

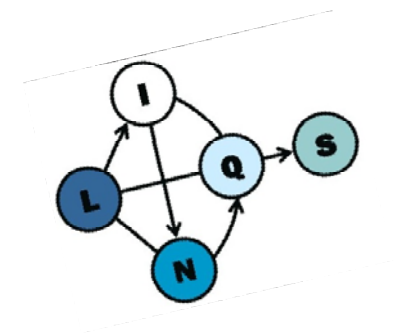
# Graph Identification & Privacy

Lise Getoor

University of Maryland, College Park

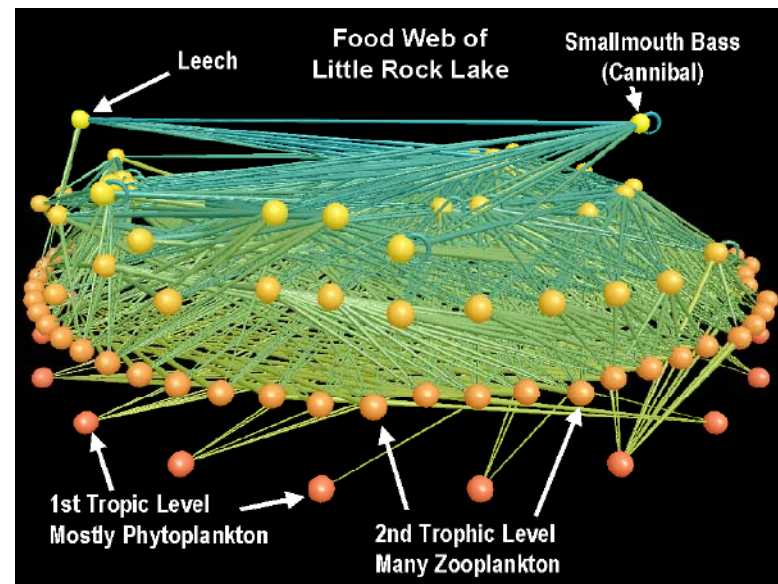
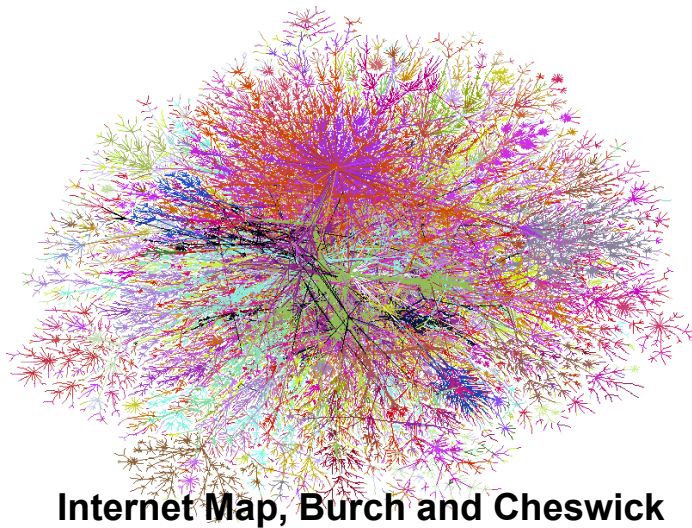


Stanford InfoSeminar  
January 16, 2009



# Graphs and Networks everywhere...

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



Food Web, Martinez et al.

# Wealth of Data

- Inundated with data describing networks
- But much of the data is noisy and incomplete and at **WRONG** level of abstraction for analysis

# Identification

- On the other hand, the data can be joined and sensitive information can be inferred

# Privacy

# Overview: Identification

- Many real world datasets are relational in nature
  - Social Networks – people related to each other by relationships like friendship, family, enemy, boss\_of, etc.
  - Biological Networks – proteins are related to each other based on if they physically interact
  - Communication Networks – email addresses related by who emailed whom
  - Citation Networks – papers linked by which other papers they cite, as well as who the authors are
- However, the observations describing the data are noisy and incomplete
- **graph identification problem** is to **infer** the appropriate **information graph** from the **data graph**

# Example: Organizational Hierarchy

Ideally:

- Know who are the criminals
- Know where the criminals stand in the organization
- Know friends and social groups belong to



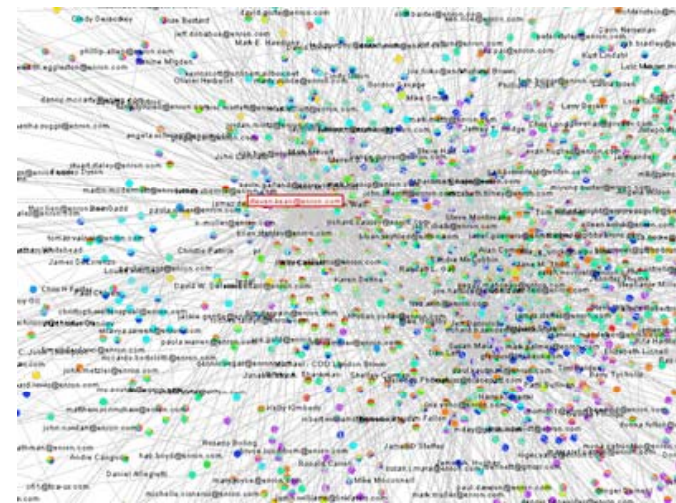
Enron Investigators

In Reality:

- Annotated only a handful of individuals
- Don't have social structure, have an email communication network which reflects that structure



Information Graph



Data Graph





# Example: Protein Interaction Network

**Ideally:**

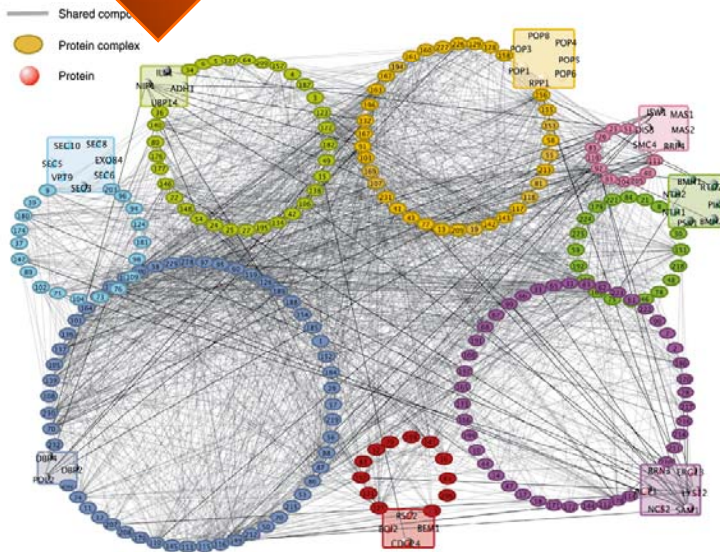
- Know which proteins interact
- Know functions of proteins
- Known complexes of proteins



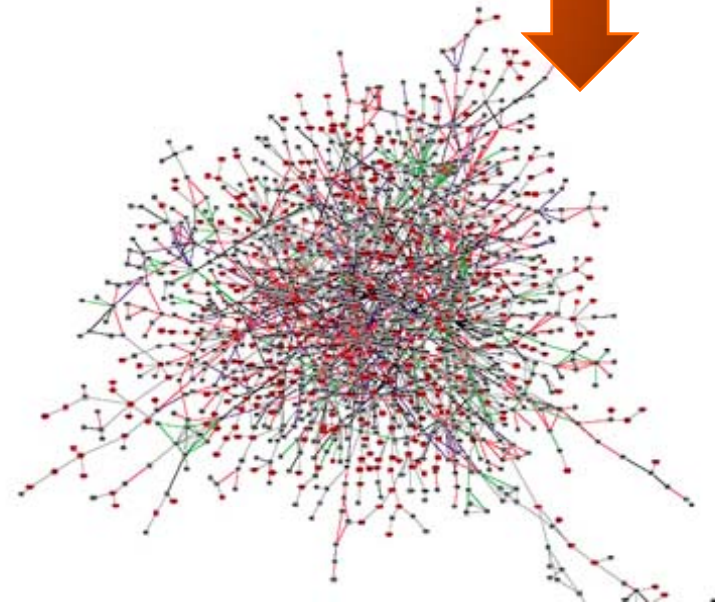
**In Reality:**

- Accurate and Complete Information is expensive
- Available information is noisy and incomplete (i.e., high throughput)

**Network Research Group**



**Information Graph**



**Data Graph**

# Example: Internet Security

**Ideally:**

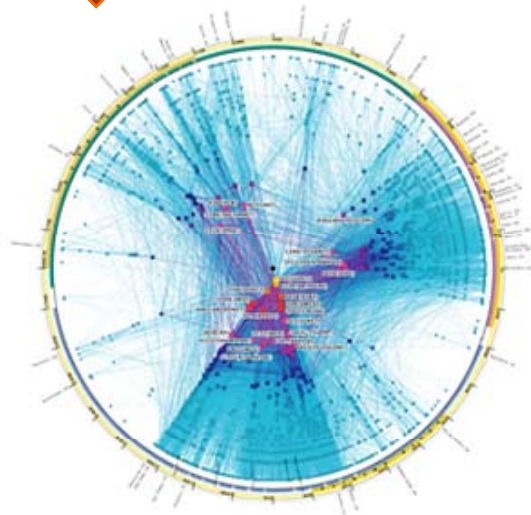
- Know the network from an AS and ISP level
- Know which computers are malicious and launching a DDOS attack



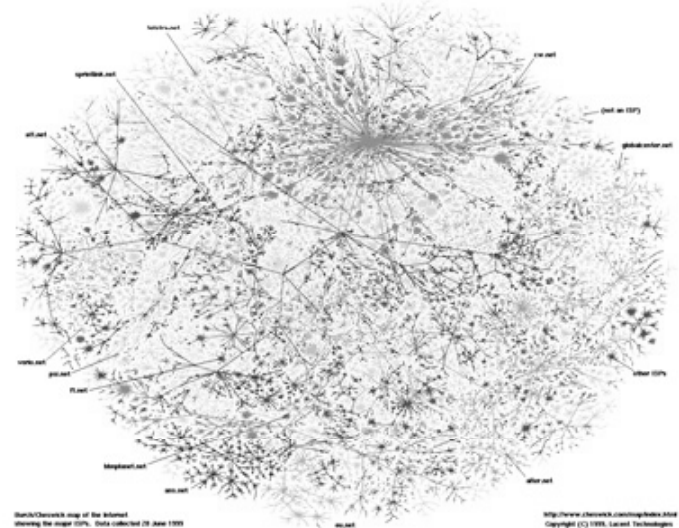
**Network Operator**

**In Reality:**

- Only have trace route information at IP address level
- Do not know legitimate traffic vs. malicious traffic



**Information Graph**

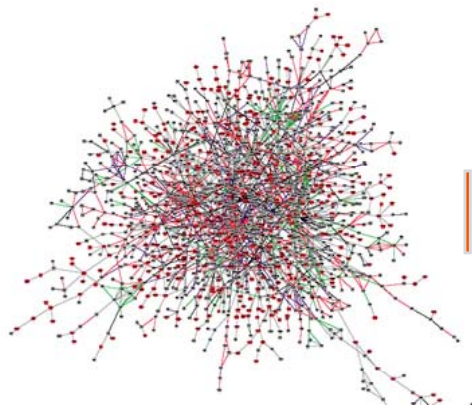


**Data Graph**

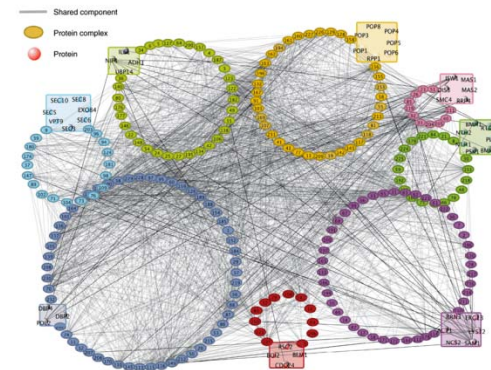
# Solution

## o Graph Identification:

- Infer the information graph that we want from the data graph that we have
- Key Assumption:
  - Dependencies exist such that knowledge of the nodes, edges, and attributes of the data graph can help us infer the nodes, edges, and attributes of the information graph



**Data Graph**

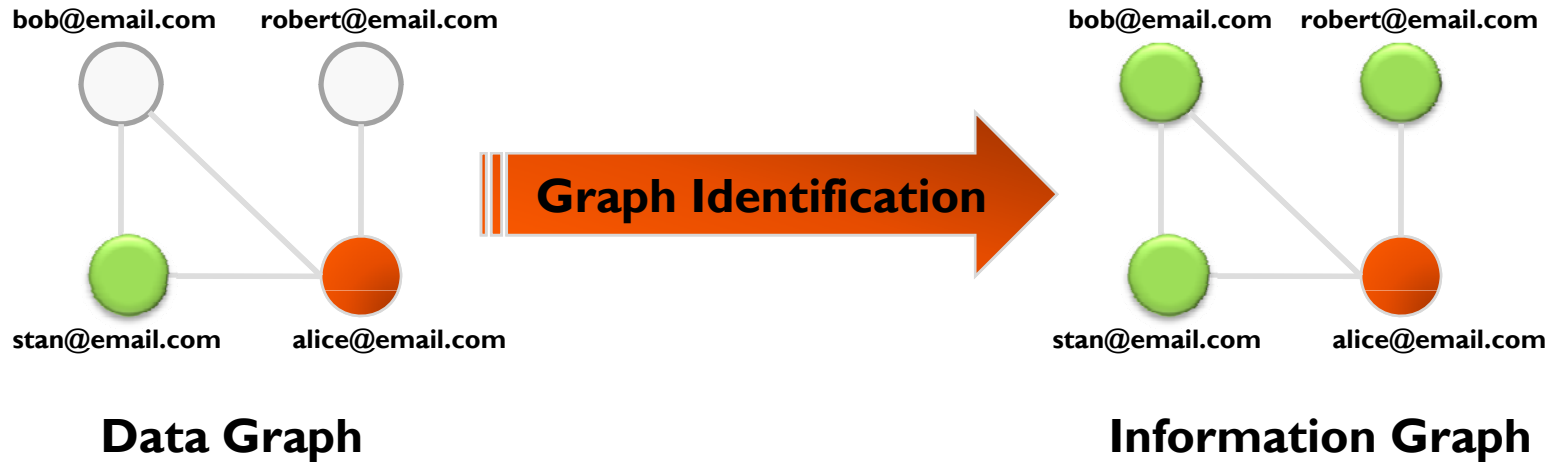


**Information Graph**



# Collective Classification

**Collective Classification (CC):** Given a set of labels (orange and green), label the objects whose label is unknown with the correct label

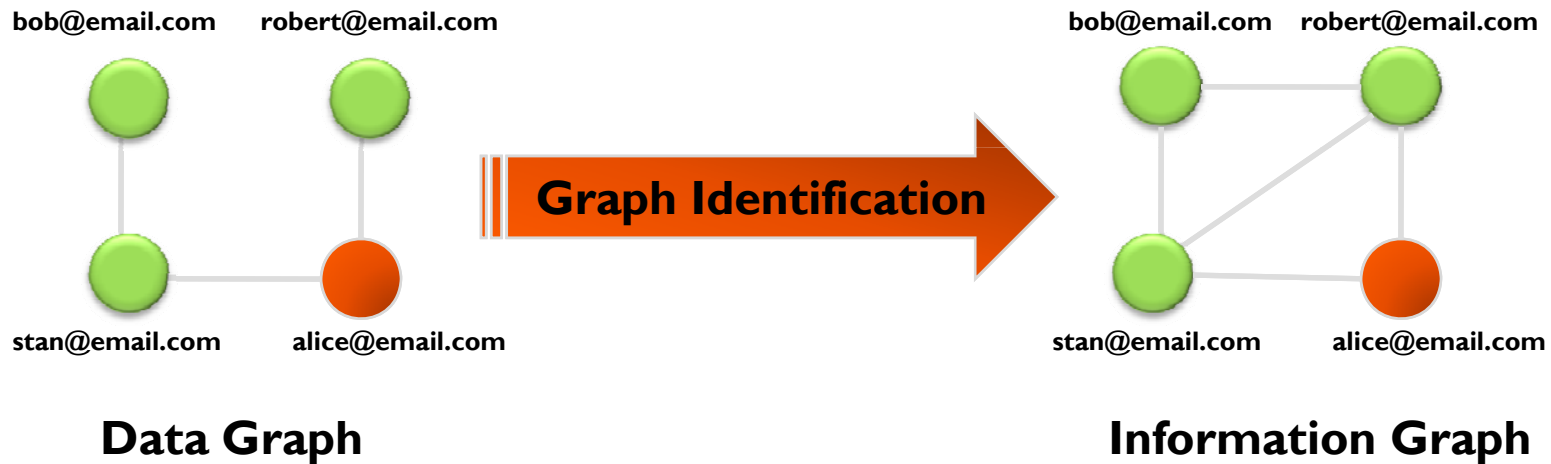


**Assumptions:**

- Set of nodes and edges in data and information graphs are the same
- Inference depends on known labels and attributes of the nodes and edges

# Link Prediction

**Link Prediction (LP):** Predict the existence of edges

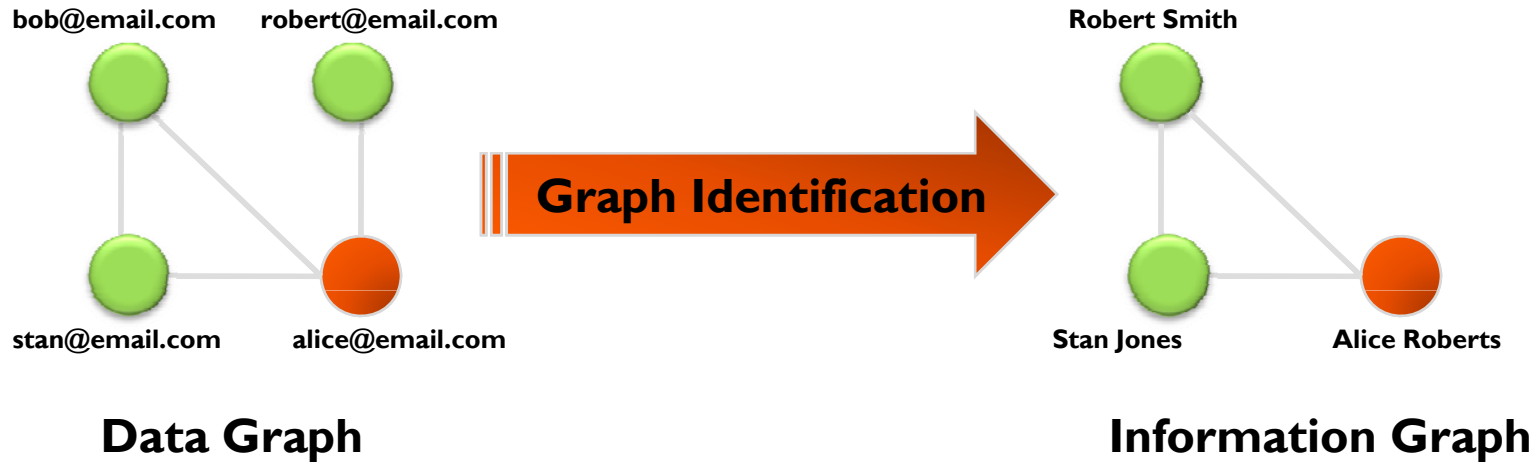


**Assumptions:**

- Set of nodes and attributes in data and information graphs are the same
- Inference depends on known labels and attributes of the nodes and edges

# Entity Resolution

**Entity Resolution (ER):** Identify the the underlying entity represented by the references

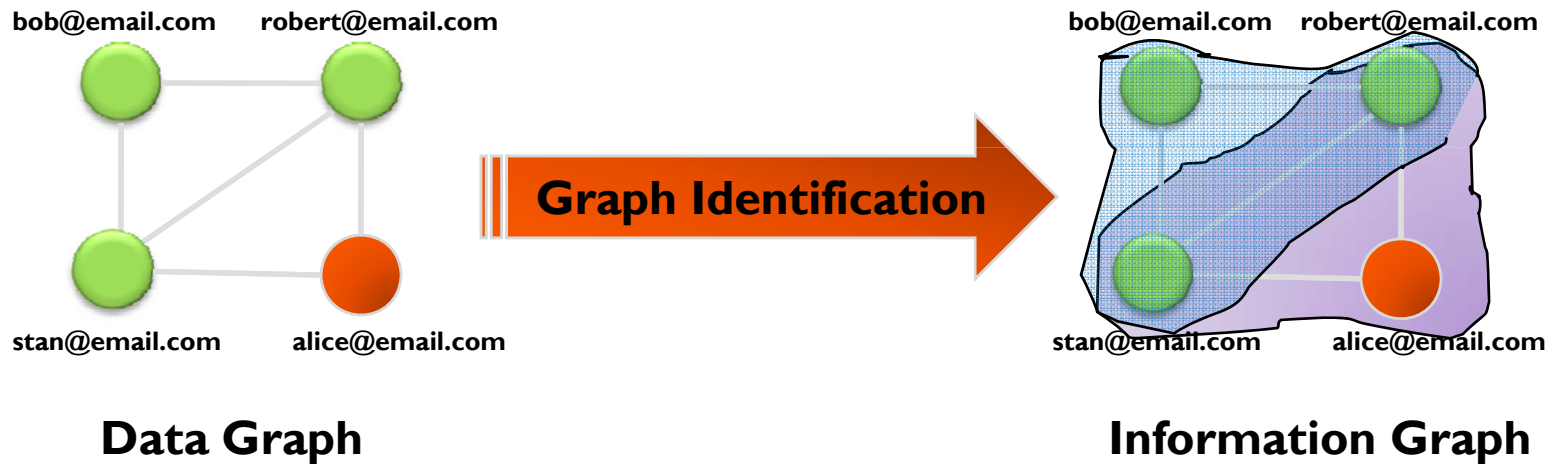


**Assumptions:**

- Edges and attributes of entities based on the edges and attributes of the merged references (if known)
- Inference only depends on known labels, nodes, and edges

# Group Detection

**Group Detection (GD):** Detect the underlying group(s) that the nodes and edges belong to



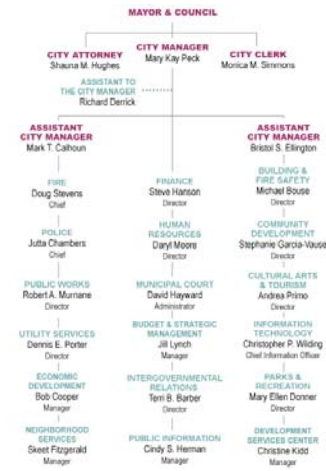
## Assumptions:

- Set of nodes, edges, and attributes in data and information graphs are the same
- Inference only depends on known labels, nodes, and edges

# Inference from Email Communications



Data Graph



Information Graph

- No direct mapping from the nodes, edges, and attributes of data to information graph
- Need to infer existence of all the nodes and edges
- Need to infer the values of attributes based on data graph, as well as the nodes, edges, and other attributes of the information graph



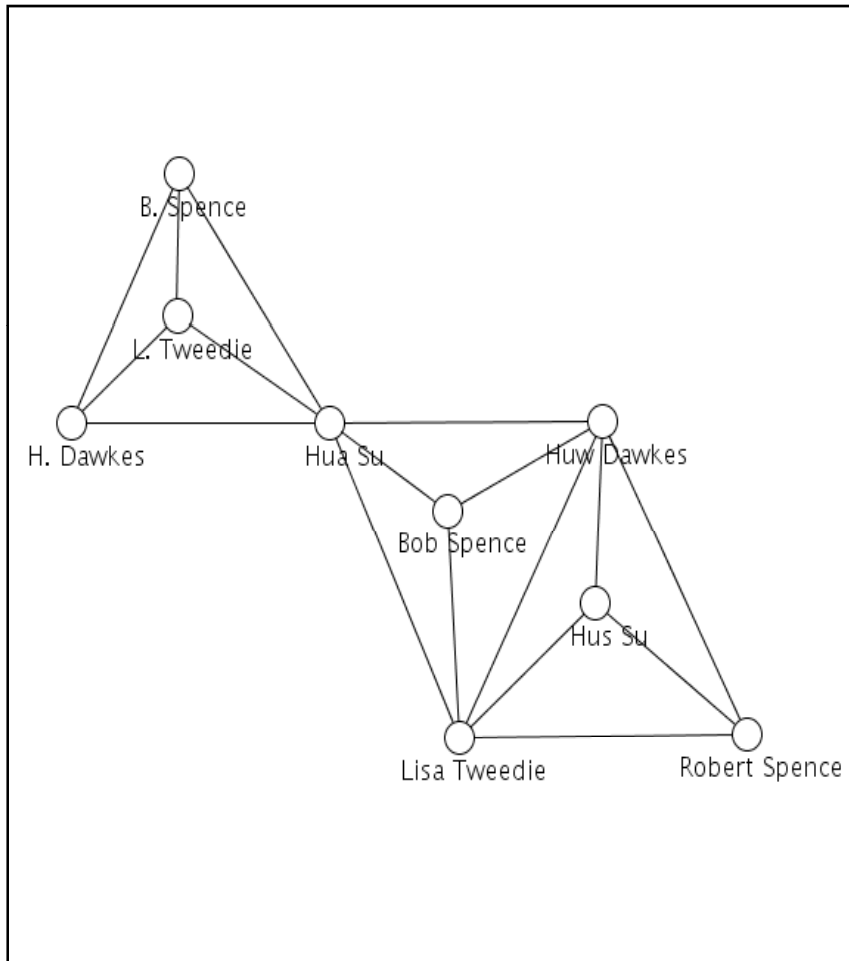
# Entity Resolution

- o **The Problem**

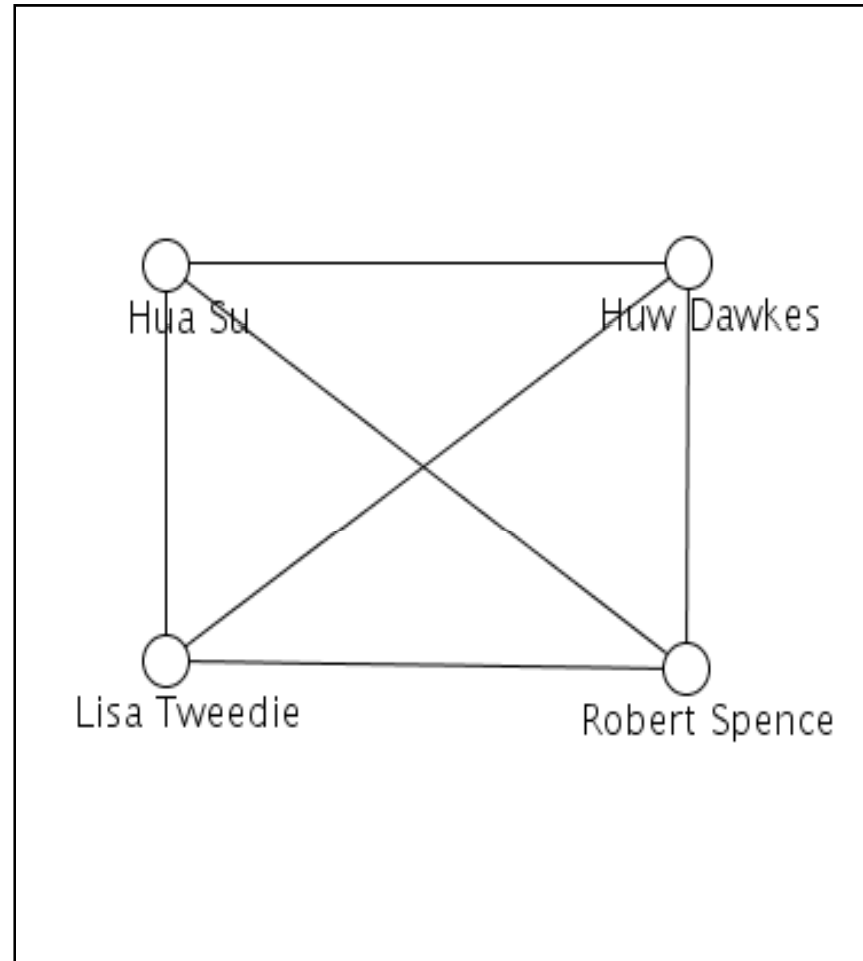
- o Relational Entity Resolution

- o Algorithms

# InfoVis Co-Author Network Fragment

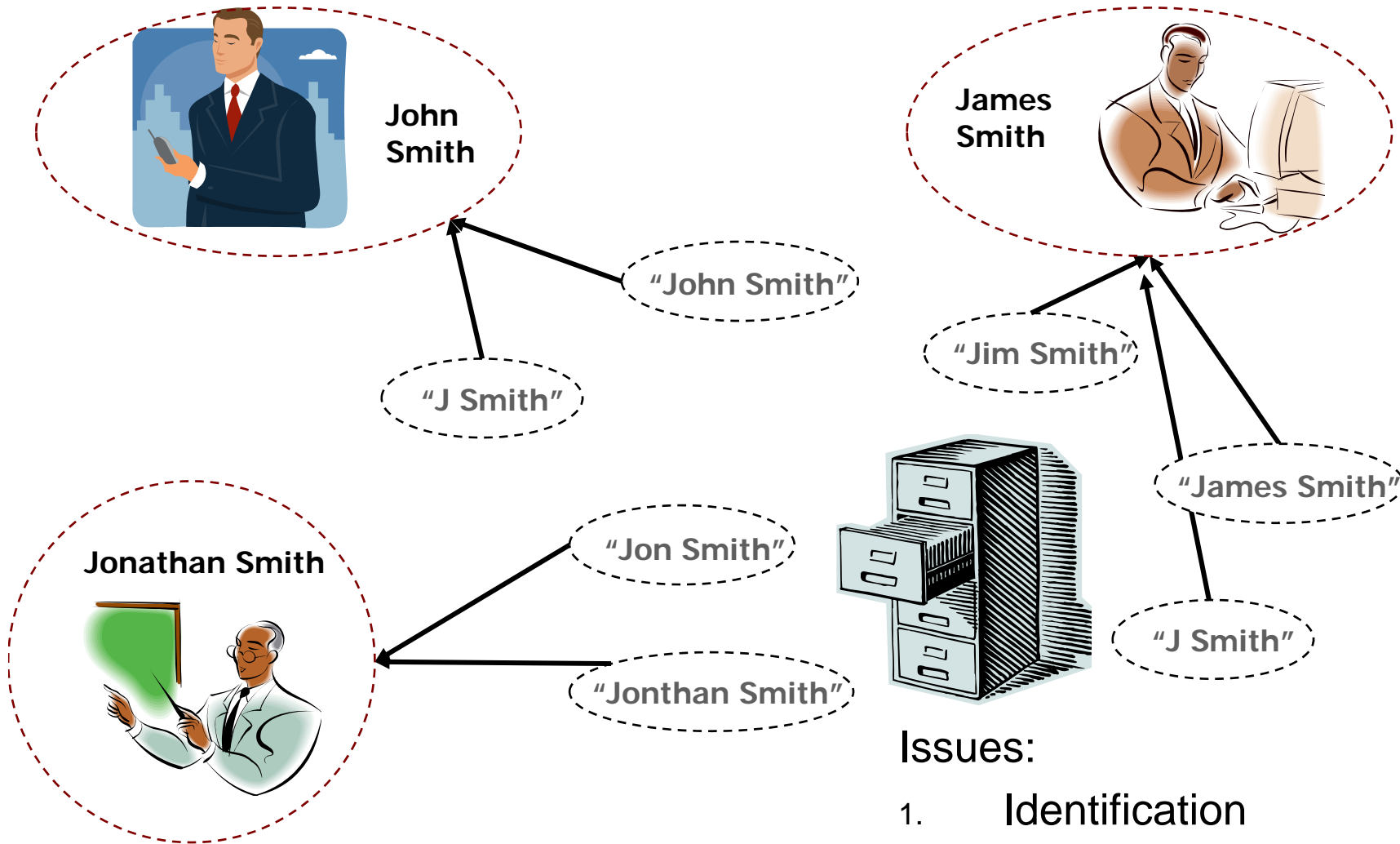


before



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

- o The Problem

- o **Relational Entity Resolution**

- o Algorithms

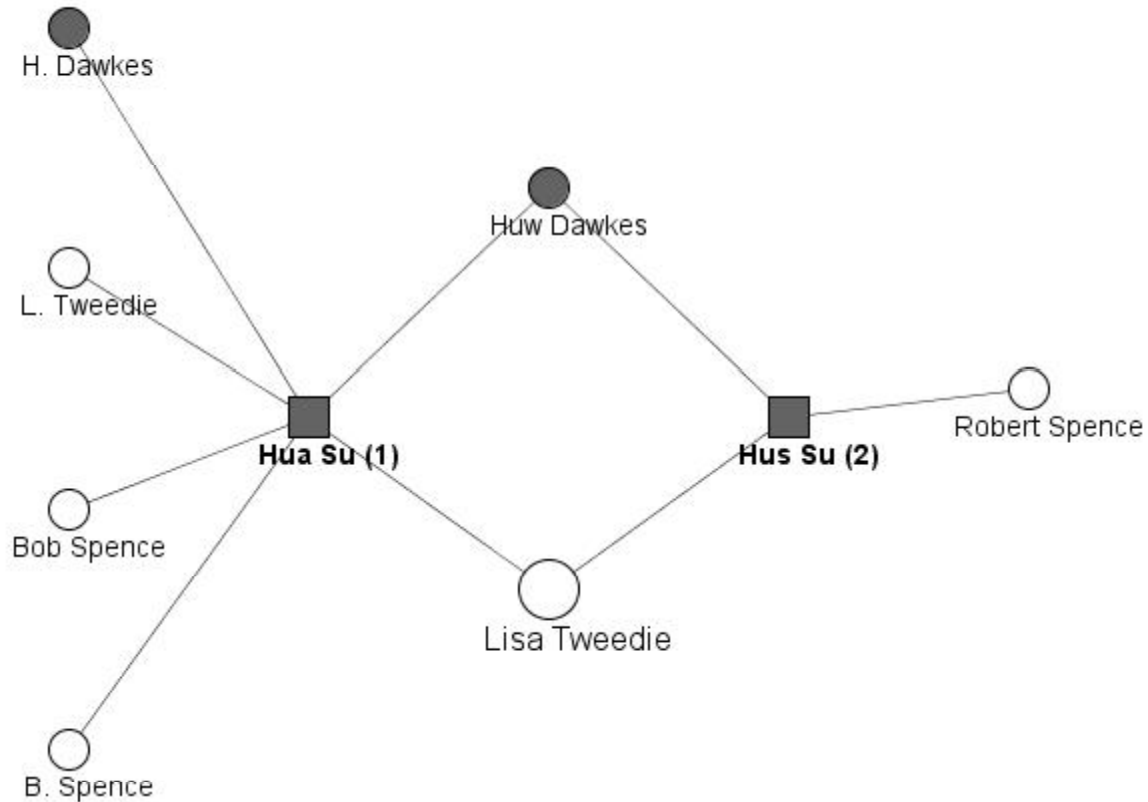


# Relational Entity Resolution

- 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

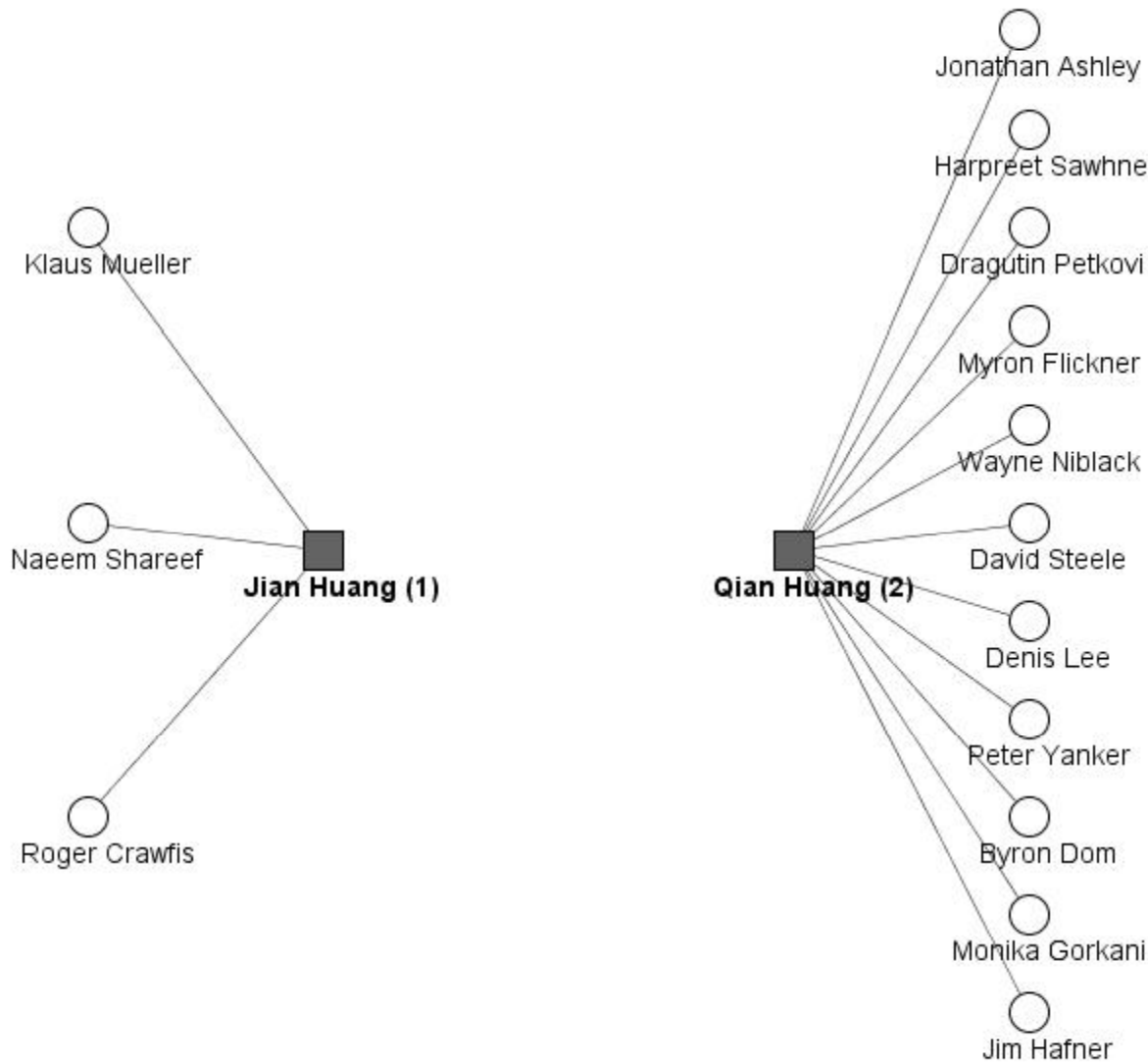
**Pasula et al. 03, Ananthakrishna et al. 02, Bhattacharya & Getoor 04,06,07, McCallum & Wellner 04, Li, Morie & Roth 05, Culotta & McCallum 05, Kalashnikov et al. 05, Chen, Li, & Doan 05, Singla & Domingos 05, Dong et al. 05**

# Relational Identification



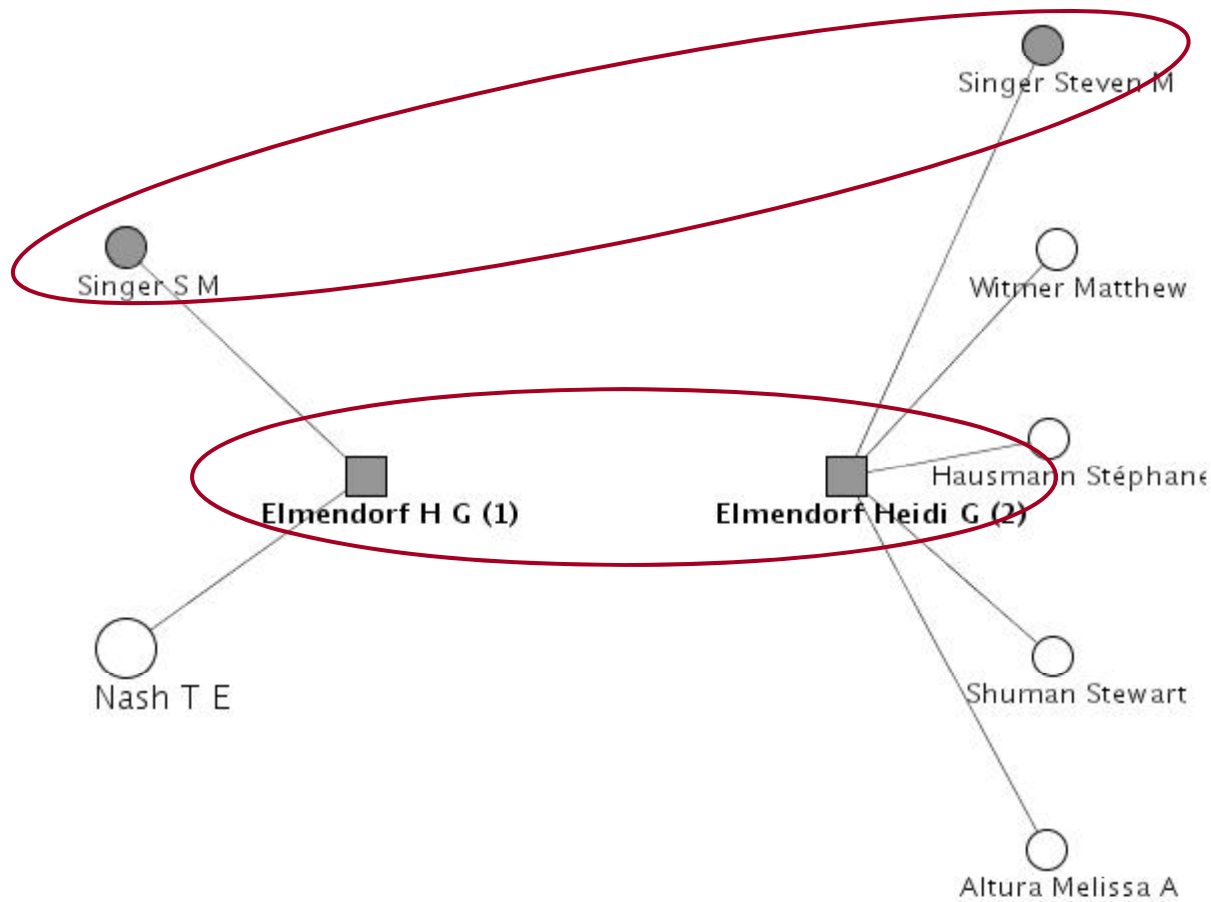
Very similar names.  
Added evidence from  
shared co-authors

# Relational Disambiguation



Very similar names  
but no shared  
collaborators

# Collective Entity Resolution



One resolution provides evidence for another => joint resolution

# Entity Resolution

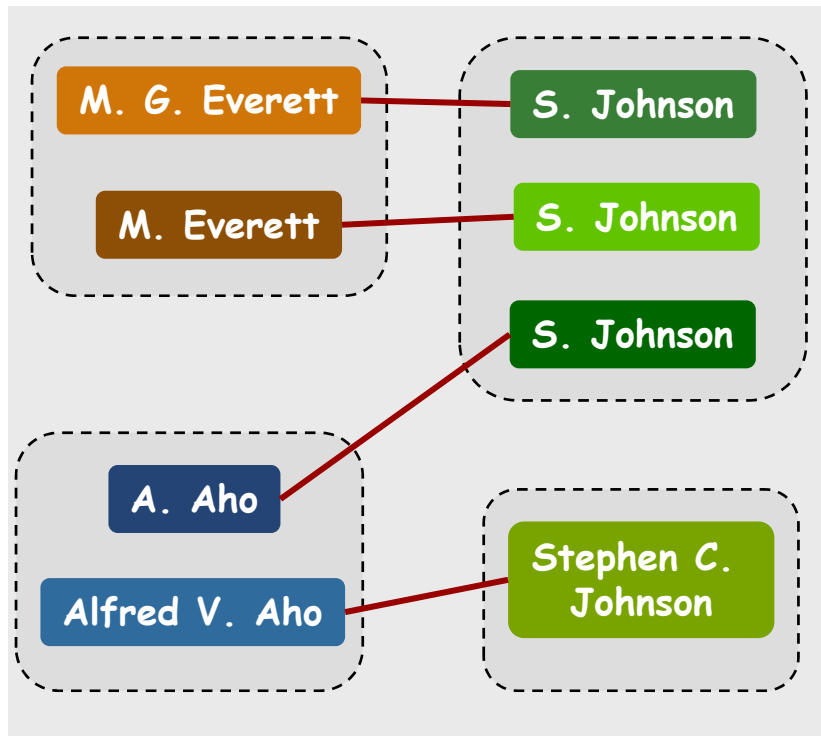
- The Problem
- Relational Entity Resolution
- **Algorithms**
  - Relational Clustering (RC-ER)
  - Probabilistic Model (LDA-ER)
  - Experimental Evaluation



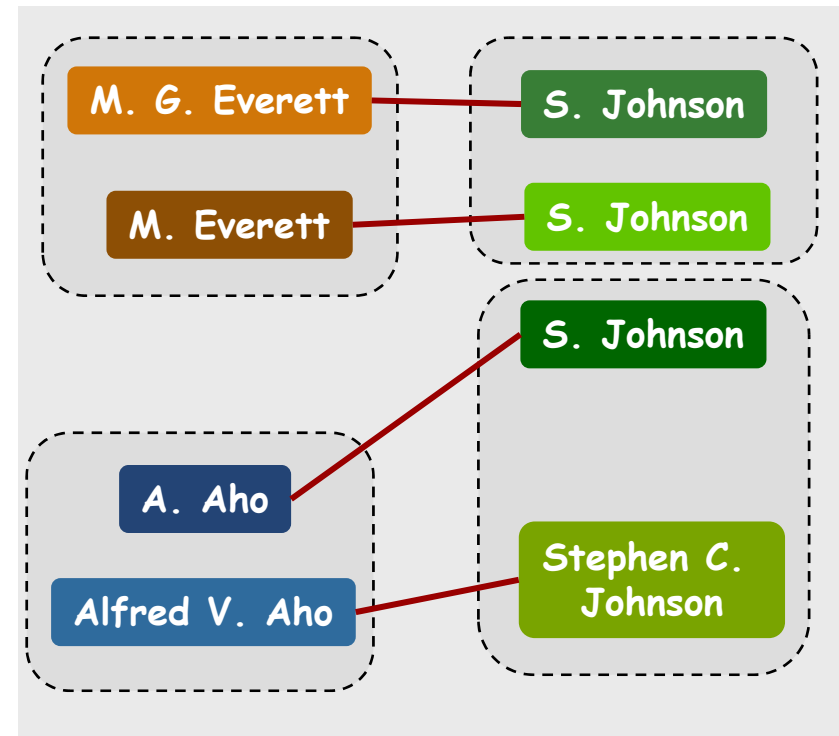
# Entity Resolution

- The Problem
- Relational Entity Resolution
- **Algorithms**
  - **Relational Clustering (RC-ER)**
  - Probabilistic Model (LDA-ER)
  - Experimental Evaluation

# 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

# 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)**
  - 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

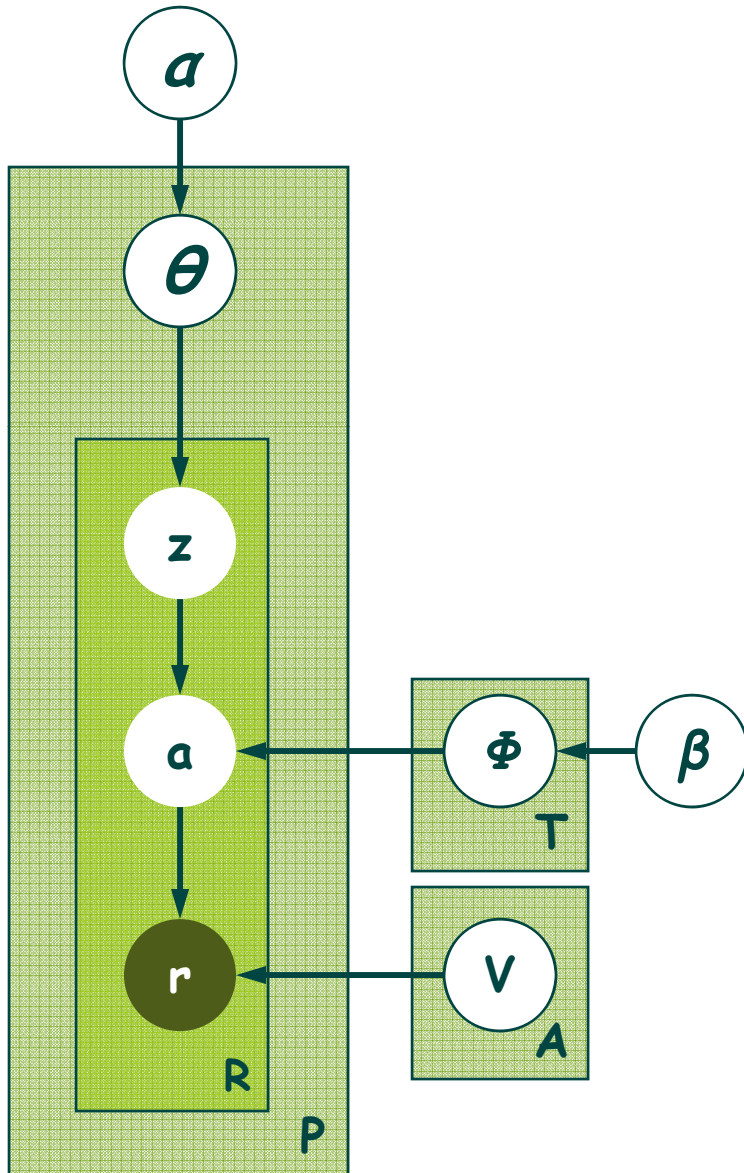


P4: Alfred V. Aho, **Stephen C. Johnson**,  
Jefferey D. Ullman

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  $r$
- $\Theta$ : 'mixture' of groups for each co-occurrence
- $\Phi_z$ : multinomial for choosing entity  $a$  for each group  $z$
- $V_a$ : multinomial for choosing reference  $r$  from entity  $a$
- Dirichlet priors with  $\alpha$  and  $\beta$

# 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	<b>0.995</b>	<b>0.985</b>	<b>0.818</b>
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 Data Mining, 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	<b>0.995</b>	<b>0.985</b>	<b>0.818</b>
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

Flipside....

**Privacy**

# Privacy in social networks

- Identity disclosure
  - Entity resolution
- Attribute disclosure
  - Collective classification
- Link re-identification
  - Link prediction
- Group membership disclosure
  - Group detection

# A public profile on Facebook



Add Emily as a Friend

View Photos of Emily (5)

Send Emily a Message

Poke Emily

## Information

Networks:

The World Bank  
Washington, DC

Birthday:

February 2

## Friends

78 friends

See All



Julia  
Bucknall



Roi  
Weitz



David  
Pollak

**Emily Schneeweis** got up at 5 and cleaned the house (laundry, floors, fridge, sheets, recycling, bills..). 6 hours ago

Wall

Info

Photos

Boxes

## Basic Information

Networks: The World Bank  
Washington, DC  
Sex: Female  
Birthday: February 2  
Hometown: Washington, DC  
Political Views: Liberal

attributes

## Personal Information

Favorite Movies: 400 Blows, Being John Malkovich, Breakfast at Tiffany's, Casablanca, The Devil Wears Prada, Diva, The Diving Bell and the Butterfly, Eternal Sunshine of the Spotless Mind, Lost in Translation, Manhattan, sex, lies and videotape, Volver

Favorite Books: Divisadero, Emma's War, Kafka on the Shore, The Interpreter of Maladies, Love in the Time of Cholera, Remains of the Day

Favorite Quotations: Normal people are people you don't know well.

## Groups

See All (12)

Member of: Bryn Mawr College Class of 1991, Dogs at the Astoria, The Trews, Sarah Palin is NOT Hillary Clinton, I have more Foreign Policy Experience than Sarah Palin, DC Foodies, Bryn Mawr College Alumna, PeaceCorpsConnect - Returned Peace Corps Volunteers, IDS Alumni: George Washington University, Thailand will always be the Kingdom of Thailand not the republic, International Finance Corporation / The World Bank Group, Peace Corps Thailand

groups

## Pages

See All (5)



World Bank Publications  
Non-Profit

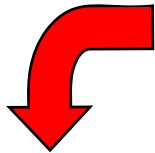


Bryn Mawr College  
Education

friends

# Emily's friends and groups

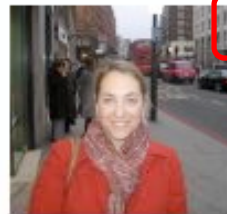
friends



group affiliation



Emily has 78 friends.



**Elise Labott**

private profile

Networks: Turner Broadcasting  
CNN



**Paul Barry**

public profile

Message | View Friends

Networks: Washington, DC



**Daniela Araujo**

Message | View Friends

Networks: The World Bank

Displaying members of Sarah Palin is NOT Hillary Clinton.

500+ Members

No Officers

5 Admins



Name:

**Kim Hennessey**

Network:

Washington, DC



Name:

**Alex Healy**

Network:

Washington, DC



Name:

**Elise Labott**

Network:

Turner Broadcasting  
CNN





# Identity disclosure

- Occurs when the adversary is able to determine the mapping from a record to a specific individual
- Privacy literature has concentrated on structural identification

bob davis  Profile Search | Friend Finder

Show results from **Washington, DC**

Displaying 1 - 10 out of over 500 people results at Washington, DC for: **bob davis**

	Name: <b>Robert Davis</b> Networks: Mary Washington Washington, DC	<a href="#">Add as Friend</a> <a href="#">Send a Message</a> <a href="#">View Friends</a>
	Name: <b>Robert Davis</b> Network: Washington, DC	<a href="#">Add as Friend</a> <a href="#">Send a Message</a>
	Name: <b>Robert Davis</b> Network: Washington, DC	<a href="#">Add as Friend</a> <a href="#">Send a Message</a> <a href="#">View Friends</a>
	Name: <b>Bob Davis</b> Network: Washington, DC	<a href="#">Add as Friend</a> <a href="#">Send a Message</a>

# Attribute disclosure

- Occurs when an adversary is able to determine the value of a user attribute that the user intended to stay private
  - Example: is someone liberal?

## private profile




**Elise Labott**

---

Networks: [Turner Broadcasting](#)  
[CNN](#)

## public profile



**Paul Barry**

[Message](#) | [View Friends](#)

---

Networks: [Washington, DC](#)



- [Add Paul as a Friend](#)
- [View Photos of Paul \(1\)](#)
- [Send Paul a Message](#)
- [Poke Paul](#)

## Paul Barry

[Wall](#) [Info](#) [Photos](#) [Boxes](#)

### Basic Information

Networks: [Washington, DC](#)  
Sex: [Male](#)  
Birthday: [June 2](#)  
Relationship Status: [Married to Kacie Goddard](#)

### Education and Work

College: [American '92 International Studies](#)  
High School: [Cooper High School '85](#)  
Employer: [WashingtonPost.Newsweek](#)  
Position: [Network Engineer](#)  
Time Period: [March 1999 - Present](#)

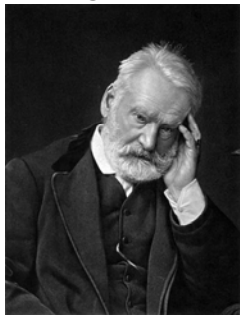


# Link re-identification

- Occurs when an adversary is able to infer that two entities participate in a particular type of sensitive relationship or communication

## Disease data

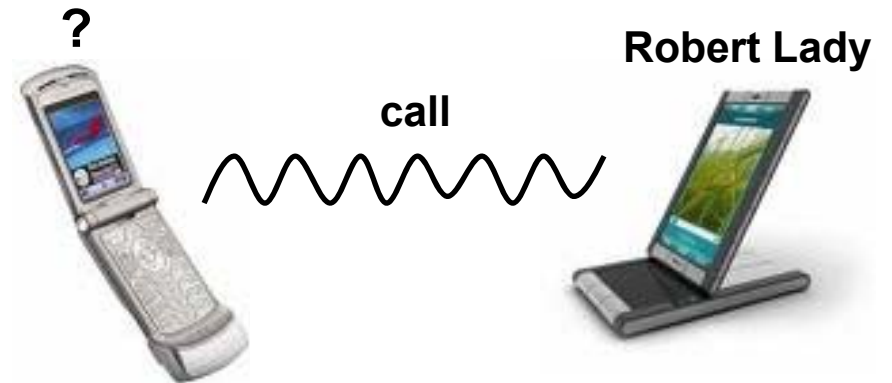
has hypertension



father-of



## Communication data



## Search data

Query 1:

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

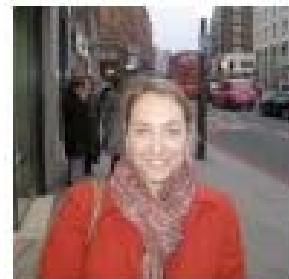
same-user

Query 2:

“myrtle beach golf course job listings”

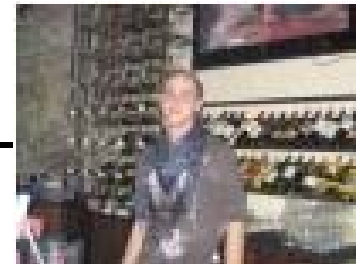
## Social network data

Elise Labott



friends

Robert Davis



# Group membership disclosure

- Occurs when an adversary is able to infer that a person affiliates with a group relevant to the classification of a sensitive attribute.
  - Example: is she liberal?

private profile



**Elise Labott**  
Networks: Turner Broadcasting  
CNN

group affiliation?



Displaying members of Sarah Palin is NOT Hillary Clinton.

500+ Members No Officers 5 Admins



Name: **Kim Hennessey**  
Network: Washington, DC

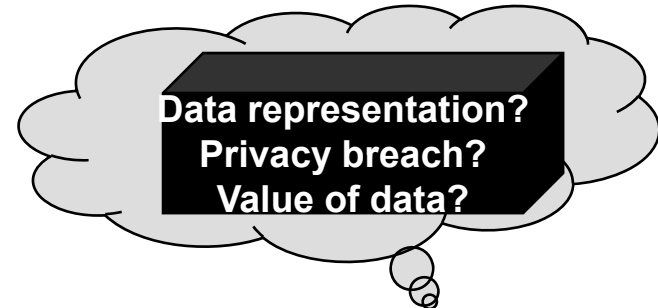
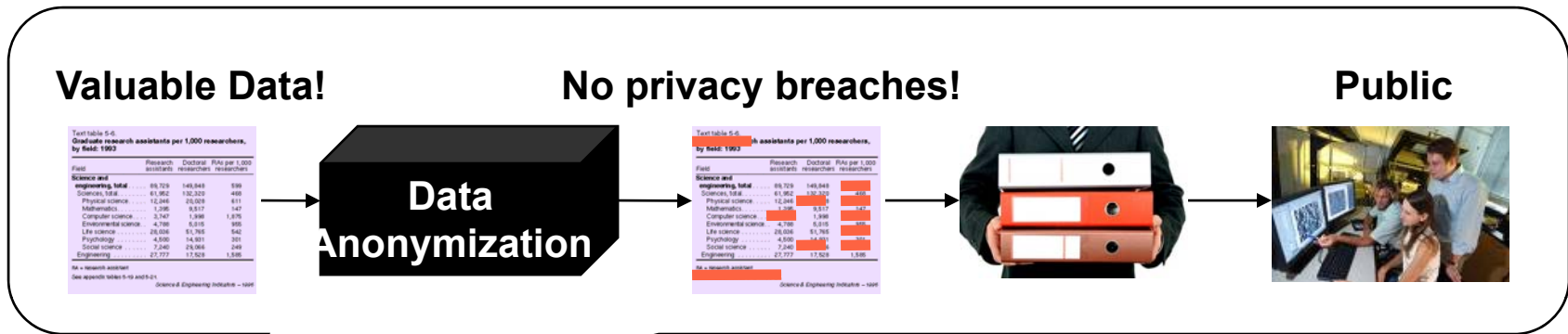


Name: **AIX Healy**  
Network: Washington, DC



Name: **Elise Labott**  
Network: Turner Broadcasting  
CNN

# Anonymization Process



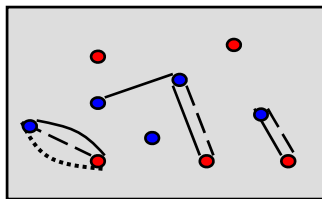
# Anonymizing nodes

Ana	21	F	20740
Bob	25	M	83201
Chris	24	M	20742
Don	29	M	83209
Emma	28	F	83230
Fabio	31	M	83222
Gia	24	F	20640
Halle	29	F	83201
Ian	23	M	20760
John	24	M	20740

*5-anonymity*  
 →  
*applied to nodes*

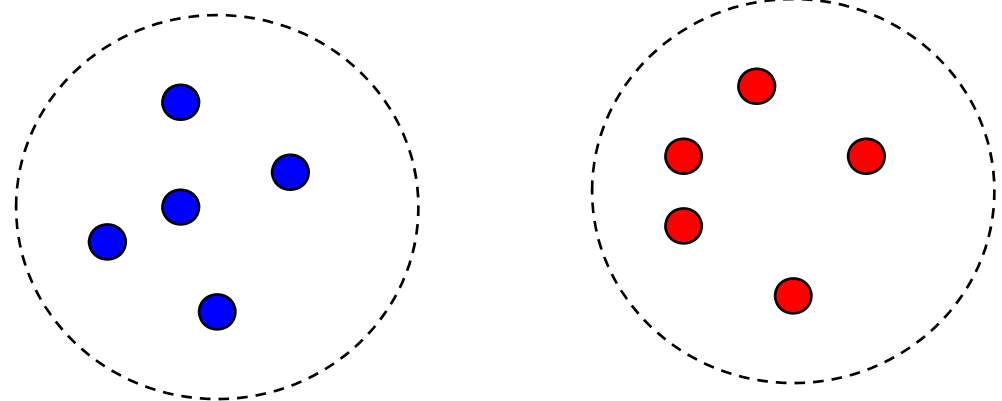
●	< 25	*	20***
●	≥ 25	*	832**
●	< 25	*	20***
●	≥ 25	*	832**
●	≥ 25	*	832**
●	≥ 25	*	832**
●	< 25	*	20***
●	≥ 25	*	832**
●	< 25	*	20***
●	< 25	*	20***

original data graph



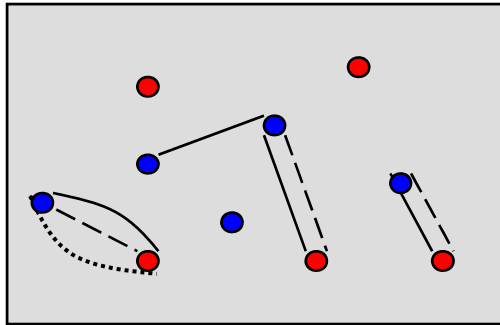
*Equivalence*  
 →  
*classes*

anonymized data graph

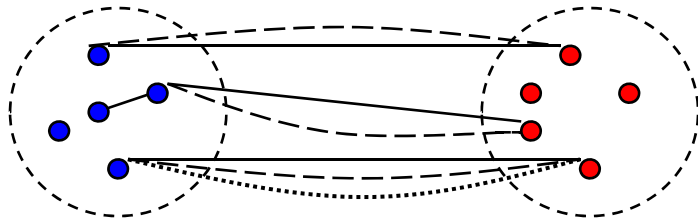


# Anonymizing links

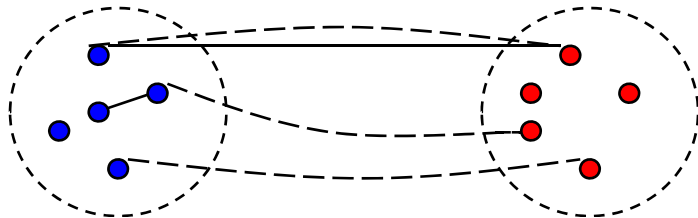
**original graph**



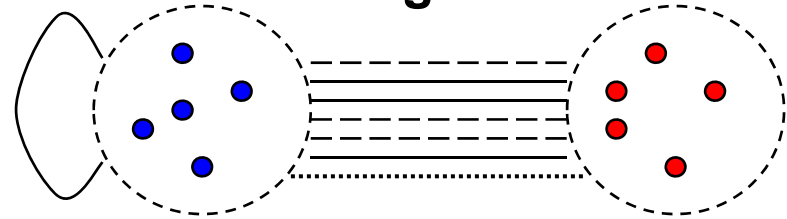
**intact links**



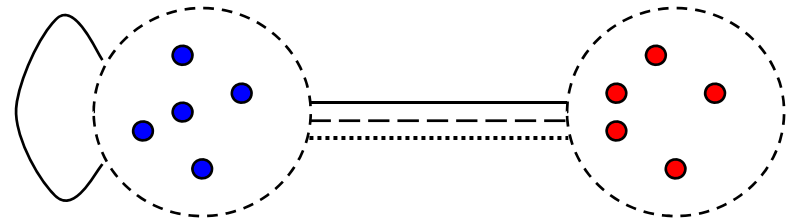
**partial link removal**



**cluster-edge method**



**constrained cluster-edge method**



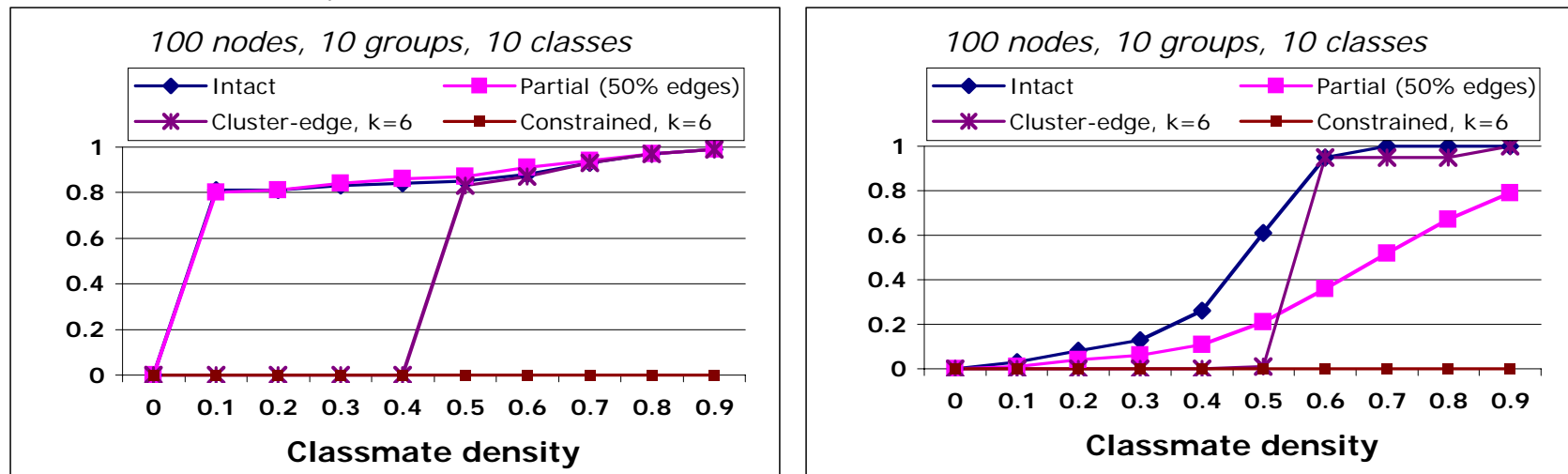
**all links removed**



# Link re-identification results

- Synthetic dataset of students
  - Class enrollment and research group information
  - Observed links - classmates and groupmates
  - Sensitive link – friends
- Anonymize the data using the proposed methods
- Compute the existence prob of sensitive edge using a Noisy-Or model

**Prediction precision and recall rates at various classmate densities**



Reference: E. Zheleva, L. Getoor. Preserving the privacy of sensitive relationships in graph data. PinKDD 2007.

# Attribute disclosure

- o In the context of online social networks

**Emily Schneeweis** got up at 5 and cleaned the house (laundry, floors, fridge, sheets, recycling, bills...) 4 hours ago

**Basic Information**

Networks: The World Bank  
Washington, DC

Sex: Female

Birthday: February 2

Home town: Washington, DC

Political views: Liberal

**Personal Information**

**Favorite Movies:** 400 Blows, Being John Malkovich, Breakfast at Tiffany's, Casablanca, The Devil Wears Prada, Drive, The Dining Bell and the Butterfly, Eternal Sunshine of the Spotless Mind, Lost in Translation, Manhattan, one, two and indelible, Volver

**Favorite Books:** Disasters, Emma's War, Kafka on the Shore, The Intergator of Maladies, Love in the Time of Cholera, Remains of the Day

**Favorite Quotations:** Normal people are people you don't know well.

**Groups** See All (12)

Member of:

- Bryn Mawr College Class of 1991, Dogs at the Astors, The Times, Sarah Palin is NOT Hillary Clinton, I have more Foreign Policy Experience than Sarah Palin, DC Foodies, Bryn Mawr College Alumni, PeaceCorpsConnect - Returned Peace Corps Volunteers, IDS Alumni - George Washington University, Thailand will always be the Kingdom of Thailand not the Republic, International Finance Corporation / The World Bank Group, Peace Corps Thailand

**Pages** See All (1)

- World Bank Publications Non-Profit
- Bryn Mawr College Education

**Friends** See All

78 Friends

- Lucy Bucknell
- Rae Weitz
- David Pollak

Displaying members of Sarah Palin is NOT Hillary Clinton.

500+ Members No Officers 5 Admins



Name: **Kim Hennessey**  
Network: Washington, DC



Name: **Aix Healy**  
Network: Washington, DC

Emily has 78 friends.



**Elise Labott**

private profile

Networks: Turner Broadcasting  
CNN



**Paul Barry**

public profile

Message | View Friends

Networks: Washington, DC



**Daniela Araujo**

Message | View Friends

Networks: The World Bank

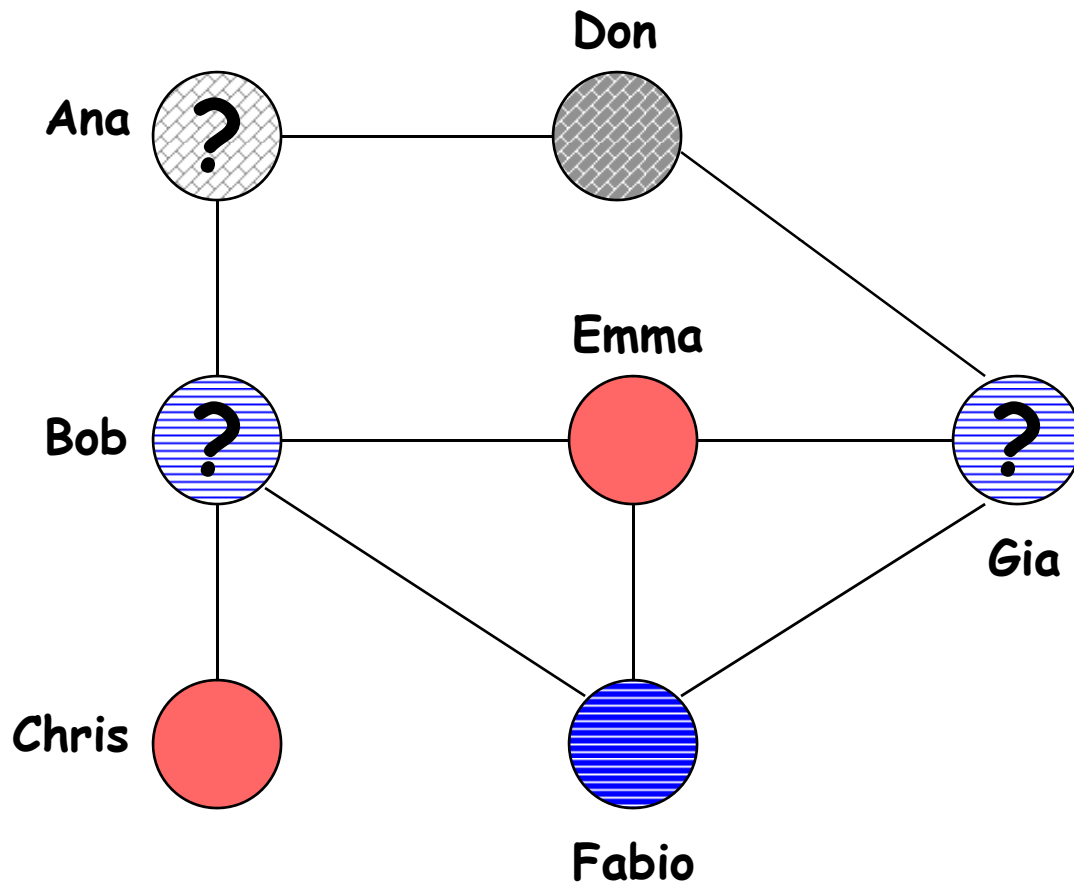


Name: **Elise Labott**  
Network: Turner Broadcasting  
CNN

# Attribute inference models

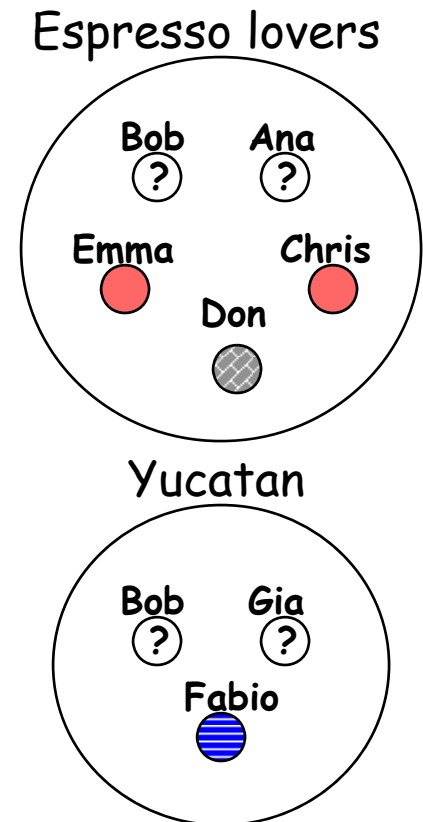
- o Toy social network

Friendship network:



- class labels (public profiles)
- unknown labels (private profiles)

Social network groups:

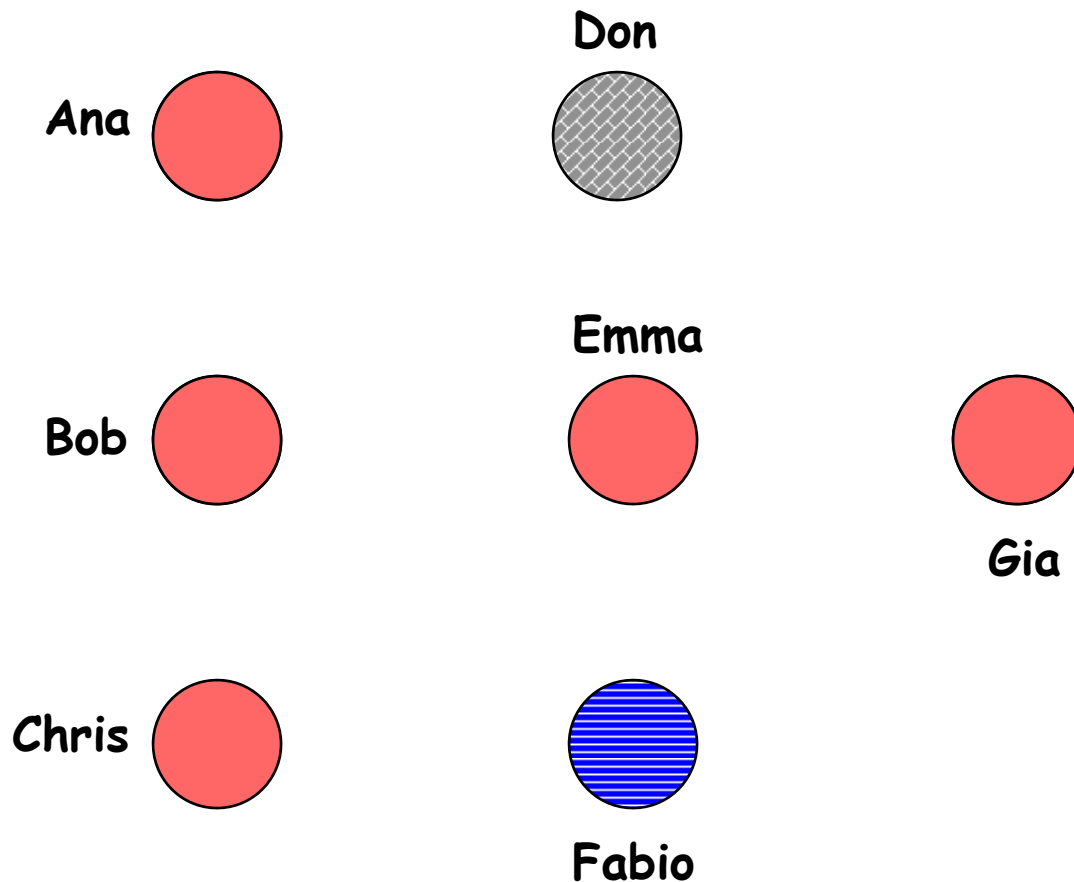








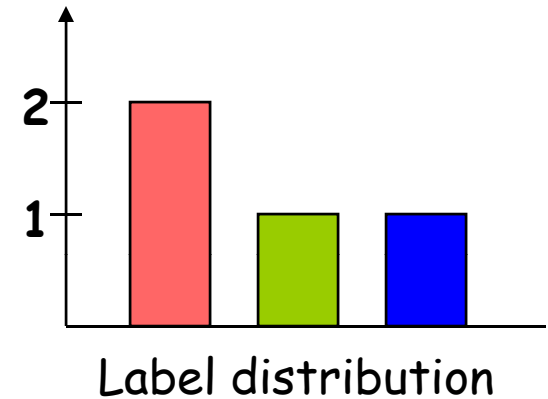
# Attribute inference models

- o In the absence of links and groups

Friendship network:



-    - class labels (public profiles)
-  - unknown labels (private profiles)

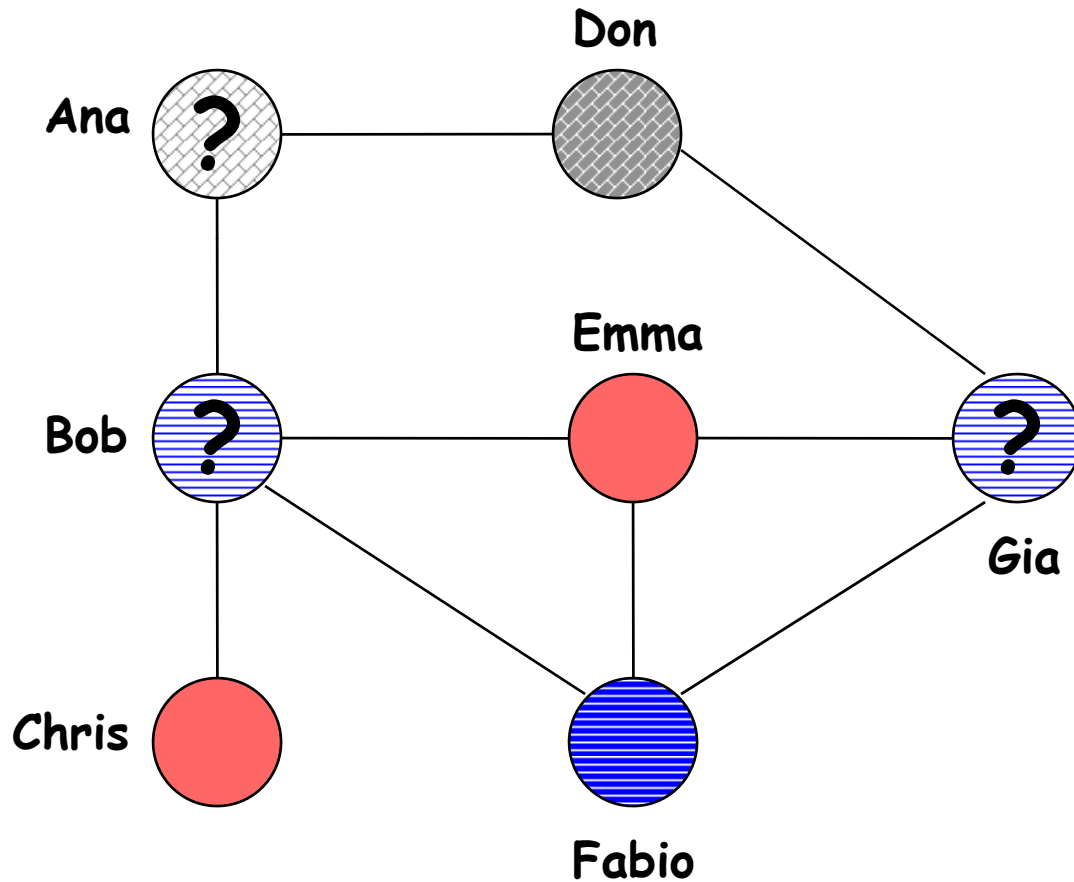


1) BASIC: assigns majority label





# Attribute inference models

- o Link-based models (in the absence of groups):

Friendship network:



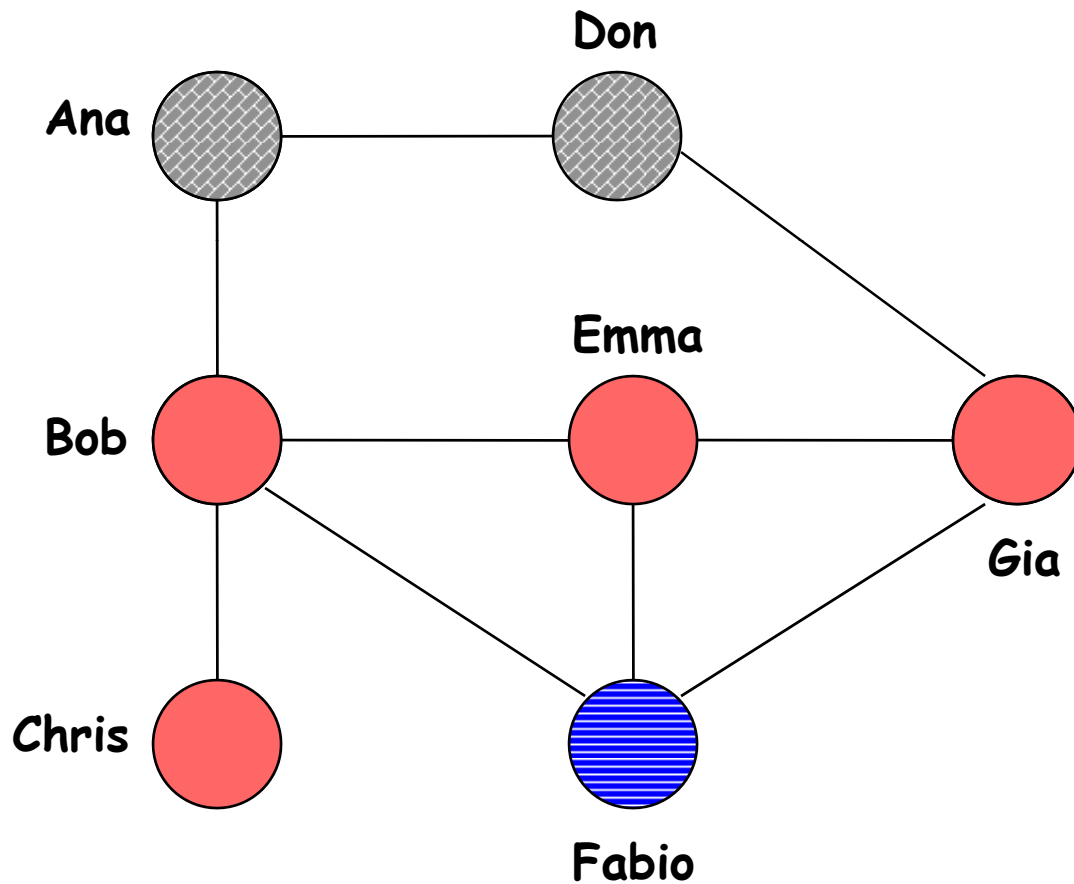
- 1) AGG
- 2) CC
- 3) LINK
- 4) BLOCK

   - class labels (public profiles)  
 - unknown labels (private profiles)

# Attribute inference models

- o Link-based models (in the absence of groups)

Friendship network:



- ● ● - class labels (public profiles)
- - unknown labels (private profiles)

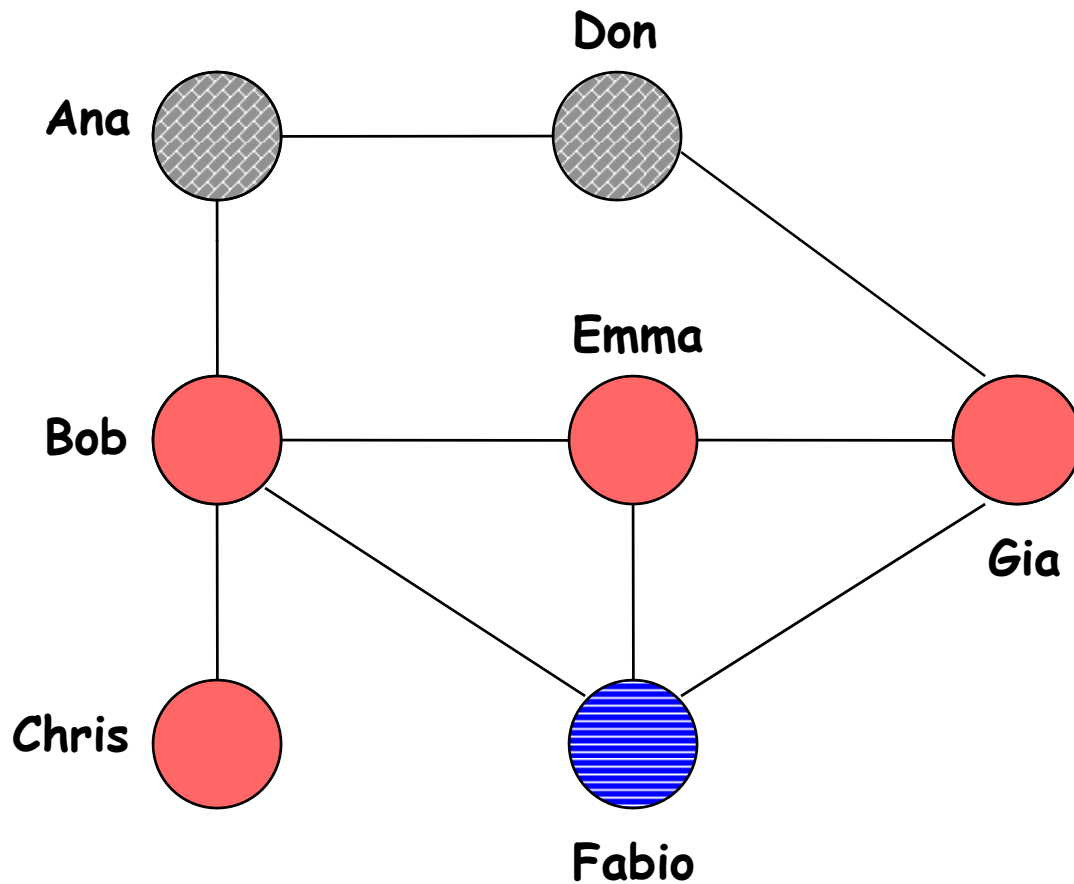
1) *AGG*:

- Take all known friends' labels and aggregate over them
- E.g., majority

# Attribute inference models

- o Link-based models (in the absence of groups)

Friendship network:



- ● ● - class labels (public profiles)
- ⊙ - unknown labels (private profiles)

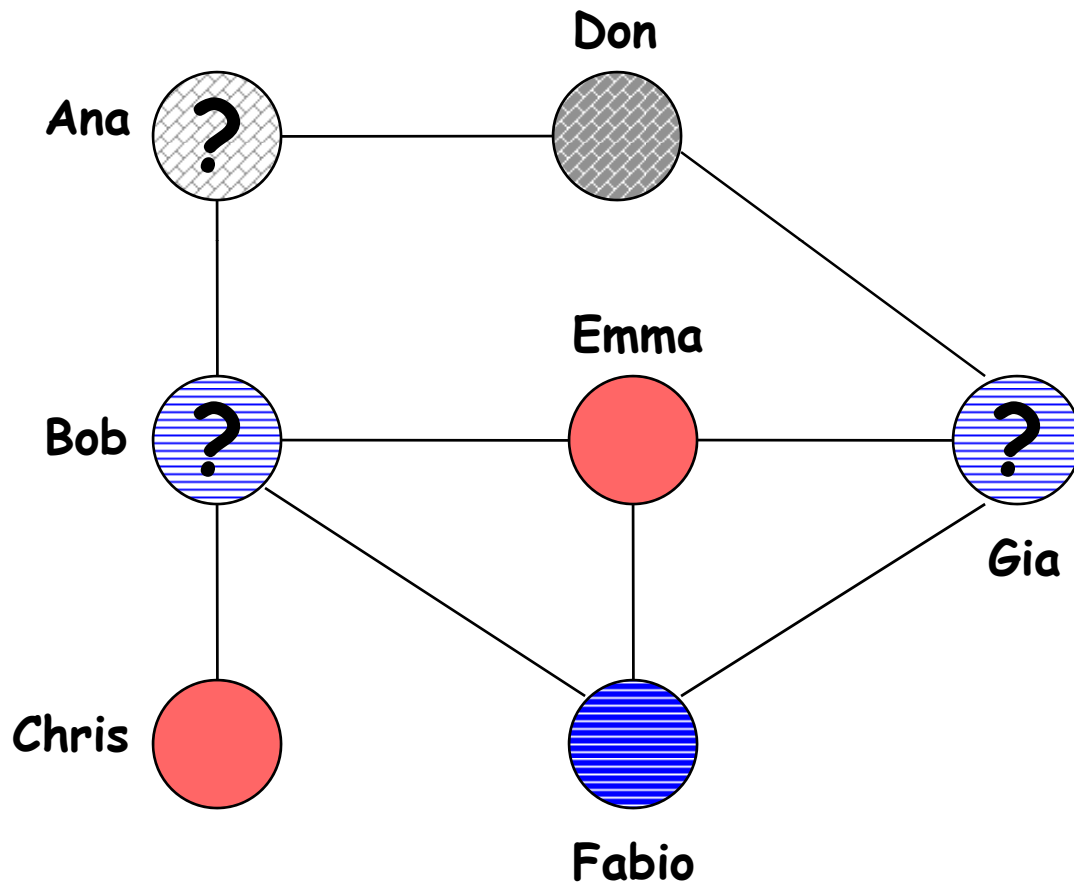
2) CC: collective classification

- Infer labels together
- Use predicted labels
- A few iterations of the nodes

# Attribute inference models

- o Link-based models (in the absence of groups)

Friendship network:



- ● ● - class labels (public profiles)
- ⊙ - unknown labels (private profiles)

3) LINK: friends as classification features

- for each user there is a vector of size N

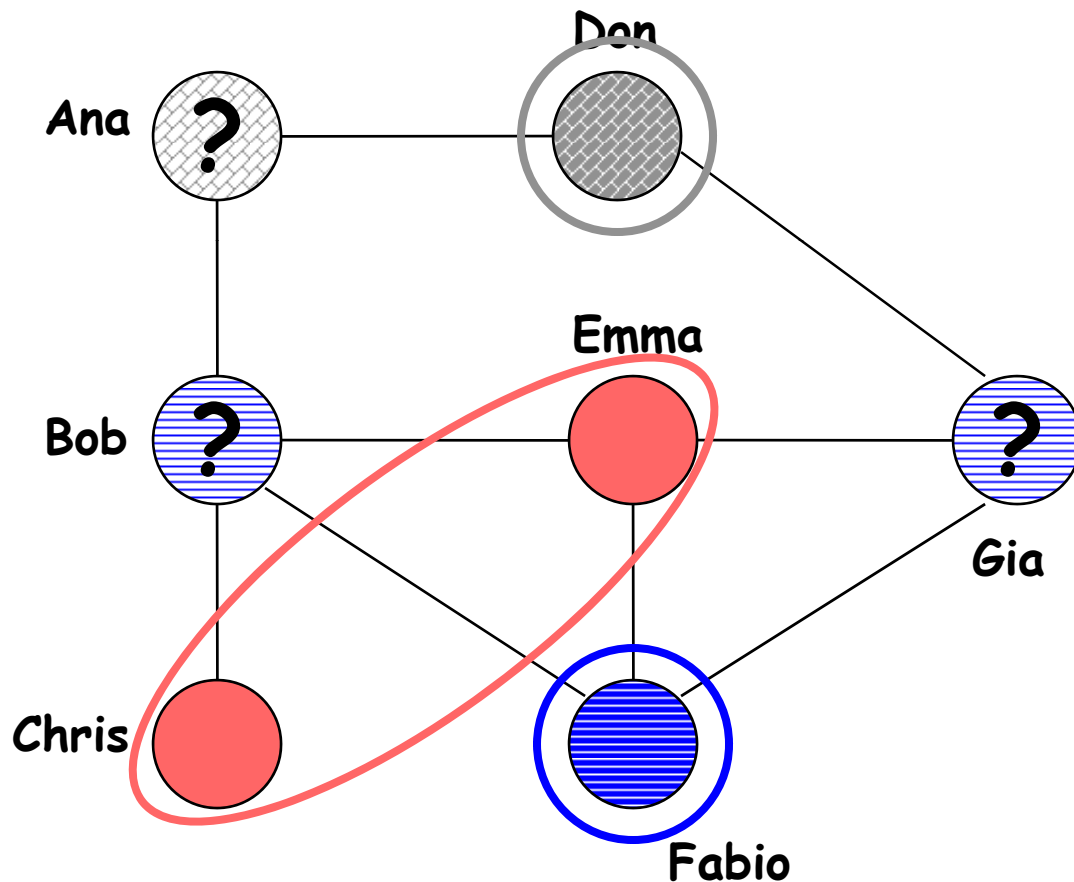
Example: Emma  
(0 1 0 0 0 1 1) ●

- train a classifier on known labels
- classify unknown labels

# Attribute inference models

- o Link-based models (in the absence of groups)

Friendship network:



- ● ● - class labels (public profiles)
- ⊙ - unknown labels (private profiles)

4) BLOCK - assume the nodes form blocks according to their labels

	●	●	●
●	0	0.5	0
●	0.5	0	0
●	0	0	0

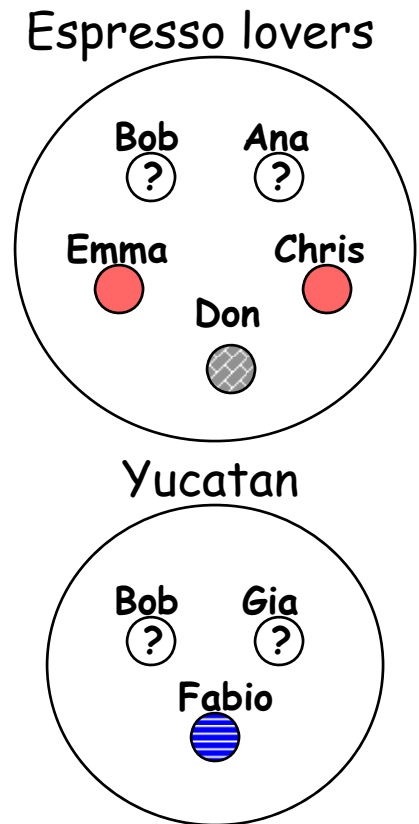
- assign probability of a node to belong to a block:
- label assignment according to most likely block

# Attribute inference models

- o Group-based models

- 1) CLIQUE
- 2) GROUP
- 3) GROUP (lower node coverage)

Social network groups:



● ● ● - class labels (public profiles)  
⊙ - unknown labels (private profiles)

# Attribute inference models

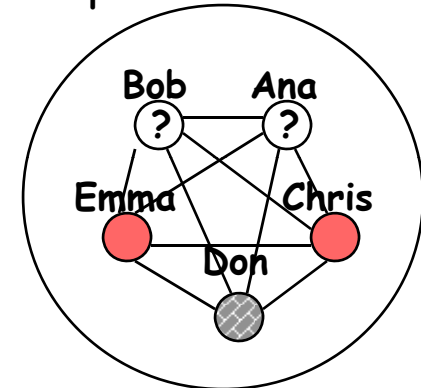
- o Group-based models

1) CLIQUE:

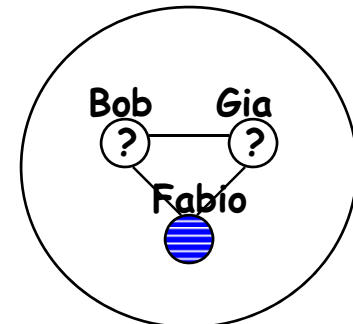
- assume friendship links between groupmates (group=clique)
- apply a link-based model

Social network groups:

Espresso lovers



Yucatan



- ● ● - class labels (public profiles)
- ⊙ - unknown labels (private profiles)



# Attribute inference models

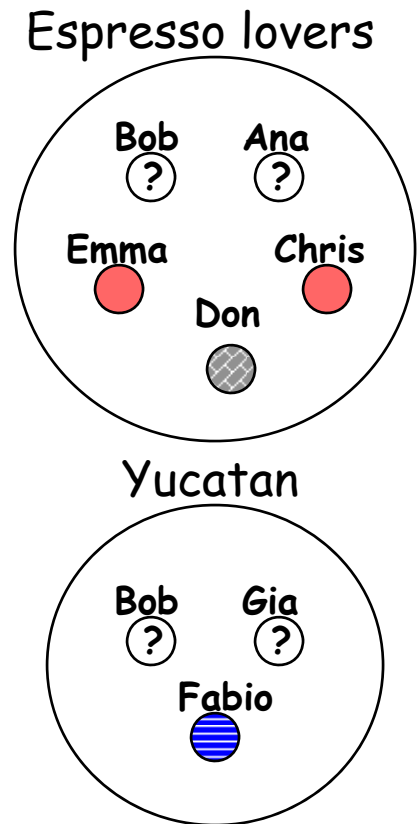
- o Group-based models

## 2) GROUP:

- use groups as classification features

Example: Emma (0 1) ●  
Ana (0 1) ?

### Social network groups:



- ● - class labels (public profiles)
- - unknown labels (private profiles)

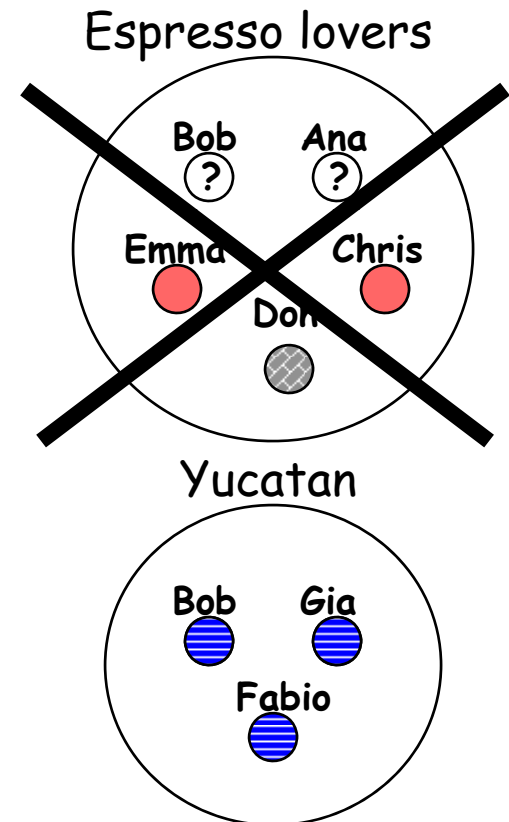
# Attribute inference models

- o Group-based models

3) GROUP (lower node coverage):

- choose *informative* groups
- apply the GROUP model

Social network groups:



- ● ● - class labels (public profiles)
- ⊙ - unknown labels (private profiles)

# Attribute disclosure results

- Given: public profiles (attribute label known), private profiles, groups, links

PROPERTY	flickr®	facebook	dogster	BibSonomy ::
Number of possible labels	55	2/6	7	2
Sensitive attribute	location	gender/polviews	breed category	spammer

**About caterina / Caterina Fake**  
← Photostream | Send FlickrMail | Buy caterina a Pro Account

Flickr Workr, friend of small dogs, art fiend, book lover. I can't come to my personal Flickr Mail account (I delete them with check out the [help page](#) or post in the forums.

<http://www.caterina.net/>  
San Francisco, CA

**caterina's contacts (575)**

- kentgokman
- heliosonnet
- clarkadial
- Pauilo Coeitho
- Indie Craft
- Julia Bucknall
- Rol Wertz
- David Polak
- rust\_sender
- etsylabs

**caterina's public groups**

- 43 Things
- mappr
- One Letter
- Parking Lot Indicator
- NEW Horror
- Maitre Goulemier et Miladus
- Ransom Note Helor
- NEW a day in the life of... 122nd
- September 2008!
- Web 2.0

**Emily Schneeweis** got up at 5 and cleaned the house (laundry, floors, fridge, sheets, recycling, bills...). 6 hours ago

**Basic Information**  
Networks: The World Bank, Washington, DC  
Sex: Female  
Birthday: February 2  
Hometown: Washington, DC  
Political Views: Liberal

**Favorite Movies:** 400 Blows, Being John Malkovich, Breakfast at Tiffany's, Casablanca, The Devil Wears Prada, Die, Die My Darling, The Butterfly, Eternat Sunrise of the Spotless Mini Translation, Manhattan, sex, lies and videotape, Vi

**Favorite Books:** Divisadero, Emma's War, Kafka on the Shore, The Maladies, Love in the Time of Cholera, Remains of

**Favorite Quotations:** Normal people are people you don't know well.

**Groups:**  
Member of:  
Bryn Mawr College Class of 1991, Dogs at the A&T Tower, Sarah Palin in NOR Hillary Clinton, I have no Policy Experience than Sarah Palin, DC Foodies, Br College Alumna, PeaceCorpsConnect - Returned P Volunteers, IDS Alumni, George Washington Unives Thailand will always be the Kingdom of Thailand - republic, International Finance Corporation / The Group, Peace Corps Thailand

**Find a Dogster**  
Find by name  
Find by town  
Find Adoptable Dogs  
Advanced Search  
Search Photo Tags  
Search Video Tags

**My Account**  
Messages  
See the Dogst  
Adoption  
Community  
Answers  
Local Listings  
Watch Videos  
Resources  
Read Diaries  
DogsterPlus  
Dogster Store  
Dogster Info  
Visit Catster

**Pi**  
Golden Retriever  
Home Silver Spring, MD  
Age: 9 Years Sex: Female Weight: 50-60 lbs

**16 Promising Content Management Systems - 16 Revisions**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**Daniel Kitzler - Softwareentwickler - XING**  
to learn something new on Nov 6, 2008, 11:31 AM

**Using CIVICRM APIs - Code Snippets - Drupal.org**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**Quantifying the accuracy of relational statements in Wikipedia: a methodology**  
Global Research and Business Intelligence and Omega Online - 2008, 2008

**Statistical Methods and Linguistics**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**Automatically Refining the Wikipedia Infobox Ontology**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**Words and what not: Happy with a Wikimania presentation**  
to see how they fit together on Nov 9, 2008, 10:57 PM

**Revision 1551: Wikipedia-tracker/track**  
to see how they fit together on Nov 9, 2008, 10:57 PM

**Andrew's MediaWiki Codebase Reference**  
to see how they fit together on Nov 9, 2008, 10:57 PM

**Using CIVICRM APIs - Code Snippets - Drupal.org**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**16 Promising Content Management Systems - 16 Revisions**  
to see how they fit together on Nov 7, 2008, 12:34 AM

**Daniel Kitzler - Softwareentwickler - XING**  
to learn something new on Nov 6, 2008, 11:31 AM

**Main Page - WikiTrust**  
to see how they fit together on Nov 5, 2008, 11:30 AM

**WIKIMANIA: The Pit, The Spotted Brown Eyed Cow-Cow of the Eastern Mandu Valley**  
Doggie Dynamics

**Sun Sign: AQUARIUS** See today's dog horoscope

**Quick Bio: purebred**  
Likes: Swimming, fetching  
Pet/Personae: Getting her ears cleaned

**Meet my family**

**Meet my Pup Pals**

Reference: E. Zheleva, L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. WWW 09.

# Attribute disclosure results

- Approaches to achieving attribute disclosure:
  - Using overall distribution - BASIC
  - Link-based - BLOCK, AGG, CC, LINK
  - Group-based - CLIQUE-LINK, GROUP

Table 2: Attack accuracy assuming 50% private profiles. The successful attacks are shown in bold.

	DEL	FLICKR	FACEBOOK (GENDER)	FACEBOOK (POLVIEWS)	DOGSTER	BIBSONOMY
BASIC		27.7%	50.0%	56.5%	28.6%	92.2%
Random guess		1.8%	50.0%	16.7%	14.3%	50%
BLOCK		8.8%	49.1%	6.1%	-	-
AGG		28.4%	50.2%	57.6%	-	-
CC		28.6%	50.4%	56.3%	-	-
LINK		<b>56.5%</b>	<b>68.6%</b>	58.1%	-	-
CLIQUE-LINK		<b>46.3%</b>	51.8%	57.1%	<b>60.2%</b>	-
GROUP		<b>63.5%</b>	<b>73.4%</b>	45.2%	<b>65.5%</b>	<b>94.0%</b>
GROUP (50% node coverage)		<b>83.6%</b>	<b>77.2%</b>	46.6%	<b>82.0%</b>	<b>96.0%</b>

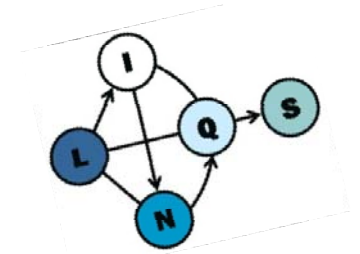
What's the connection?

Inference  $\Rightarrow$  Identification

$\neg$ Identification  $\Rightarrow$  Privacy



# LINQS Group @ UMD



- Members: myself, *Indrajit Bhattacharya*, *Mustafa Bilgic*, *Lei Guang*, *Sarn Huang*, *Rezarta Islamaj*, *Hyunmo Kang*, *Louis Licamele*, *Qing Lu*, *Walaa El-Din Mustafa*, *Galileo Namata*, *Barna Saha*, *Prithivaraj Sen*, *Vivek Sehgal*, *Hossam Sharara*, *Elena Zheleva*



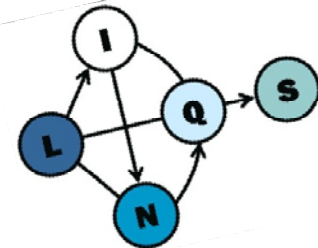
# Conclusion

- 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
  - Search: Abstractions of click and query graphs
- 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/~getoor>

Work sponsored by the National Science Foundation,  
Google, Microsoft, KDD program,  
National Geospatial Agency, Army Research Office



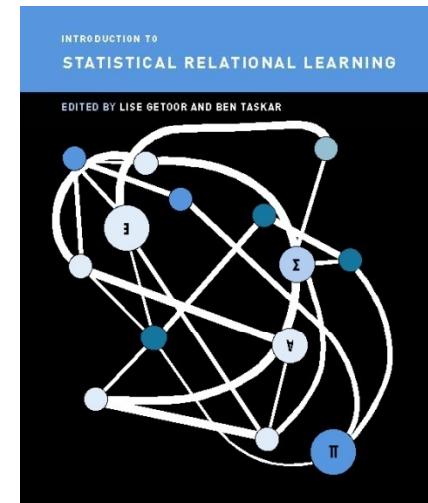


# Statistical Relational Learning (SRL)

- 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