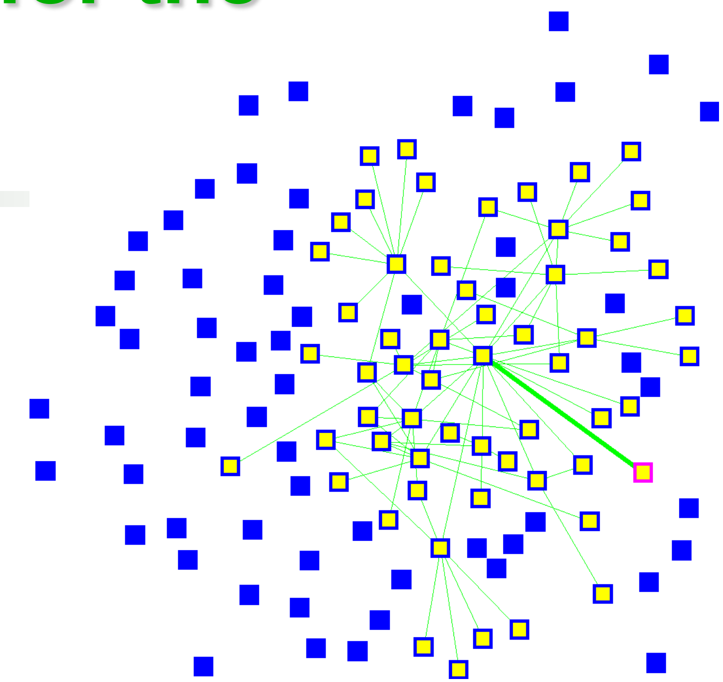
ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

UNIVERSITY OF ALBERTA

Social Network Analysis for the Assessment of Learning



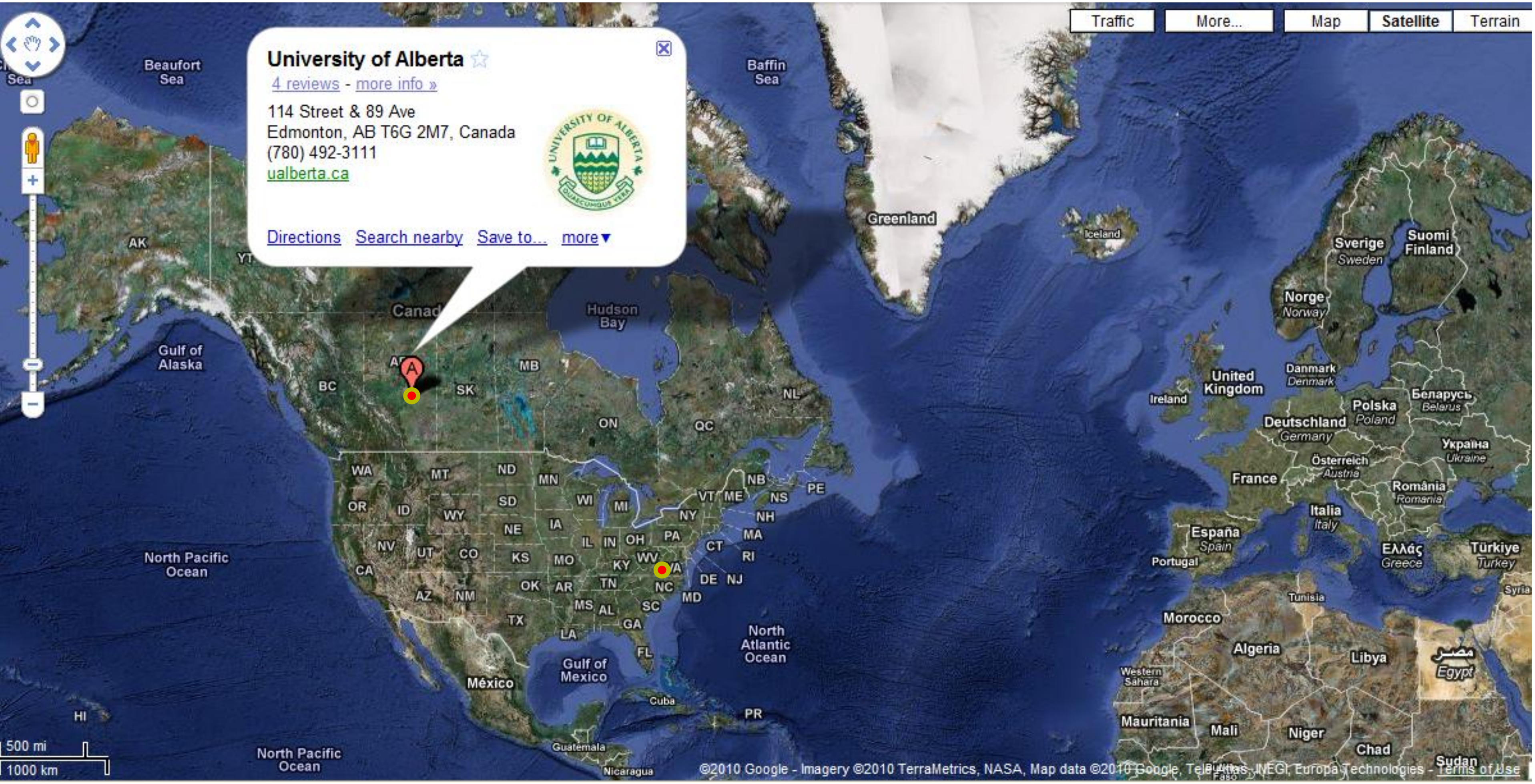
Osmar R. Zaïane

Professor & Scientific Director of AICML

Educational Data Mining 2010
Pittsburgh, USA

University of Alberta - Edmonton

<u>2,867.97 kilometres (1,782.08 miles)</u>



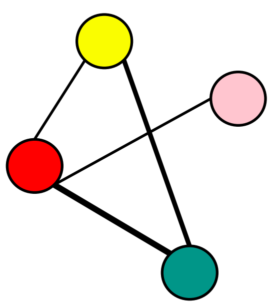
Edmonton, capital of Alberta, is the <u>5th largest city</u> in Canada with more than 1 million people.
The University of Alberta is the second largest university in the country in terms of research funding

Thank you to

- Jiyang Chen
- Justin Fagnan
- Reihaneh Rabbany
- Farzad Sangi
- Mansoureh Takaffoli

SNA vs Social Networking

Social Network Analysis Deals with Information Networks
It is NOT Social Networking



Nodes are entities
Edges are relationships

SNA = Analysing such information networks

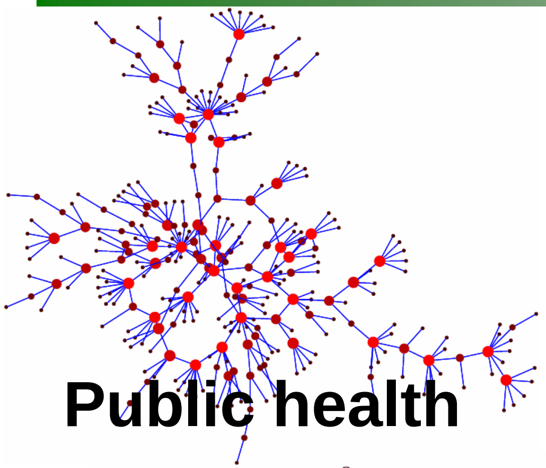
Roadmap

- Introduction of Social Network Analysis
- Some needs in understanding Educational Data
  - Interpreting a student communication network
  - Finding groups/communities
  - Finding discussion topics
  - Understanding dynamics
- Needs in EDM lead to contributions in Data Mining
  - Community Mining and Validation
  - Global versus Local Community Mining
  - Branching to other interesting applications
- Conclusion

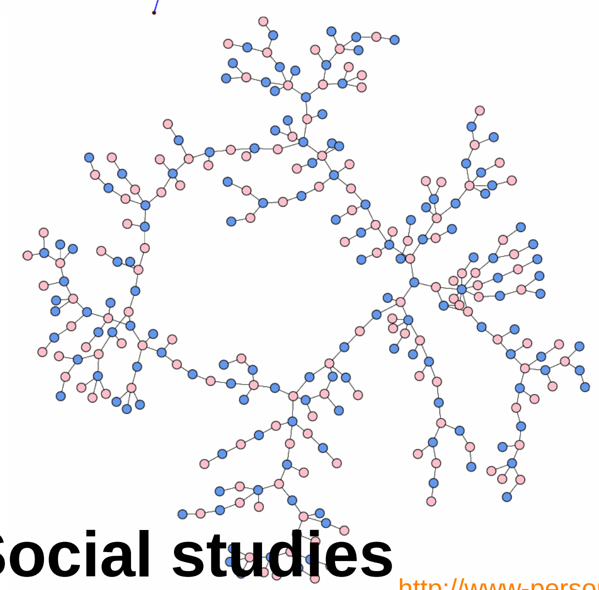
What is Social Network Analysis?

- [Wikipedia] A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friendship, sexual relationships, kinship, dislike, conflict or trade.
- Social Network Analysis (SNA) is the study of social networks to understand their structure and behaviour.
- Which node is the most influential? which one is central? What are the hubs? What are the groups? Who knows who?, What are the short paths? What is perceived by who? ...

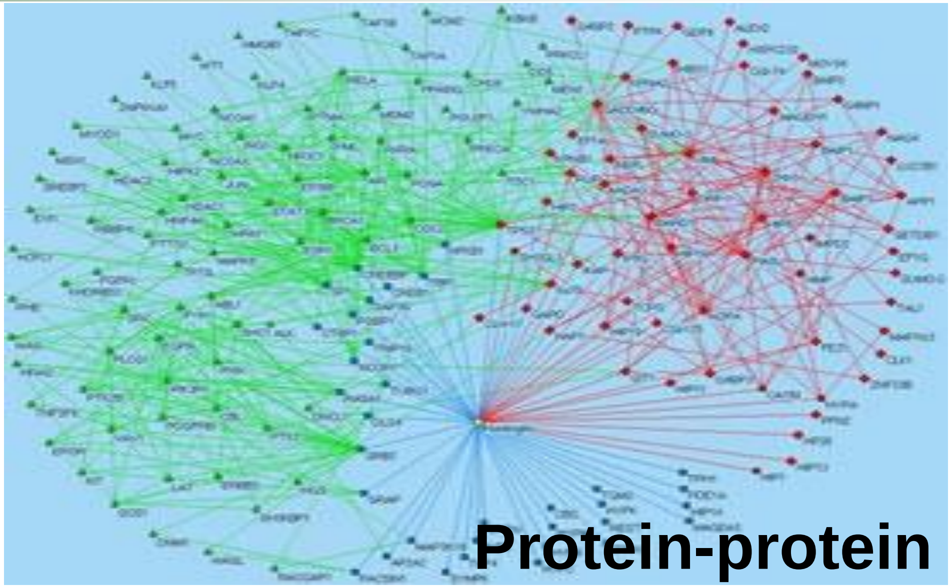
SNA, A Multidisciplinary Field



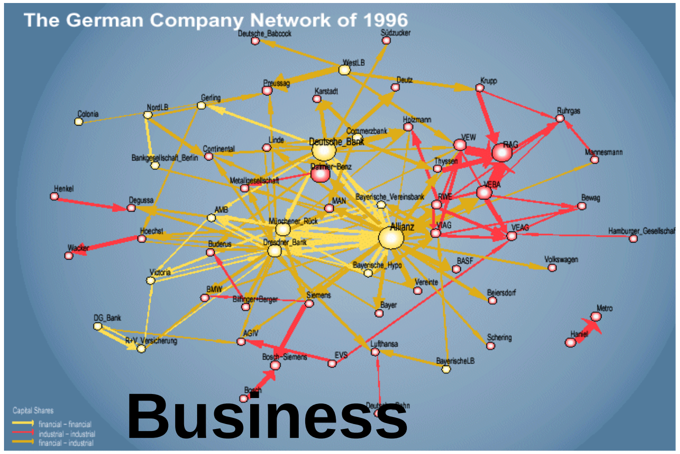
Public health



Social studies



Protein-protein



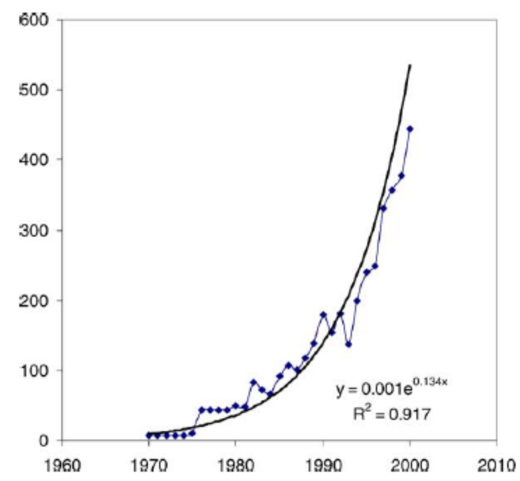
Business

http://www-personal.umich.edu/~mejn/networks

A quick History

- Social network analysis is a key technique traditionally studied in sociology, anthropology, epidemiology, sociolinguistics, psychology, etc. Today it is a modern technique in marketing, economics, intelligence gathering, criminology, medicine, computer science, etc.
- J. Barnes is credited with coining the notion of social networks (theory) in 1954.
- Precursors of social network theory date from the 19th century such as Simmel, Durkheim and Tönnies.
- Massive increase in studies of social networks (in social sciences) since the 1970s.
- The increase of available data, the Internet phenomenon, Web 2.0, etc. have only catapulted the interest in SNA research

S.P. Borgatti, P.C. Foster / Journal of Management 2003 29(6) 991–1013

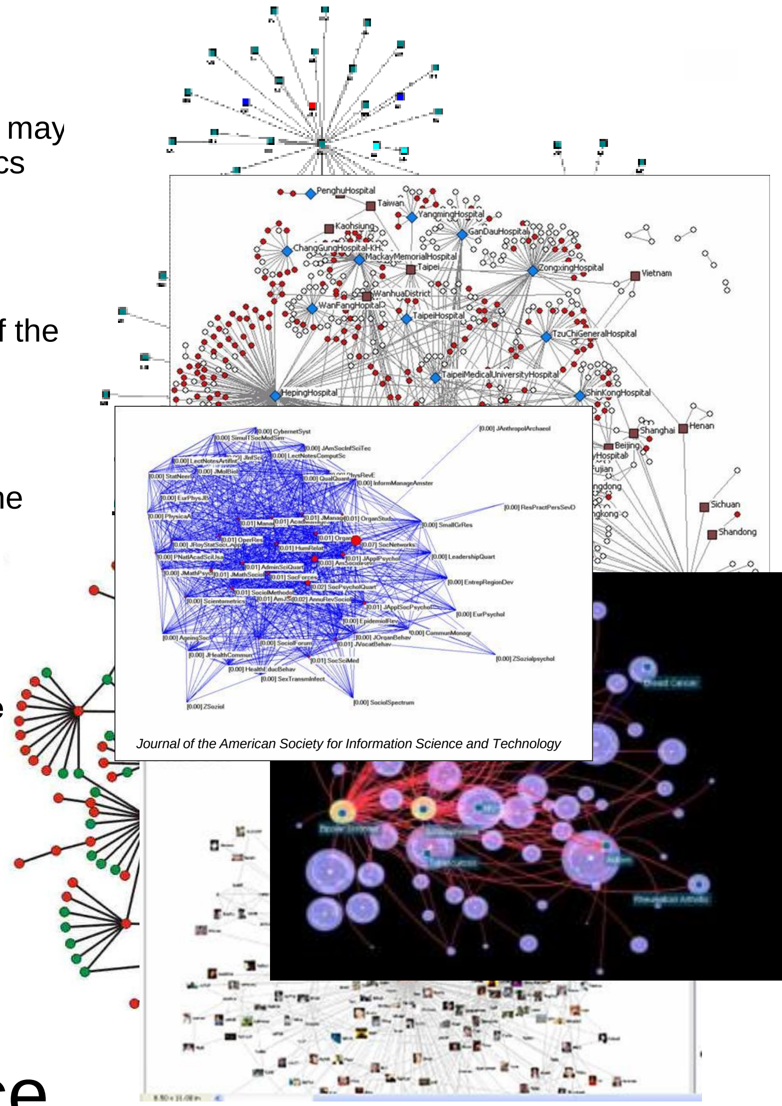


Some Key Concepts

- Edge Weight : interaction frequency, importance of information exchange, intimacy, emotional intensity, etc.
- Symmetric relation or not (directional)
- Centrality: determines the relative importance of a vertex (or edge) within a network.
  - Degree Centrality: Mesures the normalized number of edges incident upon a node *n;*
  - Betweeness Centrality: Measures how many times a node *n* occurs in a shortest path between any other 2 nodes in the graph;
  - Closeness Centrality: Mean shortest path distance between a node *n* and all other nodes reacheable from it;
  - Eigenvector Centrality: Measures importance of a node n by assigning a score to each node based on the principal that connections to high-scoring nodes contribute more to the score of a node in question than equal connections to low-scoring nodes (e.g. PageRank).

Applications of Social Network Analysis

- Terrorism and crimes
  - Social Network analysis is an important part of a conspiracy investigation and is used as an investigative tool. Group structure may be important to investigations of racketeering enterprises, narcotics operations, illegal gambling, and business frauds.
- Medicine – epidemiology
  - valuable epidemiological tool for understanding the progression of the spread of an infectious disease.
- Marketing
  - Emarketer projected that Social Network Marketing spending in the USA will reach approximately $1.3 billion in 2009. http://www.emarketer.com/Reports/All/Emarketer_2000541.aspx
- Product Recommendation
  - Current recommendation models assume all users' opinions to be independent. Use of SNA relaxes the iid assumption.
- Bio-informatics (protein interaction)
- Relevance Ranking
- Information and Library Science

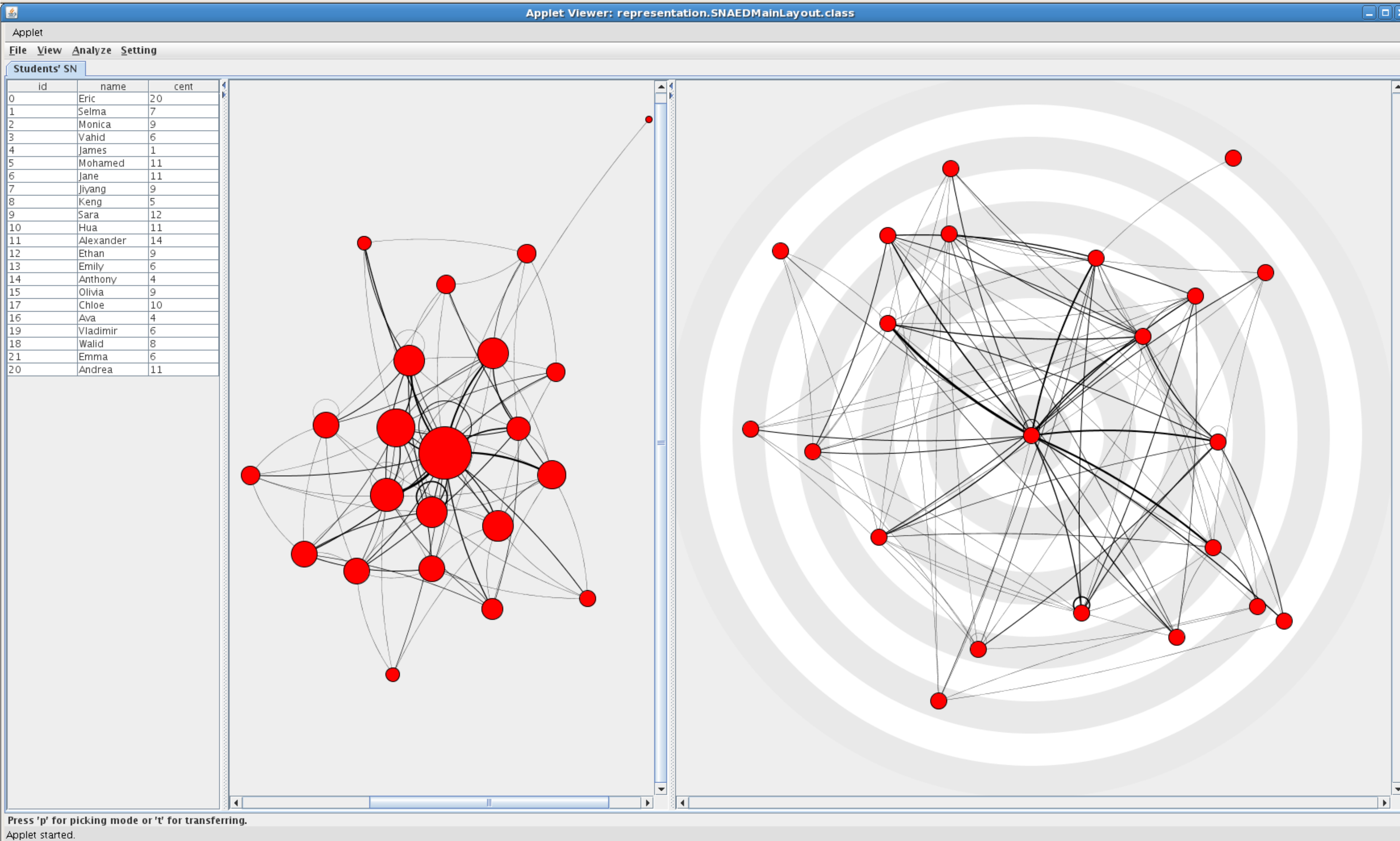


Prominent problems in SNA

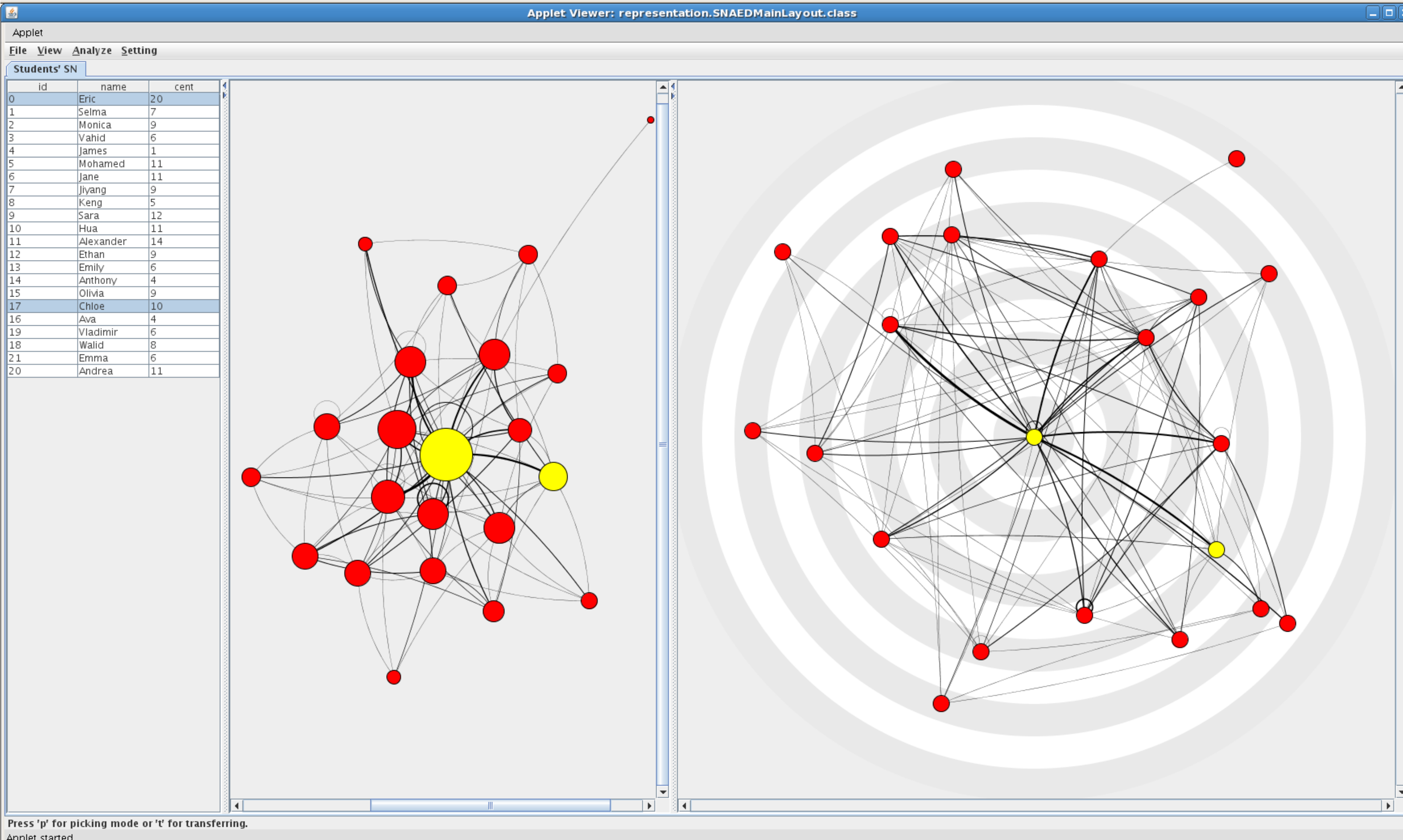
- Social network extraction/construction
- Link prediction
- Approximating large social networks
- Identifying prominent/trusted/expert actors in social networks
- Search in social networks
- Discovering communities in social networks
- Knowledge discovery from social networks
- Finding patterns in dynamic networks
- Predicting evolution

Analogy with Clustering

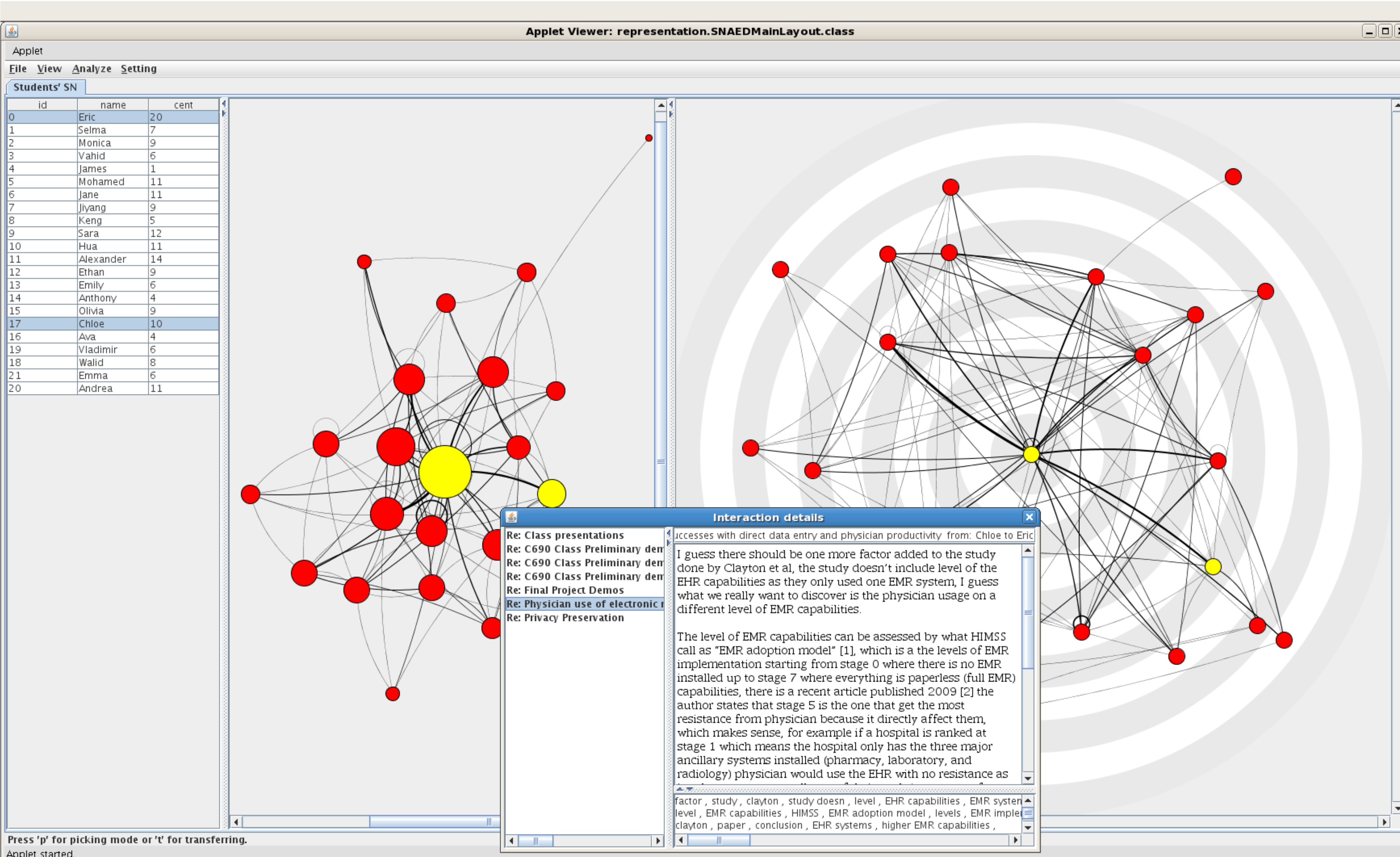
Meerkat-ED: Student Network



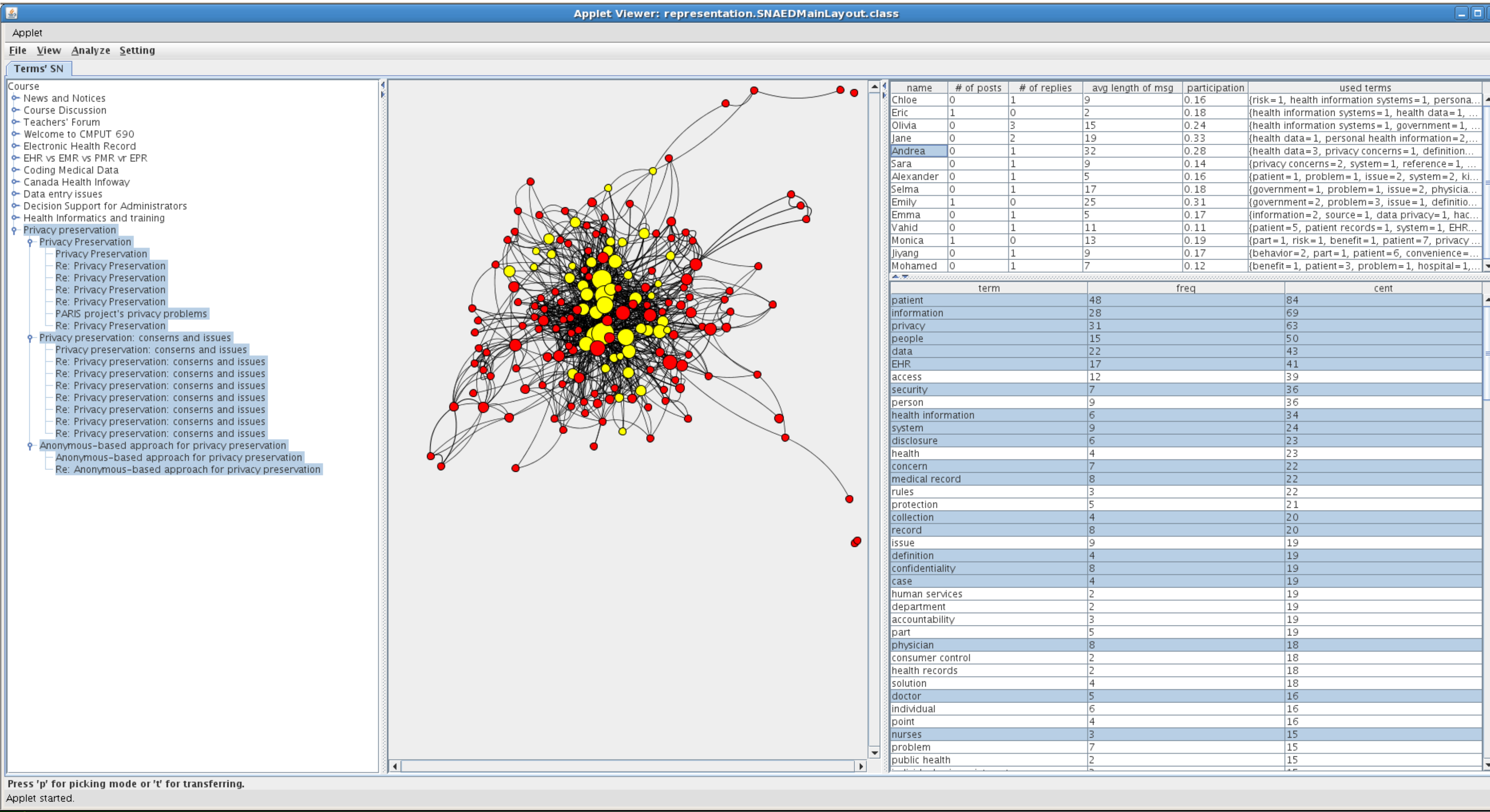
Meerkat-ED: Student Network



Meerkat-ED: Student Network



Meerkat-ED: Term Network



Meerkat-ED: Term Network

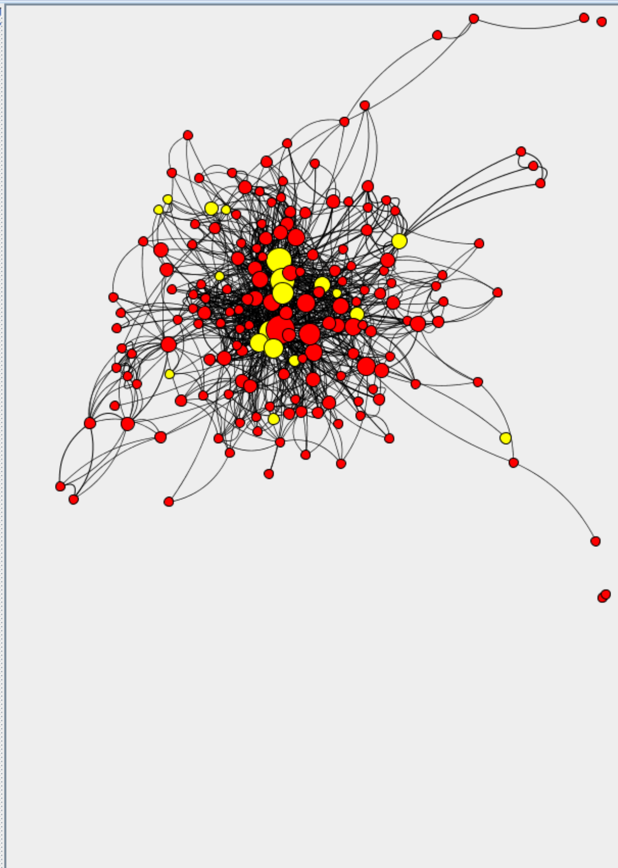
Applet Viewer: representation.SNAEDMainLayout.class

Applet

File View Analyze Setting

Terms' SN

- Course
  - News and Notices
  - Course Discussion
  - Teachers' Forum
  - Welcome to CMPUT 690
  - Electronic Health Record
  - EHR vs EMR vs PMR vr EPR
  - Coding Medical Data
  - Canada Health Infoway
  - Data entry issues
  - Decision Support for Administrators
  - Health Informatics and training
  - Privacy preservation
    - Privacy Preservation
      - Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - PARIS project's privacy problems
      - Re: Privacy Preservation
    - Privacy preservation: consens and issues
      - Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
      - Re: Privacy preservation: consens and issues
    - Anonymous-based approach for privacy preservation
      - Anonymous-based approach for privacy preservation
      - Re: Anonymous-based approach for privacy preservation



name	# of posts	# of replies	avg length of msg	participation	used terms
Chloe	0	1	9	0.16	{risk=1, health information systems=1, persona...
Eric	1	0	2	0.18	{health information systems=1, health data=1, ...
Olivia	0	3	15	0.24	{health information systems=1, government=1, ...
Jane	0	2	19	0.33	{health data=1, personal health information=2,...
Andrea	0	1	32	0.28	{health data=3, privacy concerns=1, definition...
Sara	0	1	9	0.14	{privacy concerns=2, system=1, reference=1, ...
Alexander	0	1	5	0.16	{patient=1, problem=1, issue=2, system=2, ki...
Selma	0	1	17	0.18	{government=1, problem=1, issue=2, physicia...
Emily	1	0	25	0.31	{government=2, problem=3, issue=1, definitio...
Emma	0	1	5	0.17	{information=2, source=1, data privacy=1, hac...
Vahid	0	1	11	0.11	{patient=5, patient records=1, system=1, EHR...
Monica	1	0	13	0.19	{part=1, risk=1, benefit=1, patient=7, privacy ...
Jiyang	0	1	9	0.17	{behavior=2, part=1, patient=6, convenience=...
Mohamed	0	1	7	0.12	{benefit=1, patient=3, problem=1, hospital=1,...

term	freq	cent
patient	48	84
information	28	69
privacy	31	63
people	15	50
data	22	43
EHR	17	41
access	12	39
security	7	36
person	9	36
health information	6	34
system	9	24
disclosure	6	23
health	4	23
concern	7	22
medical record	8	22
rules	3	22
protection	5	21
collection	4	20
record	8	20
issue	9	19
definition	4	19
confidentiality	8	19
case	4	19
human services	2	19
department	2	19
accountability	3	19
part	5	19
physician	8	18
consumer control	2	18
health records	2	18
solution	4	18
doctor	5	16
individual	6	16
point	4	16
nurses	3	15
problem	7	15
public health	2	15

Press 'p' for picking mode or 't' for transferring.

Applet started.

Meerkat-ED: Topic Hierarchy

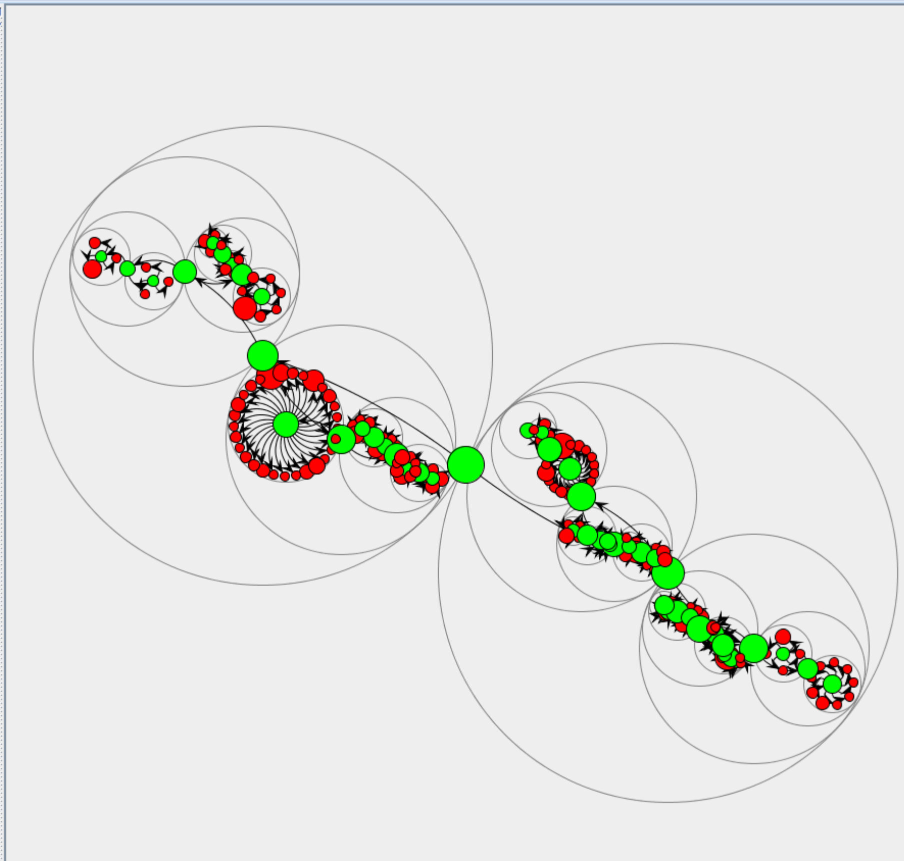
Applet Viewer: representation.SNAEDMainLayout.class

Applet

File View Analyze Setting

Terms' SN

- Course
  - News and Notices
  - Course Discussion
  - Teachers' Forum
  - Welcome to CMPUT 690
  - Electronic Health Record
  - EHR vs EMR vs PMR vr EPR
  - Coding Medical Data
  - Canada Health Infoway
  - Data entry issues
  - Decision Support for Administrators
  - Health Informatics and training
  - Privacy preservation
    - Privacy Preservation
      - Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - PARIS project's privacy problems
      - Re: Privacy Preservation
    - Privacy preservation: conserns and issues
      - Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
    - Anonymous-based approach for privacy preservation
      - Anonymous-based approach for privacy preservation
      - Re: Anonymous-based approach for privacy preservation



name	# of posts	# of replies	avg length of msg	participation	used terms
Chloe	0	1	9	0.16	{risk=1, health i...
Eric	1	0	2	0.18	{health informati...
Olivia	0	3	15	0.24	{health informati...
Jane	0	2	19	0.33	{health data=1,...
Andrea	0	1	32	0.28	{health data=3,...
Sara	0	1	9	0.14	{privacy concern...
Alexander	0	1	5	0.16	{patient=1, pro...
Selma	0	1	17	0.18	{government=1,...
Emily	1	0	25	0.31	{government=2,...
Emma	0	1	5	0.17	{information=2,...
Vahid	0	1	11	0.11	{patient=5, pati...
Monica	1	0	13	0.19	{part=1, risk=1...
Jiyang	0	1	9	0.17	{behavior=2, p...
Mohamed	0	1	7	0.12	{benefit=1, pati...

term	freq	cent
patient	48	84
information	28	69
privacy	31	63
people	15	50
data	22	43
EHR	17	41
access	12	39
security	7	36
person	9	36
health information	6	34
system	9	24
disclosure	6	23
health	4	23
concern	7	22
medical record	8	22
rules	3	22
protection	5	21
collection	4	20
record	8	20
issue	9	19
definition	4	19
confidentiality	8	19
case	4	19
human services	2	19
department	2	19
accountability	3	19
part	5	19
physician	8	18
consumer control	2	18
health records	2	18
solution	4	18
doctor	5	16
individual	6	16
point	4	16
nurses	3	15
problem	7	15
public health	2	15

Press 'p' for picking mode or 't' for transferring.

Applet started.

Meerkat-ED: Topic Hierarchy

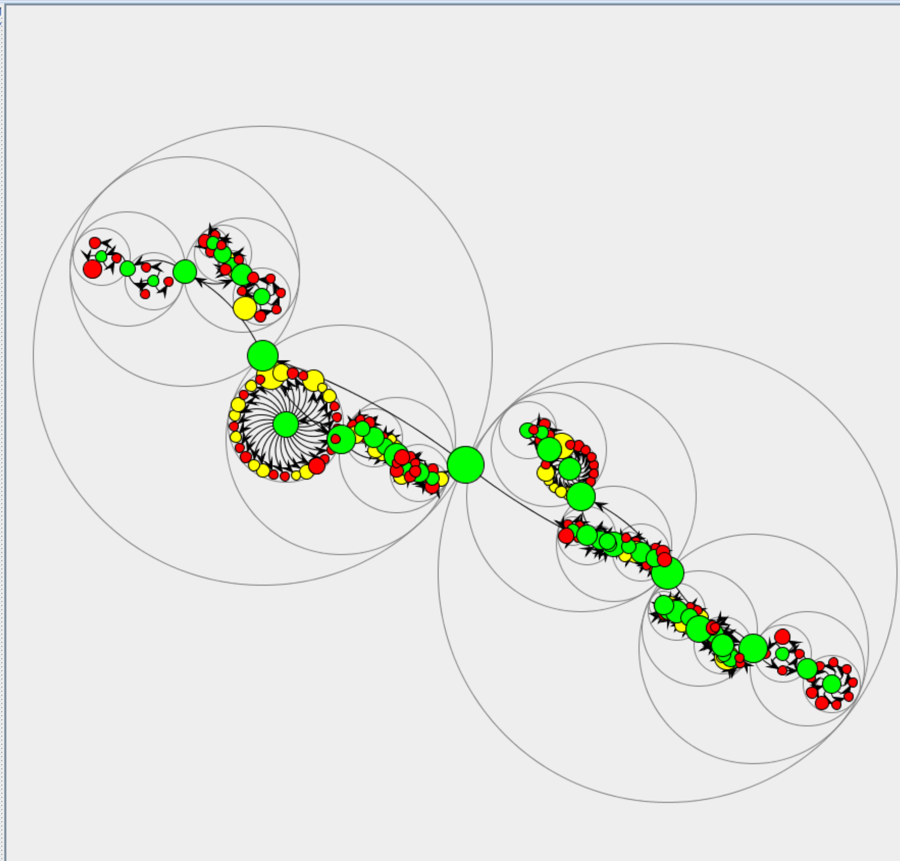
Applet Viewer: representation.SNAEDMainLayout.class

Applet

File View Analyze Setting

Terms' SN

- Course
  - News and Notices
  - Course Discussion
  - Teachers' Forum
  - Welcome to CMPUT 690
  - Electronic Health Record
  - EHR vs EMR vs PMR vr EPR
  - Coding Medical Data
  - Canada Health Infoway
  - Data entry issues
  - Decision Support for Administrators
  - Health Informatics and training
  - Privacy preservation
    - Privacy Preservation
      - Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - PARIS project's privacy problems
      - Re: Privacy Preservation
    - Privacy preservation: consems and issues
      - Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
      - Re: Privacy preservation: consems and issues
    - Anonymous-based approach for privacy preservation
      - Anonymous-based approach for privacy preservation
      - Re: Anonymous-based approach for privacy preservation



name	# of posts	# of replies	avg length of msg	participation	used terms
Chloe	0	1	9	0.16	{risk=1, health i...
Eric	1	0	2	0.18	{health informati...
Olivia	0	3	15	0.24	{health informati...
Jane	0	2	19	0.33	{health data=1,...
Andrea	0	1	32	0.28	{health data=3,...
Sara	0	1	9	0.14	{privacy concern...
Alexander	0	1	5	0.16	{patient=1, pro...
Selma	0	1	17	0.18	{government=1,...
Emily	1	0	25	0.31	{government=2,...
Emma	0	1	5	0.17	{information=2,...
Vahid	0	1	11	0.11	{patient=5, pati...
Monica	1	0	13	0.19	{part=1, risk=1...
Jiyang	0	1	9	0.17	{behavior=2, p...
Mohamed	0	1	7	0.12	{benefit=1, pati...

term	freq	cent
patient	48	84
information	28	69
privacy	31	63
people	15	50
data	22	43
EHR	17	41
access	12	39
security	7	36
person	9	36
health information	6	34
system	9	24
disclosure	6	23
health	4	23
concern	7	22
medical record	8	22
rules	3	22
protection	5	21
collection	4	20
record	8	20
issue	9	19
definition	4	19
confidentiality	8	19
case	4	19
human services	2	19
department	2	19
accountability	3	19
part	5	19
physician	8	18
consumer control	2	18
health records	2	18
solution	4	18
doctor	5	16
individual	6	16
point	4	16
nurses	3	15
problem	7	15
public health	2	15

Press 'p' for picking mode or 't' for transferring.

Applet started.

Meerkat-ED: Topic Hierarchy

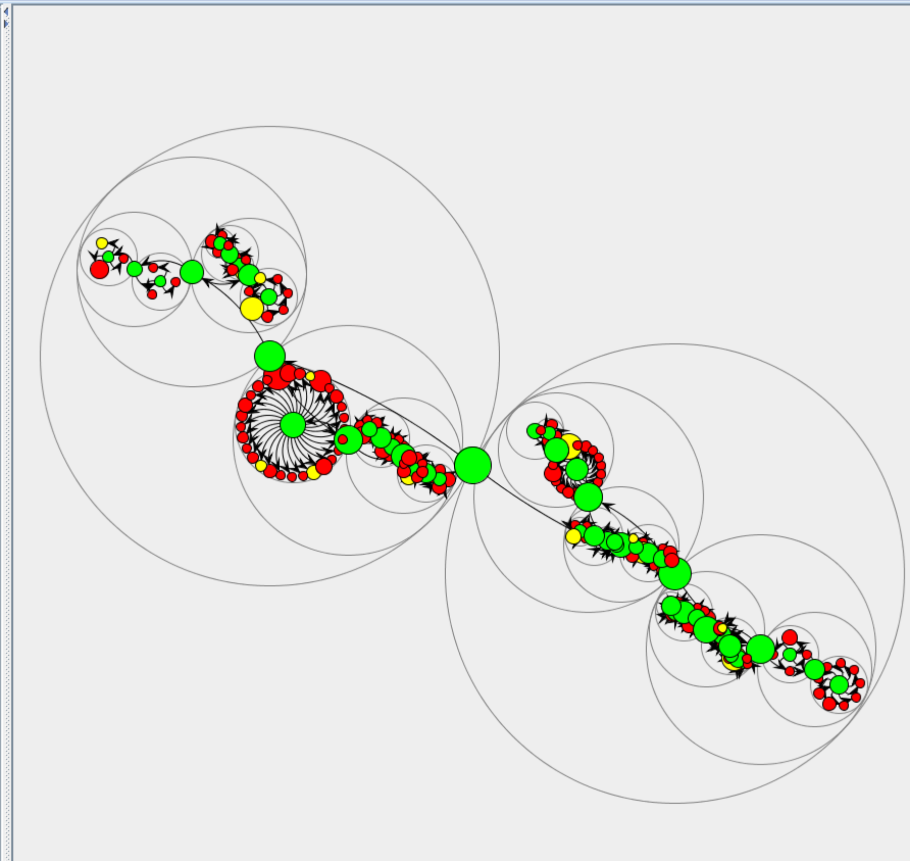
Applet Viewer: representation.SNAEDMainLayout.class

Applet

File View Analyze Setting

Terms' SN

- Course
  - News and Notices
  - Course Discussion
  - Teachers' Forum
  - Welcome to CMPUT 690
  - Electronic Health Record
  - EHR vs EMR vs PMR vr EPR
  - Coding Medical Data
  - Canada Health Infoway
  - Data entry issues
  - Decision Support for Administrators
  - Health Informatics and training
  - Privacy preservation
    - Privacy preservation
      - Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - Re: Privacy Preservation
      - PARIS project's privacy problems
      - Re: Privacy Preservation
    - Privacy preservation: conserns and issues
      - Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
      - Re: Privacy preservation: conserns and issues
    - Anonymous-based approach for privacy preservation
      - Anonymous-based approach for privacy preservation
      - Re: Anonymous-based approach for privacy preservation



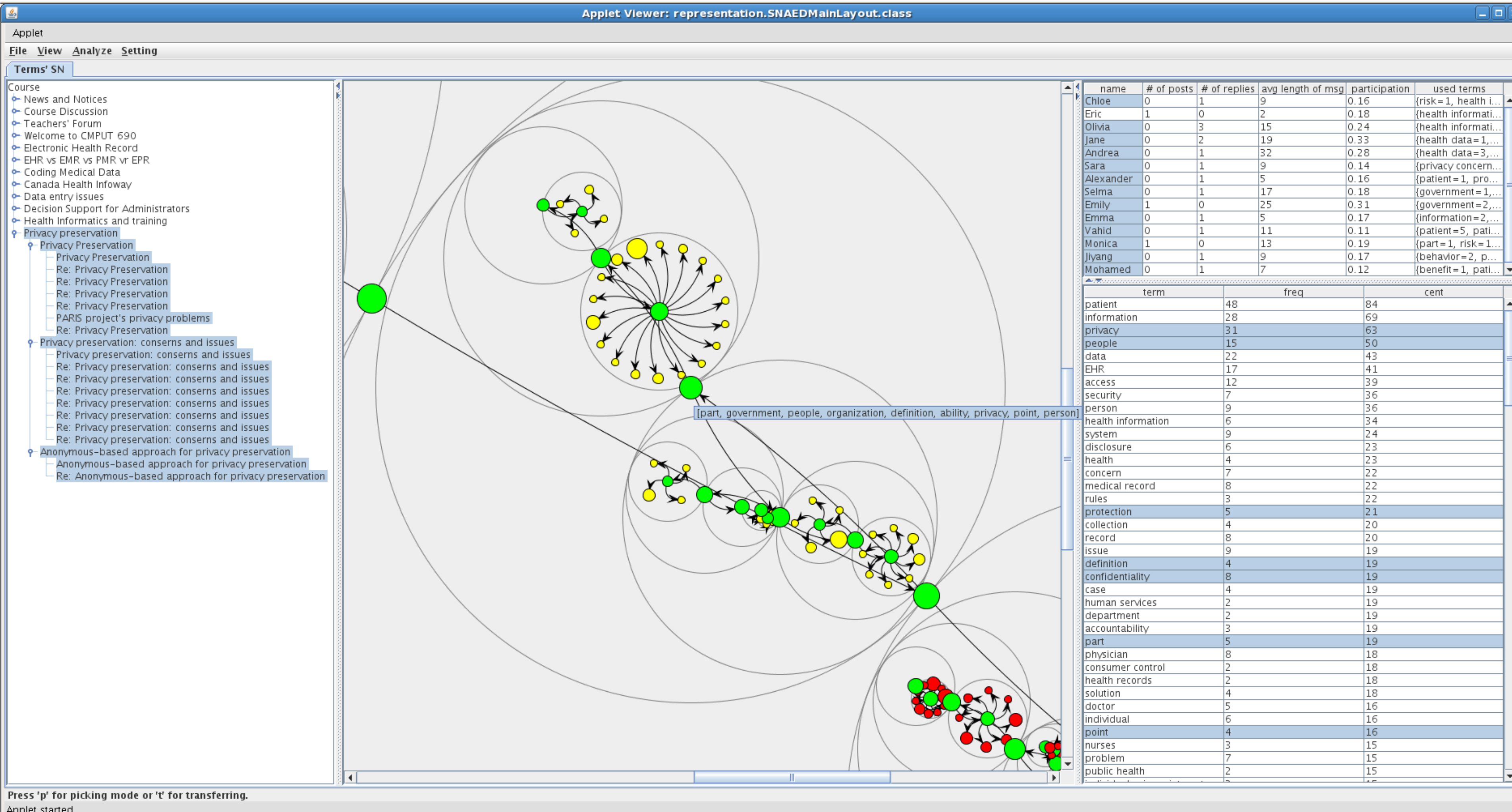
name	# of posts	# of replies	avg length of msg	participation	used terms
Chloe	0	1	9	0.16	{risk=1, health i…
Eric	1	0	2	0.18	{health informati…
Olivia	0	3	15	0.24	{health informati…
Jane	0	2	19	0.33	{health data=1,…
Andrea	0	1	32	0.28	{health data=3,…
Sara	0	1	9	0.14	{privacy concern…
Alexander	0	1	5	0.16	{patient=1, pro…
Selma	0	1	17	0.18	{government=1,…
Emily	1	0	25	0.31	{government=2,…
Emma	0	1	5	0.17	{information=2,…
Vahid	0	1	11	0.11	{patient=5, pati…
Monica	1	0	13	0.19	{part=1, risk=1…
Jiyang	0	1	9	0.17	{behavior=2, p…
Mohamed	0	1	7	0.12	{benefit=1, pati…

term	freq	cent
patient	48	84
information	28	69
privacy	31	63
people	15	50
data	22	43
EHR	17	41
access	12	39
security	7	36
person	9	36
health information	6	34
system	9	24
disclosure	6	23
health	4	23
concern	7	22
medical record	8	22
rules	3	22
protection	5	21
collection	4	20
record	8	20
issue	9	19
definition	4	19
confidentiality	8	19
case	4	19
human services	2	19
department	2	19
accountability	3	19
part	5	19
physician	8	18
consumer control	2	18
health records	2	18
solution	4	18
doctor	5	16
individual	6	16
point	4	16
nurses	3	15
problem	7	15
public health	2	15

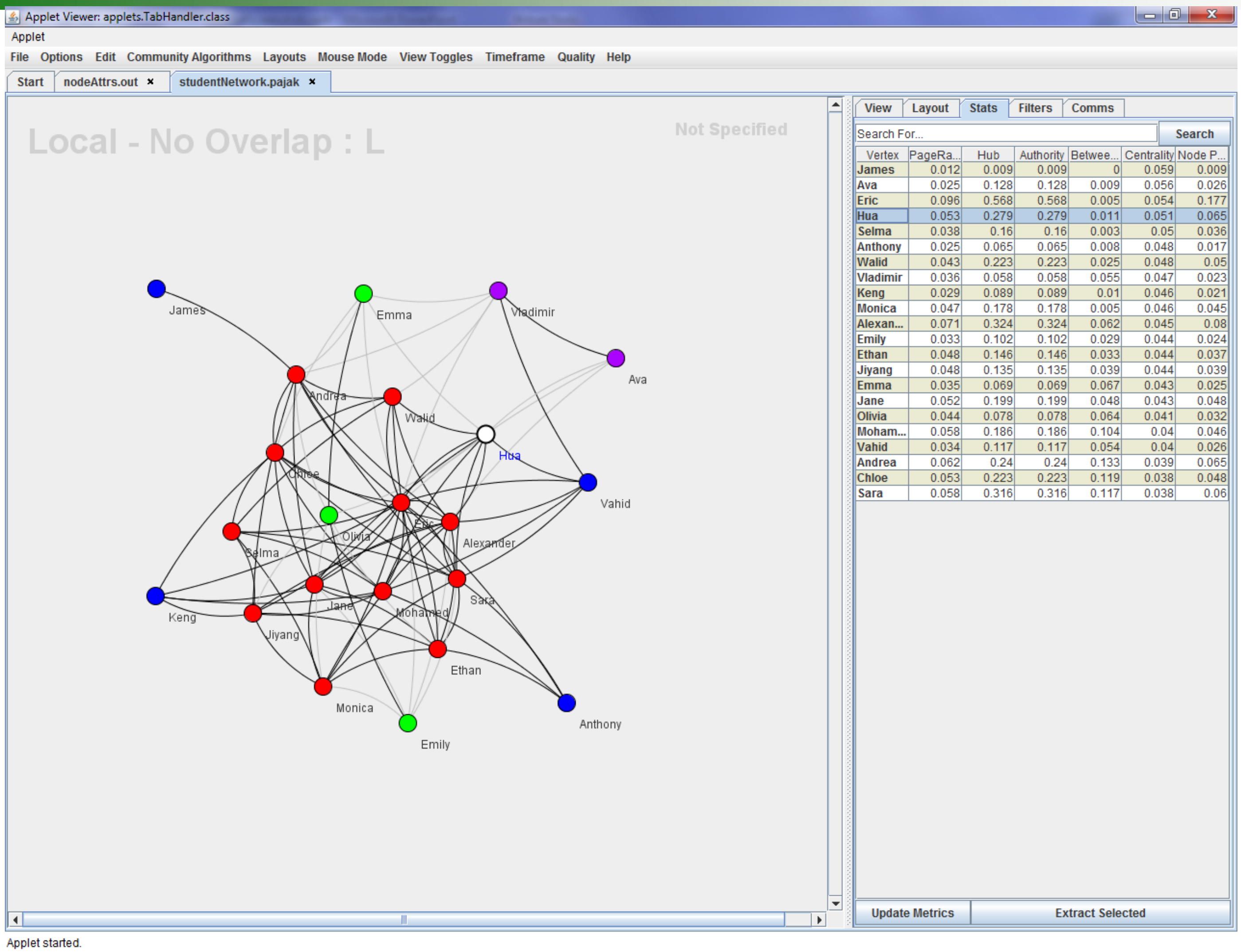
Press 'p' for picking mode or 't' for transferring.

Applet started.

Meerkat-ED: Topic Hierarchy

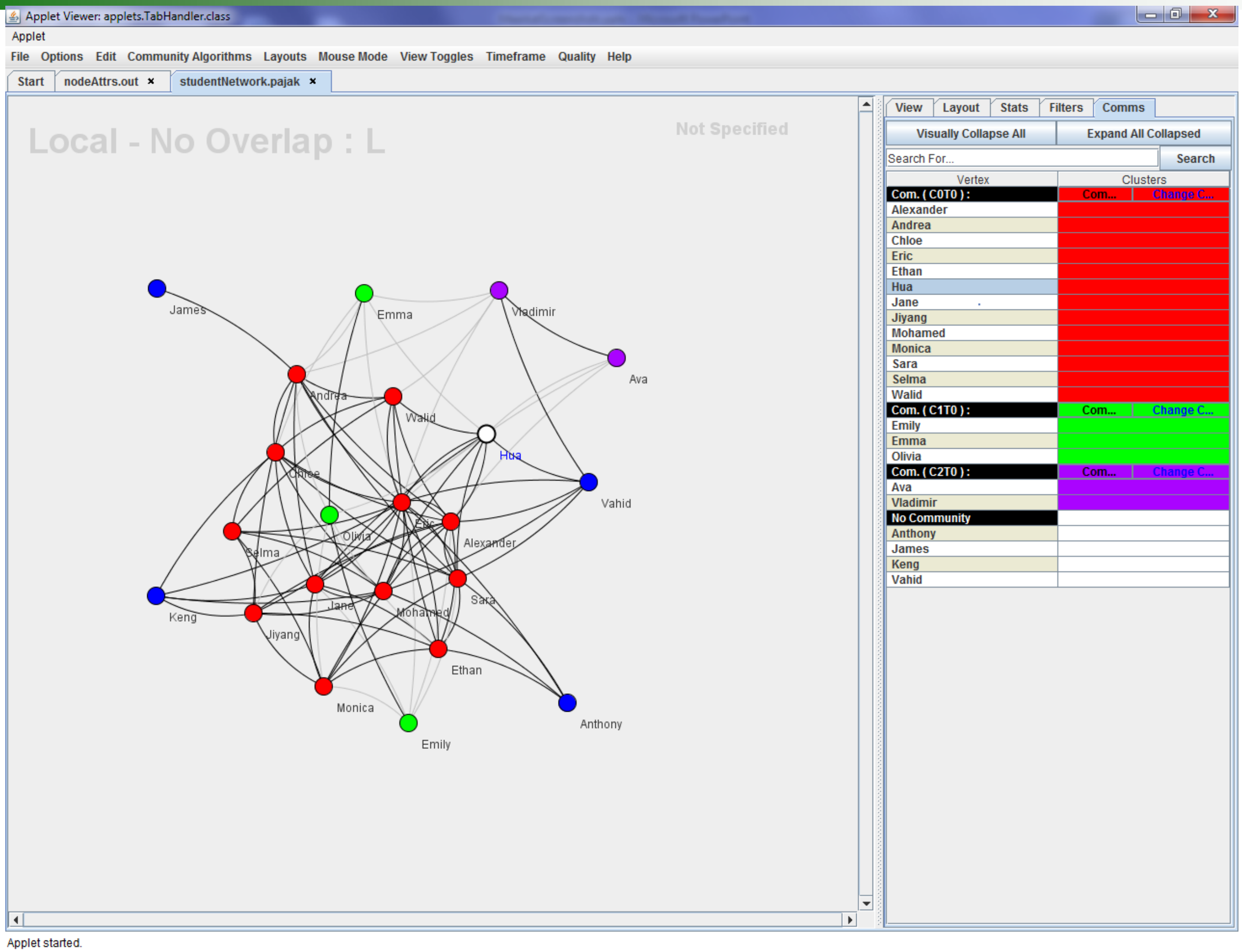


Meerkat: Relativety of Centrality

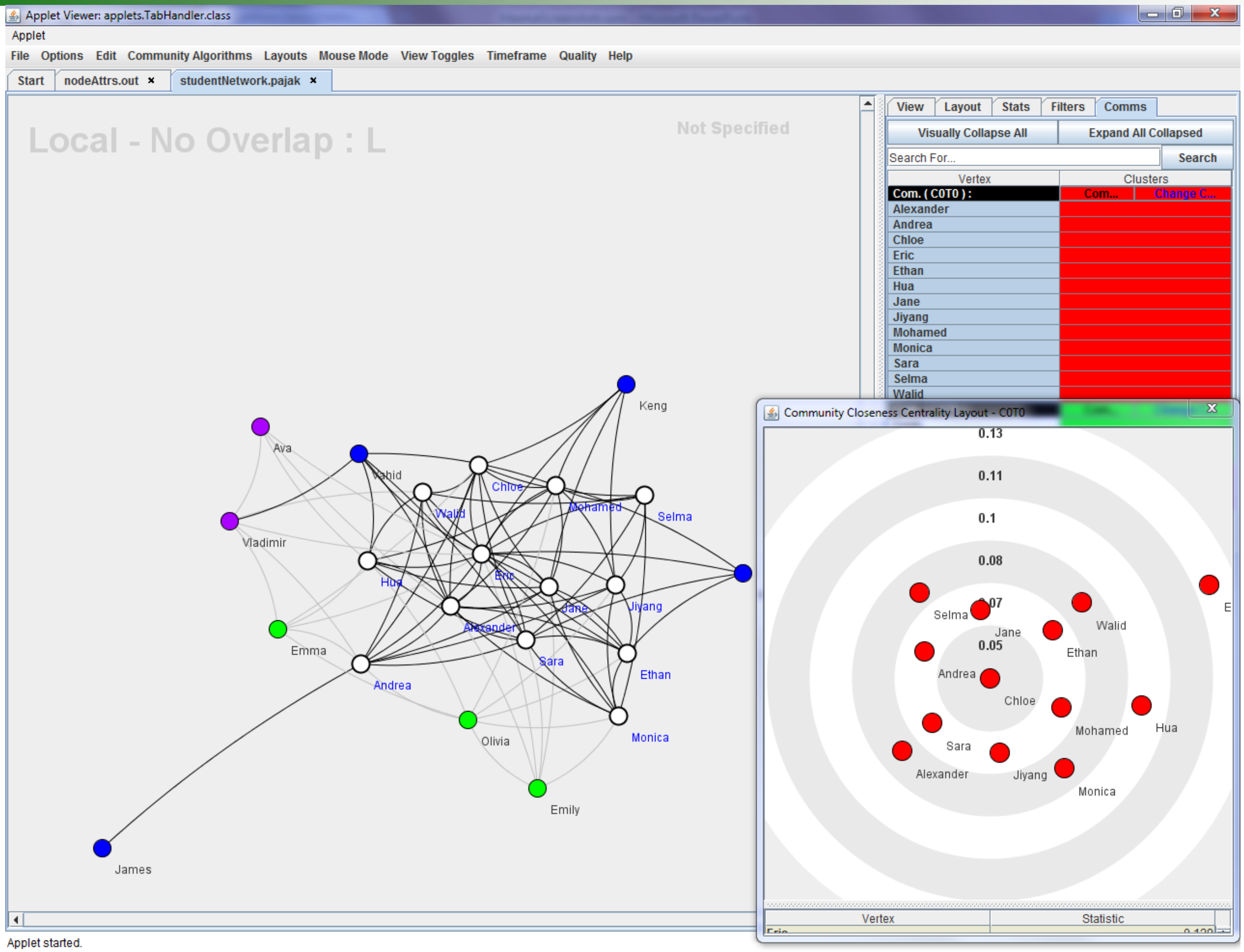


Vertex	PageRa...	Hub	Authority	Betwee...	Centrality	Node P...
James	0.012	0.009	0.009	0	0.059	0.009
Ava	0.025	0.128	0.128	0.009	0.056	0.026
Eric	0.096	0.568	0.568	0.005	0.054	0.177
Hua	0.053	0.279	0.279	0.011	0.051	0.065
Selma	0.038	0.16	0.16	0.003	0.05	0.036
Anthony	0.025	0.065	0.065	0.008	0.048	0.017
Walid	0.043	0.223	0.223	0.025	0.048	0.05
Vladimir	0.036	0.058	0.058	0.055	0.047	0.023
Keng	0.029	0.089	0.089	0.01	0.046	0.021
Monica	0.047	0.178	0.178	0.005	0.046	0.045
Alexan...	0.071	0.324	0.324	0.062	0.045	0.08
Emily	0.033	0.102	0.102	0.029	0.044	0.024
Ethan	0.048	0.146	0.146	0.033	0.044	0.037
Jiyang	0.048	0.135	0.135	0.039	0.044	0.039
Emma	0.035	0.069	0.069	0.067	0.043	0.025
Jane	0.052	0.199	0.199	0.048	0.043	0.048
Olivia	0.044	0.078	0.078	0.064	0.041	0.032
Moham...	0.058	0.186	0.186	0.104	0.04	0.046
Vahid	0.034	0.117	0.117	0.054	0.04	0.026
Andrea	0.062	0.24	0.24	0.133	0.039	0.065
Chloe	0.053	0.223	0.223	0.119	0.038	0.048
Sara	0.058	0.316	0.316	0.117	0.038	0.06

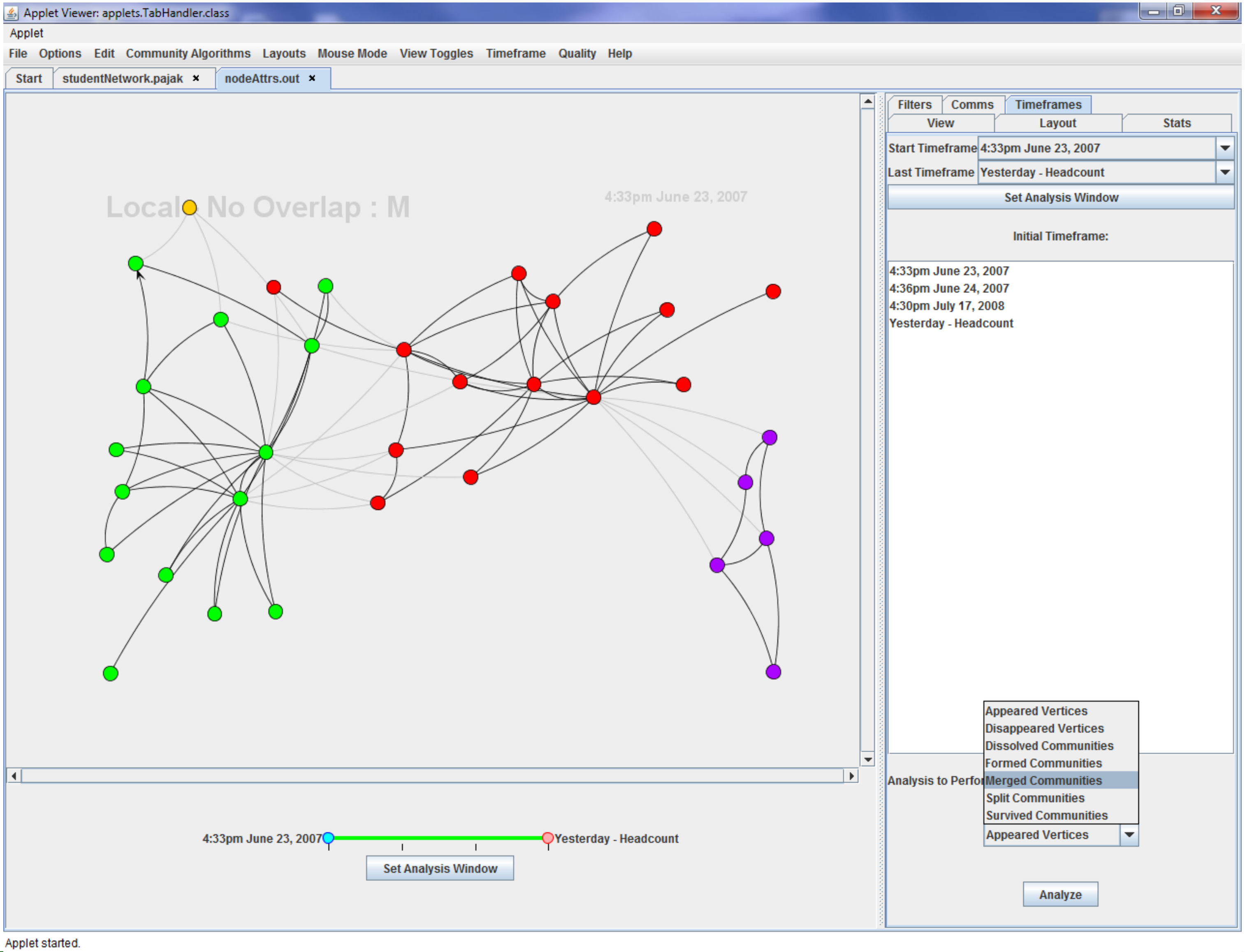
Meerkat: Relativety of Centrality



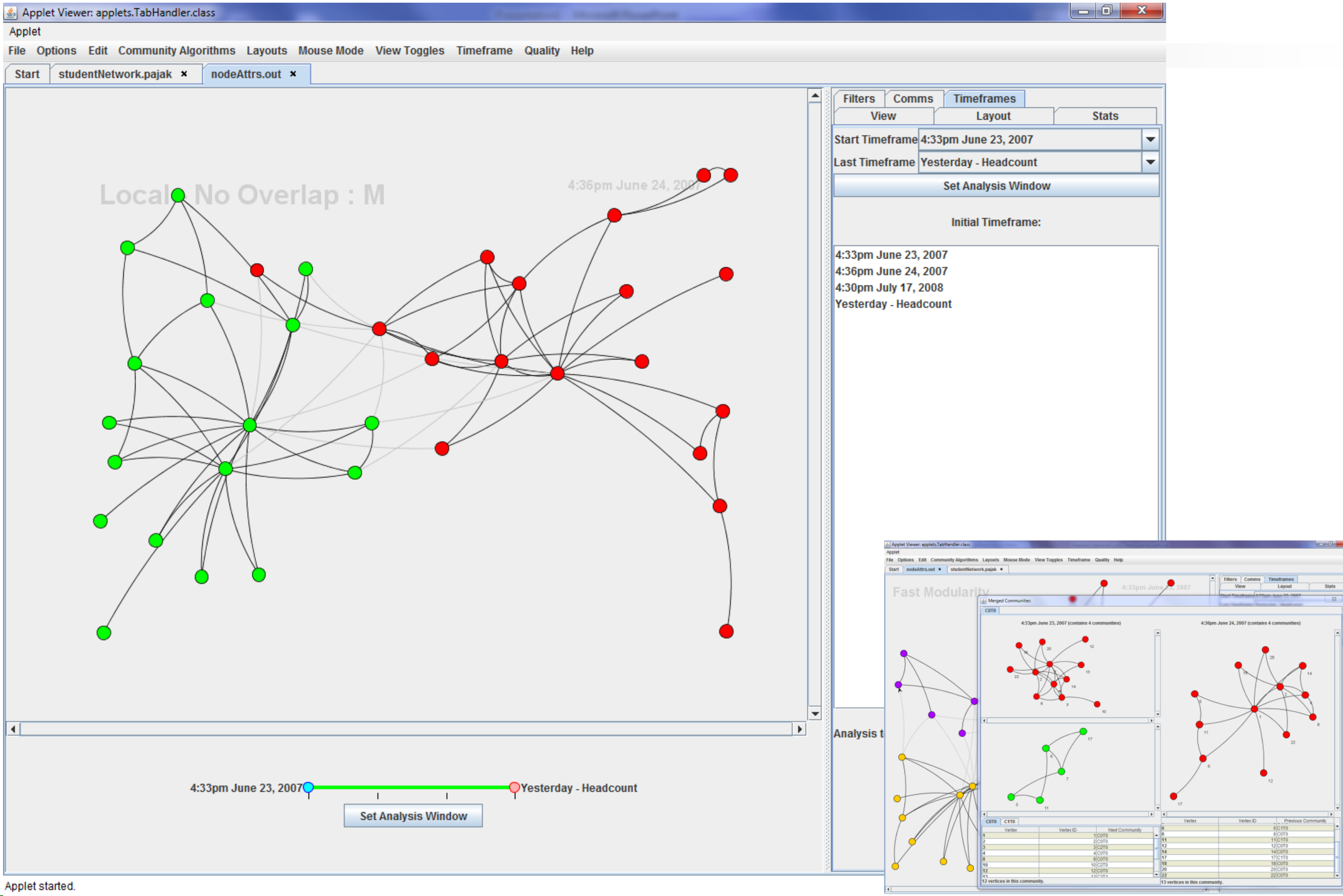
Meerkat: Relativety of Centrality



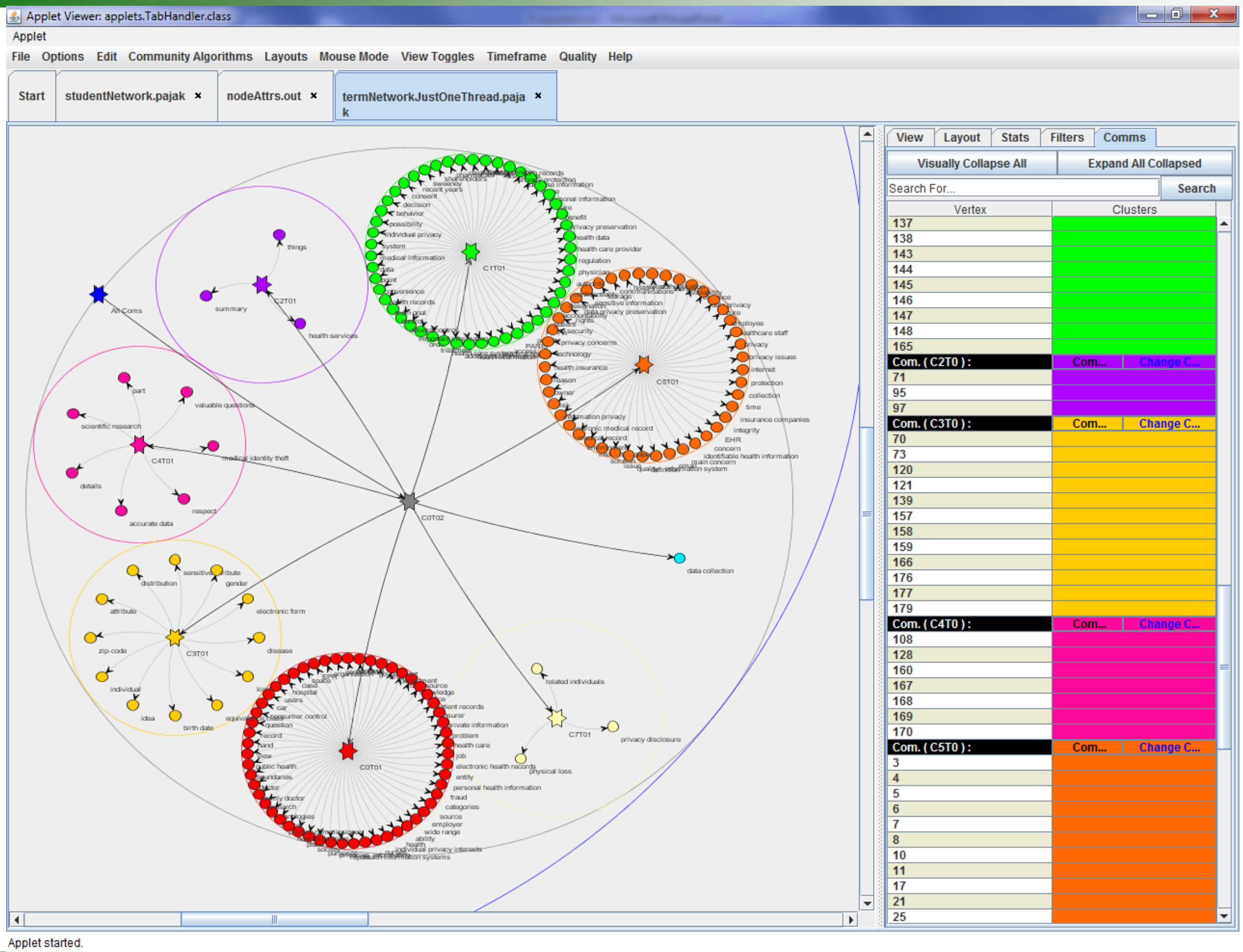
Meerkat: Community Dynamics



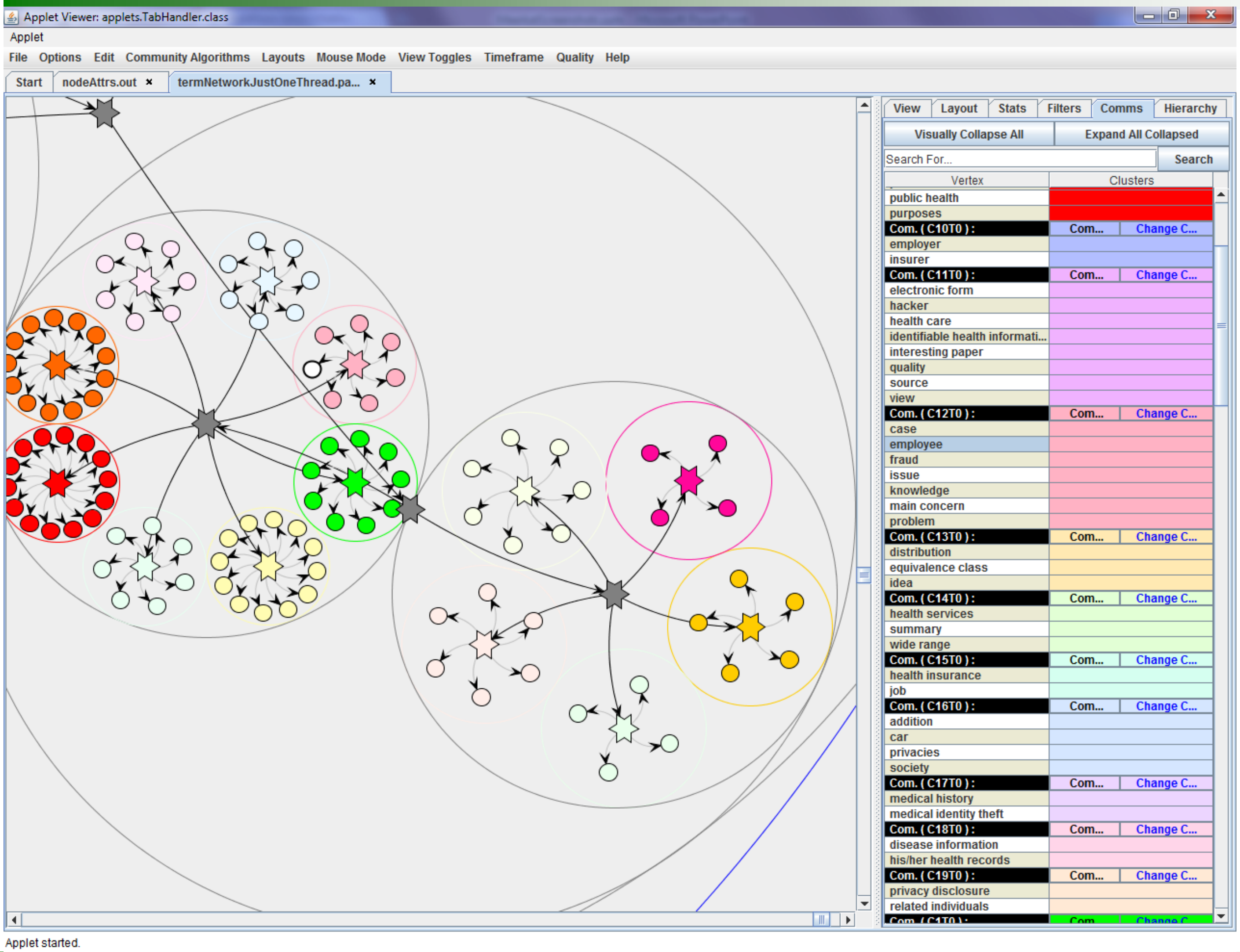
Meerkat: Community Dynamics



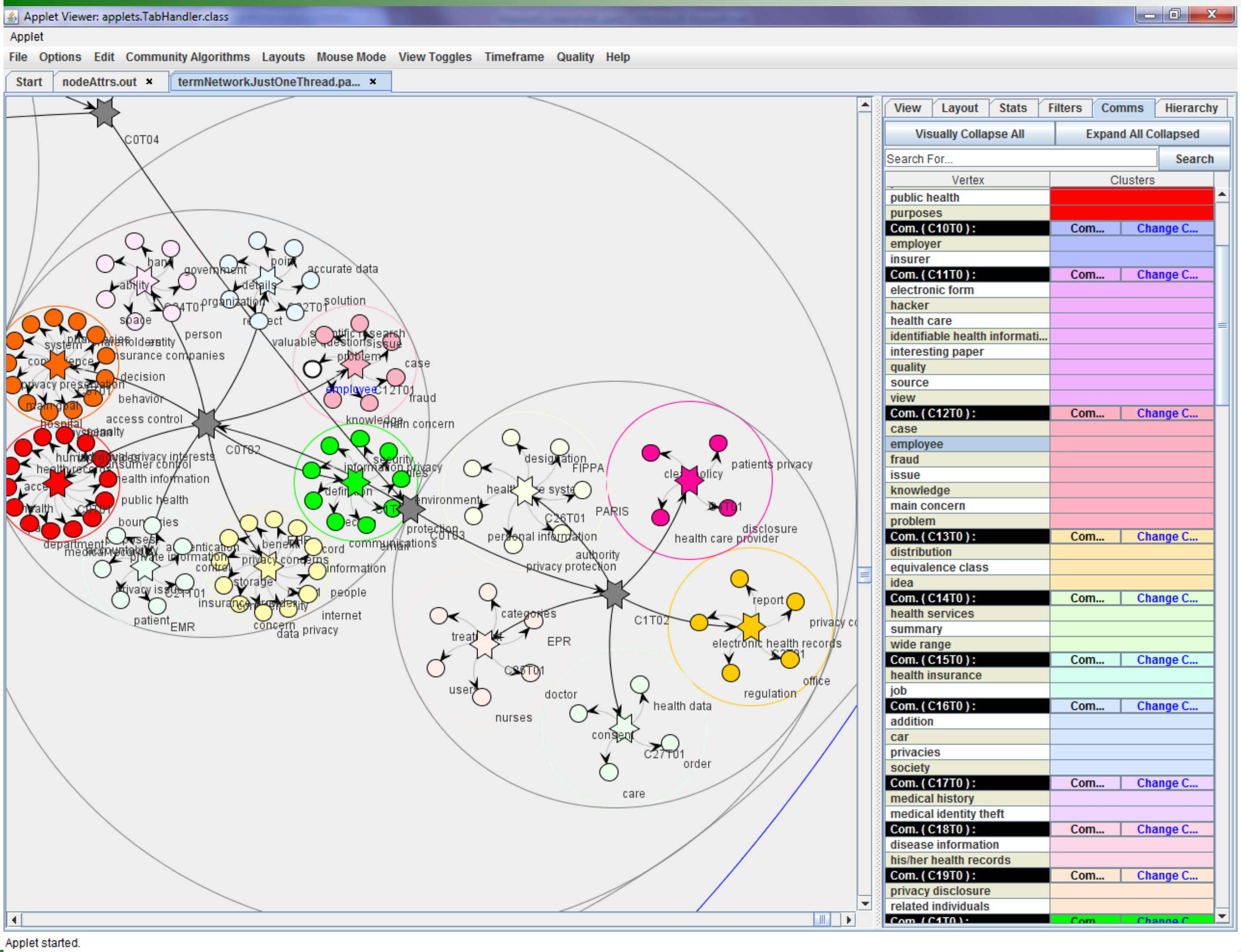
Meerkat: Topic (term community) Hierarchy



Meerkat:
Topic (term community) Hierarchy



Meerkat:
Topic (term community) Hierarchy



Challenges

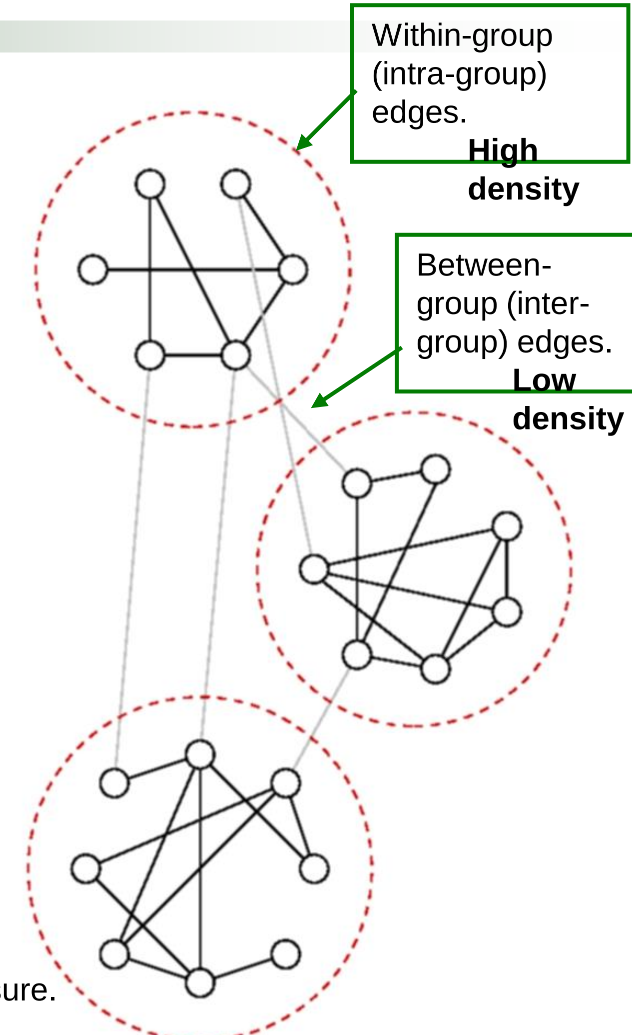
How do we find communities?

How do we find topic hierarchies?

What is Community Structure?

- *Community structure* denotes the existence of densely connected groups of nodes, with only sparser connections between groups.

- Many social networks share the property of a community structure, e.g., WWW, tele-communication networks, academic collaboration networks, friendship networks, etc.



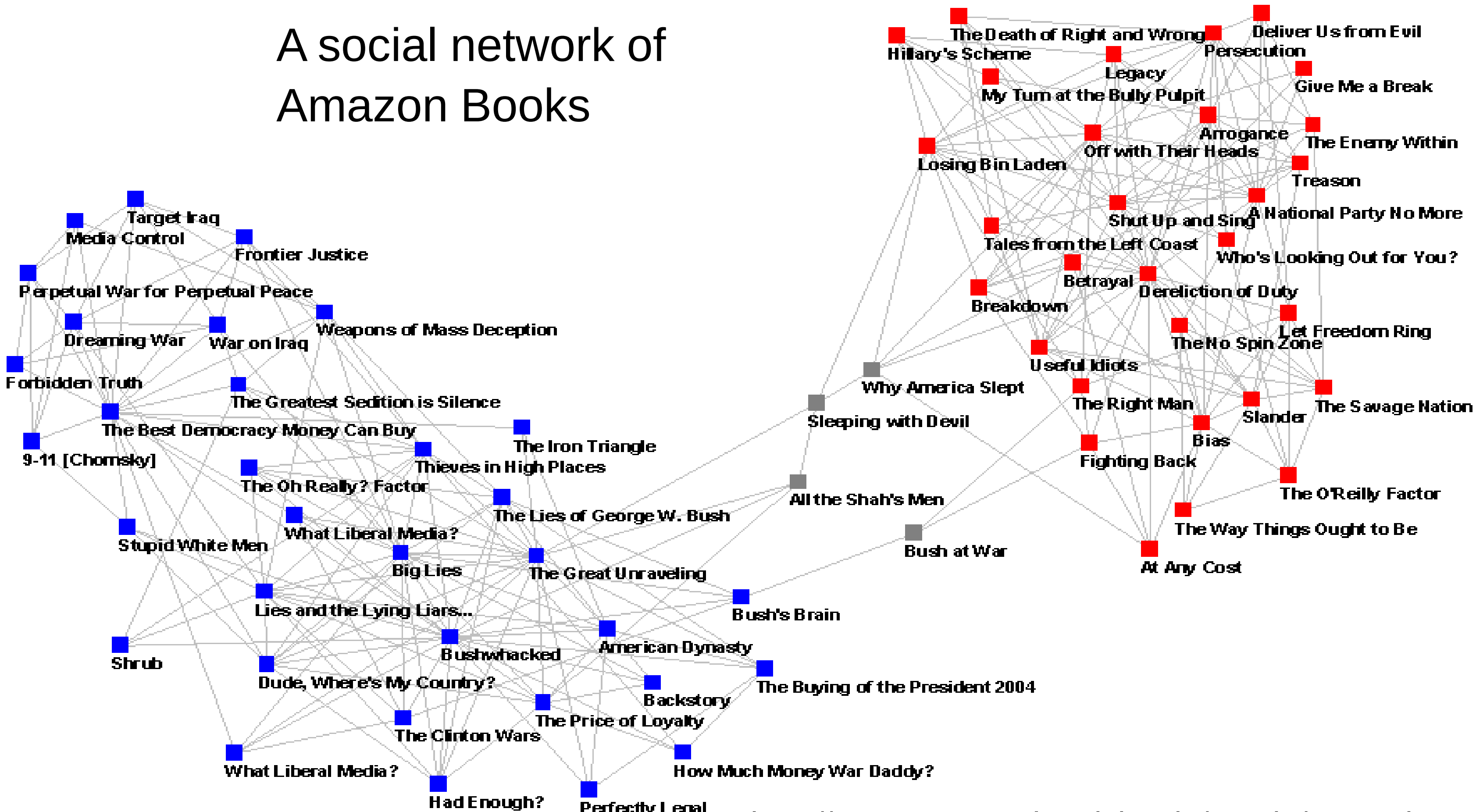
Many similarities with data **Clustering**

Clustering is dividing the data points into classes according to some similarity measure.

Community structure: dividing the network into groups according to structural info.(connectivity).

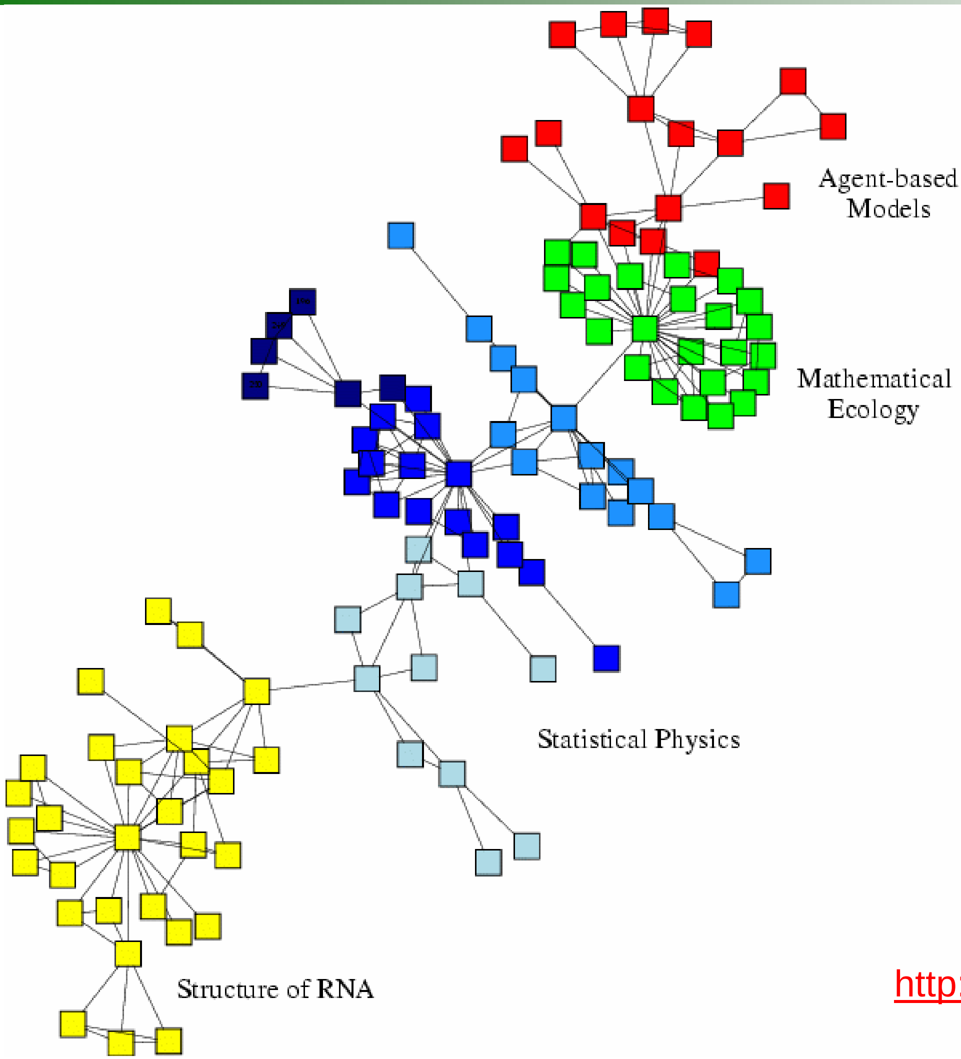
Community Structure Examples

A social network of Amazon Books



http://www-personal.umich.edu/~mejn/networks

Community Structure Examples



A academic collaboration social network

http://www-personal.umich.edu/~mejn/networks

It is important!

- Finding communities could be of significant importance.
- WWW Pages (in the same hyperlink community) might discuss related topics.
- Researchers (in the same collaboration community) might work on similar problems.
- People (in the same tele-communication community) might be close friends.
- Communities in social settings might explain or predict the spread of contagious diseases.
- And many other examples.

What is a Community?

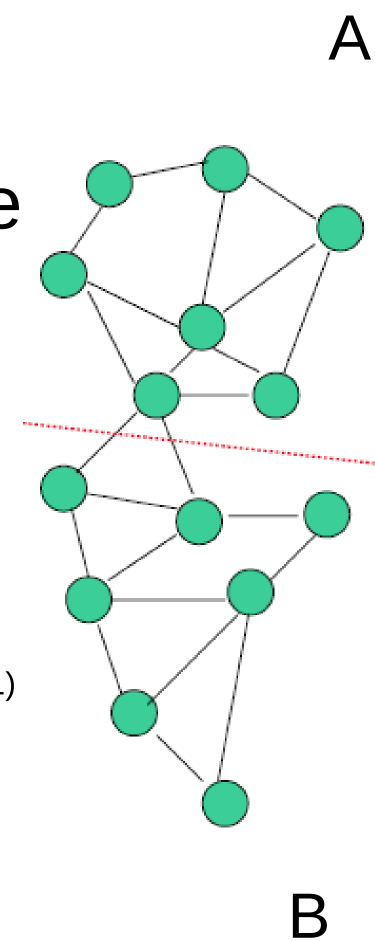
- Graph theory: Communities are those densely connected groups of vertices, with only few connections between groups.
- Sociology: Communities are social groups that entities in the same group share similar properties or connect to each other via certain relations.
- More definitions are available, however, communities are often different for different domains, even for different networks in the same domain. Thus there is no general definition.
- In community mining, the community structure found is usually a byproduct of the discovery procedure.

Graph Partitioning Approaches

- There is a long computer science tradition in graph partitioning: believed to be an NP-complete problem.
- Typical Solution: greedily optimize an objective function: the fraction between intra-community and inter-community edges.
- Iterative Bisection: find the best two-group-cut, then further sub-divide until the required community number is met.

Graph Partitioning Methods

- Graph partitioning algorithms are heavily used to find communities.
- Parameters that are difficult to decide are usually required: size of communities, number of clusters
- Spectral Clustering with benefit functions: **ratio cut** (Hagen et al. 1992), **normalized cut** (Shi et al. 1997), **min-max cut** (Ding et al. 2001)
- Unfortunately, equal-sized communities are usually favoured.



Other Problems

- Require input parameters: number of the partitions, and their sizes
- Such information would never be available for large social networks. They should be determined by the network, not the user.
- Fundamental problem: cut (sum of edge weights between communities) is simply not the right thing to optimize.

Hierarchical Clustering

- Greedily optimize a metric, which evaluates the node centrality or community quality.
- An example metric: edge betweenness, which is the number of edge occurring on the shortest path between other pair of nodes in the network.
- Up-down Algorithm: remove the edge with highest betweenness value in each step.

Modularity Q

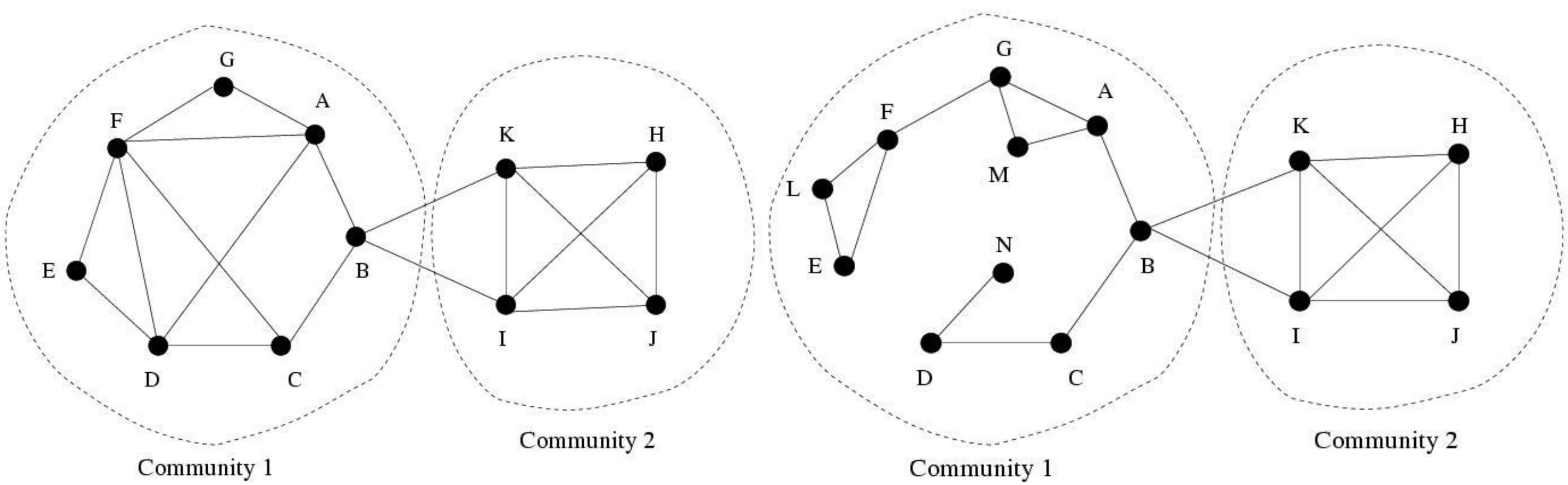
- Proposed by Newman and Girvan in 2004 as a measure of the quality of a particular division of the network.
- Q = (*number of edges within communities*) – (*expected number of such edges*)
- Intuition: compare the division to a random network with same nodes and same degrees, but edges are placed randomly.
- ➔ a good division of a network is not merely one in which the number of edges in groups is large, but it is one in which the number of edges within groups is ***larger than expected***.
- Greedily maximizing Q outperformed all other methods, in most cases by an impressive margin, for community detection.

Success of the Modularity

- Algorithm: bottom-up agglomerative hierarchical clustering to maximize Q.
- Q has proven to be highly efficient.
- Q-based methods over-perform other community mining algorithms on many networks, usually with a big margin.
- FastModularity [Clauset, Newman and Moore 2004] – use of Max Heaps and binary tree to provide an efficient O(*md* log *n*) Modularity implementation where *m* is the # of edges, *n* is the number of nodes, and *d* the depth of the dendrogram.

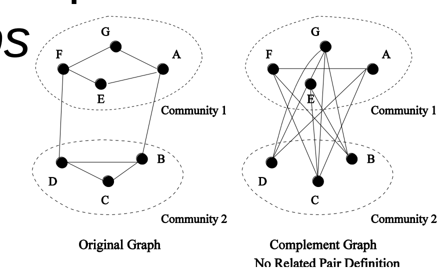
Problem Solved?

- There are three major problems for Q.
  - Q requires information of the entire network.
  - Q has a resolution limit and may fail to identify communities smaller than a certain scale.
  - Q cannot be used to compare community qualities in different networks. (Q = 0.360 for both)



Max-Min Modularity [SDM'09]

- Evaluation Metric: reward for connected pairs and penalty for disconnected ones.
- A "disconnection" can be "unobserved" in many social networks, e.g., biological network, dynamic Facebook.
- *Maximize* the edge number within groups and *minimize* the number of unrelated pairs defined by experts within groups at the same time
  ➔ *the number of unrelated pairs within groups is smaller than expected.*



- Use of complement graph

$$Q_{\max\_\min} = Q_{\max} - Q_{\min}$$

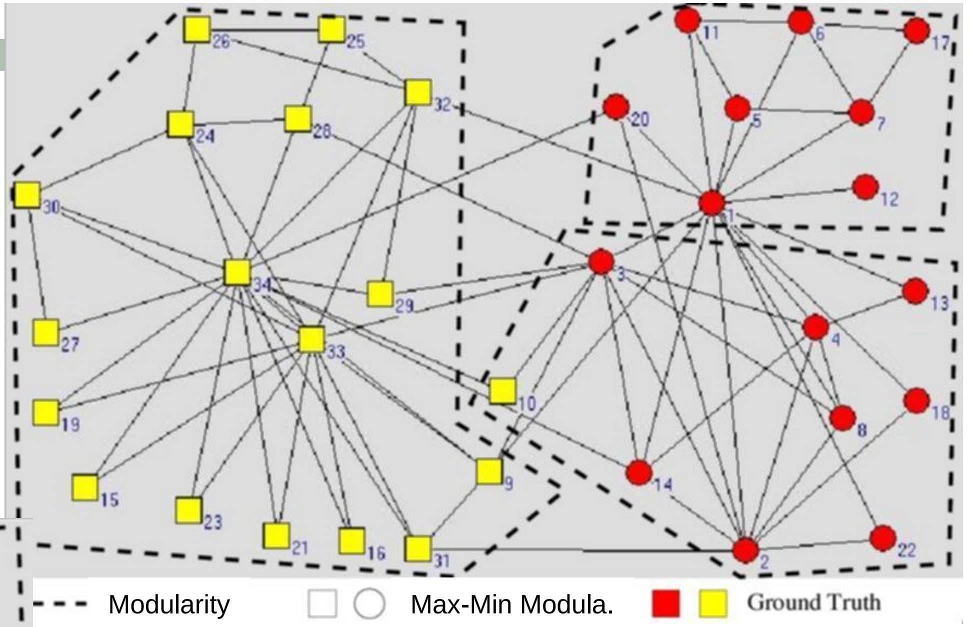
$Q_{\max}$ = Modularity Q

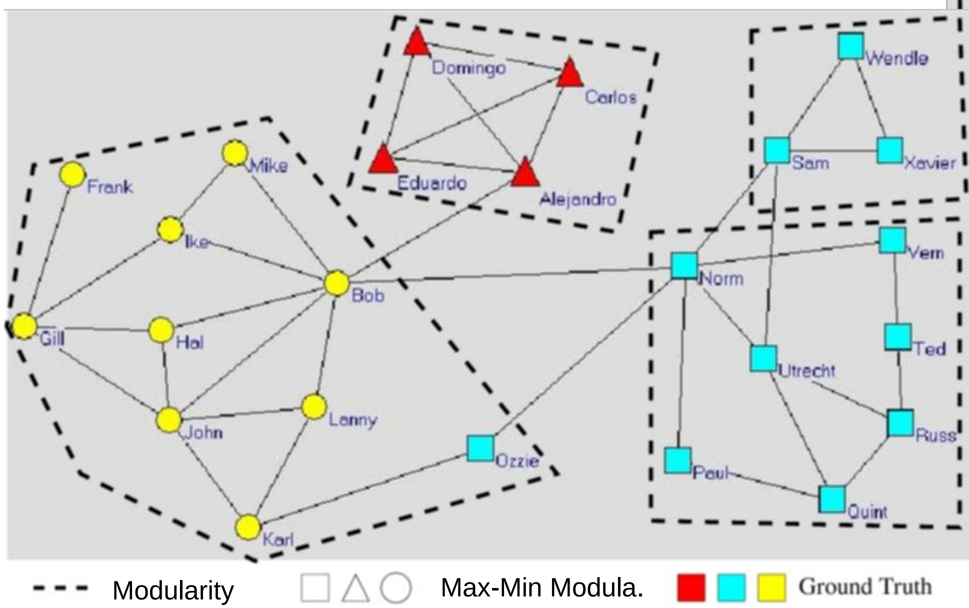
$$Q_{\min} = \frac{1}{n(n-1) - 2m - 2|U|} \sum_{xy} [A'_{xy} - P'_{xy}] \phi(C_x, C_y)$$

n is the node number.
U is the related but disconnected pair set defined by domain experts.

Example Results with Max-Min Modularity

Karate-Club dataset
34 nodes in 2 communities





Sawmill Strike dataset
24 nodes in 3 communities

node pairs are “related” if they share neighbours

On Real Networks?

- Most of these approaches require knowledge of the entire network structure, e.g., number of nodes/edges, number of communities in the network. However, this is problematic for networks which are either too large or dynamic, e.g., the WWW.

- The size of the WWW 1 trillion unique URLs. The index size of Google is about 40 billion.
  http://www.techcrunch.com/2008/07/25/googles-misleading-blog-post-on-the-size-of-the-web/
- Facebook has more than 200 million active users
  http://www.facebook.com/press/info.php?statistics
- Vodafone has 289 million customers worldwide

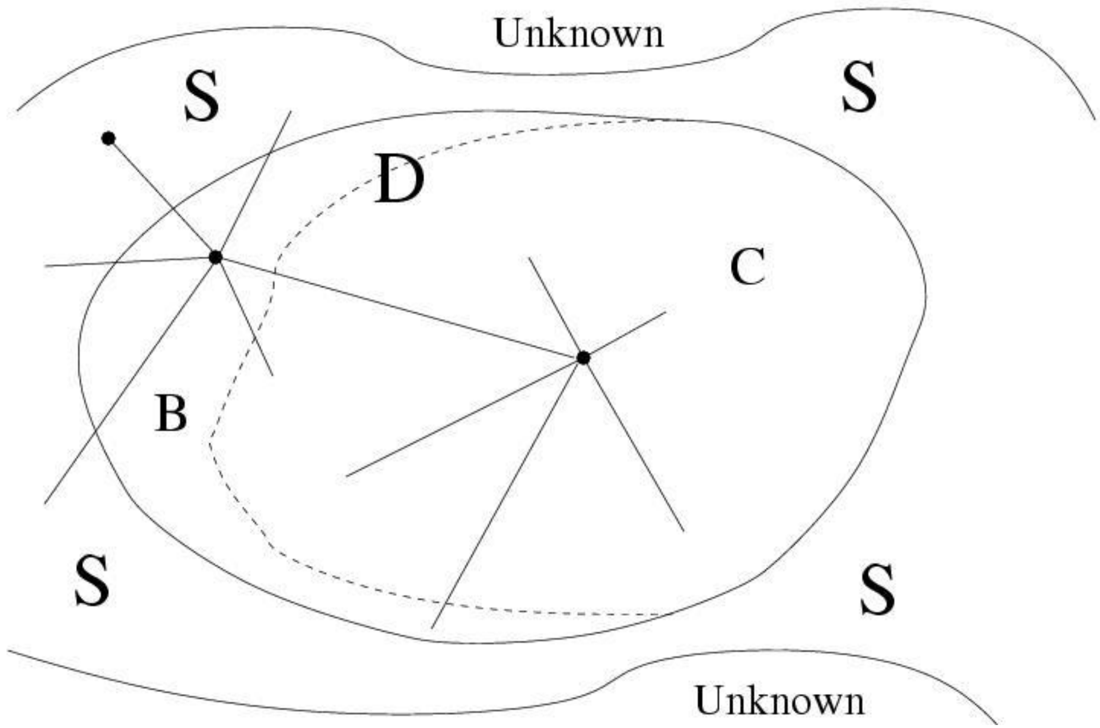
  http://www.vodafone.com/start/media_relations/news/group_press_releases/2009/mobile_internet_experience.html

Local Methods

- A common assumption for the proposed methods is that the complete global network information is always available.
- For some huge networks, e.g., WWW, global information is not always accessible.
- Scenarios: Locate a friend community of a person in Facebook or Find a page cluster of a particular page in the WWW.
- The only available information are nodes that have been visited and their neighbours. All global methods fail.

Typical Problem Definition

- A local community D includes cores (C) nodes and boundary (B) nodes.
- If one new node is merged, its neighbours are added into shell nodes (S).



Modularity in Local Network

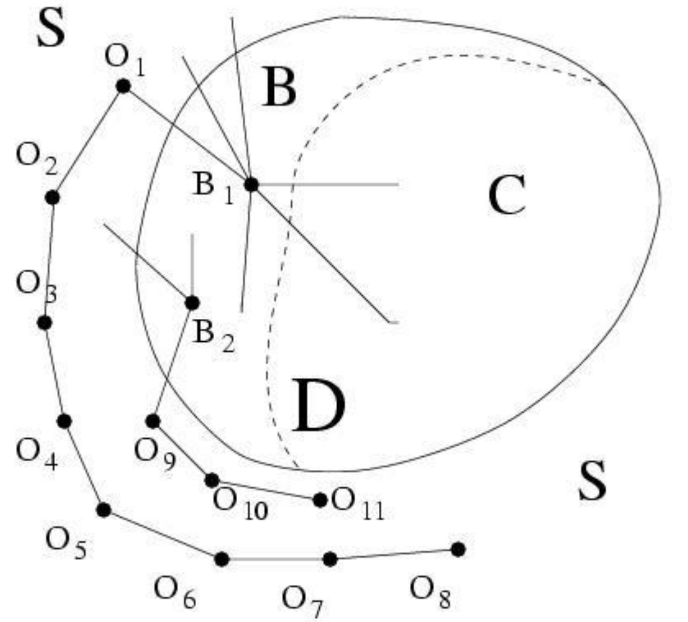
- Clauset proposed in 2005 the local modularity using the modularity methodology:
  - Measure R, the quality of communities
  - Greedily maximize the R measure to identify communities

- R = within edges of boundary nodes /
  total edges of boundary nodes

- R measures the sharpness of the boundary nodes. Identify local community by keeping merging until no merge can increase R.

Local Modularity's Problem

- Weakly linked nodes are always merged into the local community.



- In_edge / total < In_edge+1/total+1

- Outliers are merged into the local community one by one.

Measure the Local Community

- Two factors to consider in local community quality:
  - high node relations within the set
  - low relations between set and outside nodes
- R directly represent these two factors by maximizing internal degrees and minimizing external degrees
- The important missing aspect for R is the *connection density*, not the absolute number of connections, that matters in community structure evaluation.

Detecting based on Local Density

- We [ASONAM 2009] propose to measure the two factors by maximizing ***average*** internal degree (id) inside the whole community and minimizing ***average*** external degree (ed) of boundary nodes, by maximizing id/ed.

- The “density” idea solves the outlier problem and dramatically increases community detection accuracy on some datasets with ground truth.

Experiments

- We use F-measure as a metric and compare to Clauset's Local Modularity algorithm.
- We use the NCAA-2006 football network to evaluate: every conference is a community since universities in the same conferences match more often.
- The dataset: 115 conference universities, 11 conferences, 4 independent teams and 61 teams in the lower division. Teams play more games with other teams in the same conference (except Army, Navy, independent and low div)
- F-measure 0.595 -> F-measure 0.952 (on NCAA Football dataset with ground truth)

Results for the NCAA Network

2006 NCAA League		Algorithm Results						
		Greedy Algorithm R using metric R			Our Algorithm using metric L			
Conference	Size	Precision	Recall	F-measure	No Community	Precision	Recall	F-measure
Mountain West	9	0.505	0.728	0.588	0 node	0.944	1	0.963
Mid-American	12	0.392	0.570	0.463	1 nodes	0.923	1	0.96
Southeastern	12	0.331	0.541	0.410	3 nodes	1	1	1
Sun Belt	8	0.544	0.891	0.675	3 nodes	1	1	1
Western Athletic	9	0.421	0.716	0.510	4 nodes	0.6	1	0.733
Pacific-10	10	0.714	1	0.833	0 nodes	1	1	1
Big Ten	11	0.55	1	0.710	9 nodes	0.729	1	0.814
Big East	8	0.414	0.781	0.534	5 nodes	1	1	1
Atlantic Coast	12	0.524	0.924	0.668	3 nodes	1	1	1
Conference USA	12	0.661	1	0.796	1 nodes	1	1	1
Big 12	12	0.317	0.465	0.355	5 nodes	1	1	1
Total	115	0.488	0.783	0.595	34 nodes (29.6%)	0.927	1	0.952

Our approach dramatically increase the local community detection accuracy, from F-measure 0.595 → 0.952.

Amazon Data

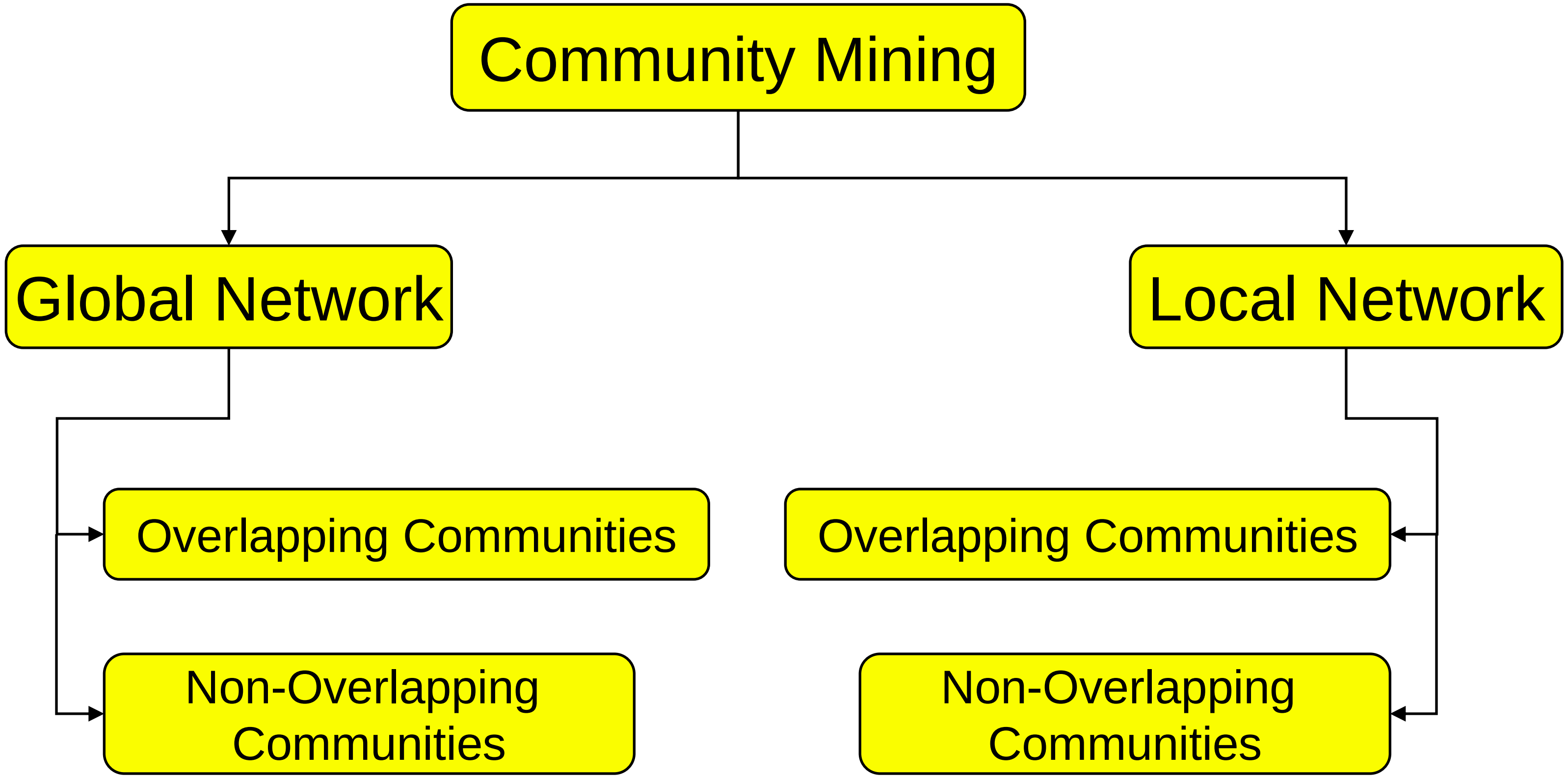
- The Amazon network (Jan. 2006) represents the purchase records of books,CDs, and DVDs.
- Edges connect items are frequently purchased together, represented by “customers who bought this item also bought these items” feature in Amazon website.
- A sparse network: 585,253 nodes, 3,448,754 undirected edges, mean degree 5.89.

Amazon Result

Algorithm	Items (Books) in the Local Community
Both	Smith of Wootton Major*
	The Lord of the Rings: A Reader's Companion#
	The Lord of the Rings: 50th Anniversary, One Vol. Edition*
	(The starting node) The Lord of the Rings [BOX SET]*
L	On Tolkien: Interviews, Reminiscences, and Other Essays#
	Tolkien Studies: An Annual Scholarly Review, Vol. 2#
	Tolkien Studies: An Annual Scholarly Review, Vol. 1#
	A Gateway To Sindarin: A Grammar of an Elvish Language from J.R.R. Tolkien's Lord of the Rings#
	J.R.R. Tolkien Companion and Guide#
	The Rise of Tolkienian Fantasy#
	Perilous Realms: Celtic And Norse in Tolkien's Middle-Earth#
R	Farmer Giles of Ham & Other Stories*
	Smith of Wootton Major & Farmer Giles of Ham*
	Roverandom*
	Letters from Father Christmas, Revised Edition*
	Bilbo's Last Song*
	Farmer Giles of Ham : The Rise and Wonderful Adventures of Farmer Giles*
	Poems from The Hobbit*
	Father Christmas Letters Mini-Book*
	Tolkien: The Hobbit Calendar 2006*

* **indicates the author is J.R.R.Tolkien while** # **is not.**

Community Mining Hierarchy



Global Overlapping Methods

- We usually assume that one node belongs to only one community. However, in the real world, it is not the case.

- One person can belong to two or more communities, thus we need to consider overlapping communities.

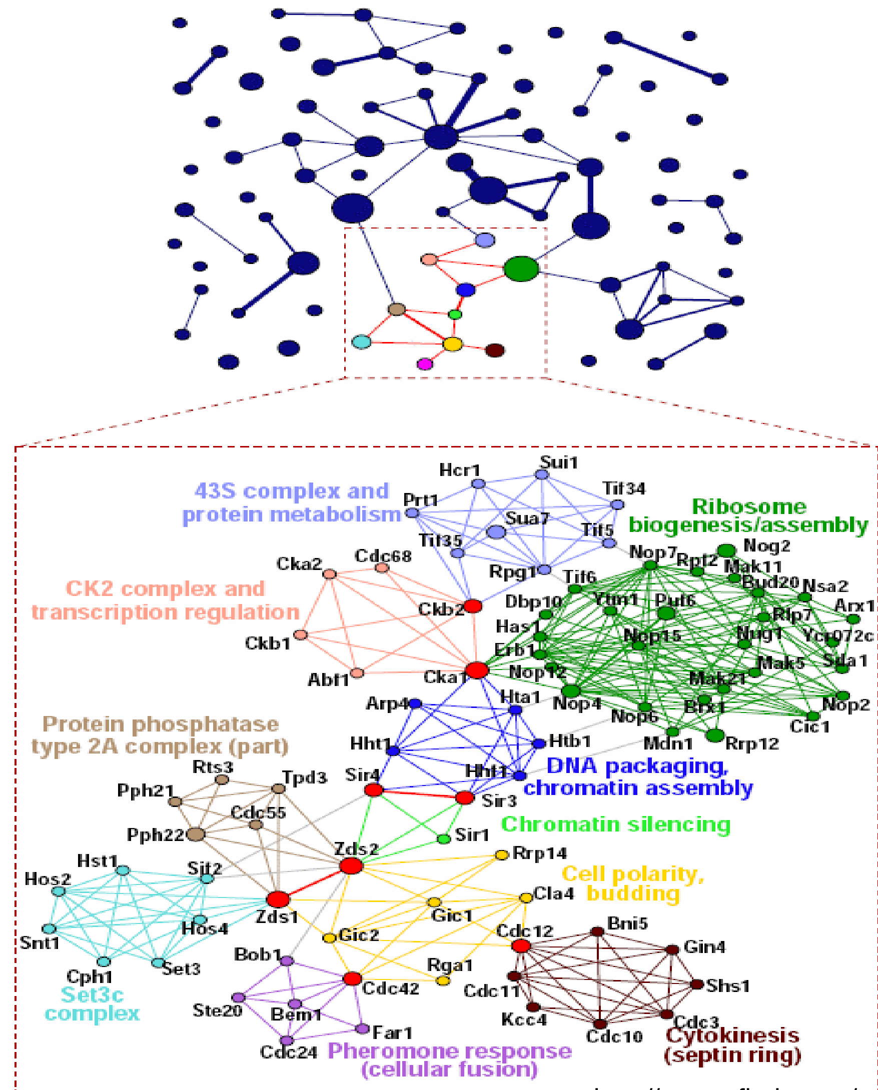
- Typical approach: find the cluster, then measure the relations of nodes in question to different clusters with arbitrary threshold.

CFinder

- Palla et al. proposed the CFinder system in Nature 2005, using a simple but efficient idea to detect overlaps based on cliques.
- Cliques are completely connected sub-graphs, representing strong communities.
- One node can belong to multiple cliques, which shows community overlaps.

CFinder

- CFinder takes a parameter k, which is the clique size.
- Two k-cliques are adjacent if they share k-1 nodes.
- Given clique size k, merge adjacent k-cliques into one community to identify the network structure.
- Problem: also depends on parameters, k = 3,4,5 usually give reasonable results.



http://www.cfinder.org/

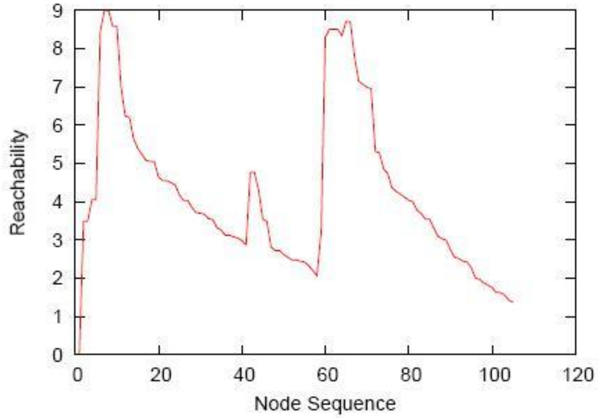
Local Overlapping Methods

- Previous works, using only local information, focus on locating the first local community given a start node.

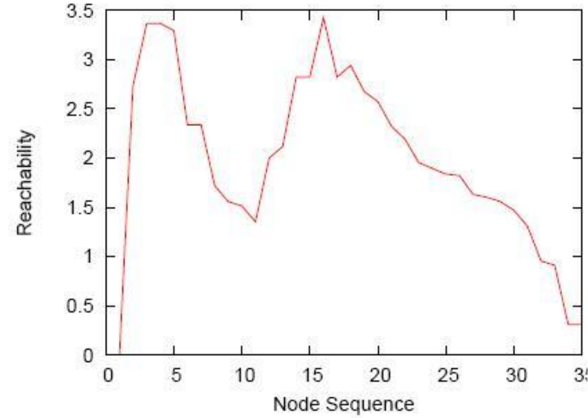
- Iteratively applying the community identification algorithm based on local modularity may be able to find local-overlapping communities (Chen et.al CASoN 2009)

Visual Community Mining

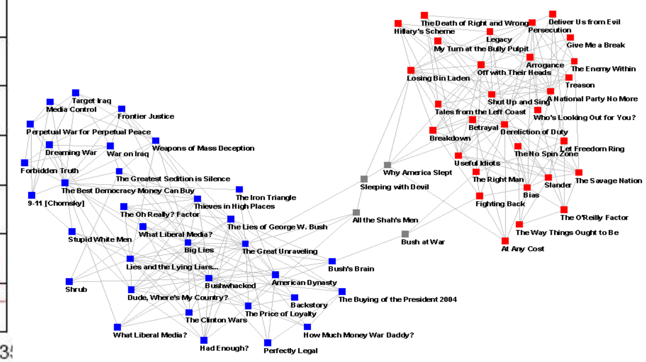
- We proposed a visual data mining approach to detect overlapping communities [Chen et al. 2009].
- Given a start node, the approach first generates a sequence of nodes with their highest "*reachability score"* to former nodes in the list.
    - similar to the well-known visual data mining approach OPTICS.
- A 2D visualization is then built to show the community structure, with "mountain" and "valley" curves.

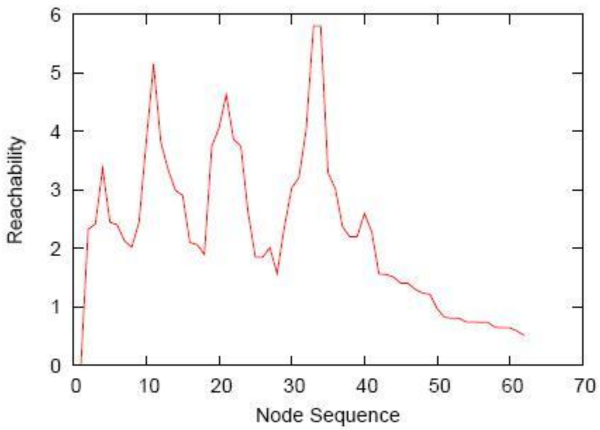


(a) Political Book Network

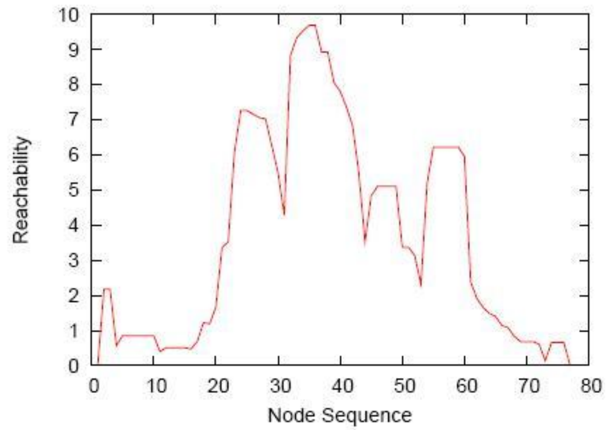


(b) Mexican Politician Network

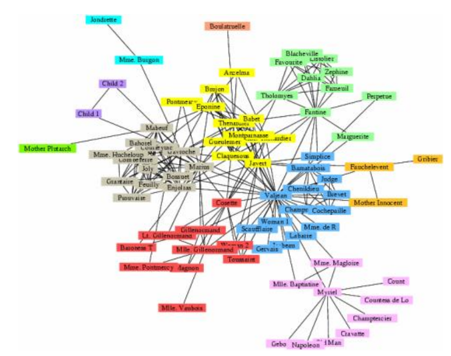




(c) Dolphin Network



(d) Les Miserables Network



Topic Hierarchy

Fact : Search engines always return a long list of pages, ranked by relevance to the query.

Problem : One query may have multiple meanings, and pages on different meanings are mixed and returned together.

Jaguar:



Car
Animal
Operating System
…

Matrix:

A Matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

In math
The movie
…

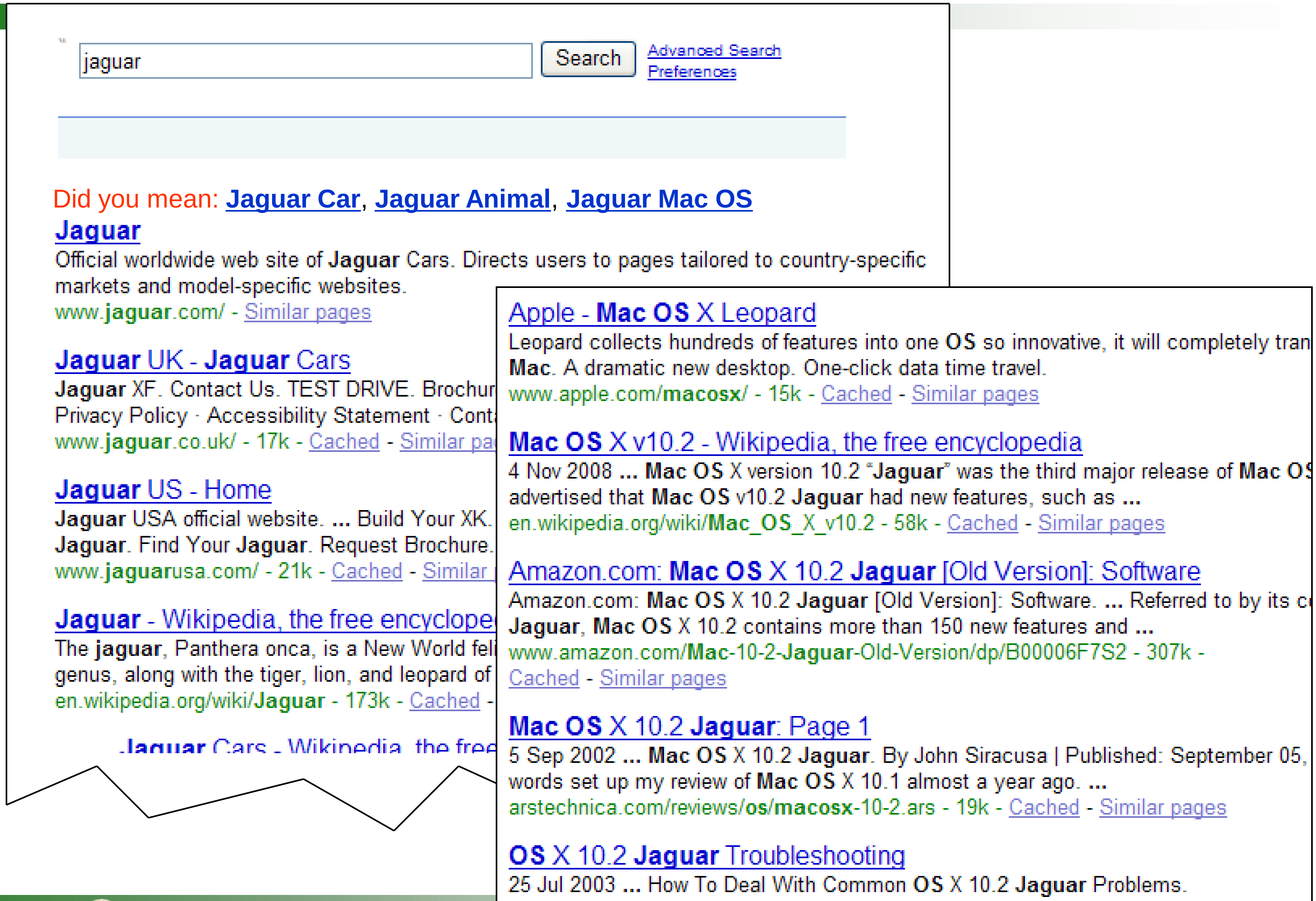
Java:


Coffee
Island
Language
…

Eclipse:



Solar Eclipse
Mitsubishi
IDE
…

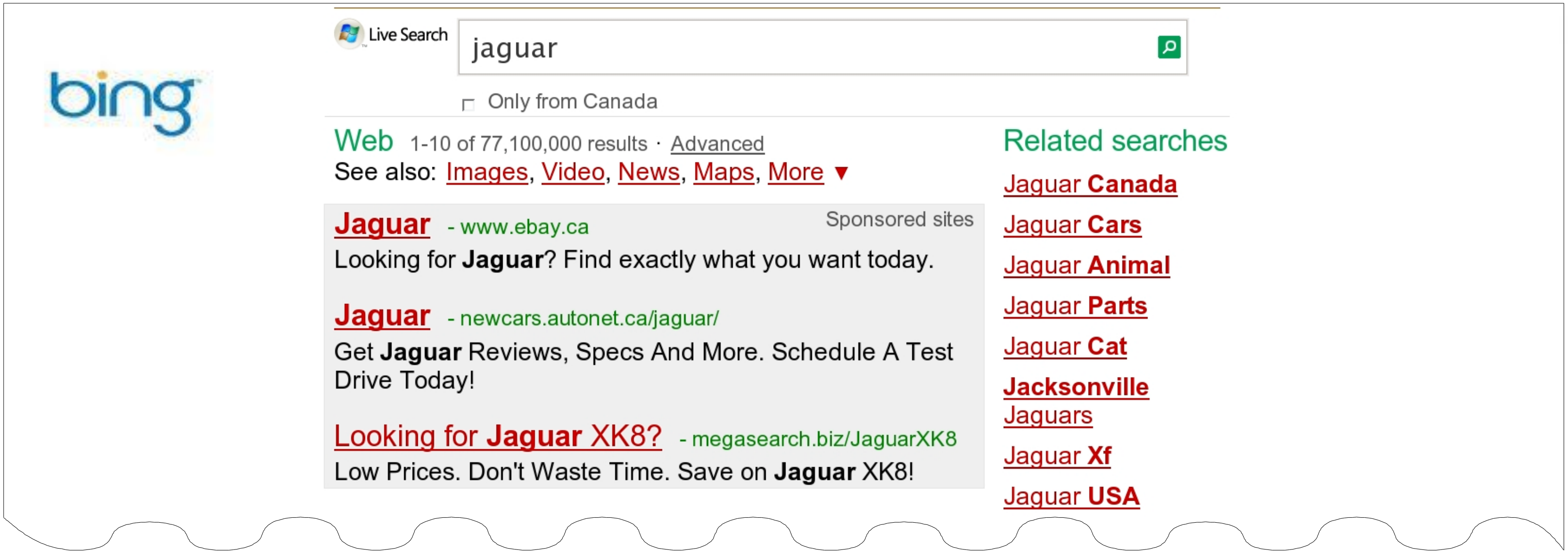

jaguar
Search
Advanced Search
Preferences
Did you mean: Jaguar Car, Jaguar Animal, Jaguar Mac OS
Jaguar
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets and model-specific websites.
www.jaguar.com/ - Similar pages
Jaguar UK - Jaguar Cars
Jaguar XF. Contact Us. TEST DRIVE. Brochur
Privacy Policy · Accessibility Statement · Cont
www.jaguar.co.uk/ - 17k - Cached - Similar pa
Jaguar US - Home
Jaguar USA official website. ... Build Your XK.
Jaguar. Find Your Jaguar. Request Brochure.
www.jaguarusa.com/ - 21k - Cached - Similar
Jaguar - Wikipedia, the free encyclope
The jaguar, Panthera onca, is a New World feli
genus, along with the tiger, lion, and leopard of
en.wikipedia.org/wiki/Jaguar - 173k - Cached -
Apple - Mac OS X Leopard
Leopard collects hundreds of features into one OS so innovative, it will completely tran
Mac. A dramatic new desktop. One-click data time travel.
www.apple.com/macosx/ - 15k - Cached - Similar pages
Mac OS X v10.2 - Wikipedia, the free encyclopedia
4 Nov 2008 ... Mac OS X version 10.2 "Jaguar" was the third major release of Mac O
advertised that Mac OS v10.2 Jaguar had new features, such as ...
en.wikipedia.org/wiki/Mac_OS_X_v10.2 - 58k - Cached - Similar pages
Amazon.com: Mac OS X 10.2 Jaguar [Old Version]: Software
Amazon.com: Mac OS X 10.2 Jaguar [Old Version]: Software. ... Referred to by its c
Jaguar, Mac OS X 10.2 contains more than 150 new features and ...
www.amazon.com/Mac-10-2-Jaguar-Old-Version/dp/B00006F7S2 - 307k -
Cached - Similar pages
Mac OS X 10.2 Jaguar: Page 1
5 Sep 2002 ... Mac OS X 10.2 Jaguar. By John Siracusa | Published: September 05,
words set up my review of Mac OS X 10.1 almost a year ago. ...
arstechnica.com/reviews/os/macosx-10-2.ars - 19k - Cached - Similar pages
OS X 10.2 Jaguar Troubleshooting
25 Jul 2003 ... How To Deal With Common OS X 10.2 Jaguar Problems.

Existing Solutions

Google

YAHOO!

Existing Solutions



Disadvantages

- Suggestions are solely based on search query logs, but the “right” query might not be frequently searched.
- Result for refined queries may still contain mixed information, i.e., pages on different topics.

Our Approach

- Intuition :

  The context in which a word appears is usually related to its sense.
- Word Sense Community:

  A group of words or phrases that co-appear frequently in a set of search result pages.
- Basic idea :

  Cluster the pages into different groups based on word sense community disambiguation.

Approach Procedure

- Phase I :

  Extract keywords from crawled documents.
- Phase II :

  Generate a frequency-based keyword network. Each edge represent the co-occurrence of two words in one sentence.
- Phase III :

  Find communities in the network by applying a hierarchical clustering algorithm which maximizes a network structure metric: Q

Approach Procedure

- Phase IV :

  Refine the communities to eliminate noise.

- Phase V :

  Assign pages to each sense communities to form clusters and return the result to the user.

- Automatic Labeling :

  A dependency-based word relation dataset is used to select the representative word of a word set.

Experiment Data and Labeling

- Evaluation datasets.

  Merged: Amazon, Java, Eclipse

  Real: Jaguar, Salsa

  Large: Reuters

- Manual labeling for ground truth.

Dataset	Manual Labels	Page Set Size
amazon	river, warrior, company	114
java	software, island, coffee	119
eclipse	car, solar, java	125
jaguar	car, animal, mac	101
salsa	dance, sauce	85
Reuters*	Trade, Crude, Money-fx	946

Experiment Results

DataSet	Manual Label	Dependency-based Keyword	ARI score			Q score
			Our Method	K-means	Human h	
Amazon	River	lake, river, water, ocean, forest	0.888	0.693	1	0.367
	Warrior	girl, battle, woman, artist, writer				
	Company	computer, consumer, rate, database				
Java	Coffee	coffee, fruit, tea, vegetable, sugar	0.889	0.728	0.964	0.403
	Island	island, mountain, city, coast, resort				
	Software	software, interface, graphic, application				
Eclipse	Car	engine, car, video, audio, vehicle	0.931	0.765	0.955	0.428
	Solar	sun, picture, moon, earth, light				
	Java	software, interface, server, application				
Jaguar	Animal	animal, wildlife, forest, tiger, bird	0.785	0.114	0.961	0.471
	Car	car, vehicle, truck engine, sedan				
	Mac	database, software, interface, file, server				
Salsa	Dance	music, dance, teacher, jazz, musician	0.642	0.605	0.974	0.405
	Sauce	garlic, tomato, onion, sauce, lemon				
Reuter	Trade	budget, tax, tariff, export, import	0.618	0.504	1	0.222
	Crude	oil, crude, supply, price, output				
	Money-fx	currency, market, dollar, rate, franc				

Use Q to Measure Clustering Result

Query	Extracted Cluster Label	Q Score	Refined Query	Q Score
columbia	student, professor, institute park, town, area order, court, request river, water, lake state, district, county market, film, product season, mph, game	0.369	british columbia	0
saturn	vehicle, technology, model heat, water, hydrogen	0.259	saturn car	0.012
matrix	system, tool, database character, film, movie order, equation, rule	0.330	matrix movie	0.033
blizzard	snow, wind, weather software, product, computer	0.234	blizzard game	0.006
latex	file, format, user patient, treatment, hospital	0.401	latex allergy	0
trailblazer	student, professor, director boy, kid, book network, phone, technology car, vehicle, engine	0.324	trailblazer chevrolet	0.076
mouse	model, study, cell interface, keyboard, device animal, rat, cat film, art, character	0.316	mouse keyboard	0
tiger	animal, wildlife, habitat business, market, industry game, team, player software, user, version	0.356	tiger woods	0.209
tiger woods	game, mode, player tournament, career, record daughter, kid, child	0.209	tiger woods daughter	0.033
casablanca	movie, film, theater city, service, hotel	0.270	casablanca city	0.145
casablanca city	capital, town, region restaurant, hotel, park	0.145	casablanca city hotel	0.004

Conclusions

- Educational applications (on-line applications, CMS, ITS, collaborative tools, forums, etc.) collect a large amount of data.
- This large collection of data is a gold mine to extract patterns to help improve (personalize, make more intelligent…) the applications, to help assess learners' activities.
- In particular, there is a significant opportunity for SNA with synchronous and asynchronous collaborative tools data collection
- Existing DM tools may help, but some problems may require new tools
- These data mining challenges are not uniquely germane to educational applications and the data mining field as a whole can benefit from provided solutions. ➔ Think out of the box.
- Social network analysis, while a century old, in computer science it is still in its infancy. There are myriad open problems for which solutions would be relevant to countless applications beyond EDM.

Thank you – Questions?

UNIVERSITY OF
ALBERTA

Osmar R. Zaïane, Ph.D.
MacCalla-Killam Professor
Department of Computing Science

352 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: zaiane@cs.ualberta.ca
http://www.cs.ualberta.ca/~zaiane/