



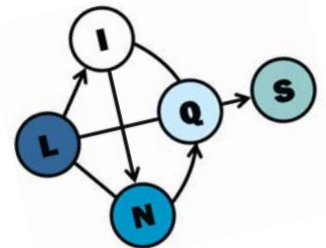
Making Sense of Social Networks using Pairwise Visualization and Analysis

Lise Getoor

University of Maryland, College Park



HCIL SNA Workshop
June 1, 2007



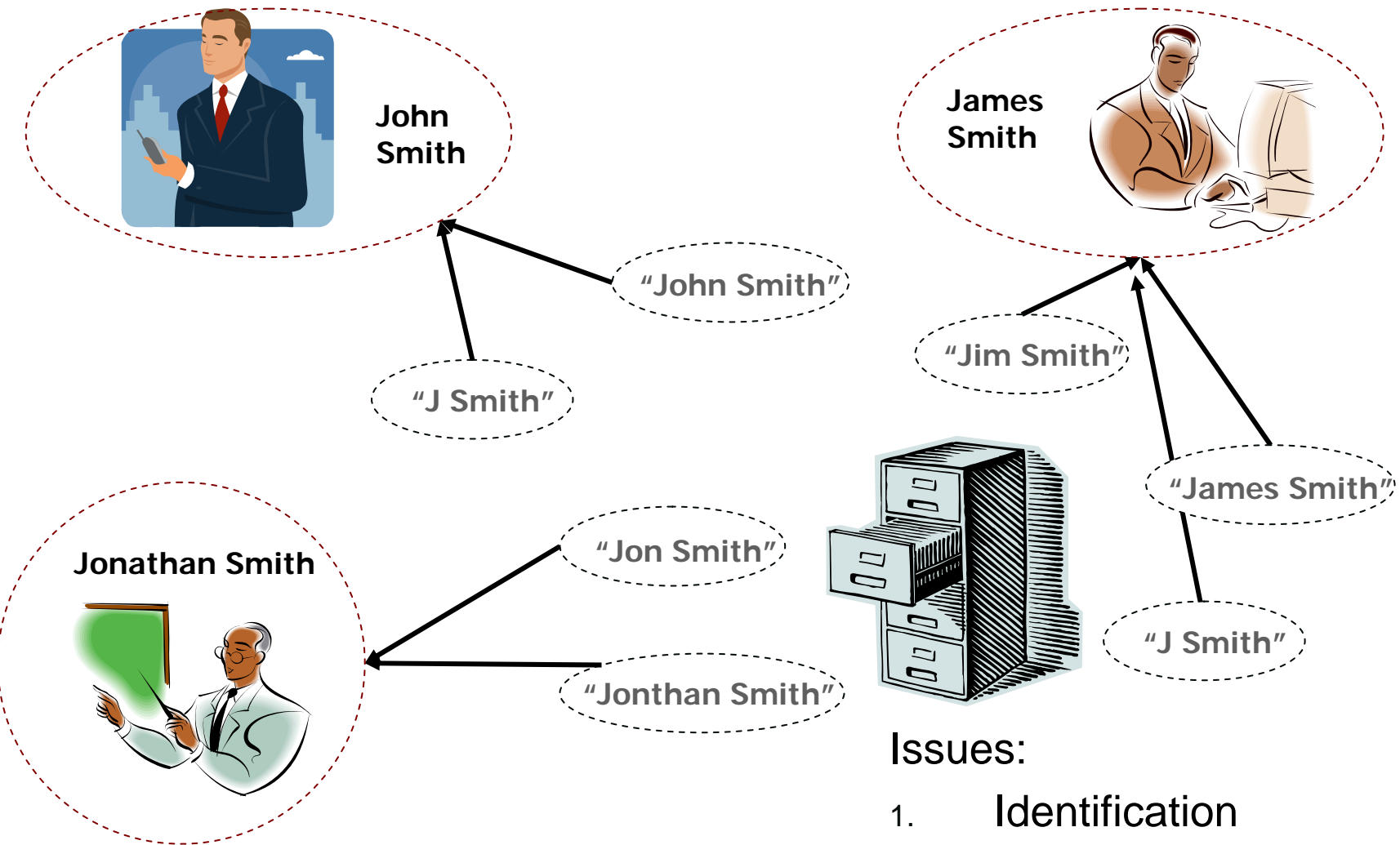
● ● ● Motivation

- Growing need for tools which help us understand and make sense of social networks
 - Many tools allow flexible overview and navigation the entire network
 - Some tools focus on a single actor-centric view of the network
- The tools I will describe today focus on a **pairwise** paradigm well-suited to certain visual analytic tasks

● ● ● Outline

- Problem #1
 - **Task: Entity Resolution**
 - Tool: D-Dupe
 - URL: <http://www.cs.umd.edu/lings/ddupe>
 - *H. Kang, M. Bilgic, L. Licamele, B. Shneiderman, VAST06*
- Problem #2
 - Task: Geospatial Data Integration
 - Tool: GeoD-Dupe
 - URL: <http://www.cs.umd.edu/lings/geoddupe>
 - *H. Kang, V. Sehgal, IV07*
- Problem #3
 - Task: Dynamic Group Membership
 - Tool: C-Dupe
 - URL: <http://www.cs.umd.edu/lings/cgroup>
 - H. Kang, L. Singh, J. Mann, E. Krzyszczyk, under review
- The Big Picture

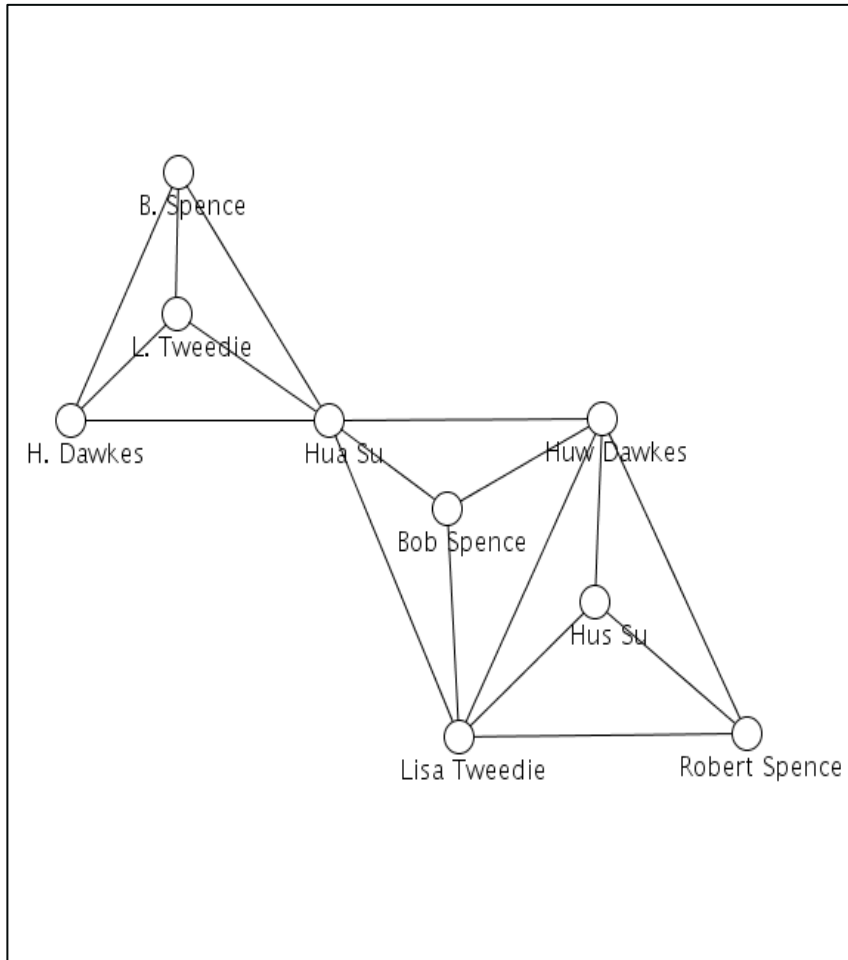
The Entity Resolution Problem



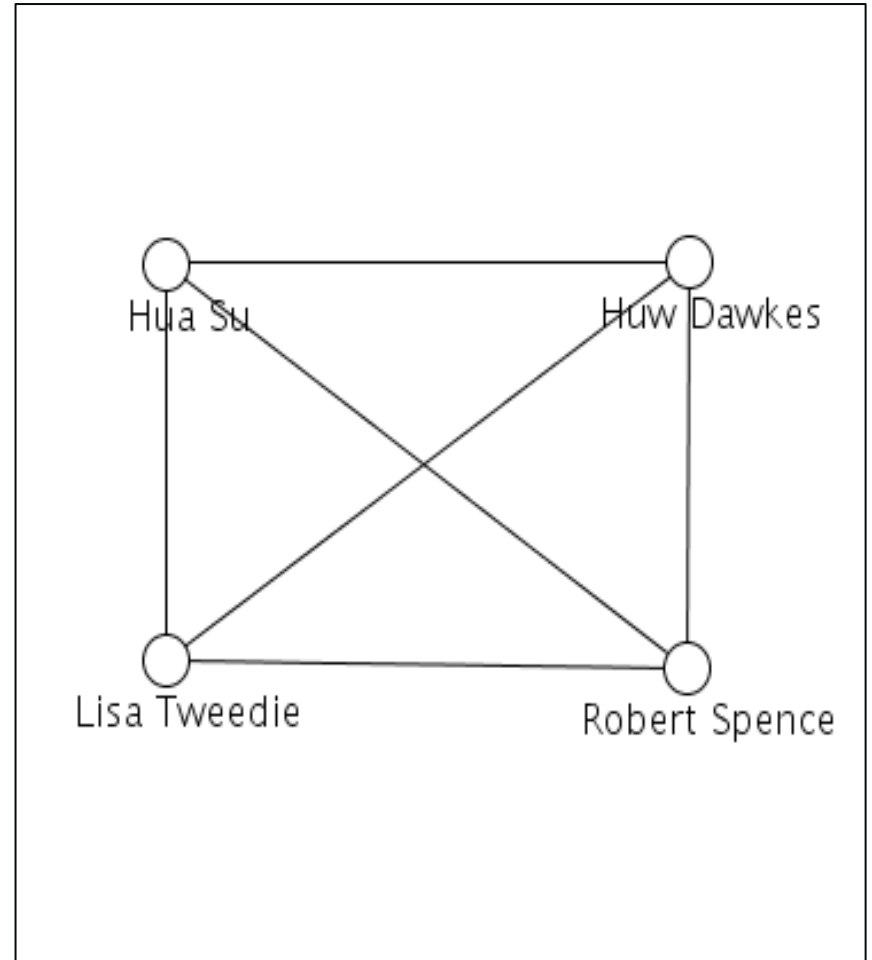
Issues:

1. Identification
2. Disambiguation

● ● ● InfoVis Co-Author Network Fragment



before

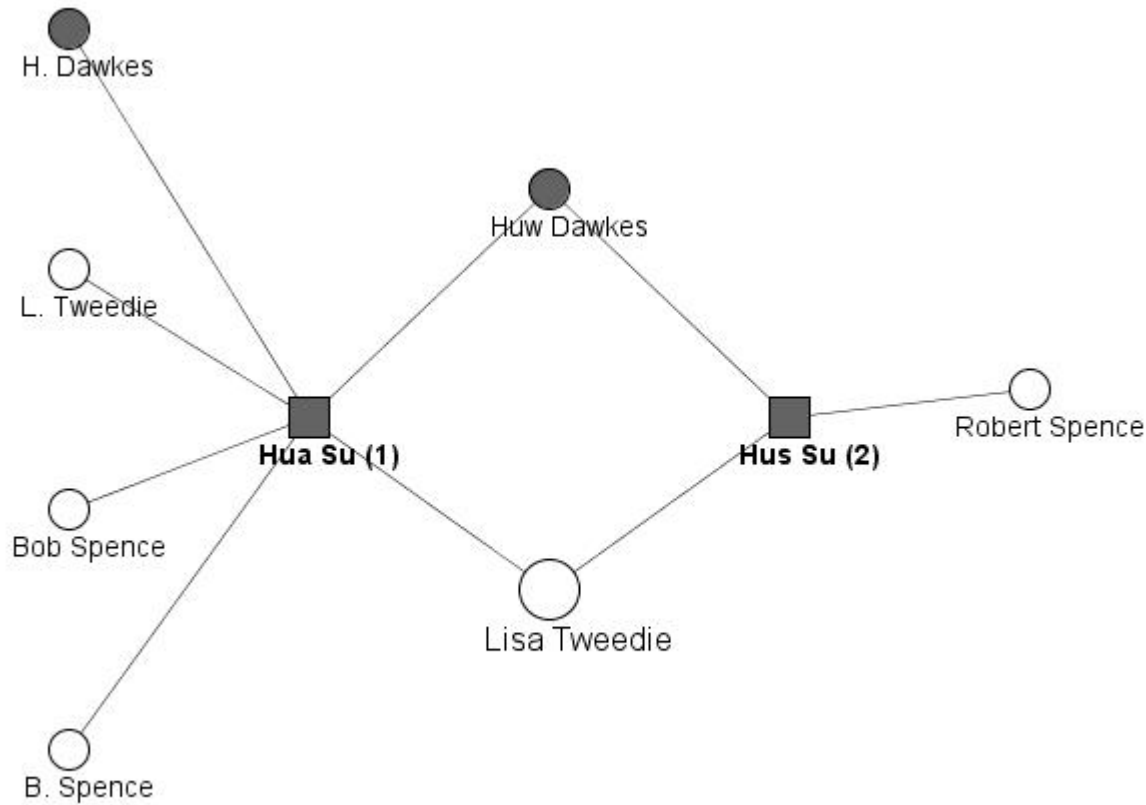


after

● ● ● Entity Resolution in Networks

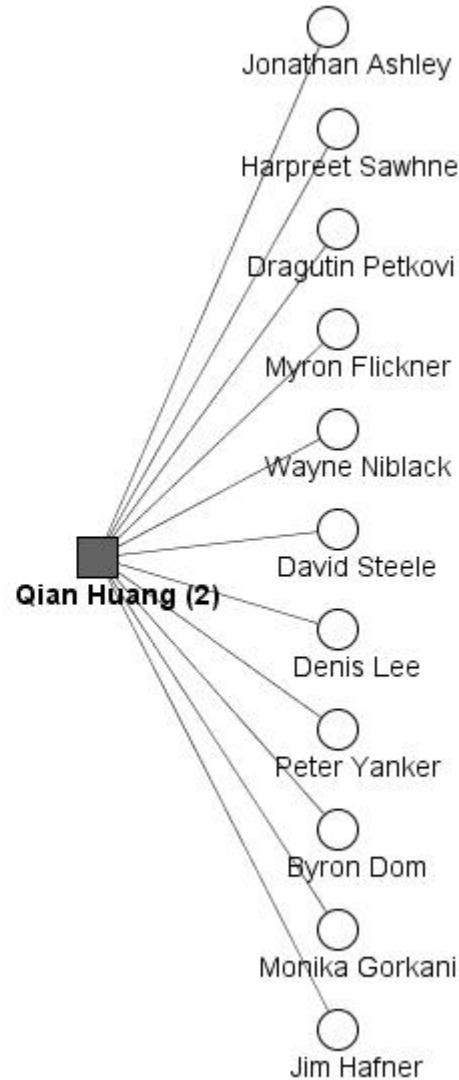
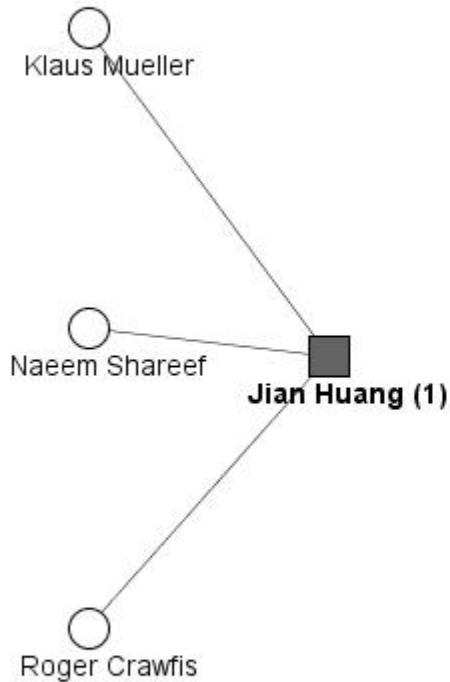
- References not observed independently
 - Links between references indicate relations between the entities
 - Co-author relations for bibliographic data
 - To, cc: lists for email
- Use relations to improve identification and disambiguation

● ● ● Relational Identification



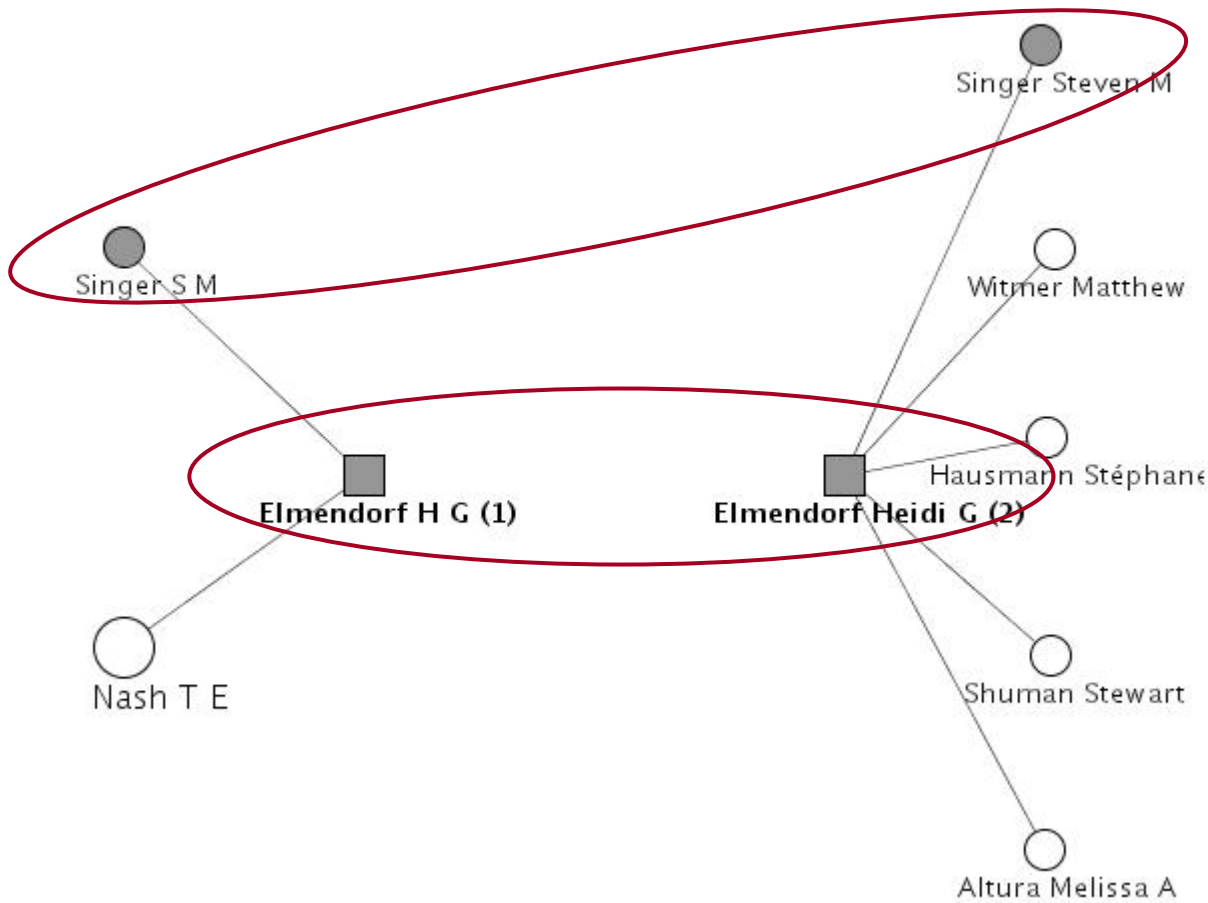
Very similar names.
Added evidence from
shared co-authors

● ● ● Relational Disambiguation



Very similar names
but no shared
collaborators

● ● ● Collective Entity Resolution



One resolution provides evidence for another => joint resolution

● ● ● Entity Resolution Algorithms

○ Relational Clustering

- Agglomerative clustering algorithm which uses attribute and relational similarity
- *Bhattacharya and Getoor, DMKD'04, Wiley'06, TKDD'07*

○ Generative Probabilistic Model

- Nonparametric Bayesian Model
- Discovers underlying group structure
- Allows overlapping groups
- *Bhattacharya and Getoor, SIAM SDM'06, Best Paper Award*

Objective Function

- Minimize:

$$\sum_i \sum_j w_A \text{sim}_A(c_i, c_j) + w_R \delta(c_i, c_j)$$

weight for
attributes

similarity of
attributes

weight for
relations

1 iff relational edge
exists between c_i and c_j

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

$$\Delta(c_i, c_j) = w_A \text{sim}_A(c_i, c_j) + w_R (|N(c_i) \cap N(c_j)|)$$

Similarity of attributes

Common cluster neighborhood

● ● ● Relational Clustering Algorithm

1. Find similar references using 'blocking'
2. Bootstrap clusters using attributes and relations
3. Compute similarities for cluster pairs and insert in 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

○

CODE AND DATA AND DATA GENERATOR AVAILABLE HERE:
<http://www.cs.umd.edu/~indrajit/ER/>

● ● ● Outline

- Problem #1
 - Task: Entity Resolution
 - **Tool: D-Dupe**
 - URL: <http://www.cs.umd.edu/lings/ddupe>
 - *H. Kang, M. Bilgic, L. Licamele, B. Shneiderman, VAST06*
- Problem #2
 - Task: Geospatial Data Integration
 - Tool: GeoD-Dupe
 - URL: <http://www.cs.umd.edu/lings/geoddupe>
 - *H. Kang, V. Sehgal IV07*
- Problem #3
 - Task: Dynamic Group Membership
 - Tool: C-Dupe
 - URL: <http://www.cs.umd.edu/lings/cgroup>
 - H. Kang, L. Singh, J. Mann, E. Krzyszczyk, under review
- The Big Picture

D-Dupe: An Interactive Tool for Entity Resolution

The screenshot displays the D-Dupe application window. On the left, a 'Find Duplicates' panel shows a table of potential duplicate pairs with similarity scores. The main area features a network graph with nodes representing authors and edges representing relationships. A search bar at the top left contains the text 'L. Tweedie'. Below the graph, there are buttons for 'Merge Duplicates' and 'Mark Distinct'. At the bottom, there are two detail viewers: 'Node Detail Viewer (7 items)' and 'Edge Detail Viewer (3 items)'. The status bar at the bottom right indicates 'Finding possible duplicates completed!'.

Similarity	Node1	Node2
0.888888888888889	Hua Su	Hus Su
0.746031746031746	Hua Su	Alan Su
0.650793650793651	Hua Su	Stuart Shieber
0.6		
0.6		
0.6		
0.6		
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		
0.590909090909091		
0.590909090909091		
0.590909090909091		
0.590909090909091		
0.590909090909091		
0.590909090909091		

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

Novel combination of network visualization and statistical relational models well-suited to the visual analytic task at hand

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 Hwai-hsin Chi
P139655	Jian Huang

AuthorID	AuthorName
P573115	H. Dawkes
P572966	B. Spence
P113087	Huw Dawkes
P172581	Lisa Tweedie
P573241	L. Tweedie
P31332	Bob Spence
P246545	Robert Spence

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!

● ● ● Outline

- Problem #1
 - Task: Entity Resolution
 - Tool: D-Dupe
 - URL: <http://www.cs.umd.edu/lings/ddupe>
 - *H. Kang, M. Bilgic, L. Licamele, B. Shneiderman, VAST06*
- **Problem #2**
 - **Task: Geospatial Data Integration**
 - **Tool: GeoD-Dupe**
 - URL: <http://www.cs.umd.edu/lings/geoddupe>
 - *H. Kang, V. Sehgal IV07*
- Problem #3
 - Task: Dynamic Group Membership
 - Tool: C-Dupe
 - URL: <http://www.cs.umd.edu/lings/cgroup>
 - H. Kang, L. Singh, J. Mann, E. Krzyszczyk, under review
- The Big Picture

● ● ● Outline

- Problem #1
 - Task: Entity Resolution
 - Tool: D-Dupe
 - URL: <http://www.cs.umd.edu/lings/ddupe>
 - *H. Kang, M. Bilgic, L. Licamele, B. Shneiderman, VAST06*
- Problem #2
 - Task: Geospatial Data Integration
 - Tool: GeoD-Dupe
 - URL: <http://www.cs.umd.edu/lings/geoddupe>
 - *H. Kang, V. Sehgal IV07*
- **Problem #3**
 - **Task: Dynamic Group Membership**
 - Tool: C-Dupe
 - URL: <http://www.cs.umd.edu/lings/cgroup>
 - H. Kang, L. Singh, J. Mann, E. Krzyszczyk, under review
- The Big Picture

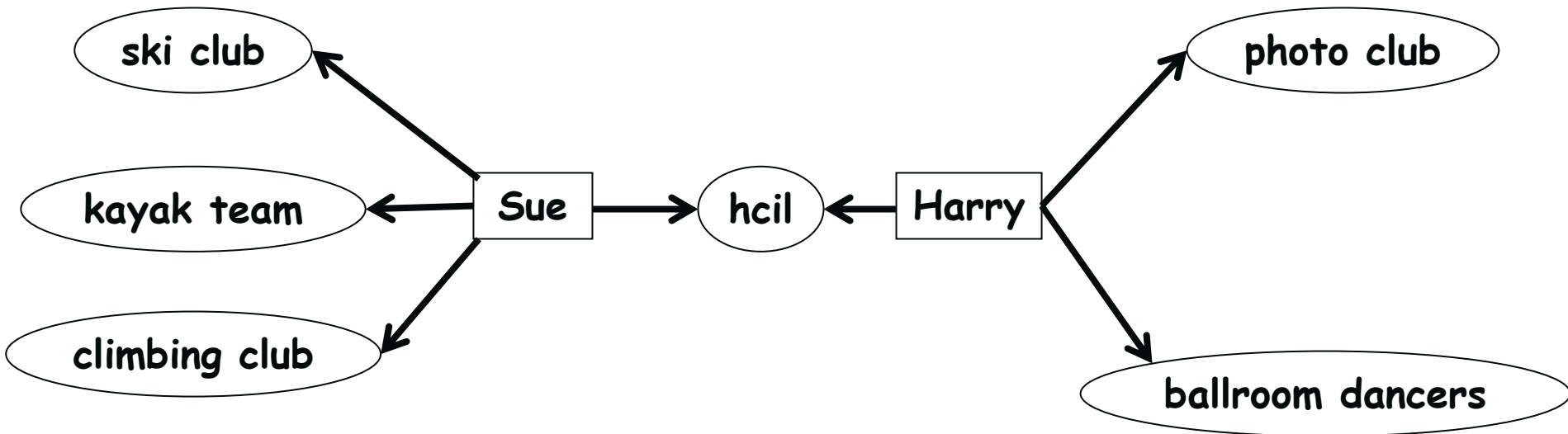
● ● ● Finding Groups in Social Networks

- Topic of intense interest
- Many methods proposed

- Here, we focus on **dynamic** group membership, how group membership changes over time
- And, we focus on comparing the group memberships of a **pair** of actors

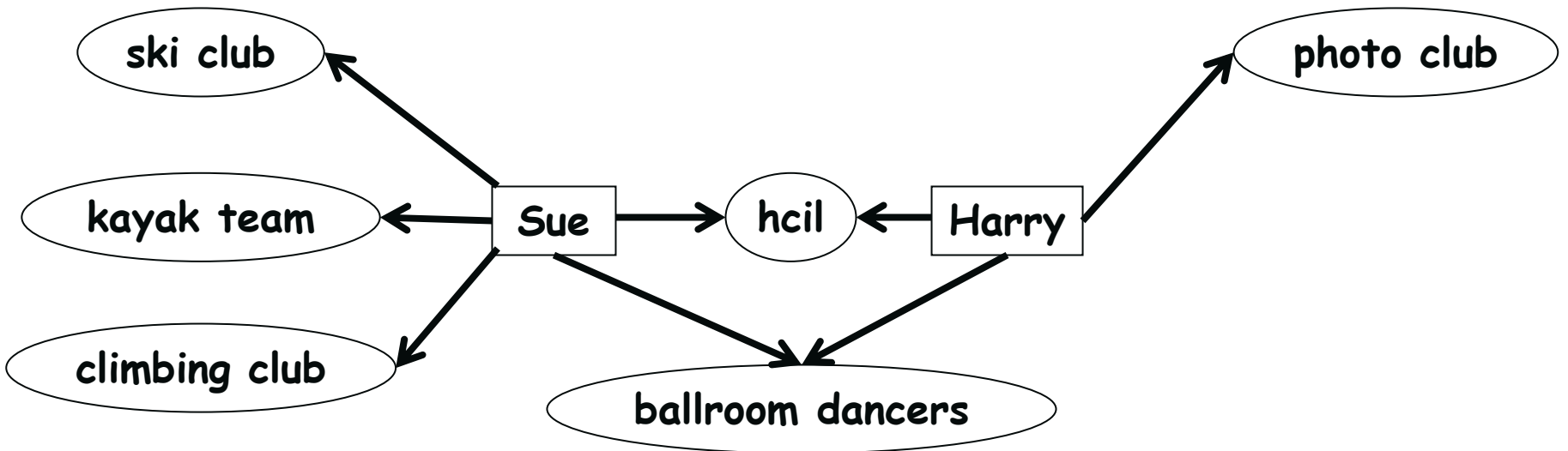
Dynamic Group Membership

Year 1...



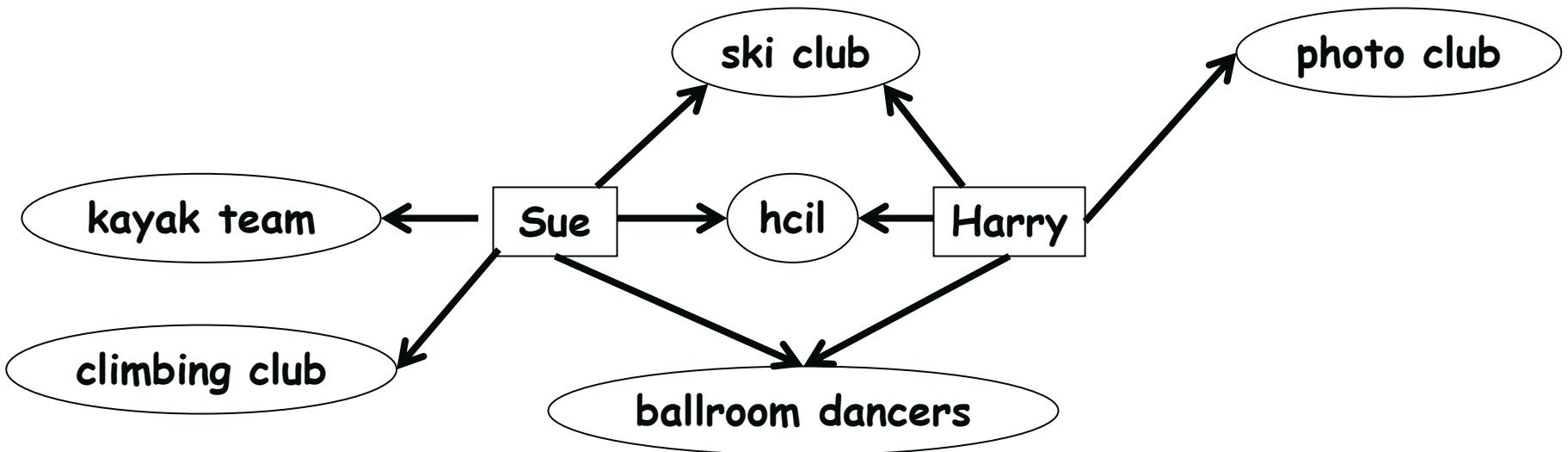
Dynamic Group Membership

Year 2....



Dynamic Group Membership

Year 3...



● ● ● Outline

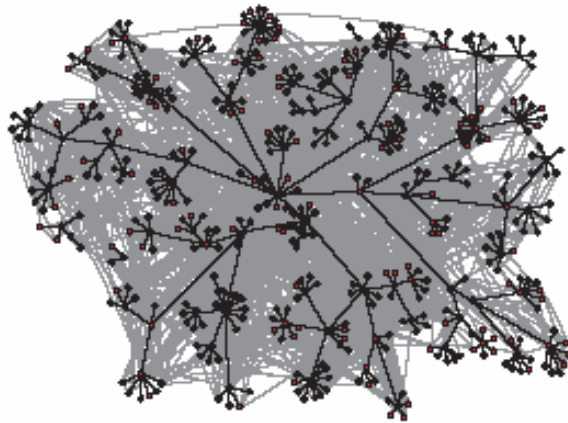
- Problem #1
 - Task: Entity Resolution
 - Tool: D-Dupe
 - URL: <http://www.cs.umd.edu/lings/ddupe>
 - *H. Kang, M. Bilgic, L. Licamele, B. Shneiderman, VAST06*
- Problem #2
 - Task: Geospatial Data Integration
 - Tool: GeoD-Dupe
 - URL: <http://www.cs.umd.edu/lings/geoddupe>
 - *H. Kang, V. Sehgal, IV07*
- Problem #3
 - Task: Dynamic Group Membership
 - **Tool: C-Dupe**
 - URL: <http://www.cs.umd.edu/lings/cgroup>
 - H. Kang, L. Singh, J. Mann, E. Krzyszczyk, under review
- The Big Picture

● ● ● Putting Everything together....

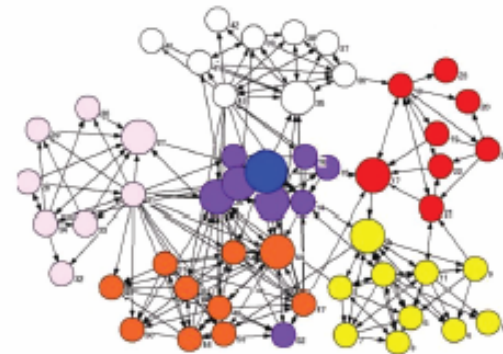
Pairwise useful for certain tasks, but still want to see bigger picture



Collaborative Social
Network Discovery
Entity Resolution
Relationship Identification



Communications Graph
Nodes: Network References
Edges: Communications Events



Network Graph
Nodes: Entities
Edges: Social Relationships

Currently working on analytic and visual tools that support interactive, exploratory, dual view of data network and information network



Thanks!

<http://www.cs.umd.edu/~getoor>

Work sponsored by the National Science Foundation,
Google, KDD program and National Geospatial Agency

