The edges are represented by lines connecting the nodes.

<http://www.cs.umd.edu/projects/lings/cgroup>

HOMER: Tool for Ontology Alignment

The screenshot displays the HOMER Ontology Alignment Analysis tool interface, which is divided into three main sections:

- Top Left: Main Alignment View**
 - Algorithm: ILIADS
 - Options: Synchronized navigation
 - Navigation controls: back, forward, search, and other icons.
 - Legend: Instance (pink), Edges (grey), Prospective (orange), Changes (red); Class (yellow), Matches (blue), User defined (green), Axioms (purple).
 - Diagram: A graph showing nodes like Bacteria, Salmonella, Escherichi, and Bacterial connected by colored arcs representing alignment relationships.
 - Search bar: search >> [input] All results
- Top Right: Alignment Information Panel**
 - Tabbed interface: Matches, Execution trace, Axioms, Console.
 - Content: A list of alignment results with various scores and logical inferences. For example, one match shows: (Salmonella enterica typhirium, owl:sameAs, SalmonellaEnterica) with a final score of 0.414.
 - Buttons: Focus, Differences, Delete match.
- Bottom: Comparative View**
 - Algorithm: ILIADS
 - Options: Synchronized navigation
 - Navigation controls: same as the top left.
 - Legend: same as the top left.
 - Diagram: A graph showing nodes like FoodPoisoning, EColiPoiso..., Botulism, TheodorEsc..., Cholera, Acute Gast..., Salmonella, E-col, and Food Borne..
 - Search bar: search >> [input] All results
 - Text below diagram: <(E-coli, owl:sameAs, EColiPoisoning), 0.698>

<http://www.cs.umd.edu/projects/lings/iliads>

SplicePort: Motif Explorer

Return to SplicePort Homepage
Browse Donor Features
Predict Splice Sites

select:

start: end:

Found 945 features in this interval

| | |
|------------------|----------|
| --ctctgcAG----- | 0.397023 |
| ---tccccAGg----- | 0.388358 |
| ---cccacAGg----- | 0.373086 |
| --tgtttcAG----- | 0.367203 |
| --tcttgcAG----- | 0.365459 |
| ---cctgcAGg----- | 0.364709 |
| -ttttt--AGg----- | 0.359494 |
| --ccttgcAG----- | 0.3534 |
| t--ctttcAG----- | 0.339176 |
| --tccccAG----- | 0.330497 |
| --ccctgcAG----- | 0.325713 |

Motif:

Weight:

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

<http://www.cs.umd.edu/projects/spliceport>

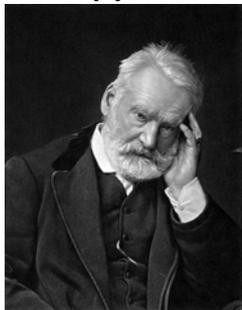
● ● ● 3. GI & Privacy

- 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:

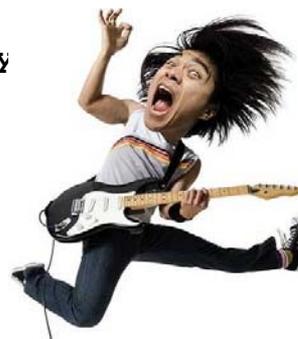
"how to tell if your wife is cheating on y

same-user

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

● ● ● Summary: GIA & AI

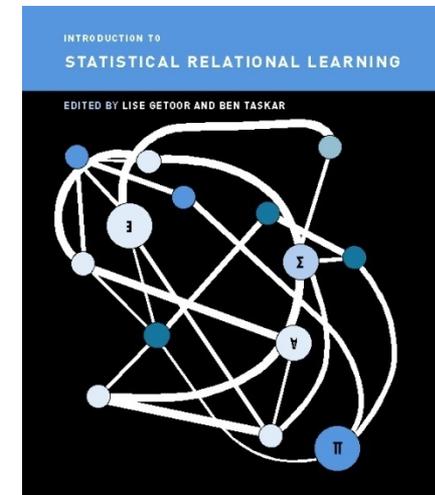
- Graph Identification can be seen as a process of **knowledge reformulation**
- In the context where we have some statistical information to help us **learn** which reformulations are more promising than others
- **Inference** is the process of transferring the learned knowledge to new situations

● ● ● 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

