Graph Identification & Alignment

Lise Getoor

University of Maryland, College Park

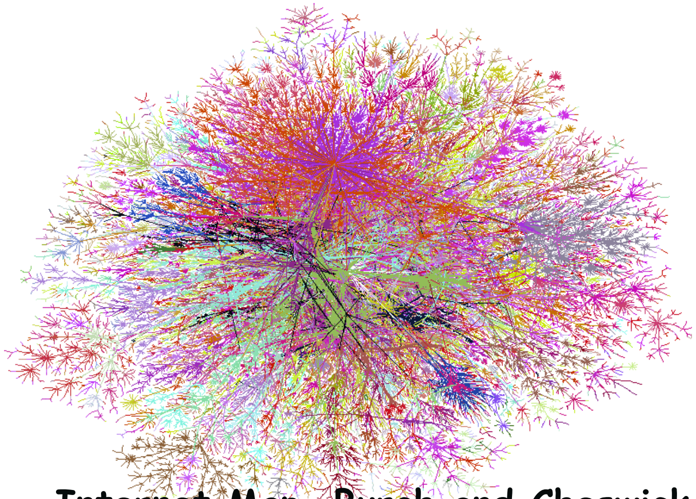
DOE/DOD Workshop on Emerging High Performance Architectures and Applications
November 29, 2007

Roadmap

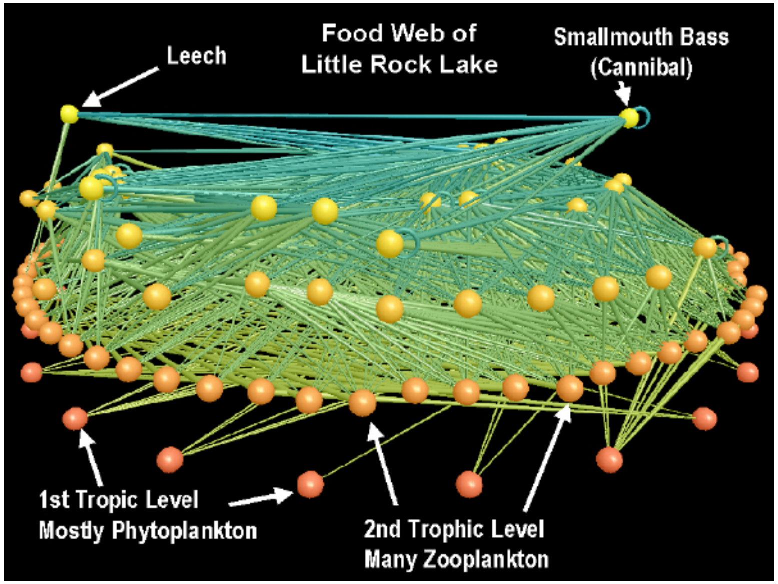
- Motivating Applications
- Algorithms
- Challenges and Opportunities

Graphs and Networks *everywhere*…

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



Internet Map, Burch and Cheswick



Food Web, Martinez

Others available at Mark Newman's gallery:
http://www-personal.umich.edu/~mejn/networks/

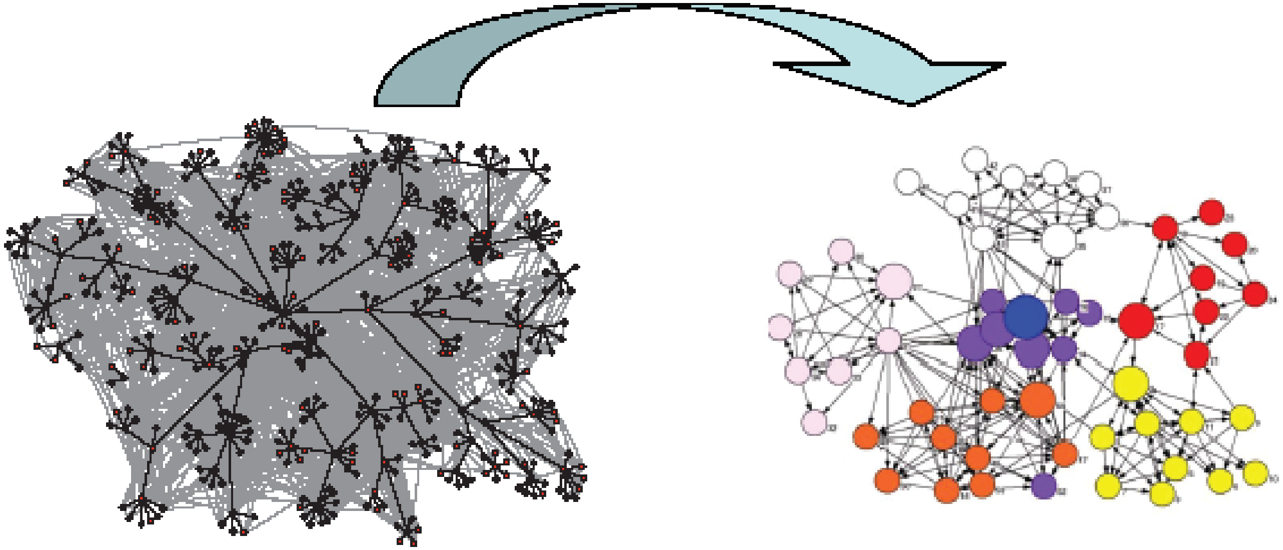
Wealth of Data

- 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



Communications Graph
Nodes: Network References
Edges: Communications Events

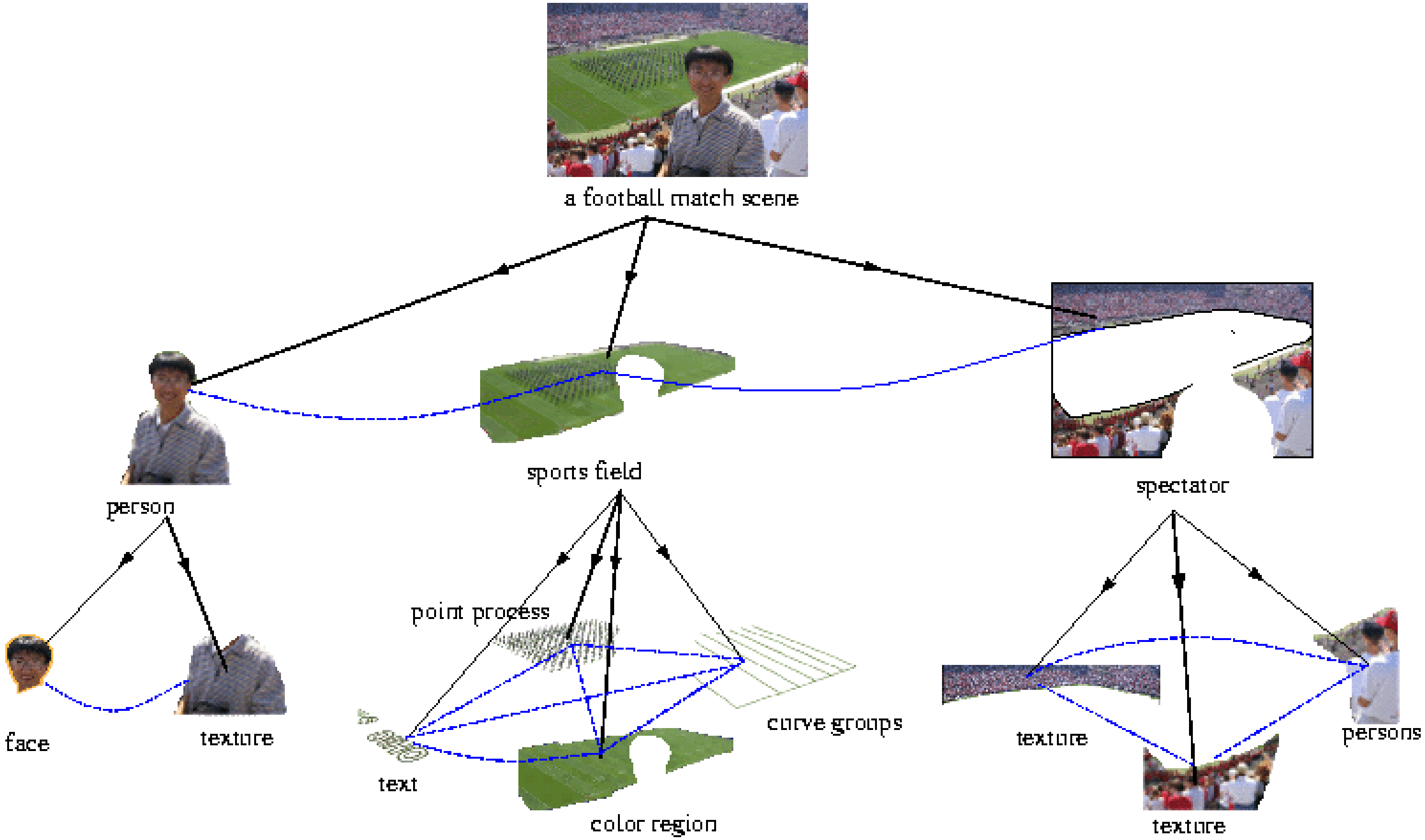
Network Graph
Nodes: Entities
Edges: Social Relationships

Data Graph ⇒ 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

HP Labs, Huberman & Adamic

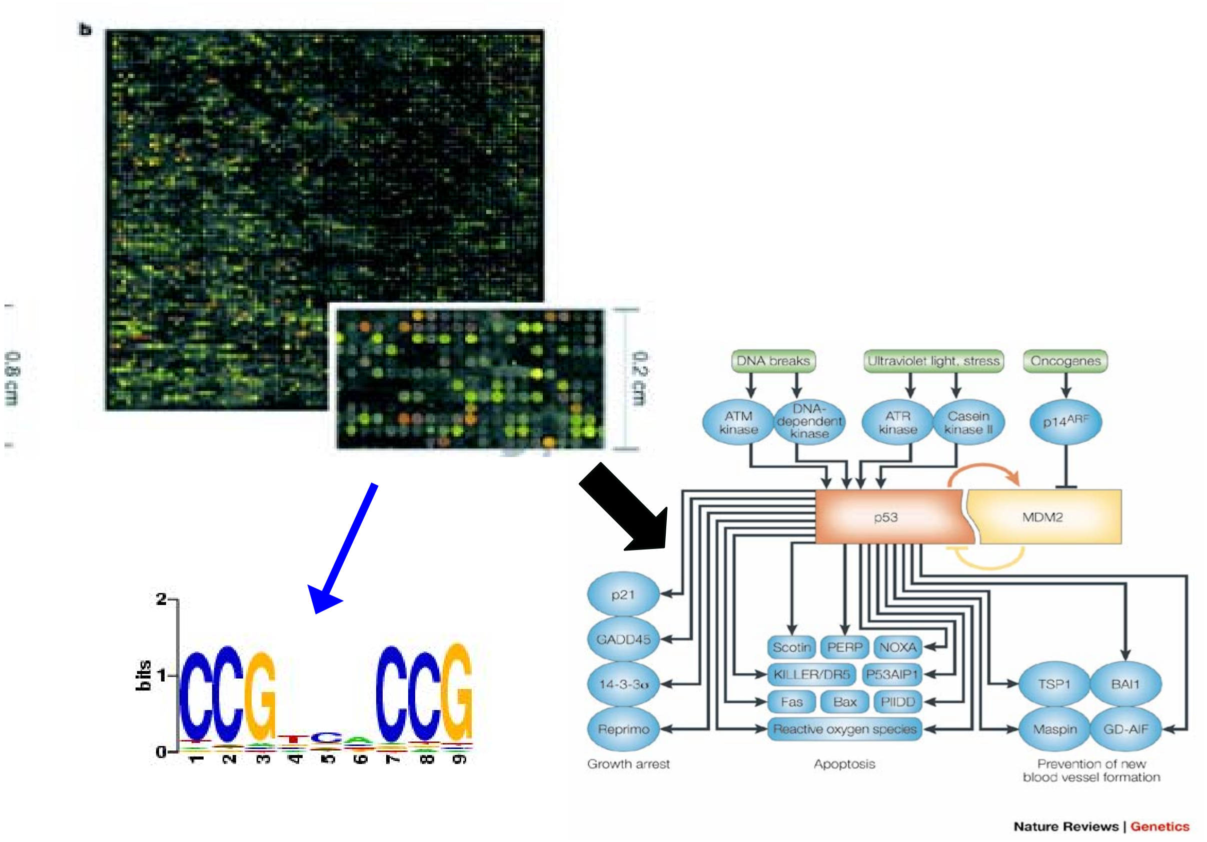
Vision: Image Parsing



Graph Partitioning + Graph Matching

Z.W. Tu, X.R. Chen, A.L. Yuille, and S.C. Zhu, IV05; Lin, Zhu and Wang, IV07

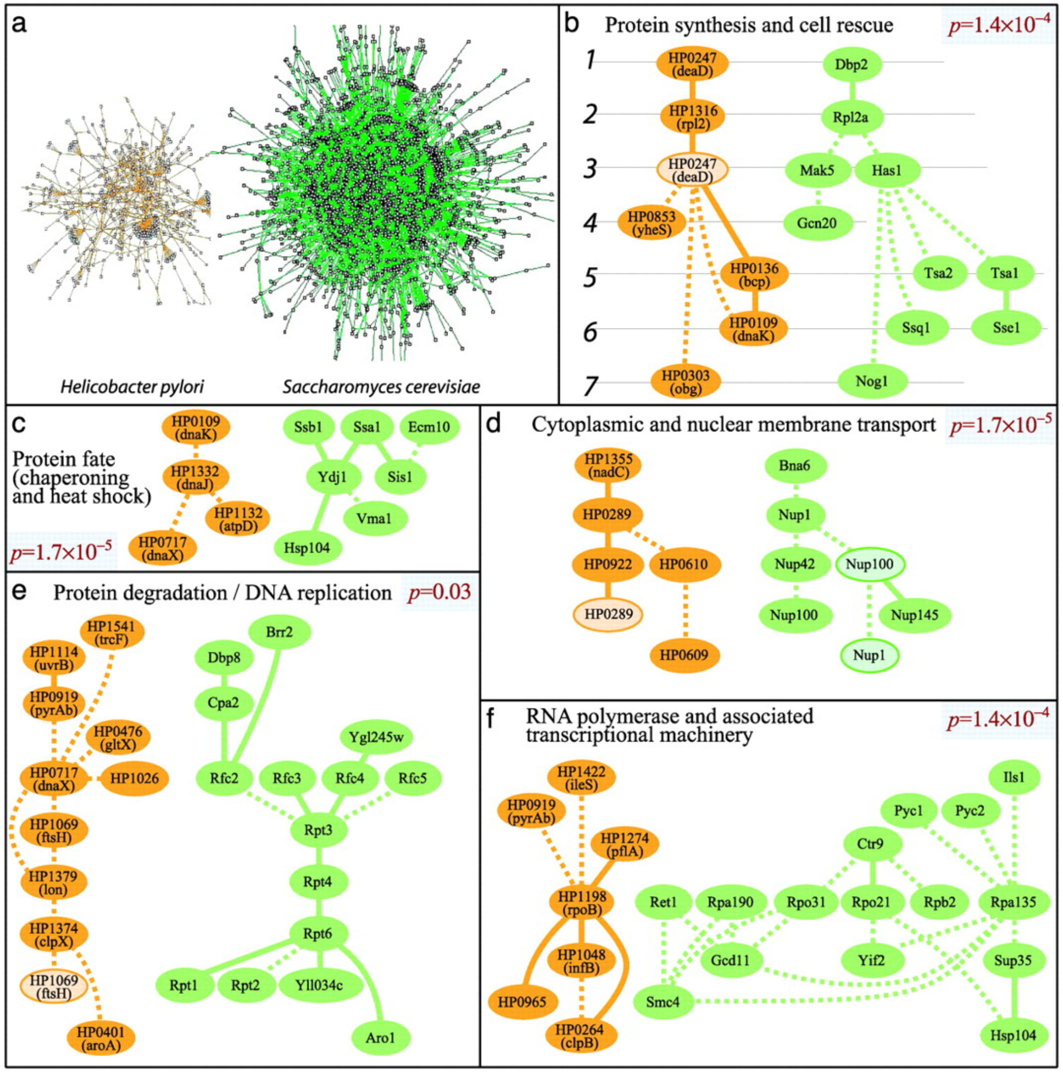
Bio: Graph Identification



Biological Networks: protein-protein, transcriptional regulation, signaling

Courtesy of Chris Wiggins, Columbia

Bio: Graph Alignment

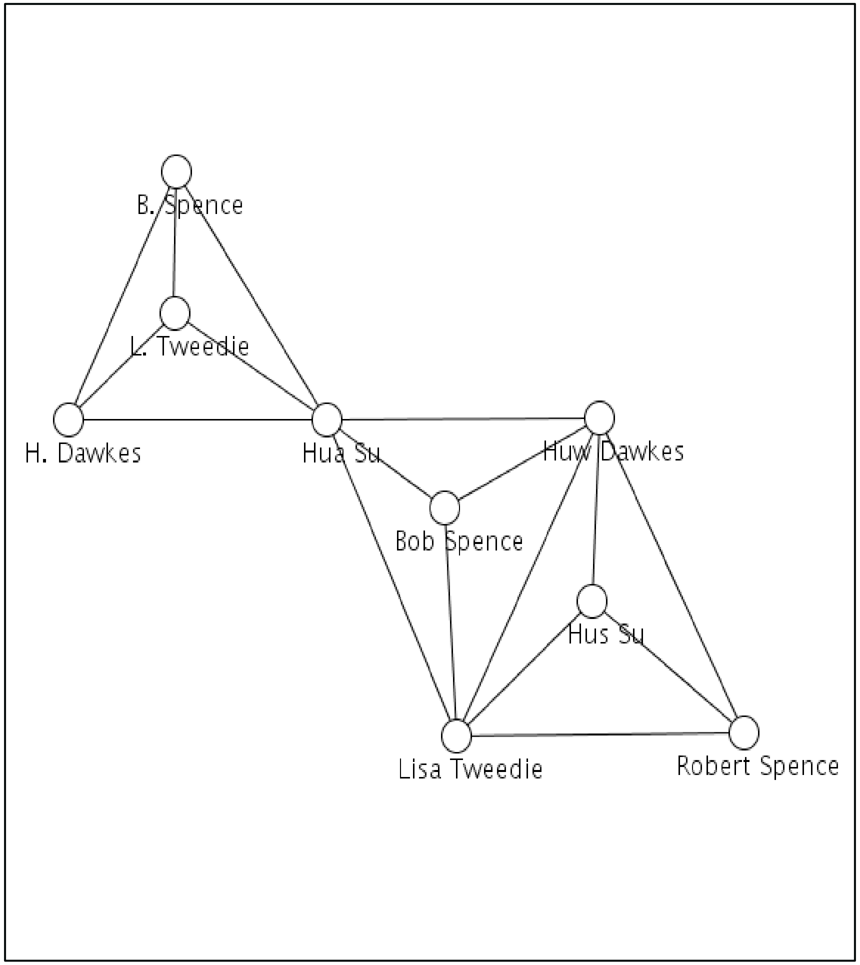


Kelley, Brian P. et al. PNAS03

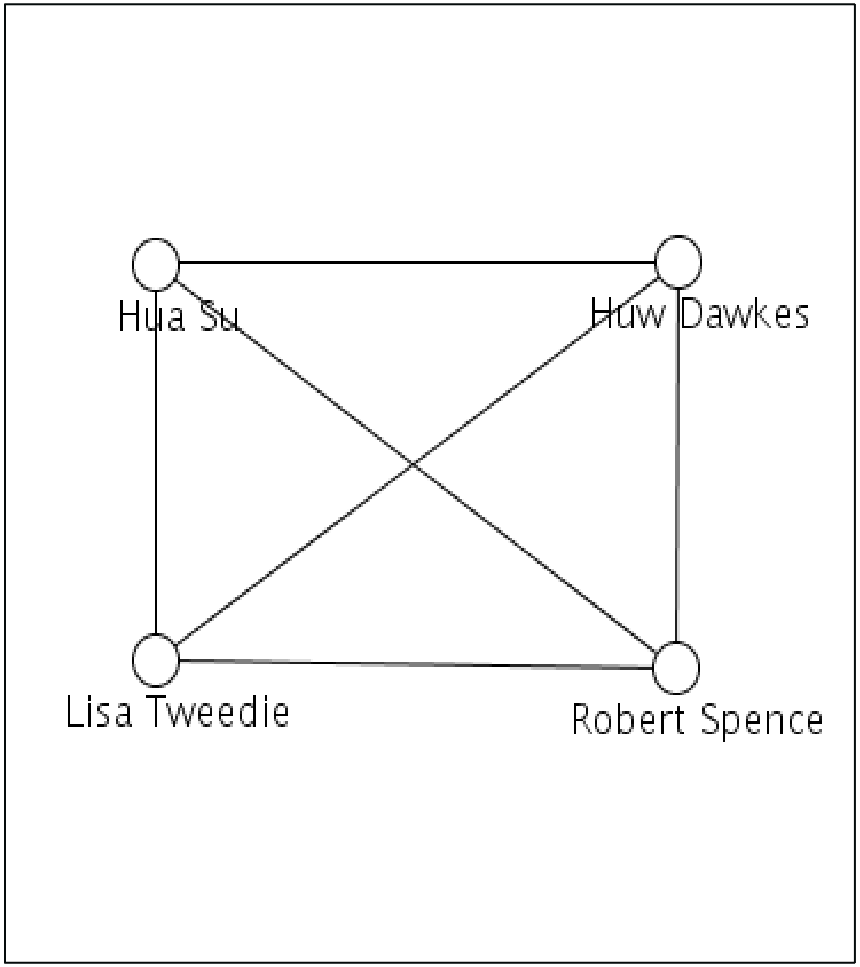
Algorithms

- **The Components**
  - #1: Entity Resolution
  - #2: Collective Classification
  - #3: Link Prediction
  - Putting It All Together
- Challenges and Opportunities

#1: Entity Resolution



before



after

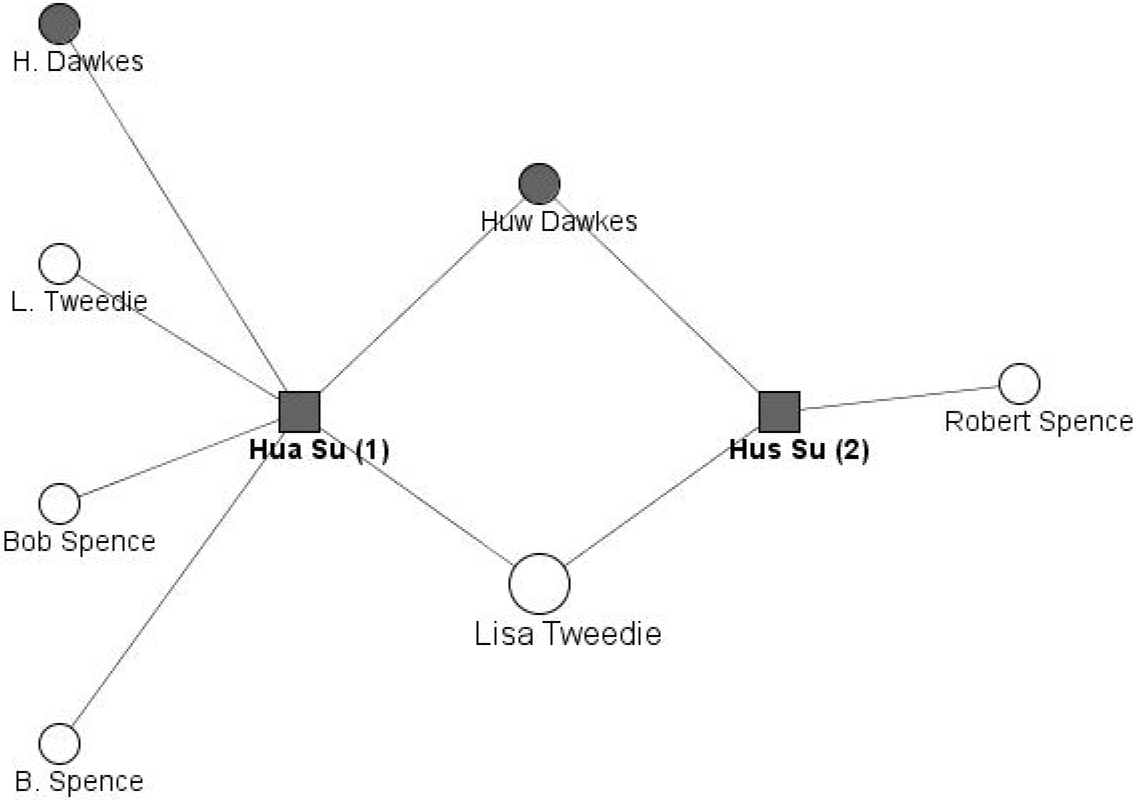
Relational Entity Resolution

- 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

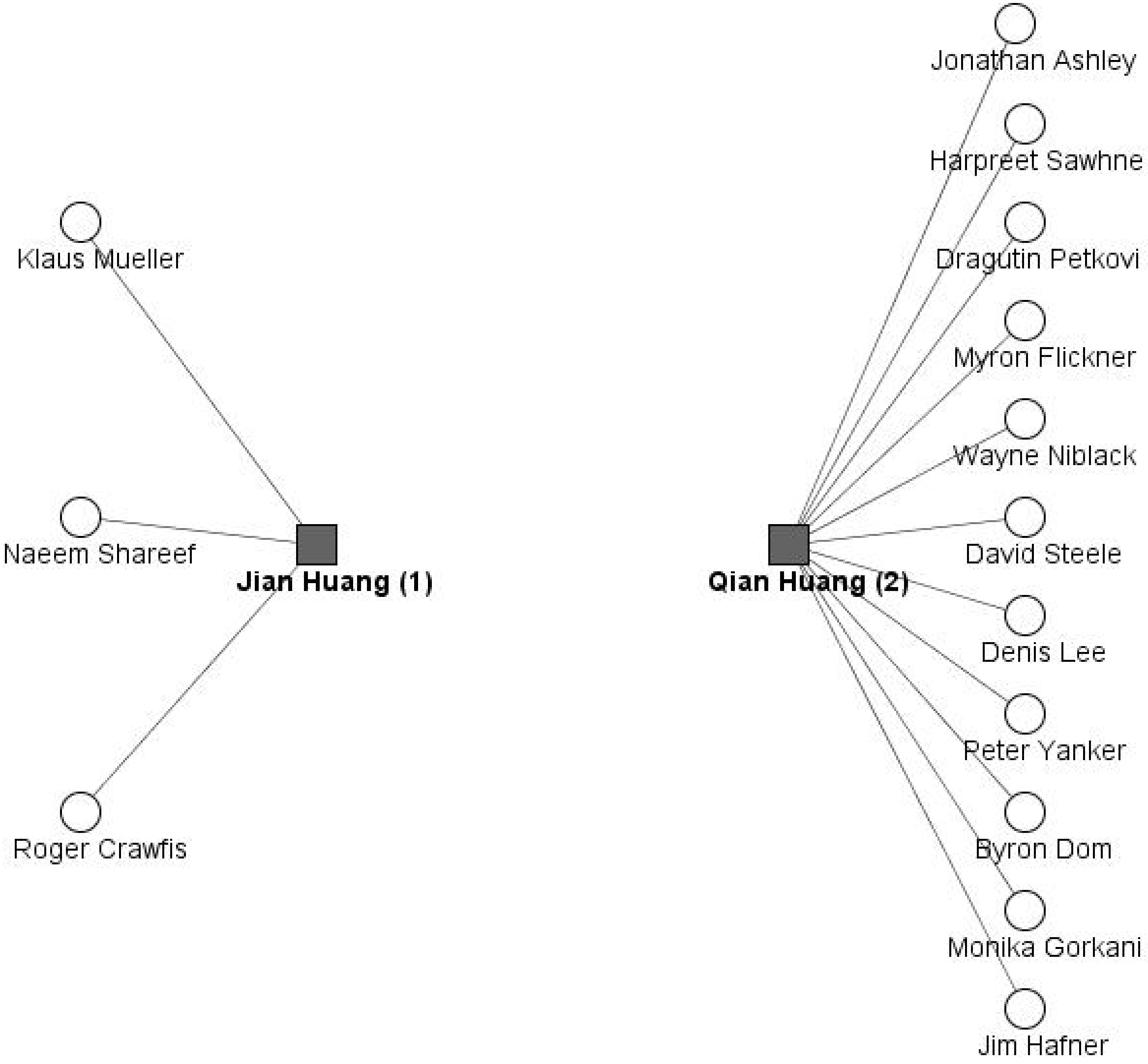
Relational Identification



Very similar names.

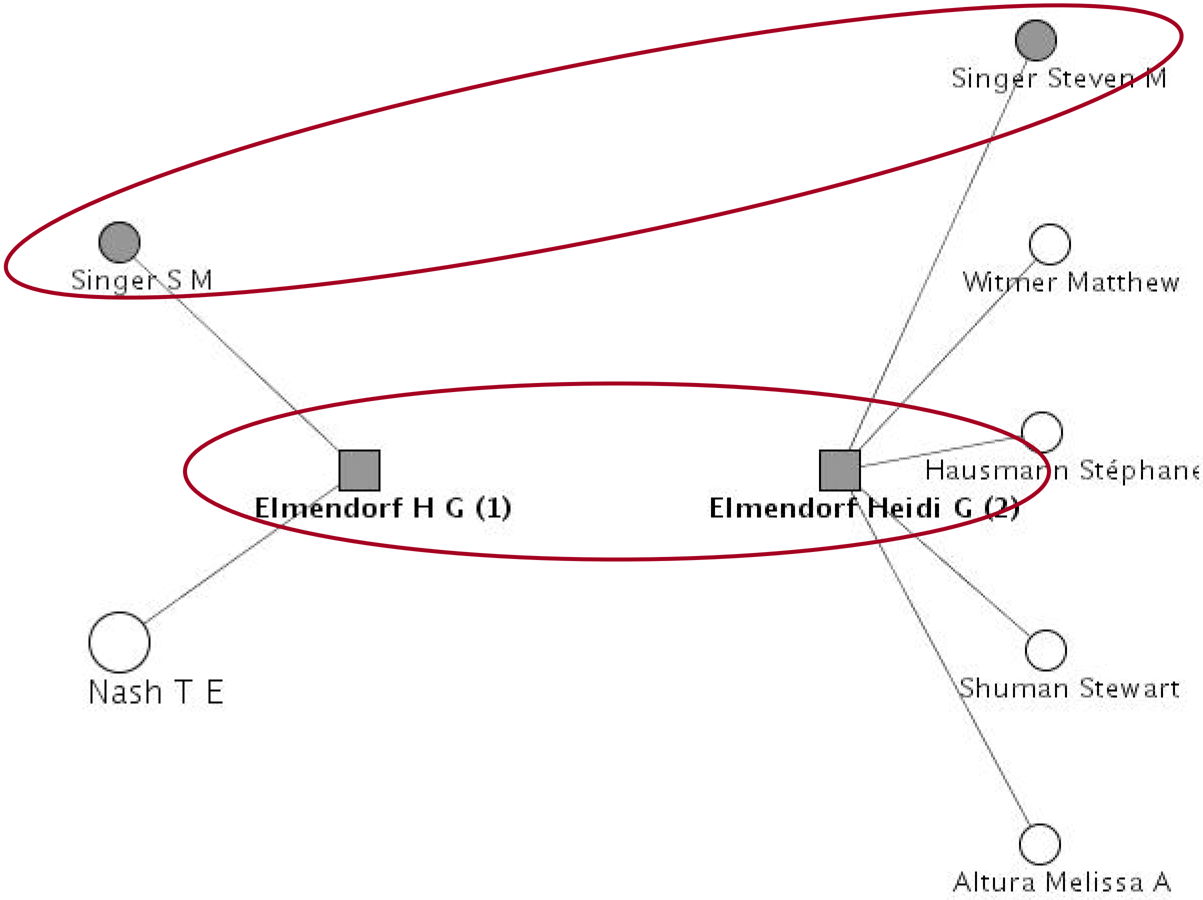
Added evidence from shared co-authors

Relational Disambiguation



Very similar names but no shared collaborators

Collective Entity Resolution



One resolution provides evidence for another => joint resolution

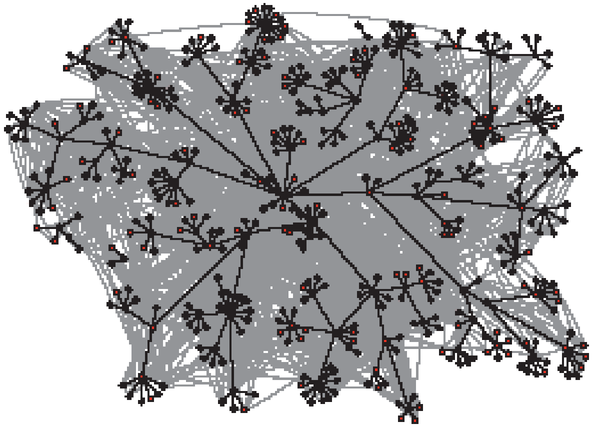
#2: Collective Classification

- 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

#3: Link Prediction: Links in Data Graph



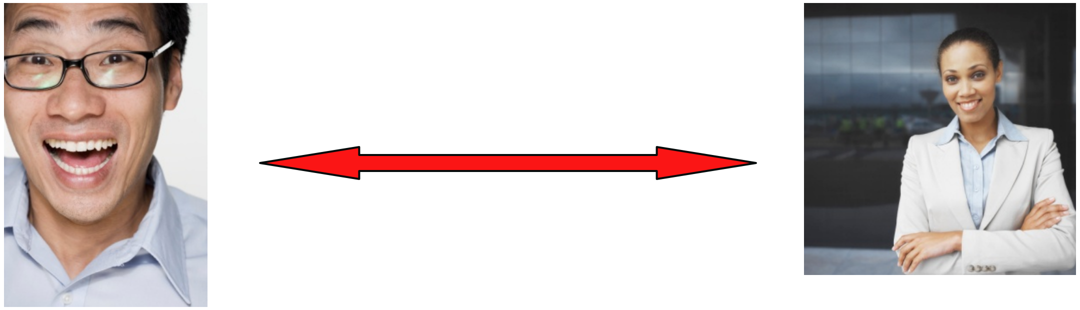
Communications Graph

Nodes: Network References

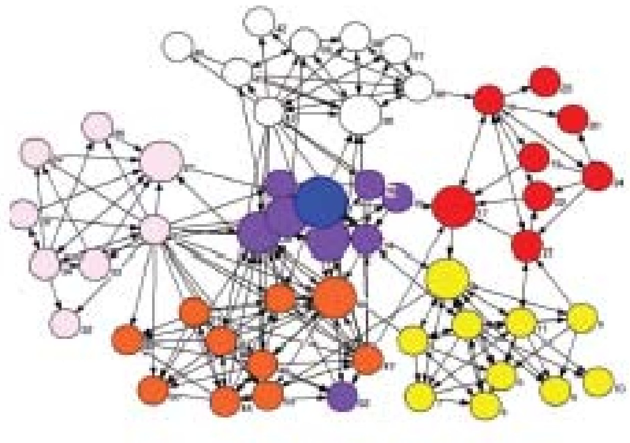
Edges: Communications Events

Node 1
Node 2

chris@enron.com	Email	liz@enron.com
chris37	IM	lizs22
555-450-0981	TXT	555-901-8812



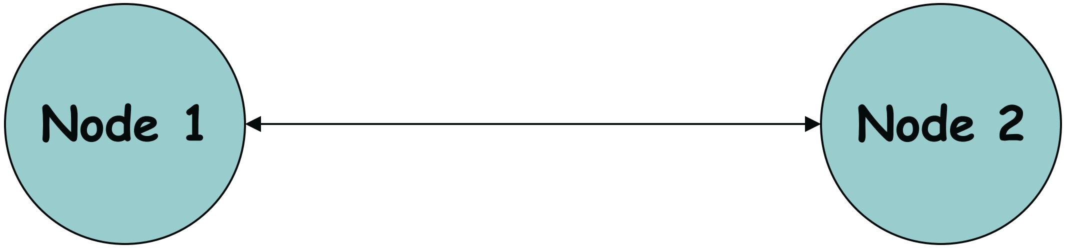
⇒ Links in Information Graph



Network Graph

Nodes: Entities

Edges: Social Relationships



Chris

Manager

Elizabeth

Steve

Family

Tim

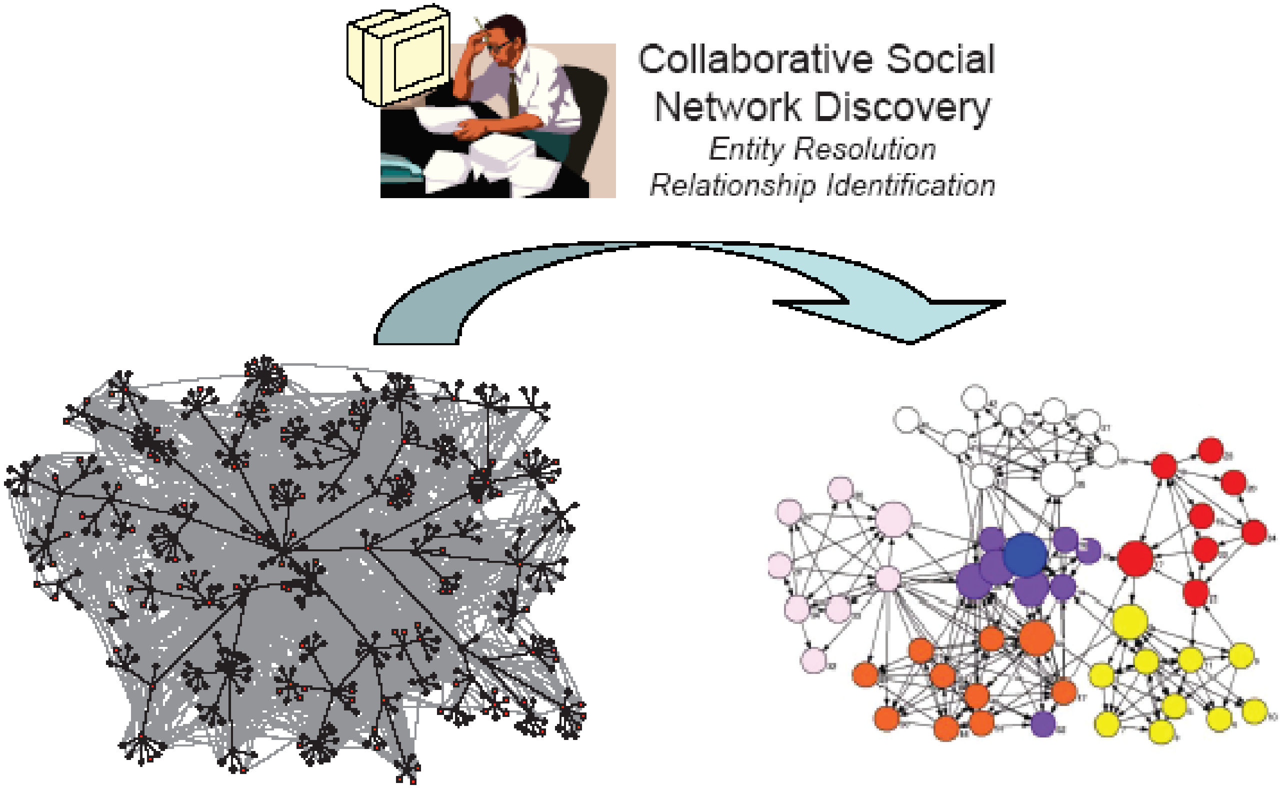
Algorithm Foundations

- Directed Models
  - Directed Graphical Models (aka Bayesian Networks)
    - Inference Algorithms:
      - Loopy Belief Propagation
      - Markov Chain Monte Carlo
  - Collection of Local Conditional Models
    - Inference Algorithms:
      - Iterative Classification Algorithm
      - Gibbs Sampling
- Undirected Models
  - (Pairwise) Markov Random Fields
    - Inference Algorithms:
      - Loopy Belief Propagation
      - Gibbs Sampling
      - Mean Field Relaxation Labeling

Algorithms

- The Components
  - Entity Resolution
  - Collective Classification
  - Link Prediction
  - **Putting It All Together**
- Challenges and Opportunities

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

Algorithms

- The Components
  - Entity Resolution
  - Collective Classification
  - Link Prediction
  - Putting It All Together
- **Challenges and Opportunities**

Challenges

- Graph/Network Data
  - Irregular Structure - not a regular grid, not fixed degree
  - Heterogeneity – different node types, relationships, etc.
  - Graph statistics – betweeneness, clique finding, subgraph isomorphism
- Inference Algorithms
  - Iterative, approximate, understanding sensitivity and robustness
- Scaling - streaming, dynamic data
- Maintaining Lineage - both data and inferences
- Access Control - privacy, security, collaboration

Opportunities for HPA

- Exploit irregularity and heterogeneity

- Approximations => fault tolerance
  - Xuanhua Li & Donald Yeung, *Application-level Correctness and its impact of Fault Tolerance*, Proceedings of the 18th International Symposium on High-Performance Computer Architectures, 2007.s

- Limited/flexible need for synchronization

- **Dirty data + approximate algorithms => great HPA opportunities!**

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

Thanks!

http://www.cs.umd.edu/linqs

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