Link Mining

Lise Getoor

University of Maryland, College Park

ACTIVE Summer School
4 Sep, 2009

Roadmap

- Intro to Link Mining
  - Link Mining Tasks
  - Link Mining Challenges
- Some Link Mining Algorithms
  - Collective Classification
  - Entity Resolution
  - Link Prediction
- Conclusion

Link Mining

- Traditional machine learning and data mining approaches assume:
  - A random sample of homogeneous objects from single relation
- Real world data sets:
  - Multi-relational, heterogeneous and semi-structured
- Link Mining
  - newly emerging research area at the intersection of research in social network and link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming

Linked Data

- Heterogeneous, multi-relational data represented as a graph or network
  - Nodes are objects
    - May have different kinds of objects
    - Objects have attributes
    - Objects may have labels or classes
  - Edges are links
    - May have different kinds of links
    - Links may have attributes
    - Links may be directed, are not required to be binary

Sample Domains

- web data (**web**)
- bibliographic data (**cite**)
- epidimiological data (**epi**)
- communication data (**comm**)
- customer networks (**cust**)
- collaborative filtering problems (**cf**)
- trust networks (**trust**)
- biological data (**bio**)

Link Mining Tasks

- Object Classification
- Object Type Prediction
- Link Type Prediction
- Link Prediction
- Link Cardinality Estimation
- Entity Resolution
- Group Detection
- Subgraph Discovery
- Graph Alignment

Object Classification

- Predicting the category of an object based on its attributes *and* its links *and* attributes of linked objects

- **web**: Predict the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags, etc.
- **cite**: Predict the topic of a paper, based on word occurrence, citations, co-citations
- **epi**: Predict disease type based on characteristics of the patients infected by the disease

Object Class Prediction

- Predicting the type of an object based on its attributes and its links and attributes of linked objects

- **comm**: Predict whether a communication contact is by email, phone call or mail.
- **cite**: Predict the venue type of a publication (conference, journal, workshop)

Link Type Classification

- Predicting type or purpose of link based on properties of the participating objects

- **web**: predict advertising link or navigational link; predict an advisor-advisee relationship
- **epi**: predicting whether contact is familial, co-worker or acquaintance

Predicting Link Existence

- Predicting whether a link exists between two objects

- **web**: predict whether there will be a link between two pages
- **cite**: predicting whether a paper will cite another paper
- **epi**: predicting who a patient’s contacts are

Link Cardinality Estimation I

- Predicting the number of links to an object

- **web**: predict the authoratativeness of a page based on the number of in-links; identifying hubs based on the number of out-links
- **cite**: predicting the impact of a paper based on the number of citations
- **epi**: predicting the number of people that will be infected based on the infectiousness of a disease.

Link Cardinality Estimation II

- Predicting the number of objects reached along a path from an object
- Important for estimating the number of objects that will be returned by a query

- **web**: predicting number of pages retrieved by crawling a site
- **cite**: predicting the number of citations of a particular author in a specific journal

Entity Resolution

- Predicting when two objects are the same, based on their attributes *and* their links
- aka: record linkage, duplicate elimination, identity uncertainty

- **web**: predict when two sites are mirrors of each other.
- **cite**: predicting when two citations are referring to the same paper.
- **epi**: predicting when two disease strains are the same
- **bio:** learning when two names refer to the same protein

Group Detection

- Predicting when a set of entities belong to the same group based on clustering both object attribute values and link structure

- **web –** identifying communities
- **cite** – identifying research communities

Subgraph Discovery

- Find characteristic subgraphs
- Focus of graph-based data mining (Cook & Holder, Inokuchi, Washio & Motoda, Kuramochi & Karypis, Yan & Han)

- **bio –** protein structure discovery
- **comm** – legitimate vs. illegitimate groups
- **chem** – chemical substructure discovery

Graph Alignment

- Schema mapping, schema discovery, schema reformulation

- **cite** – matching between two bibliographic sources
- **web** - discovering schema from unstructured or semi-structured data
- **bio –** mapping between two medical ontologies

Link Mining Tasks

- Object Classification
- Object Type Prediction
- Link Type Prediction
- Link Prediction
- Link Cardinality Estimation
- Entity Resolution
- Group Detection
- Subgraph Discovery
- Graph Alignment

Link Mining Challenges

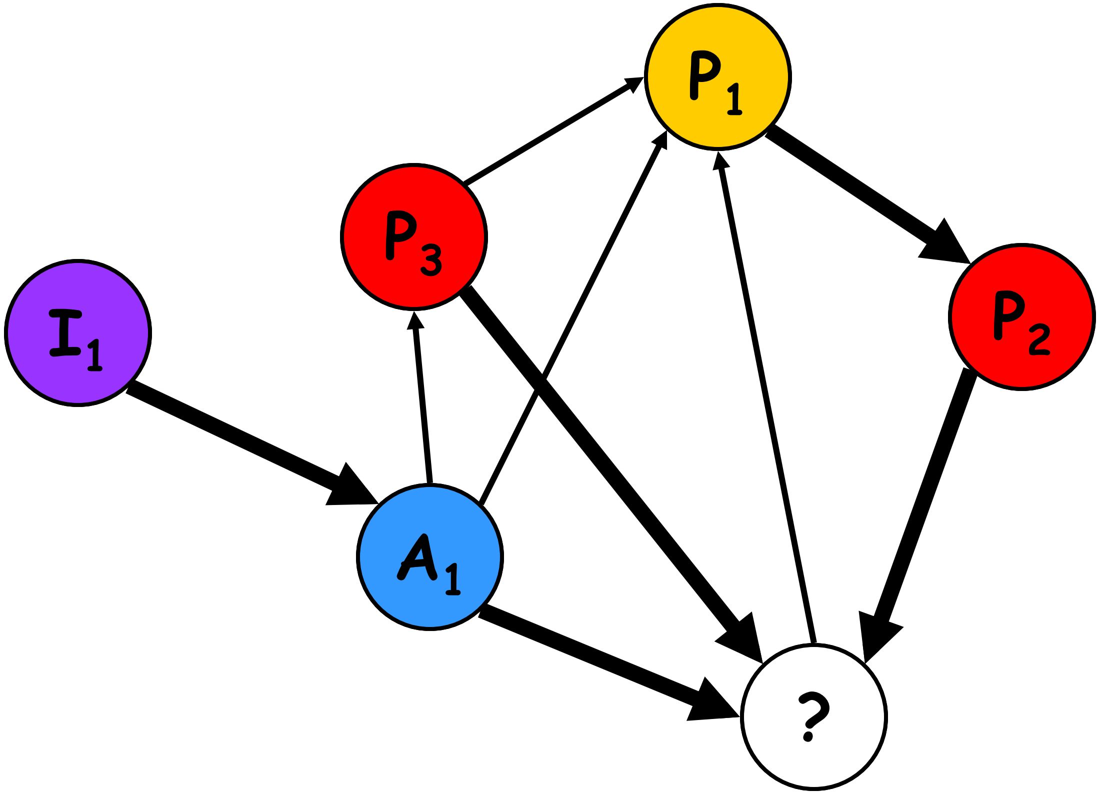
- Logical vs. Statistical dependencies
- Feature construction
- Instances vs. Classes
- Collective Classification
- Collective Consolidation
- Effective Use of Labeled & Unlabeled Data
- Link Prior Probability
- Closed vs. Open World

Challenges common to any link-based statistical model (Bayesian Logic Programs, Conditional Random Fields, Probabilistic Relational Models, Markov Logic, Relational Probability Trees, Stochastic Logic Programming to name a few)

Logical vs. Statistical Dependence

- Coherently handling two types of dependence structures:
  - Link structure - the logical relationships between objects
  - Probabilistic dependence - statistical relationships between attributes
- Challenge: statistical models that support rich logical relationships
- Model search complicated by the fact that attributes can depend on arbitrarily linked attributes -- issue: how to search this huge space

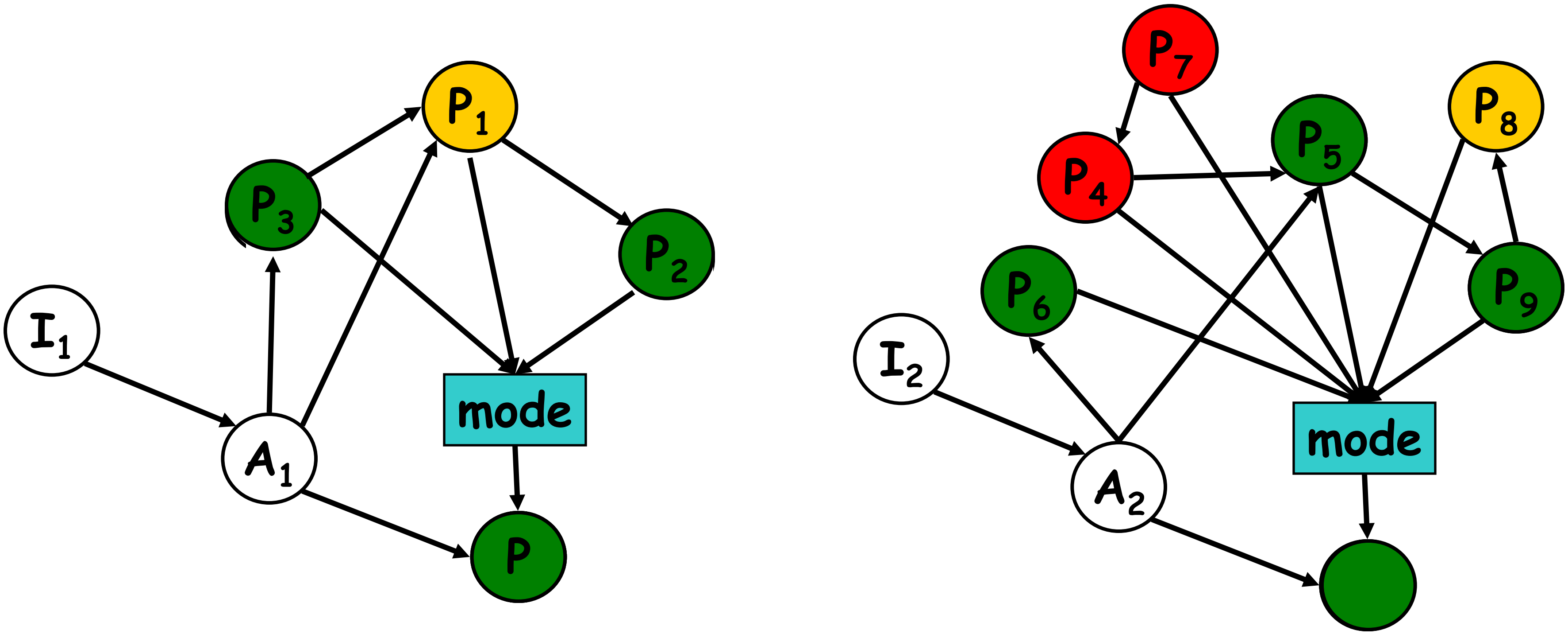
Model Search



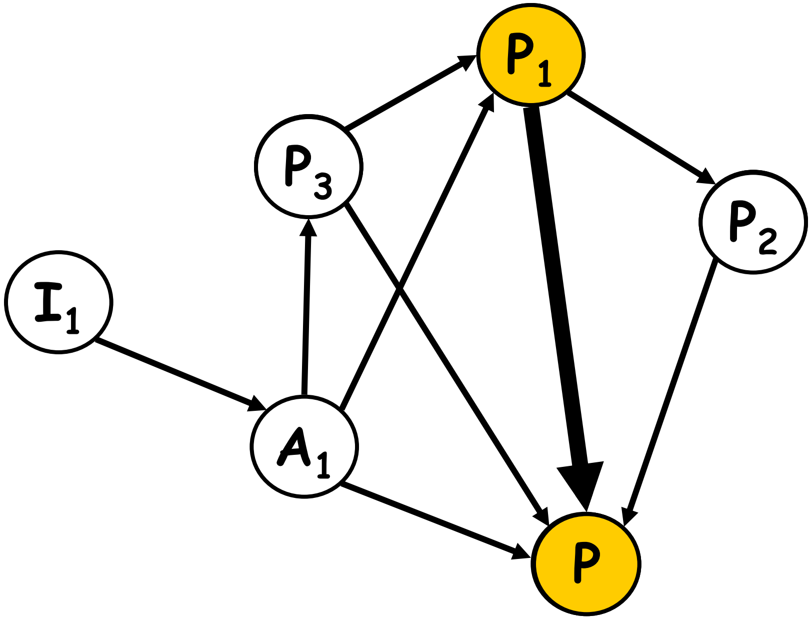
Feature Construction

- In many cases, objects are linked to a **set** of objects. To construct a single feature from this set of objects, we may either use:
  - Aggregation
  - Selection

Aggregation



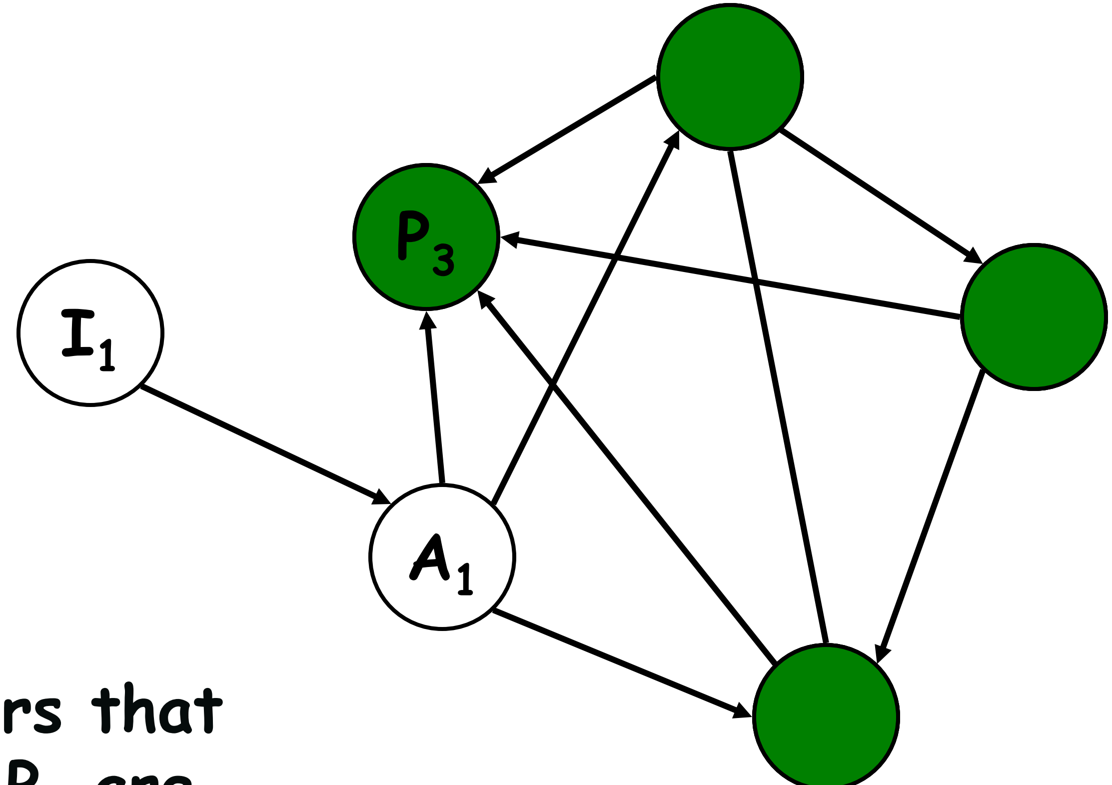
Selection

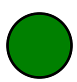


Individuals vs. Classes

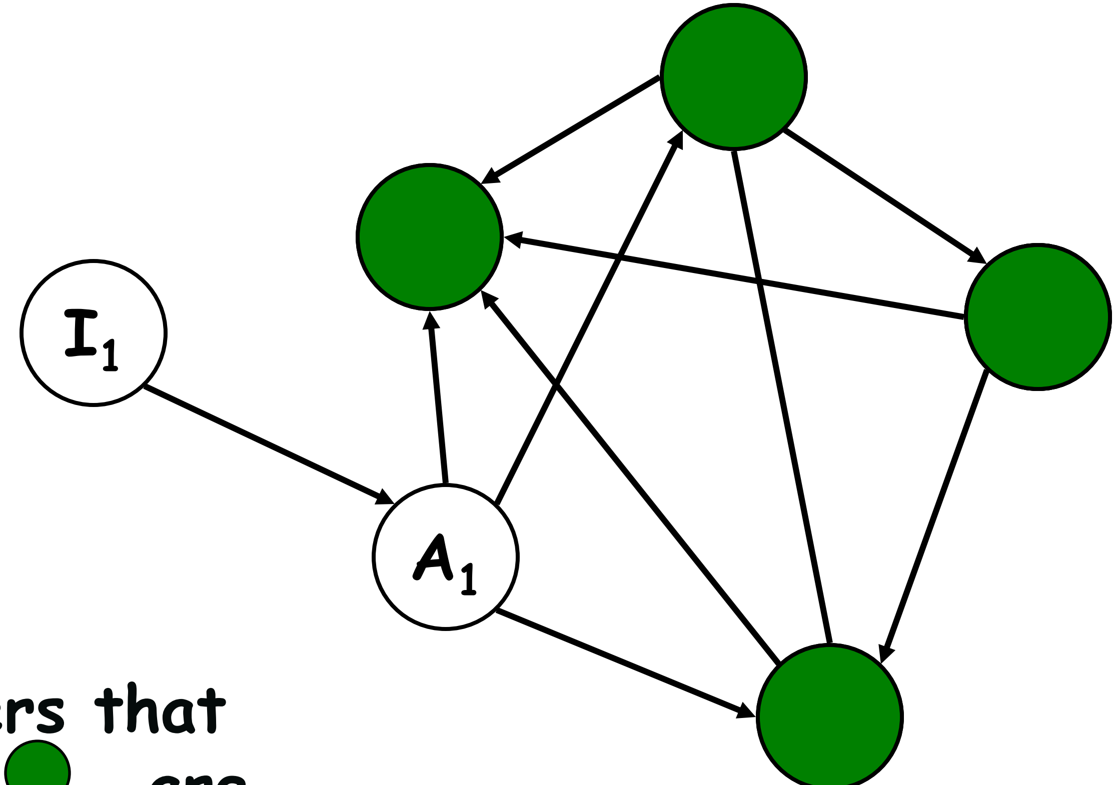
- Does model refer
  - explicitly to individuals
  - classes or generic categories of individuals
- On one hand, we'd like to be able to model that a connection to a particular individual may be highly predictive
- On the other hand, we'd like our models to generalize to new situations, with different individuals

Instance-based Dependencies



Papers that cite P_3 are likely to be 

Class-based Dependencies



Papers that cite ● are likely to be ●

Collective classification

- Using a link-based statistical model for classification
- Inference using learned model is complicated by the fact that there is correlation between the object labels

Collective Resolution

- Using a link-based statistical model for entity resolution
- Consolidation decisions should not be made independently

Labeled & Unlabeled Data

- In link-based domains, unlabeled data provide three sources of information:
  - Helps us infer object attribute distribution
  - Links between unlabeled data allow us to make use of attributes of linked objects
  - Links between labeled data and unlabeled data (training data and test data) help us make more accurate inferences

Link Prior Probability

- The prior probability of any particular link is typically extraordinarily low
- For medium-sized data sets, we have had success with building explicit models of link existence
- It may be more effective to model links at higher level--required for large data sets!

Closed World vs. Open World

- The majority of SRL approaches make a closed world assumption, which assumes that we know all the potential entities in the domain
- In many cases, this is unrealistic

Link Mining Summary

- **Link Mining Tasks**
  - Object Classification
  - Object Type Prediction
  - Link Type Prediction
  - Link Prediction
  - Link Cardinality Estimation
  - Entity Resolution
  - Group Detection
  - Subgraph Discovery
  - Graph Alignment
- **Link Mining Challenges**
  - Logical vs. Statistical dependencies
  - Feature construction
  - Instances vs. Classes
  - Collective Classification
  - Collective Resolution
  - Effective Use of Labeled & Unlabeled Data
  - Link Prior Probability
  - Closed vs. Open World

Roadmap

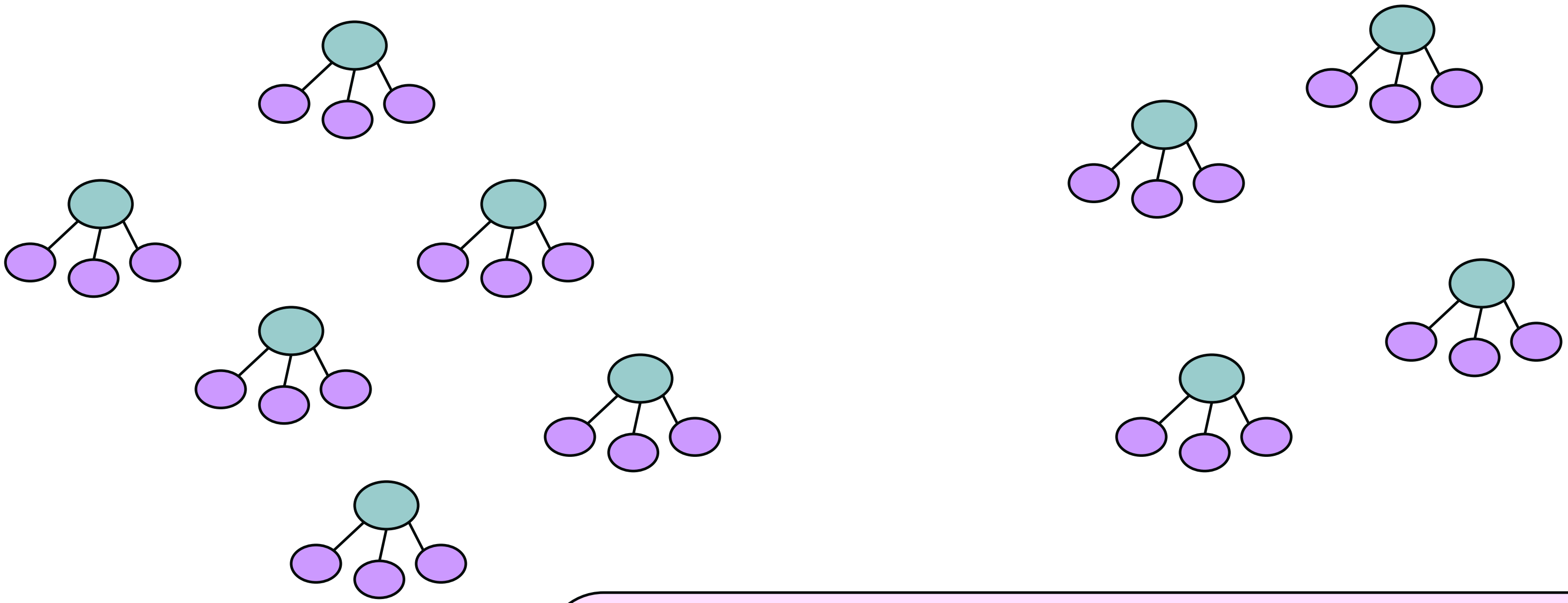
- Intro to Link Mining
  - Link Mining Tasks
  - Link Mining Challenges
- **Some Link Mining Algorithms**
  - Collective Classification
  - Entity Resolution
  - Link Prediction
- Conclusion

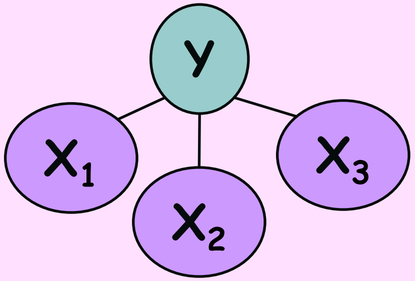
Collective Classification

- **The Problem**
- Collective Relational Classification
- Algorithms

Traditional Classification

Training Data **Test Data**

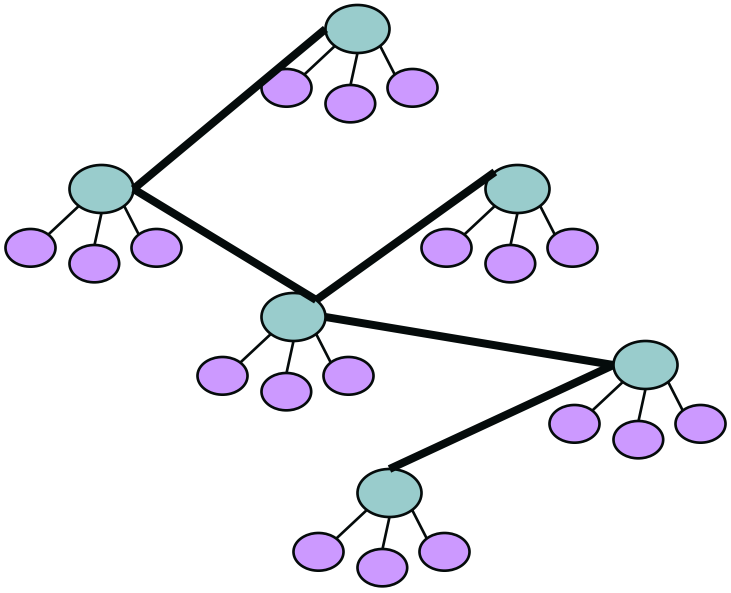




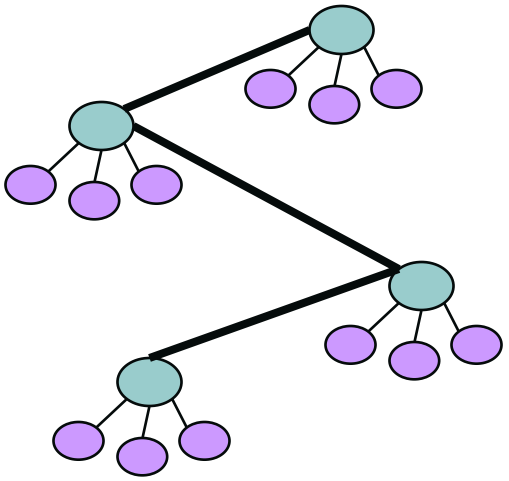
Predict Y based on attributes X_i

Relational Classification (1)

Training Data



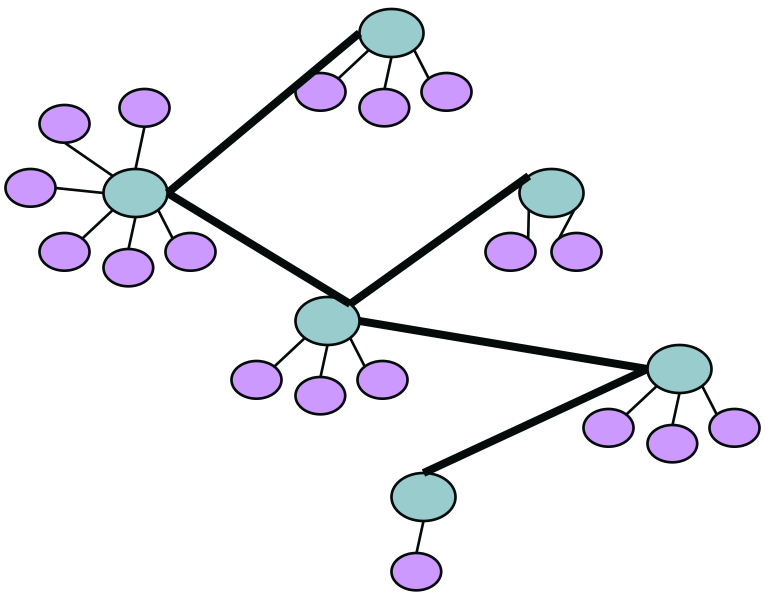
Test Data



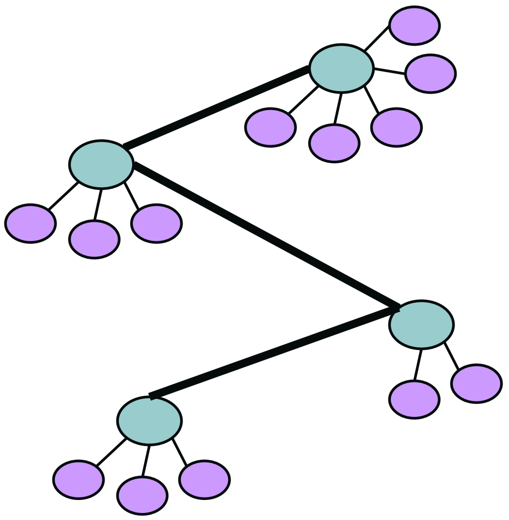
Correlations among linked instances
autocorrelation: labels are likely to be the same
homophily: similar nodes are more likely to be linked

Relational Classification (2)

Training Data

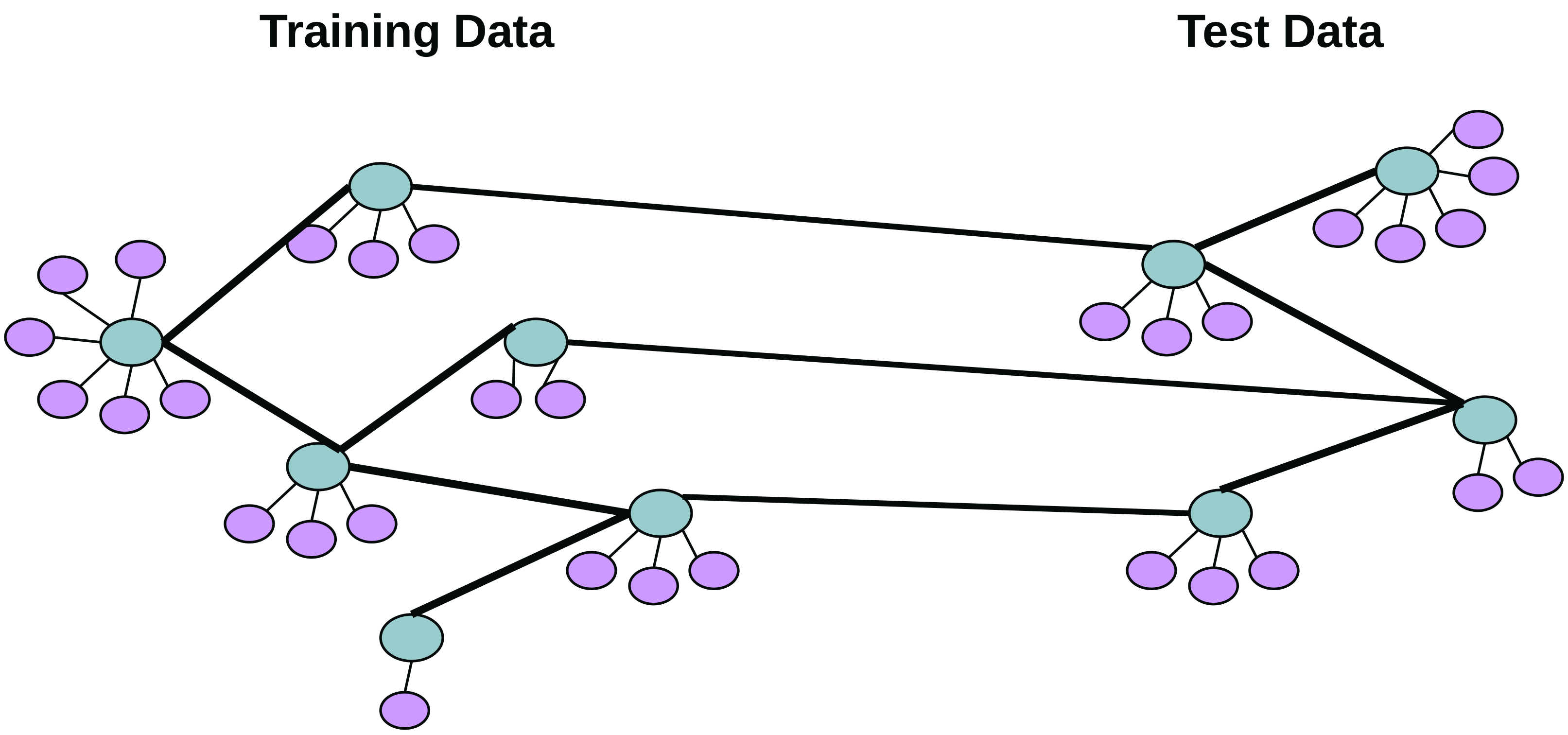


Test Data



Irregular graph structure

Relational Classification (3)



Links between training set & test set
learning with partial labels or within network classification

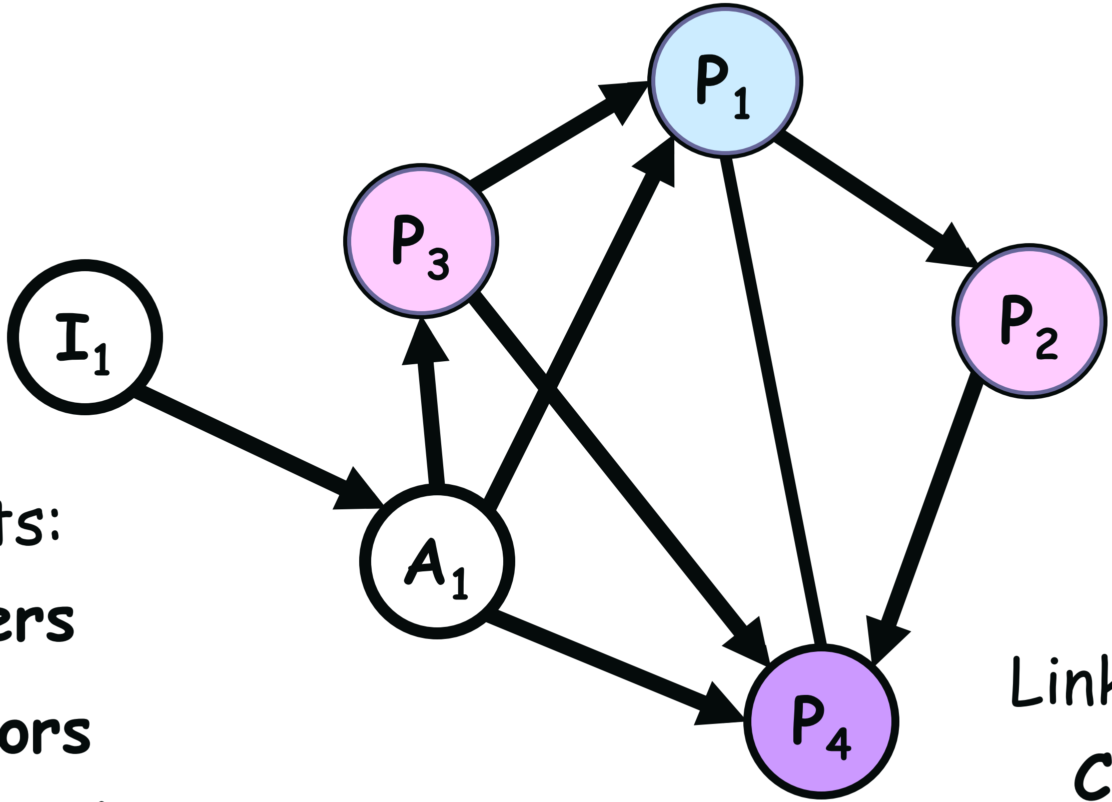
The Problem

- Relational Classification: predicting the category of an object based on its attributes *and* its links *and* attributes of linked objects

- Collective Classification: jointly predicting the categories for a collection of connected, unlabelled objects

Neville & Jensen 00, Taskar , Abbeel & Koller 02, Lu & Getoor 03, Neville, Jensen & Galliger 04, Sen & Getoor TR07, Macskassy & Provost 07, Gupta, Diwam & Sarawagi 07, Macskassy 07, McDowell, Gupta & Aha 07

Example: Linked Bibliographic Data



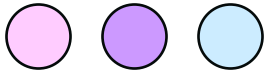
Objects:

- Papers
- Authors
- Institutions

Links:

- Citation
- Co-Citation
- Author-of
- Author-affiliation

Labels:

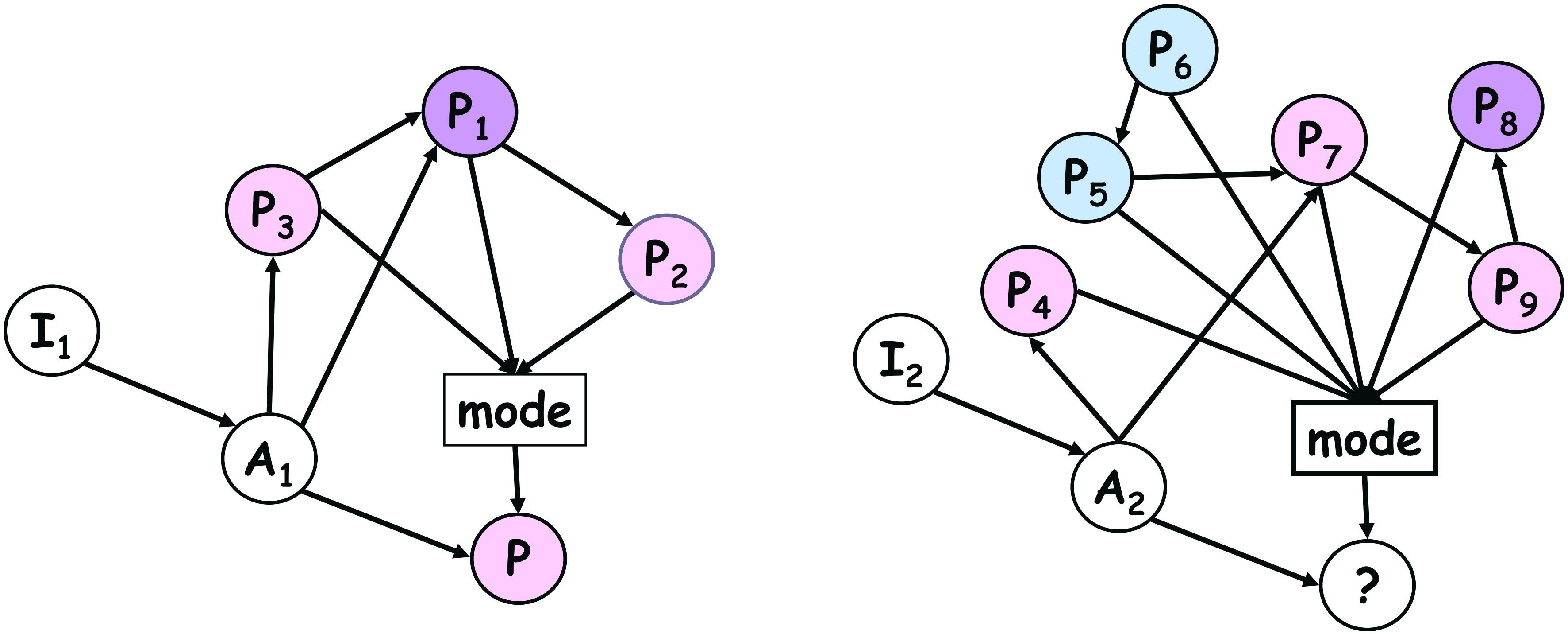


Feature Construction

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

Kramer, Lavrac & Flach 01, Perlich & Provost 03, 04, 05, Popescul & Ungar 03, 05, 06, Lu & Getoor 03, Gupta, Diwam & Sarawagi 07

Simple Aggregation



Other aggregates: count, min, max, prop, exists, selection

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

Aggregate Features Used

	Mode	Prop	Count	Exists	SQL	FOL
PRMs, Koller et al.	X				X	
RMNs, Taskar et al.					X	
MLNs, Domingos et al.						X
RDNs, Neville et al.						X
Lu & Getoor, ICML03	X		X	X		
Sen & Getoor, TR07	X		X	X		
Maskassy & Provost, JMLR07		X				
Gupta et al,. ICML07	X		X			
McDowell et al., AAAI07		X				

Formulation

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

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

CC Inference Algorithms

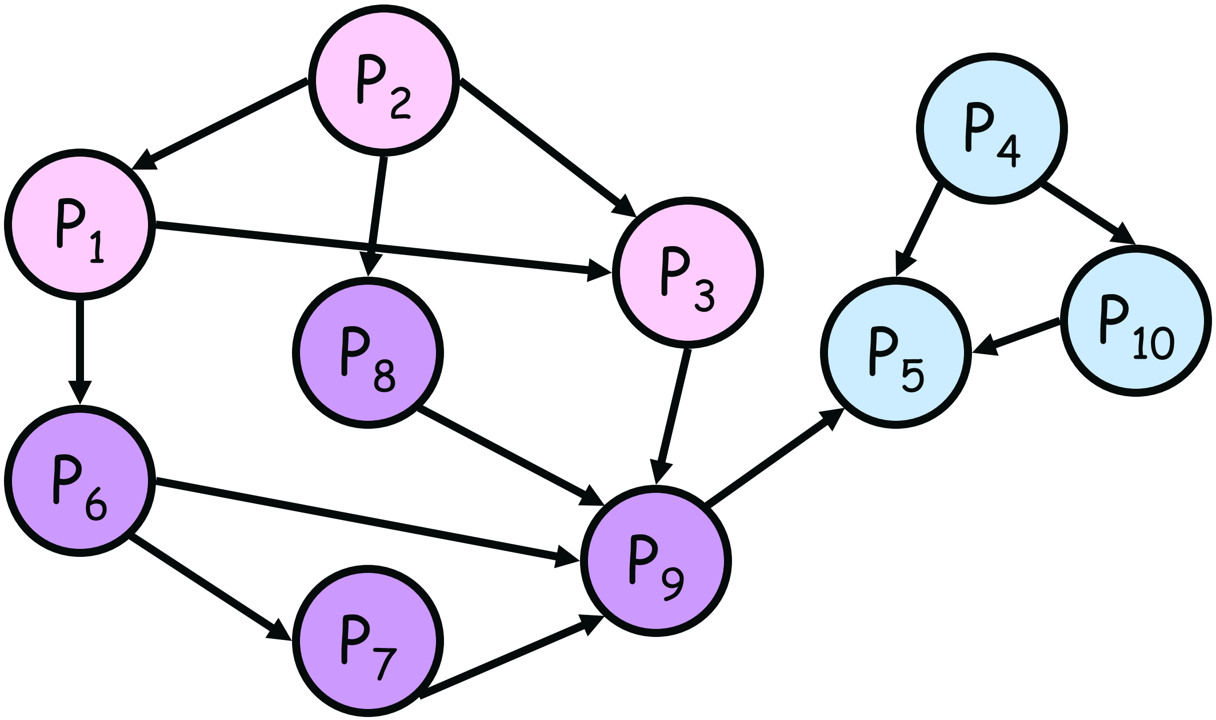
	MF	LBP	Gibbs	ICA
Chakrabarti et al SIGMOD98	X			
Jensen & Neville SRL00				X
Getoor et al. IJCAI01 WS		X		
Taskar et al. UAI02		X		
Lu & Getoor ICML03				X
Neville & Jensen KDD04			X	
Sen & Getoor TR07	X	X		X
Maskassy & Provost JMLR07	X		X	X
Gupta et al. ICML07		X		X
McDowell et al. AAAI07			X	X

Local Classifiers Used in ICA

	NB	LR	DT	kNN	wvRN
Chakrabarti et al. 1998	X				
Jensen & Neville 2000	X				
Lu & Getoor ICML03	X	X			
Neville et al. KDD04	X		X		
Macskassy & Provost JMLR07					X
McDowell et al. AAAI07	X			X	

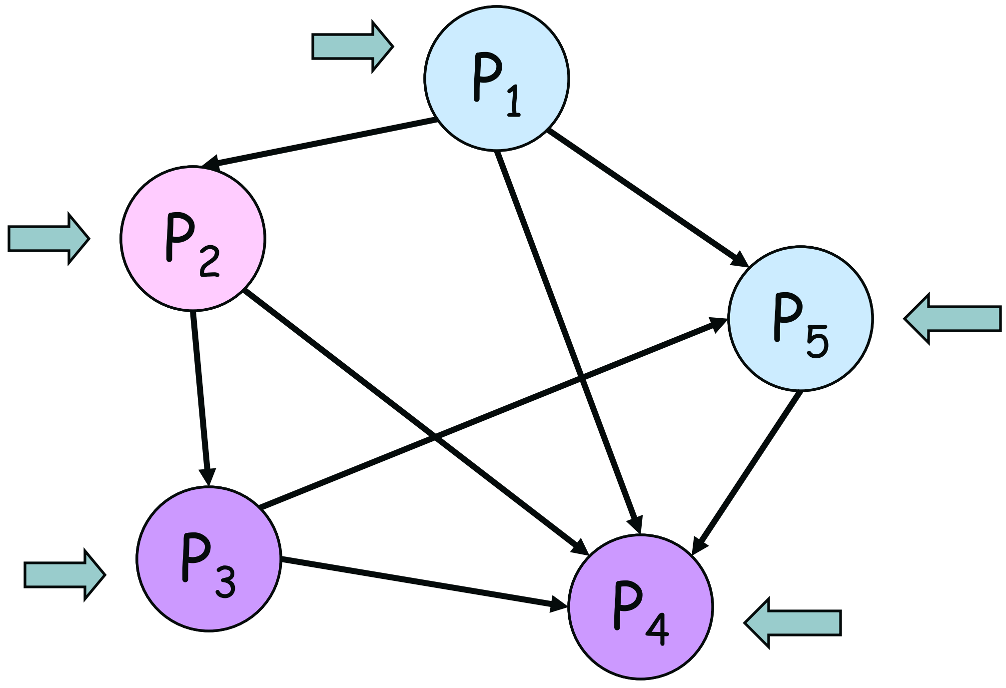
ICA: Learning

- label set:



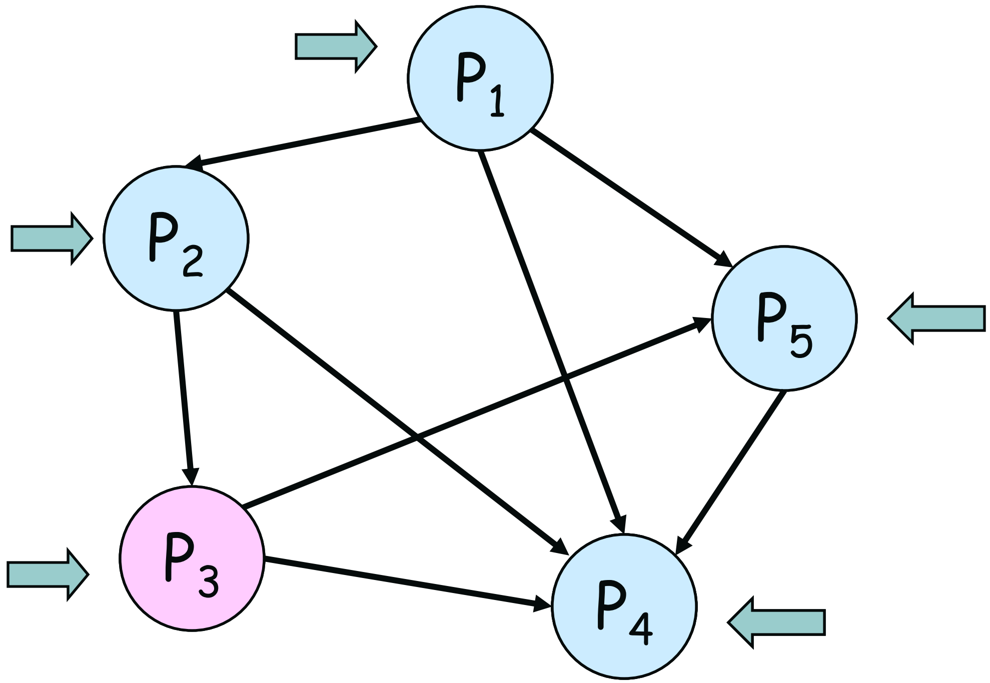
Learn model from fully labeled training set

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

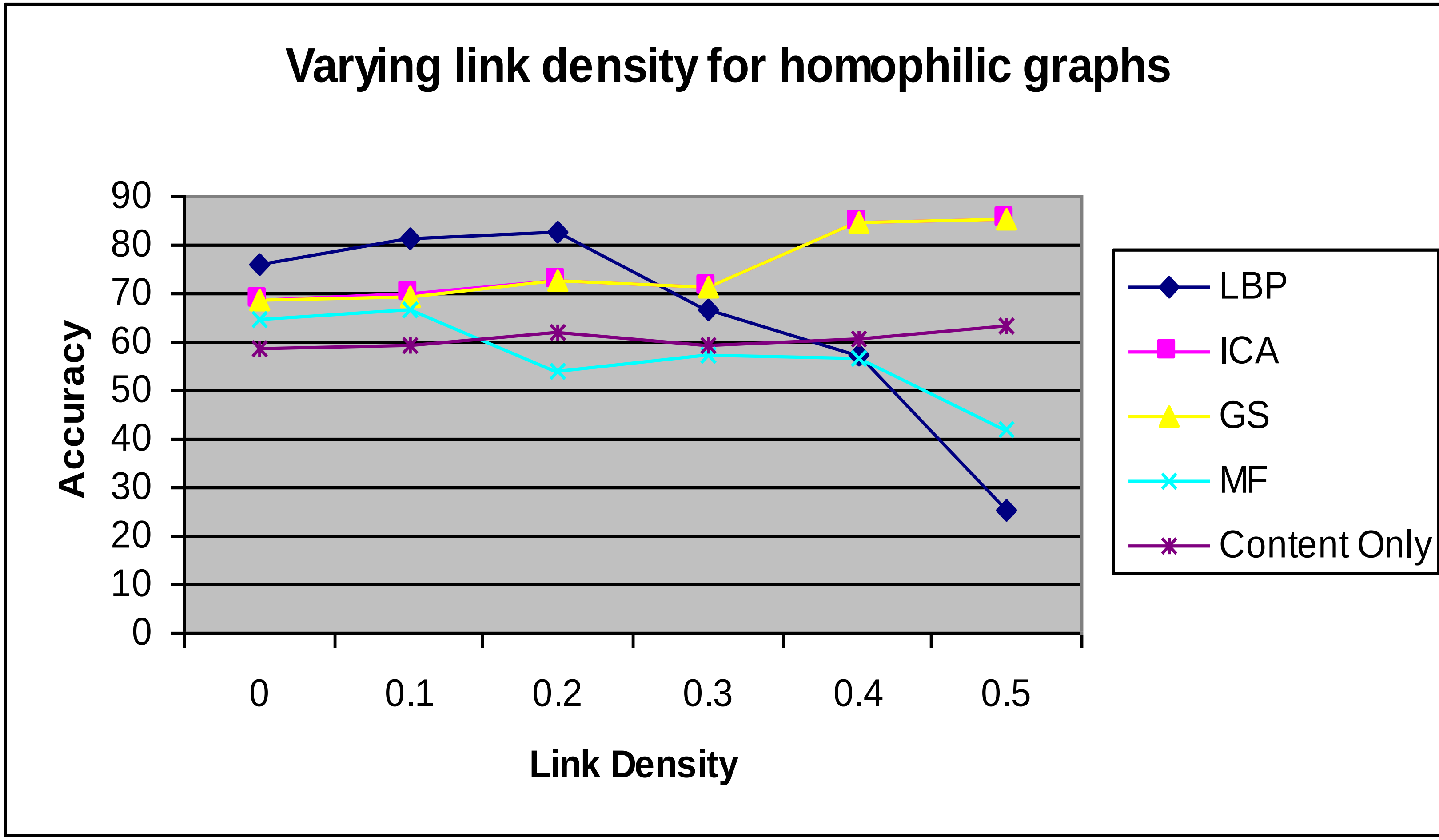
- Datasets
  - Real Data
    - Bibliographic Data (Cora & Citeseer), WebKB, etc.
  - Synthetic Data
    - Data generator which can vary the class label correlations (homophily), attribute noise, and link density

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

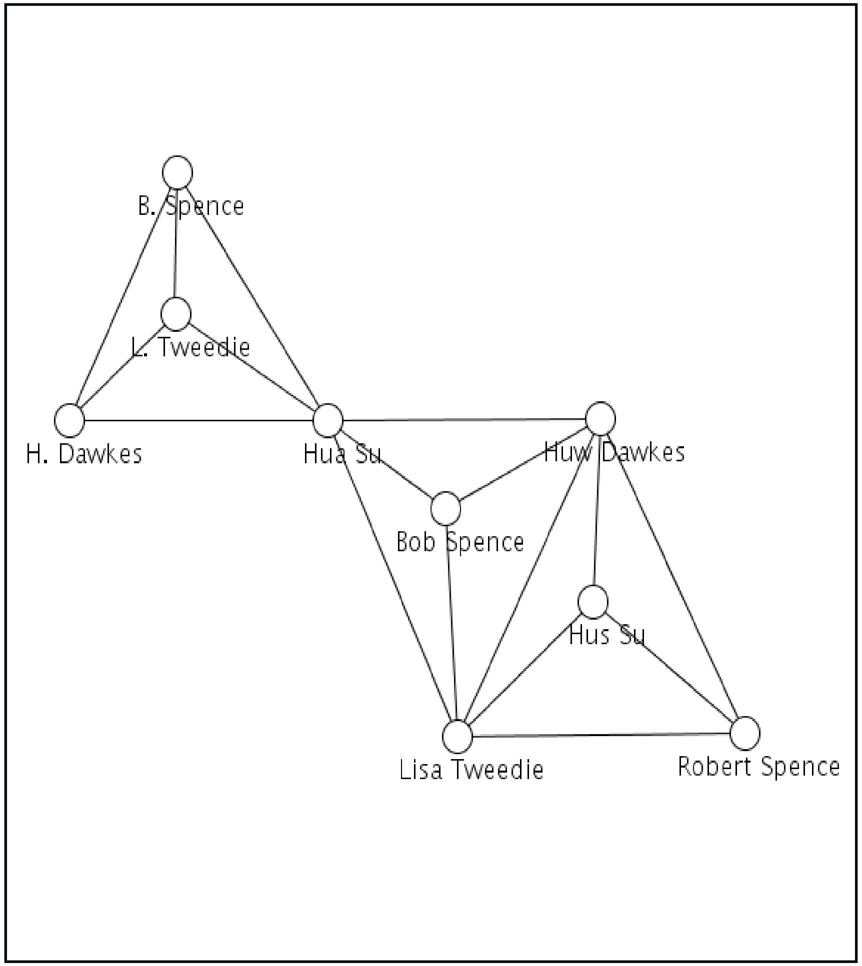
Roadmap

- Intro to Link Mining
  - Link Mining Tasks
  - Link Mining Challenges
- Some Link Mining Algorithms
  - Collective Classification
  - **Entity Resolution**
  - Link Prediction
- Conclusion

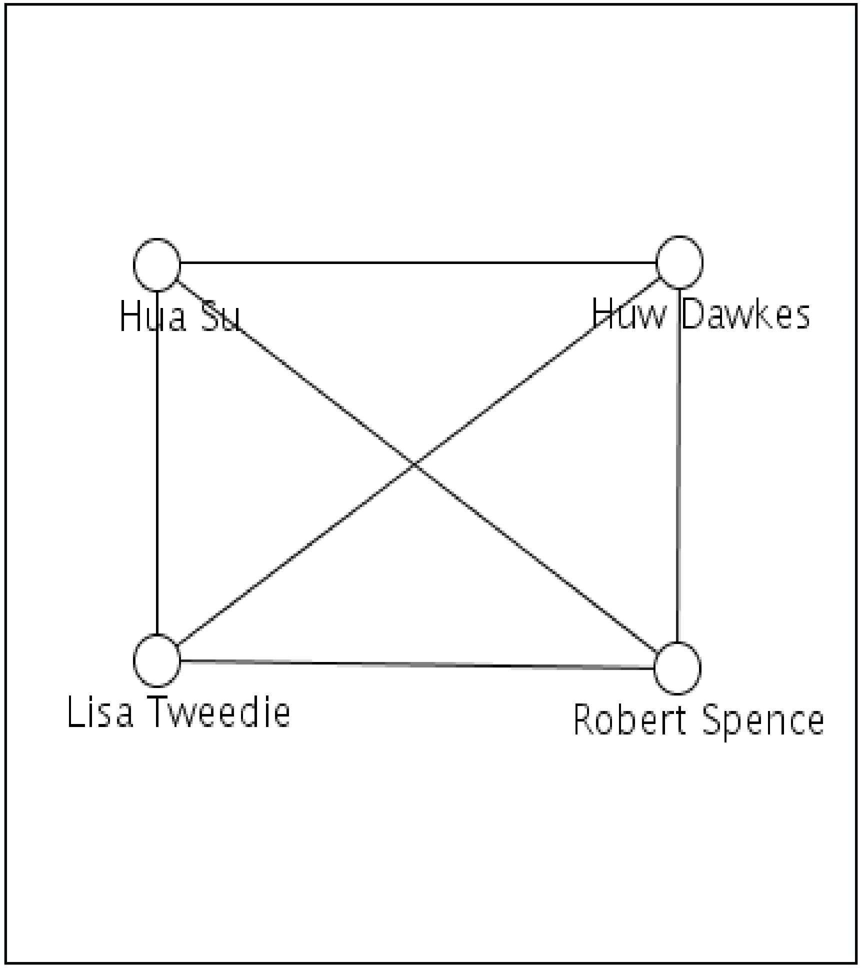
Entity Resolution

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

InfoVis Co-Author Network Fragment

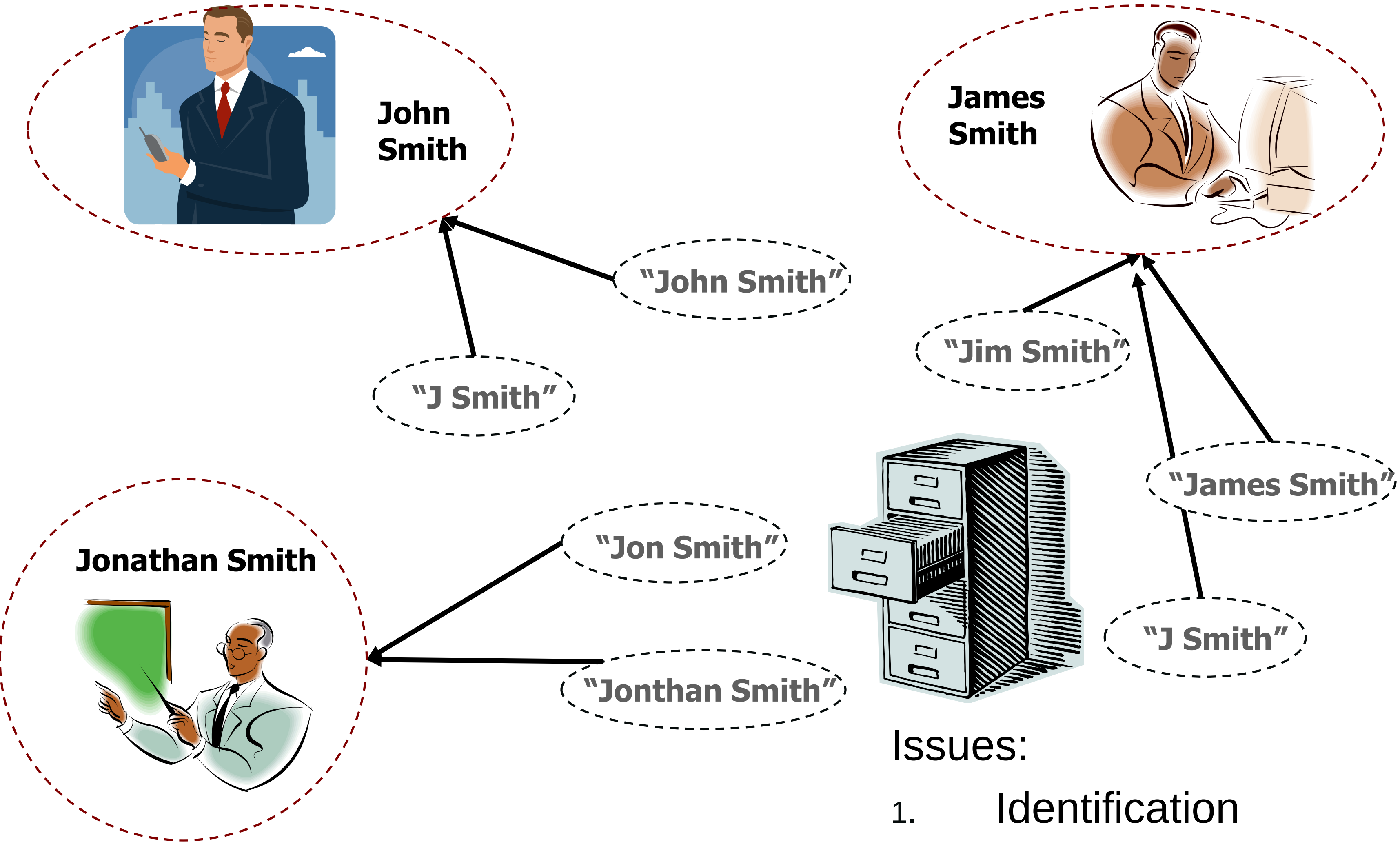


before



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

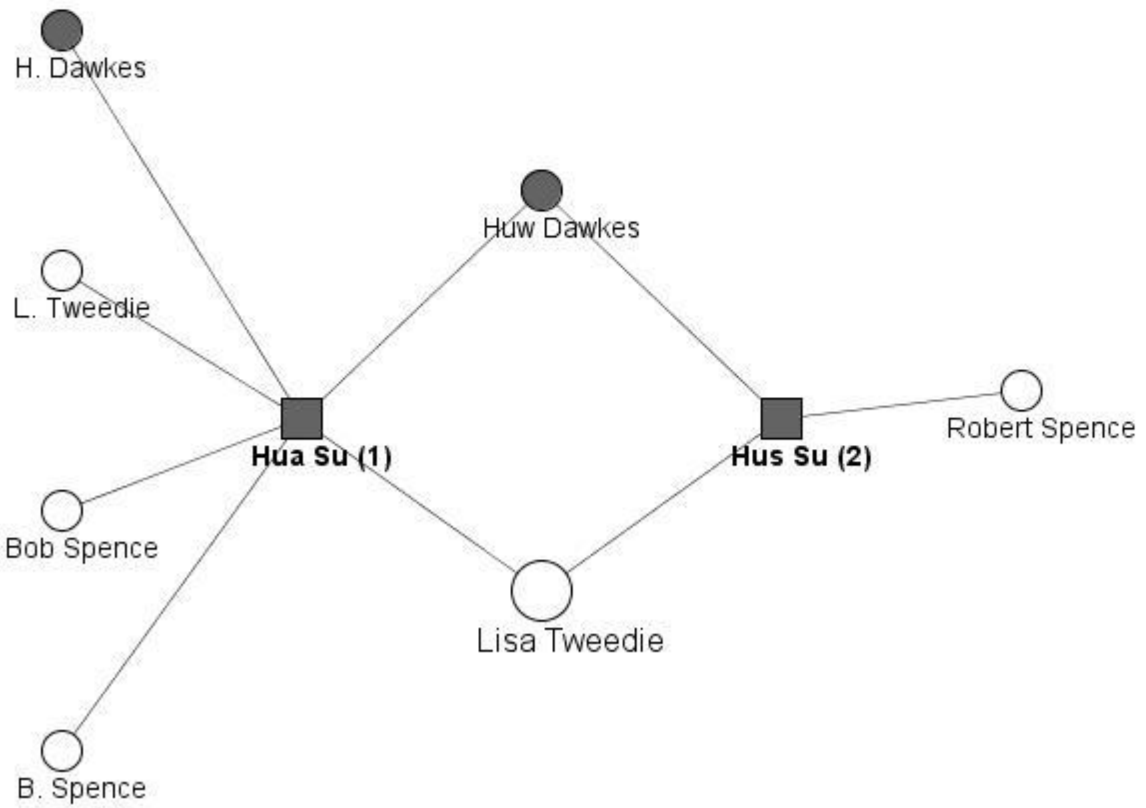
Relational Entity Resolution

- 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

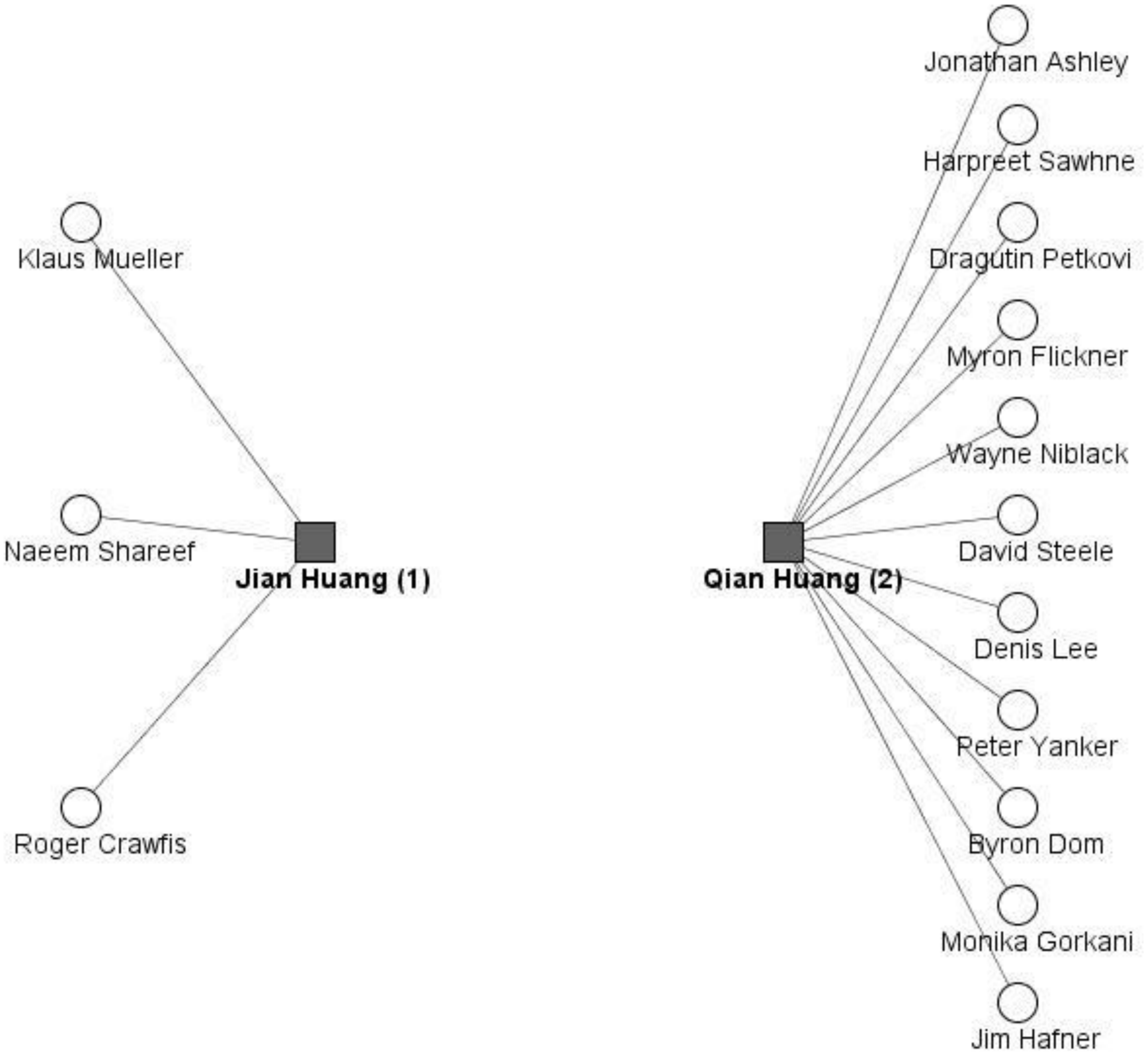
Relational Identification



Very similar names.

Added evidence from shared co-authors

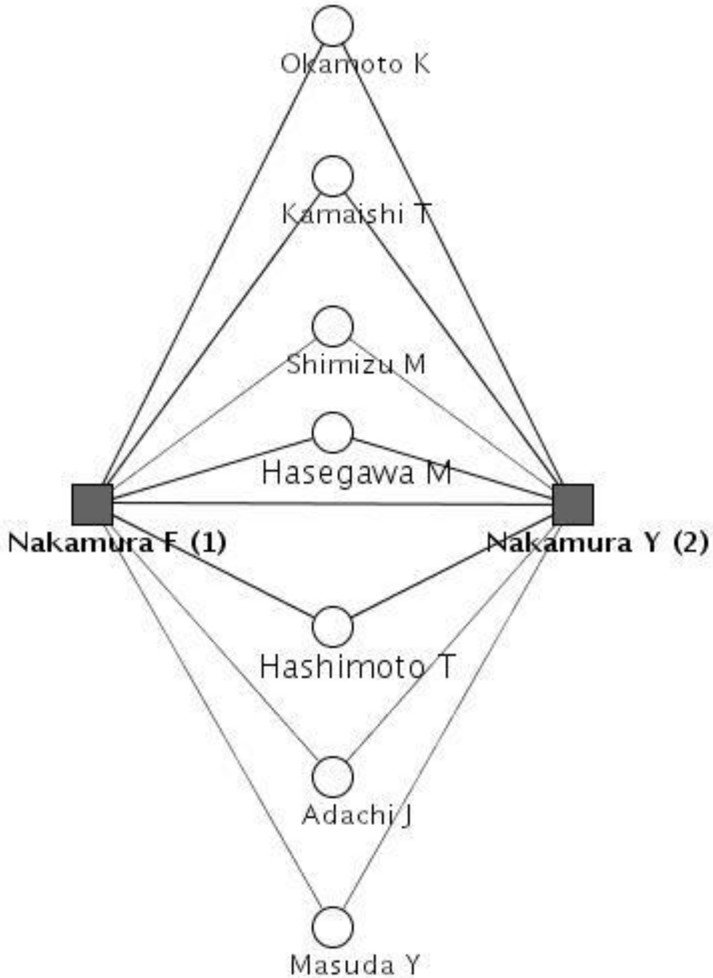
Relational Disambiguation



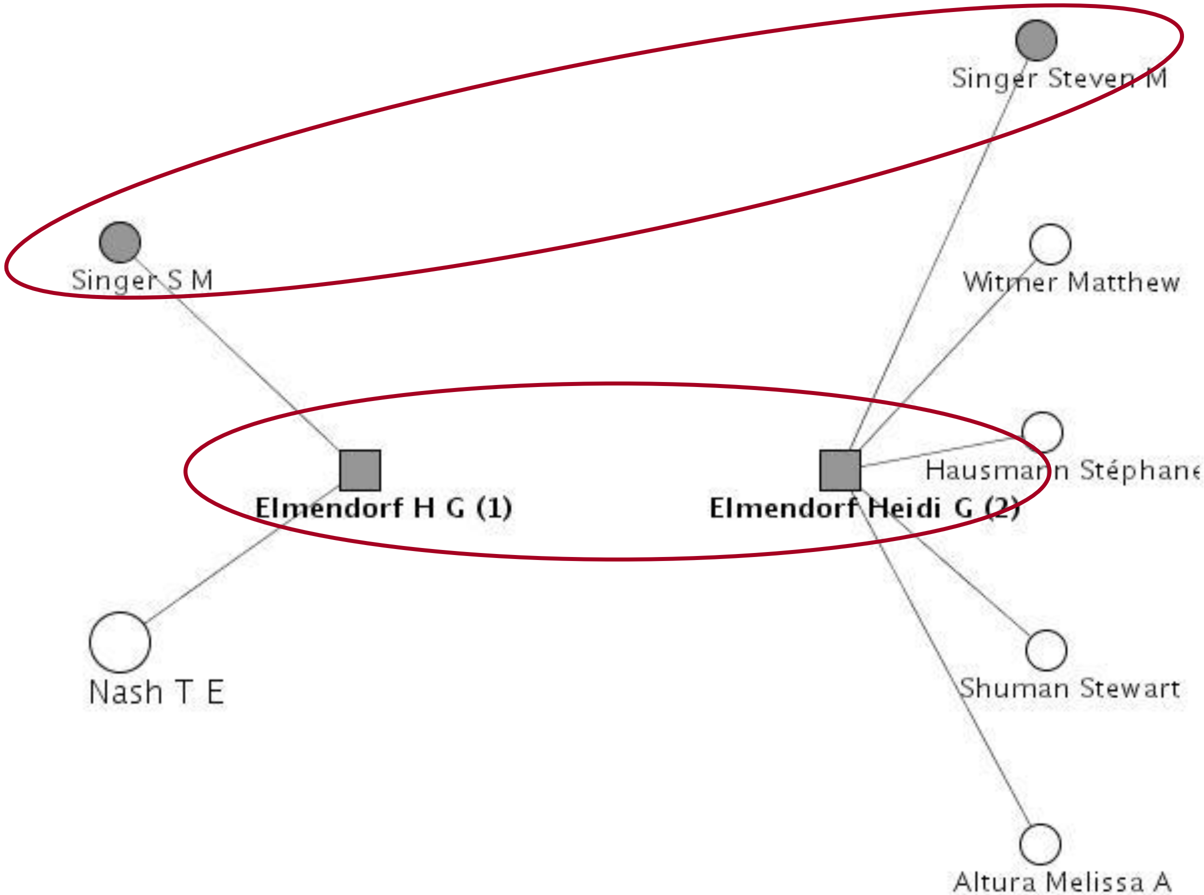
Very similar names but no shared collaborators

Relational Constraints

Co-authors are typically distinct



Collective Entity Resolution



One resolution provides evidence for another => joint resolution

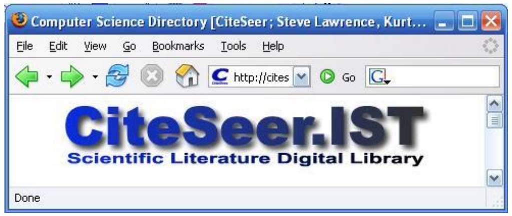
Entity Resolution with Relations

- 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

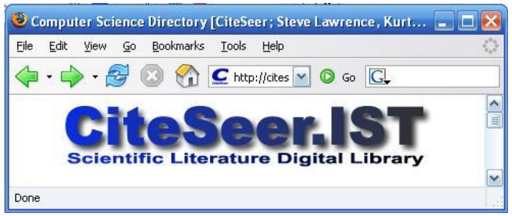
P2: "*Partitioning Mapping of Unstructured Meshes to Parallel Machine Topologies*", C. Walshaw, M. Cross, M. G. Everett, S. Johnson, K. McManus

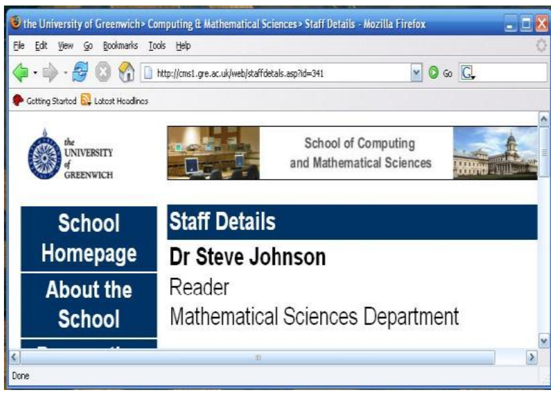
P3: "*Dynamic Mesh Partitioning: A Unied Optimisation and Load-Balancing Algorithm*", C. Walshaw, M. Cross, M. G. Everett

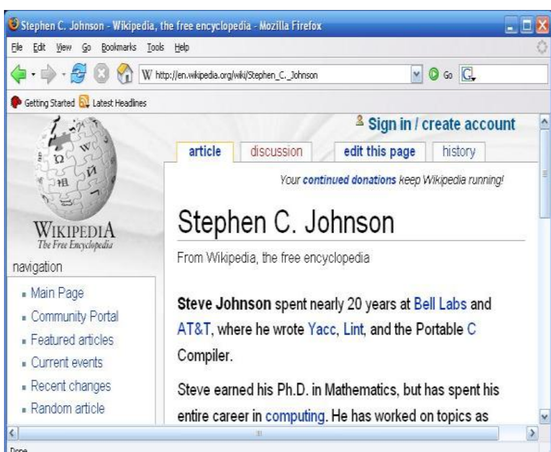
P4: "*Code Generation for Machines with Multiregister Operations*", Alfred V. Aho, Stephen C. Johnson, Jefferey D. Ullman

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)

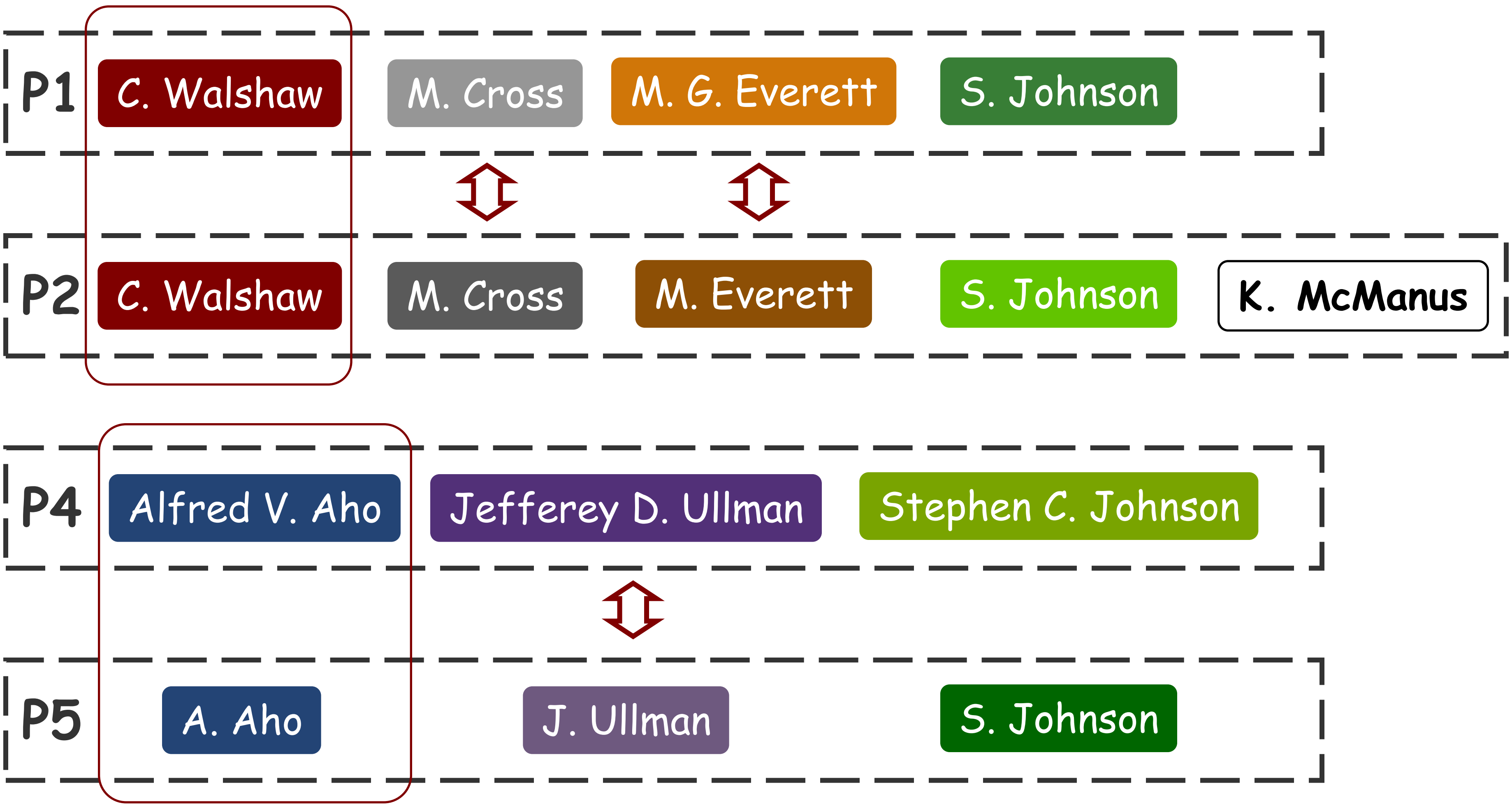
P1 C. Walshaw M. Cross M. G. Everett S. Johnson

P2 C. Walshaw M. Cross M. Everett S. Johnson K. McManus

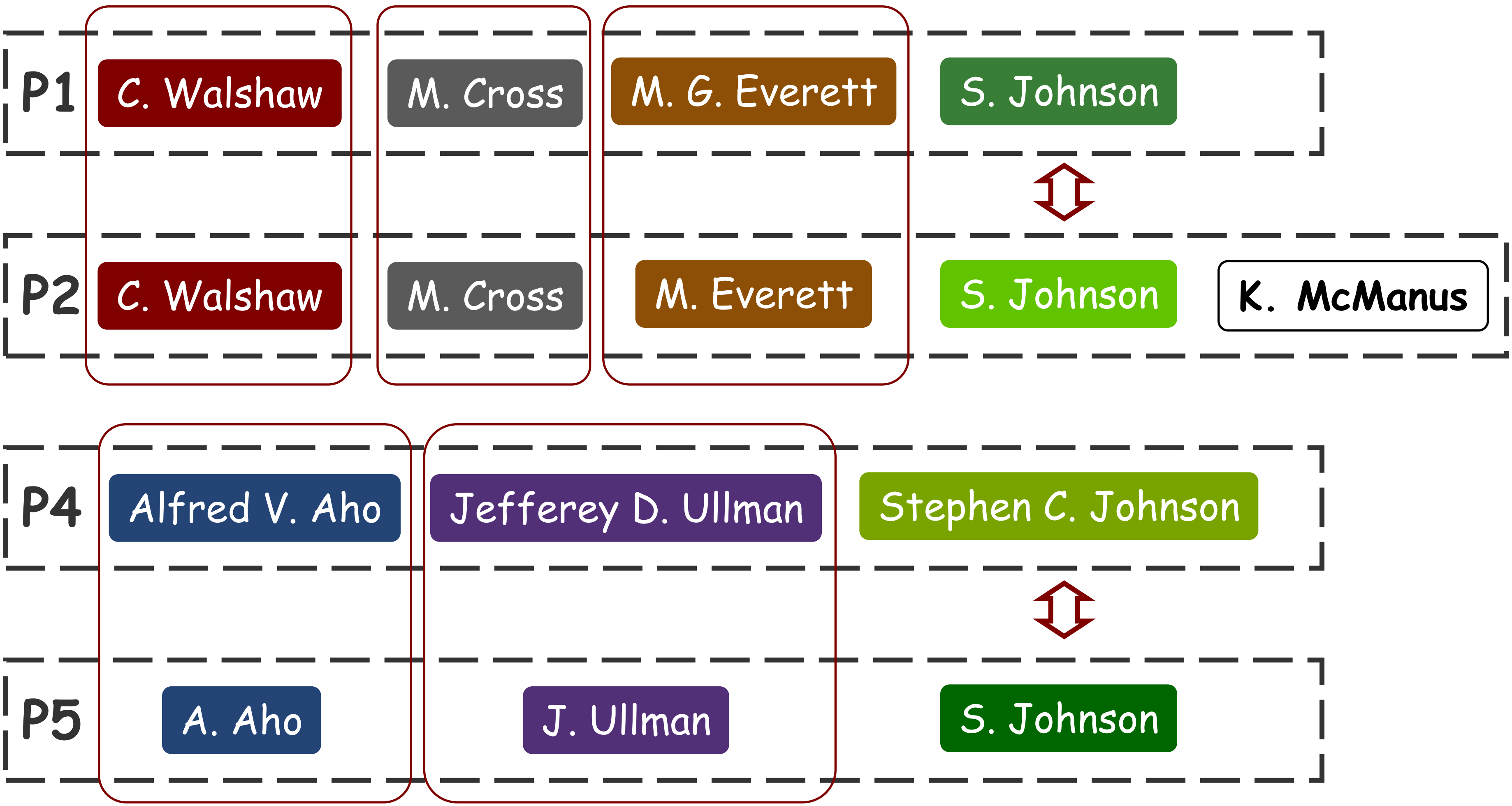
P4 Alfred V. Aho Jefferey D. Ullman Stephen C. Johnson

P5 A. Aho J. Ullman S. Johnson

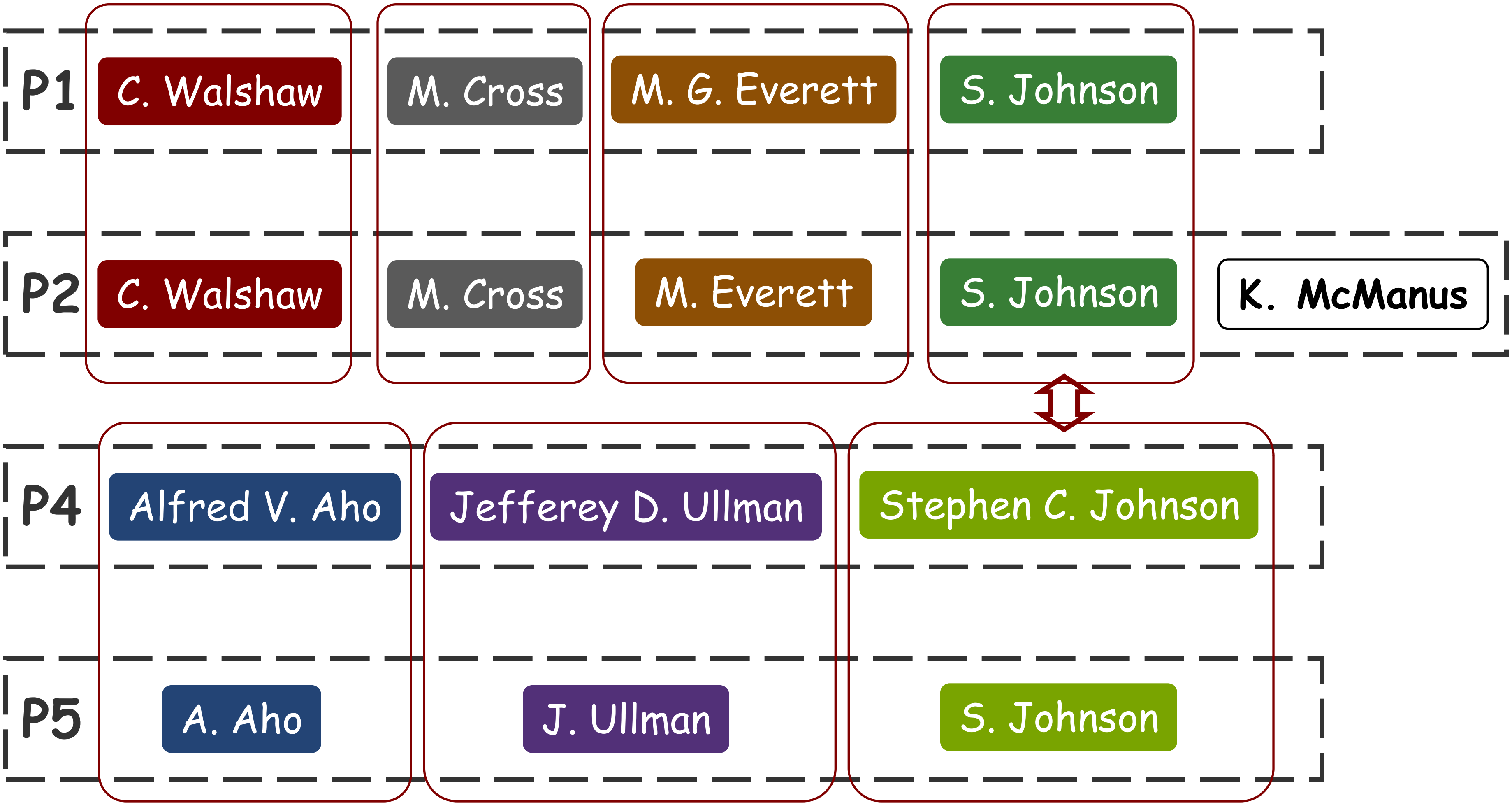
Relational Clustering (RC-ER)



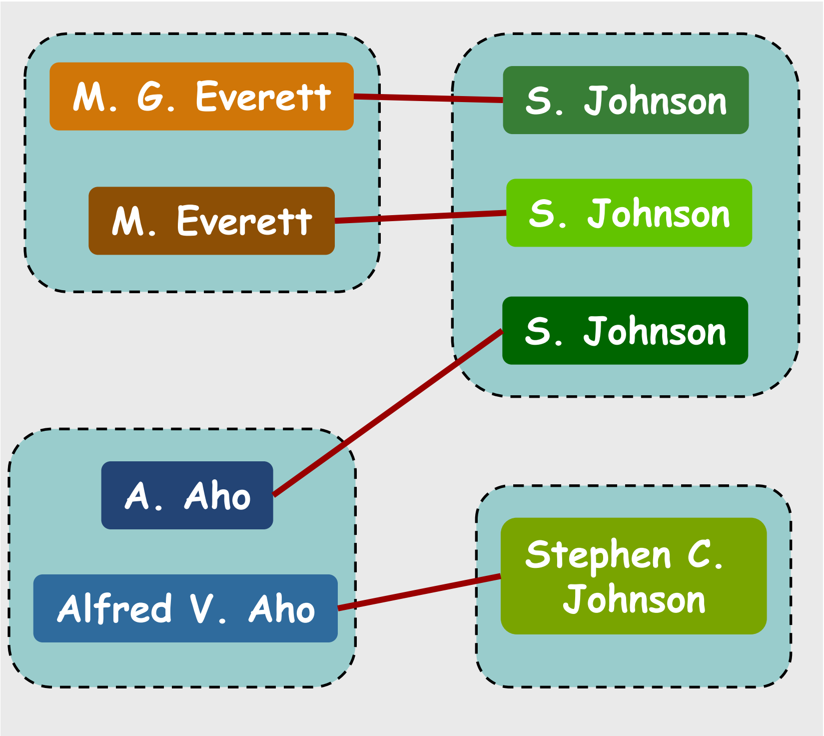
Relational Clustering (RC-ER)

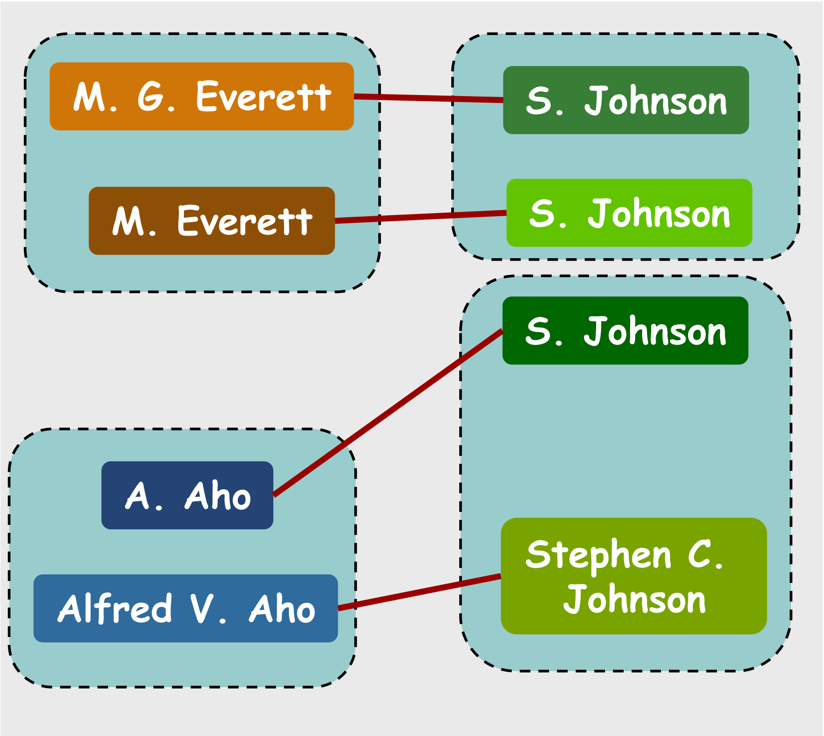


Relational Clustering (RC-ER)



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

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

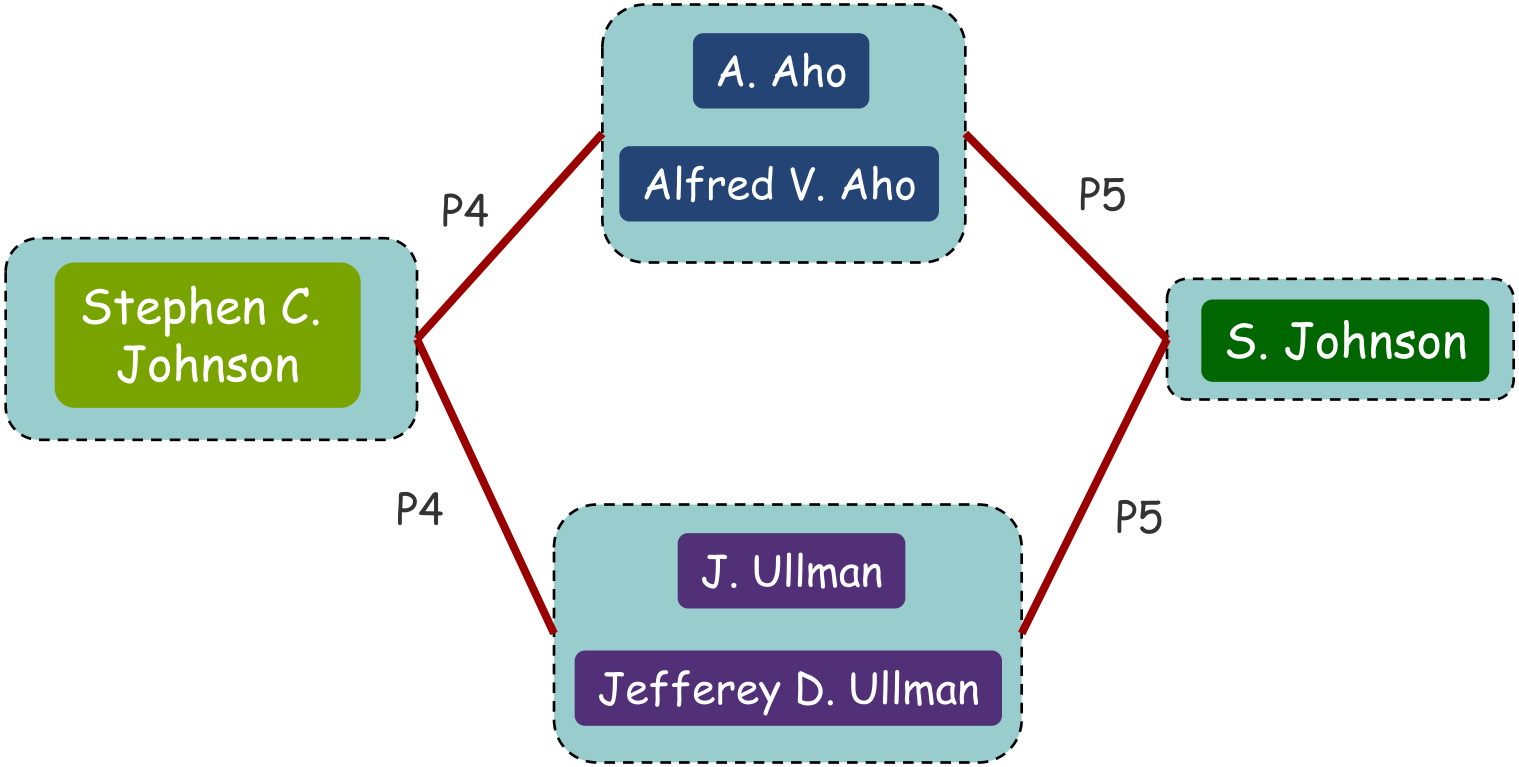
Similarity of attributes

Common cluster neighborhood

Measures for Attribute Similarity

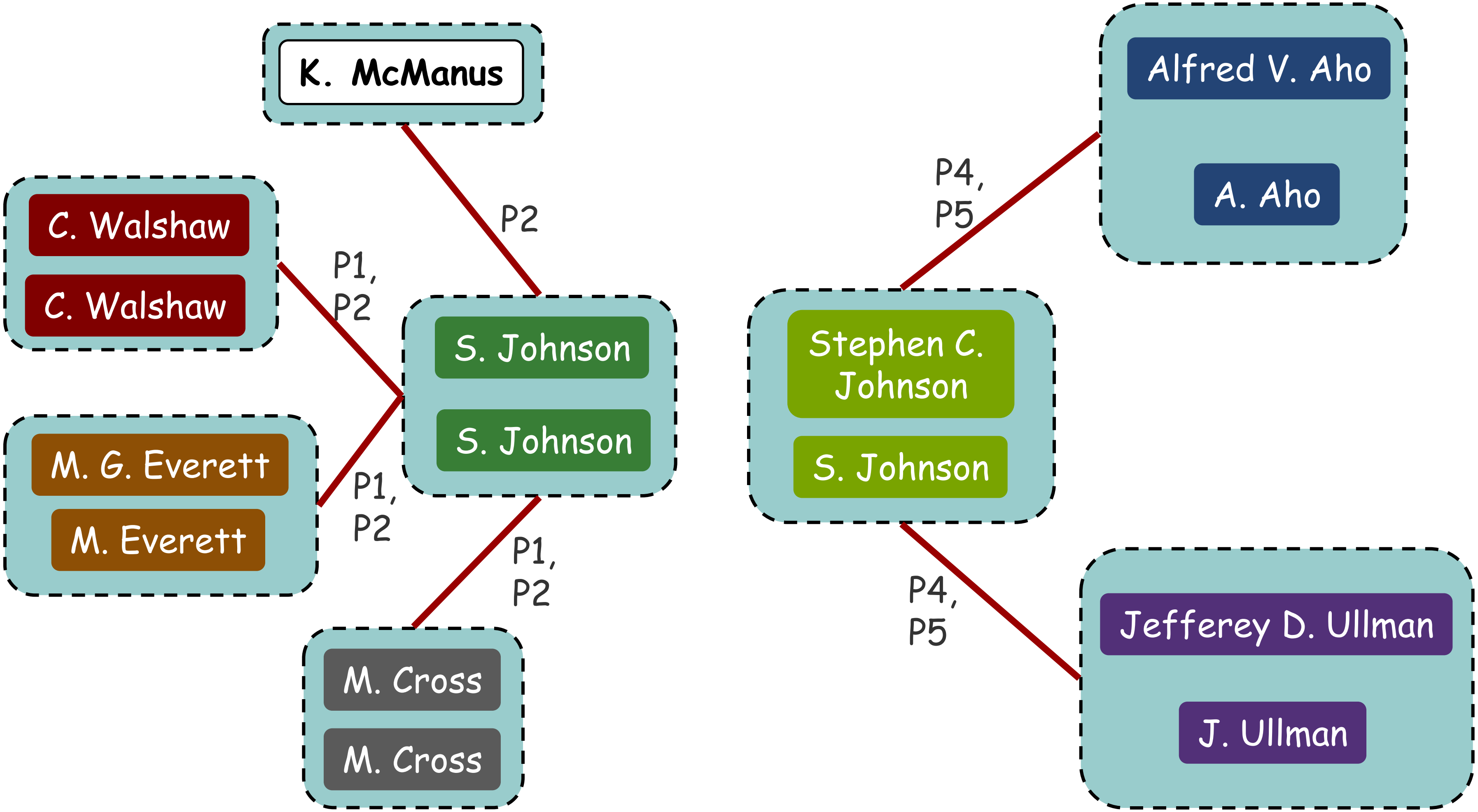
- Use best available measure for each attribute
  - Name Strings: *Soft TF-IDF, Levenstein, Jaro*
  - Textual Attributes: *TF-IDF*
- Aggregate to find similarity between clusters
  - Single link, Average link, Complete link
  - Cluster representative

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)
  - **Probabilistic Model (LDA-ER)**
    - *SIAM SDM'06, Best Paper Award*
  - 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

Parallel Processing Research Group

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

Bell Labs Group

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

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

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

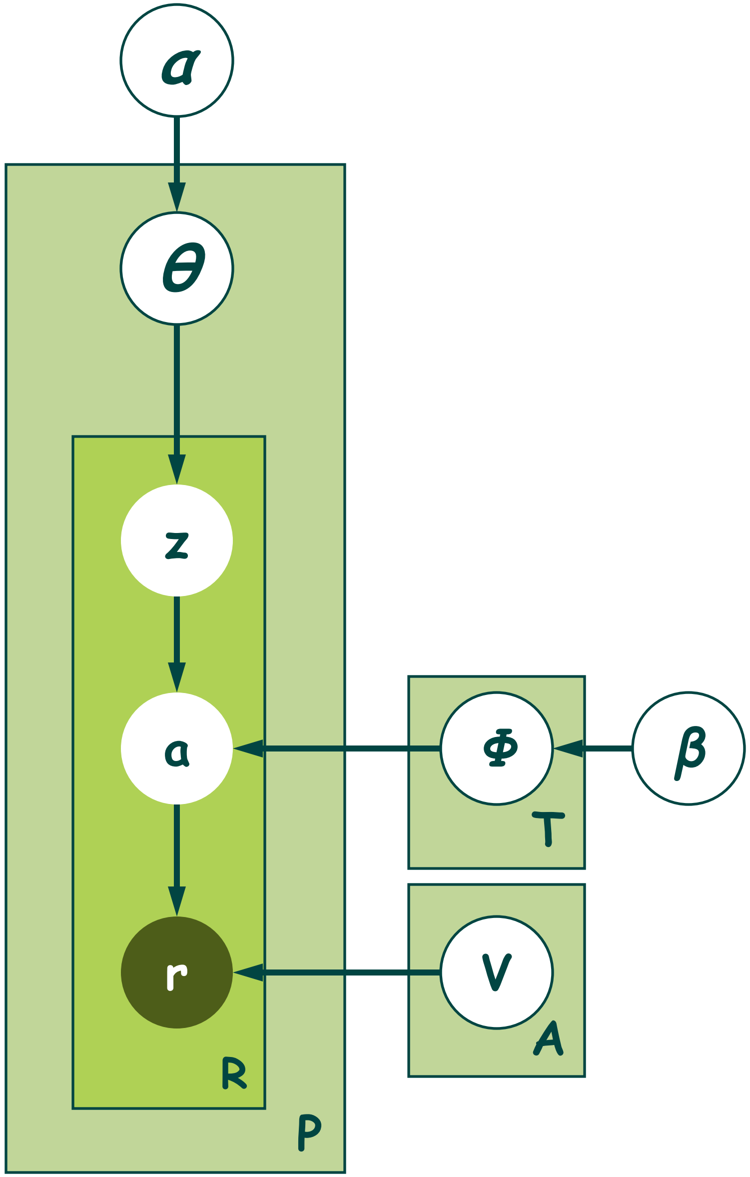
P3: C. Walshaw, M. Cross, M. G. Everett

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

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

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

Latent Dirichlet Allocation ER

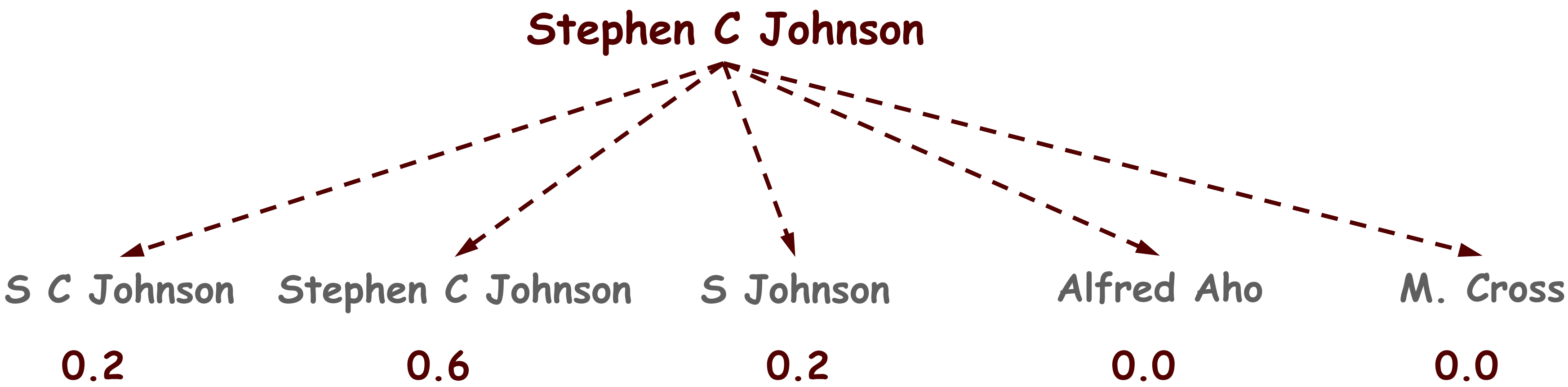


- Entity label ***a*** and group label ***z*** for each reference ***r***
- ***Θ*:** 'mixture' of groups for each co-occurrence
- $\boldsymbol{\Phi_z}$**:** multinomial for choosing entity ***a*** for each group ***z***
- $\boldsymbol{V_a}$**:** multinomial for choosing reference ***r*** from entity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

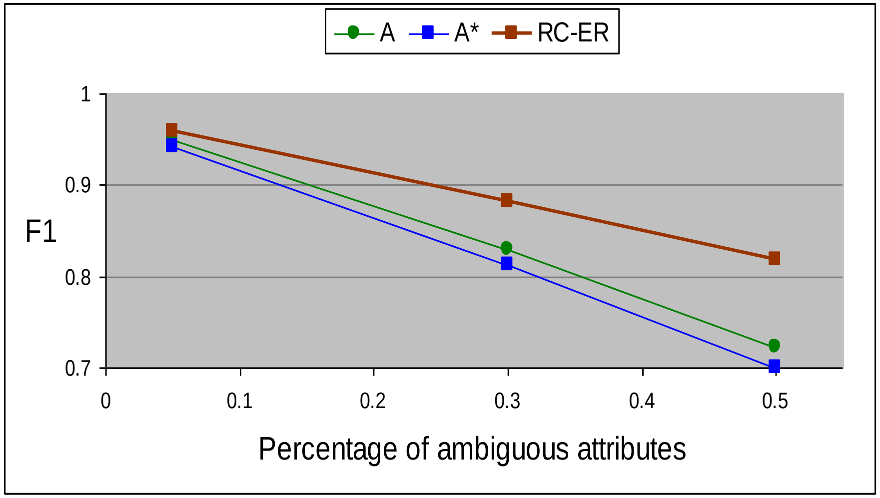
Performance for Specific Names

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

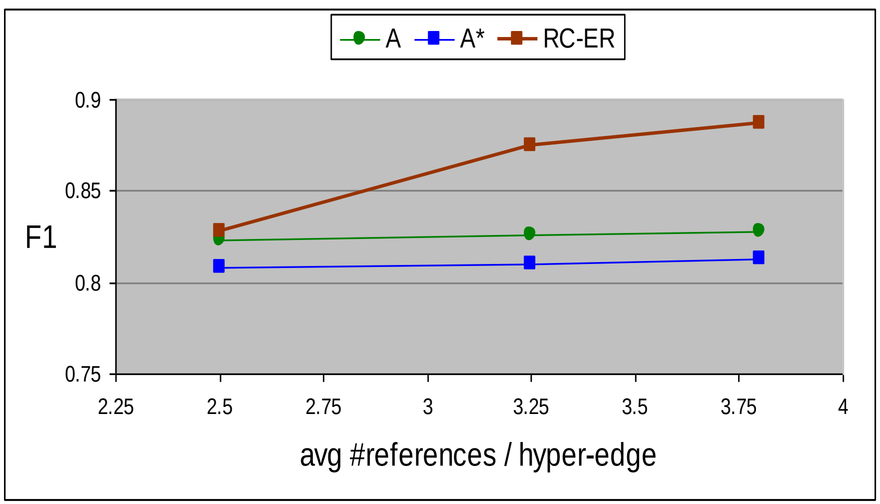
Significantly larger improvements for ‘ambiguous names’

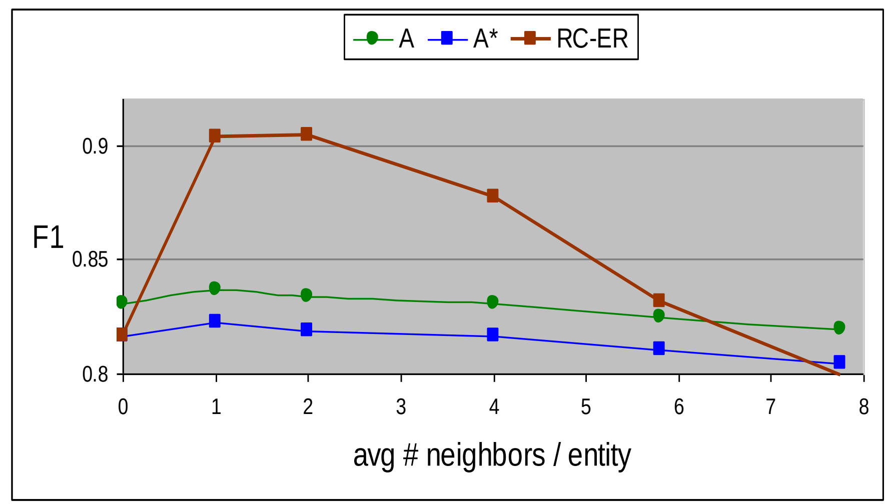
Trends in Synthetic Data



Bigger improvement with

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





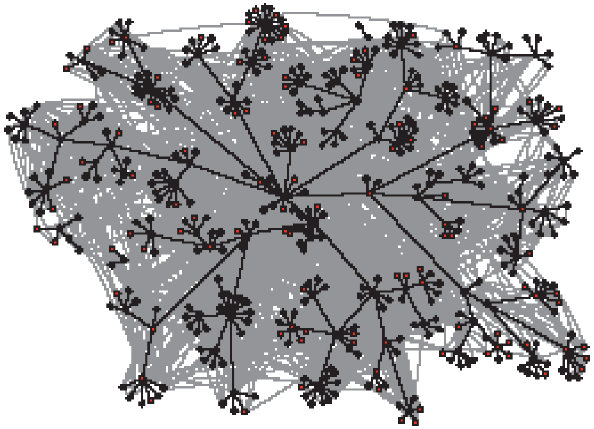
Roadmap

- Intro to Link Mining
  - Link Mining Tasks
  - Link Mining Challenges
- Some Link Mining Algorithms
  - Collective Classification
  - Entity Resolution
  - **Link Prediction**
- Conclusion

Link Prediction

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

Links in Data Graph

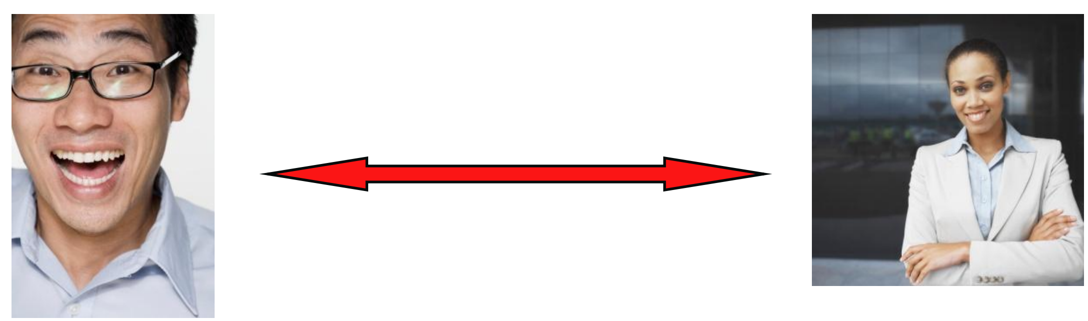


Communications Graph
Nodes: Network References
Edges: Communications Events

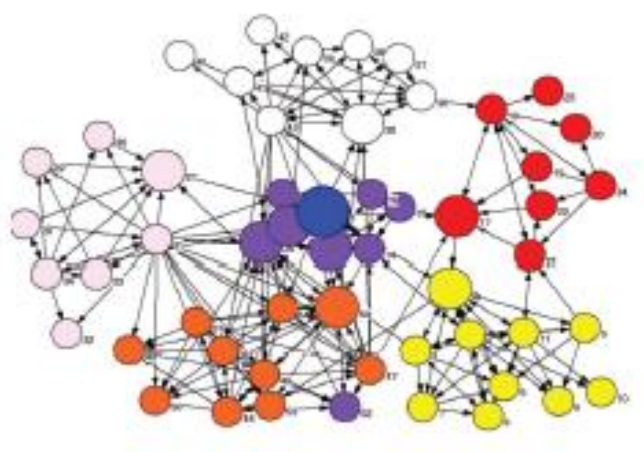
chris@enron.com ↔ Email ↔ liz@enron.com

chris37 ↔ IM ↔ lizs22

555-450-0981 ↔ TXT ↔ 555-901-8812



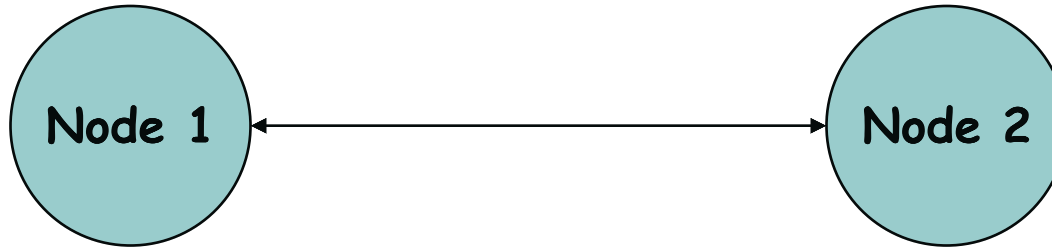
⇒ Links in Information Graph



Network Graph

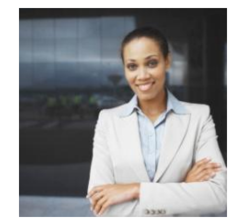
Nodes: Entities

Edges: Social Relationships



Chris

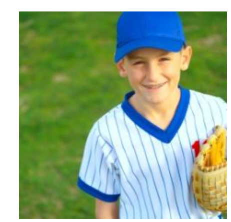
Manager



Elizabeth

Steve

Family



Tim

Predicting Relations

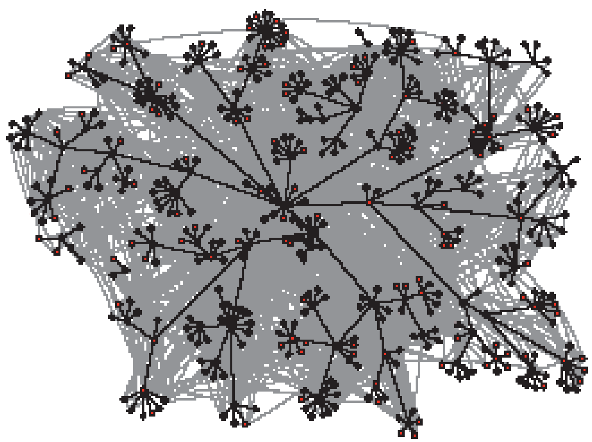
- Link Labeling
  - Can use similar approaches to collective classification
- Link Ranking
  - Many variations
    - Diehl, Namata, Getoor, *Relationship Identification for Social Network Discovery*, AAAI07
  - ‘Leak detection’
    - Carvalho & Cohen, SDM07
- Link Existence
  - HARD!
  - Huge class skew problem
  - Variations: Link completion, find missing link

Roadmap

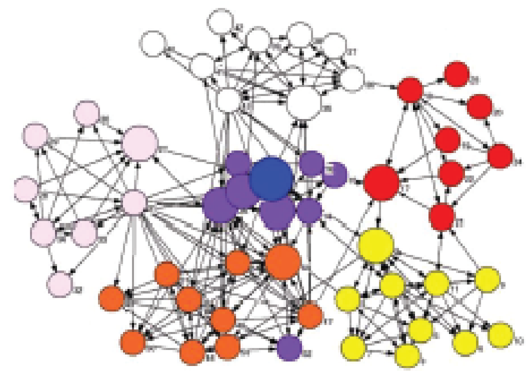
- Intro to Link Mining
  - Link Mining Tasks
  - Link Mining Challenges
- Some Link Mining Algorithms
  - Collective Classification
  - Entity Resolution
  - Link Prediction
- **Conclusion**

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!!

Caveat: Link Mining & Privacy

- Obvious privacy concerns that need to be taken into account!!!

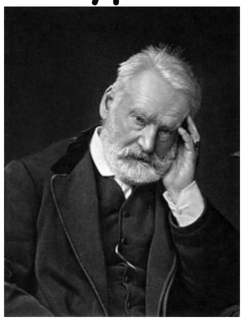
- A better theoretical understanding of when prediction is feasible will also help us understand what must be done to maintain privacy of graph data

- … 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:

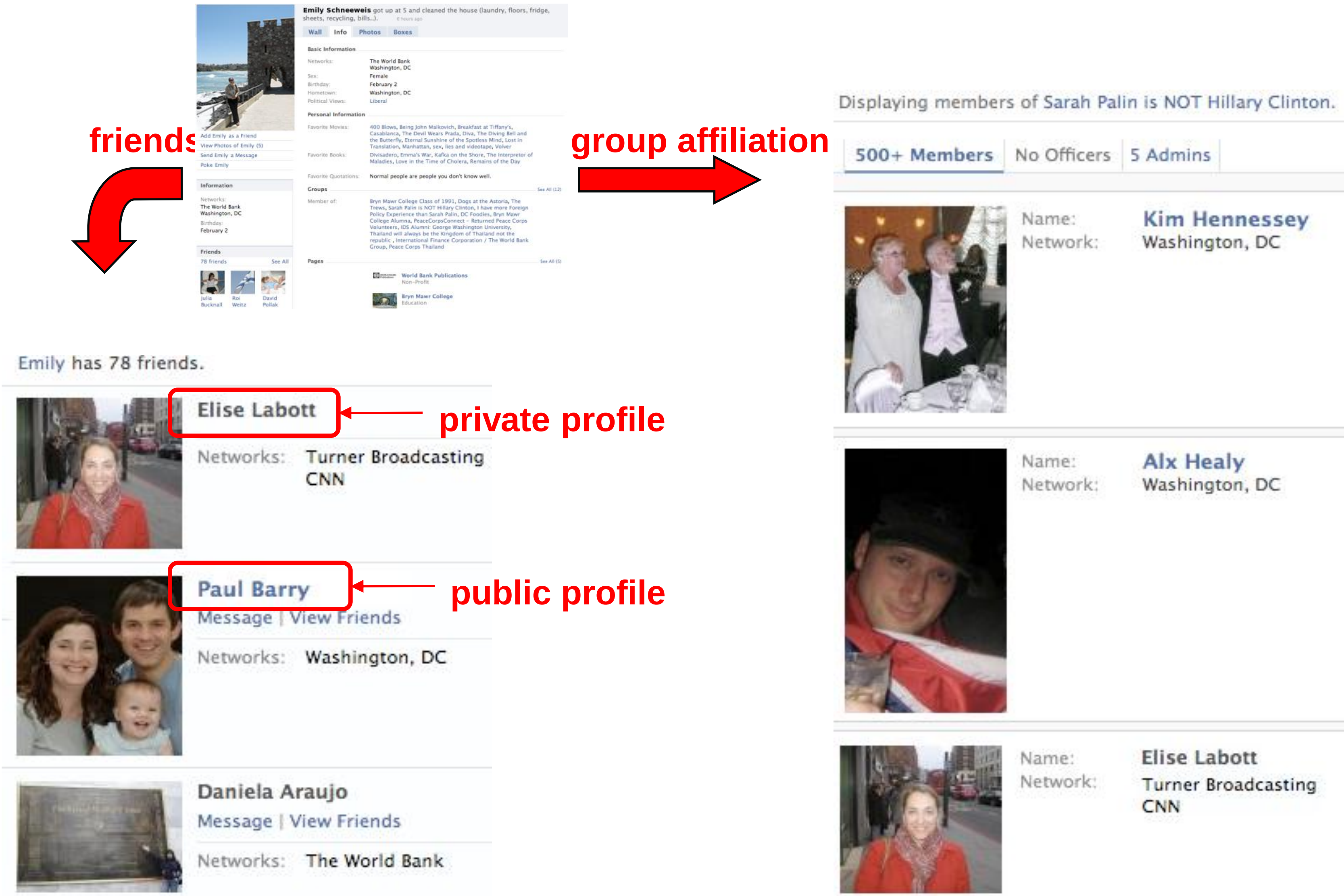
"myrtle beach golf course job listings"

Social network data

friends

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

Attribute disclosure in OSNs



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

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