



Graph Identification

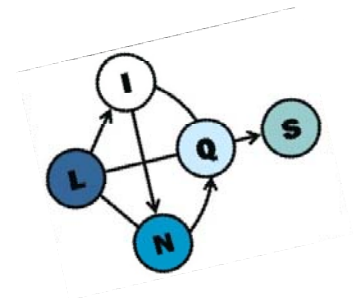
Lise Getoor

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

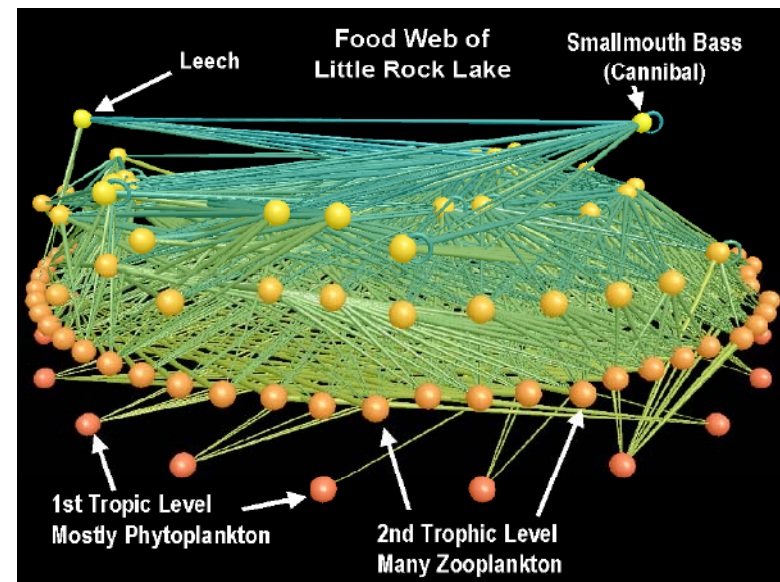
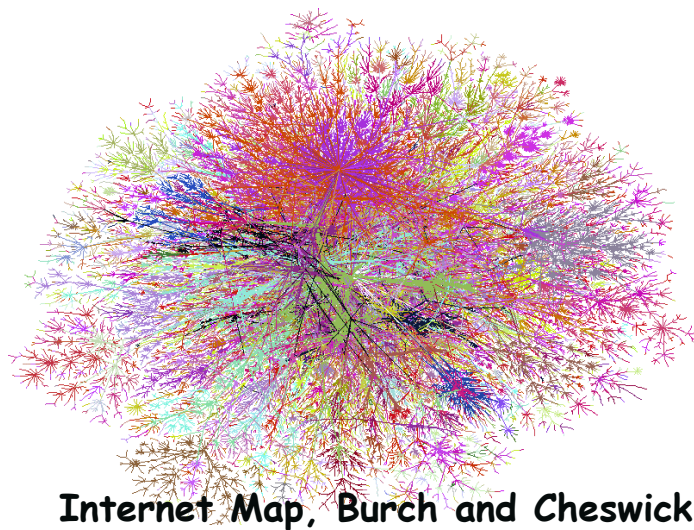
○ Funding Sources:



KDD Program

● ● ● Graphs and Networks *everywhere...*

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



Food Web, Martinez

Others available at Mark Newman's gallery:

<http://www-personal.umich.edu/~mejn/networks/>

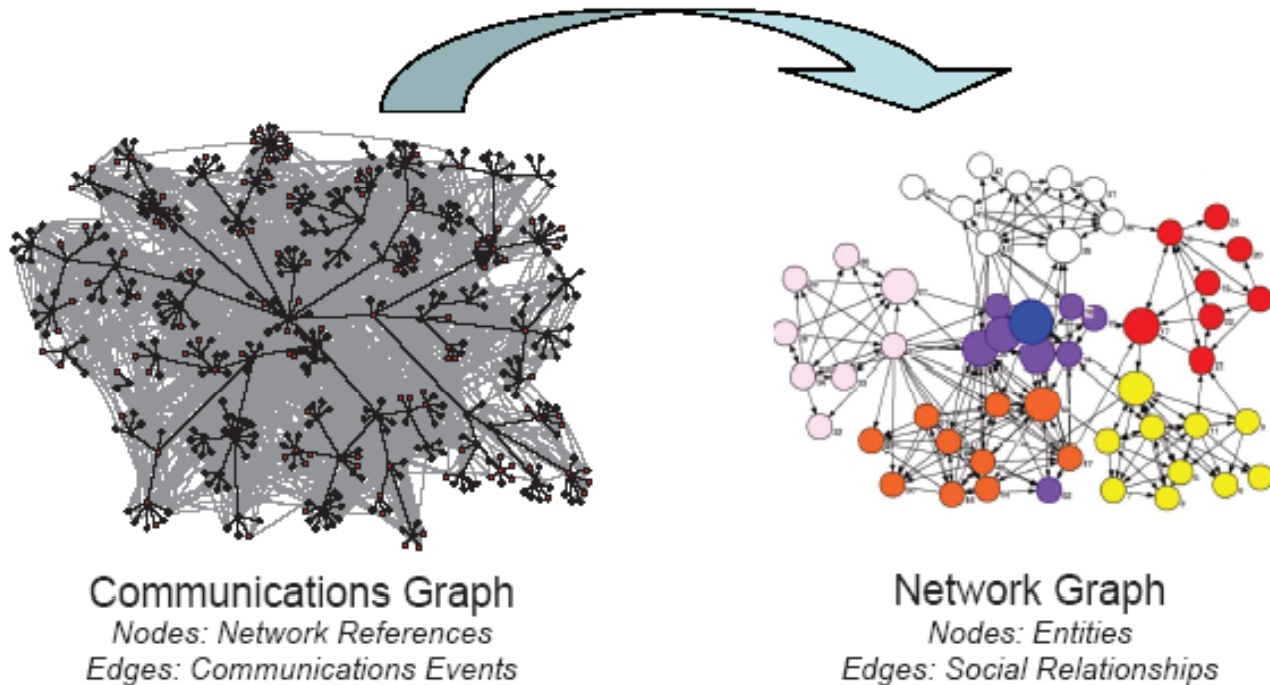
● ● ● Wealth of Data

- 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



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

● ● ● 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

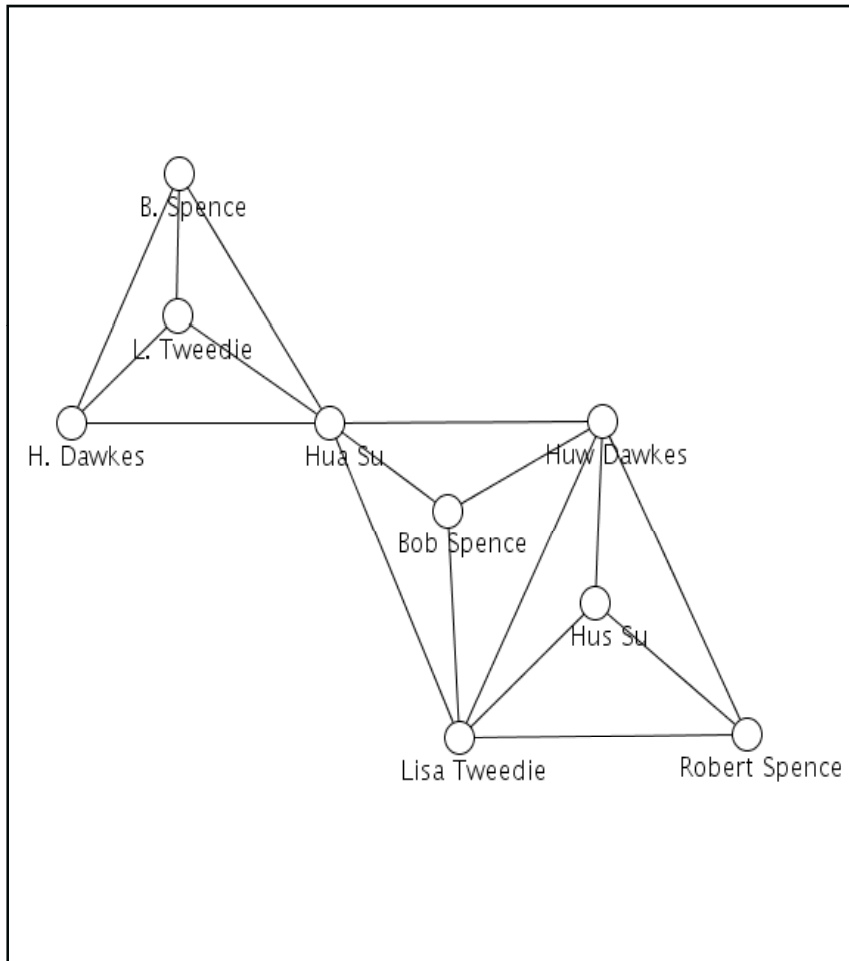
● ● ● Entity Resolution

- **The Problem**

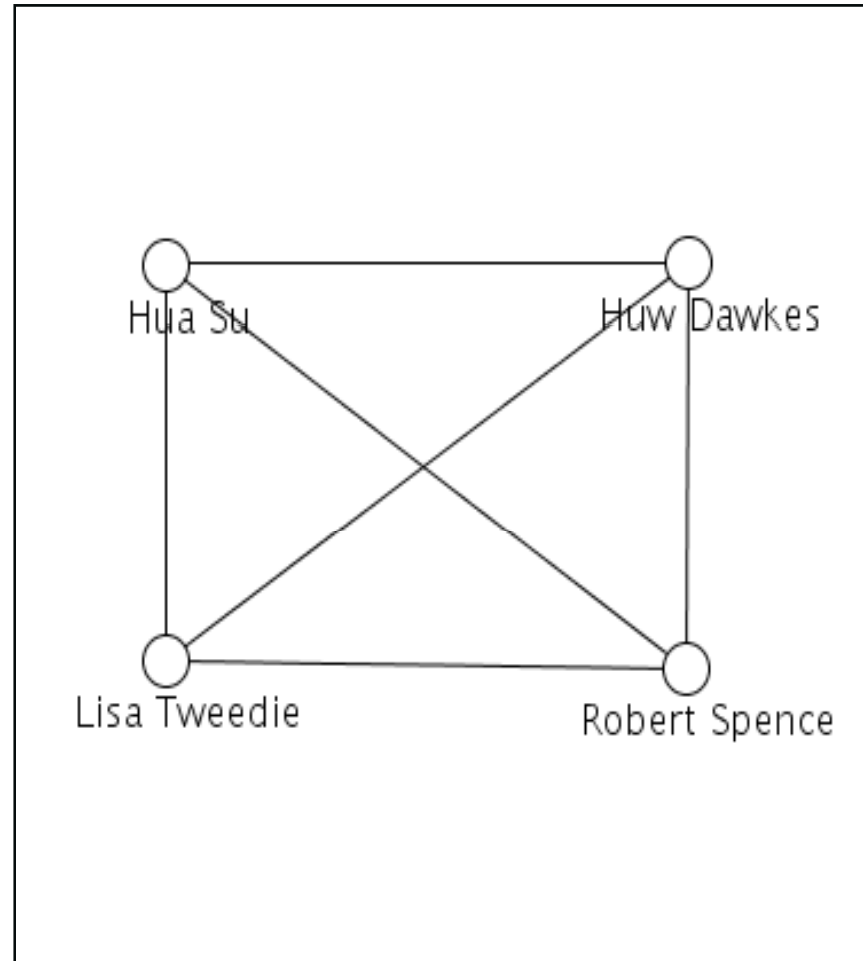
- Relational Entity Resolution

- Algorithms

● ● ● InfoVis Co-Author Network Fragment

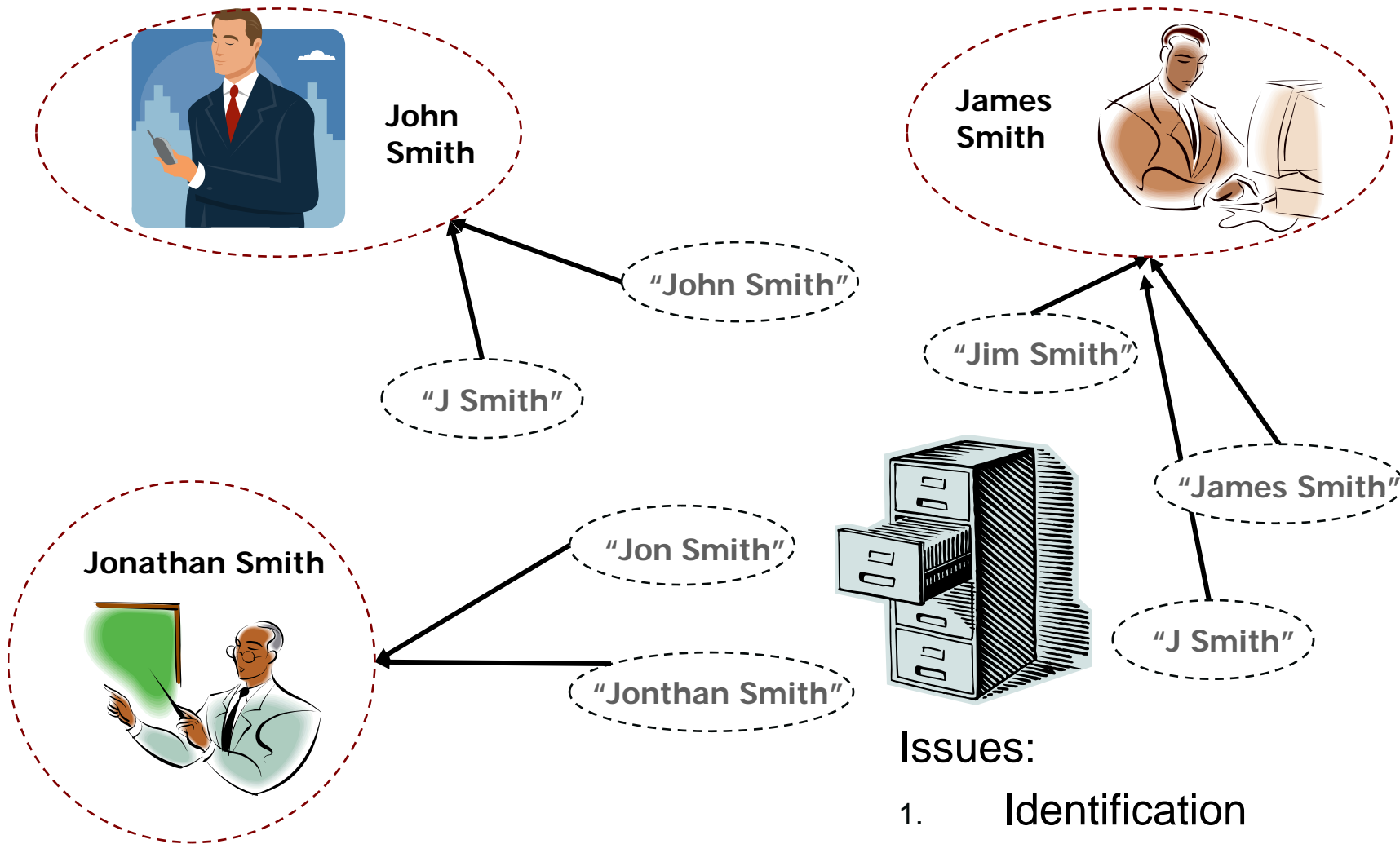


before



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**

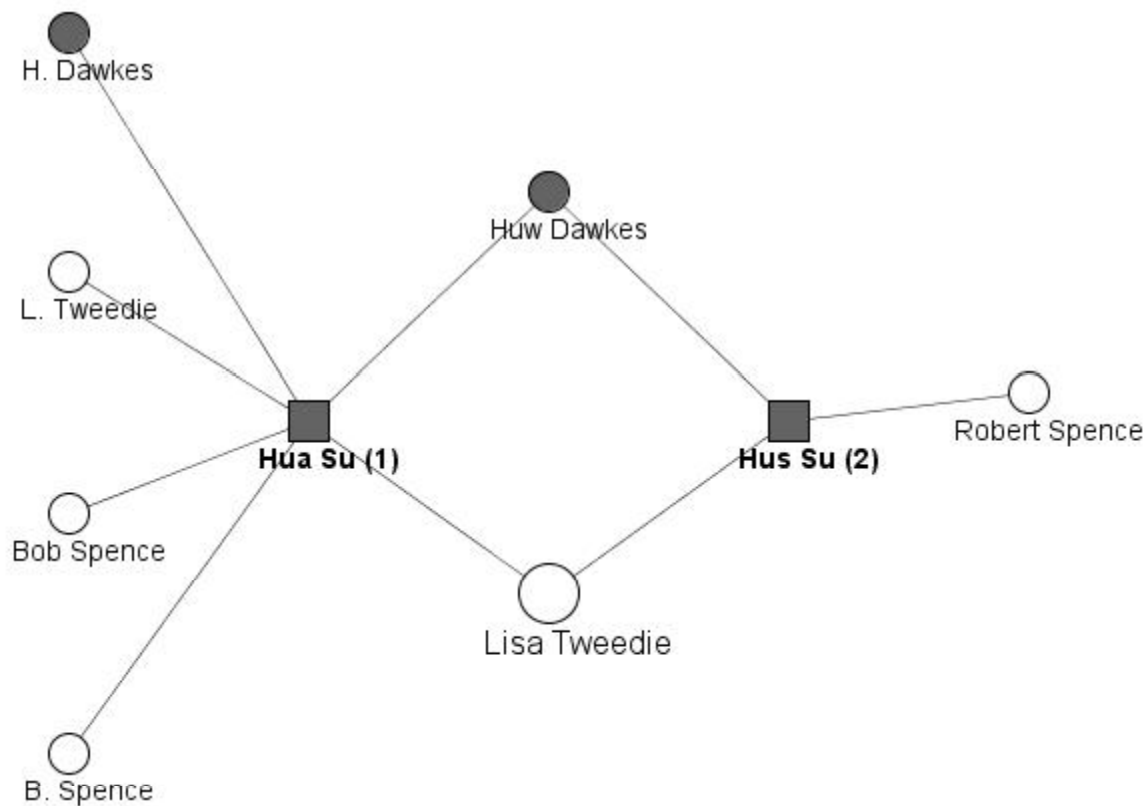
- Algorithms

● ● ● 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

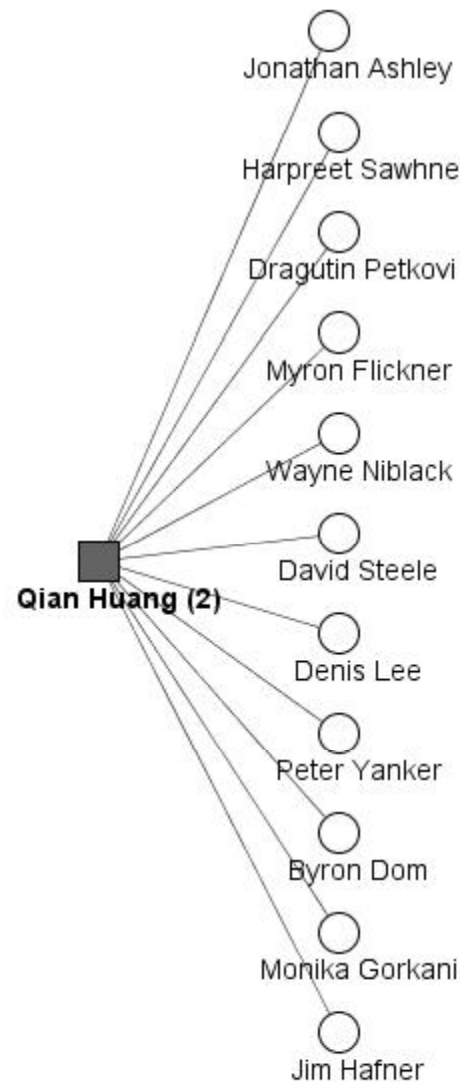
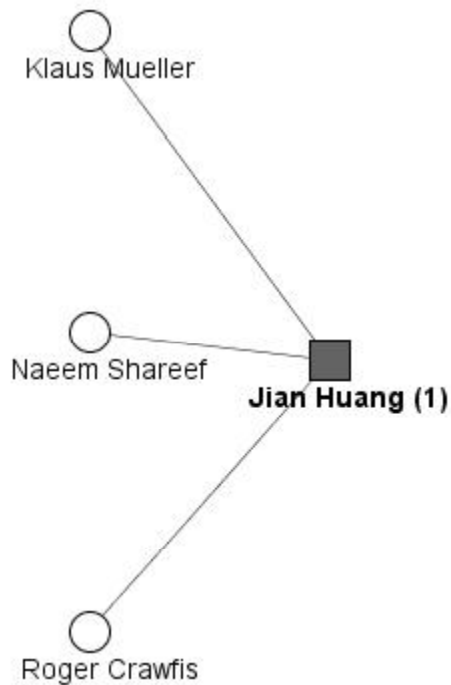
Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05

● ● ● Relational Identification



Very similar names.
Added evidence from
shared co-authors

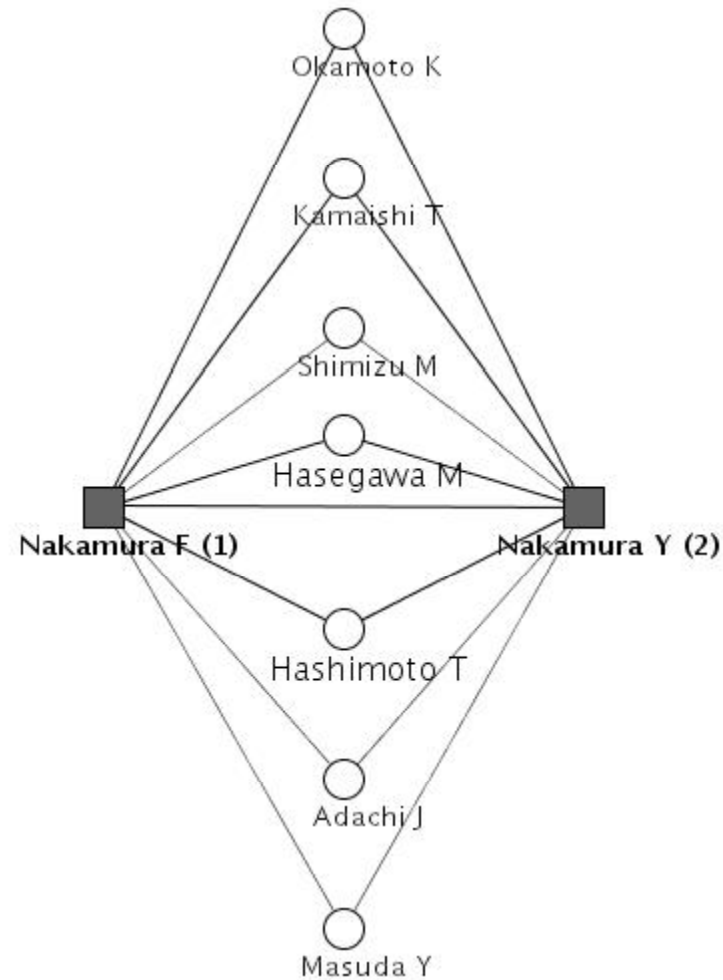
● ● ● Relational Disambiguation



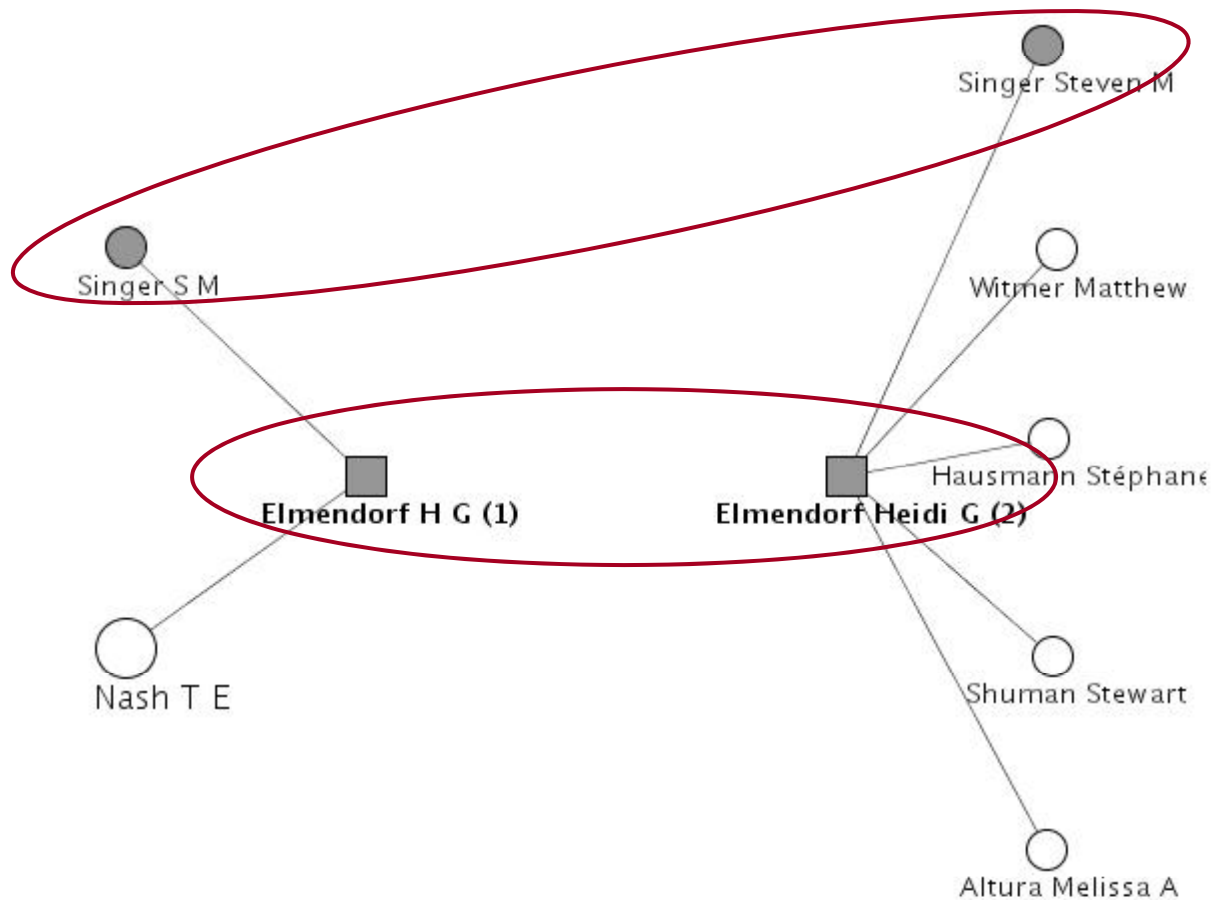
Very similar names
but no shared
collaborators

● ● ● 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

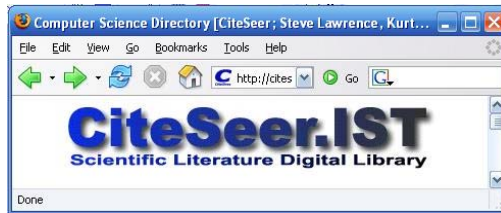
- The Problem

- 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

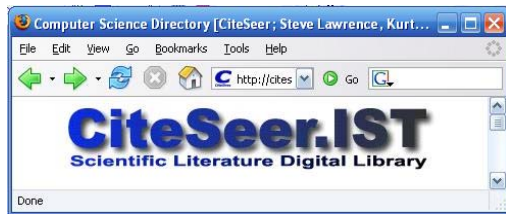
P2: “*Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies*”, C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus

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

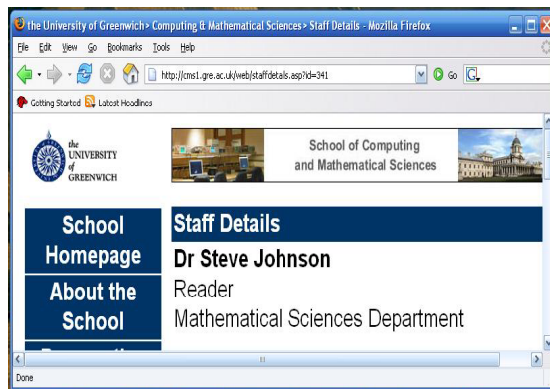
P4: “*Code Generation for Machines with Multiregister Operations*”, Alfred V. Aho, Stephen C. Johnson, Jefferey D. Ullman

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 Unified Optimisation and Load-Balancing Algorithm*”, C. Walshaw, M. Cross, M. G. Everett

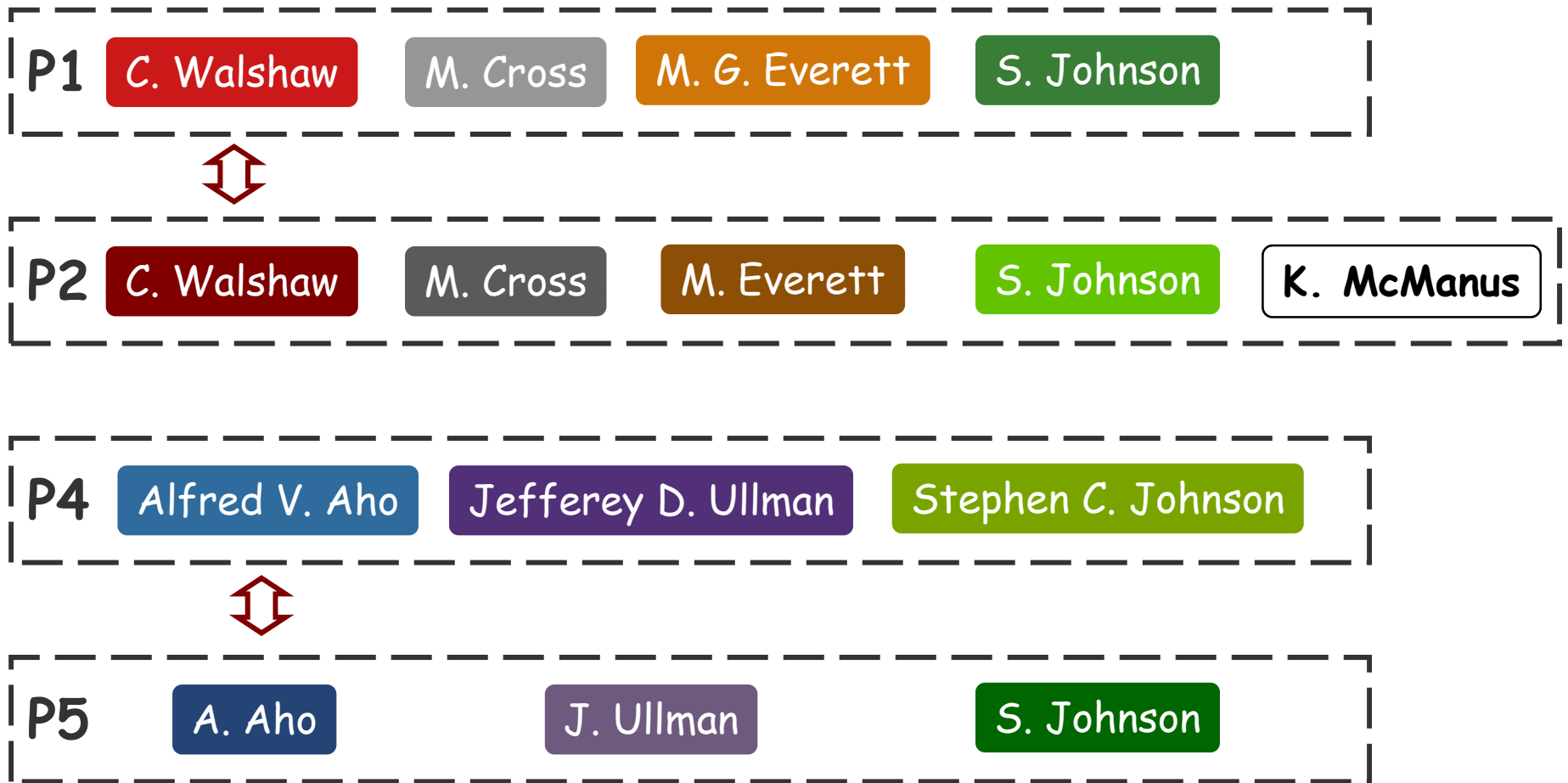
P4: “*Code Generation for Machines with Multiregister Operations*”, Alfred V. Aho, **Stephen C. Johnson**, Jefferey D. Ullman



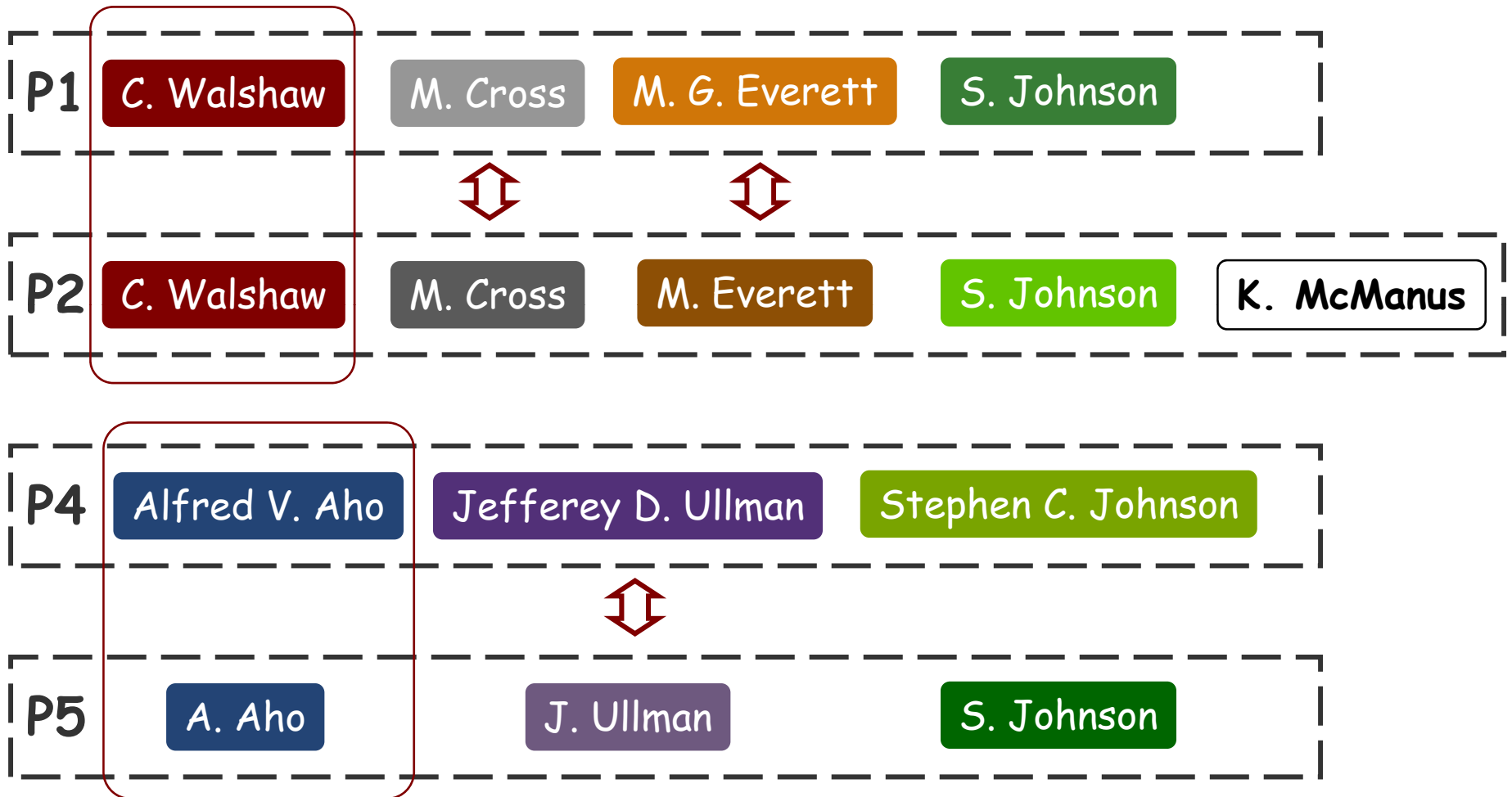
P5: “*Deterministic Parsing of Ambiguous Grammars*”, A. Aho, **S. Johnson**, J. Ullman

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

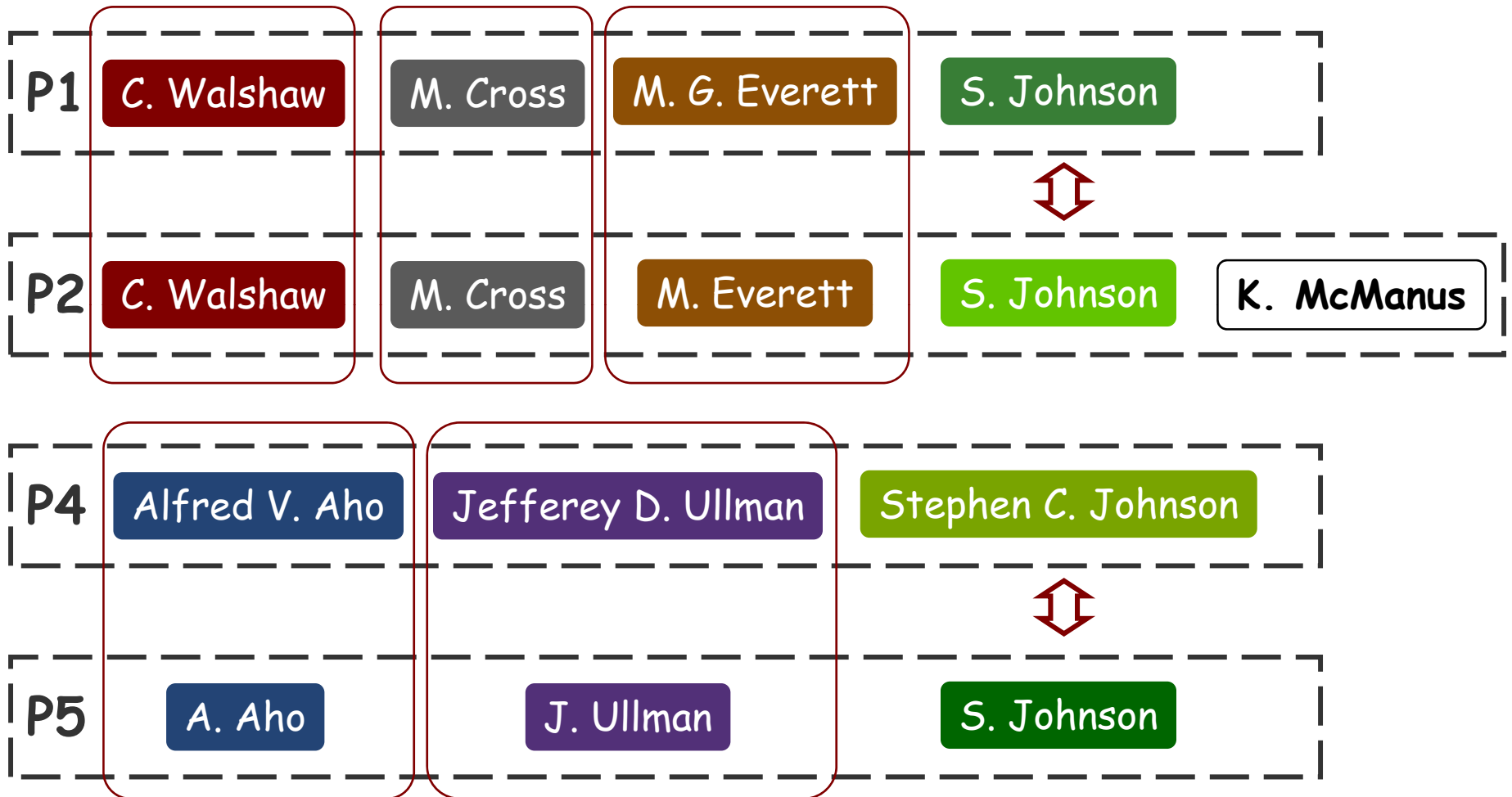
● ● ● Relational Clustering (RC-ER)



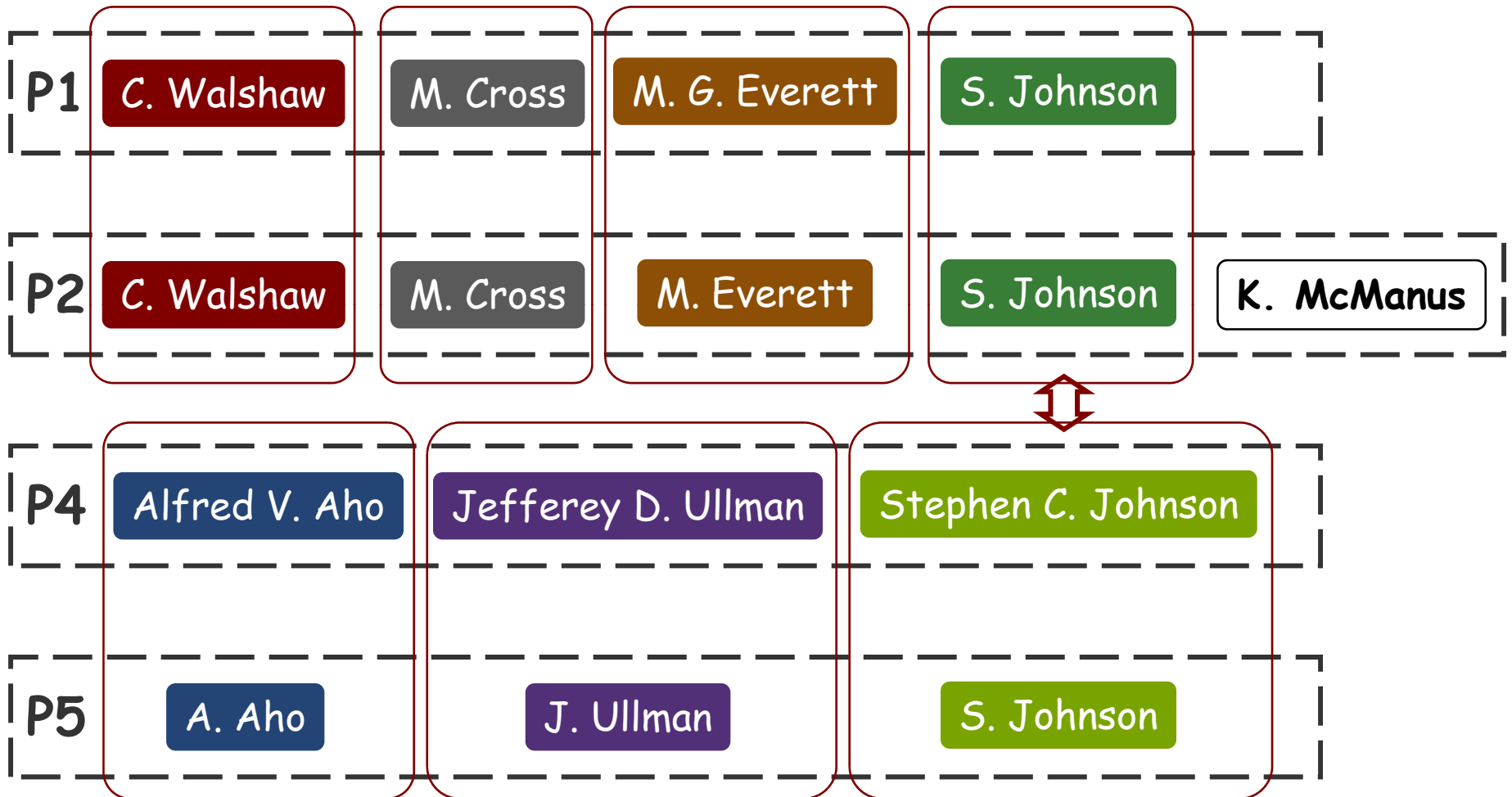
● ● ● Relational Clustering (RC-ER)



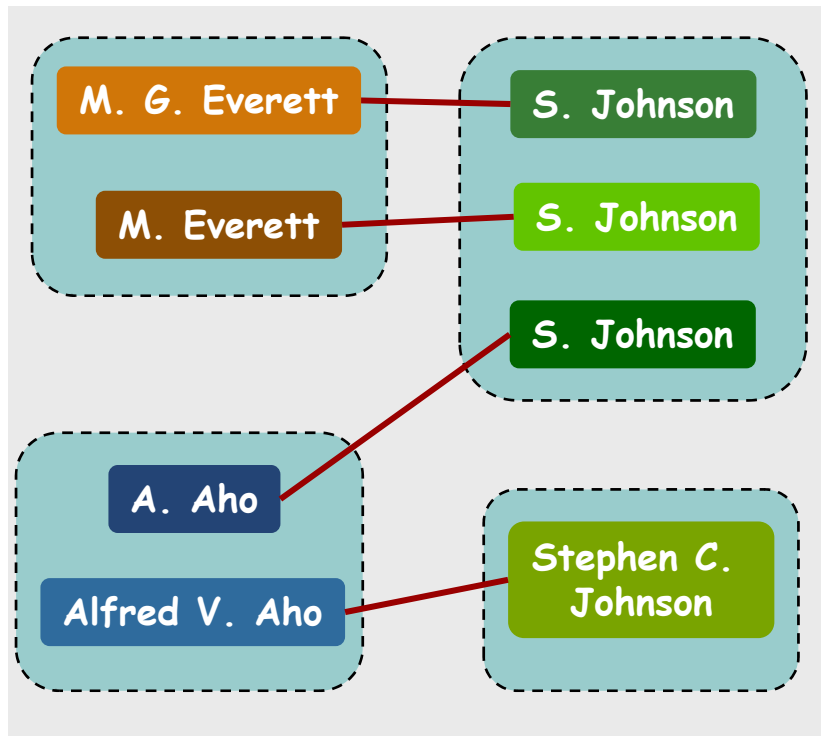
● ● ● Relational Clustering (RC-ER)



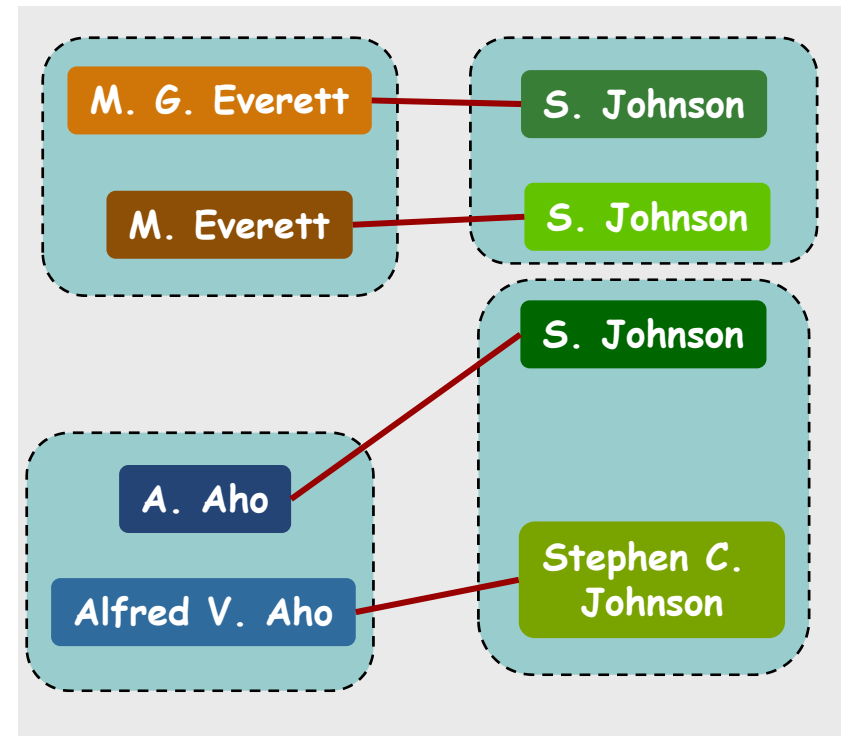
● ● ● Relational Clustering (RC-ER)



● ● ● 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

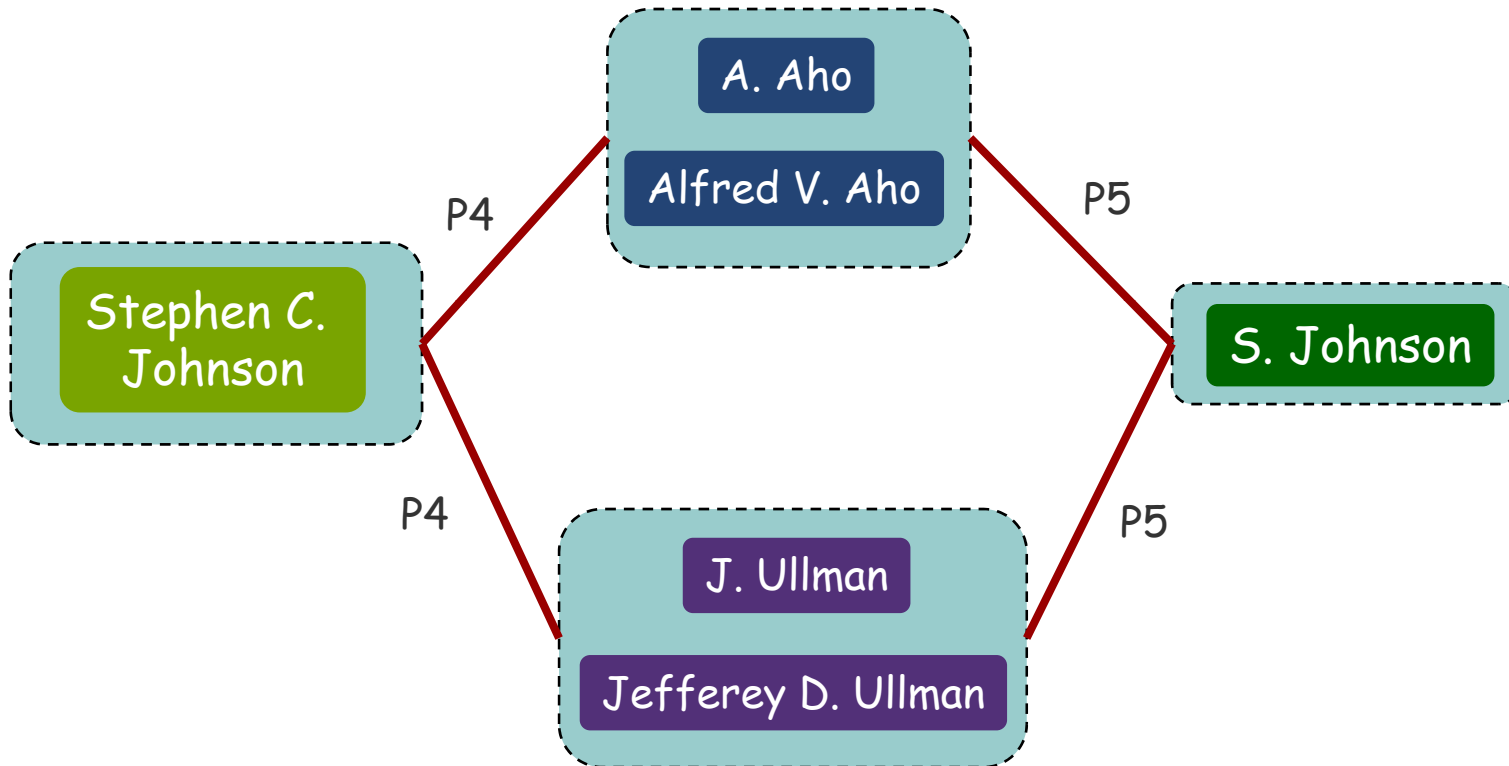
Similarity based on relational
edges between c_i and c_j

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

● ● ● Measures for Attribute Similarity

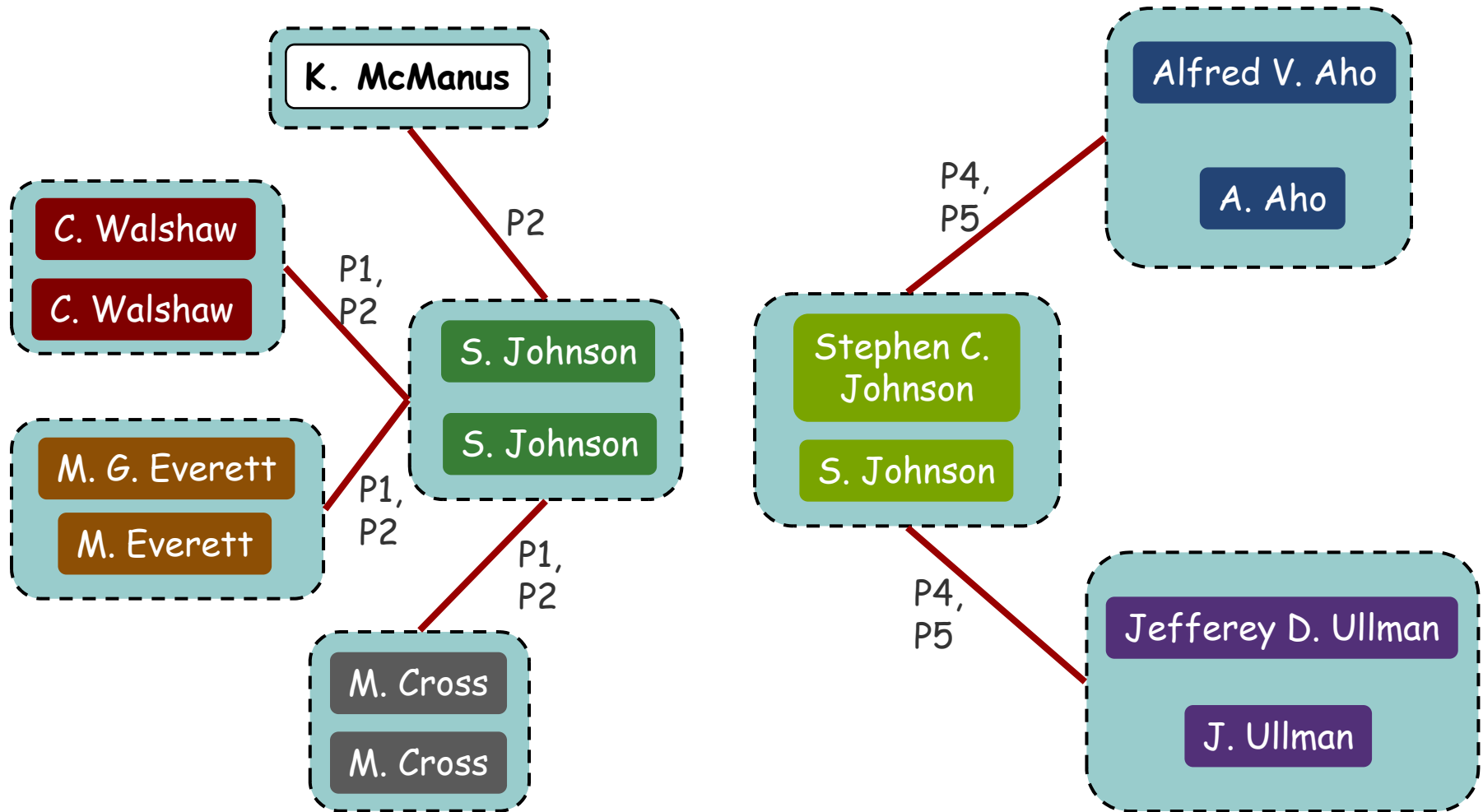
- Use best available measure for each attribute
 - Name Strings: *Soft TF-IDF, Levenstein, Jaro*
 - Textual Attributes: *TF-IDF*
- Aggregate to find similarity between clusters
 - Single link, Average link, Complete link
 - Cluster representative

● ● ● 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



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

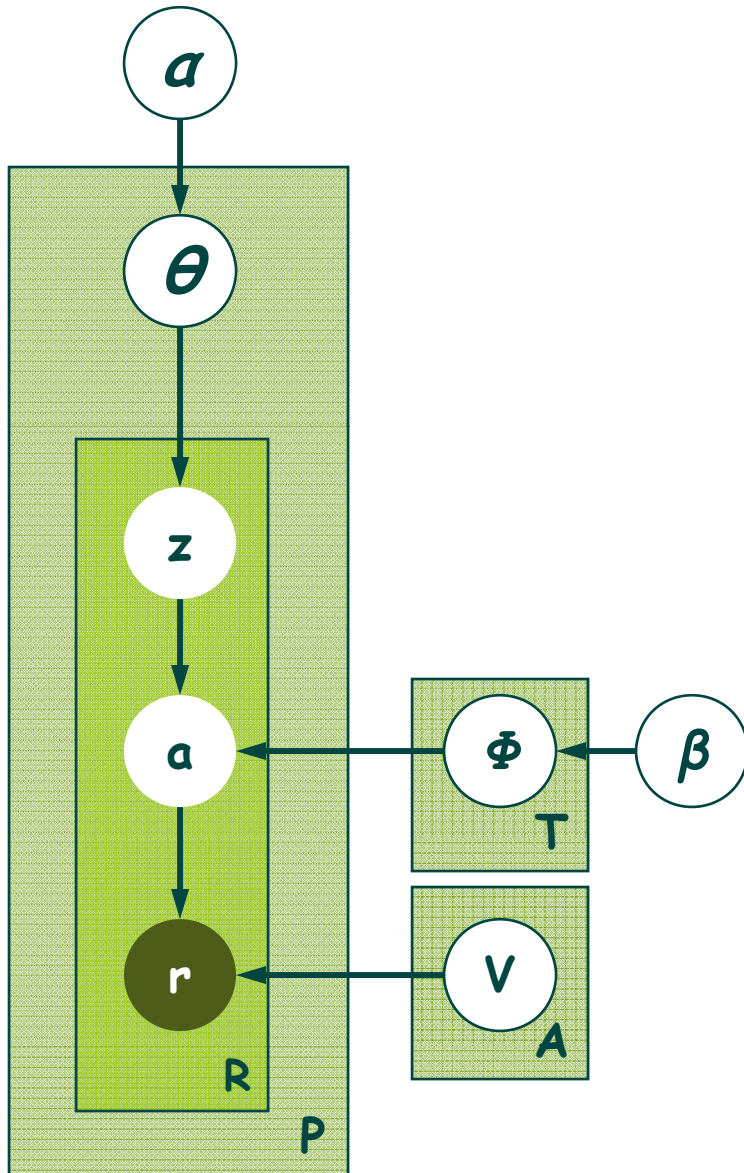


P4: Alfred V. Aho, **Stephen C. Johnson**,
Jefferey D. Ullman

P5: A. Aho, **S. Johnson**, J. Ullman

P6: A. Aho, R. Sethi, J. Ullman

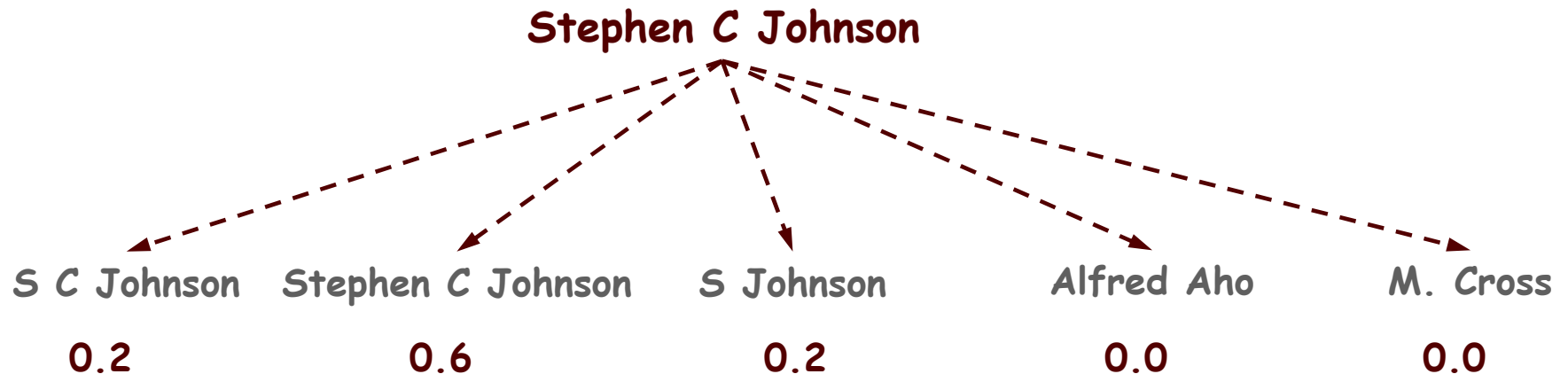
Latent Dirichlet Allocation ER



- 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

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

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

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

- 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

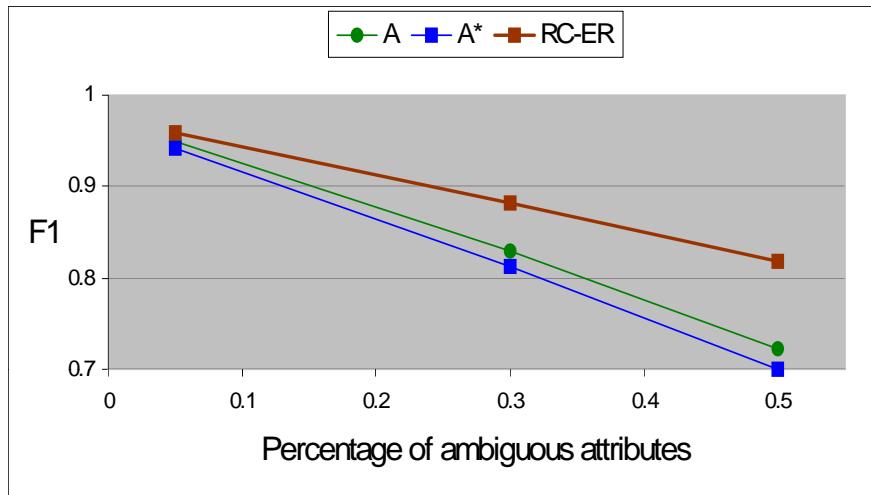
- **CiteSeer**: Near perfect resolution; 22% error reduction
- **arXiv**: 6,500 additional correct resolutions; 20% error reduction
- **BioBase**: Biggest improvement over baselines

● ● ● 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

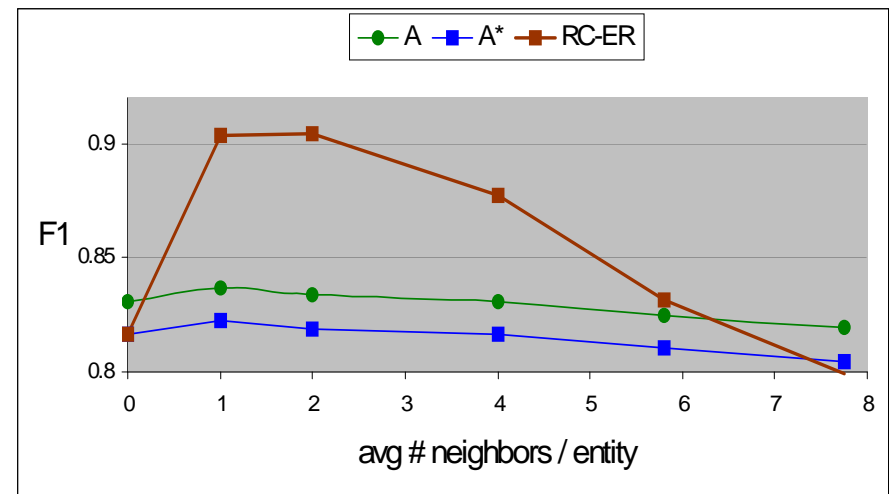
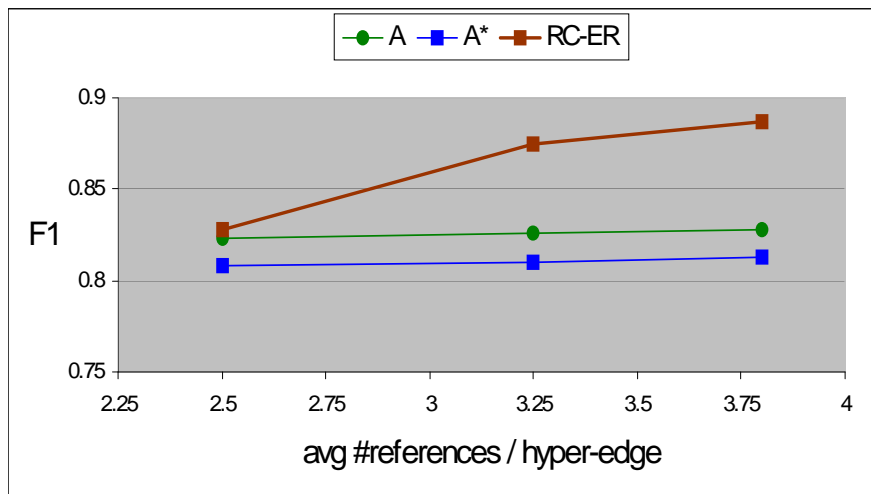
arXiv
Significantly larger improvements for 'ambiguous names'

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

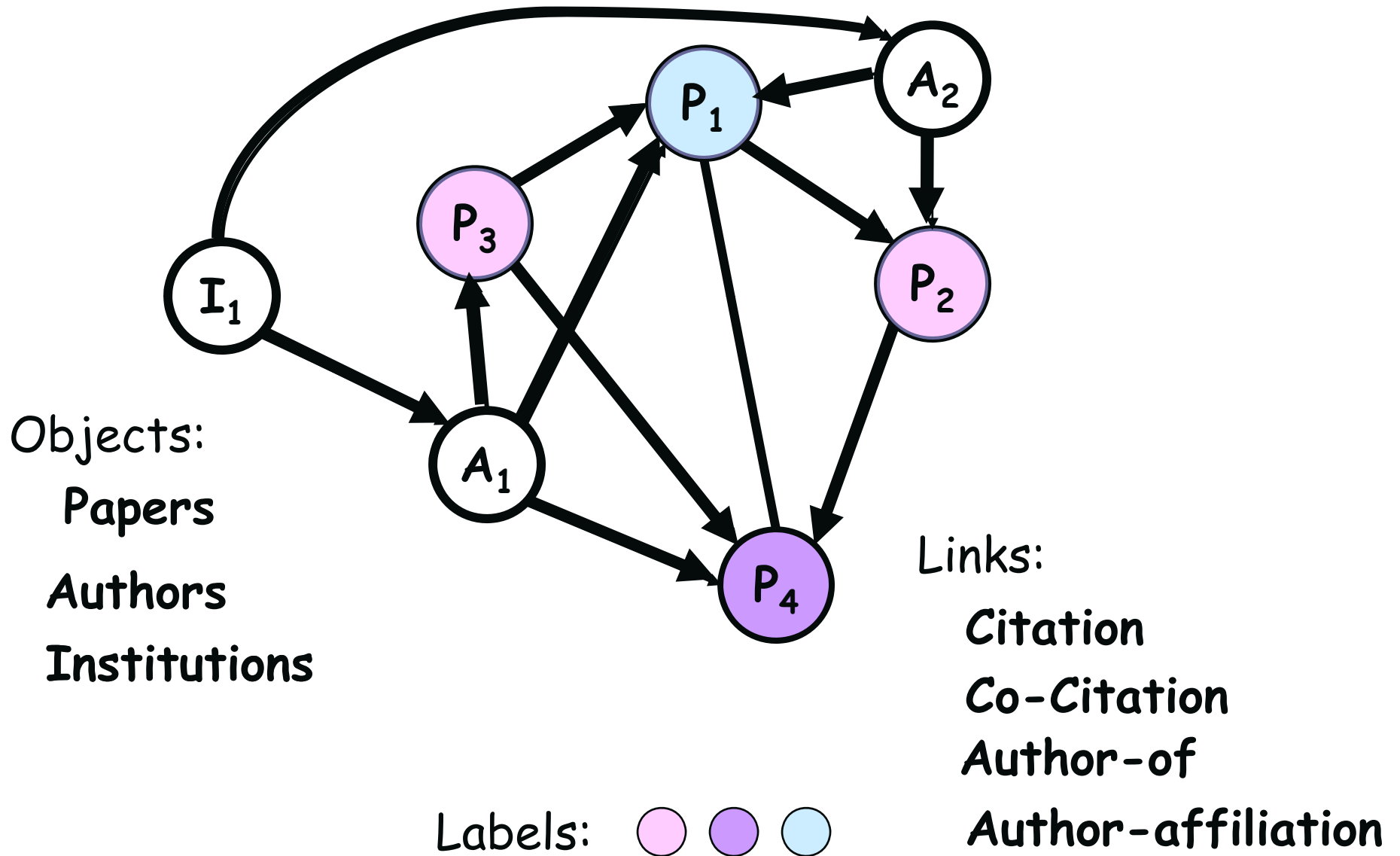
- Algorithms

● ● ● 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



● ● ● 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

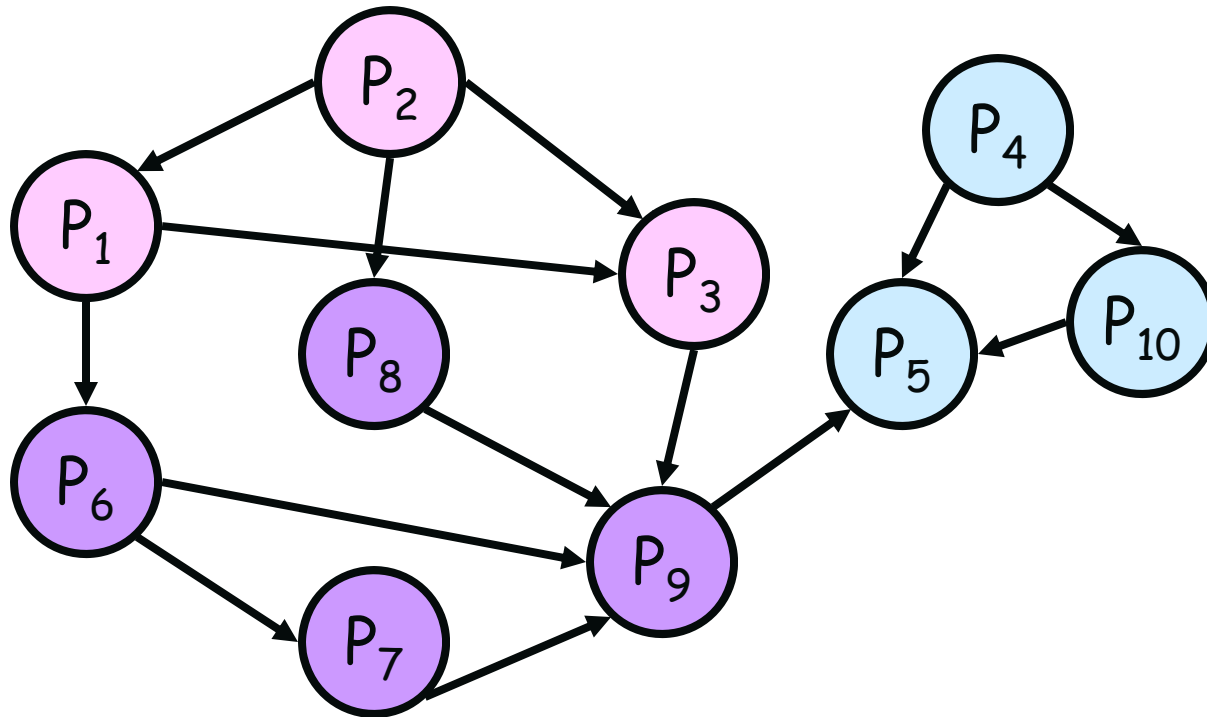
- Collection of Local Conditional Models
- Inference Algorithms:
 - Iterative Classification Algorithm (ICA)
 - Gibbs Sampling (Gibbs)

○ Undirected Models

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

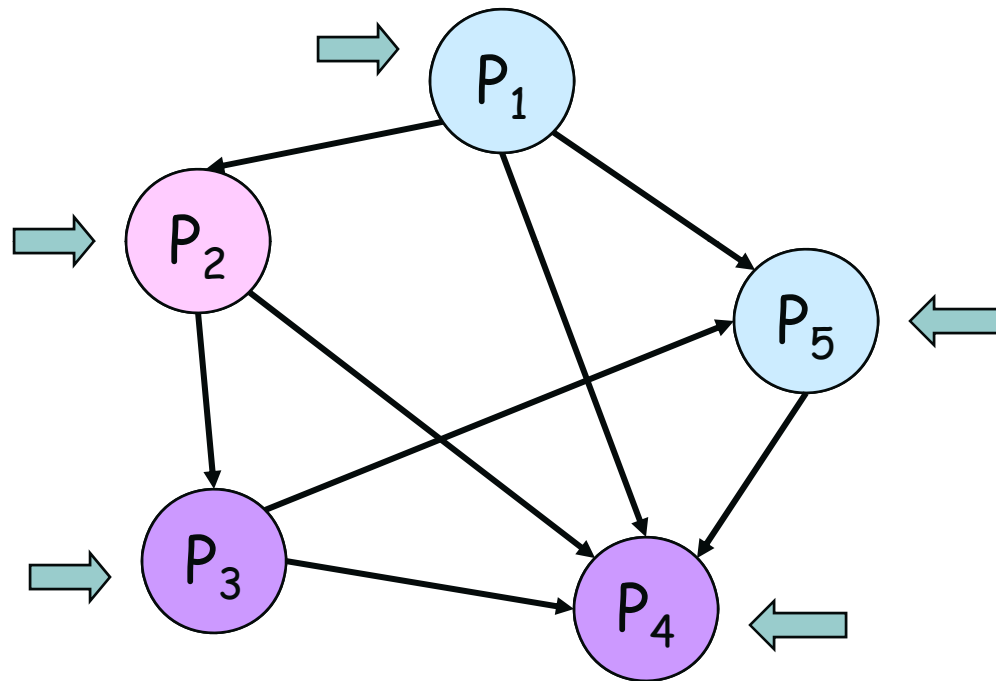
ICA: Learning

o label set:   



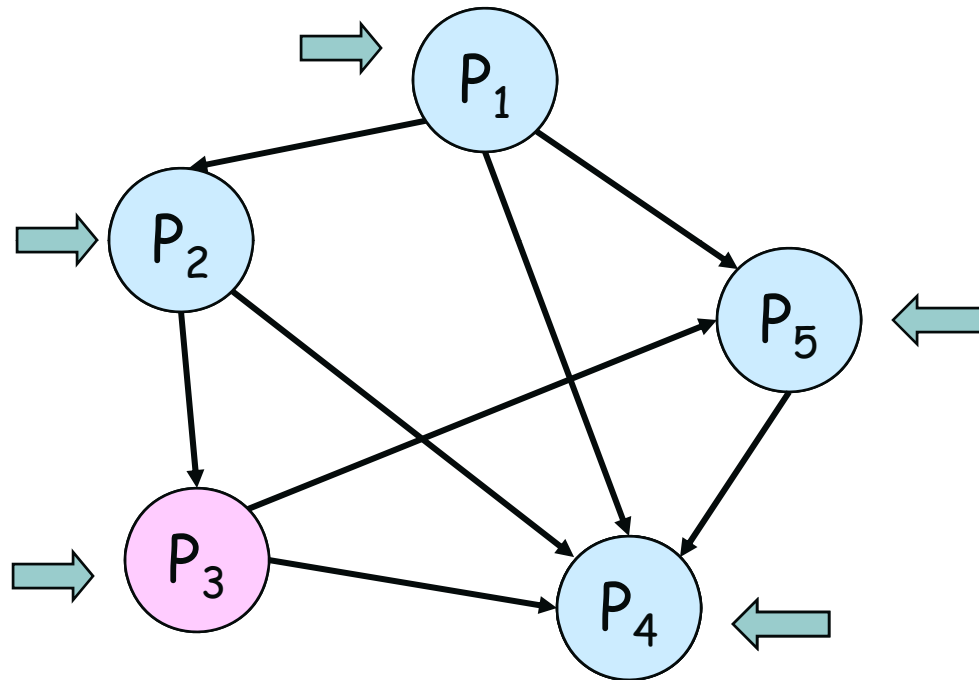
Learn model from fully labeled training set

- ● ● 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

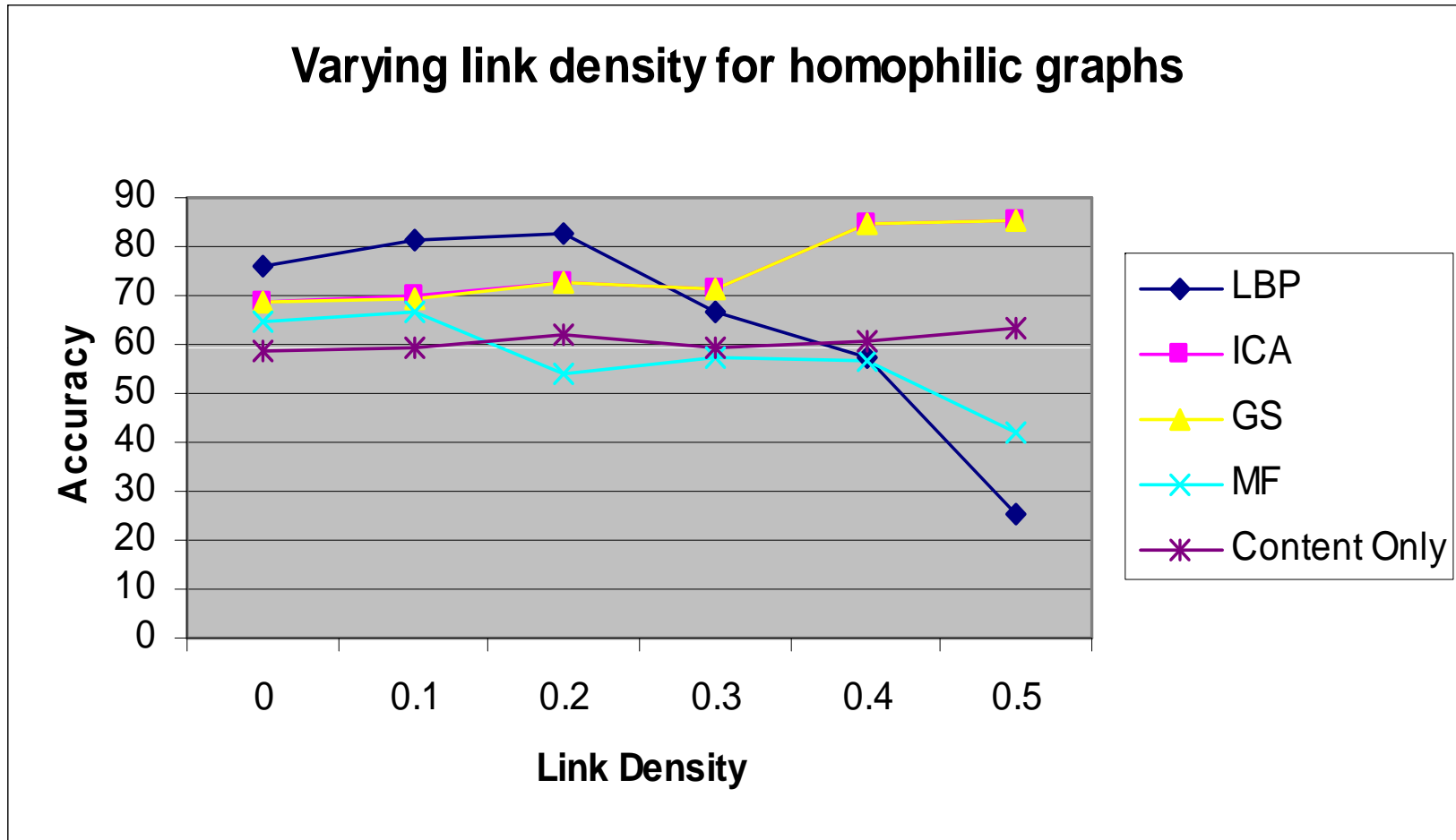
- Comparison of Collective Classification Algorithms
 - Mean Field Relaxation Labeling (MF)
 - Iterative Classification Algorithm (ICA)
 - Loopy Belief Propagation (LBP)
 - Baseline: Content Only
- Datasets
 - Real Data
 - Bibliographic Data (Cora & Citeseer), WebKB, etc.
 - Synthetic Data
 - Data generator 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**

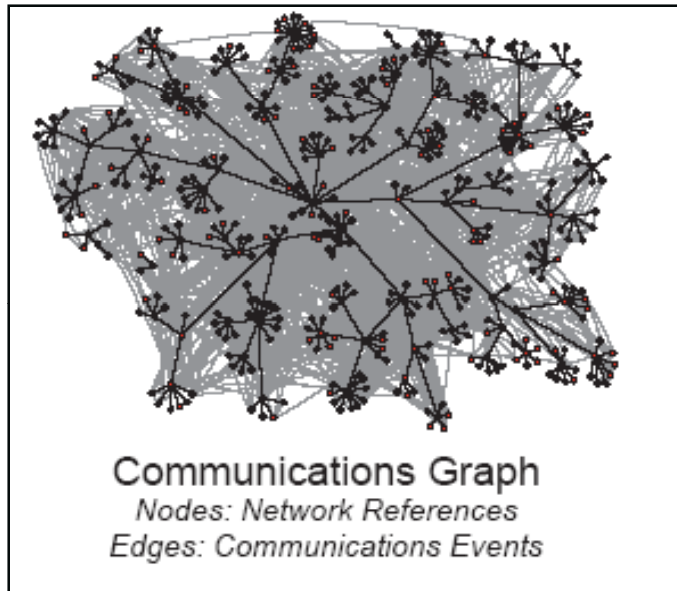
- Putting It All Together

- Open Questions

● ● ● Link Prediction

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

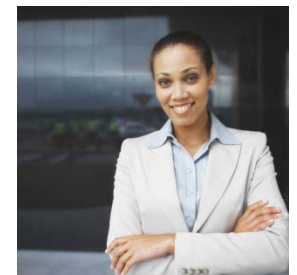
● ● ● Links in Data Graph



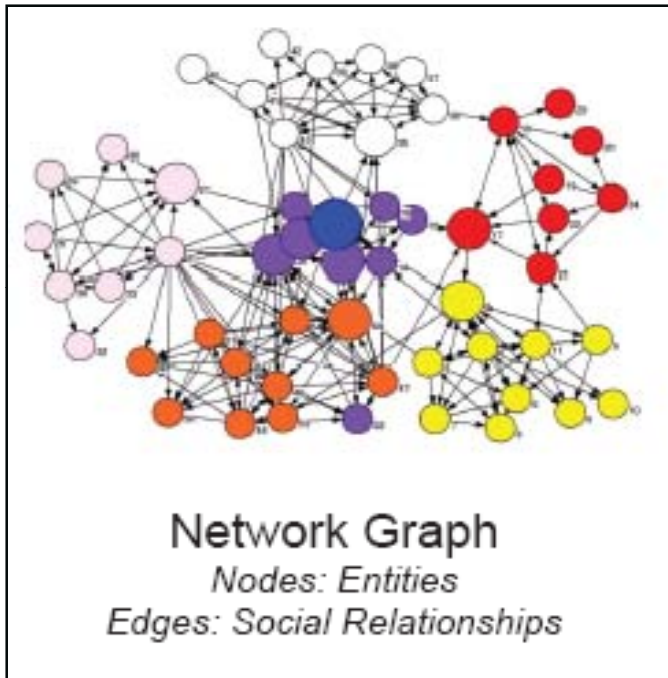
chris@enron.com ← Email → liz@enron.com

chris37 ← IM → lizs22

555-450-0981 ← TXT → 555-901-8812



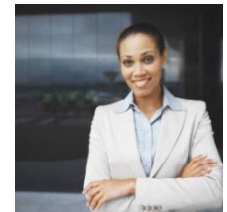
● ● ● ⇒ Links in Information Graph



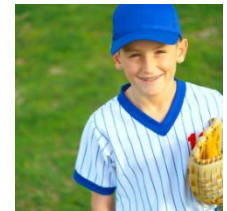
Chris



Steve



Elizabeth



Tim



● ● ● Predicting Relations

○ Link Labeling

- Can use similar approaches to collective classification

○ Link Ranking

- Many variations

- Diehl, Namata, Getoor, *Relationship Identification for Social Network Discovery*, AAAI07

- 'Leak detection'

- Carvalho & Cohen, SDM07

○ Link Existence

- HARD!

- Huge class skew problem

- Variations: Link completion, find missing link

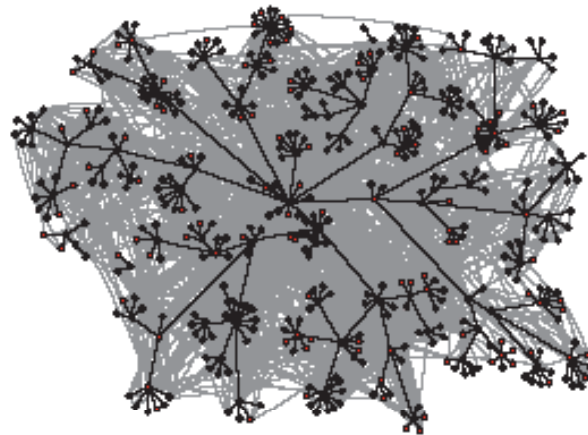
● ● ● Roadmap

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

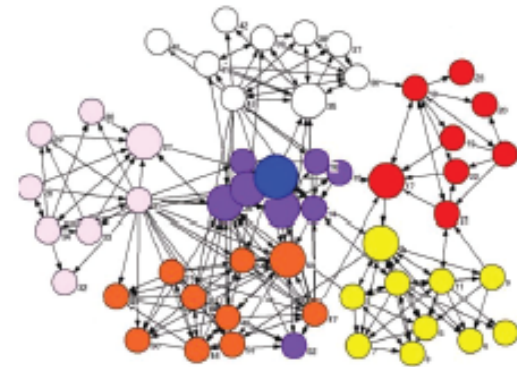
● ● ● Putting Everything together....



**Collaborative Social
Network Discovery**
Entity Resolution
Relationship Identification



Communications Graph
Nodes: Network References
Edges: Communications Events



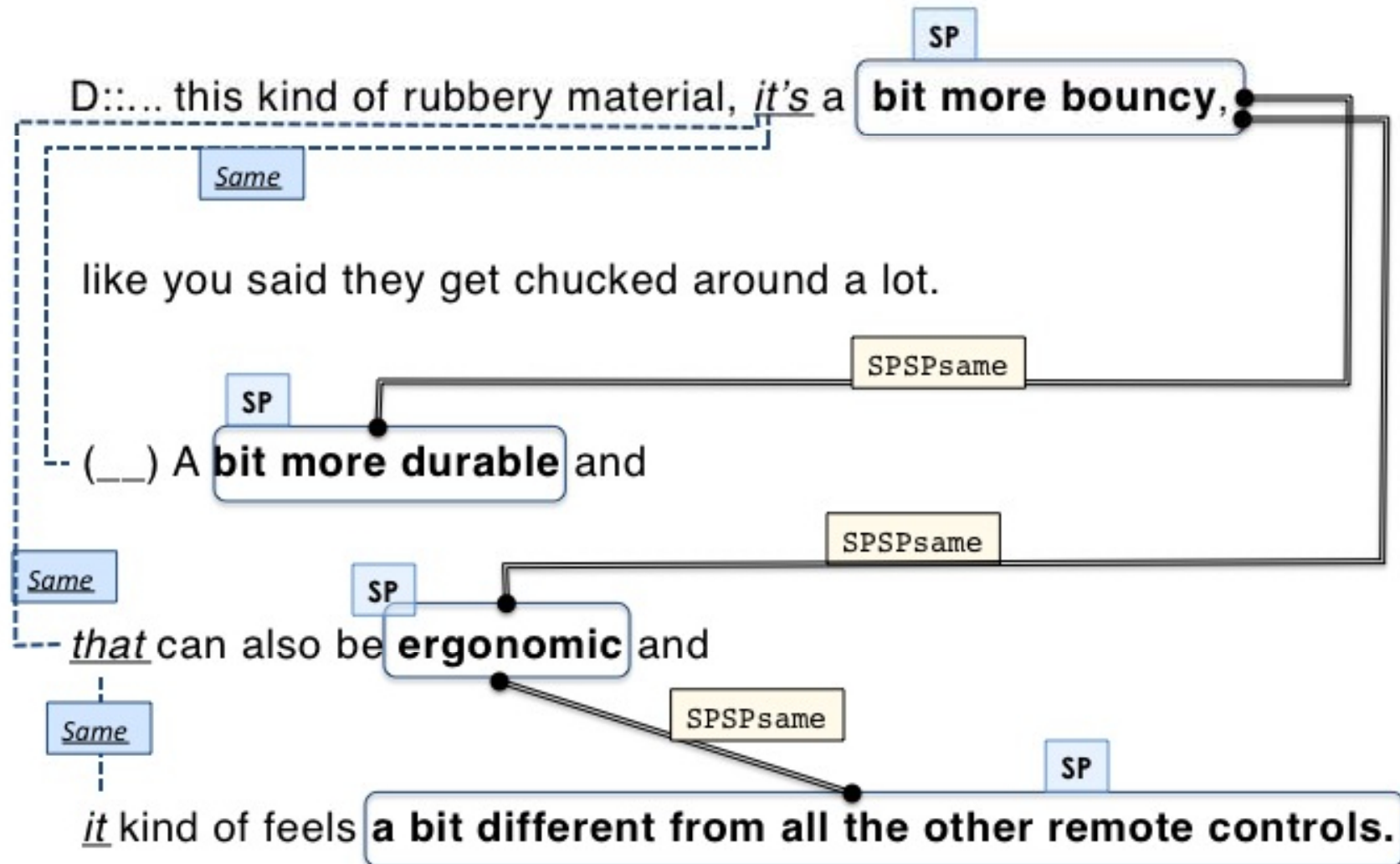
Network Graph
Nodes: Entities
Edges: Social Relationships

● ● ● Learning and Inference Hard

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

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

- Similarity Panel:** 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 central node 'Hua Su' (green square) is connected to several peripheral nodes (white circles): L. Tweedie, Bob Spence, H. Dawkes, B. Spence, Lisa Tweedie, and Huw Dawkes. 'Huw Dawkes' is further connected to 'Hus Su' (green square), which is connected to 'Robert Spence'.
- Possible Duplicates Viewer:**

AuthorID	AuthorName
P112532	Hua Su
P113040	Hus Su
Jaro (Similarity: 0.888888888888889, Weight: 1)	
- Search (9 nodes found):**

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		

Finding possible duplicates completed!

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

GeoDDupe: Tool for Interactive Entity Resolution in Geospatial Data

The screenshot displays the GeoDDupe 2.0 application window, which is used for interactive entity resolution in geospatial data. The interface is divided into several panels:

- Search Duplicate Nodes:** A panel on the left containing a search bar and a list of possible duplicates. The list includes columns for Similarity, Left Node, and Right Node. For example, it shows '0.99907411 Qaryat al Hasa Qaryat al Hasa'.
- Network Graph:** A central panel showing a network of nodes representing place names. Nodes are connected by lines, indicating relationships or similarities. Key nodes include 'Bustan Zi'al', 'Qaryat Abu Hawa', 'Qaryat al 'Ankur', 'Qaryat al Tharthar', 'Qaryat al Hasa', 'Qaryat al Khadr', 'Qaryat Al Yunis', 'Qaryat al Haqqi', 'Qaryat an Na'imiyah', and 'Qaryat an Nu'aymiyah'.
- Possible Duplicates Viewer:** A table below the graph showing details for selected duplicates. It includes columns for ID, Name, Type, and Latitude.

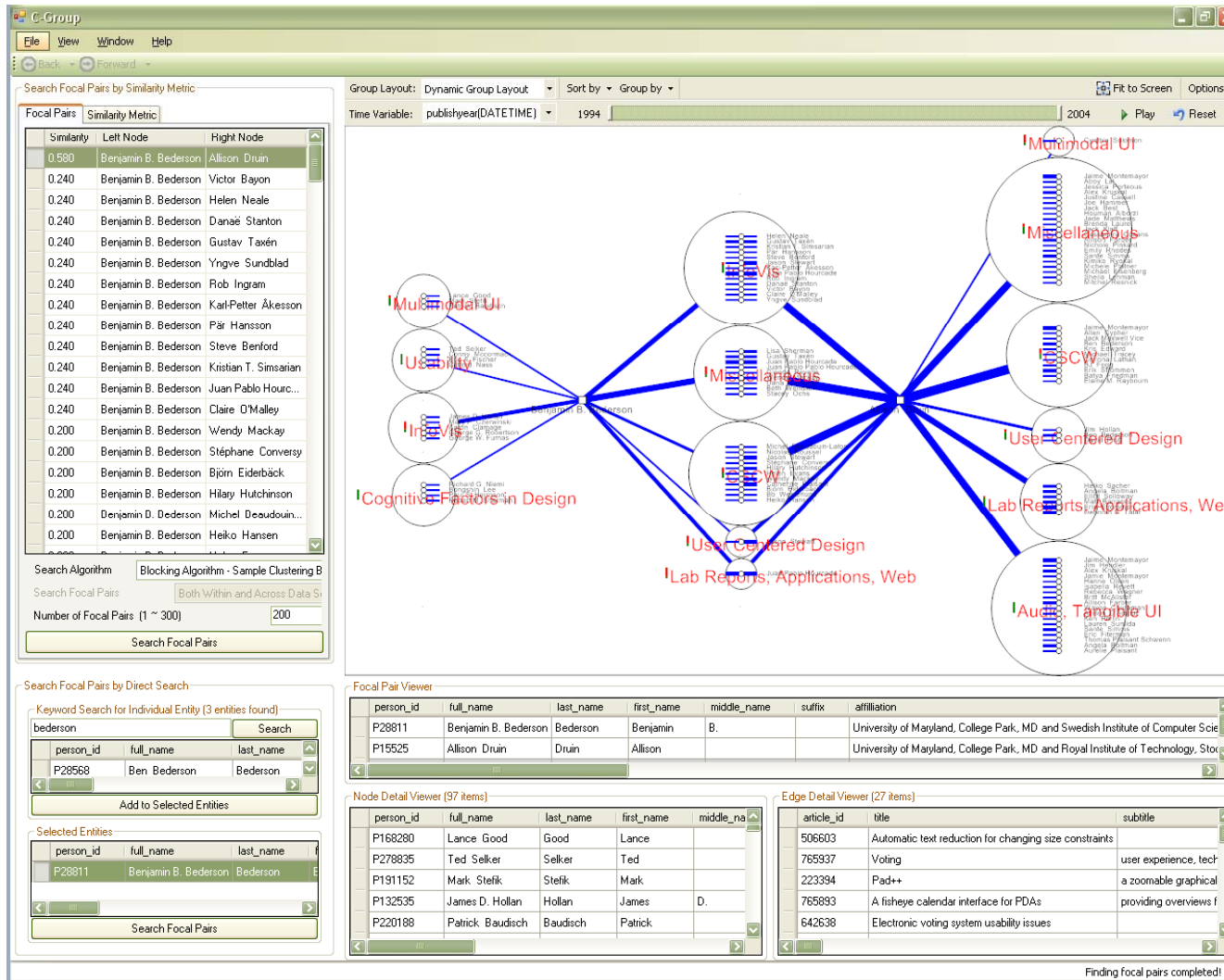
ID(STRING)	NAME(STRING)	TYPE(STRING)	LATITUDE(NUMBER)
T20583	Qaryat al Hasa	P	33.2325
D683	Qaryat al Hasa	Pop. Place	33.23261111111111
Jaro (Similarity: 1, Weight: 0.3333333333333333)			ABS (Similarity: 0.9877777777777778, Weight: 0.3333333333333333)
- Node Detail Viewer (16 rows):** A table at the bottom showing detailed information for 16 nodes, including ID, Name, Type, Latitude, Longitude, and Dataset.

ID(STRING)	NAME(STRING)	TYPE(STRING)	LATITUDE(NUMBER)	LONGITUDE(NUMBER)	DATASET(STRING)
T20627	Qaryat al Khadr	P	33.2260556	43.8602778	T
T20886	Qaryat Tall as Sultan	P	33.2488889	43.8447222	T
T20681	Qaryat an Na'imiyah	P	33.2697222	43.8383333	T
T20673	Qaryat Al Yunis	P	33.2494444	43.8722222	T
T28621	Bustan Zi'al	V	33.2195556	43.8595556	T
T20733	Qaryat al Tharthar	P	33.2613889	43.8561111	T
T20543	Qaryat Abu Hawa	P	33.2402778	43.8419444	T
T20581	Qaryat al Haqq	P	33.2527778	43.8195556	T
T20519	Qaryat al 'Ankur	P	33.1908333	43.8363889	T
T11923	Husayn Yusuf	P	33.2438889	43.8011111	T
D692	Qaryat al Haqqi	Pop. Place	33.25694444444444	43.81444444444444	O
O422	Al Hasa	Pop. Place	33.19305555555556	43.8325	O
D683	Qaryat al 'Ankur	Pop. Place	33.19222222222222	43.82693333333333	O
- Satellite Map:** A large satellite map on the right showing the geographical context of the places. Various locations are marked with colored pins (green, yellow, red) corresponding to the nodes in the graph. The map includes a compass rose and a scale bar.

Kang, Sehgal, Getoor, IV 07

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

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



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

HOMER: Tool for Ontology Alignment

The screenshot displays the HOMER Ontology Alignment Analysis tool interface, showing two comparative views and a detailed match information panel.

Top View (HOMER Ontology Alignment Analysis): This view shows a graph with nodes representing classes and instances. Nodes include "Bacteria", "Salmonella...", "Escherichi...", "EColi O157:H7", and "Bacterial". Edges connect these nodes, representing relationships. The interface includes a toolbar with navigation controls (back, forward, search, etc.) and a legend for "Instance", "Class", "Edges", "Prospective", and "Changes". A search bar is present at the bottom.

Bottom View (HOMER Ontology Alignment Analysis - comparative view): This view shows a graph with nodes representing classes and instances. Nodes include "Bacterial...", "FoodPoisoning", "SalmonelaP...", "73,000", "Cholera", "Acute Gast...", "Salmonella", "E coli", "Gastroente...", "TheodorEsc...", "Botulism", and "E coli". Edges connect these nodes. A detailed match information panel is visible, showing scores for a match: "Lexical similarity score: 0.45", "Structural similarity score: 0.488", "Extensional similarity score: 0.24", and "Final score: (before logical inference) 0.465, (after logical inference) 0.698". The interface includes a toolbar and a search bar.

Right Panel (Alignment information): This panel displays a list of matches with their respective scores. The matches are organized into a tree structure. The top match is "(Salmonella enterica typhirium, owl:sameAs, SalmonellaEnterica)" with a final score of 0.414. The second match is "(Gastroenteritis, owl:sameAs, Acute Gastroenteritis)" with a final score of 0.728. The third match is "(BacterialInfection, owl:equivalentClass, Ba)" with a final score of 0.3. The bottom match is "(SalmonetaPoisoning, owl:sameAs, Salmc)" with a final score of 0.275. The panel includes buttons for "Focus", "Differences", and "Delete match".

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

● ● ● SplicePort: Motif Explorer

Return to SplicePort Homepage
Browse Donor Features
Predict Splice Sites

Acceptor Splice Site

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>

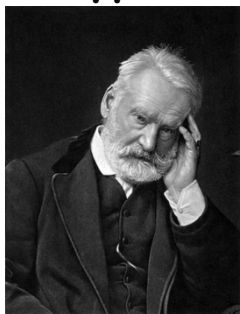
● ● ● 3. GI & Privacy

- 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

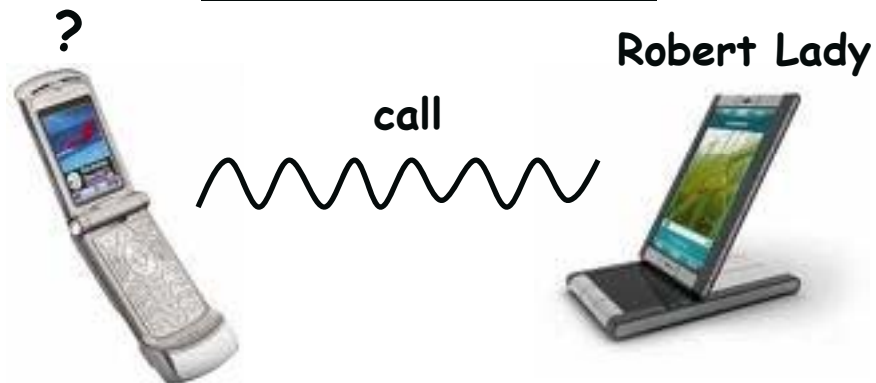
has hypertension



father-of



Communication data



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



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

● ● ● Summary: GIA & AI

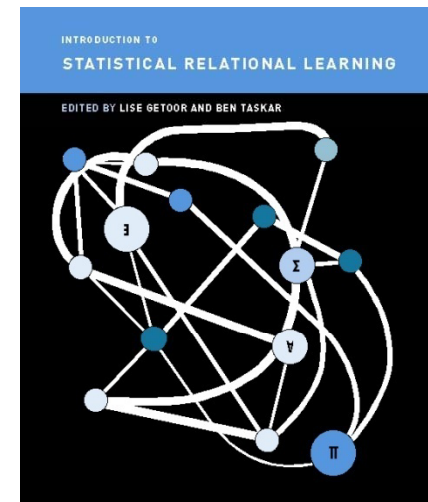
- Graph Identification can be seen as a process of **knowledge reformulation**
- In the context where we have some statistical information to help us **learn** which reformulations are more promising than others
- **Inference** is the process of transferring the learned knowledge to new situations

● ● ● 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

