UNIVERSITY OF ALBERTA

FACULTY OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

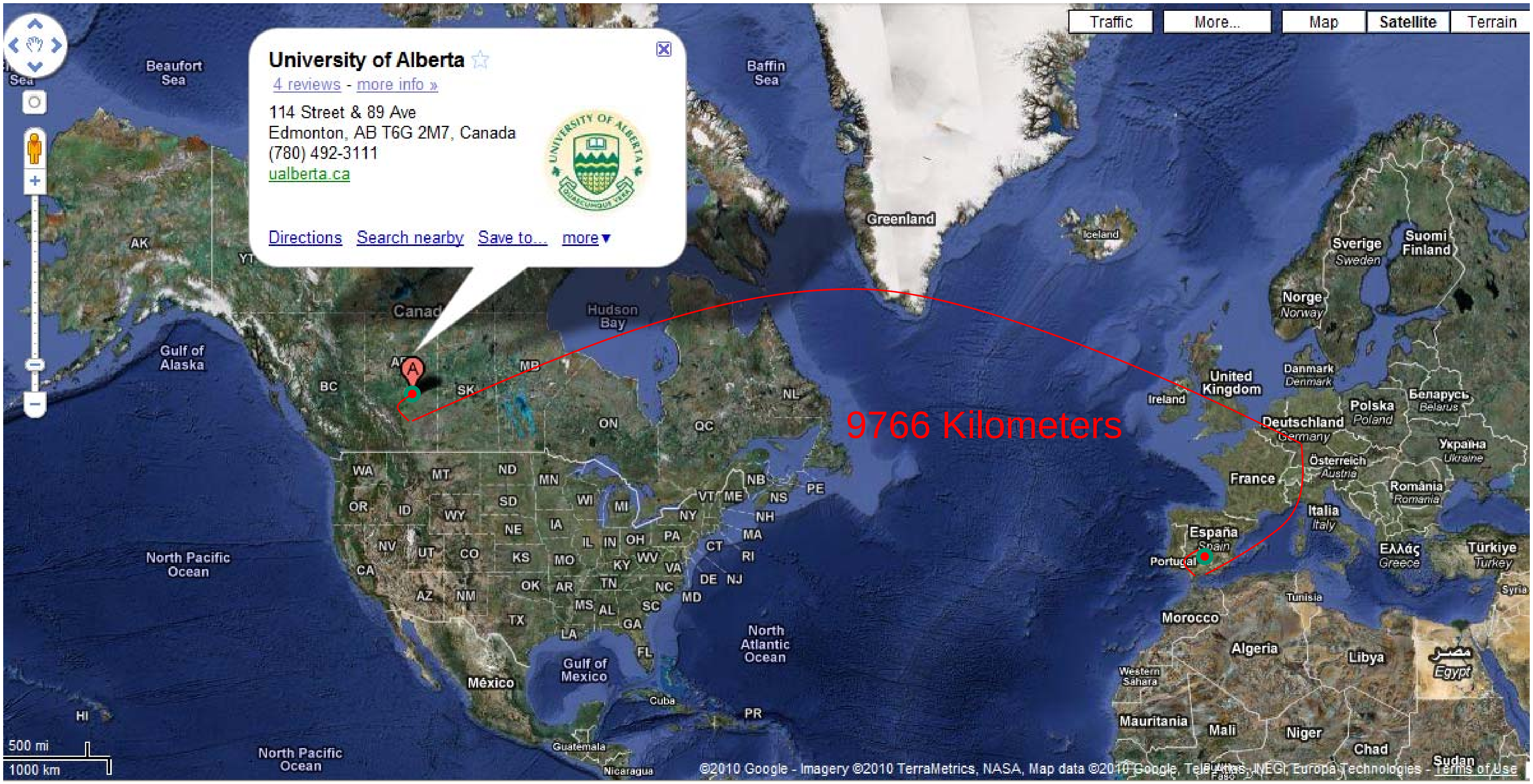
Information Network Analysis: Applications and Challenges

Osmar R. Zaïane

Professor and Scientific Director
Alberta Innovates Centre for
Machine Learning

International Conference on
Intelligent Systems Design and Applications
Cordoba, Spain, November 2011

University of Alberta - Edmonton



Edmonton, capital of Alberta, is the 5th largest city 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

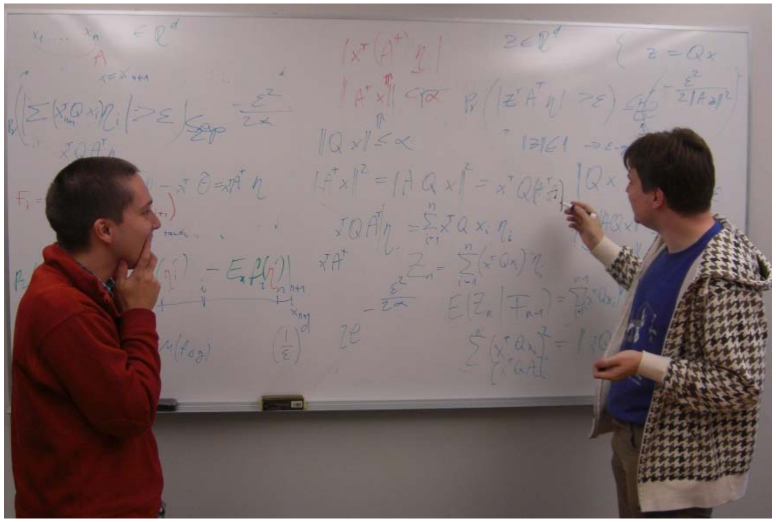
AICML Members

Founded at the University of Alberta in 2002
10 Principal Investigators (academic researchers)

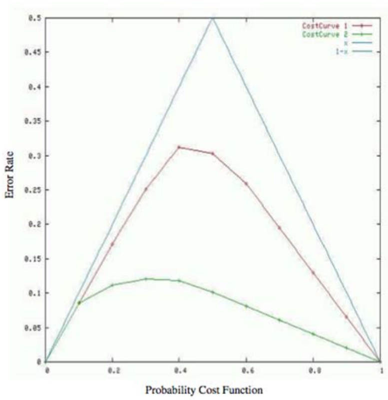
Computing Science Department
124 PhD, 96 MSc

2010-2011: 45 PhD students – 16 PDF – 37 MSc students
24 research and development staff.

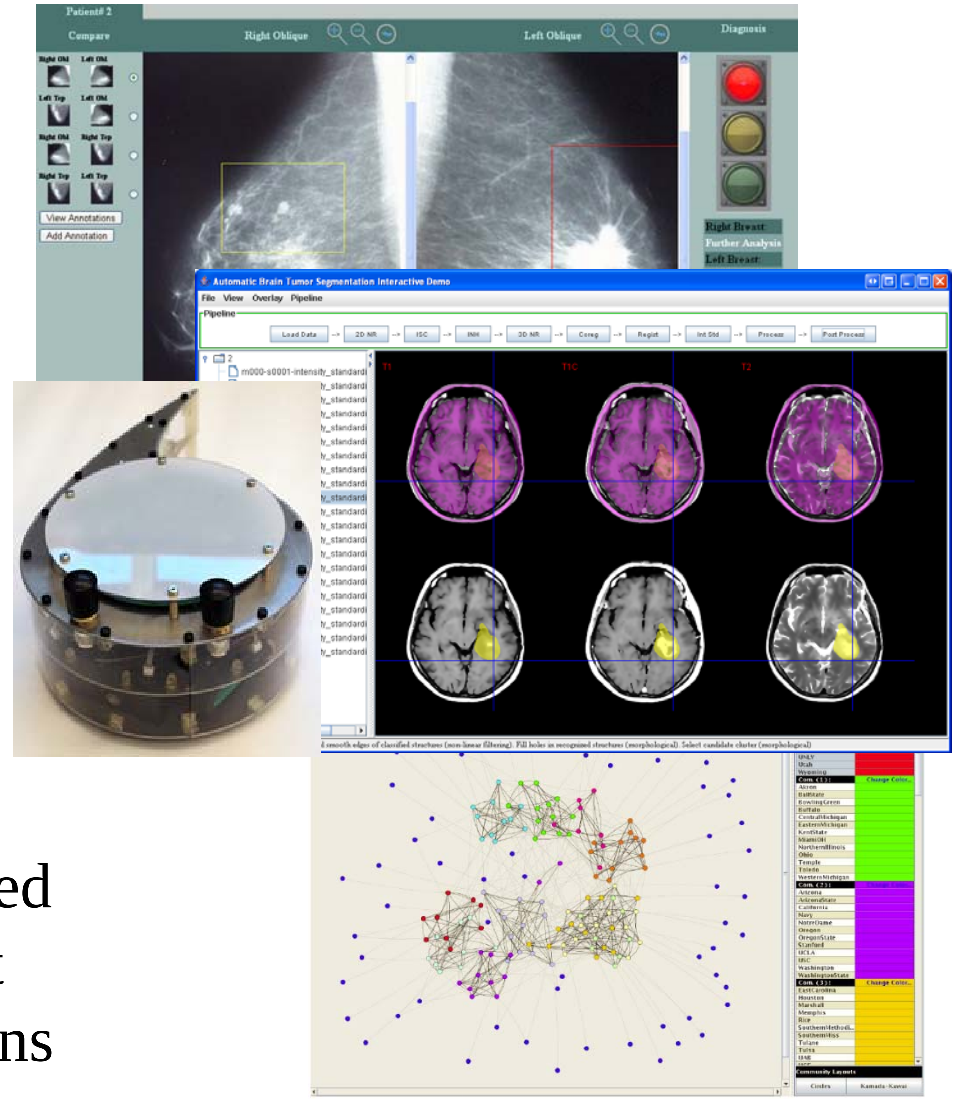
Research at AICML



From fundamental
and practical research



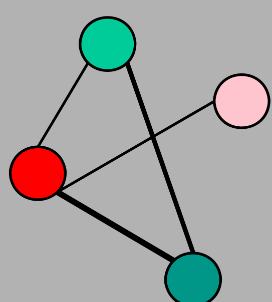
to advanced
intelligent
applications



SNA vs Social Networking

Social Network Analysis Deals with **Information Networks**.

It is NOT **Social Networking**



Nodes are entities
Edges are relationships
Nodes and edges may have attributes

SNA = Analysing such information networks

Hypothetical telecom data

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
1	John Smith	647 225 8085	Toronto	2y	($12)
2	Joe Burns	416 345 6060	Toronto	3y	$724.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
5	Jane Smith	780 233 5645	Edmonton	2y	$673.38
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
9	Wanda Rhymes	403 441 2534	Calgary	3y	$92.00
10	Julie Austinshaur	403 223 7654	Calgary	3y	$983.12
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
13	Megan Potink	780 432 5623	Edmonton	3y	$802.00
14	Kim Cho	780 434 2399	Edmonton	3y	$542.00
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
16	Brian Olso	403 939 7574	Calgary	3y	$430.78
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
19	Greg Aderan	403 332 7468	Calgary	3y	$746.82
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
23	Ryan Waters	403 715 7550	Calgary	3y	$75.50
24	Ben Rikon	403 26			
25	Jun Liu	226 69			
26	Maggie Wong	226 88			
27	Joe Garther	416 22			
28	Karen Pollonts	403 75			
29	Iris Cristle	403 64			
30	Gunther Twallaby	403 77			
31	Monica Kwalshuck	403 21			
32	Fred Couros	416 77			
33	Natalie May	403 40			
34	Aly Huffington	403 255 0304	Calgary	3y	$55.03

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
1	John Smith	647 225 8085	Toronto	2y	($12)
33	Natalie May	403 409 6223	Calgary	3y	$0.96
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
34	Aly Huffington	403 255 0304	Calgary	3y	$55.03
29	Iris Cristle	403 644 1423	Calgary	3y	$64.14
32	Fred Couros	416 773 2234	Toronto	3y	$73.22
23	Ryan Waters	403 715 7550	Calgary	3y	$75.50
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
30	Gunther Twallaby	403 778 6040	Calgary	3y	$78.31
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
25	Jun Liu	226 690 4241	Toronto	3y	$90.42
9	Wanda Rhymes	403 441 2534	Calgary	3y	$92.00
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
16	Brian Olso	403 939 7574	Calgary	3y	$430.78
			on	3y	$459.37
			on	3y	$542.00
			on	3y	$628.01
			on	2y	$673.38
				3y	$724.00
				3y	$746.82
			on	3y	$802.00
			on	3y	$830.00
				3y	$983.12
				3y	$1,044.48
27	Joe Garther	416 224 1109	Toronto	3y	$1,100.10

Not enough profit

Plan	Avg. 3m Profit
3y	($26.23)
2y	($12)
3y	$0.96
3y	$8.00
3y	$33.79
3y	$38.78
1y	$50.18

<u>Assumption</u>:
Customers are **i**ndependent
Values are **i**dentically **d**istributed

34 customers up for plan renewal

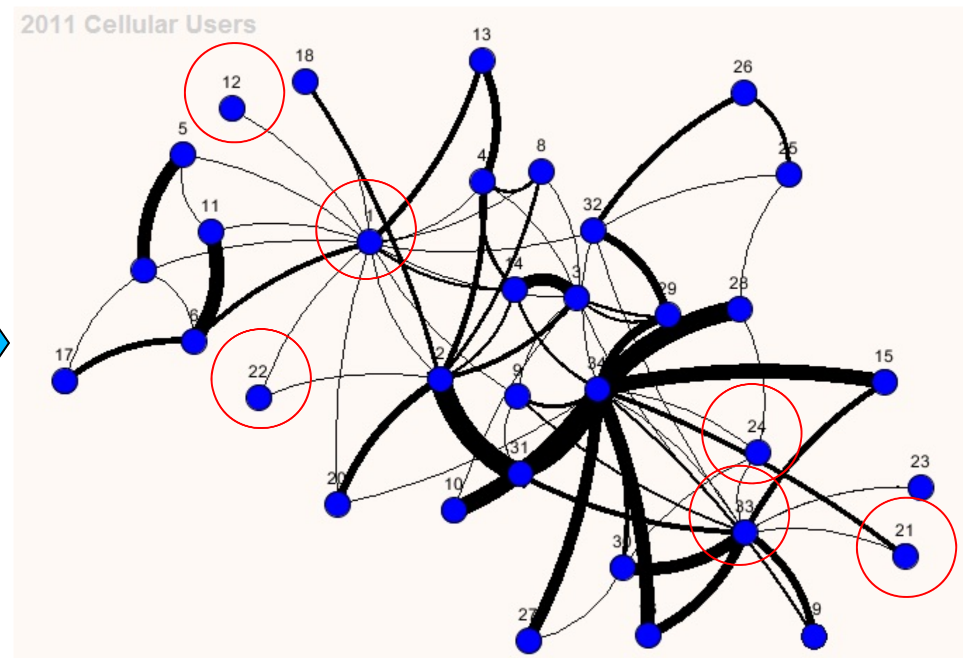
Which one to renew?

Which one to give incentive to stay?

Sort by profit in the last 3 months

Do not renew or give incentive if profit < $50 (?)

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
1	John Smith	647 225 8085	Toronto	2y	($12)
33	Natalie May	403 409 6223	Calgary	3y	$0.96
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
34	Aly Huffington	403 255 0304	Calgary	3y	$55.03
29	Iris Cristle	403 644 1423	Calgary	3y	$64.14
32	Fred Couros	416 773 2234	Toronto	3y	$73.22
23	Ryan Waters	403 715 7550	Calgary	3y	$75.50
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
30	Gunther Twallaby	403 778 6040	Calgary	3y	$78.31
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
25	Jun Liu	226 690 4241	Toronto	3y	$90.42
9	Wanda Rhymes	403 441 2534	Calgary	3y	$92.00
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
16	Brian Olso	403 939 7574	Calgary	3y	$430.78
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
14	Kim Cho	780 434 2399	Edmonton	3y	$542.00
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
5	Jane Smith	780 233 5645	Edmonton	2y	$673.38
2	Joe Burns	416 345 6060	Toronto	3y	$724.00
19	Greg Aderan	403 332 7468	Calgary	3y	$746.82
13	Megan Potink	780 432 5623	Edmonton	3y	$802.00
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
10	Julie Austinshaur	403 223 7654	Calgary	3y	$983.12
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48
27	Joe Garther	416 224 1109	Toronto	3y	$1,100.10

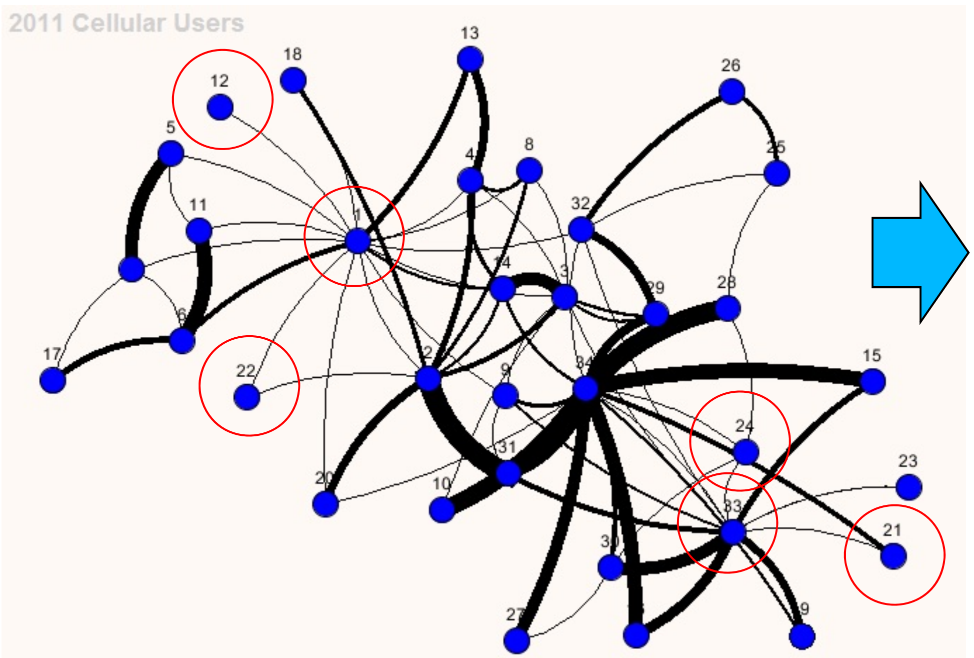


34 customers up for plan renewal

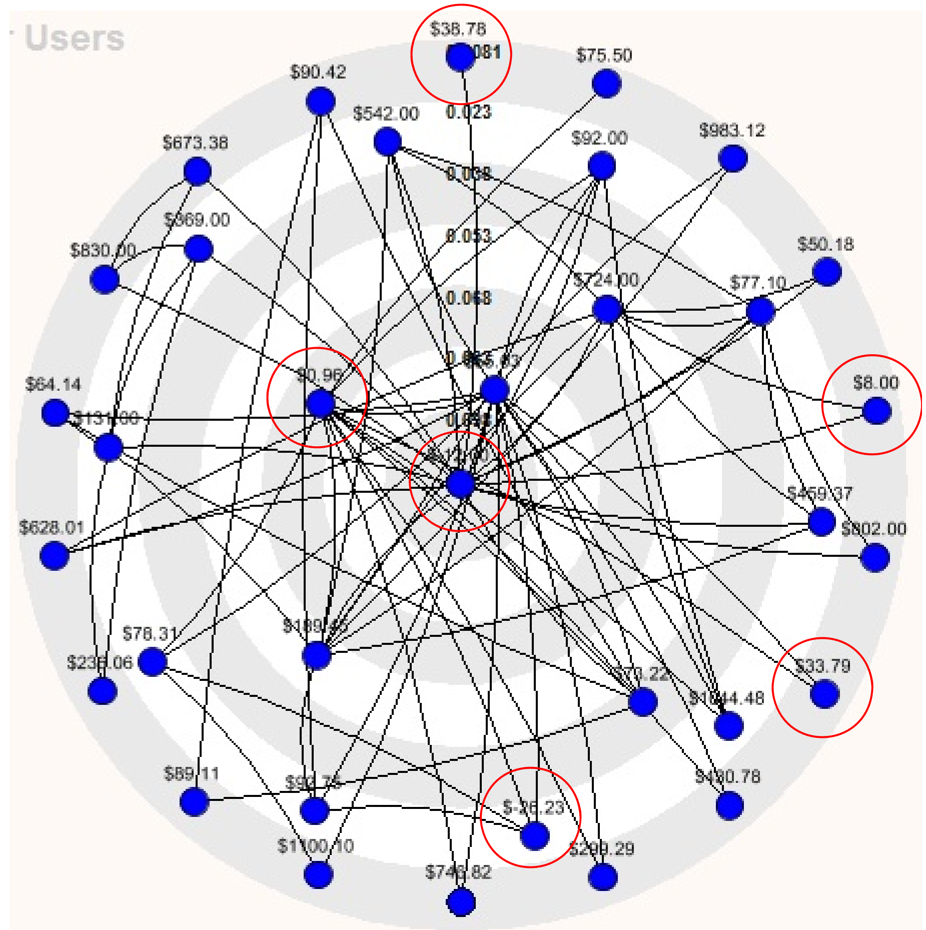
Which one to renew?

Which one to give incentive to stay?

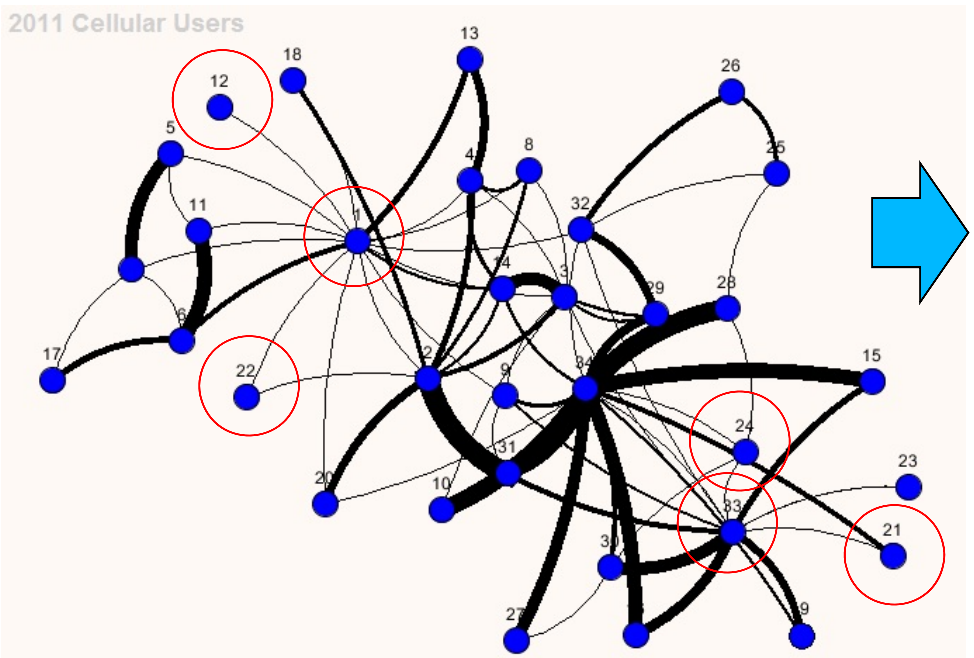
Inter-call network with call frequency



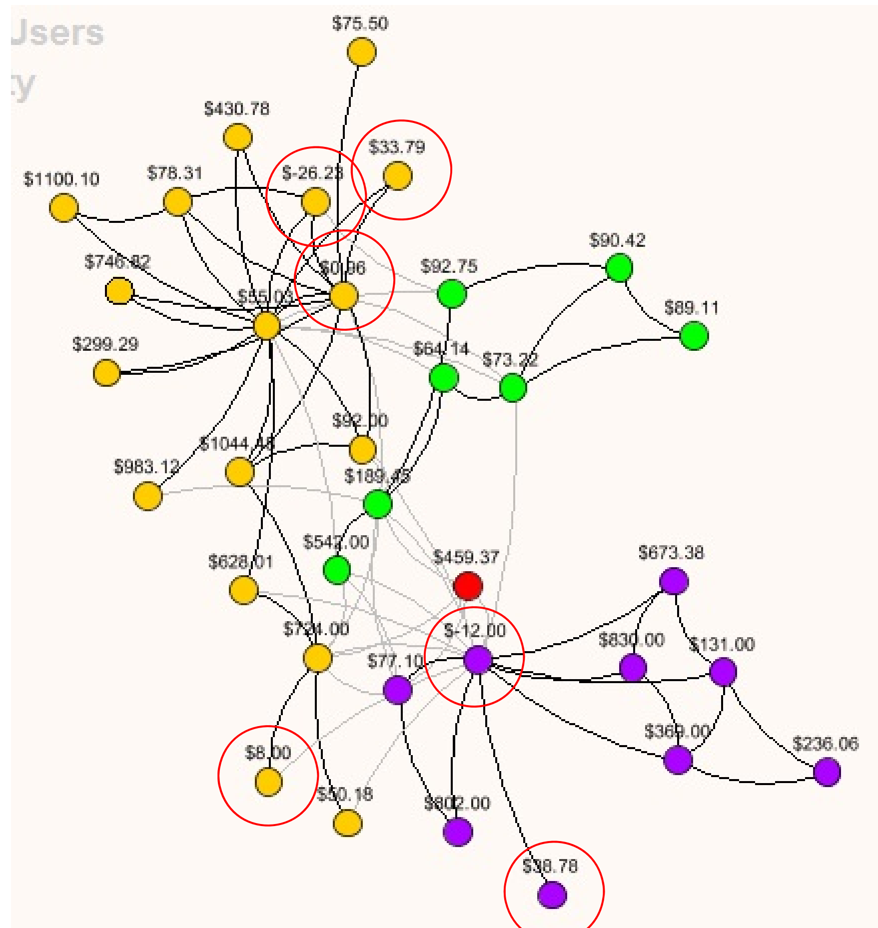
Inter-call network with call frequency



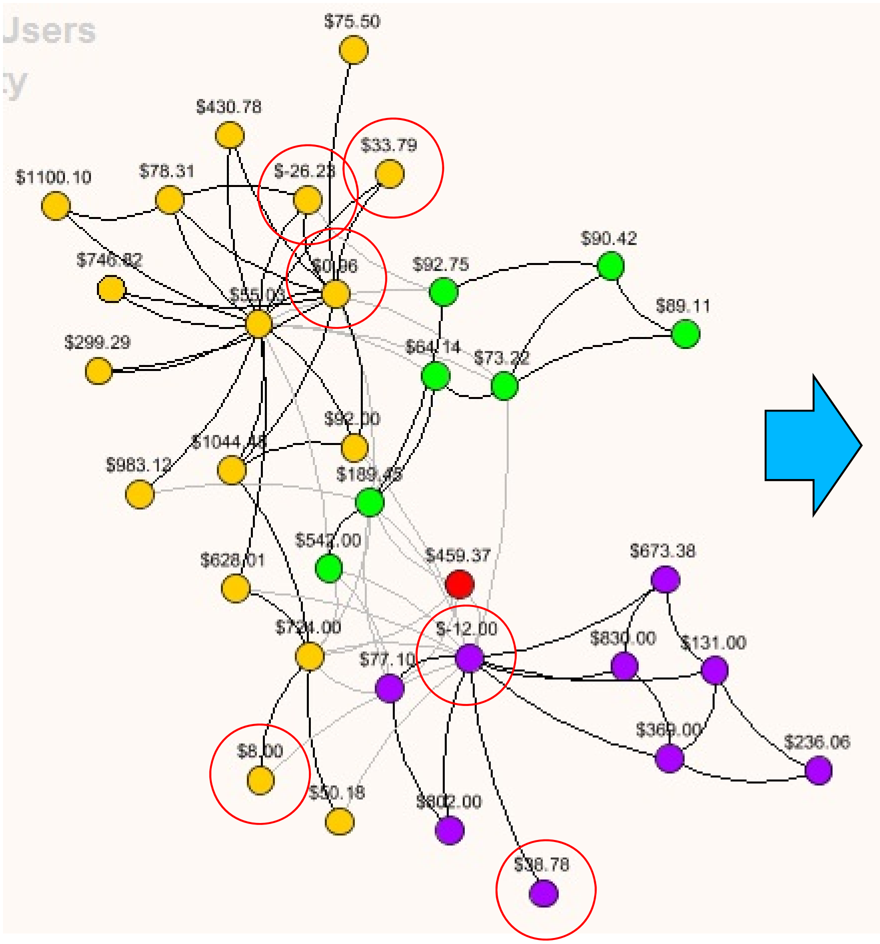
Global centrality based PageRank



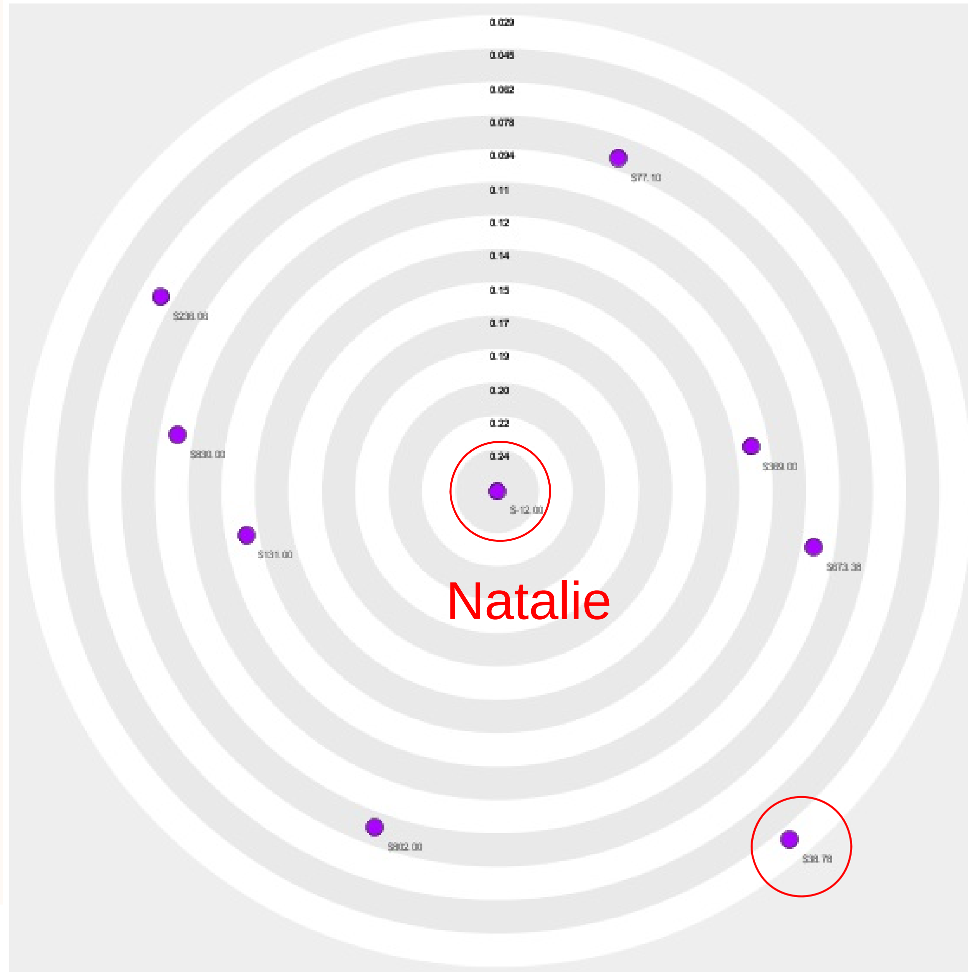
Inter-call network with call frequency



Community Mining

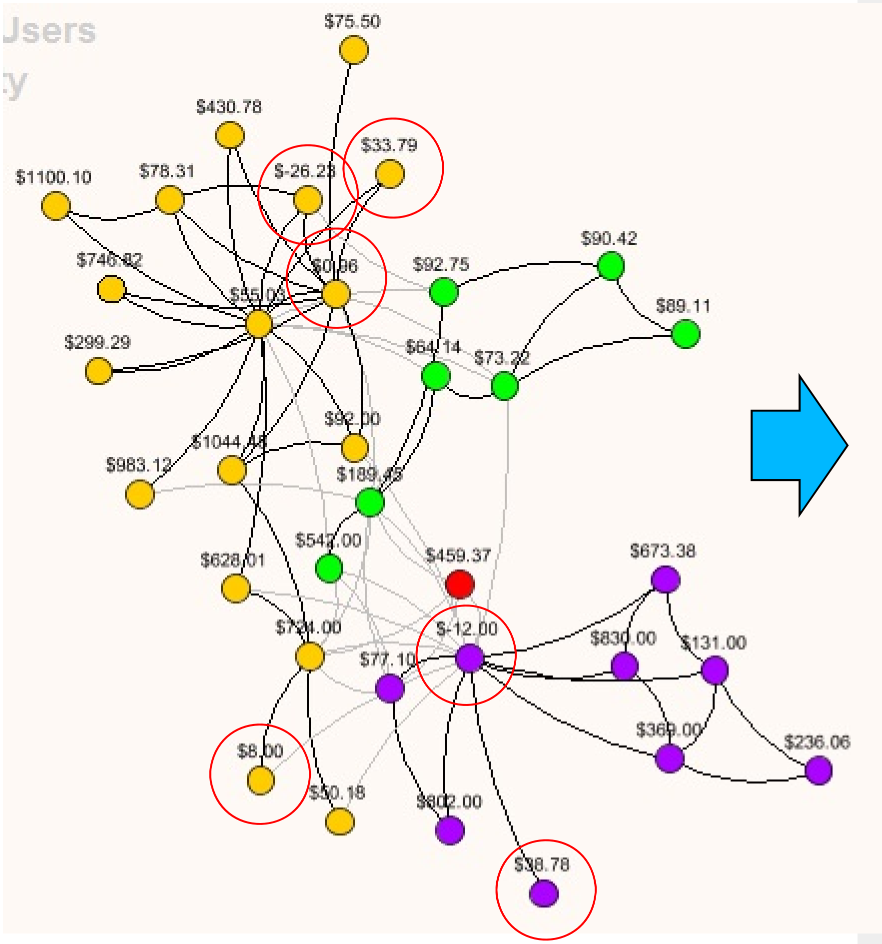


Community Mining

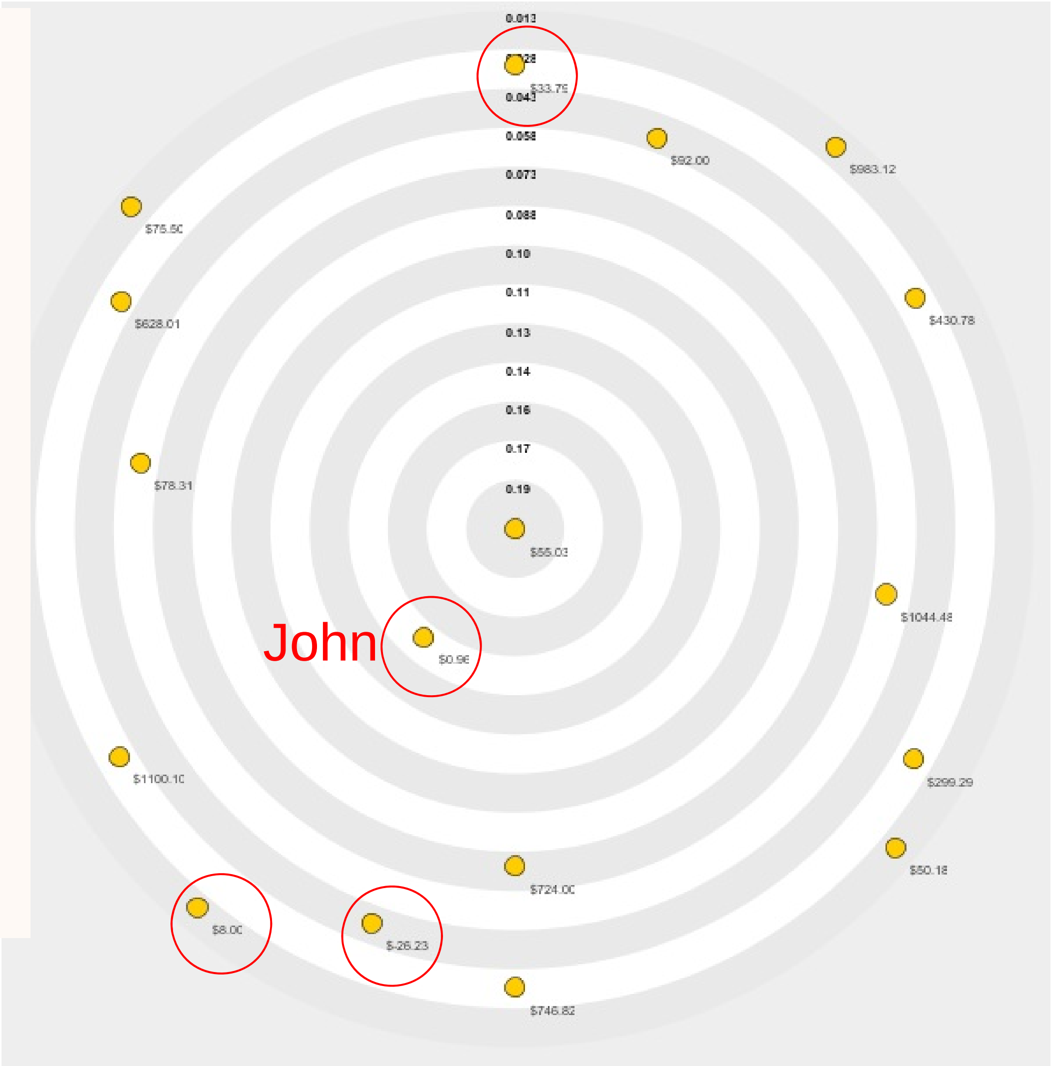


Centrality per community

Dropping Natalie: Risk = $3145.32



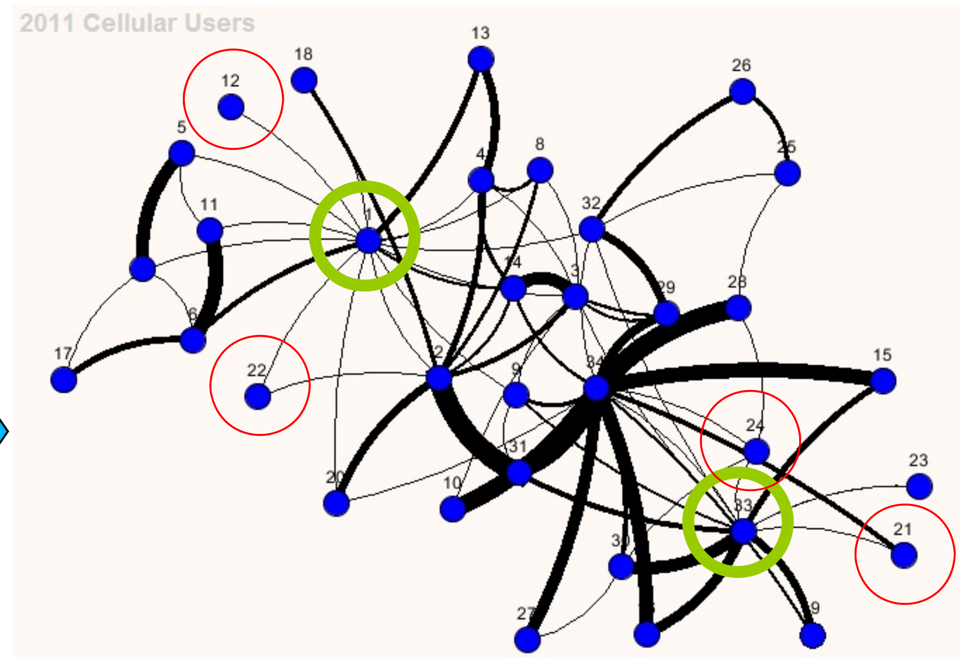
Community Mining



Centrality per community

Dropping John: Risk = $6324.14

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
1	John Smith	647 225 8085	Toronto	2y	($12)
33	Natalie May	403 409 6223	Calgary	3y	$0.96
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
34	Aly Huffington	403 255 0304	Calgary	3y	$55.03
29	Iris Cristle	403 644 1423	Calgary	3y	$64.14
32	Fred Couros	416 773 2234	Toronto	3y	$73.22
23	Ryan Waters	403 715 7550	Calgary	3y	$75.50
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
30	Gunther Twallaby	403 778 6040	Calgary	3y	$78.31
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
25	Jun Liu	226 690 4241	Toronto	3y	$90.42
9	Wanda Rhymes	403 441 2534	Calgary	3y	$92.00
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
16	Brian Olso	403 939 7574	Calgary	3y	$430.78
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
14	Kim Cho	780 434 2399	Edmonton	3y	$542.00
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
5	Jane Smith	780 233 5645	Edmonton	2y	$673.38
2	Joe Burns	416 345 6060	Toronto	3y	$724.00
19	Greg Aderan	403 332 7468	Calgary	3y	$746.82
13	Megan Potink	780 432 5623	Edmonton	3y	$802.00
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
10	Julie Austinshaur	403 223 7654	Calgary	3y	$983.12
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48
27	Joe Garther	416 224 1109	Toronto	3y	$1,100.10



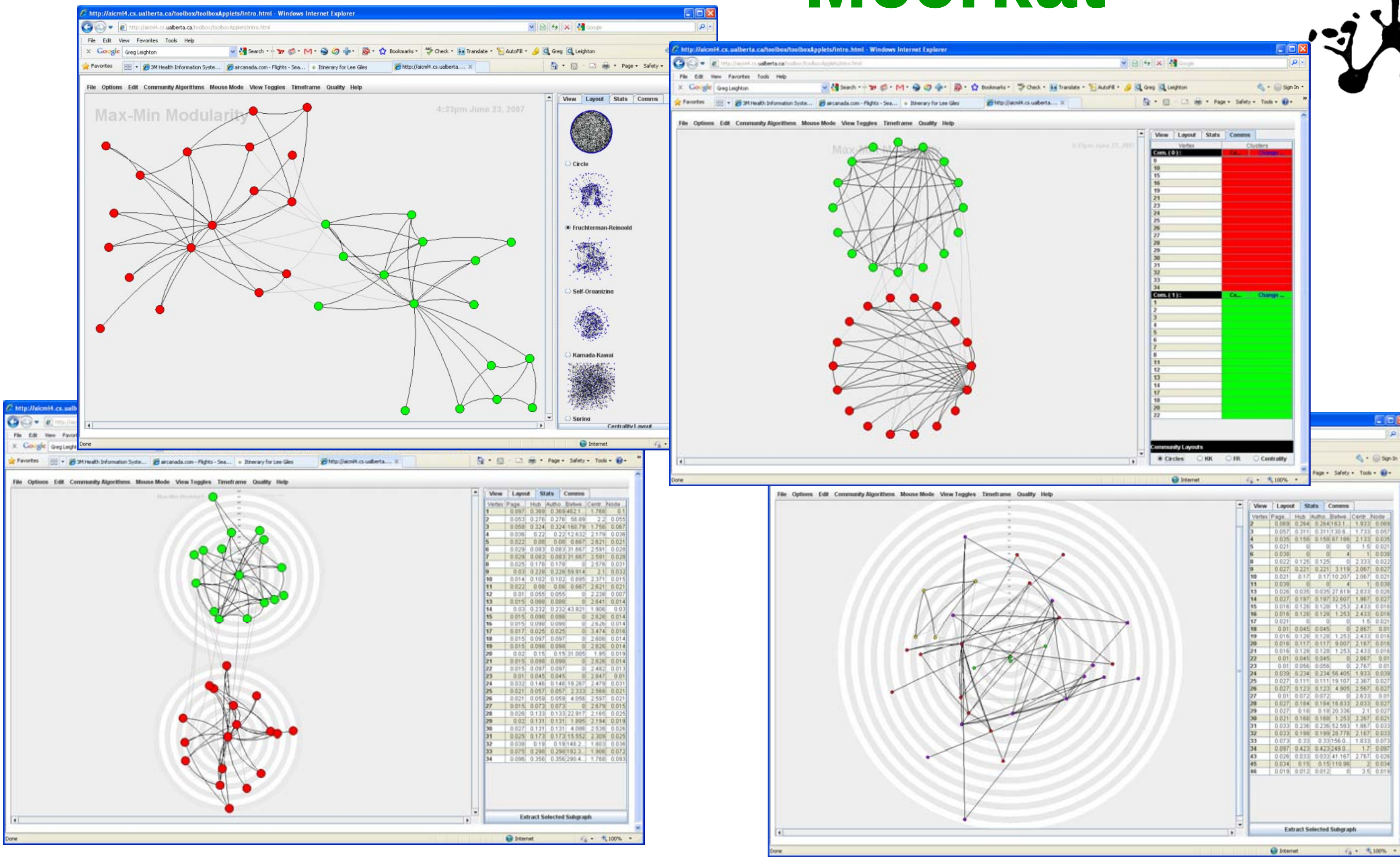
34 customers up for plan renewal

Which one to renew?

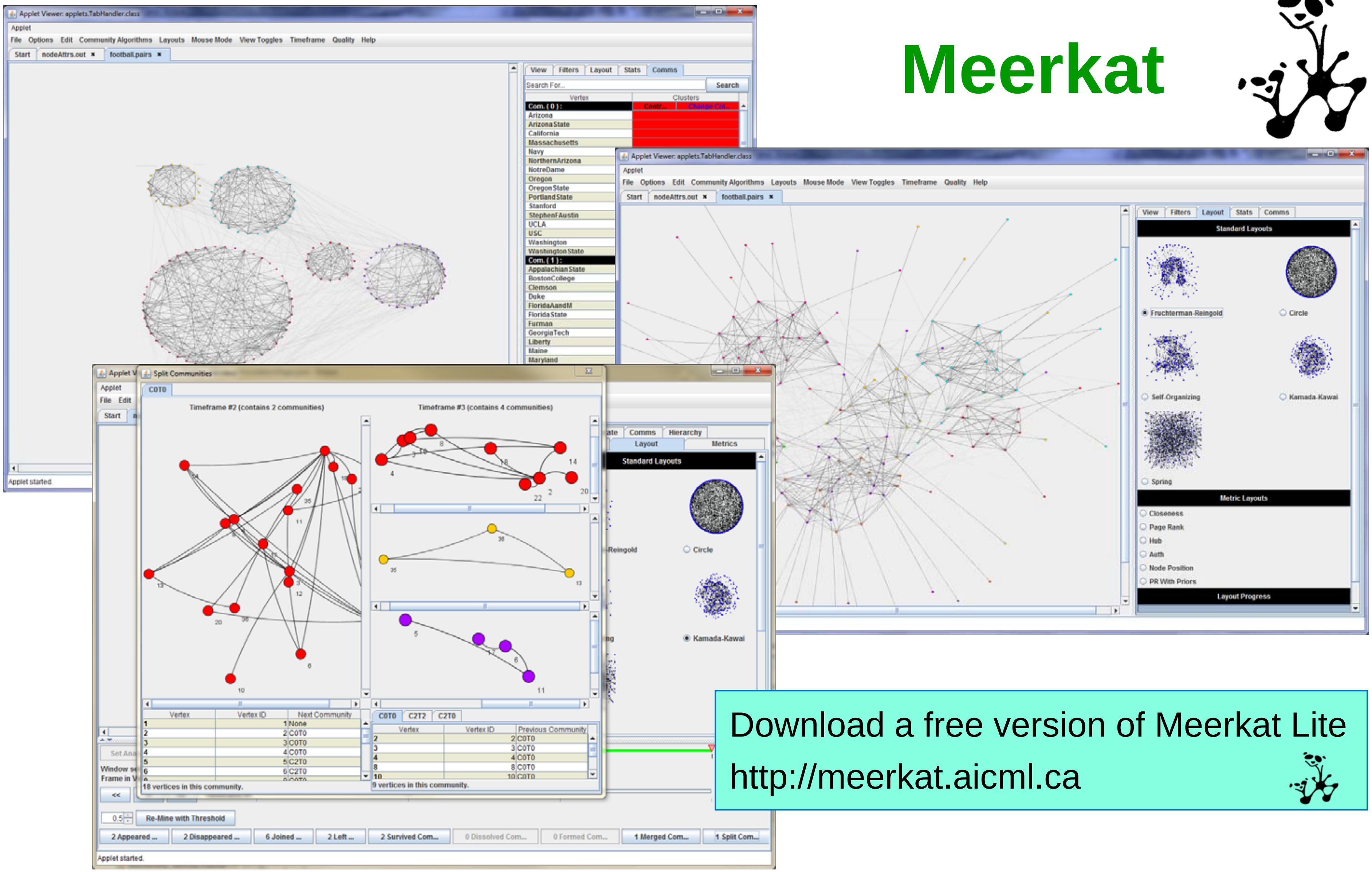
Which one to give incentive to stay?

Give incentives to 1 (John Smith -$12) and 33 (Natalie May $0.96) to stay but let the others go.

Meerkat


Max-Min Modularity
Circle
Fruchterman-Reingold
Self-Organizing
Kamada-Kawai
Extract Selected Subgraph

Meerkat



Download a free version of Meerkat Lite
http://meerkat.aicml.ca

What is Social Network Analysis?

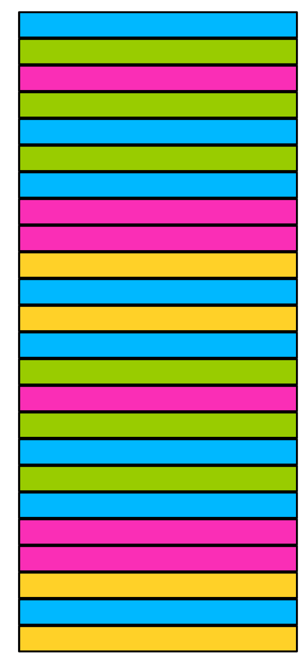
- **[Wikipedia] A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friendship, sexual relationships, kinship, dislike, conflict or trade.**
- **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? ...**

Example of How SNA can Improve Existing Applications: Recommending a Book

- ***Collaborative filtering: The basic idea is that people are recommending items to one another*.**

Example of How SNA can Improve Existing Applications: Recommending a Book

- Build a user profile for user u;
- Predictions for unseen (target) items are computed based the other users' with similar interests on items in user *u*'s profile



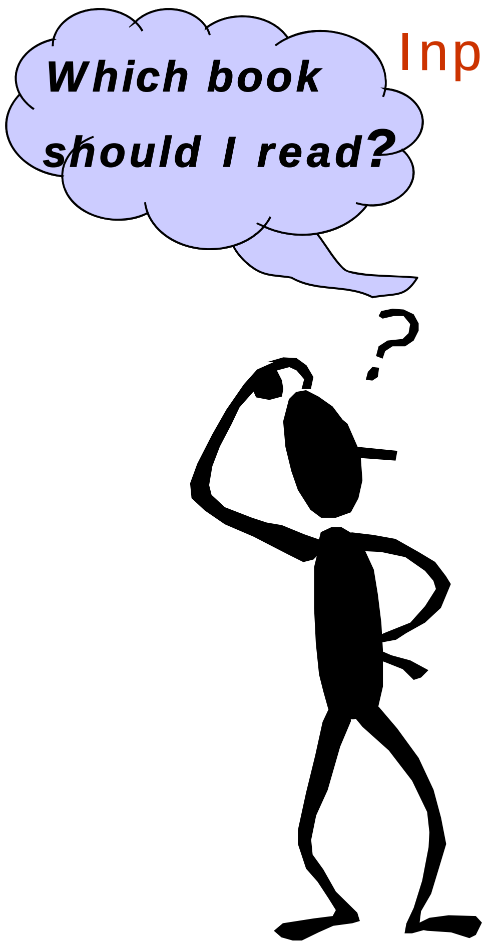
At the heart of Recommender Systems are Collaborative Filtering Algorithms that rely on correlation between individuals

Ratings of Books	1	2	3	4	5	6	7	8
Jane	5	3	3	4	2	1		
Alexander	3	4	2	3	4	5	1	3
Amelia	4	3	1	2	4	2	4	1
Duncan	4	2	1	3	4	1	5	2

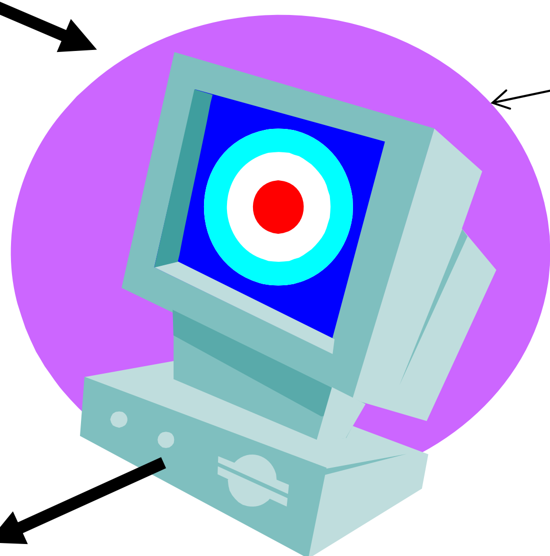
- **Jane & Duncan: correlation = .52**
- **Jane & Alexander: correlation = -.67**
- **Jane & Amelia: correlation = .23**

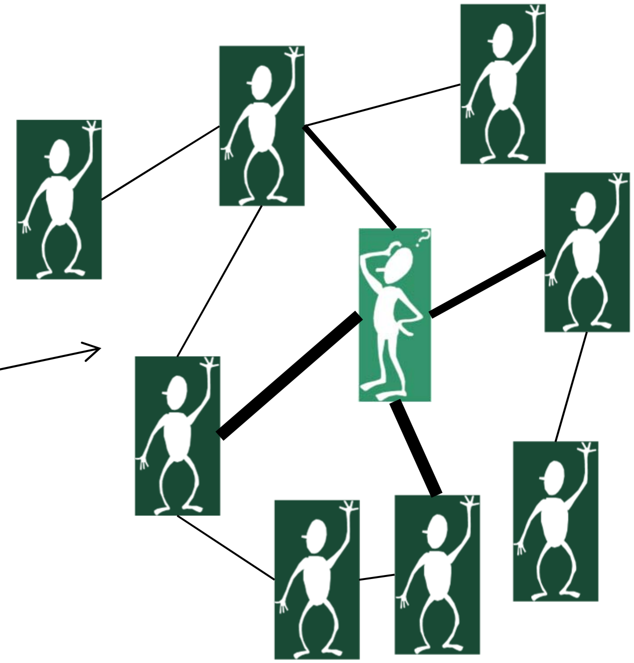
Recommendations for Jane: **Book 7**

Interaction Paradigm of Recommender Systems with SNA



Input (Ratings of Books)





Output (Recommendations):

Books you might enjoy are...

More accurate Recommendation

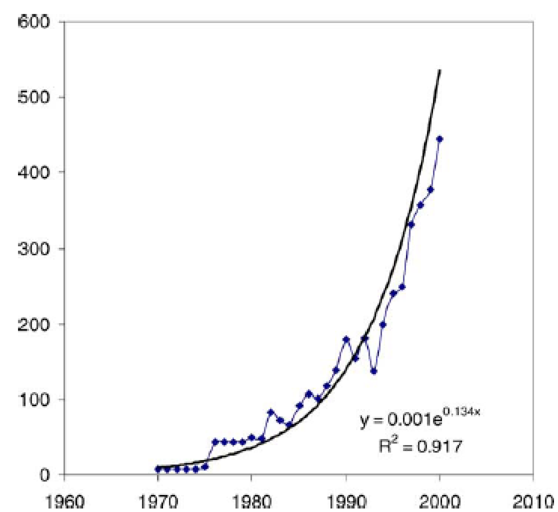
Narrow down neighbourhood
Prioritize similarities
Define similarity

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 (sociometry, sociograms).**

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



- **Precursors of social network theory date from the century such as Simmel, Durkheim and Tönnies.**

- **Massive increase in studies of social networks 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**

Networks in Social and Behavioral Sciences

[Monge, and Contractor, 2003]

- **Social Networks**
  – Who knows who?
- **Socio-cognitive Networks**
  – Who thinks who knows who?
- **Knowledge Networks**
  – Who knows what?
- **Cognitive Knowledge Networks**
  – Who thinks who knows what?

Reality	Social Network	Knowledge Network
Perception	Socio-cognitive Network	Cognitive knowledge Network
	Acquaintance	knowledge

- Socio-centric Analysis
  - Emerged in sociology: quantification of interaction among a group of people. Focus on Identifying global structural patterns in a network.
- Ego-centric Analysis
  - Emerged in psychology and anthropology: quantification of interaction between an individual (ego) and others (alters) directly or indirectly related to ego.

Popularization

- **Six degrees of separation** **(Chains by Frigyes Karinthy 1929)**
  Hypothesized: modern world was 'shrinking' due to the ever-increasing connectedness of human beings. Used the idea of six degrees of freedom in mechanics.

- **Milgram's Paradox: Small world effect** **(Stanley Milgram, 1967)**
  Famous experiment in 1970 sending letters from Omaha to Boston 64/296 arrived. Average path 5.5~6.

- **Google's PageRank (1998) uses a network of web page « citations » to estimate the importance of pages and rank them.**

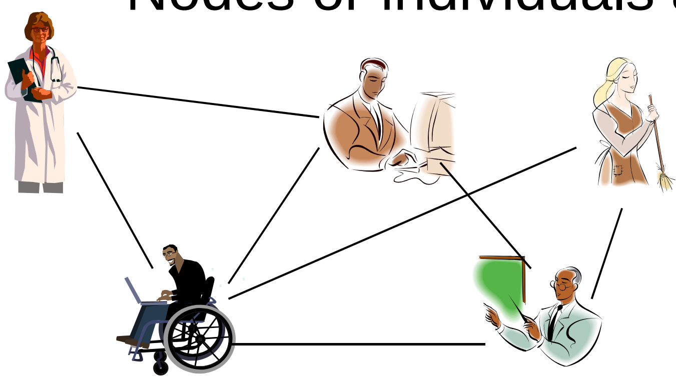
- **Internet social networking tools**

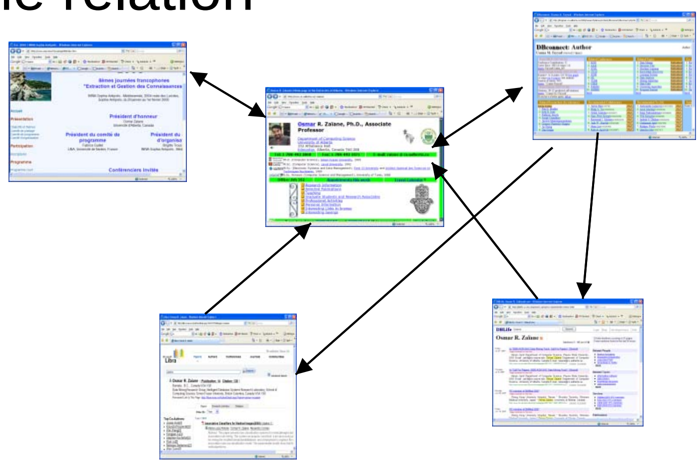
- **Research team in Milan finds degree of separation = 4.74 using 721 million FB users (4.37 in USA) . NYT Nov. 2011**

http://www.nytimes.com/2011/11/22/technology/between-you-and-me-4-74-degrees.html?_r=1

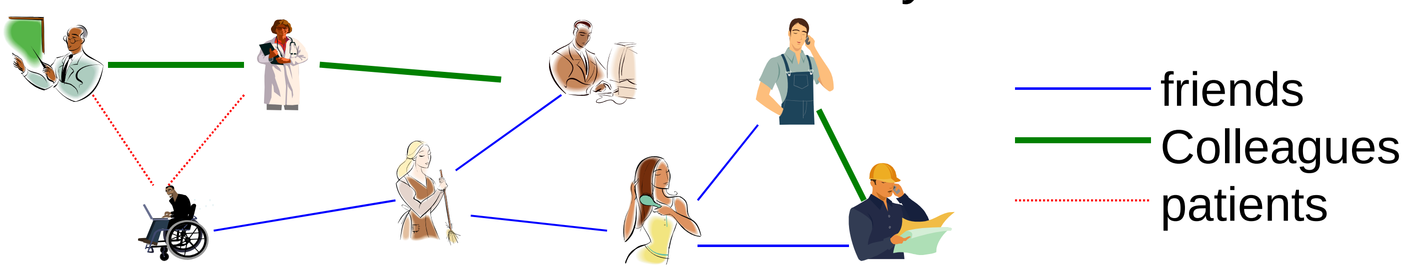
Types of Relations and Networks (1)

- **Unique relation networks**
  – Nodes or individuals are tied by the same relation



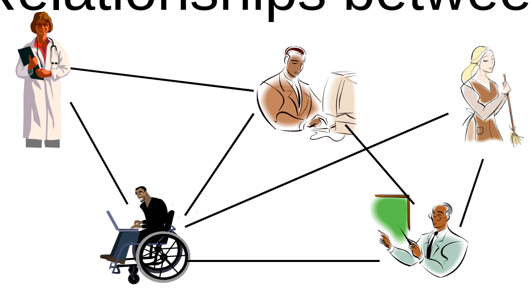


- **Multiple relation networks**
  – Nodes or individuals are tied by different kinds of relationships



Types of Relations and Networks (2)

- **Homogineous relationship**
  – Relationships between nodes of the same type



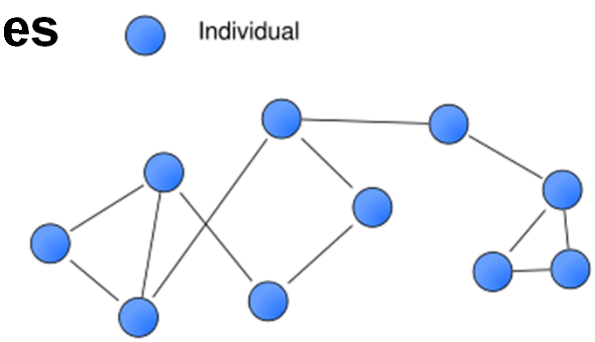
- **Heterogineous relationships**
  – Relationships between nodes of different types

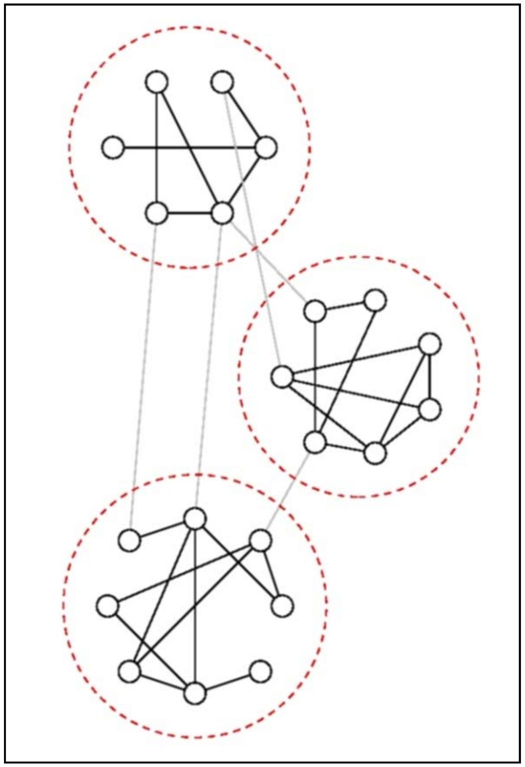
Some Key Concepts

- **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).
- **Peripheral nodes and outliers**
- **Communities**

Some Typical Operations

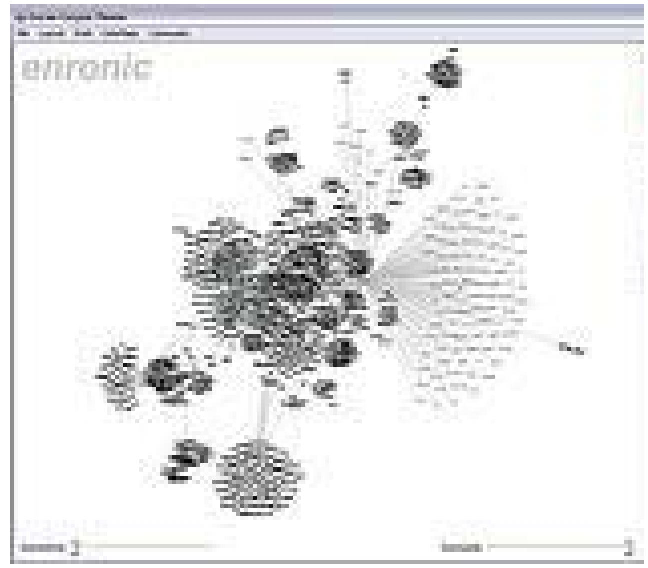
- Visualization of networks
- Filtering/Querying (selecting specific nodes and or edges)
- Finding central nodes (Centrality)
- Ranking nodes
- Finding peripheral nodes
- Community mining
- Discovering outliers
- Predicting unobserved edges
- Discovering dynamics in time





The famous case of Enron

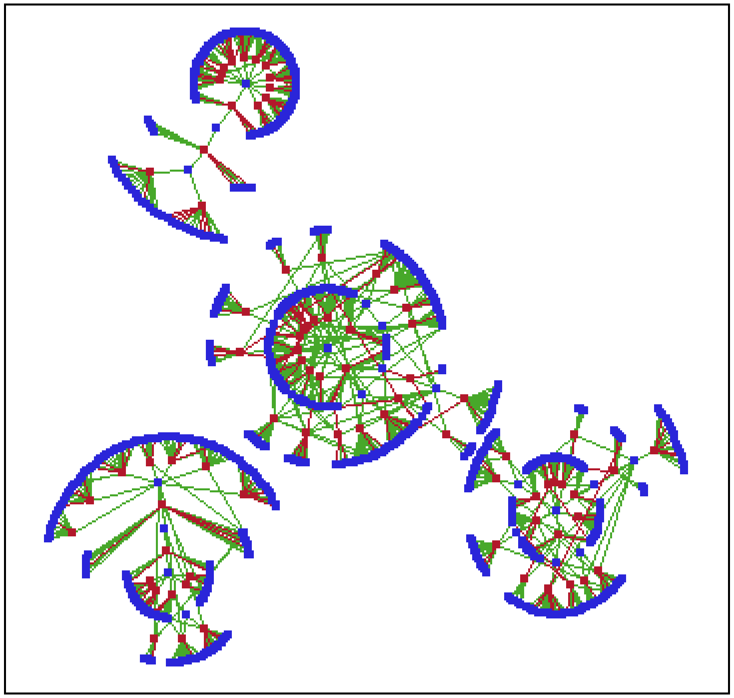
- **Enron E-mail data made public**
  - 151 users
  - 200,399 e-mail messages



Visualization of Enron’s email network, Jeffrey Heer, 2005

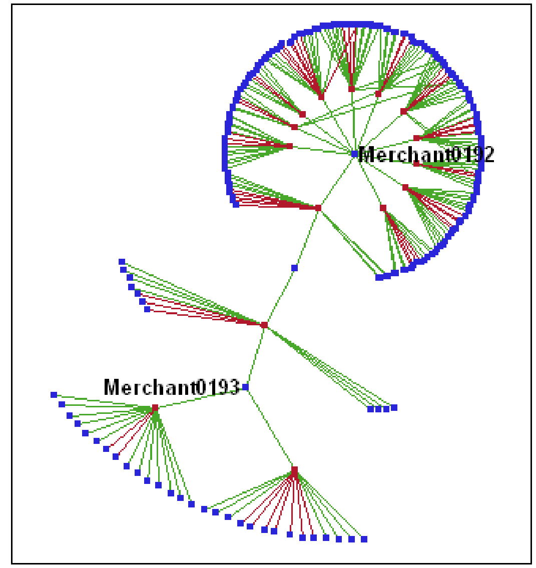
  - Modeling a Socio-Cognitive Network
  - Quantitative Measures for Perceptual Closeness
  - Automatic Extraction of Concealed Relations
  - …

Social Network Analysis to Identify Suspicious Merchants

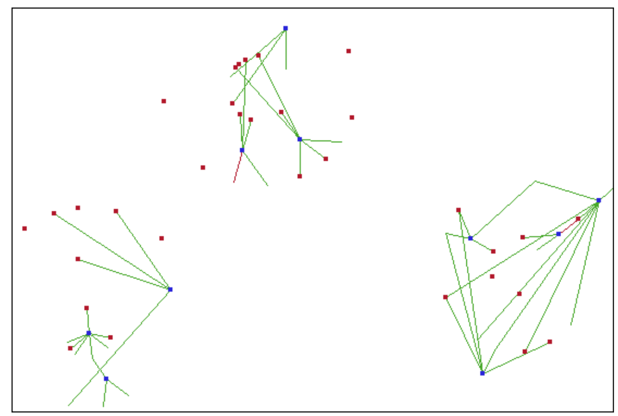


Blue nodes correspond to merchants, red nodes correspond to customers. Each link represents a transaction between a customer and a merchant. Green links correspond to valid transactions, red links correspond to fraudulent transactions

Degree centrality



Filtering



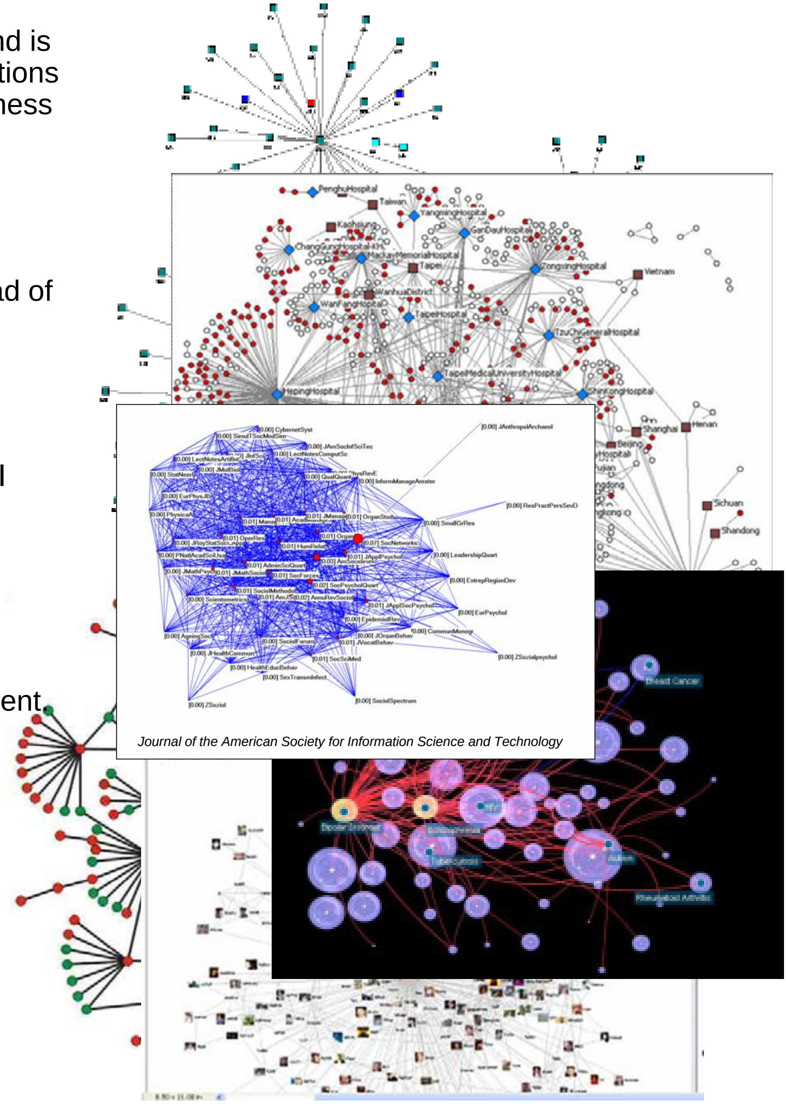
Detect patterns of credit card fraud

Identify merchants that warrant additional scrutiny with regard to fraudulent credit card transactions

Example from SAS

Applications of SNA

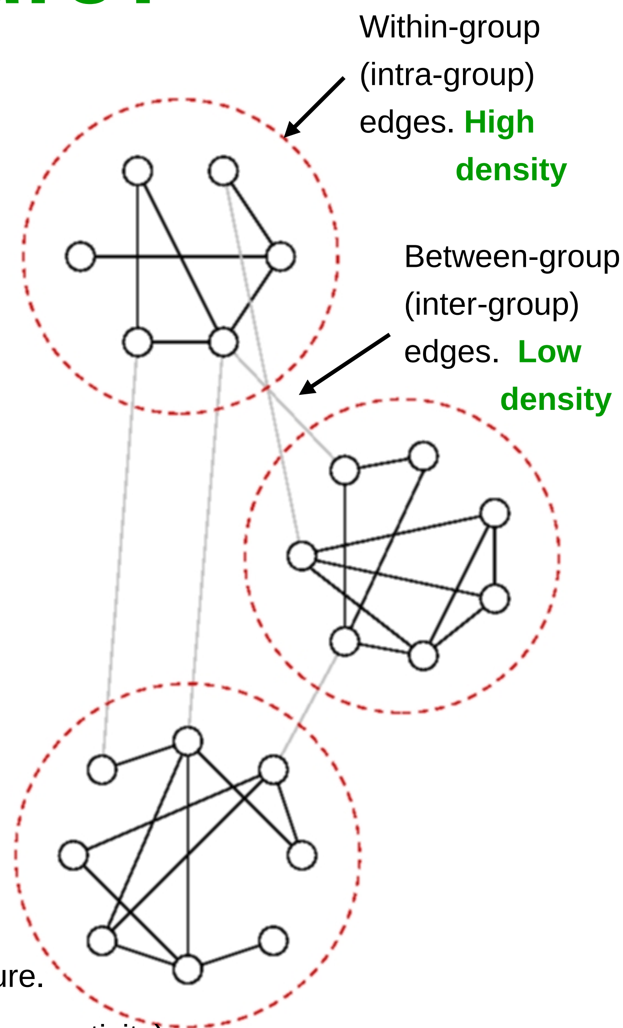
- **Terrorism and crimes**
  - Social Network analysis is an important part of a conspiracy investigation and is used as an investigative tool. Group structure may be important to investigations of racketeering enterprises, narcotics operations, illegal gambling, and business frauds.
- **Medicine – epidemiology**
  - 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**



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**

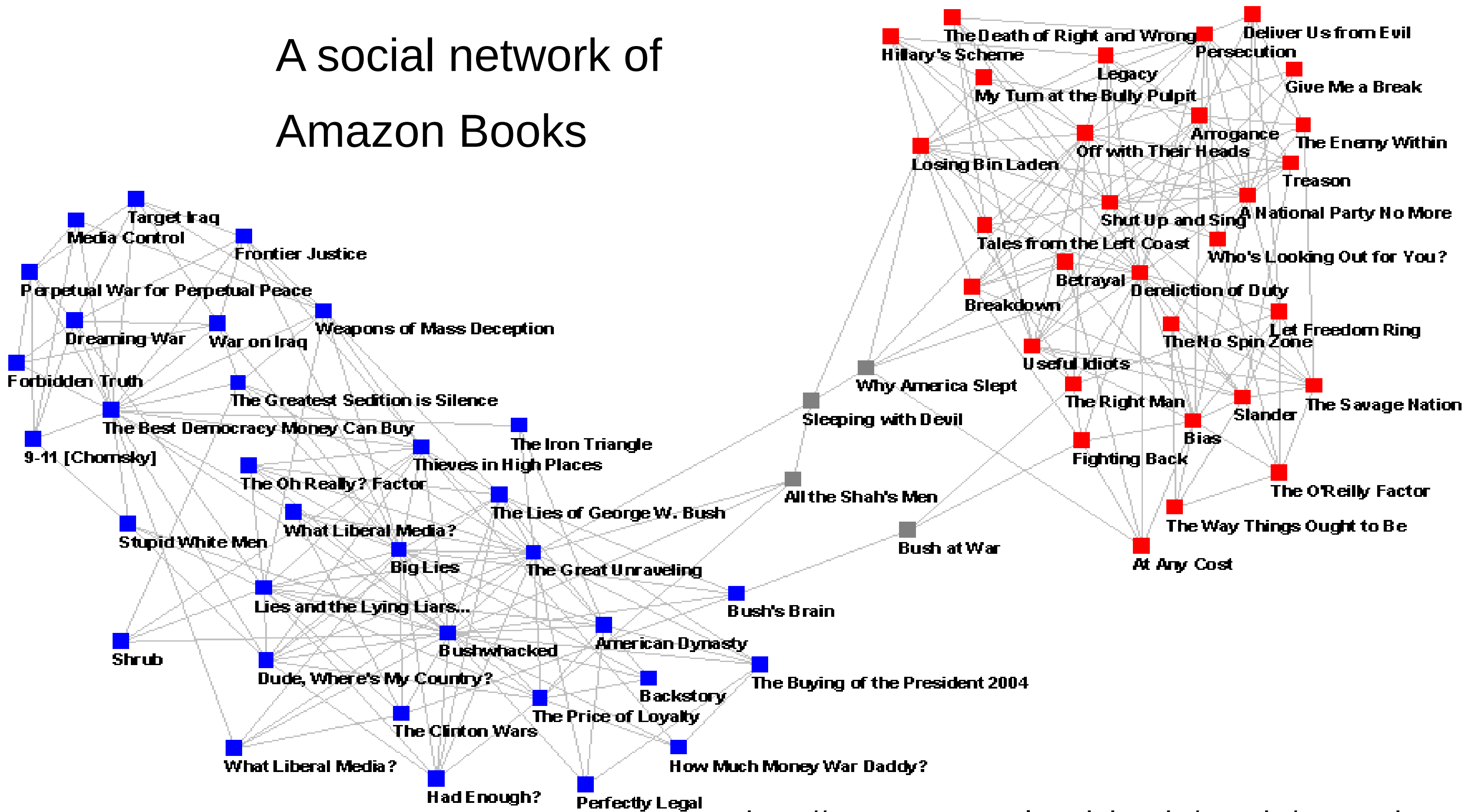
Clustering is dividing the data points into classes according to some similarity measure.

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

Modularity Q

- **Proposed by Newman and Girvan in 2004 as a measure of the quality of a particular division of the network.**
- **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*.**
- **Q is the *number of edges within communities* minus the *expected number of such edges***
- **Intuition: compare the division to a random network with same nodes and same degrees, but edges are placed randomly.**
- **Greedily maximizing Q outperformed all other methods, in most cases by an impressive margin, for community detection.**

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. (2008 stats)**

  http://www.techcrunch.com/2008/07/25/googles-misleading-blog-post-on-the-size-of-the-web/

- **Facebook has more than 500 million active users. (2010 stats)**

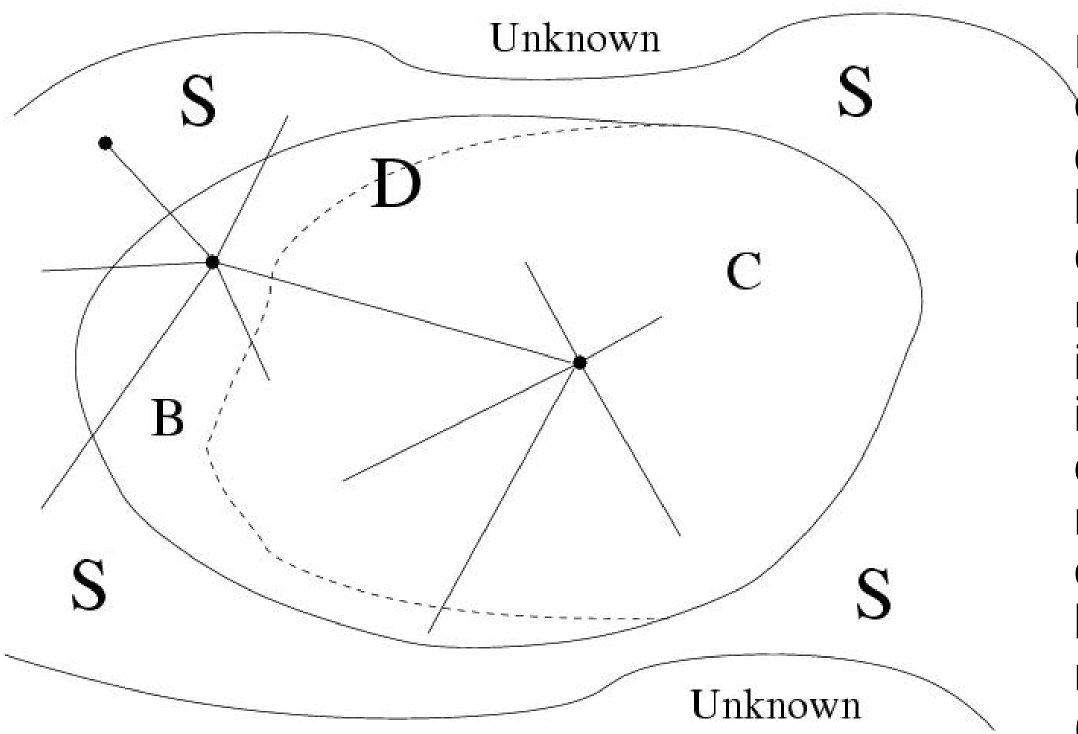
  http://www.facebook.com/press/info.php?statistics

- **Vodafone has 289 million customers worldwide. (2009 stats)**

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

Local Methods
Typical Problem Definition

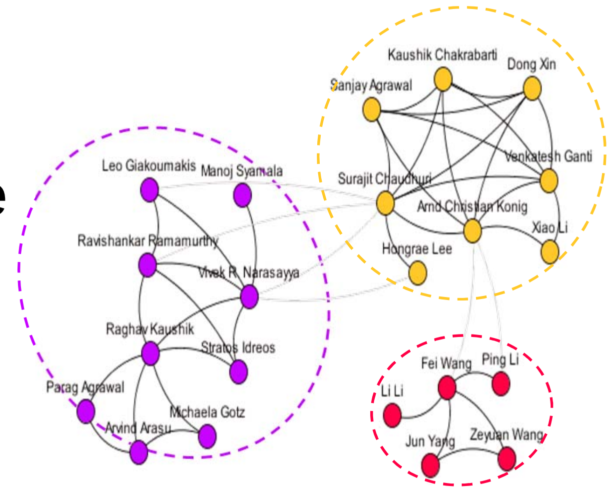
- **A local community D includes cores (C) nodes and boundary (B) nodes.**
- **If one new node is merged, its neighbours are added into shell nodes (S).**



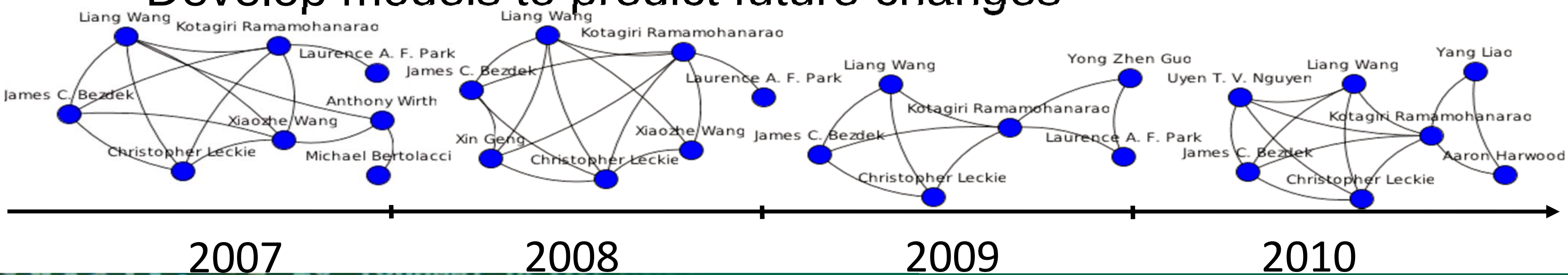
Maximize within edges of boundary nodes divided by total edges of boundary nodes
Or
maximize ***average*** internal degree (id) inside the whole community and minimize ***average*** external degree (ed) of boundary nodes, by maximizing id/ed (density)

Dynamic Networks

- **Many real-world social networks are dynamic**
  - Nodes and interactions change over time
  - Structure of communities evolves over time

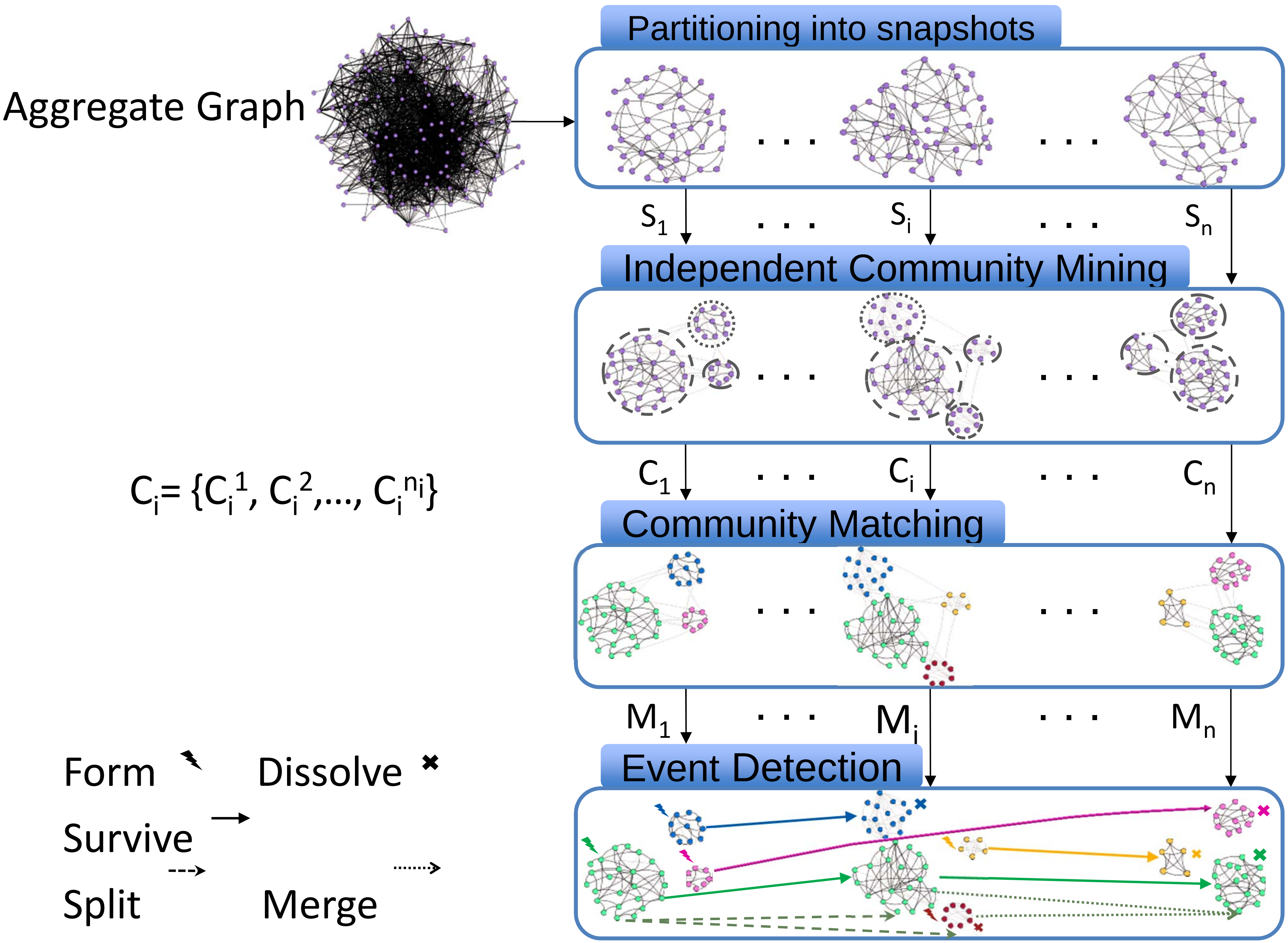


- **Dynamic Social Network Analysis**
  - Model network using time series graphs
  - Characterize evolution of communities and entities
  - Develop models to predict future changes



MODEC Framework

- <u>Mo</u>deling and <u>D</u>etecting the <u>E</u>volutions of <u>C</u>ommunities
- Communities are independently extracted in each snapshot
- A one-to-one matching algorithm is applied to match communities at different snapshots
- Significant events are identified to track the evolution of communities and individuals


Aggregate Graph
Partitioning into snapshots
S1 . . . Si . . . Sn
Independent Community Mining
C1 . . . Ci . . . Cn
Ci= {Ci1, Ci2,…, Cini}
Community Matching
M1 . . . Mi . . . Mn
Event Detection
Form
Dissolve
Survive
Split
Merge

Community Similarity

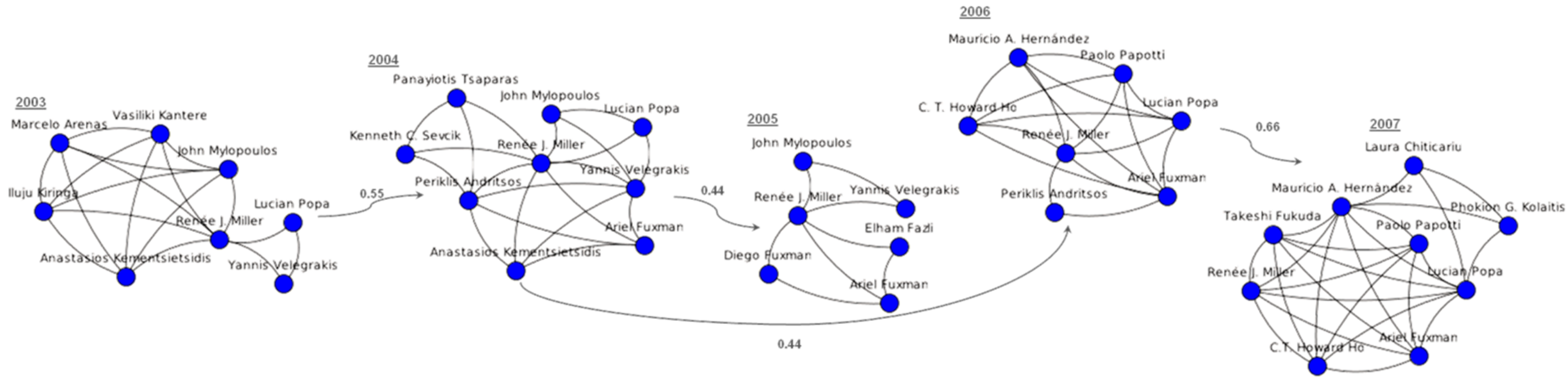
- Two communities at different snapshots are similar if the percentage of their mutual members exceed a given threshold *k∈[0, 1]*
- $sim(C^p, C^q) =$

$$\begin{cases} \frac{|V^p \cap V^q|}{\max(|V^p|,|V^q|)} & if \quad \frac{|V^p \cap V^q|}{\max(|V^p|,|V^q|)} \geq k \\ 0 & otherwise \end{cases}$$

- The similarity threshold *k* captures the tolerance of member fluctuation

Community vs. Meta Community

- **Community**
  - Densely connected individuals at a particular snapshot
  - Result of any static community mining algorithm
- **Meta community**
  - Series of similar communities from different snapshots
  - Represents the evolution of its constituent communities



Events Involving Communities

- **A community forms**
  - if there is no similar community at a previous snapshot
- **A community survives**
  - if there exists a similar community in a future snapshot
- **A community dissolves**
  - if there is no similar community at a later snapshot
- **A community splits**
  - if it fractures into multiple communities at a later snapshot
- **Two or more communities merge together**
  - if they integrate into one community in a future snapshot

Transitions Involving Communities

- **Size Transition**
  - A community **shrinks** if its number of nodes decreases
  - A community **expands** if its number of nodes increases
- **Compactness Transition**
  - A community **compacts** if its normalized number of edges increases
  - A community **diffuses** if its normalized number of edges decreases
- **Persistence Transition**
  - A community **persists** if its number of nodes and edges remains the same
- **Leader Transition**
  - A community experiences **leader shift** if its most central member shifts from one node to the other

Events Involving Individuals

- **A node appears**
  - if it was not present in a previous snapshot
- **A node disappears**
  - if it will not occur in a later snapshot
- **A node joins to a community**
  - if it did not belong to a similar community in a previous snapshot
- **A node leaves a community**
  - if it will not belong to a similar community in a later snapshot

Optimal Bipartite Matching

```
for all snapshots i
    remaining_communities ← communities at snapshot i
    clear selected_meta_communities
    j ← i-1
    while j >= 0 && size of remaining_communities > 0
        Construct weighted bipartite graph with remaining_communities
        and communities at snapshot j whose meta community is not in
        selected_meta_communities
        Match communities by the maximum weight bipartite matching
        for all communities c with detected match m
            Add c to meta community of m
            Remove c from remaining_communities
            Add meta community of m to selected_meta_communities
        end
        j ← j -1
    end
    for all communities c at remaining_communities
        Create meta community m
        Add c to m
    end
end
```

Weighted bipartite Matching based on community similarity

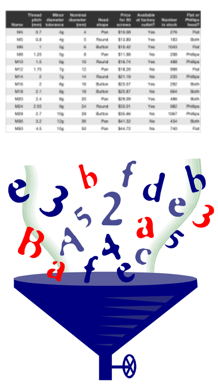
Results of matching are used to update meta communities

New meta communities are created for communities at snapshot 0 or communities left with no match

New Challenges

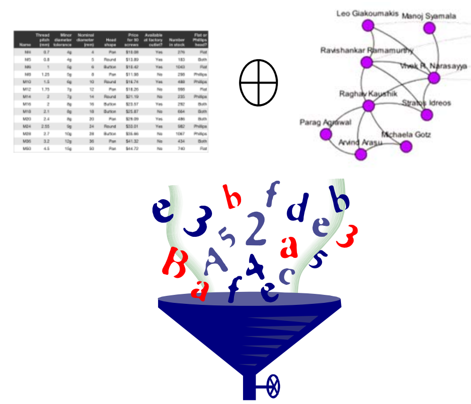
- **Machine Learning with relationships**

We know how to do this



i.i.d. data

- Classification
- Clustering
- Outlier detection
- Nearest Neighbour
- Etc.



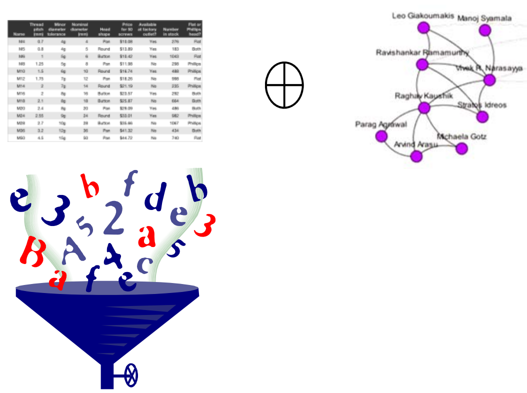
Non i.i.d. data

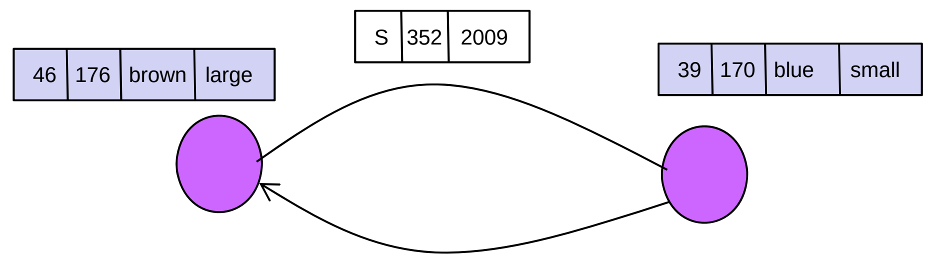
- Classification
- Clustering
- Outlier detection
- Nearest Neighbour
- Etc.

We DO NOT know how to do this

New Challenges

- **Machine Learning with relationships**





- ❑ Classification
- ❑ Clustering
- ❑ Outlier detection
- ❑ Nearest Neighbour
- ❑ Etc.

Not only nodes have attributes but relationships may have attributes.

Relationships may be directional.

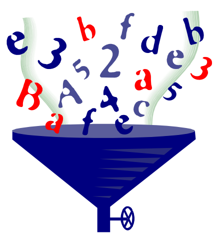
New Challenges

- **Probabilistic Databases and Probabilistic Information Networks**

We do not know how to do this

$\oplus$

Uncertainty



- Classification
- Clustering
- Outlier detection
- Nearest Neighbour
- Etc.

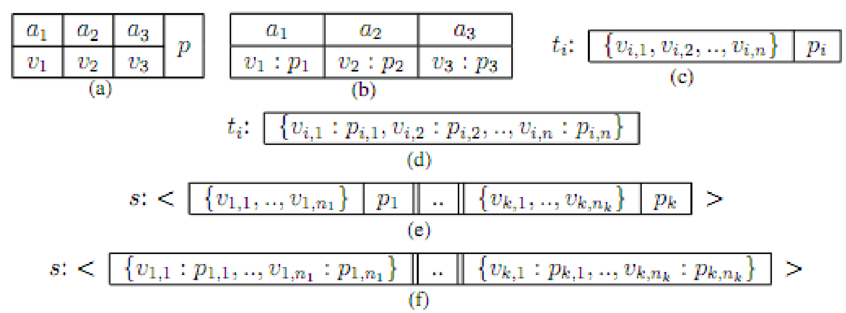
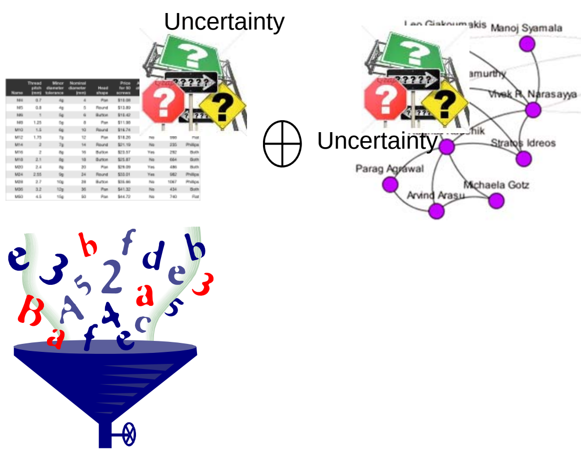


Figure 1.2: Some of the possible models for uncertainty in databases and sequential datasets: (a) A tuple with record-level uncertainty; (b) A tuple with attribute level uncertainty; (c) A transaction with transaction-level uncertainty; (d) A transactions with item-level uncertainty; (e) A sequence with transaction-level uncertainty; (f) A sequence with item-level uncertainty.

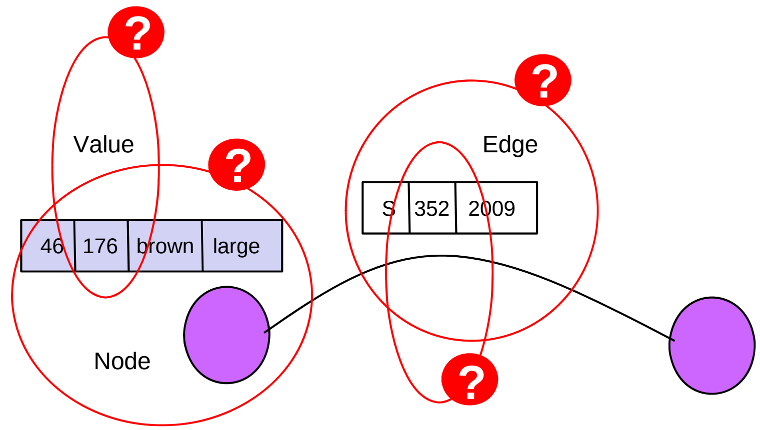
New Challenges

- **Probabilistic Databases and Probabilistic Information Networks**

We do not know how to do this



- ❑ Classification
- ❑ Clustering
- ❑ Outlier detection
- ❑ Nearest Neighbour
- ❑ Etc.



How to compute

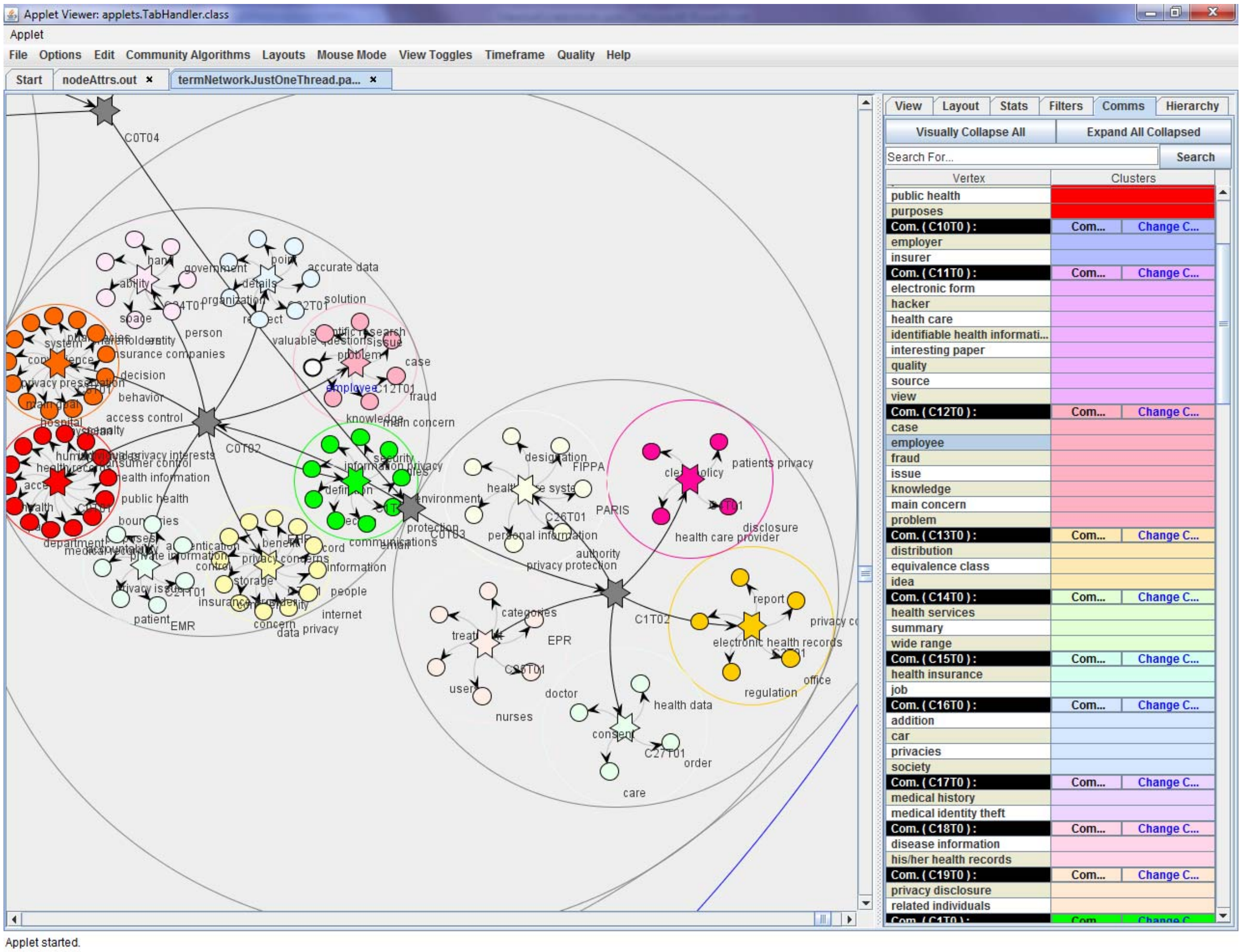
- Network diameter
- Shortest path
- Centrality
- Find communities
- ….

Conclusions

- Social Network Analysis is not a new science and is even more useful nowadays given the inter-related and complex data we are collecting.
- Applications in epidemiology, biomedicine, security, marketing, Psychology, Animal behavior, etc.
- 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.
- Opportunities for research in SNA with heterogeneous as well as homogeneous information networks.
- Opportunities for research in probabilistic information networks
- Opportunities for research in SNA for discovering patterns in dynamic networks

Meerkat: Topic (term community) Hierarchy

MeerkatED



Thank you to

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







IEEE
ICDM 2011 IEEE International Conference on Data Mining
Vancouver, Canada | December 11th to 14th, 2011

Thank you – Questions?

UNIVERSITY OF
ALBERTA

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

443 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/