Collaborative Social Network Discovery from Online Communications

Chris Diehl

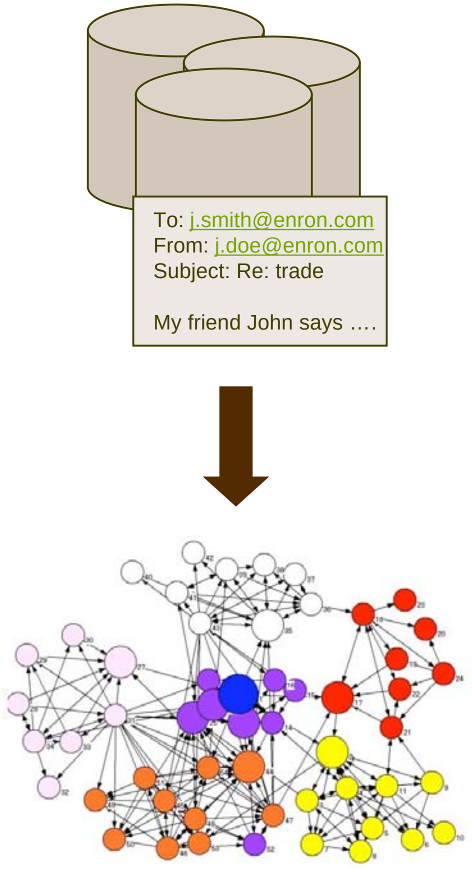
USMA-ARI Network Science Workshop

Collaboration with Lise Getoor and Galileo Namata, University of Maryland – College Park

APL
The Johns Hopkins University
APPLIED PHYSICS LABORATORY

The Question

- **Organizations today utilize a number of communication channels**
  - Email, Instant Messaging, Text Messaging, Wikis, Blogs
- **Given access to an organization's online communications, how does one infer relationship and role types within the organization from the data?**



To: j.smith@enron.com
From: j.doe@enron.com
Subject: Re: trade

My friend John says ….

Data Attributes

- ***Structured Data (Metadata)***
  - Sender and recipient(s), datetime
  - Can identify patterns of communication from metadata
  - Metadata provides no relationship context
- ***Unstructured Data (Content)***
  - Message subject and body, attachments
  - Content may provide relationship and role information
  - Additional context may be needed to clarify the message
- ***Goal is to exploit complimentary cues offered by the metadata and content***

Identifying Key Actors – A Motivating Example

From: Jennifer Fraser

Subject: john arnold bid for 20,000?

true? and when do you plan on selling them?

From: John Arnold

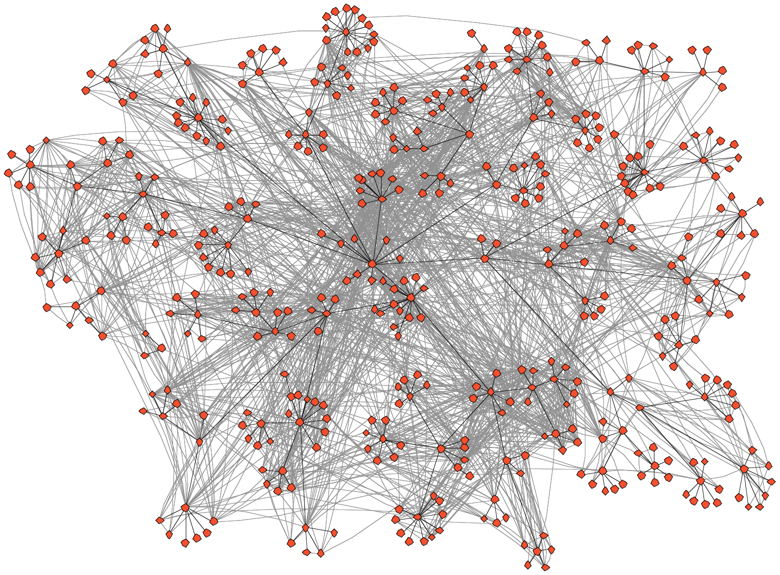
exaggerations...word travels everywhere doesnt it? how'd you hear?

From: Jennifer Fraser

johnny johhny johnny-- ***there is no secrecy when one is the king of ng*** **.. your brokers have the biggest moves in the world…**

Representations: Data and Network

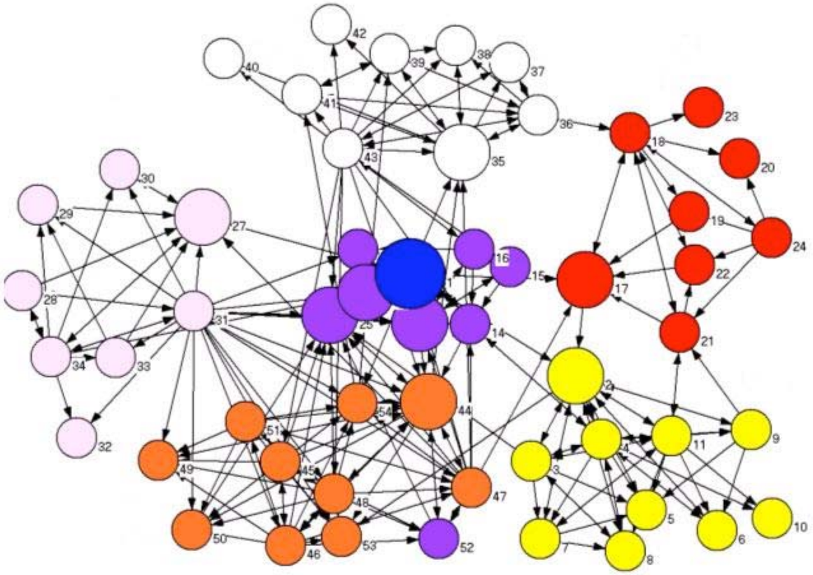
Communication (Hyper)Graph



HP Labs Communication Graph
(Adamic and Adar, 2003)

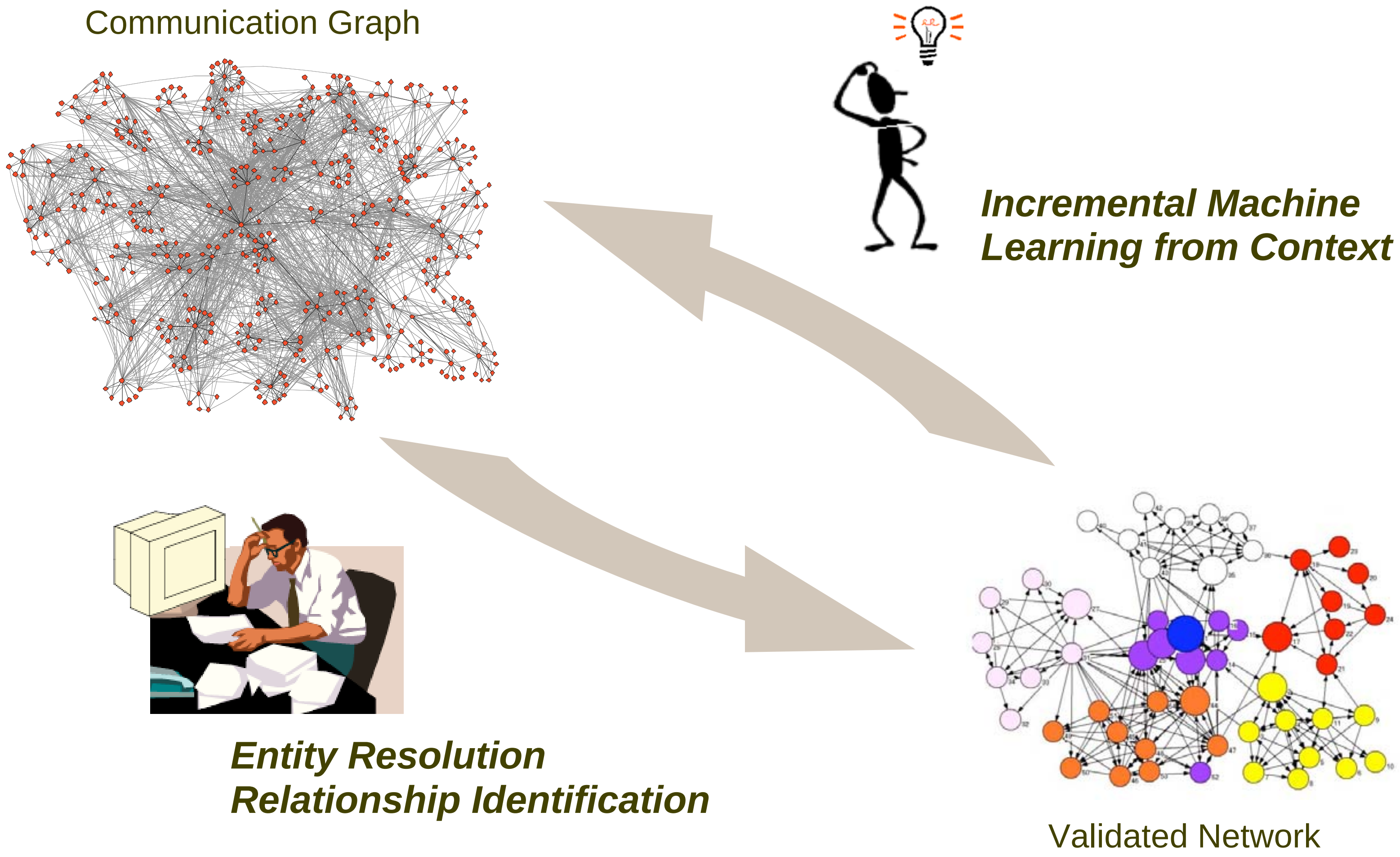
Nodes: Network References
Edges: Communication Events

Network (Hyper)Graph



Nodes: Entities
Edges: Social Relationships

Collaborative Social Network Discovery



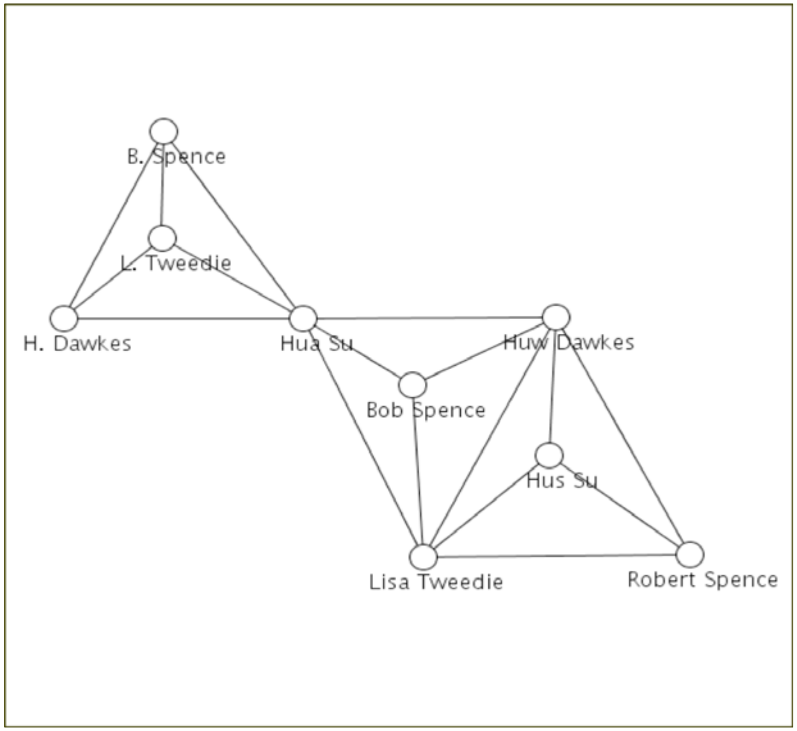
Communication Graph

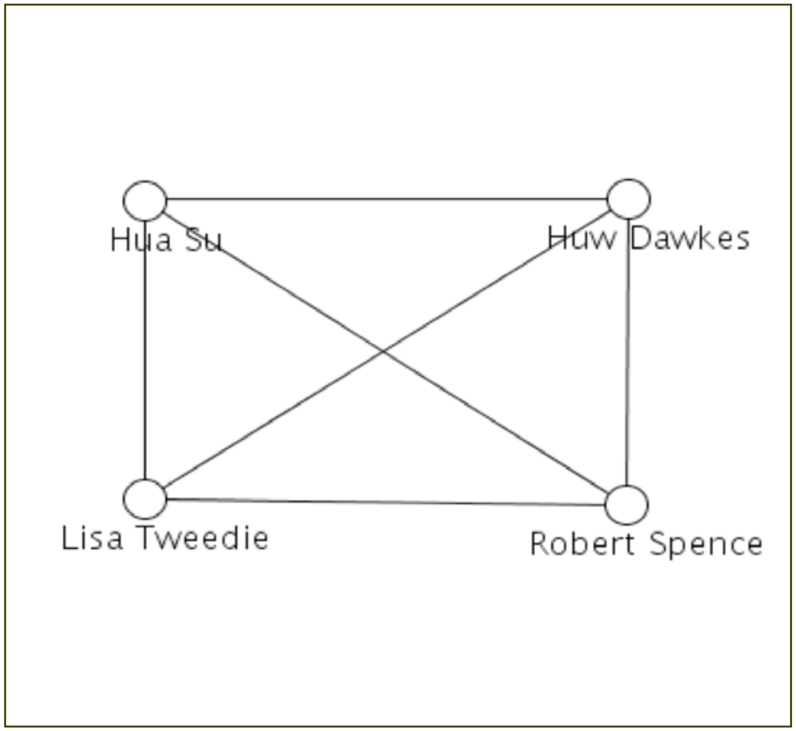
Incremental Machine Learning from Context

Entity Resolution
Relationship Identification

Validated Network

Entity Resolution:
InfoVis Co-Author Network Fragment

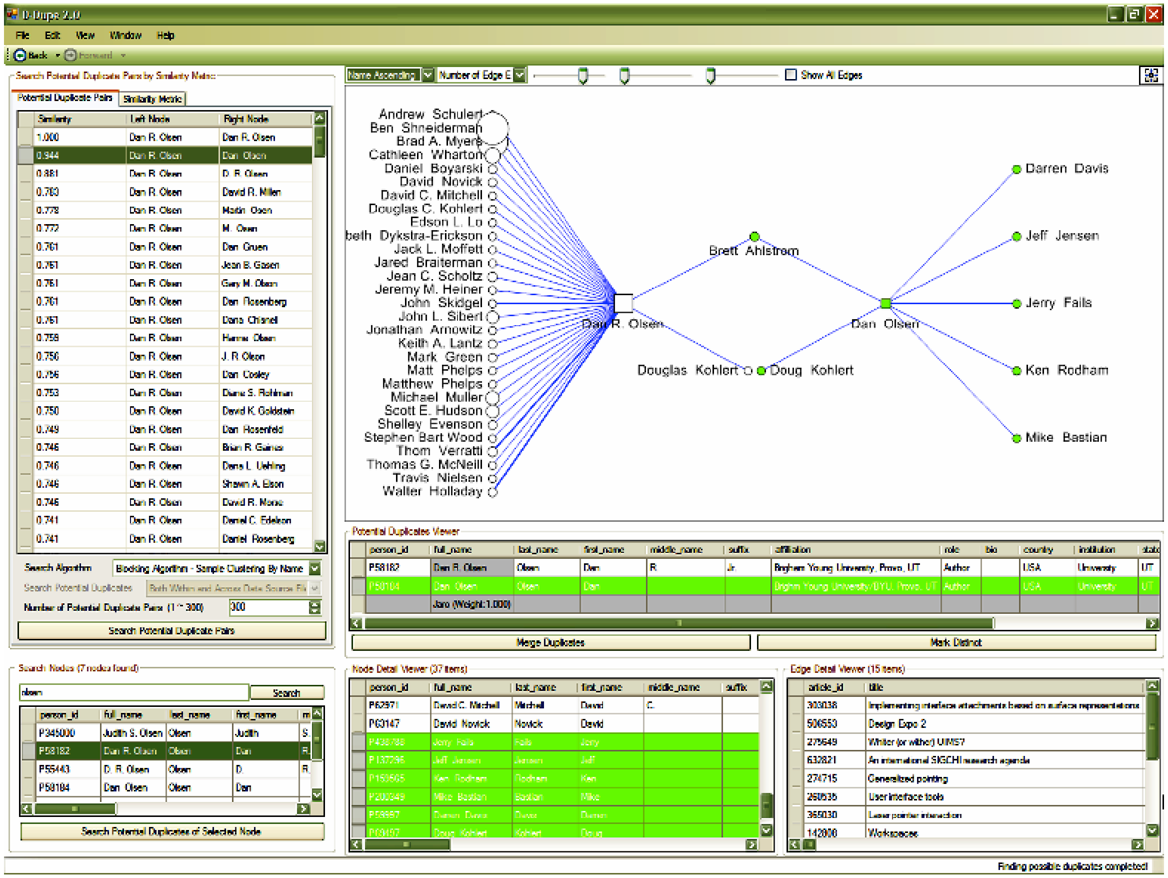




Before

After

D-Dupe: An Interactive Tool for Entity Resolution



Search Potential Duplicate Pairs by Similarity Metric

Potential Duplicate Pairs | Similarity Metric

Similarity	Left Node	Right Node
1.000	Dan R. Olsen	Dan R. Olsen
0.944	Dan R. Olsen	Dan Olsen
0.881	Dan R. Olsen	D. R. Olsen
0.783	Dan R. Olsen	David R. Millen
0.778	Dan R. Olsen	Martin Olsen
0.772	Dan R. Olsen	M. Olsen
0.761	Dan R. Olsen	Dan Gruen
0.761	Dan R. Olsen	Jean B. Gasen
0.761	Dan R. Olsen	Gary M. Olson
0.761	Dan R. Olsen	Dan Rosenberg
0.761	Dan R. Olsen	Dana Chisnell
0.758	Dan R. Olsen	Hanna Olsen
0.756	Dan R. Olsen	J. R. Olson
0.756	Dan R. Olsen	Dan Cosley
0.753	Dan R. Olsen	Diane S. Rohlman
0.750	Dan R. Olsen	David K. Goldstein
0.749	Dan R. Olsen	Dan Rosenfeld
0.746	Dan R. Olsen	Brian R. Gaines
0.746	Dan R. Olsen	Dana L. Uehling
0.746	Dan R. Olsen	Shawn A. Elson
0.746	Dan R. Olsen	David R. Morse
0.741	Dan R. Olsen	Daniel C. Edelson
0.741	Dan R. Olsen	Daniel Rosenberg

Search Algorithm: Blocking Algorithm - Sample Clustering By Name

Search Potential Duplicates: Both Within and Across Data Source Fil

Number of Potential Duplicate Pairs (1 ~ 300): 300

Search Potential Duplicate Pairs

Search Nodes (7 nodes found)

olsen Search

person_id	full_name	last_name	first_name	m
P345000	Judith S. Olsen	Olsen	Judith	S.
P58182	Dan R. Olsen	Olsen	Dan	R.
P55443	D. R. Olsen	Olsen	D.	R.
P58184	Dan Olsen	Olsen	Dan	

Search Potential Duplicates of Selected Node

Potential Duplicates Viewer

person_id	full_name	last_name	first_name	middle_name	suffix	affiliation	role	bio	country	institution	state
P58182	Dan R. Olsen	Olsen	Dan	R.	Jr.	Brigham Young University, Provo, UT	Author		USA	University	UT
P58184	Dan Olsen	Olsen	Dan			Brighm Young University/BYU, Provo, UT	Author		USA	University	UT
	Jaro (Weight:1.000)										

Merge Duplicates | Mark Distinct

Node Detail Viewer (37 items)

person_id	full_name	last_name	first_name	middle_name	suffix
P62971	David C. Mitchell	Mitchell	David	C.	
P63147	David Novick	Novick	David		
P438788	Jerry Fails	Fails	Jerry		
P137296	Jeff Jensen	Jensen	Jeff		
P159565	Ken Rodham	Rodham	Ken		
P200349	Mike Bastian	Bastian	Mike		
P59997	Darren Davis	Davis	Darren		
P69497	Doug Kohlert	Kohlert	Doug		

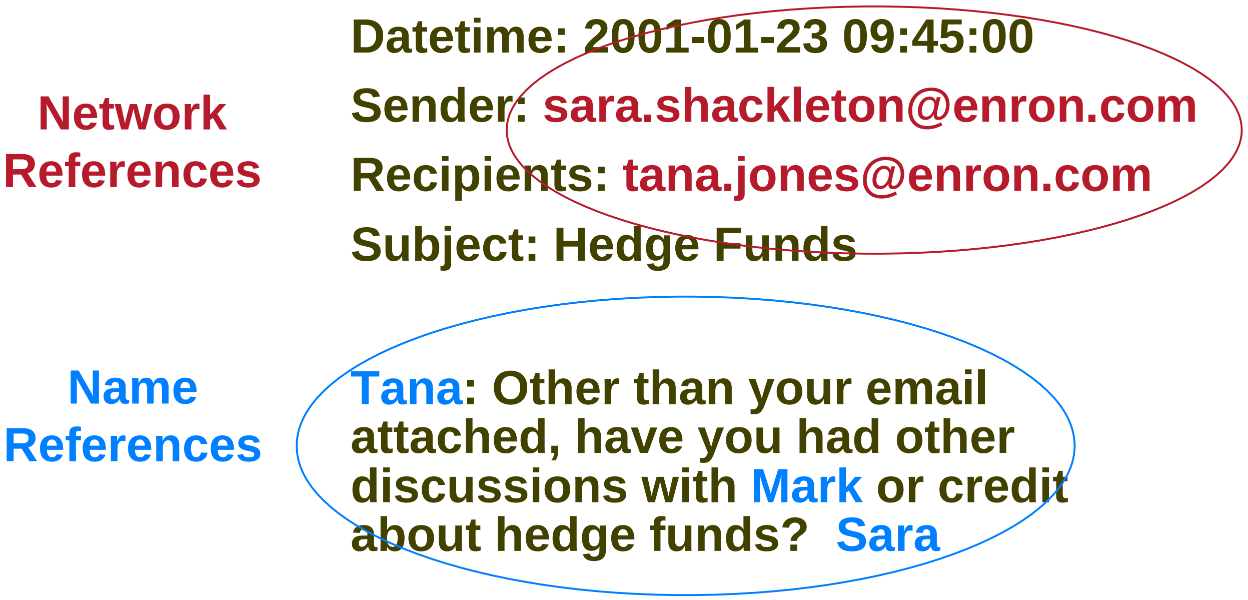
Edge Detail Viewer (15 items)

article_id	title
303038	Implementing interface attachments based on surface representations
506553	Design Expo 2
275649	Whiter (or wither) UIMS?
632821	An international SIGCHI research agenda
274715	Generalized pointing
260535	User interface tools
365030	Laser pointer interaction
142000	Workspaces

Finding possible duplicates completed!

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

Entity Resolution:
Name and Network References



Network References

Datetime: 2001-01-23 09:45:00

Sender: sara.shackleton@enron.com

Recipients: tana.jones@enron.com

Subject: Hedge Funds

Name References

Tana: Other than your email attached, have you had other discussions with Mark or credit about hedge funds? Sara

- Every individual has two classes of references
- To define an individual’s identity and draw broader connections across emails, we need to first associate name and network references

Reference: C. P. Diehl, L. Getoor, G. Namata, "Name Reference Resolution in Organizational Email Archives," SIAM Data Mining 2006

Context Challenges

Datetime: 2000-06-19 09:52:00

Sender: tana.jones@enron.com

Recipients: marie.heard@enron.com

Subject: Just a tease!!!

Wouldn t you like to know which of the two Susan s gave her notice today

Datetime: 2001-02-28 09:32:00

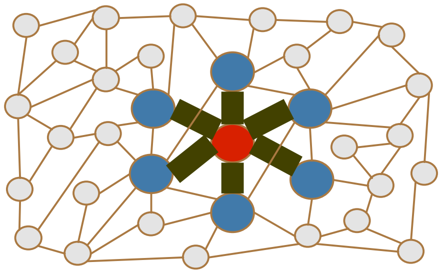
Sender: liz.taylor@enron.com

Recipients: john.arnold@enron.com

Subject: Greg s Bill

Johnny, What does Greg owe you for the champagne? Is it $896.00? Liz

Relationship Identification - Incremental Ego Network Exploration



Evidence Discovery

From: Christian Yoder [christian.yoder@enron.com]

To: Elizabeth Sager [elizabeth.sager@enron.com], Genia Fitzgerald [genia.fitzgerald@enron.com]

Subject: Happiness

Happiness is looking at the new legal org chart (which Jan just now dropped on my desk). I always approach these dry documents as though they were trigrams resulting from throwing the coins and consulting the I-Ching. At the top of the trigram which I find myself listed in I see a single name: Elizabeth Sager, and at the bottom I see the name Genia FitzGerald. ... cgy

Relationship Ranking

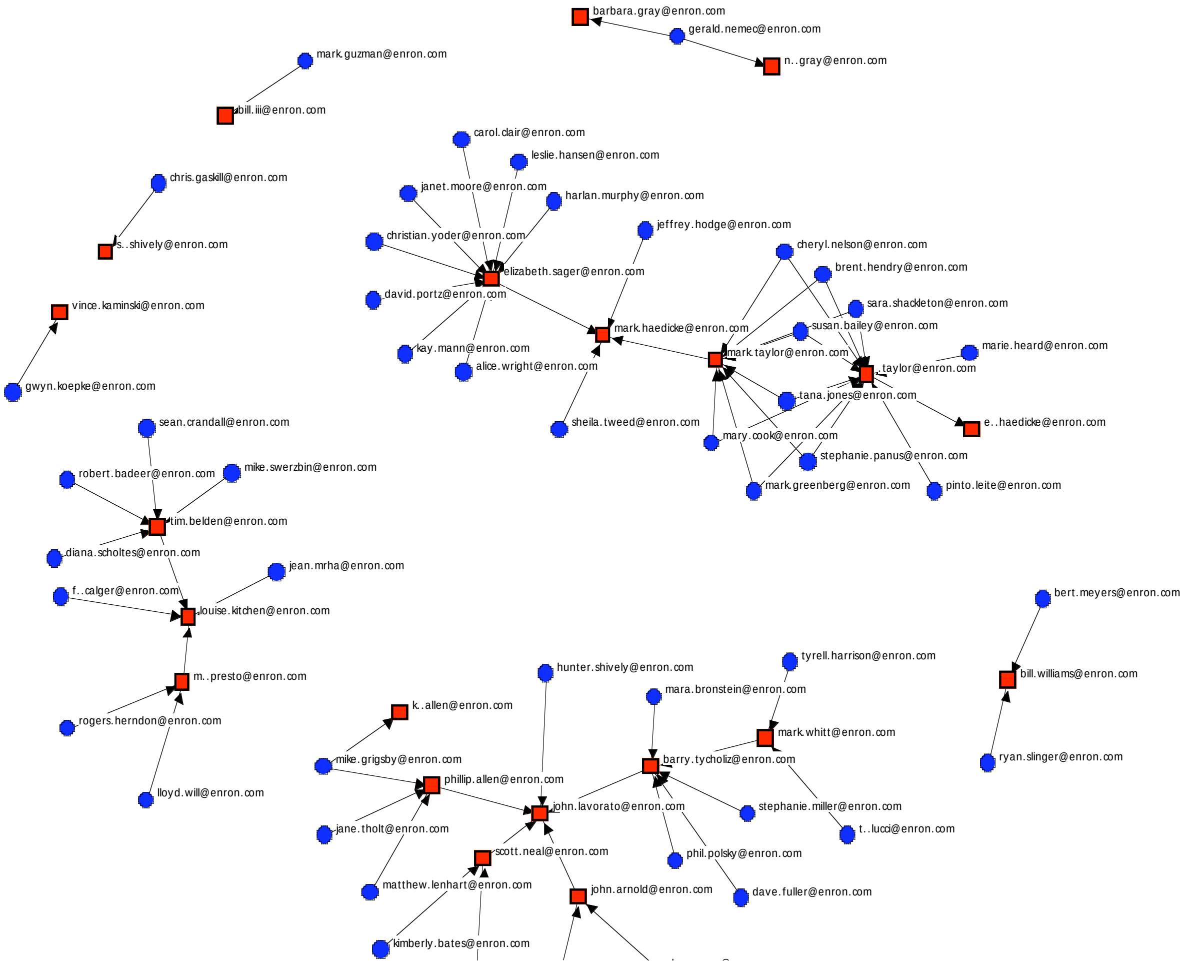
Rank	*Relationship with Ego (Christian Yoder)*
1	Elizabeth Sager
2	Richard Sanders
3	Steve Hall
4	Mark Haedicke
5	Dave Fuller
6	Tracy Ngo

Message Ranking

Rank	*Message Subject*
1	Happiness
2	System Outage Risk
3	Mark Taylor Visit
4	Question about a deal we did

Reference: C. P. Diehl, G. Namata, L. Getoor, ”Relationship Identification for Social Network Discovery," AAAI 2007

Enron Manager-Subordinate Communications Relationships



Relationship Identification - Manager-Subordinate Relations

- ***Preference Learning***
  - Supervised learning of relationship ranker
  - Given initial set of labeled ego networks
  - Ranking dyadic relationships
- ***Traffic-Based Approach***
  - Message frequency
  - Number of recipients
  - Exchanges between relationship participants and common recipients
- ***Content-Based Approach***
  - Term frequency vector for set of messages corresponding to the relationship
  - Exploits text from sender to recipient

Approach	***Mean Reciprocal Rank***
Content-Based with Attribute Selection	**0.719**
Content-Based	**0.660**
Traffic-Based	**0.518**
Random Selection	**0.211**
Worst Case	**0.141**

Future Directions

- ***Incremental, Active Learning***
  - Relationship-Level and Message-Level Annotations
  - Automated Model Selection
  - Automated Feature Selection
- ***Visualization***
  - Communications Graph Exploration
  - Network Graph Construction
- ***Interaction Paradigms***
  - Unified Workflow for Entity Resolution and Relationship Identification