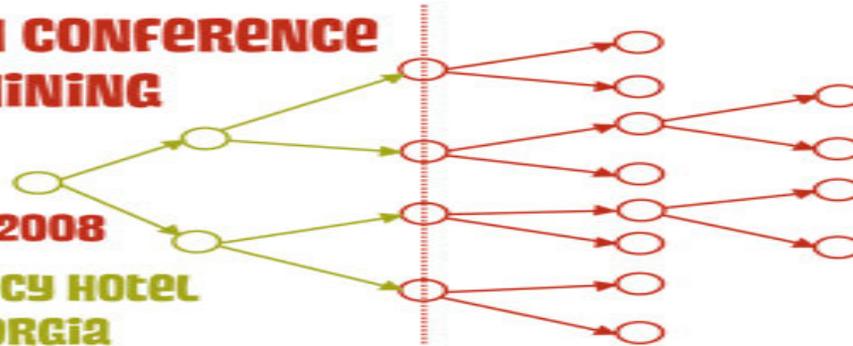


**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Data Mining Based Social Network Analysis from Online Behaviour

Jaideep Srivastava, Muhammad A. Ahmad, Nishith  
Pathak, David Kuo-Wei Hsu

University of Minnesota

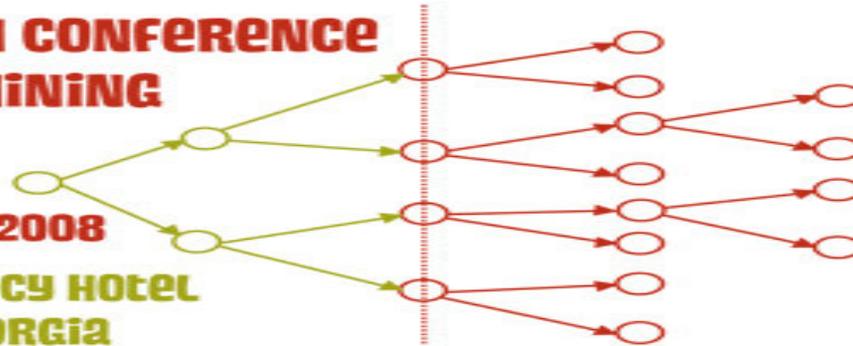
# Outline

- Introduction
- Framework for Social Network Analysis
- Classical Social Network Analysis
- Social Networks in the Online Age
- Data Mining for Social Network Analysis
- Application of Data Mining based Social Network Analysis Techniques
- Emerging Applications
- Conclusion
- References

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Introduction to Social Network Analysis

# Social Networks

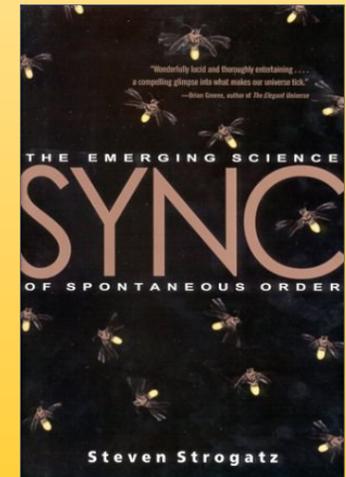
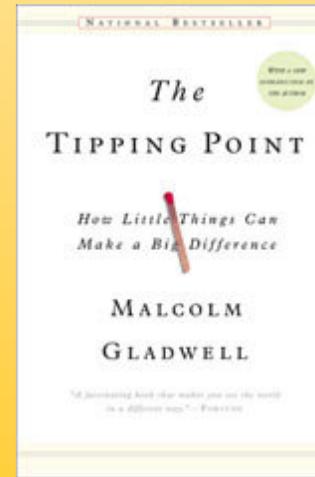
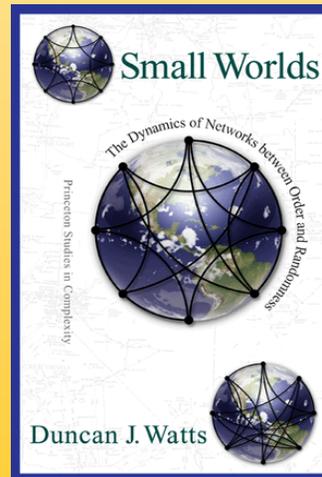
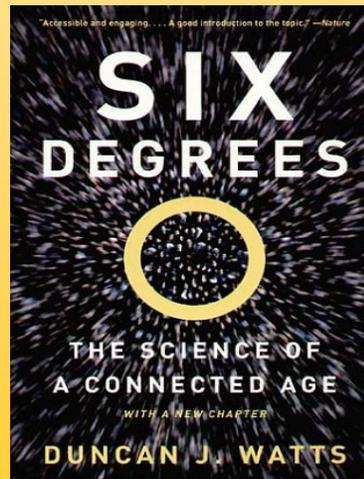
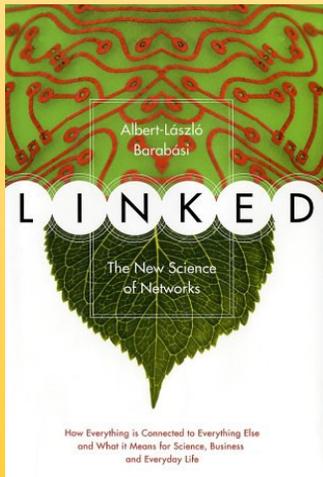
- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior



(Source: Freeman, 2000)

# SNA in Popular Science Press

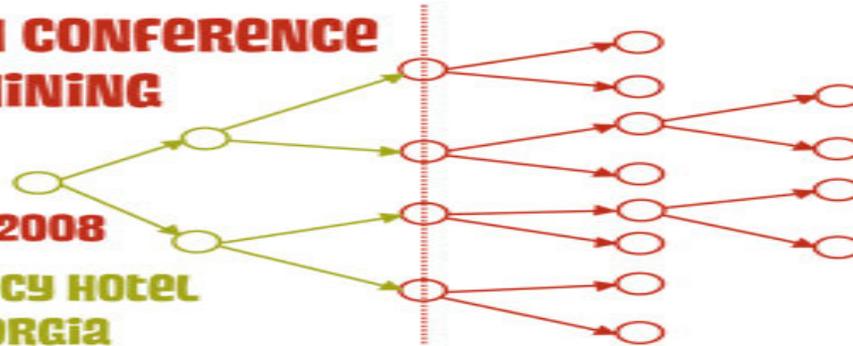
Social Networks have captured the public imagination in recent years as evident in the number of popular science treatment of the subject



**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Framework for Social Network Analysis

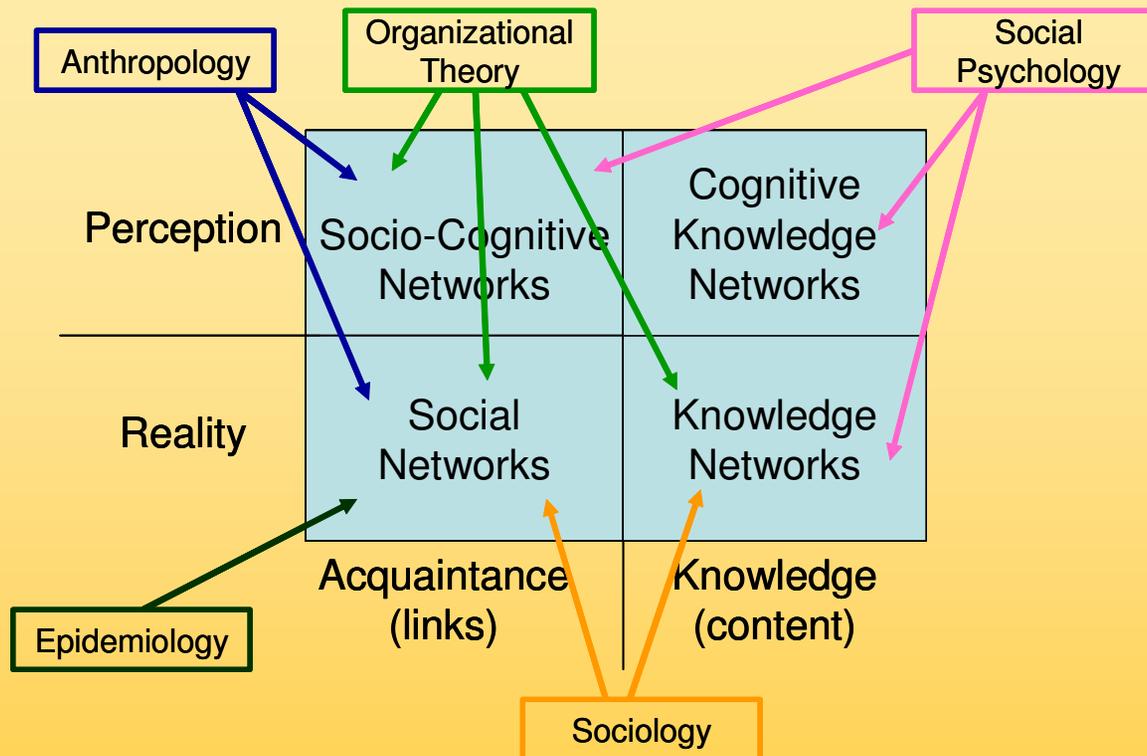
# Types of Social Network Analysis

- **Sociocentric (whole) network analysis**
  - Emerged in sociology
  - Involves quantification of interaction among a socially well-defined group of people
  - Focus on identifying global structural patterns
  - Most SNA research in organizations concentrates on sociometric approach
- **Egocentric (personal) network analysis**
  - Emerged in anthropology and psychology
  - Involves quantification of interactions between an individual (called *ego*) and all other persons (called *alters*) related (directly or indirectly) to ego
  - Make generalizations of features found in personal networks
  - Difficult to collect data, so till now studies have been rare

# Types of Social Network Analysis

- Knowledge Based Network Analysis
  - Emerged in Computer Science
  - Involves quantification of interaction between individuals, groups and other entities
  - Knowledge discovery based on entities associated with actors in the social network

# Networks Research in Social Sciences

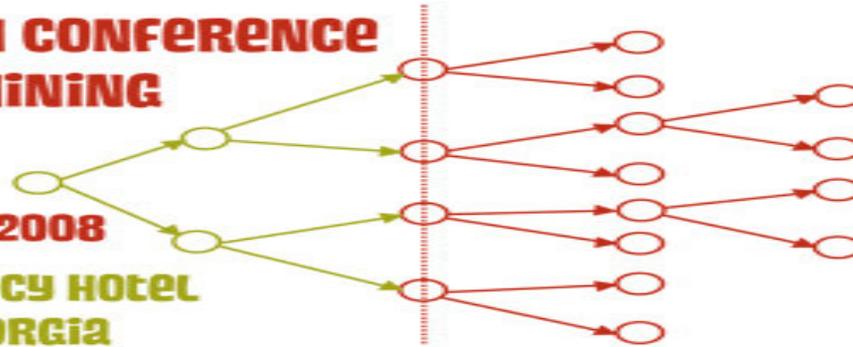


- Social science networks have widespread application in various fields
- Most of the analyses techniques have come from Sociology, Statistics and Mathematics
- See (Wasserman and Faust, 1994) for a comprehensive introduction to social network analysis
- Classification based on Contractor 2006

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

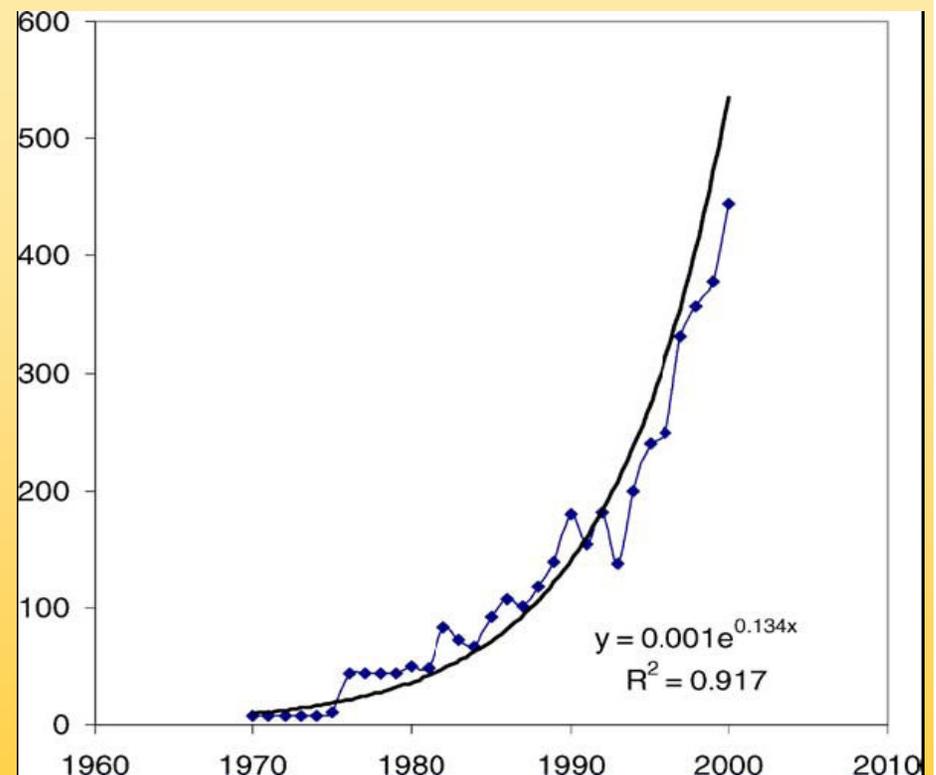
**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Classical Social Network Analysis

# Historical Trends

- Historically, social networks have been widely studied in the social sciences
- Massive increase in study of social networks since late 1990s, spurred by the availability of large amounts of data
- **Actors:** Nodes in a social network
- **Social Capital:** value of connections in a network
- **Embeddedness:** All behaviour is located in a larger context
- **Social Cognition:** Perception of the network
- **Group Processes:** Interrelatedness of physical proximity, belief similarity and affective ties



Exponential growth of publications indexed by Sociological Abstracts containing "social network" in the abstract or title.  
(Source: Borgatti and Foster, 2005)

# Terms & Key Concepts

- **Dyad:** A pair of actors (connected by a relationship) in the network
- **Triad:** A subset of three actors or nodes connected to each other by the social relationship
- **Degree Centrality:** Degree of a node normalized to the interval  $\{0 \dots 1\}$
- **Clustering Coefficient:** If a vertex  $v_i$  has  $k_i$  neighbors,  $k_i(k_i-1)/2$  edges can exist among the vertices within the neighborhood. The clustering coefficient is defined as

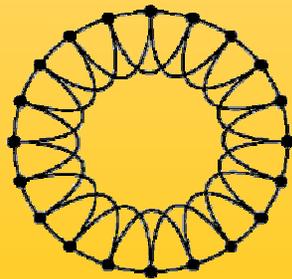
$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{ij} \in E$$

(M. E. J. Newman 2003, Watts, D. J. and Strogatz 1998)

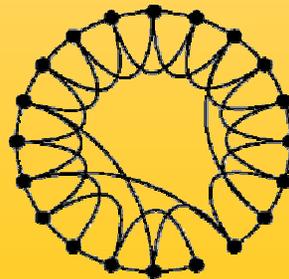
(Jon Kleinberg 1999, 2001) (D Watts, S Strogatz 1998), (D Watts 1999, 2003), (P. Marsden 2002)  
(Barabasi and Albert, 1999)

## Terms & Key Concepts

- **Six-degrees of separation:** Seminal experiment by Stanley Milgram
- **Scale Free Networks:** Networks that exhibit power law distribution for edge degrees
- **Preferential Attachment:** A model of network growth where a new node creates an edge to an extant node with a probability proportional to the current in-degree of the node being connected to
- **Small world phenomenon:** Most pairs of nodes in the network are reachable by a short chain of intermediates; usually the average pair-wise path length is bound by a polynomial in  $\log n$



(i) Regular Network



(ii) Small World Network  
University of Minnesota

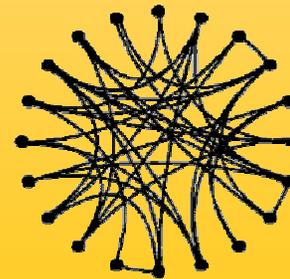


Figure Source: D Watts, S Strogatz (1998)

(iii) Random Network

# Measures of network centrality

- **Betweenness Centrality:** Measures how many times a node occurs in a shortest path; measure of 'social brokerage power'
  - Most popular measure of centrality
  - Efficient computation is important, best technique is  $O(mn)$
- **Closeness Centrality:** The total graph-theoretic distance of a given node from all other nodes
- **Degree centrality:** Degree of a node normalized to the interval  $\{0 .. 1\}$ 
  - is in principle identical for egocentric and sociocentric network data
- **Eigenvector centrality:** Score assigned to a node based on the principle that a high scoring neighbour contributes more weight to it
  - Google's PageRank is a special case of this
- **Other measures**
  - Information centrality
- **All of the above measures have directed counterparts**

# Statistical Models of Social Networks

- **P\* Models (Wasserman and Pattison, 1996)**

- Exponentially parametrized random graph models
- Given a set of  $n$  nodes, and  $X$  a random graph on these nodes and let  $x$  be a particular graph on these nodes  $P_{\theta}(X = x) \propto \exp\{\theta^t s(x)\}$
- Fitting the model refers to estimating the parameter  $\theta$  given the observed graph. MCMC-MLE techniques are used for estimation

- **Stochastic Actor Oriented Model (Tom Snijder, 2005)**

- Modelling Evolution of a social network over time
- Networks observed at specific points in time follow a continuous Markov Process
- Evolution is governed by actors rearranging their links in order to maximize a utility function

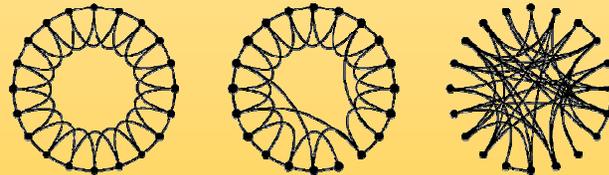
- **Latent Space Model (Hoff, Raftery and Handcock, 2002)**

- Probability of a relation between actors depends upon the position of individuals in an unobserved “social space”
- Inference for social space is developed within a maximum likelihood and Bayesian framework and is done via MCMC

# Models for Small World Phenomenon

- **Watts-Strogatz Network Model (1998)**

- Starts with a set  $V$  of  $n$  points spaced uniformly on a circle
- Join each vertex by an edge to each of its  $k$  nearest neighbors ("local contacts")
- Add small number of edges such that vertices are chosen randomly from  $V$  with probability  $p$  ("long-range contacts")
- Different values of  $p$  yield different types of networks



- **Kleinberg (2001) generalized the Watts-Strogatz Network Model**

- Start with two-dimensional grid and allow for edges to be directed
- A node  $u$  has a directed edge to every other node within lattice distance  $p$  - these are its local contacts
- Using independent random trial construct directed edges from  $u$  to  $q$  other nodes (long-range contacts)
- Expected diameter of the graph is  $\theta(\log n)$

# Models of Social Networks from Physics

- **Reka and Barabasi's model (Reka & Barabasi, 2000)**
  - Networks evolve because of local processes
  - Addition of new nodes, new links or rewiring of old links
  - Preferential attachment is used for link changes
  - The relative frequency of these factors determine whether the network topology has a power-law tail or is exponential
  - A phase transition in the topology was also determined
- **Characteristics of Collaboration Networks (Newman, 2001, 2003, 2004)**
  - Degree distribution follows a power-law
  - Average separation decreases in time
  - Clustering coefficient decays with time
  - Relative size of the largest cluster increases
  - Average degree increases
  - Node selection is governed by preferential attachment
- Newman (2003) provides an extensive survey of various networks, their properties and models

# SNA and Epidemiology

- **SIR Model (Morris, 2004; Kermack and McKendrick 1927)**

- Population is divided into three groups
  - **Susceptible (S):** People not infected, can be infected if exposed
  - **Infected (I):** People infected, can also infect others
  - **Recovered (R):** People recovered, have immunity
- The model consists of a system of three coupled nonlinear ordinary differential equations

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

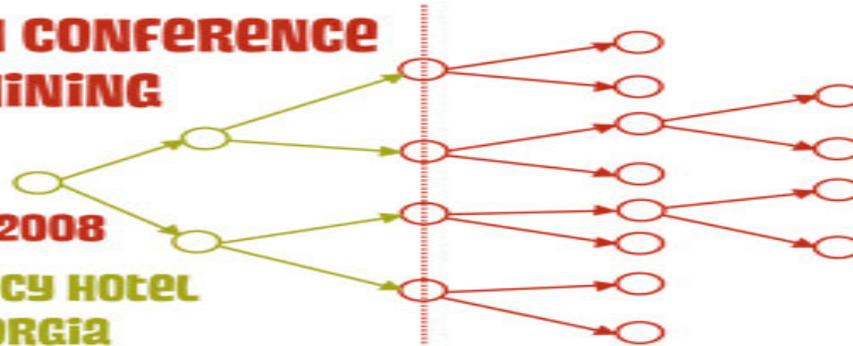
- where  $t$  is time,  $S(t)$  is the number of susceptible people,  $I(t)$  is the number of people infected,  $R(t)$  is the number of people who have recovered and developed immunity to the infection,  $\beta$  is the infection rate, and  $\gamma$  is the recovery rate.

- **SEIR Model:** Similar to the SIR model but there is a period of time during which the infected person is not infectious
- **SIS Model:** Used to model diseases where long lasting immunity is not present

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Social Networks in the Online Age

# Computer networks as social networks

- “Computer networks are inherently social networks, linking people, organizations, and knowledge” (Wellman, 2001)
- Data sources include newsgroups like USENET; instant messenger logs like AIM; e-mail messages; social networks like Orkut and Yahoo groups; weblogs like Blogger; and online gaming communities

USENET

YAHOO! GROUPS



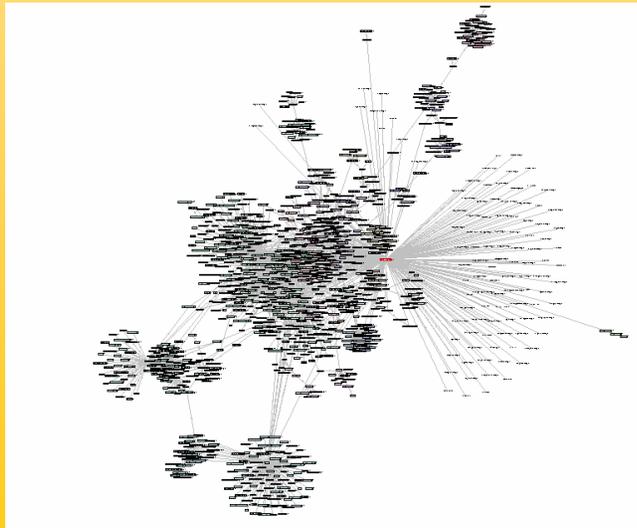
orkut



YAHOO! GAMES

# Example: Enron email dataset

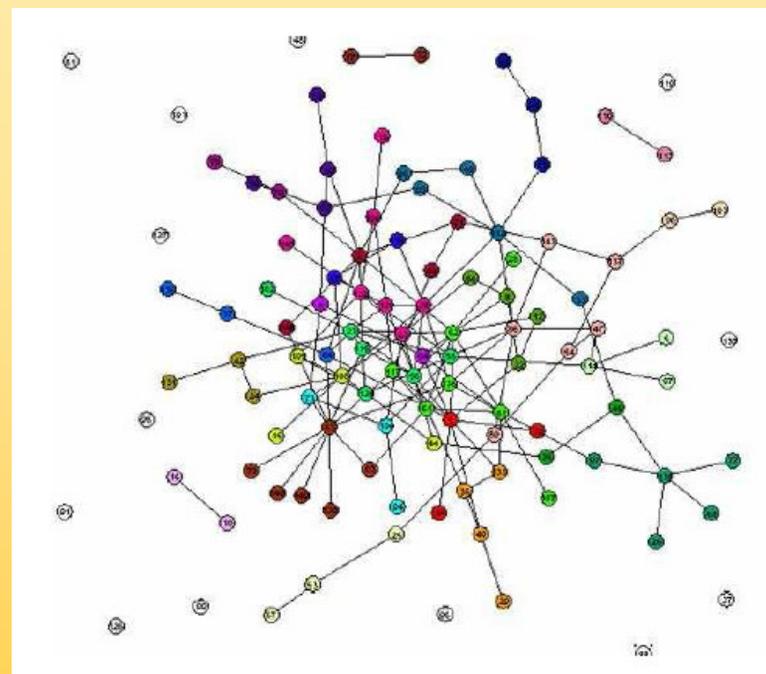
- Publicly available: <http://www.cs.cmu.edu/~enron/>
- Cleaned version of data
  - 151 users, mostly senior management of Enron
  - Approximately 200,399 email messages
  - Almost all users use folders to organize their emails
  - The upper bound for number of folders for a user was approximately the log of the number of messages for that user



A visualization of Enron email network  
(Source: Heer, 2005)

# Spectral and graph theoretic analysis

- **Chapanond et al (2005)**
  - Spectral and graph theoretic analysis of the Enron email dataset
  - Enron email network follows a power law distribution
  - A giant component with 62% of nodes
  - Spectral analysis reveals that the Enron data's adjacency matrix is approximately of rank 2
  - Since most of the structure is captured by first 2 singular values, the paper presents a visual picture of the Enron graph



(Source: Chapanond et al, 2005)

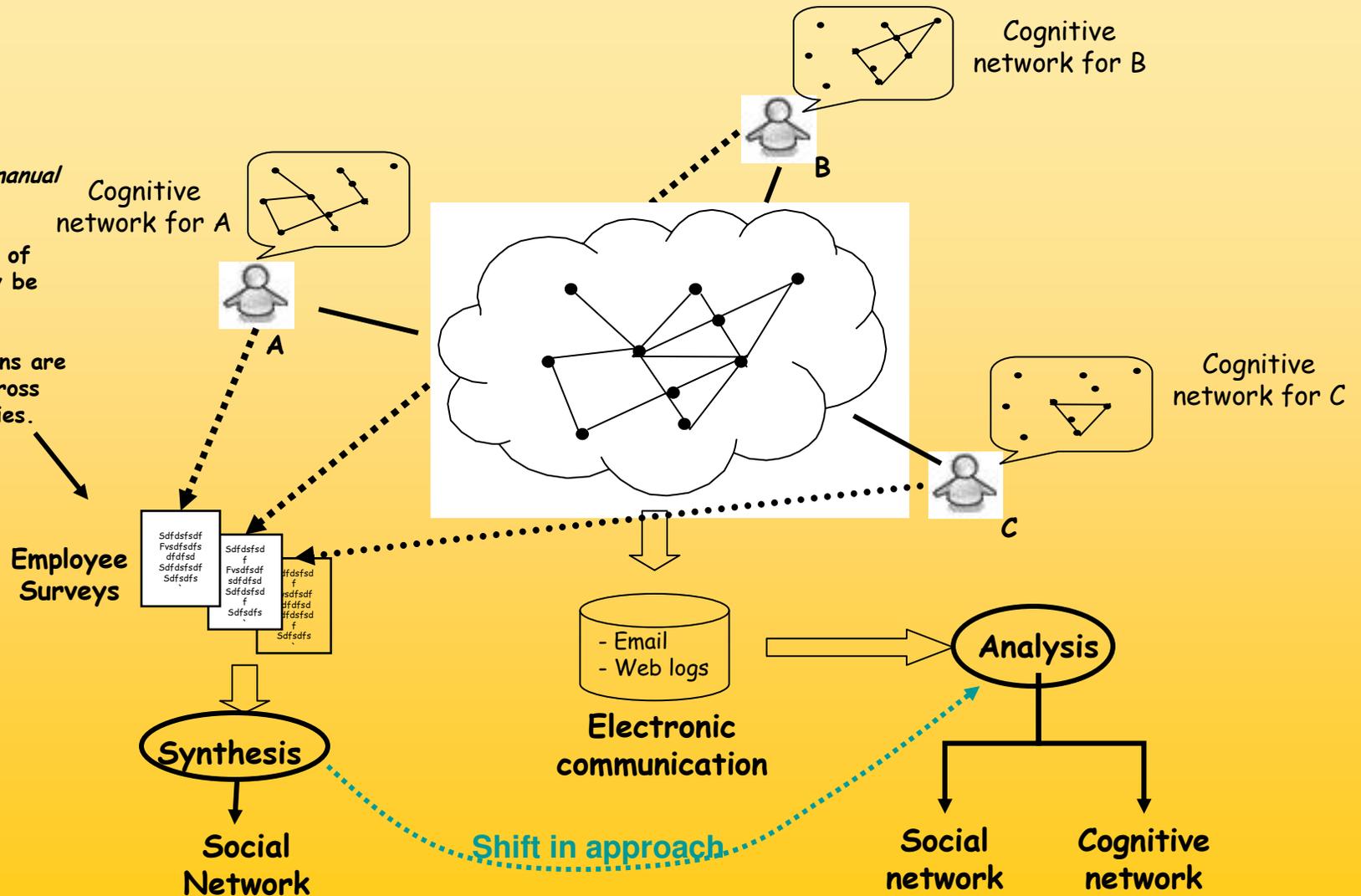
# Other analyses of Enron dataset

- Shetty and Adibi (2004)
  - Introduction to the dataset
  - Present basic statistics on Enron e-mail data such as email frequency, indegree, outdegree w.r.t time etc.
- Diesner and Carley (2005)
  - Compare the social network for the crisis period (Oct, 2001) to that of a normal time period (Oct, 2000)
  - The network in Oct, 2001 was more dense, connected and centralized compared to that of Oct, 2000
  - Half of the key actors in Oct, 2000 remained important in Oct, 2001
  - During crisis, the communication among employees did not necessarily follow the organization structure/hierarchy
  - During the crisis period the top executives formed a tight clique indicating mutual support

# Understanding the Network: a new approach

## Problems

- *High cost of manual surveys*
- *Survey bias*
  - Perceptions of individuals may be incorrect
- *Logistics*
  - Organizations are now spread across several countries.



# Key Drivers for CS Research in SNA

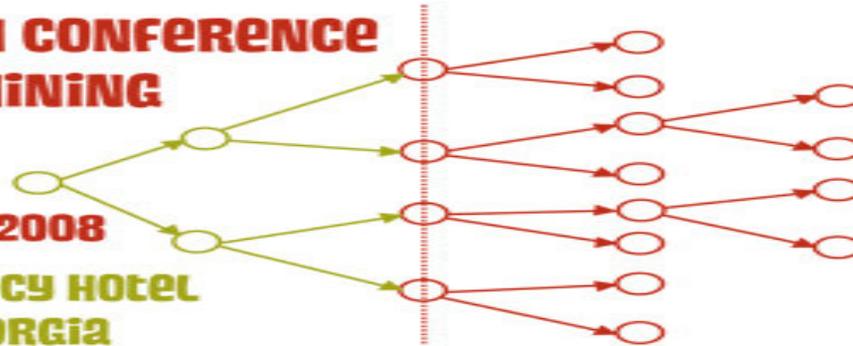
- Computer Science has created the über-cyber-infrastructure for
  - Social Interaction
  - Knowledge Exchange
  - Knowledge Discovery
- Ability to capture
  - different about various types of social interactions
  - at a very fine granularity
  - with practically no reporting bias
- Data mining techniques can be used for building descriptive and predictive models of social interactions

**→ Fertile research area for data mining research**

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Data Mining for Social Network Analysis

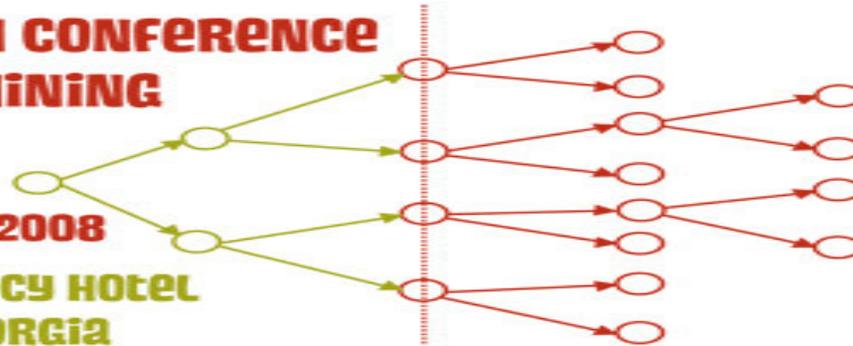
# DM for SNA (Overview)

- Community Extraction
- Link Prediction
- Cascading Behavior
- Identifying Prominent Actors and Experts in Social Networks
- Search in Social Networks
- Trust in Social Networks
- Characterization of Social Networks
- Anonymity in Social Networks
- Other Research

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

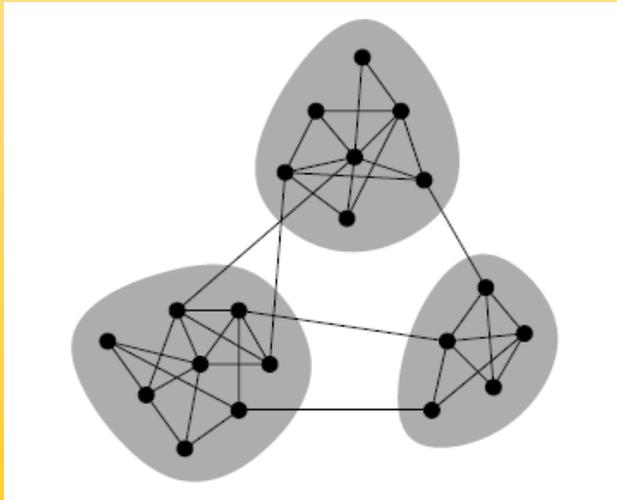
**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Community Extraction

# Community Extraction

- Discovering communities of users in a social network
- Possible to use popular link analysis techniques
  - HITS algorithm
  - Graph Clustering techniques



**Community structure in networks**  
(Source: Newman, 2006)

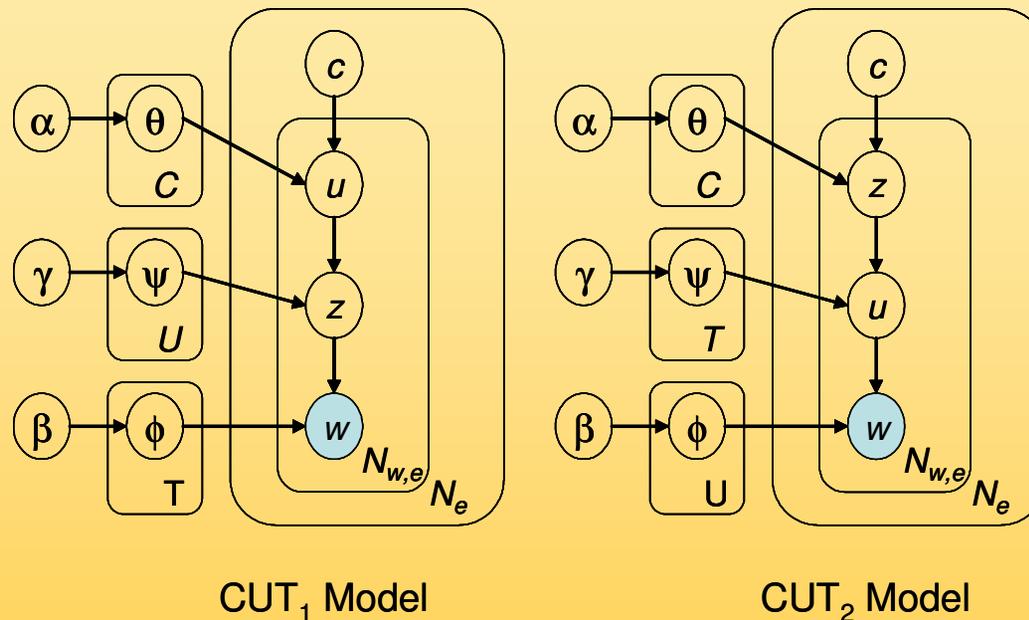
# Community Extraction

- Tyler et al (2003)
  - A graph theoretic algorithm for discovering communities
  - The graph is broken into connected components and each component is checked to see if it is a community
  - If a component is not a community then iteratively remove edges with highest betweenness till component splits
    - Betweenness is recomputed each time an edge is removed
  - The order of in which edges are removed affects the final community structure
  - Since ties are broken arbitrarily, this affects the final community structure
  - In order to ensure stability of results, the entire procedure is repeated  $i$  times and the results from each iteration are aggregated to produce the final set of communities
- Girvan and Newman (2002) use a similar algorithm to analyze community structure in social and biological networks

# Community Extraction

- Newman (2004)
  - Efficient algorithm for community extraction from large graphs
  - The algorithm is agglomerative hierarchical in nature
  - The two communities whose amalgamation produces the largest change in modularity are merged
  - Modularity for a given division of nodes into communities  $C_1$  to  $C_k$  is defined as
    - $Q = \sum_i (e_{ii} - a_i^2)$
    - Where  $e_{ii}$  is the fraction of edges that join a vertex in  $C_i$  to another vertex in  $C_i$  and  $a_i$  is the fraction of edges that are attached to a vertex in  $C_i$
    - Measure of difference between intra-community strength for given communities and a random network
- Clauset et al (2004) provide an efficient implementation for the above algorithm based on Max Heaps
  - The algorithm has  $O(md \log n)$  where  $m$ ,  $n$  and  $d$  are the number of edges, number of nodes and the depth of the dendrogram respectively

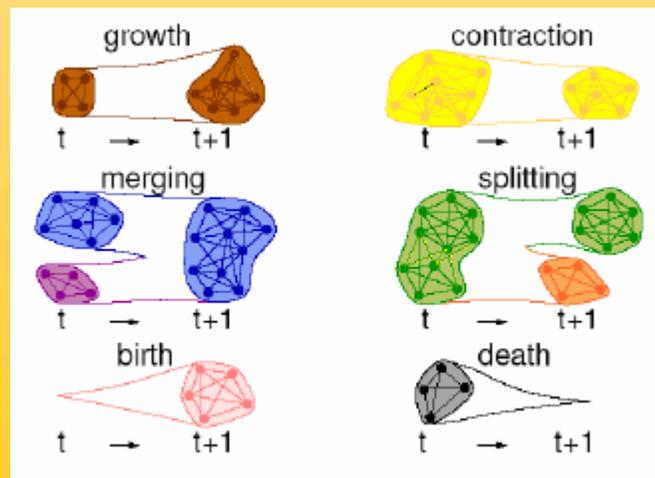
# Community Extraction



- Zhou et al (2006) present Bayesian models for discovering communities in email networks
  - Takes into account the topics of discussion along with the social links while discovering communities

# Community Extraction

- Clique Percolation Method (CPM), Palla et al. (2007)
  - Locating communities
    - Community = Union of adjacent  $k$ -cliques
    - Two  $k$ -cliques are adjacent if they share  $(k-1)$  nodes
    - $k$  is a parameter
  - Identifying evolving communities
    - Consider sizes and ages of communities
    - Match evolving communities for relatively distant points in time



Possible events in the community evolution

Source: Palla et al. (2007)

## CPM (Palla et al., 2007)

- Study of Evolution revealed four types of communities in both co-authorship and mobile phone networks:
  - (a) a small and stationary community
    - Stable as members are tight knit and available
  - (b) a small and non-stationary community
    - Unstable
  - (c) a large and stationary community
    - Unstable
  - (d) a large and non-stationary community.
    - Stable as the community is adaptable

# Community Detection in Large Networks

- Community detection algorithm based on label propagation (Albert et al., 2007)
  - One's label is determined based on the majority of labels of its neighbors
  - Algorithm gives near-linear time complexity

1. Initialize the labels at all nodes in the network. For a given node  $x$ ,  $C_x(0) = x$ .
2. Set  $t = 1$ .
3. Arrange the nodes in the network in a random order and set it to  $X$ .
4. For each  $x \in X$  chosen in that specific order, let  $C_x(t) = f(C_{x_{i1}}(t), \dots, C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), \dots, C_{x_{ik}}(t-1))$ .  $f$  here returns the label occurring with the highest frequency among neighbors and ties are broken uniformly randomly.
5. If every node has a label that the maximum number of their neighbors have, then stop the algorithm. Else, set  $t = t + 1$  and go to (3).

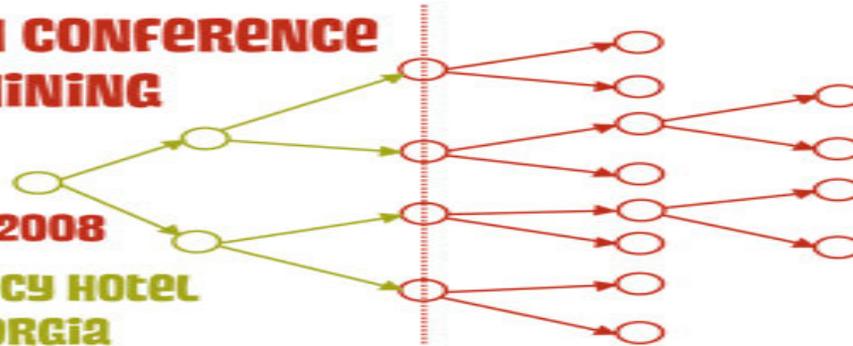


Asynchronous updating  
avoids oscillations of labels

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**

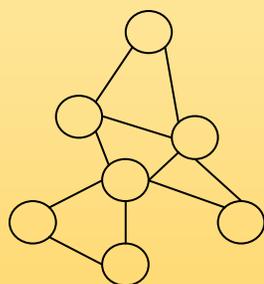


## Link Prediction

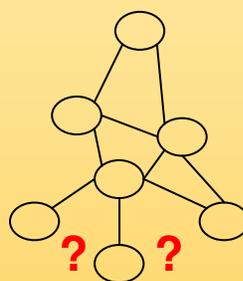
# Link Prediction

- **Different versions**

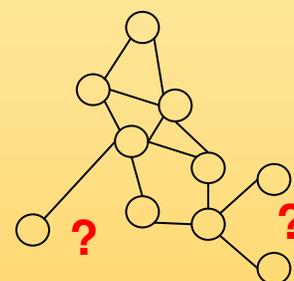
- Given a social network at time  $t_i$  predict the social link between actors at time  $t_{i+1}$
- Given a social network with an *incomplete* set of social links between a *complete* set of actors, predict the unobserved social links
- Given information about actors, predict the social link between them (this is quite similar to social network extraction)



Time  $t$



Time  $(t+1)$



Incomplete Network

- **Classical approach for link prediction is to fit the social network on a model and then use it for link prediction**
  - Latent Space model (Hoff et al, 2002), Dynamic Latent Space model (Sarkar and Moore, 2005),  $p^*$  model (Wasserman and Pattison, 1996)
- **Link Mining - encompassing a range of tasks including descriptive and predictive modelling (Getoor, 2003)**

# Link Prediction

- Predictive powers of the various proximity features for predicting links between authors in the future (Liben-Nowell and Kleinberg, 2003)
  - Link prediction as a means to gauge the usefulness of a model
  - Proximity Features: Common Neighbors, Katz, Jaccard, etc
  - No single predictor consistently outperforms the others
    - However all perform better than random
- Link Prediction using supervised learning (Hasan et al, 2006)
  - Citation Network (BIOBASE, DBLP)
  - Use machine learning algorithms to predict future co-authorship (decision tree, k-NN, multilayer perceptron, SVM, RBF network)
  - Identify a group of features that are most helpful in prediction
  - Best Predictor Features: Keyword Match count, Sum of neighbors, Sum of Papers, Shortest Distance

# Link Prediction

- Prediction of Link Attachments by Estimating Probabilities of Information Propagation (Saito et al 2007)
- Problem: Given a network at time  $t$ , the goal is to predict  $k$  potential links that are most likely to be converted to real links after a certain period of time.
- A ranking method: Top  $k$  links are predicted to be the real links.
- Pick two nodes  $v$  and  $w$  such that edge  $(v,w)$  does not exist and  $d(v,w) = 2$
- An edge is created between  $v$  and the adjacent nodes of  $w$  if information propagation between the two is successful.
- The probability of information propagation between  $v$  and  $w$  is given by –

$$q_{\{v,w\}} = 1 - \prod_{u \in A(v) \cap A(w)} (1 - p_{\{u,v\}})(1 - p_{\{u,w\}}).$$

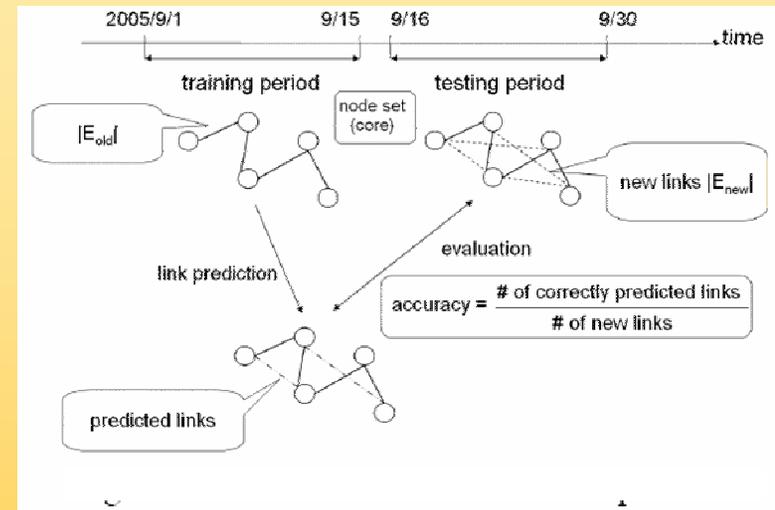
Where  $p_{\{x,y\}}$  is the ratio of number of common neighbors to the total number of neighbors belonging to  $x$  and  $y$ .

- In the dataset only a small fraction (0.0002) of the potential links are converted to real links. The proposed method outperformed all the other comparison methods.

# Link Prediction

- Link Prediction of Social Networks Based on Weighted Proximity Measures (Murata et al 2007)
  - Link Prediction in (Question Answering Bulletin Boards) QABB
  - Weight on edges depends upon past encounters between the nodes

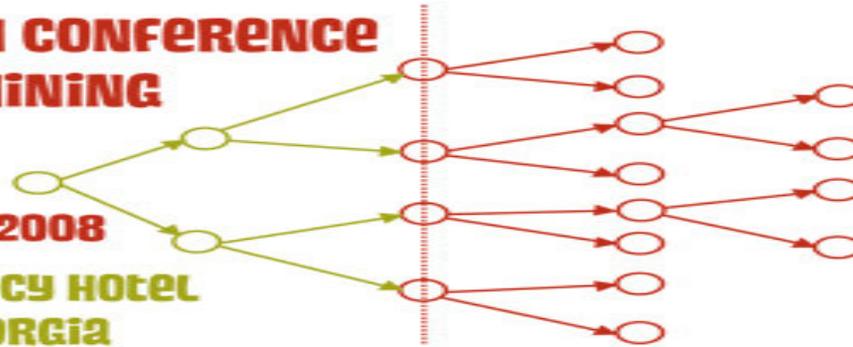
$$score(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(y, z)}{2}$$



**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

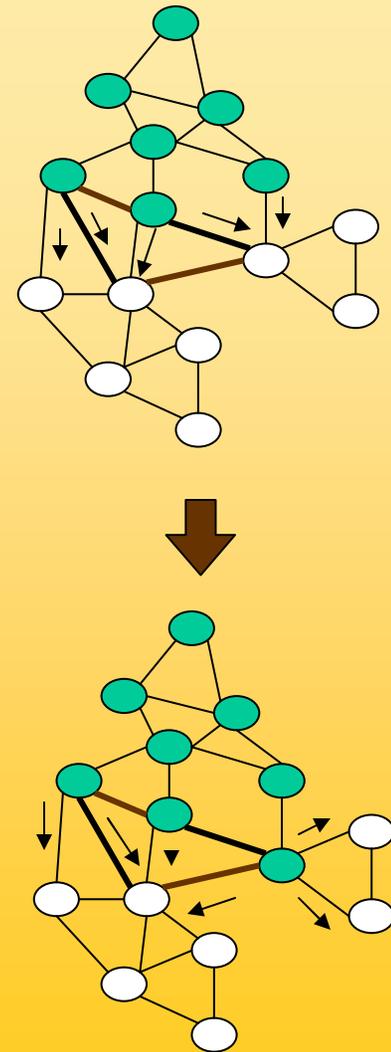
**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Cascading Behavior

# Cascading Models

- **Model of Diffusion of Innovation (Young, 2000)**
  - Interactions between the agents are weighted
  - Directed edges represent influence of one agent on the other
  - Agents have to choose between outcomes
    - The choice is based on a utility function which has an individual and a social component
  - The social component depends upon the choices made by the neighbours
  - Under the assumption of a logistic response, diffusion time is independent of number of actors and initial state and a final stable state will be reached
- **Related work:** Schelling (1978), Granovetter (1978), Domingos (2005), Watts (2004), Kempe et al (2003), Leskovec et al (2007)



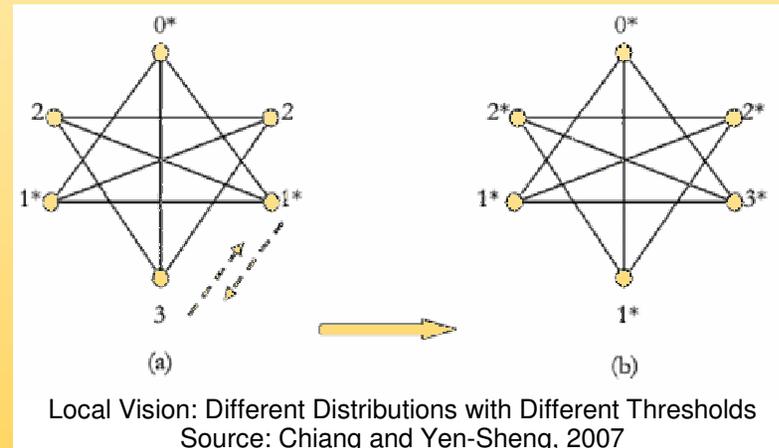
# Cascading Behavior

**TABLE 1** A Summary of Network Threshold Models

Criteria Works	Network setting	Supplemental mechanisms	Main findings: High levels of bandwagon effects are associated with:
Macy (1991) Hedstrom (1994) Abrahamson & Rosenkopf (1997)	Random networks Random networks "Core-Periphery" network	Stochastic learning Individual traits Profitability of innovation	Structure of weak ties Smaller geographical space Boundary agents: Low threshold and low connectivity (termed boundary weakness) and High threshold and high connectivity (termed boundary pressure point)
Krackhardt (1997)	"Chain" and "Star" network	Competition between positive and negative bandwagons	Moderate immigration between groups of adopters and non-adopters
Chwe (1999) Watts (2002)	Random networks Random networks	Common knowledge	Structure of strong ties Low average threshold propensity and low connectivity

# Cascading Behavior

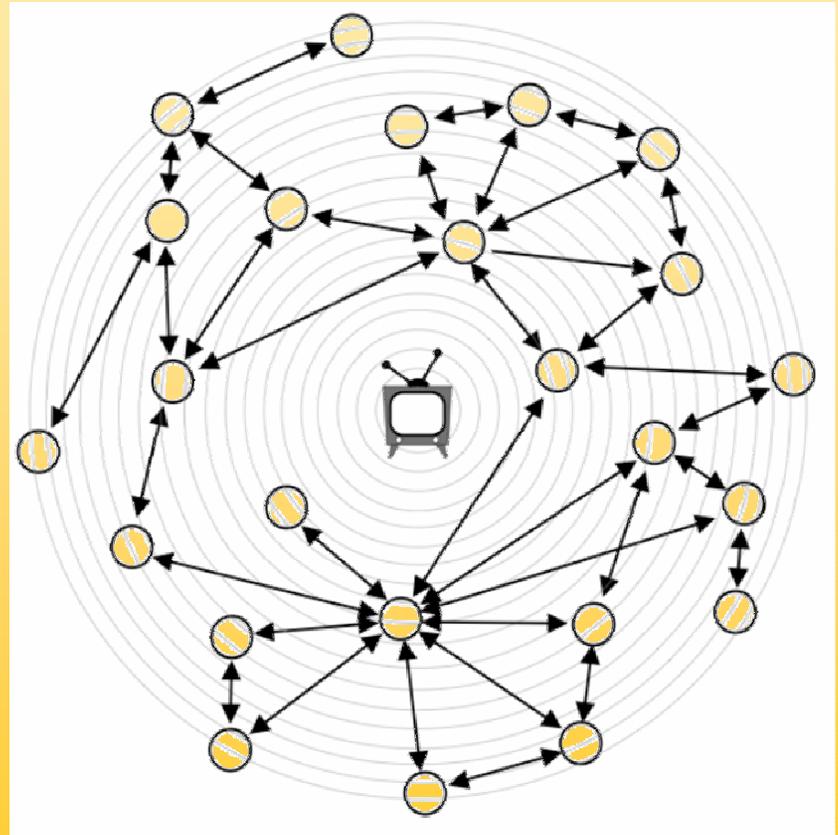
- Bandwagon Dynamics and the Threshold Heterogeneity of Network Neighbors (Chiang, Yen-Sheng 2007)
- Bandwagon Dynamics: Adopt a trait in one's neighborhood because other people have done so.
- Granovetter's threshold model
  - Different Threshold for adoption
  - Disadvantage:  
Assumes Global Vision for thresholds



- Assume 'local vision' – Each node looks at its neighbors
- Participation levels increase as network departs from pure homophily
- Participation decreases as heterogeneity among neighbors increases
- Simulation Results: The optimal distribution of thresholds in a network is one where a balance of homophily and heterogeneity between actors' thresholds exists

# Cascading Behavior

- Influentials, Networks, and Public Opinion Formation (Watts et al 2007)
- It is usually assumed that a small group of influential actors greatly influence the diffusion of information in the network.
- This assumption is challenged.
- A number of models are simulated
- Observation: In most cases information cascades are driven not by a small group of influentials but by a critical mass of easily influenced individuals.
- Conclusion: The spread of influence is more complex than previously thought.



Network Influence Model

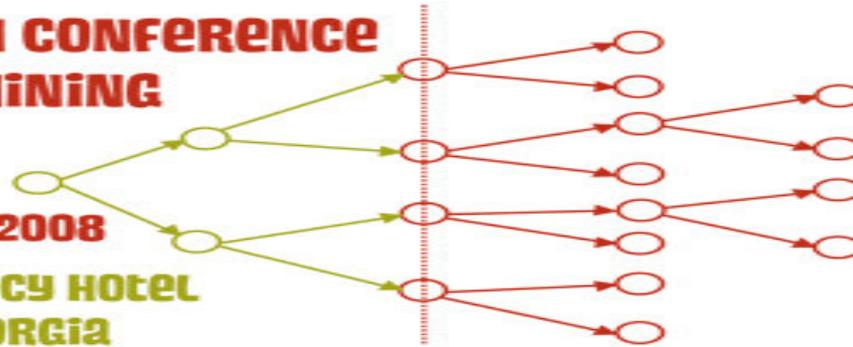
# Cascading Behavior

- Maximizing Influence in a Competitive Social Network: A Follower's Perspective (Tim Carnes et al 2007)
- Previous models of influence maximization in a social network assumed that there is only one entity.
- How does one introduce a product in the market when a similar product is already being introduced in the market by a rival.
- Assumptions:
  - A limited marketing budget.
  - Knowledge of early adopters of rival's product.
  - A node will not change to another technology
- Determining the set which contains the users that are most likely to adopt the product is an NP hard problem (in the model setting.)
- An approximate solution is given is given to determine a subset of such most influential users.
- This subset can be varied based upon the cost and also on the size.
- Two models for diffusion of two competing technologies are described.
- Model 1: Technology only diffuses from the set of initial adapters.
- Model 2: Model a node can interested in a technology and get it from a neighbor.

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



# Identifying Prominent Actors and Experts in Social Networks

# Link mining

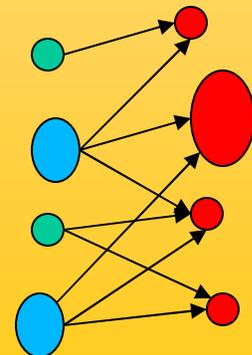
- Availability of rich data on link structure between objects
- Link Mining - new emerging field encompassing a range of tasks including descriptive and predictive modeling (Getoor, 2003)
- Extending classical data mining tasks
  - Link-based classification – predict an object's category based not only on its attributes but also the links it participates in
  - Link-based clustering – techniques grouping objects (or linked objects)
- Special cases of link-based classification/clustering
  - Identifying link type
  - Predicting link strength
  - Link cardinality
  - Record linkage
- Getoor et al (2002)
  - Two mechanisms to represent probabilistic distributions over link structures
  - Apply resulting model to predict link structure

# Link Mining (Getoor & Diehl, 2005)

- **Link Mining:** Data Mining techniques that take into account the links between objects and entities while building predictive or descriptive models
- Link based object ranking, Group Detection, Entity Resolution, Link Prediction
- **Applications:** Hyperlink Mining, Relational Learning, Inductive Logic Programming, Graph Mining

## Hubs and Authorities (Kleinberg, 1997)

- Being Authority depends upon in-edges; an authority has a large number of edges pointing towards it
- Being a Hub depends upon out-edges; a hub links to a large number of nodes
- Nodes can be both hubs and authorities at the same time



# Identifying Prominent Actors in a Social Network

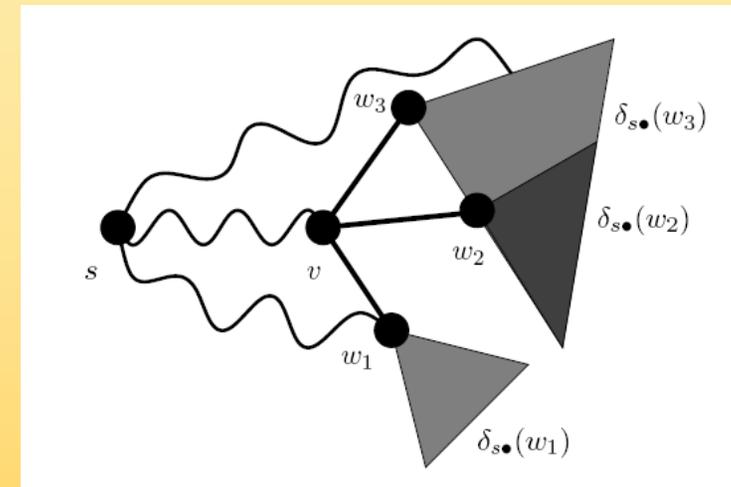
- A common approach is to compute scores/rankings over the set (or a subset) of actors in the social network which indicate degree of importance/expertise/influence
  - E.g. Pagerank, HITS, centrality measures
- Various algorithms from the link analysis domain
  - PageRank and its many variants
  - HITS algorithm for determining authoritative sources
- Kleinberg (1999)
  - Discusses different prominence measures in the social science, citation analysis and computer science domains
- Centrality measures exist in the social science domain for measuring importance of actors in a social network
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality

# Identifying Prominent Actors in a Social Network

- Brandes, (2001)
  - Prominence → high betweenness value
  - An efficient algorithm for computing for betweenness centrality
  - Betweenness centrality requires computation of number of shortest paths passing through each node
  - Compute shortest paths between all pairs of vertices
  - Trivial solution of counting all shortest paths for all nodes takes  $O(n^3)$  time
  - A recursive formula is derived for the total number of shortest paths originating from source  $s$  and passing through a node  $v$ 

$$\delta_s(v) = \sum_{\{w_i\}} [1 + \delta_s(w_i)] (\sigma_{sv} / \sigma_{sw})$$

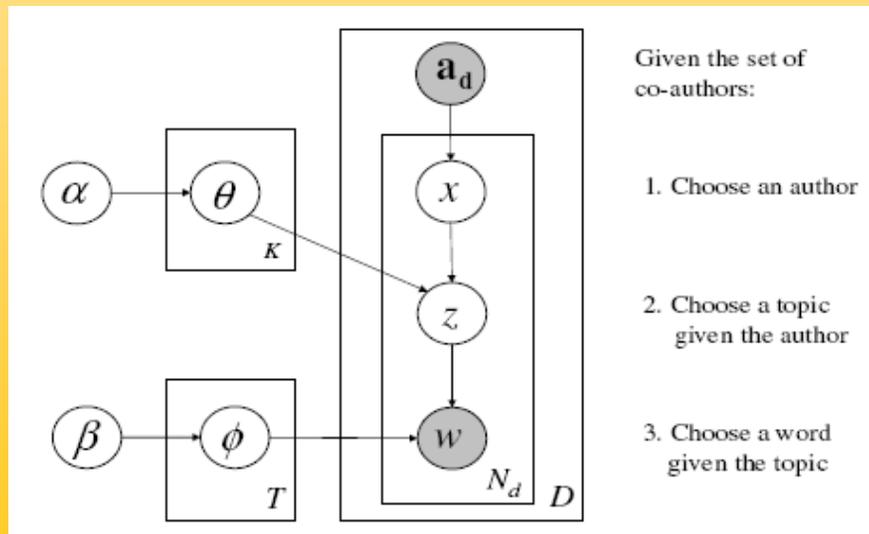
$\sigma_{ij}$  is the number of shortest paths between  $i$  and  $j$   
 $w_i$  is a node which has node  $v$  preceding itself on some shortest path from  $s$  to itself
  - The time complexity reduces to  $O(mn)$  for unweighted graphs and  $O(mn + \log^2 n)$  for weighted graphs
  - The space complexity decreases from  $O(n^2)$  to  $O(n+m)$



Nodes  $s$ ,  $v$  and  $\{w_i\}$   
 Source: (Brandes, 2001)

# Identifying Experts in a Social Network

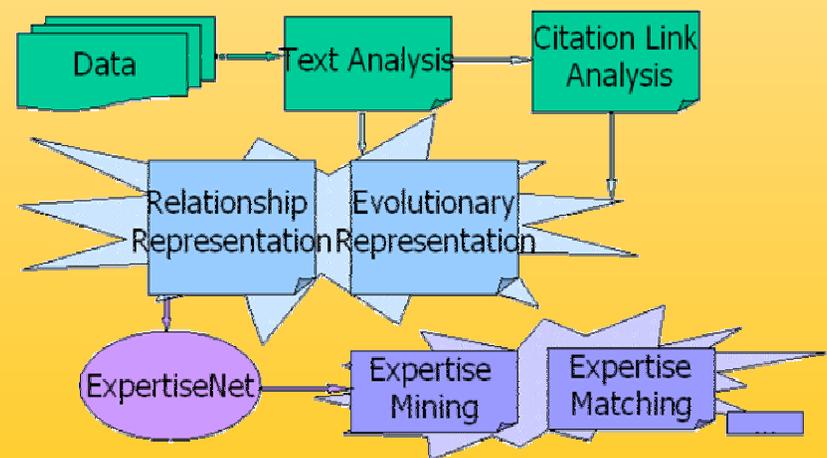
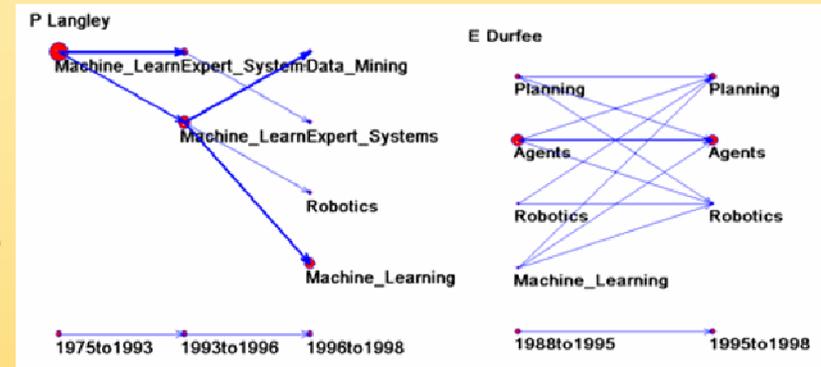
- Apart from link analysis there are other approaches for expert identification
  - Steyvers et al (2004) propose a Bayesian model to assign topic distributions to users which can be used for ranking them w.r.t. to the topics
  - Harada et al (2004) use a search engine to retrieve top  $k$  pages for a particular topic query and then extract the users present in them
- **Assumption:** *existence implies knowledge*



(Source: Steyvers et al, 2004)

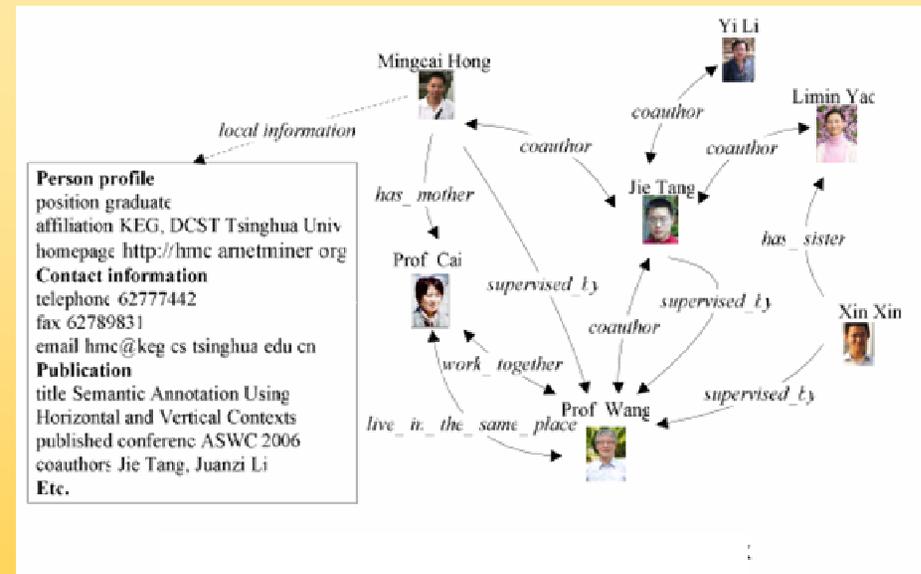
# Expertise in Social Networks

- Expertise Net (Song et al. 2005)
- Text classification to determine the expertise of individuals in the network.
- Use citations analysis in text classification.
- Papers in the corpus are classified into pre-defined categories.
- Relational ExpertiseNet: For each author, construct an expertise graph where each node represents the expertise of a person for a topic.
- Evolutionary ExpertiseNet: Use sliding windows to get the expertise and use the  $p^*$  model to determine the structure of the expertise graph over time.
- Changes in the network are modeled as the stochastic result of network effects like density, reciprocity The network



# Expertise in Social Networks

- Expertise oriented search using social networks (Jing Zhang 2007, Juan-Zi Lee 2007)
- A social network is constructed the co-authorship between authors
- Expert Identification
  - First compute relevancy based on documents associated with the author for a given topic.
  - Secondly propagate the topic relevancy of the researcher to his/her neighbors.
  - Thus the expertise depends upon authored documents and the expertise of one's neighbors.
  - Alternatively, compute expertise and then get the experts relevant to the query and then construct the social network and then propagate expertise.
- (Yupeng Fu 2007) is based on a similar idea



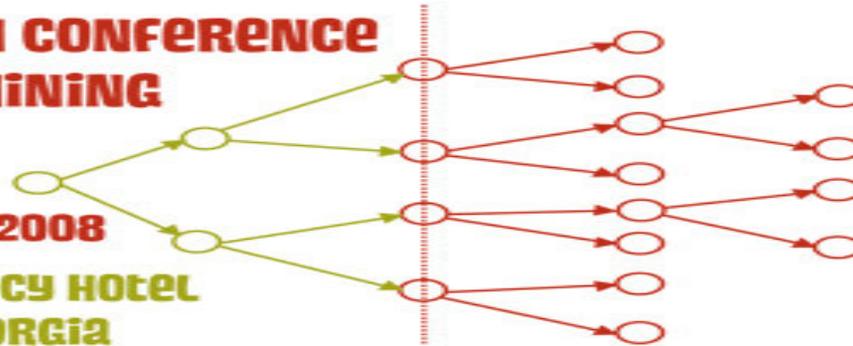
# Expertise in Social Networks

- Social Networks for Expertise Identification (Yupeng Fu 2007)
  - Identify experts based on documents for only a subset of the data.
  - Use this initial seed of people to propagate expertise to other agents in the rest of the network
  - The probability of a non-seed node being an expert depends upon the expertise of the associated nodes. The social network can be built based on co-occurrence on web pages or co-occurrence in e-mails for the whole corpus.
  - Alternatively the social network can also be built based on the web pages and e-mails which are relevant to the query.
- Related Work: (Jing Zhang and Juan-Zi Lee 2007)

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Search in Social Networks

# Search in Social Networks

- Searching/Querying for information in a social network
- Query routing in a network
  - A user can send out queries to neighbors
  - If the neighbor knows the answer then he/she replies else forward it to their neighbors. Thus a query propagates through a network
  - Develop schemes for efficient routing through a network
- Adamic et al (2001)
  - Present a greedy traversal algorithm for search in power law graphs
  - At each step the query is passed to the neighbor with the most number of neighbors
  - A large portion of the graph is examined in a small number of hops
- Kleinberg and Raghavan (2005) present a game theoretic model for routing queries in a network along with incentives for people who provide answers to the queries

# Search in Social Networks

- Watts-Dodds-Newman's Model (Watts et al, 2002)

- Individuals in a social network are marked by distinguishing characteristics

- Groups of individuals can be grouped under groups of groups

- Group membership is the primary basis for social interaction

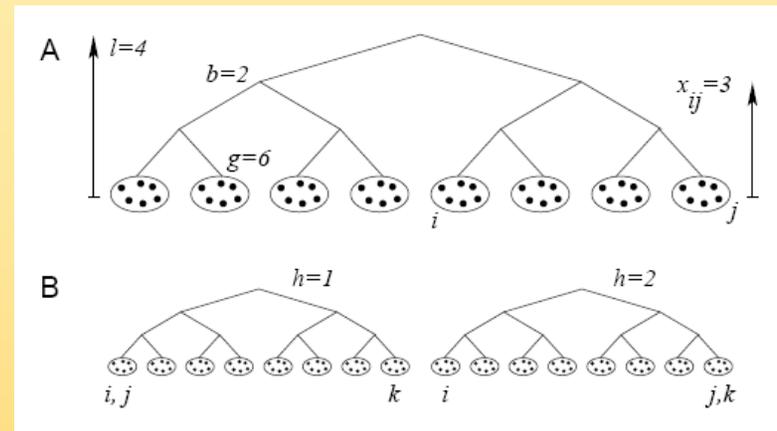
- Individuals hierarchically cluster the social world in multiple ways based on different attributes

- Perceived similarity between individuals determine 'social distance' between them

- Recreate Stanley Milgram's experiment: Message routing in a network is based only on local information

- Results

- Searchability is a generic property of real-world social networks



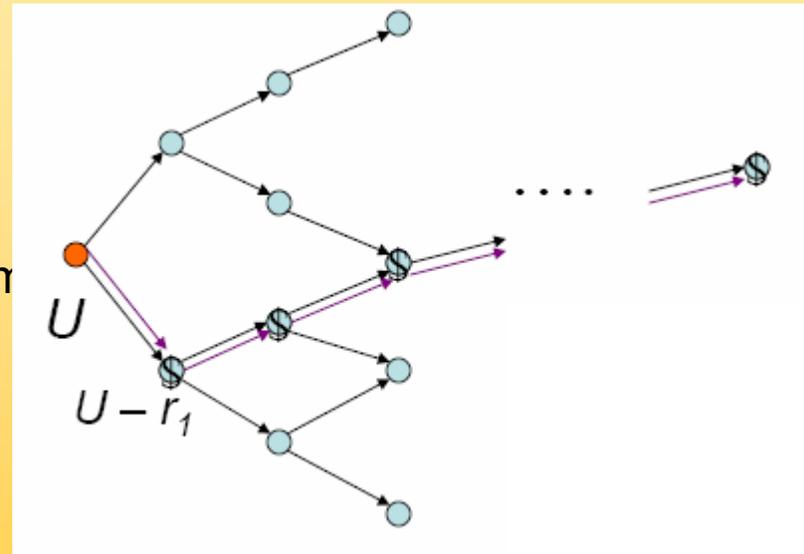
Source: Watts et al, 2002

# Search in Social Networks

- Yu and Singh (2003)
  - Each actor has a vector over all terms and every actor stores the vectors and immediate neighborhoods of his/her neighbors
  - Individual vector entries indicate actor's familiarity/knowledge about the various terms
  - Each neighbor is assigned a relevance score
  - The score is a weighted linear combination of the similarity between query and term vectors (cosine similarity based measure) and the sociability of that neighbor
  - Sociability is a measure of that neighbor knowing other people who might know the answer
  - The expert and sociability ratings maintained by a user are updated based on answers provided by various users in the network

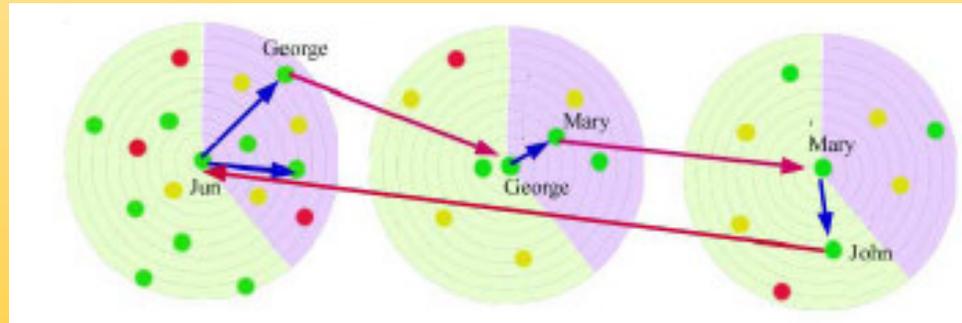
# Query Incentive Networks

- Kleinberg and Raghavan (2005)
- **Setting:** Need for something say  $T$  e.g., information, goods etc.
- Initiate a request for  $T$  with a corresponding reward, to some person  $X$
- $X$  can
  - Answer the query
  - Do nothing
  - Forward the query to another person
- **Problem:** How much should  $X$  “skim off” from the reward, before propagating the request?
- A Game Theoretic Model of Networks
  - **query routing in the social network is described as a game**
- Nodes can use strategies for deciding amongst offers
- All nodes are assumed to be rational
- A node will receive the incentive after the answer has been found
- Thus maximize one's incentive offering part of the incentive to others
- **Convex Strategy Space:** Nash Equilibrium exists



# Information Search in Social Network

- Zhang and Alstyne (2004) provide a small world instant messenger (SWIM) to incorporate social network search functionalities into instant messenger
  - Each actor's profile information (e.g. expertise) is maintained
  - Actor issues query → forward it to his/her network → return list of experts to actor → actor chats with a selected expert to obtain required information



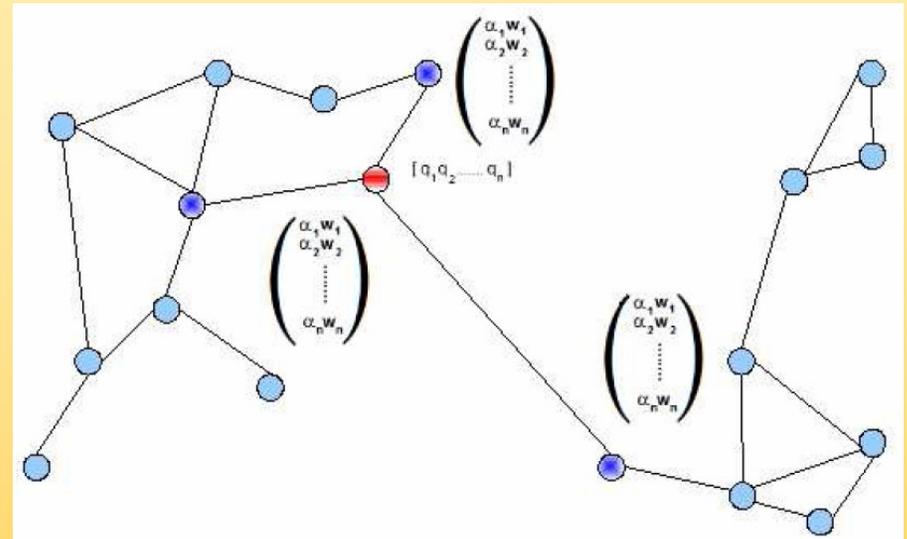
**SWIM search and refer process** (Source: Zhang and Alstyne 2004)

# Search and Expert Identification

- Setting: A decentralized knowledge market in the form of a social network
- Goal: Identify experts and route queries in the network
- Solution: An ant colony optimization technique that keeps track of the history of past queries.
- Advantages:
  - No need to keep track of ‘topics’ as topics can evolve.
  - Takes into account the dynamic nature of the network.
  - Track the changes in expertise of nodes over time.

(Ahmad and Srivastava 2008)

Related Work: (Yu and Singh, 2003)

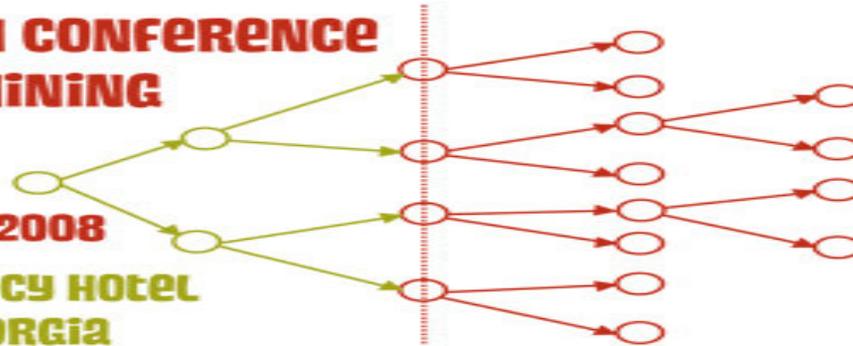


Source: (Ahmad and Srivastava 2008)

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



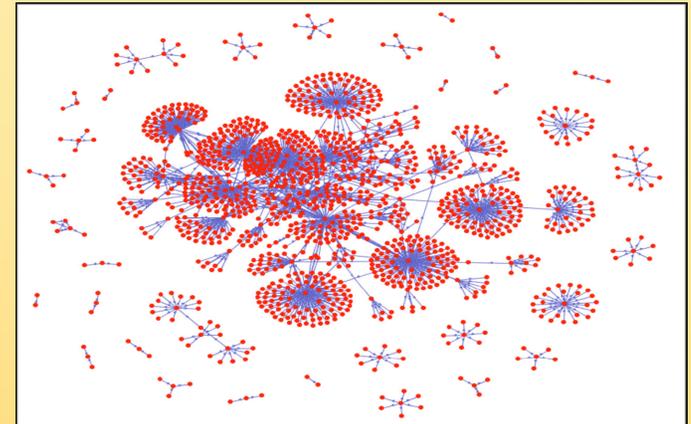
## Trust in Social Networks

# Trust in Social Networks

- **Trust propagation:** An approach for inferring trust values in a network
  - A user trusts some of his friends, his/her friends trust their friends and so on...
  - Given trust and/or distrust values between a handful of pairs of users, can one predict unknown trust/distrust values between any two users
- Golbeck et al (2003) discusses trust propagation and its usefulness for the semantic web
- TrustMail
  - Consider research groups X and Y headed by two professors such that each professor knows the students in their respective group
  - If a student from group X sends a mail to the professor of group Y then how will the student be rated?
  - Use the rating of professor from group X who is in professor Y's list of trusted list and propagate the rating
- Example of a real life trust model – [www.ebay.com](http://www.ebay.com)

# Trust in Social Networks

- **TidalTrust Algorithm** (Golbeck, 2005)
  - A source is more likely to believe the trust ratings, regarding a third person (sink), from a *close and highly trusted* actor
  - Using BFS all paths with the minimum length from source to sink are determined
  - Trust rating for a path is the minimum trust rating along that path
  - Use weighted average of trust ratings only from those paths on which source trusts its neighbour  $> \max \{\text{trust score of all paths}\}$
- **Propagation of Trust and Distrust in Networks** (Guha et al, 2004)
  - Propose a framework of trust propagation schemes
  - Modelled via a matrix of Beliefs  $B = T$  (Trust matrix) or  $B = T-D$  (Trust – Distrust)
  - Applications of atomic propagations are used to propagate trust values
    - E.g. Trust is transitive -  $B*B$ , co-citation -  $B*BTB$
  - Various schemes for chaining atomic propagations
  - **Goal:** Produce a final matrix  $F$  from which one can read off the computed trust or distrust of any two users

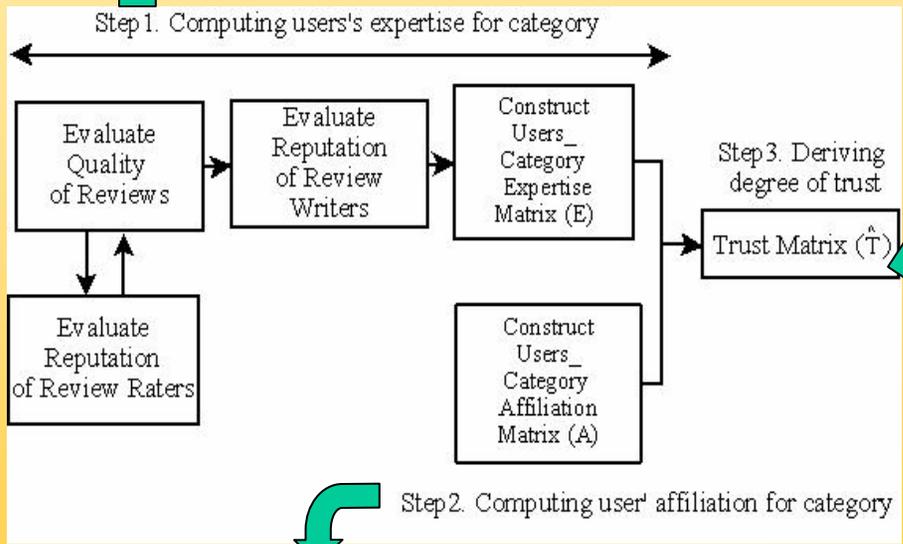
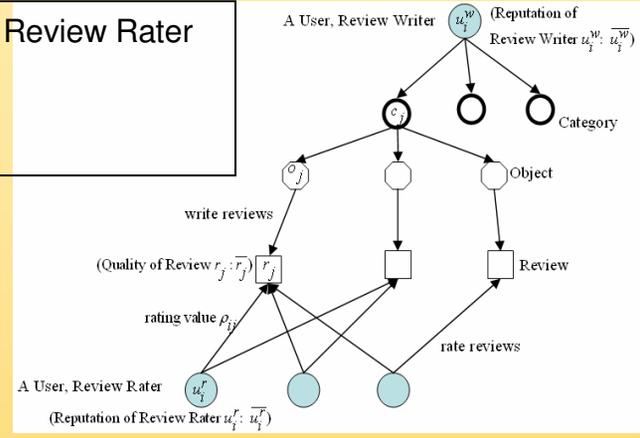


(Source: Golbeck, 2005)

# Building a Web of Trust w/o Explicit Trust Ratings<sup>1</sup>

A framework for deriving degree of trust

- 1-1: Calculating Quality of a Review and Reputation of a Review Rater
- 1-2: Calculating Reputation of a Review Writer
- 1-3: Constructing Users\_Category Expertise Matrix E



The relationship between a review writer and a review rater

$$A_{ij} = \left( \frac{a_{ij}^r}{\max_{j \in \text{all category}} (a_{ij}^r)} + \frac{a_{ij}^w}{\max_{j \in \text{all category}} (a_{ij}^w)} \right) \times \frac{1}{2}$$

$A_{ij}$  is the affiliation of a user  $i$  for category  $j$  ( $0 \leq A_{ij} \leq 1$ )  
 $a_{ij}^r$  is the total number of review that a user  $i$  rates on a category  $j$   
 $a_{ij}^w$  is the total number of review that a user  $i$  writes on a category  $j$

$$\hat{T}_{ij} = \frac{\sum_{\text{category } c} A_{ic} E_{cj}^T}{\sum_{\text{category } c} A_{ic}}$$

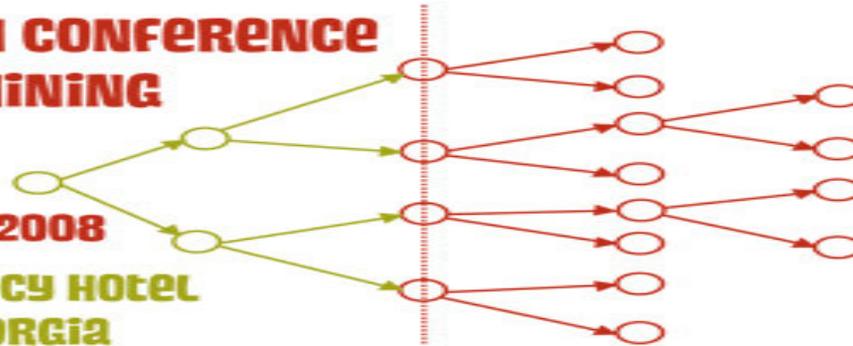
$\hat{T}_{ij}$  is the degree of trust that a user  $i$  holds for a user  $j$  ( $0 \leq \hat{T}_{ij} \leq 1$ )  
 $A_{ic}$  is an affiliation level of user  $i$  for a category  $c$   
 $E_{jc}$  is expertise value of user  $j$  on a category  $c$

1. Young Ae Kim, Hady W. Lauw, Ee-Peng Lim, Jaideep Srivastava, Building a Web of Trust without Explicit Trust Ratings, ICDE 2008 Workshop.

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Characterization of Social Networks

# Social Network Characterization

- Both the large size of social network data and the high complexity of social network models make SNA even harder today. Mislove et al. (2007)

	Flickr	LiveJournal	Orkut	YouTube
Number of users	1,846,198	5,284,457	3,072,441	1,157,827
Estimated fraction of user population crawled	26.9%	95.4%	11.3%	unknown
Dates of crawl	Jan 9, 2007	Dec 9 - 11, 2006	Oct 3 - Nov 11, 2006	Jan 15, 2007
Number of friend links	22,613,981	77,402,652	223,534,301	4,945,382
Average number of friends per user	12.24	16.97	106.1	4.29
Fraction of links symmetric	62.0%	73.5%	100.0%	79.1%
Number of user groups	103,648	7,489,073	8,730,859	30,087
Average number of groups memberships per user	4.62	21.25	106.44	0.25

Table 1: High-level statistics of our social networking site crawls.

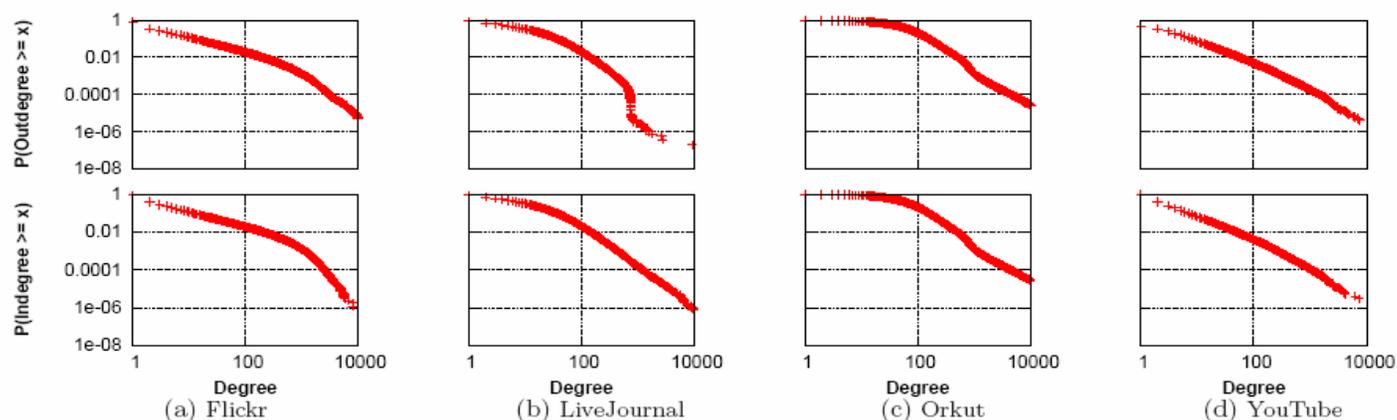


Figure 2: Log-log plot of outdegree (top) and indegree (bottom) complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

# Social Network Characterization

- Mislove et al. (2007)

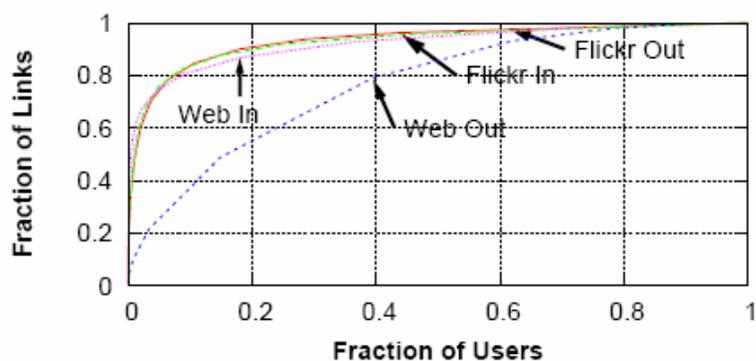


Figure 3: Plot of the distribution of links across nodes. Social networks show similar distributions for outgoing and incoming links, whereas the Web links shows different distributions.

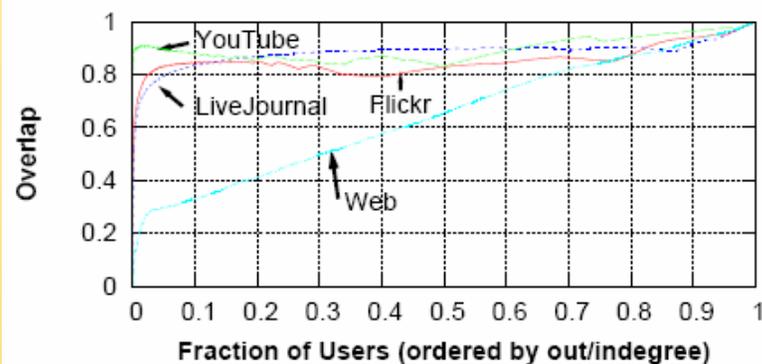


Figure 4: Plot of the overlap between top  $x\%$  of nodes ranked by outdegree and indegree. The high-indegree and high-outdegree nodes are often the same in social networks, but not in the Web.

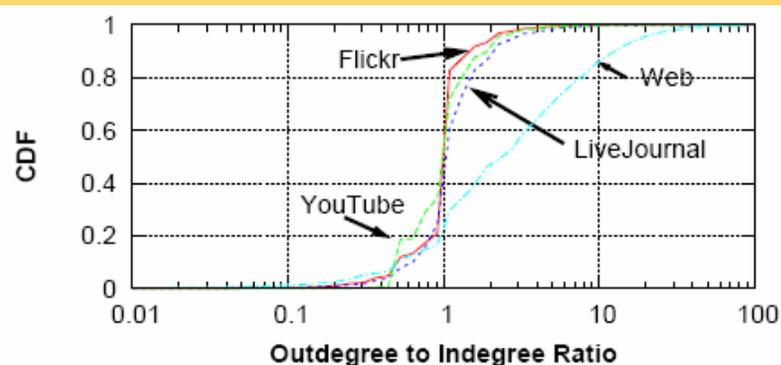


Figure 5: CDF of outdegree to indegree ratio. Social networks show much stronger correlation between indegree and outdegree than the Web.

# Social Network Characterization

- Mislove et al. (2007)

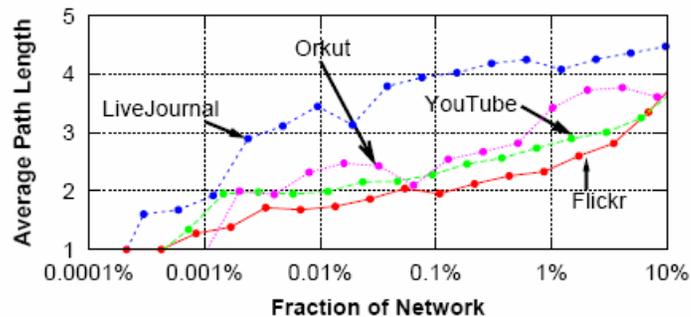


Figure 8: Average path length among the most well-connected nodes. The path length increases sub-logarithmically.

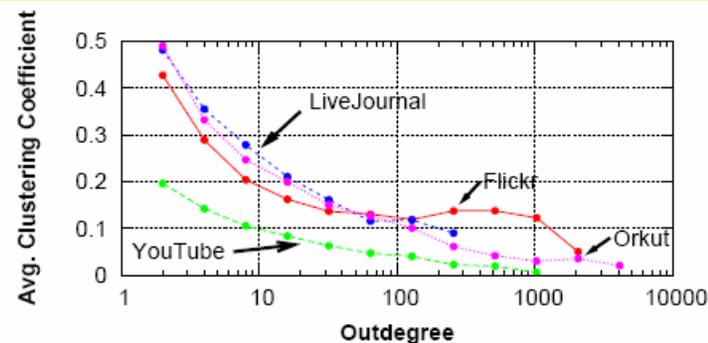


Figure 9: Clustering coefficient of users with different outdegrees. The users with few “friends” are tightly clustered.

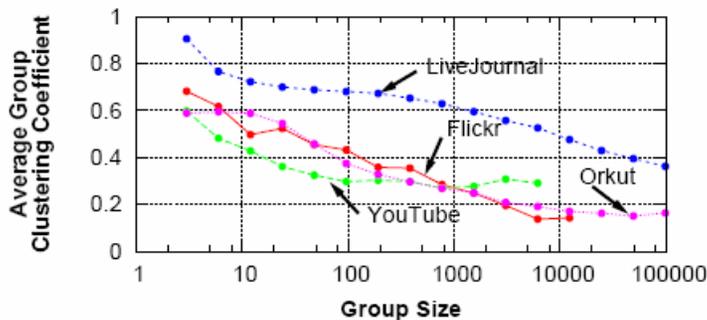


Figure 10: Plot of group size and average group clustering coefficient. Many small groups are almost cliques.

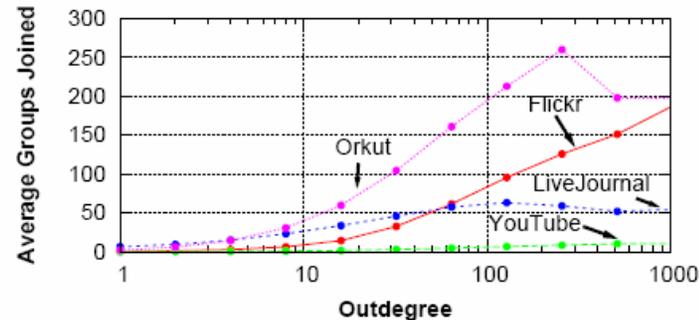
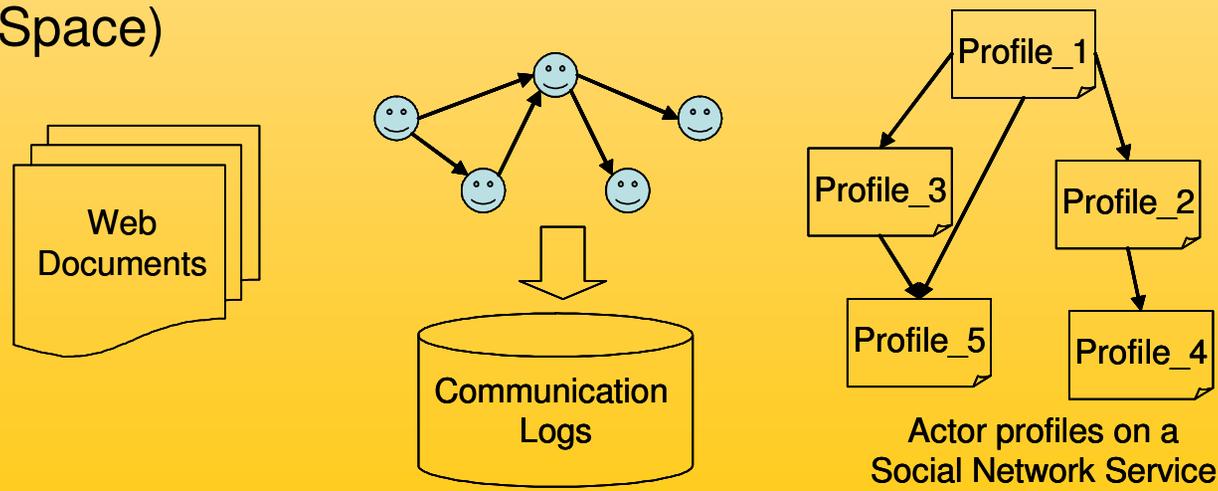


Figure 11: Outdegree versus average number of groups joined by users. Users with more links tend to be members of many groups.

# Social Network Extraction

- Mining a social network from data sources
- Hope et al (2006) identify three sources of social network data on the web
  - Content available on web pages (e.g. user homepages, message threads etc.)
  - User interaction logs (e.g. email and messenger chat logs)
  - Social interaction information provided by users (e.g. social network service websites such as Orkut, Friendster and MySpace)

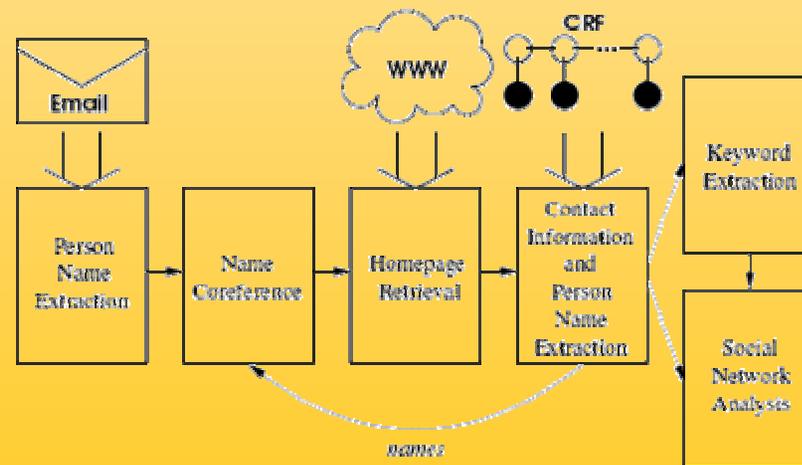


# Social Network Extraction

- **IR based extraction from web documents: Adamic and Ader (2003), Makrehchi and Kamel (2005), Matsumura et al, (2005)**
  - **Construct an “actor-by-term” matrix**
  - **The terms associated with an actor come from web pages/documents created by or associated with that actor**
  - **IR techniques such as tf-idf, LSI and cosine matching or other intuitive heuristic measures are used to quantify similarity between two actors’ term vectors**
  - **The similarity scores are the edge label in the network**
    - **Thresholds on the similarity measure can be used in order to work with binary or categorical edge labels**
    - **Include edges between an actor and its k-nearest neighbors**
- **Co-occurrence based extraction from web documents Matsuo et al (2006), Kautz et al (1997), Mika (2005)**
  - **For each pair of actors X and Y, issue queries of the form “X and Y”, “X or Y”, “X” and “Y” using a search engine (such as Google) and record corresponding number of hits**
  - **Use the number of hits to quantify strength of social relation between X & Y**
    - **Jaccard Coefficient –  $J(x,y) = (\text{hits}_{X \text{ and } Y}) / (\text{hits}_{X \text{ or } Y})$**
    - **Overlap Coefficient –  $OC(x,y) = (\text{hits}_{X \text{ and } Y}) / \min\{\text{hits}_X, \text{hits}_Y\}$**
    - **See (Matsuo 2006) for a discussion on other measures**
  - **Expand the social network by iteratively adding more actors**
    - **Query known actor X and extract unknown actors from first k hits**

# Social Network Extraction

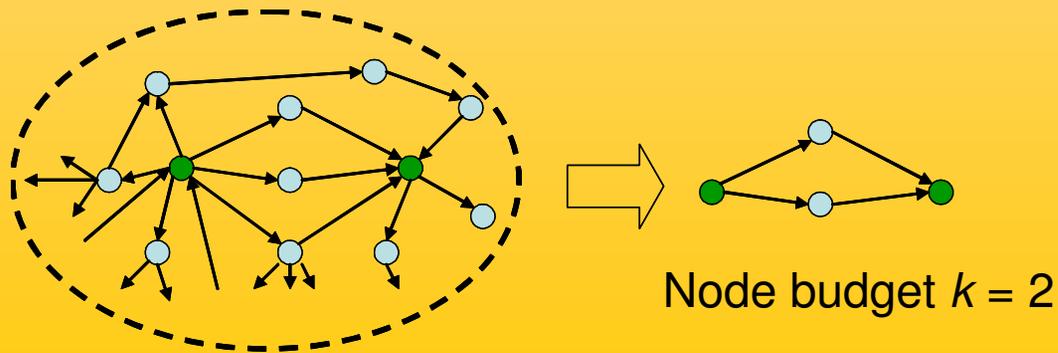
- **Lauw et al (2005) discuss a co-occurrence based approach for mining social networks from spatio-temporal events**
  - Logs of actors' movements over various locations are available
  - Events can occur at irregular time intervals
  - Co-occurrence of actors in the space-time domain are mined and correspondingly a social network graph is generated
- **Culotta et al (2004) present an end-to-end system for constructing a social network from email inboxes as well as web documents**
- **Validation of results is generally ad-hoc in nature due to lack of actual social network**



(Source: Culotta et al, 2004)

# Approximating Large Social Networks

- **Approximating a large social network allows for easier analyses, visualization and pattern detection**
- **Faloutsos et al (2004)**
  - Extracting a “connection subgraph” from a large graph
  - A connection subgraph is a small subgraph that best captures the relation between two given nodes in the graph using at most  $k$  nodes
  - Used to focus on and summarize the relation between any two nodes in the network
  - The node “budget”  $k$  is specified by the user
  - Optimize a goodness function based on an ‘electrical circuit’ model
    - The goodness function is the quantity of current flowing between the two given nodes
    - Edge weights between nodes are used as conductance values
    - A universal sink is attached to every node in order to penalize high degree nodes and longer paths



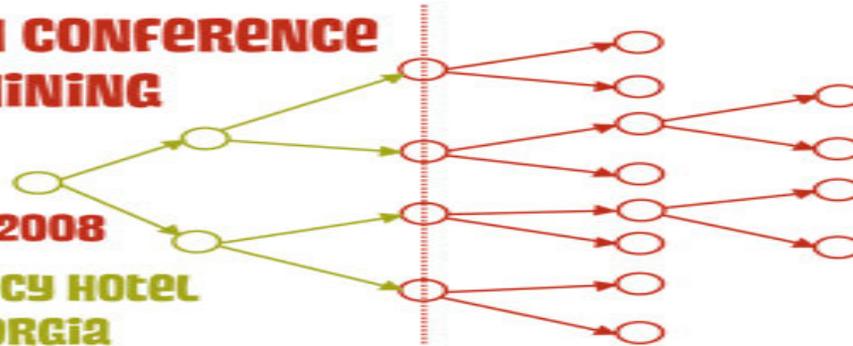
# Approximating Large Social Networks

- **Leskovic and Faloutsos (2006)** compare various strategies for sampling a small representative graph from a large graph
  - Strategies: Random Node, Random Edge, Random Degree Node, Random Edge Node, Random Walk etc.
  - Various graph distribution properties are compared between samples and original graph
  - Random Walk performs best for sampling from large static graph
  - Also discuss sampling history of evolution of a graph
- **Wu et al (2004)** presents an approach for summarizing scale-free networks based on shortest paths between vertices
  - Determine  $k$  number of “median” vertices such that the average shortest path from any vertex to its closest median vertex is minimized
  - Length of shortest path  $p$  between any two vertices is approximated by the sum of
    - shortest distance between median vertices for the clusters of the two vertices + sum of shortest distance between the vertices and their respective medians
  - Median vertices are chosen as the nodes with highest degree, HITS score, betweenness centrality and random selection
  - Further efficiency can be achieved by recursively clustering a graph and working with a hierarchy of simplified graphs
  - Used for approximating shortest paths and their lengths for large graphs
  - Around 20% error observed for querying a graph one magnitude smaller

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Anonymity in Social Networks

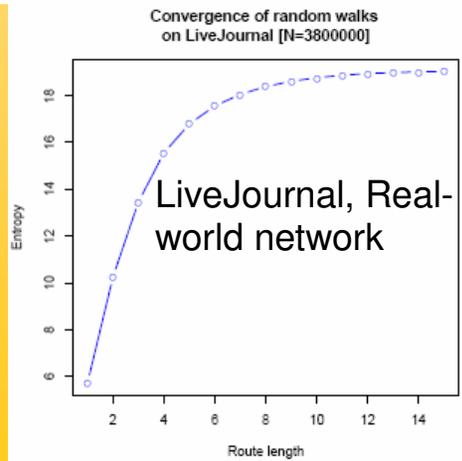
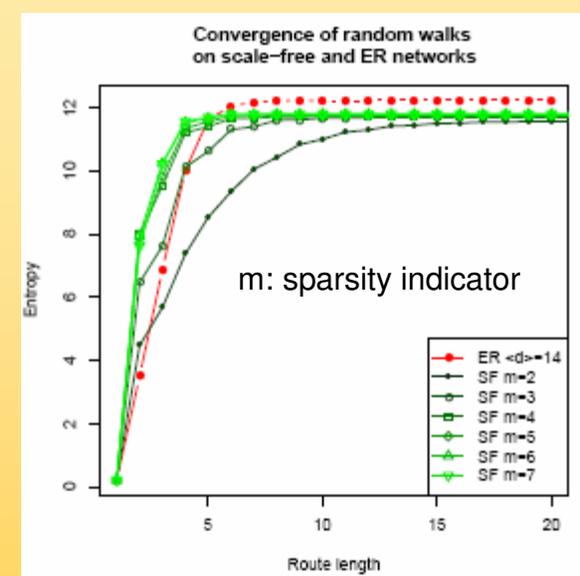
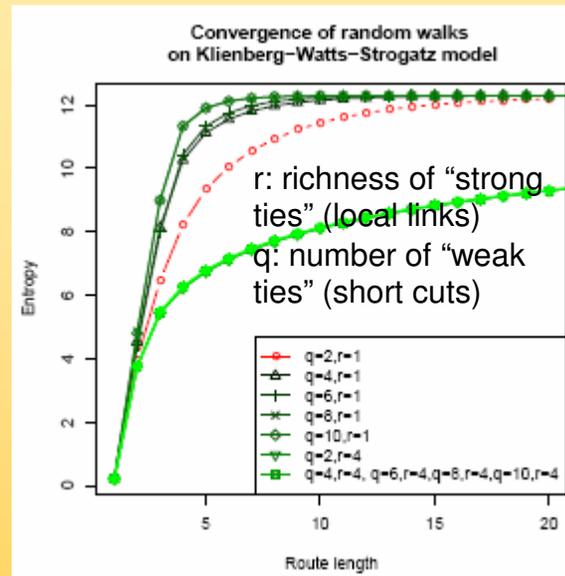
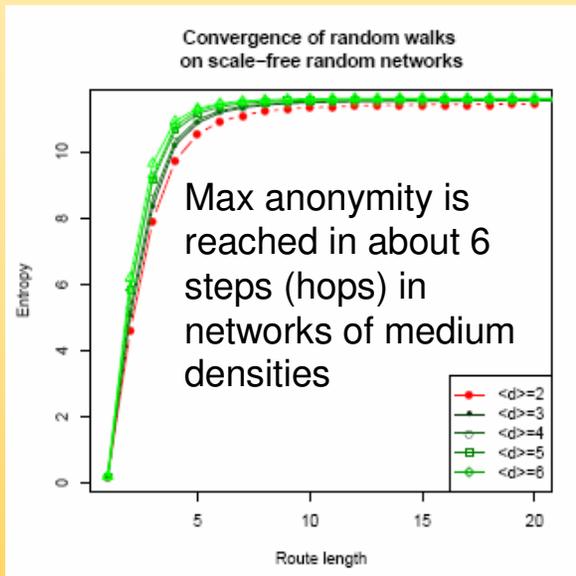
# Anonymity in Social Networks

- Anonymity in the Wild: Mixes on unstructured networks (Nagaraja, 2007)
  - Message anonymity: The ability to protect message communication (i.e., “who sent the message to whom”) from being identified by attackers
  - “The anonymity of a system is the entropy of the probability distribution over all the actors that they committed a specific action”
    - $\epsilon$  is the entropy,  $\alpha_i$  is an actor (sender or receiver)

$$\mathcal{A} = \mathcal{E}[\alpha_i] = - \sum_i Pr[\alpha_i] \log_2 Pr[\alpha_i]$$

# Anonymity in Social Networks

- Simulation results (Nagaraja, 2007)



When the local links are not "strong" enough ( $r=1$ ), it converges in about 7/8 steps, especially with enough short cuts ( $q$  is large). When there are "strong" local links ( $r=4$ ), there are strong community structures, and it converges very slowly, regardless of the amount of shortcuts.

Sparse networks make it converge slowly, while in dense networks it converges within 6 steps.

# Anonymity in Social Networks

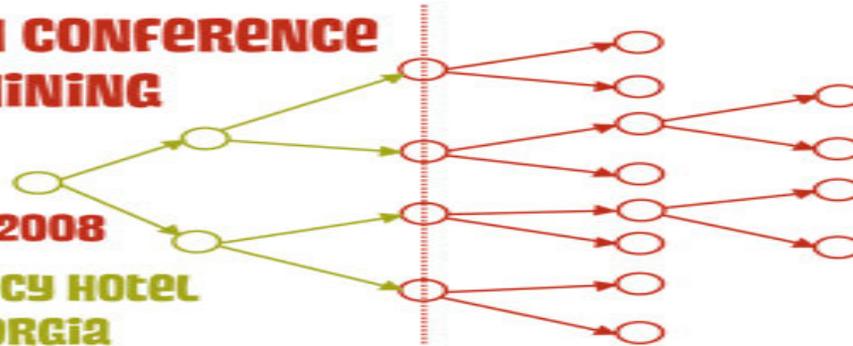
- Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, (Backstrom et al., 2007)
  - **Anonymized copy of social network: Replacing names/identities with random unique codenames**
  - **Attacks: Recovery of original and private names/identities from an anonymized copy of social network**

Type of attack	Steps
Active, e.g., (1) walked-based, (2) cut-based	1. Create a small set of new nodes; $O((\log n)^{1/2})$ 2. Connect new nodes to targeted nodes by adding new edges 3. Build edges with special patterns among these new nodes 4. Discover edges and nodes with the special patterns and then edges and the targeted nodes
Passive	1. Create nothing; no new nodes and edges are added 2. Discover nodes representing attackers themselves, or locate themselves with local structure 3. Identify existence of edges (subgraphs) among nodes linked by attackers (“coalition” network) 4. Find (private) edges and nodes around the subgraphs; not any arbitrary node can be targeted
Semi-passive	Create no new nodes but a set of new edges to targeted nodes; it is possible to attack specific nodes by using algorithms based on passive attack

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

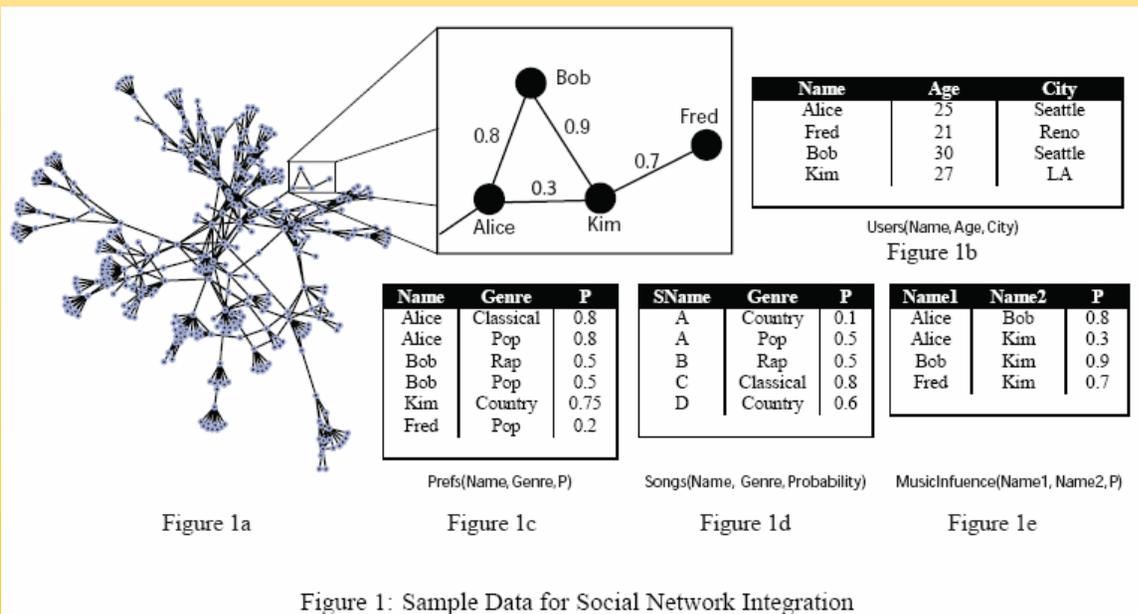
**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Other Research in Social Networks

# Uncertainty in Social Networks

- Adar and R'e (2007)
  - The ability to process and manage a large quantities of uncertain data (i.e., the imprecision in data) is critical in commercializing research projects.



```

SELECT U.name
FROM Users U, Prefs P, Songs S
WHERE U.name = P.name
AND P.Genre = S.genre
AND S.name = 'A'
(a)

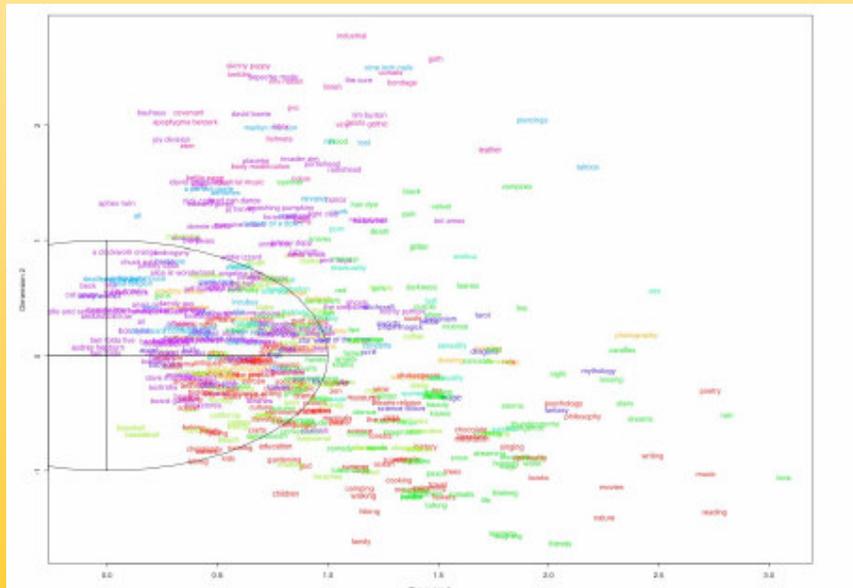
SELECT U.Name
FROM Users U , Prefs P, Song S,
MusicInfluence M, Recommends R
WHERE U.name = P.name AND P.Genre = S.genre
AND M.name2 = U.name AND M.name1 = R.from
AND R.To = U.name AND R.song = S.sname
AND S.sname = 'D'
(b)
    
```

Figure 2: Sample Queries

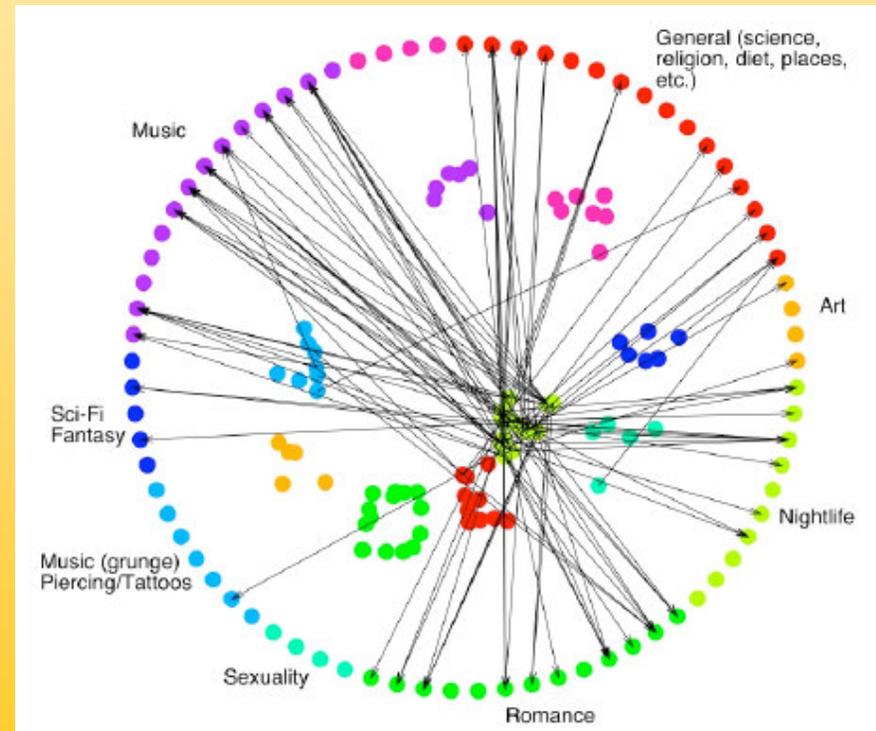
Simple queries in a relational database (or OLAP)

# Visualization

- **Semantic web and social network analysis**
  - **Paolillo and Wright (2005) provide an approach to visualizing FOAF data that employs techniques of quantitative Social Network Analysis to reveal the workings of a large-scale blogging site, LiveJournal**



Plot of nine interest clusters along the first two principal clusters (Paolillo and Wright, 2005)

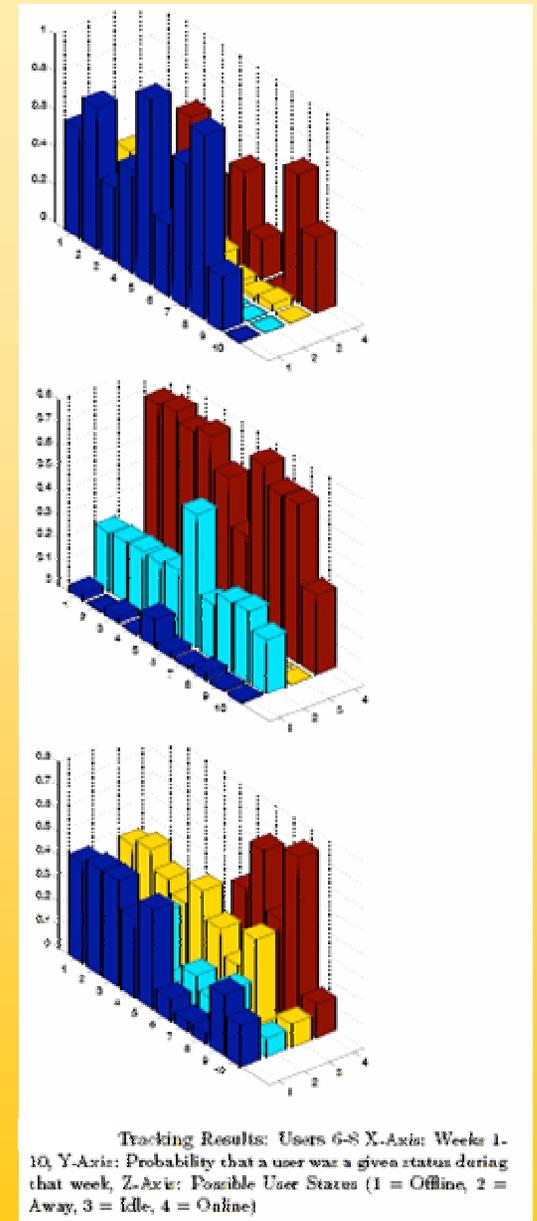
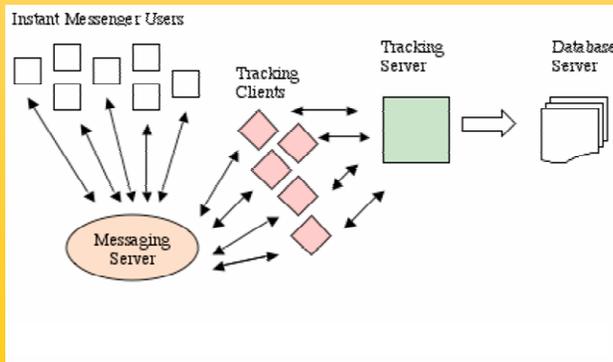


Relation of interest clusters to groups of actors with shared interests (Paolillo and Wright, 2005)

# SNA from IM Networks

## IMSCAN: A Framework for IM Mining (Resig et al 2004, Teredesai et al 2004)

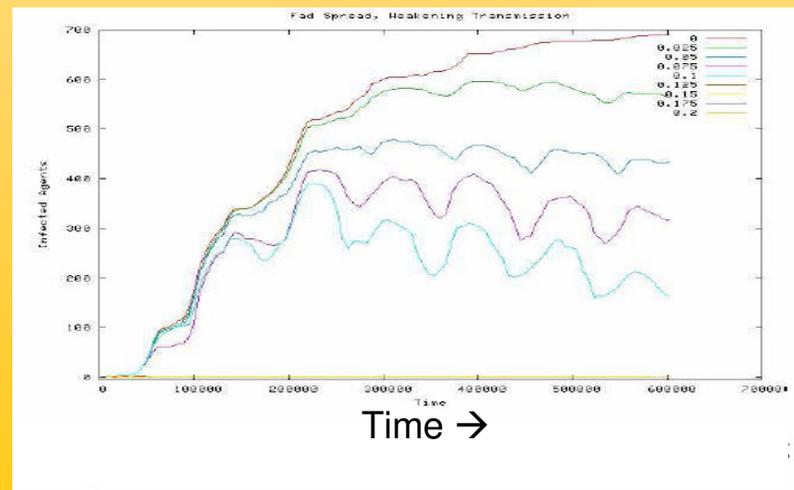
- Infer social networks based on user status behavior.
- Do communities of users behave in a similar manner?
- User LiveJournal for validation
- No correlation between amount of time spent online and LiveJournal popularity
- Use similarity in context to cluster users



# SNA from IM Networks

- Reginald studied the structure of an instant messaging network and determined it to be a scale free network. (Reginald 2004)
- Simulation of how fads and non-fads proliferate in an instant messaging behavior.
  - After a certain saturation point, fads can exhibit a periodic spreading behavior. (Ahmad and Teredesai 2006)

# of Infected Actors →



# SNA from Online Networks

- (Golder et al 2007) studied 362 million messages exchanged by 4.2 million users on the Facebook.

## Key Observations:

- Poking and messaging patterns are extremely similar.
- Activity on the online social network varies depending upon the time of the day.
- Different patterns are observed in a corporate messaging network as compared to Facebook suggesting different nature of interaction.
- Interaction on the Facebook does not represent leisure time but rather interaction in parallel with other work.
- Most messages are sent to friends but vice versa is not true.

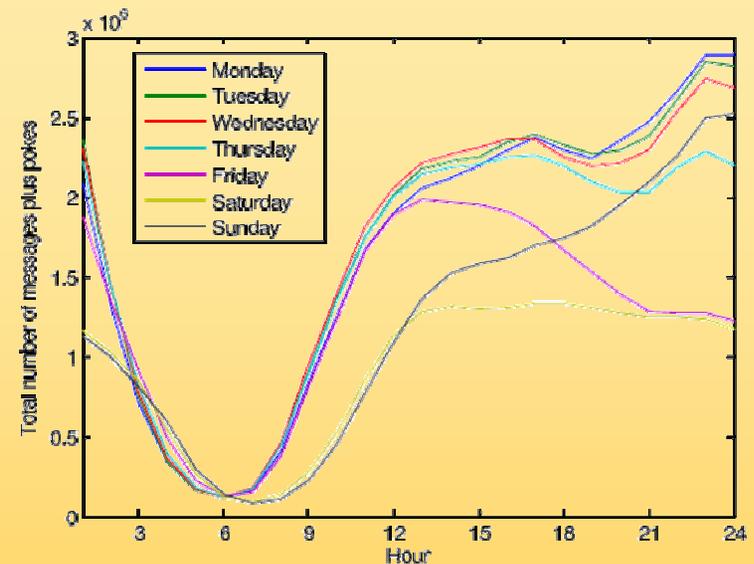


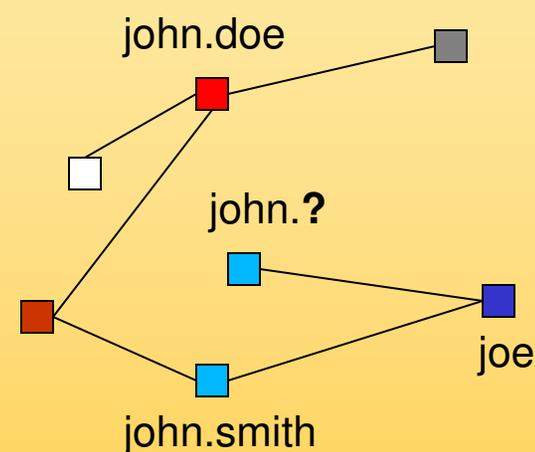
Fig. 1.3. Message plus pokes sent by hour in Facebook (color)

# Learning Social Networks

- Learning Social Networks Using Multiple Resampling Method (Makrehchi and Kamel, 2007)
  - Goal: To learn unknown relations
    - Application: Learning relations in an FOAF (Friend Of A Friend) network
  - Learning social network  $\leftrightarrow$  text classification
    - SVM with linear kernel is employed (to handle high dimensionality)
  - Actor modeling: vector of documents, e.g., homepage, blog, CV, etc.
  - Relation modeling: a relation vector is a vector of words from documents associated with actors
    - Aggregate document vectors between two actors using MAX operator
    - Known relation is the label for a relation vector
  - Social networks are so sparse that training data become imbalance
    - To re-balance training data: undersampling the majority, oversampling the minority

# Alias Detection

- Alias detection (or identity resolution)
  - Online users assume multiple aliases (e.g. email addresses)
  - Map multiple aliases to same entity
  - Approaches leverage information about communication in a social network to determine such aliases
- Bhattacharya and Getoor, (2006)
  - Bayesian modelling approach for identity resolution
  - Model maps multiple aliases as well as social links for them to a user
- Related Work: Hill (2003), Malin (2005), Holzer et al (2005)



Using social information  
for entity resolution

# Relationship Labeling in Social Networks

- Entity and relationship labeling in affiliation networks (Zhao et al 2006)
- Affiliation network: A social network made up of two types of nodes e.g., actors and events.
- Relational Markov Networks (RMN) were used for the experiments.
- The Profiles in Terror (PIT) dataset is use for experiments.
- Task: Predict labels between the various entities
- Relations are represented by random variables and the edges represent correlations.
- Since two terrorists can be related in multiple ways e.g., part of the same family and organization, multi-label classification is considered.

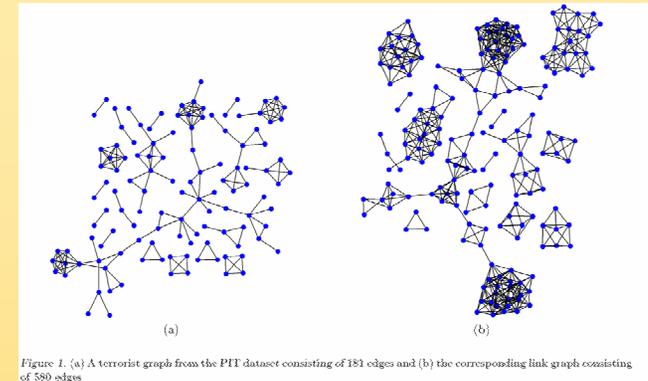


Figure 1. (a) A terrorist graph from the PIT dataset consisting of 181 edges and (b) the corresponding link graph consisting of 580 edges

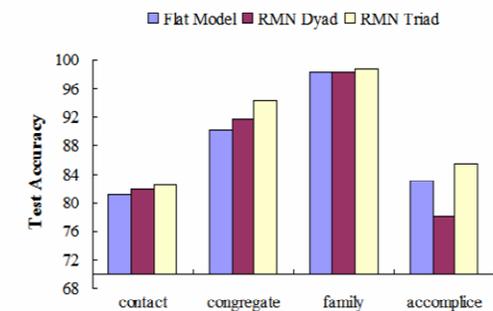


Figure 2. The average classification accuracy for binary terrorist relationship labeling.

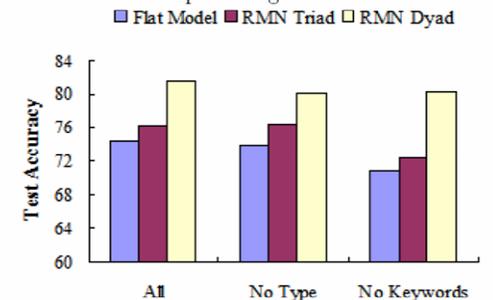


Figure 3. Average classification accuracy of terrorist relation labeling.

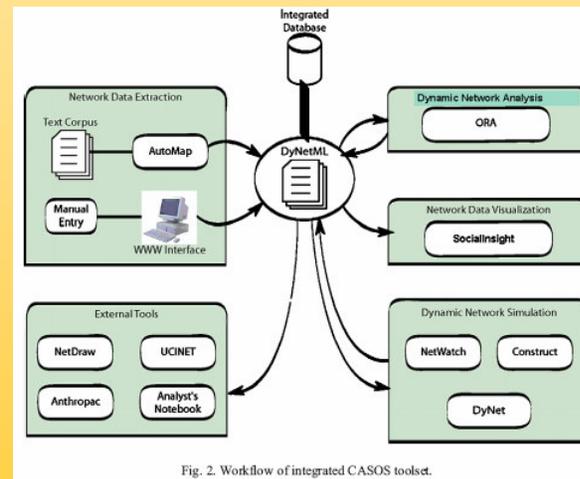
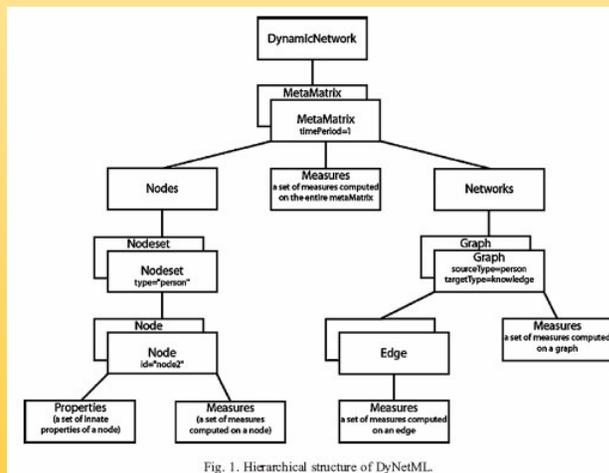
# Quadratic Assignment Procedure

- Problem: Are people more likely to be friends if they share similar characteristics, such as being about the same age?
- Observation: A significant correlation exists between the two.
- Hypothesis: Is the correlation observed because proximity constrains their friendship networks and advice networks
- How to check for spurious correlations in dyadic data (in Social Networks)?
- QAP (Quadratic Assignment Procedure) check if two variables are spuriously correlated with one another
- Combines least squares estimation with Hubert's non-parametric test

(Krackhardt 1987)

# SNA Toolkit

- Tsvetovat et al. (2004)
  - Store social structure data in the well-developed relational schema
- Carley et al. (2007)
  - Requirements of dynamic/social network analysis toolkits:
  - Extensibility, Interoperability, Ontologies, XML interchange language, Data storage and management, Scalability, Robustness of tools

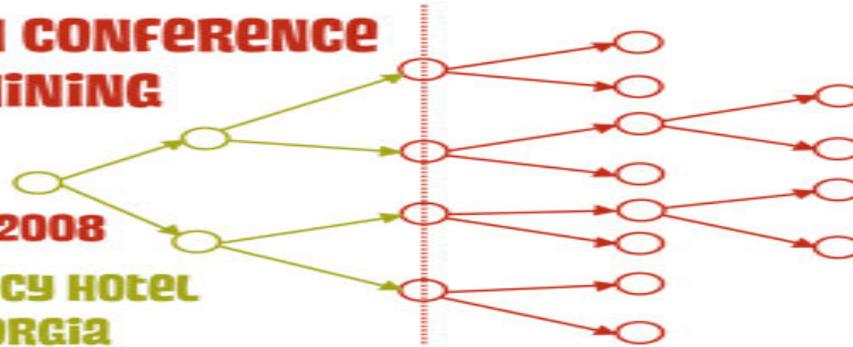


- Scalability is an issue, especially for data too large to fit in memory
  - (Hsu et al., 2008) Relational db and OLAP techniques are supportive
    - Algorithm, e.g., Bregman co-clustering algorithm, intensively (re-)compute summary statistics for analysis of underlying characteristics of data
    - Computing and managing summary statistics are the strength of OLAP

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



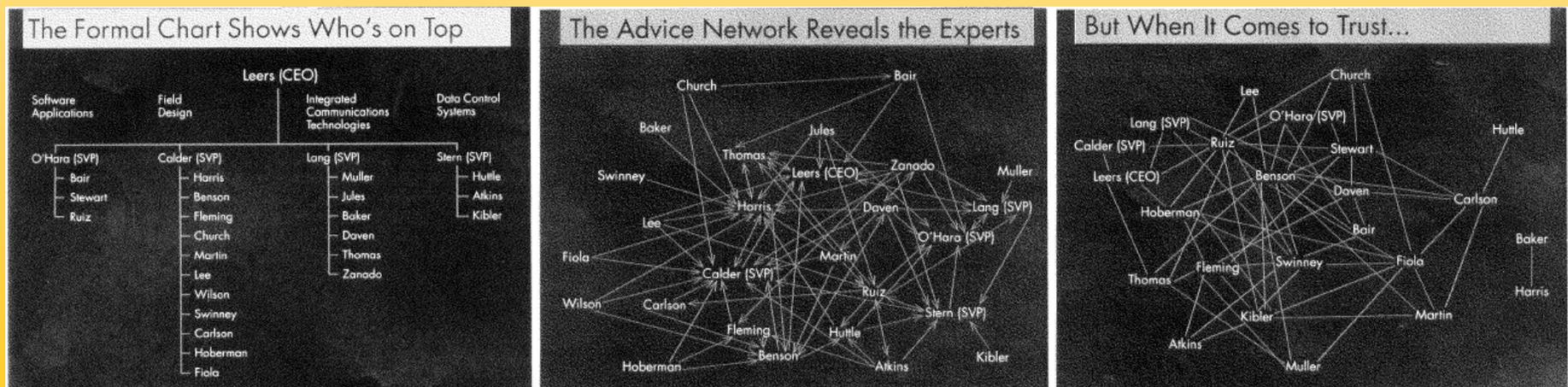
# Application of Data Mining Based Social Network Analysis Techniques

# Applications (Outline)

- Organization Theory
- Semantic Web
- Viral Marketing
- Social Influence and E-Commerce
- Social Computing
- Criminal Network Analysis
- Newsgroup Message Classification
- Social Recommendation Systems
- Terrorism and Crime Related Weblog Social Network

# Organization Theory

- **Krackhardt and Hanson (1993)**
  - Informal (social) networks present in an enterprise are different from formal networks
  - Different patterns exist in such networks like imploded relationships, irregular communication patterns, fragile structures, holes in network and bow ties
- **Lonier and Matthews (2004)**
  - Survey as well as study the impact of informal networks on an enterprise



(Source: Krackhardt and Hanson, 1993)

# Extracting Co-appearance Networks among Organizations

- Extracting Inter-Firm Networks from WWW (Jin et al., 2007)

Results from a search engine can be estimated in a more robust way (Matsuo et al., 2007)

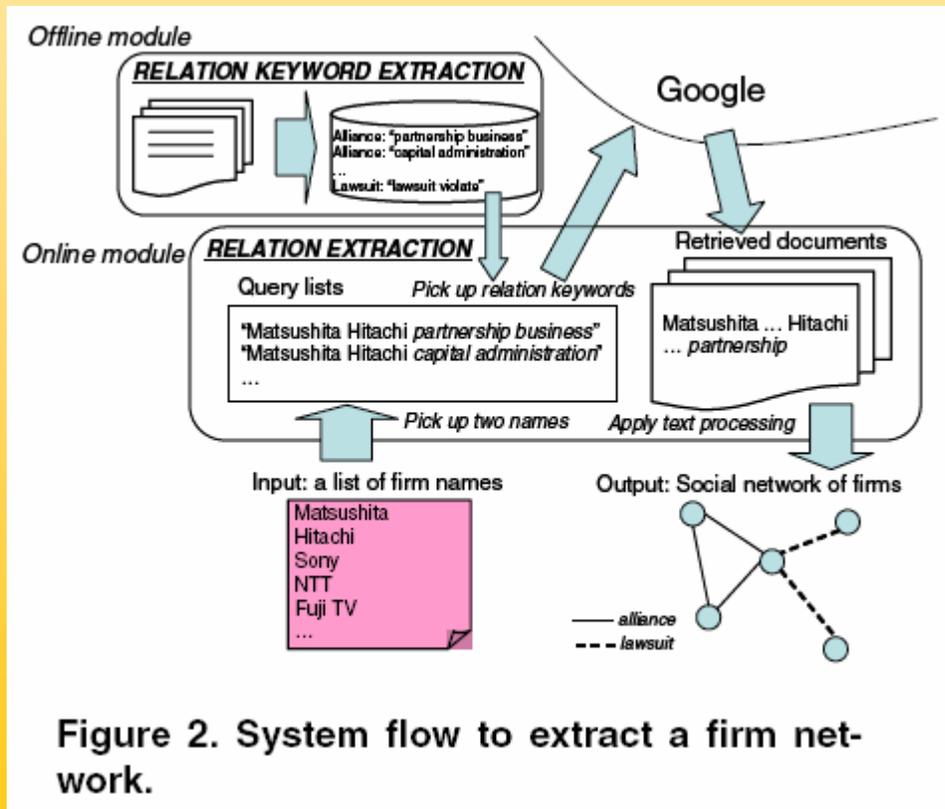
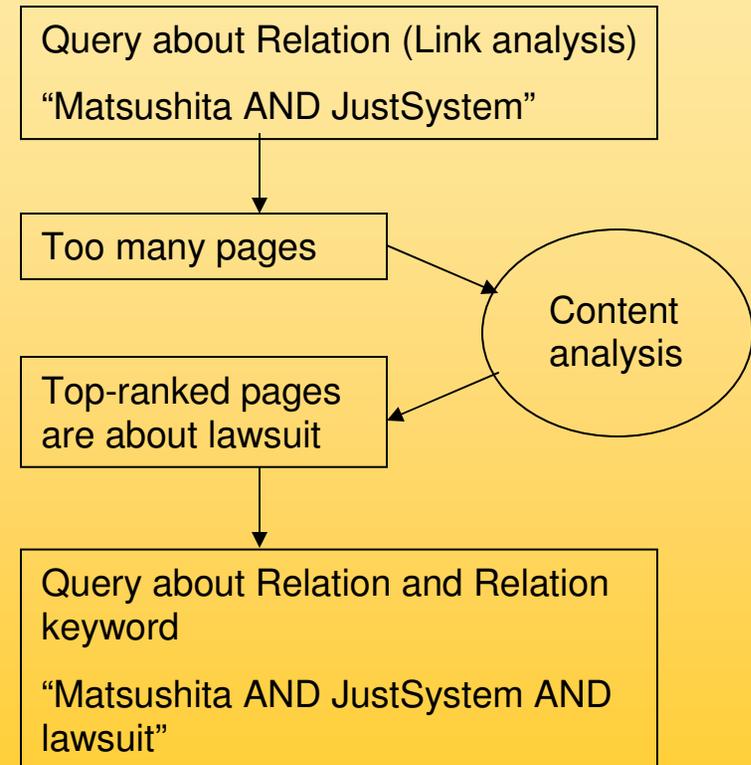


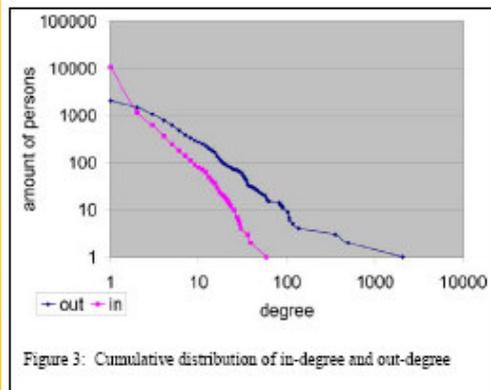
Figure 2. System flow to extract a firm network.



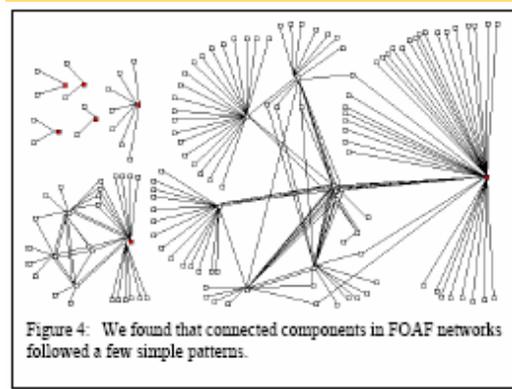
# Semantic Web Community

- Ding et al (2005)
  - Semantic web enables explicit, online representation of social information while social networks provide a new paradigm for knowledge management e.g. Friend-of-a-friend (FOAF) project (<http://www.foaf-project.org>)
  - Applied SNA techniques to study this FOAF data (DS-FOAF)

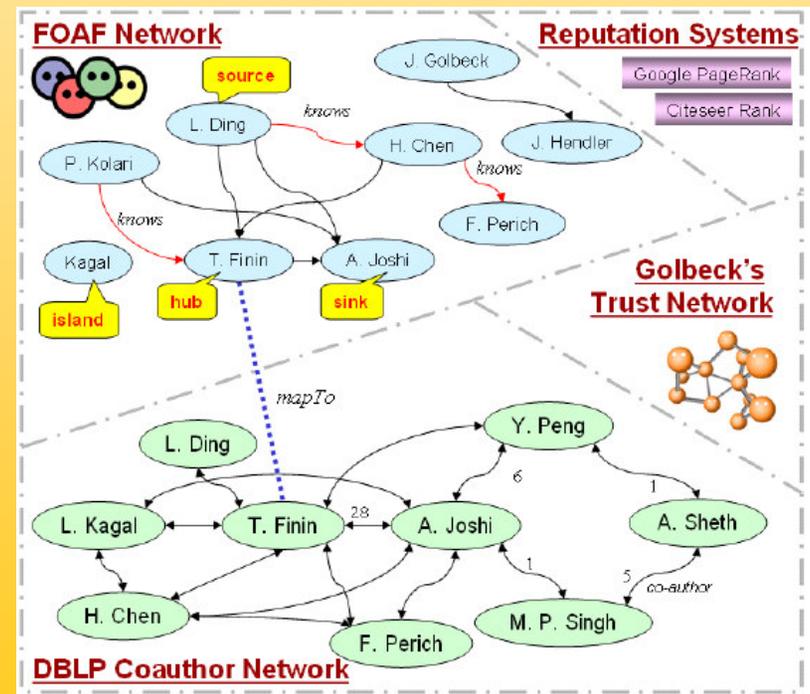
## Preliminary analysis of DS-FOAF data (Ding et al, 2005)



**Degree distribution**



**Connected components**



**Trust across multiple sources (Ding et al, 2005)**

# Semantic Web and SNA

- The friend of a friend (FOAF) project has enabled collection of machine readable data on online social interactions between individuals. <http://www.foaf-project.org>
- Mika (2005) illustrates Flink system (<http://flink.semanticweb.org/>) for extraction, aggregation and visualization of online social network.



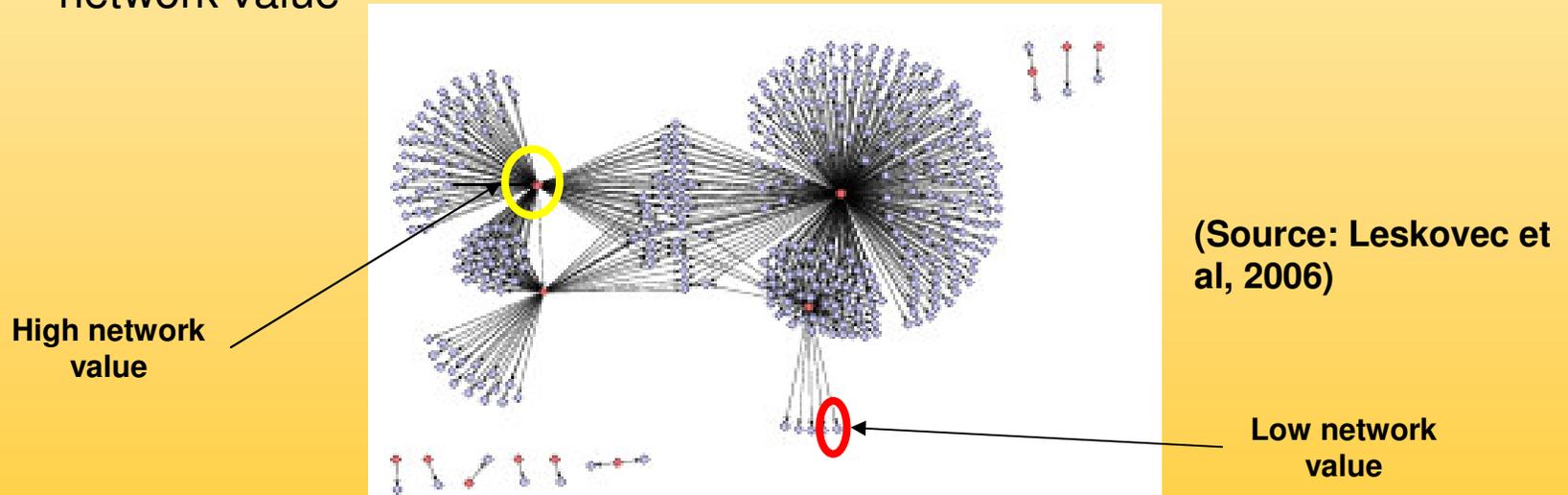
The Sun never sets under the Semantic Web: the network of semantic web researchers across globe (Mika, 2005)



Snapshot of clusters  
(<http://flink.semanticweb.org/>)

# Viral Marketing

- Domingos(2005), Domingos and Richardson (2001, 2002)
  - *Network value* of a customer is the expected profit from marketing a product to a customer, accounting for the customer's influence on the buying decisions of other customers
  - Propose a greedy strategy for identifying customers with maximum network value



- Kempe et al (2003)
  - For a general class of cascading models, the problem of identifying customers with maximum network value is NP-hard
  - A greedy strategy provides a solution within 63% of the optimal

# Social Influence and E-Commerce<sup>1</sup>

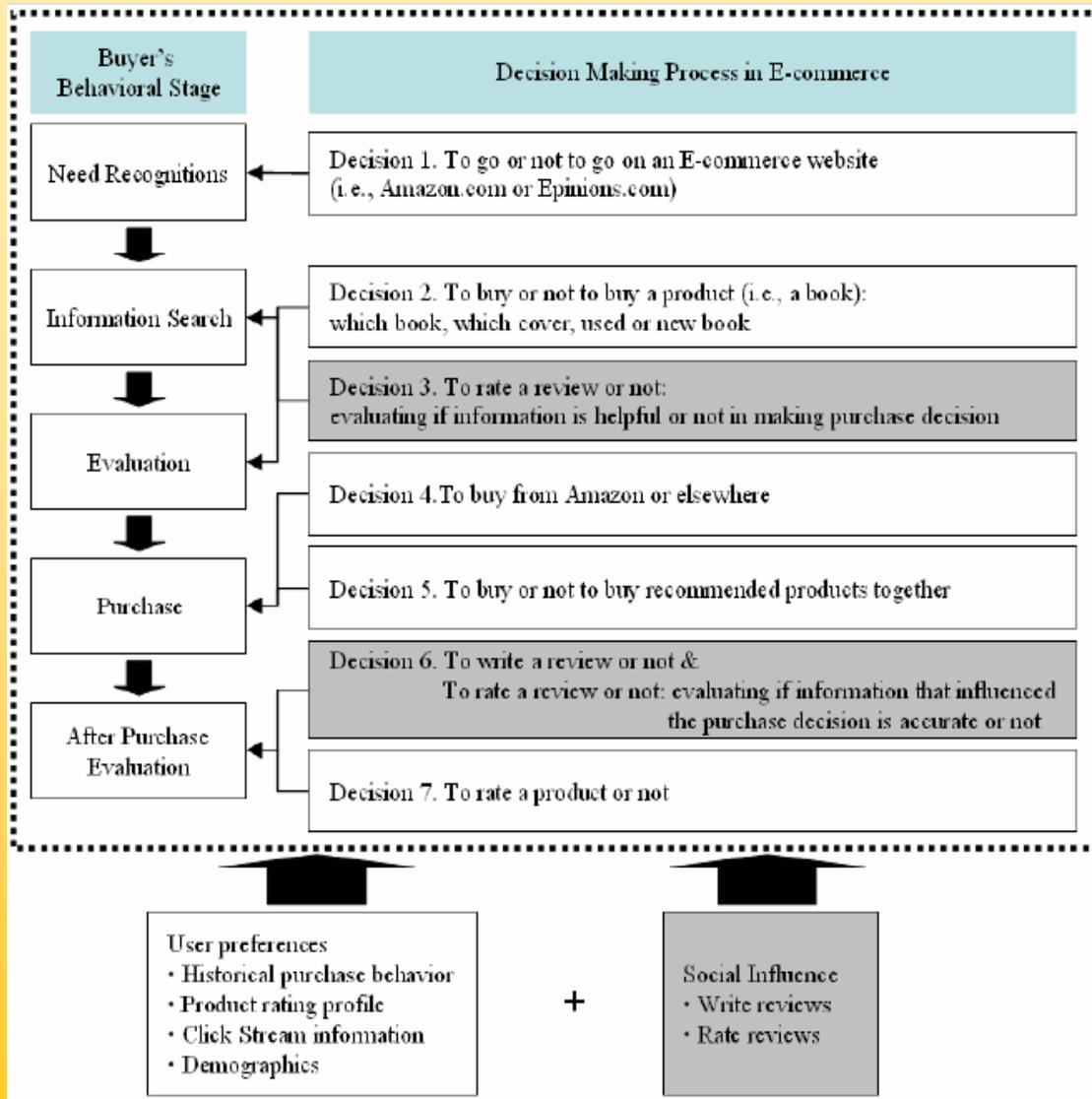


Figure 1. A decision making process in E-commerce

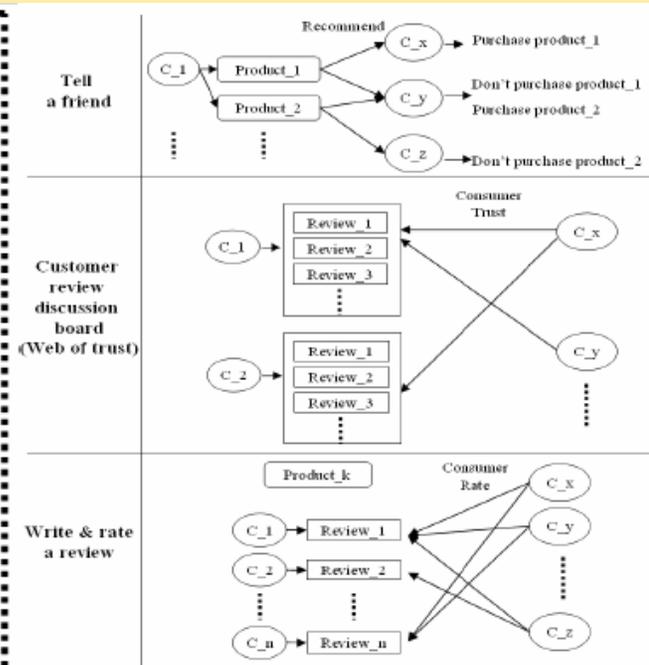


Figure 2. The data about social interaction

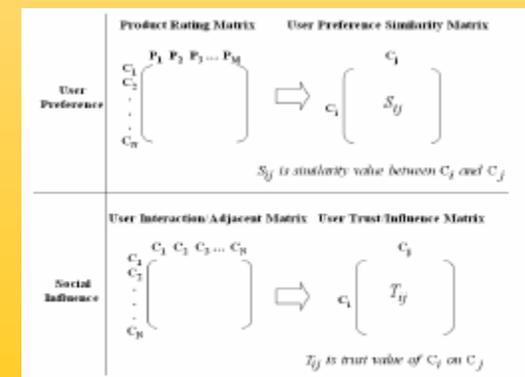


Figure 3. E-commerce information source

1. Young Ae Kim, Jaideep Srivastava: Impact of social influence in e-commerce decision making. ICEC 2007: 293-302

# Social Computing

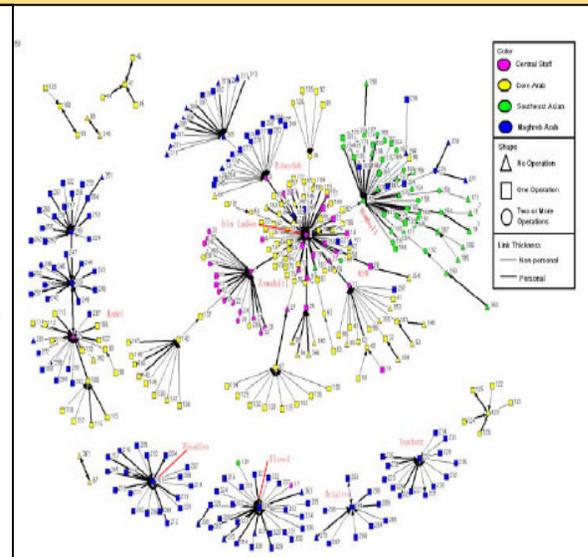
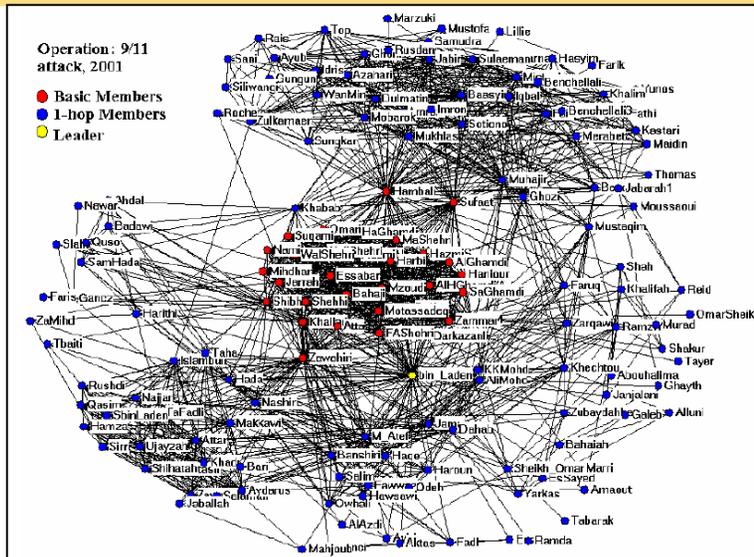
- Combining social computing and ubiquitous computing
  - iBand: A bracelet like device used for exchanging personal and relationship info.  
(Kanis et al. 2005)



# Criminal Network Analysis

- **Example (Qin et al, 2005)**
  - Information collected on social relations between members of Global Salafi Jihad (GSJ) network from multiple sources (e.g. reports of court proceedings)
  - Applied social network analysis as well as Web structural mining to this network
  - Authority derivation graph (ADG) captures (directed) authority in the criminal network

Ranking	Leader	Gatekeeper	Outlier
<b>Central Member</b>			
1	Zawahiri	bin Laden	Khalifah
2	Makkawi	Zawahiri	SbinLaden
3	Islambuli	Khadr	Ghayth
4	bin Laden	Sirri	M Atef
5	Attar	Zubaydah	Sheikh Omar
<b>Core Arab</b>			
1	Khallad	Harithi	Elbaneh
2	Shibh	Nashiri	Khadr4
3	Jarrah	Khallad	Janjalani
4	Atta	Johani	Dahab
5	Mihdhar	ZaMihd	Mehdi
<b>Maghreb Arab</b>			
1	Hambali	Baasyir	Siliwangi
2	Baasyir	Hambali	Fathi
3	Mukhlas	Gungun	Naharudin
4	Iqbal	Muhajir	Yunos2
5	Azahari	Setiono	Maidin
<b>Southeast Asian</b>			
1	Doha	Yarkas	Mujati
2	Benyaich2	Zaoui	Parlin
3	Fateh	Chaib	Mahdjoub
4	Chaib	DavidC	Zinedine
5	Benyaich1	Maaroufi	Ziyad



Terrorists with top centrality ranks in each clump

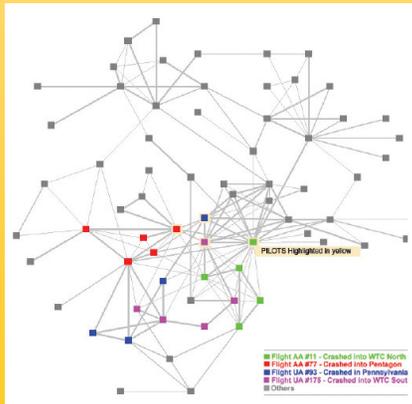
1-hop network of 9/11 attack

ADG of GSJ network

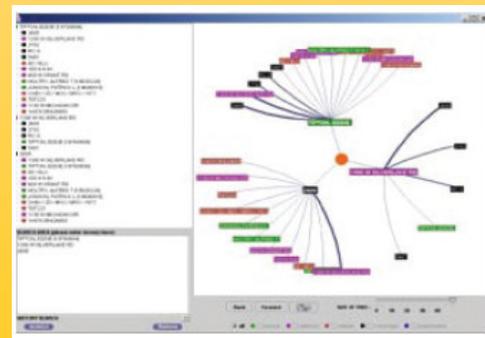
# Criminal Network Analysis

- Knowledge gained by applying SNA to criminal network aids law enforcement agencies to fight crime proactively
- Criminal networks are large, dynamic and characterized by uncertainty.
- Need to integrate information from multiple sources (criminal incidents) to discover regular patterns of structure, operation and information flow (Xu and Chen, 2005)
- Computing SNA measures like centrality is NP-hard
  - **Approximation techniques (Carpenter et al 2002)**
- Visualization techniques for such criminal networks are needed

*Figure: Terrorist network of 9/11 hijackers (Krebs, 2001/ Xu and Chen, 2005)*



Example of 1<sup>st</sup> generation visualization tool.



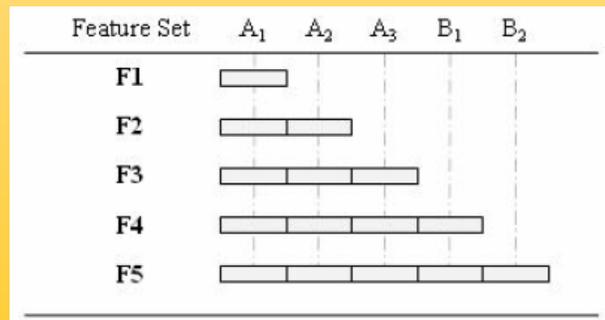
Example of 2<sup>nd</sup> generation visualization tool

# Newsgroup Message Classification

- Using SNA to help classify newsgroup messages (Fortuna et. Al, 2007)
  - SVM classifier
  - Rich feature set from “networks”



Networks where users socially interact with others through posting and replying



Networks where similarities between two nodes are determined by authors or contents

# Social Recommendation Systems

- **Initial approaches**
  - **Anonymous recommendations: treat individuals preferences as independent of each other**
  - **Failure to account for influence of individual's social network on his/her preferences**
- **Kautz et al (1997)**
  - **Incorporate information of social networks into recommendation systems**
  - **Enables more focused and effective search**
- **McDonald (2003)**
  - **Analyzes the use of social networks in recommendation systems**
  - **Highlights the need to balance between purely social match vs. expert match**
  - **Aggregate social networks may not work best for individuals**
- **Palau et al, (2004)**
  - **Apply social network analysis techniques to represent & analyze collaboration in recommender systems**
- **Lam (2004)**
  - **SNACK - an automated collaborative system that incorporates social information for recommendations**
  - **Mitigates the problem of cold-start, i.e. recommending to a user who not yet specified preferences**

# Social Recommendation Systems

## Deriving Ratings Through Social Network Structures

(Alshabib et al 2006)

- **Motivation: Sparsity problem in recommendation systems**
- **Using social networks to aggregate ratings in a recommendation system**
- **Compare rating based at the level of product categories instead of products**
- **A user with many ratings should have more weight than a user with fewer ratings.**
- **Recommendation based on the social network built from trust and reputation.**

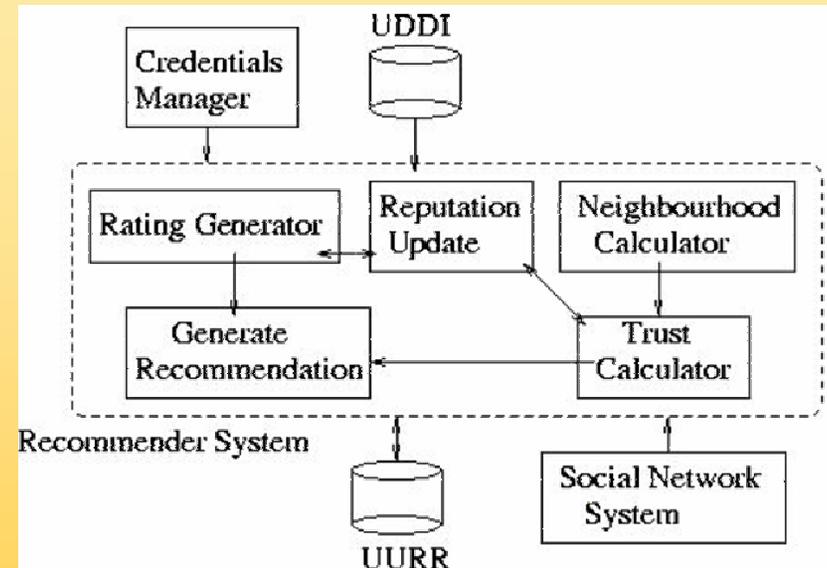


Fig. 4. Internal Architecture of the Recommender Component

# Terrorism and Crime Related Weblog Social Network

- Yang and Ng, 2007

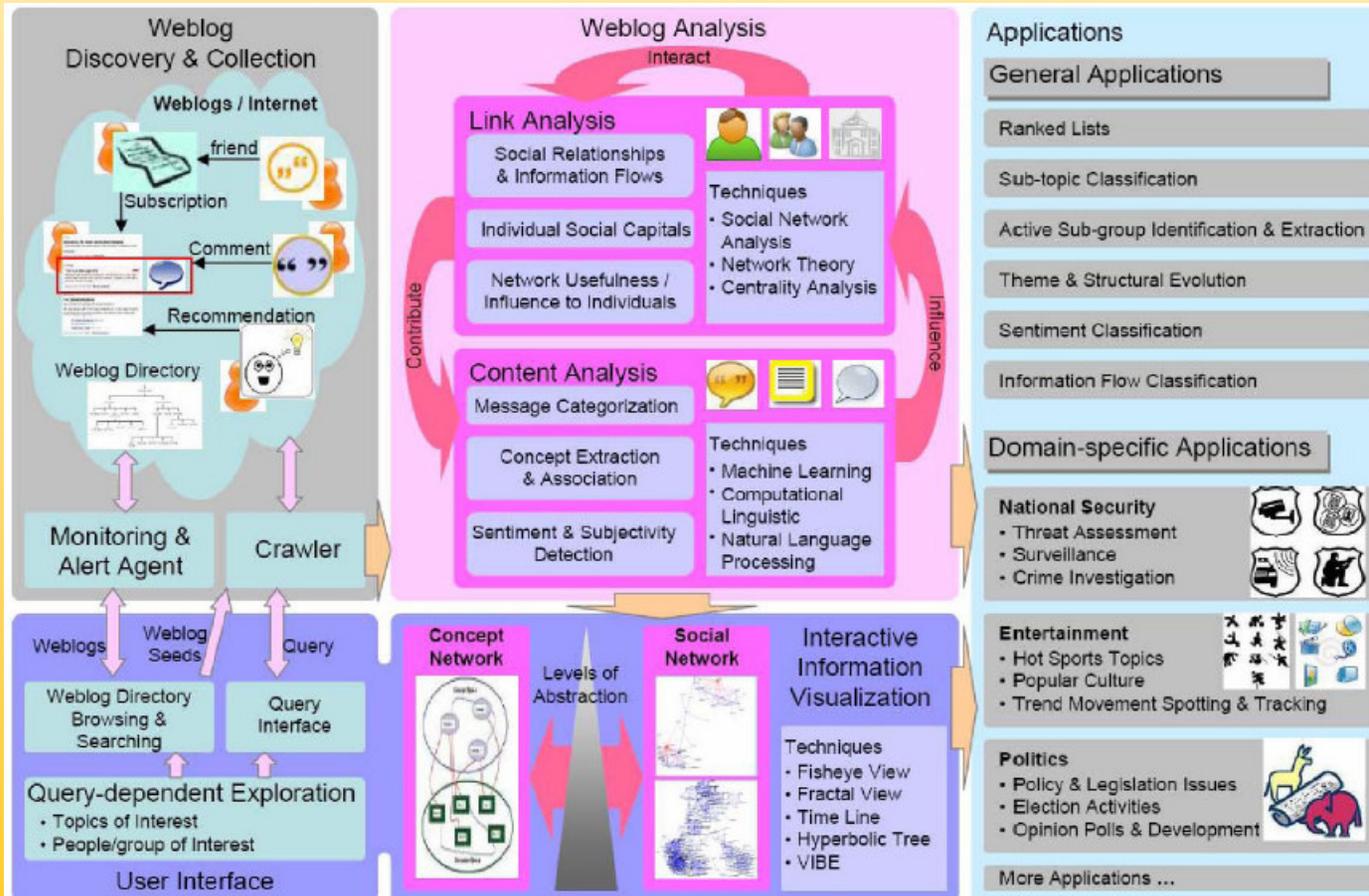
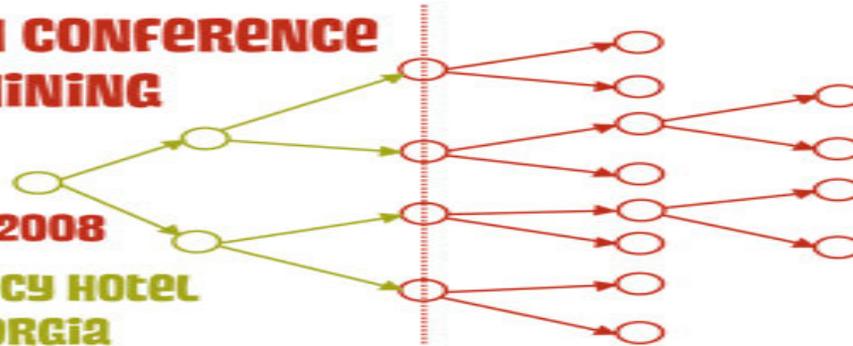


Figure 1. Framework of the proposed Terrorism and Crime Related Weblog Social Network project

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## Emerging Applications in SNA

# Example of E-mail Communication

- A sends an e-mail to B
  - With Cc to C
  - And Bcc to D
- C forwards this e-mail to E
- From analyzing the header, we can infer
  - A and D know that A, B, C and D know about this e-mail
  - B and C know that A, B and C know about this e-mail
  - C also knows that E knows about this e-mail
  - D also knows that B and C do not know that it knows about this e-mail; and that A knows this fact
  - E knows that A, B and C exchanged this e-mail; and that neither A nor B know that it knows about it
  - and so on and so forth ...

# Modeling Pair-wise Communication

- Modeling pair-wise communication between actors
  - Consider the pair of actors  $(A_x, A_y)$
  - Communication *from*  $A_x$  *to*  $A_y$  is modeled using the Bernoulli distribution  $L(x,y)=[p,1-p]$
  - Where,
    - $p = (\# \text{ of emails from } A_x \text{ with } A_y \text{ as recipient}) / (\text{total } \# \text{ of emails exchanged in the network})$
- For  $N$  actors there are  $N(N-1)$  such pairs and therefore  $N(N-1)$  Bernoulli distributions
- Every email is a Bernoulli trial where success for  $L(x,y)$  is realized if  $A_x$  is the sender and  $A_y$  is a recipient

## Modeling an agent's belief about global communication

- Based on its observations, each actor entertains certain beliefs about the communication strength between all actors in the network
- A belief about the communication expressed by  $L(x,y)$  is modeled as the Beta distribution,  $J(x,y)$ , over the parameter of  $L(x,y)$
- Thus, belief is a probability distribution over all possible communication strengths for a given ordered pair of actors  $(A_x, A_y)$

# Measures for Perceptual Closeness

- We analyze the following aspects
  - Closeness between an actor's belief and reality, i.e. “true knowledge” of an actor
  - Closeness between the beliefs of two actors, i.e. the “agreement” between two actors
- We define two measures, *r-closeness* and *a-closeness* for measuring the closeness to reality and closeness in the belief states of two actors respectively

# Perceptual Closeness Measures

- The a-closeness measure is defined as the level of agreement between two given actors  $A_x$  and  $A_y$  with belief states  $B_{x,t}$  and  $B_{y,t}$  respectively, at a given time  $t$  and is given by,

$$a - closeness(B_{x,t}, B_{y,t}) = \frac{1}{1 + div(B_{x,t}, B_{y,t}) + div(B_{y,t}, B_{x,t})} \dots (6)$$

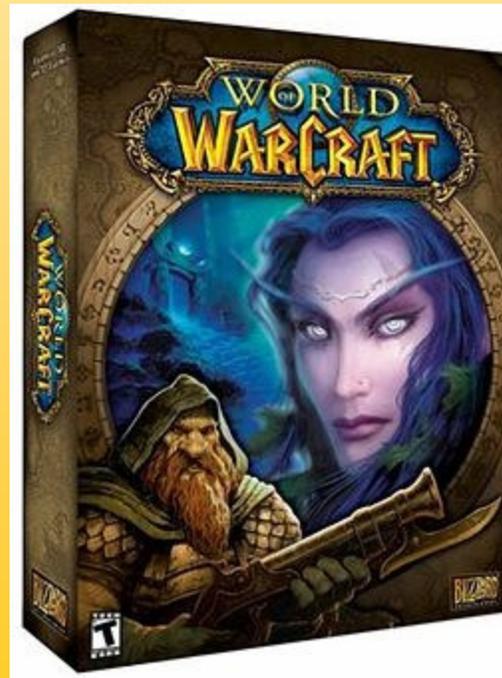
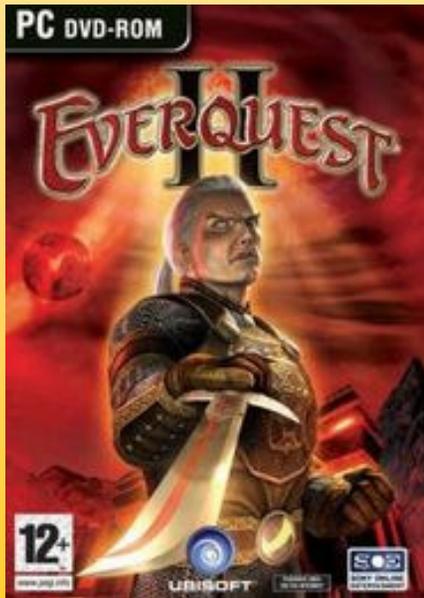
- The r-closeness measure is defined as the closeness of the given actor  $A_k$ 's belief state  $B_{k,t}$  to reality at a given time  $t$  and it is given by,

$$r - closeness(A_k) = \frac{1}{1 + div(B_{s,t}, B_{k,t})} \dots (7)$$

Where  $B_{s,t}$  is the belief state of the super-actor  $A_s$  at time  $t$

# Online Games

- Massively Multiplayer Online Role Playing Games (MMORPG) are computer games that allow hundreds to thousands of players to interact and play together in a persistent online world



Popular MMO Games- Everquest 2, World of Warcraft and Second Life

# MMORPG – Everquest 2

- MMORPGs (MMO Role Playing Games) are the most popular of MMO Games
  - Examples: World of Warcraft by Blizzard and Everquest 2 by Sony Online Entertainment
- Various logs of players' behavior are maintained
- Player activity in the environment as well his/her chat is recorded at regular time instances, each such record carries a time stamp and a location ID
- Some of the logs capture different aspects of player behavior
  - Guild membership history (member of, kicked out of, joined, left)
  - Achievements (Quests completed, experience gained)
  - Items exchanged and sold/bought between players
  - Economy (Items/properties possessed/sold/bought, banking activity, looting, items found/crafted)
  - Faction membership (faction affiliation, record of actions affecting faction affiliation)

# Impact on Social Science

- Interactions in MMO Gaming environments are real
- MMO Games provide sociologists with a unique source of data allowing them to observe real interactions in the context of a complete environment on a very fine granularity
- Gets around the serious issue of unbiased complete data collection
- Analysis of such data presents novel computational challenges
  - The scale of data is much larger than normally encountered in traditional social network analysis
  - The number of environment variables captured is greater
  - Player interaction data is captured at a much finer granularity
- MMORPG data requires models capable of handling large amounts of data as well as accounting for the many environment variables impacting the social structure

# Social Science Research with Everquest 2 Data

- Objective of our research from a social science point of view is to improve understanding of the dynamics of group behavior
- Traditional analysis of dynamics of group behavior works with a *fixed* and *isolated* set of individuals
- MMORPG data enables us to look at dynamics of groups in a new way
  - Multiple groups are part of a large social network
  - Individuals from the social network can join or leave groups
  - Groups are not isolated and some of them can be related i.e. they may be geared towards specific objectives, each of which works towards a larger goal (e.g. different teams working towards disaster recovery)
  - The emergence, destruction as well as dynamic memberships of the groups depend on the underlying social network as well as the environment

# DM Challenges for Social Science Research with Everquest 2 Data

- Inferring player relationships and group memberships from game logs
  - Basic elements of the underlying social network such player-player and layer-group relationships need to be extracted from the game logs
- Developing measures for studying player and group characteristics
  - Novel measures need to be developed that measure individual and group relationships for dynamic groups
  - Novel metrics must also be developed for quantifying relationships between the groups themselves, the groups and the underlying social network as well as the groups and the environment
- Efficient computational models for analyzing group behavior
  - Extend existing group analysis techniques from the social science domain to handle large datasets
  - Develop novel group analysis techniques that account for the dynamic multiple group scenario as well as the data scale

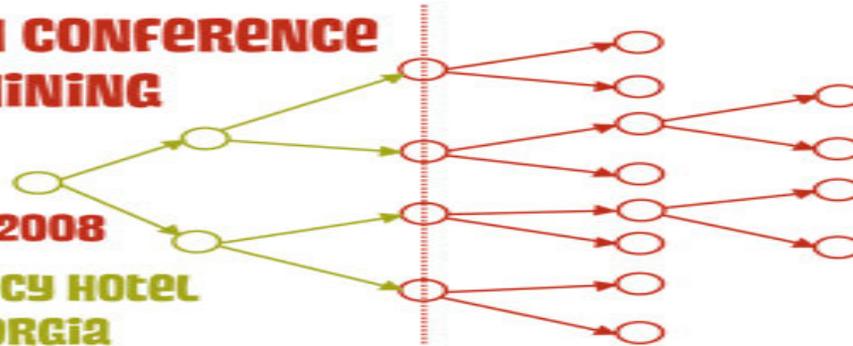
# Conclusion

- Computers have provided the ideal infrastructure for
  - Fostering social interaction
  - Capture it at a very fine granularity
  - Practically no reporting bias
- **→ Fertile research area for data mining research**
- The emerging field of computational social science has the potential to revolutionize social sciences much as
  - Gene Sequencing revolutionized study of genetics
  - The electron microscope revolutionized chemistry

**2008 SIAM CONFERENCE  
ON DATA MINING**

**APRIL 24-26, 2008**

**HYATT REGENCY HOTEL  
ATLANTA, GEORGIA**



## References

# References

1. L. Adamic, R.M.Lukose, A.R.Puniyani and B.A.Huberman. Search in power law networks. Phys. Rev. E 64, 046135(2001).
2. L. Adamic and E. Ader. Friends and Neighbors on the web. Social Netowrks, 25(3), pp 211-230, 2003.
3. Ahmad, Muhammad A., Teredesai, Ankur., Modeling Proliferation of Ideas in Online Social Networks, Proceedings of the 5th Australasian Data Mining Conference, November 29-30 2006, Held in conjunction with the 19th Australian Joint Conference on Artificial Intelligence, Sydney, AUS.
4. Ahmad, Muhammad A., Srivastava, Jaideep An Ant Colony Optimization Approach to Expert Identification in Social Networks First International Workshop on Social Computing, Behavioral Modeling, and Prediction Arizona April, 2008
5. Réka Albert; Albert-László Barabási, Topology of Evolving Networks: Local Events and Universality Physical Review Letters, Volume 85, Issue 24, December 11, 2000, pp.5234-5237
6. Hameeda Alshabib, Omer F. Rana, and Ali Shaikh Ali, "Deriving Ratings Through Social Network Structures," Proceedings First International Conference on Availability, Reliability and Security (ARES'06), 2006.
7. B.W.Bader, R. Harshman and T. G. Kolda. Temporal Analysis of social networks using three way DEDICOM. (Technical Report), SAND2006-2161, Sandia National Laboratories, 2006.
8. Backstrom , L., Dwork, C., Kleinberg, J.: Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, WWW, 2007
9. A L Barabasi, R Albert. Emergence of Scaling in Random networks. Science 15 October 1999:Vol. 286. no. 5439, pp. 509 - 512
10. Indrajit Bhattacharya, Lise Getoor: A Latent Dirichlet Model for Unsupervised Entity Resolution. SDM 2006
11. S.P. Borgatti, and P. Foster., P. 2003. The network paradigm in organizational research: A review and typology. Journal of Management. 29(6): 991-1013
12. U. Brandes. A Faster Algorithm for Betweenness Centrality. Journal of Mathematical Sociology 25(2):163-177, 2001.
13. G.G. Van De Bunt, M.A.J. Van Duijn, T.A.B Snijders Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Organization Theory, Volume 5, Number 2, July 1999, pp. 167-192(26).
14. Tim Carnes, Chandrashekhar Nagarajan, Stefan Wild, and Anke van Zuylen. Maximizing Influence in a Competitive Social Network: A Follower's Perspective
15. Cassi, Lorenzo. Information, Knowledge and Social Networks: Is a New Buzzword coming up? DRUID Academy Winter 2003 PhD Conference.

# References

1. Chiang, Yen-Sheng (2007) 'Birds of Moderately Different Feathers: Bandwagon Dynamics and the Threshold Heterogeneity of Network Neighbors', *The Journal of Mathematical Sociology*, 31:1, 47 – 69
15. A. Clauset, M.E.J.Newman and C.Moore. Finding community structure in very large networks. *Phys. Rev. E* 70, 0066111(2004).
16. N. Contractor. Multi-theoretical multilevel MTML models to study the emergence of networks, NetSci conference, 2006.
17. A. Culotta, R.Bekkerman and A.McCallum. Extracting social networks and contact information from email and the web. Conference on Email and Spam (CEAS), 2004.
18. J. Diesner and K.M. Carley. Exploration of Communication Networks from the Enron Email Corpus. Workshop on Link Analysis, Counterterrorism and Security , In SIAM International Conference on Data Mining, 2005.
19. Li Ding, Tim Finin and Anupam Joshi. Analyzing Social Networks on the Semantic Web. *IEEE Intelligent Systems*, 2005.
20. Pedro Domingos and Matthew Richardson. Mining the Network Value of Customers. *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining* (pp. 57-66), 2001. San Francisco, CA: ACM Press.
21. Pedro Domingos. Mining Social Networks for Viral Marketing (short paper). *IEEE Intelligent Systems*, 20(1), 80-82, 2005.
22. C. Faloutsos, K.S. McCurley and A. Tomkins. Fast discovery of connection subgraphs. *ACM SIGKDD 2004*: 118-127.
23. Fortuna, B., Rodrigues, E. M., and Milic-Frayling, N.: Improving the classification of newsgroup messages through social network analysis, *CIKM '07*, 877-880, 2007
24. L.C. Freeman, Visualizing Social Networks. *Journal of Social Structure*, 2000.
25. Linton C. Freeman. See you in the Funny Papers: Cartoons and Social Networks. *CONNECTIONS*, 23(1): 32-42, 2000.
26. Yupeng Fu, Rongjing Xiang, Yiqun Liu, Min Zhang, Shaoping Ma Finding Experts Using Social Network Analysis 2007 IEEE/WIC/ACM International Conference on Web Intelligence
27. L. Getoor, N. Friedman, D. Koller and B. Taskar. Learning Probabilistic Models of Link Structure. *Journal of Machine Learning Research*, 2002.
28. L. Getoor. Link mining: a new data mining challenge. *SIGKDD Explorations*, 5(1):84–89, 2003.
29. M.Girvan and M.E.J.Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 7821-7826, 2002.

# References

30. J. Golbeck, B. Parsia, J. Hendler. Trust Networks on the Semantic Web. In Proceedings of Cooperative Intelligent Agents 2003, Helsinki, Finland, August 27-29.
31. Golder S A, Wilkinson D and Huberman B A (2007) Rhythms of social interaction: Messaging within a massive online network, 3rd International Conference on Communities and Technologies (CT2007), East Lansing, MI.
32. M Granovetter. Threshold models of collective behaviour. American Journal of Sociology, 83 (6): 1420-1443, 1978.
33. R. Guha, R. Kumar, P. Raghavan and A. Tomkins. Propagation of Trust and Distrust. In Proceedings of 13th International World Wide Web Conference, 2004.
34. M. Harada, S. Sato and K. Kazama. Finding authoritative people from the web. In proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries. pp 306-313, 2004.
35. M.A.Hasan, V.Chaoji, S.Salem and M.Zaki. Link prediction using supervised learning. In Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining, 2006.
36. Heer, J. Exploring Enron: Visual Data Mining. <http://www.cs.berkeley.edu/~jheer/anlp/final/>, 2005.
37. S. Hill. Social network relational vectors for anonymous identity matching. In Proceedings of the IJCAI Workshop on Learning Statistical Models from Relational Data, Acapulco, Mexico, 2003.
38. P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. Journal of the American Statistical Association, 97:1090--1098, 2002.
39. Holzer R, Malin B, and Sweeney L. Email Alias Detection Using Network Analysis. In Proceedings of the ACM SIGKDD Workshop on Link Discovery: Issues, Approaches, and Applications. Chicago, IL. August 2005.
40. T.Hope, T. Nishimura and H.Takeda. An integrated method for social network extraction. In WWW'06: Proceedings of 15th international conference on World Wide Web, pp 845-846, New York, NY, USA, 2006, ACM Press.
41. Hsu, K.-W., Banerjee, A., Srivastava, J.: I/O Scalable Bregman Co-clustering, 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008), Osaka, Japan, May 2008.
42. Jackson, Matthew O., 2003. "A survey of models of network formation: Stability and efficiency," Working Papers 1161, California Institute of Technology, Division of the Humanities and Social Sciences.
43. Jin, Y., Matsuo, Y., Ishizuka, M.: Extracting Inter-Firm Networks from WorldWideWeb, CEC/EEE, 635-642, 2007
44. Marije Kanis, Niall Winters, Stefan Agamanolis, Anna Gavin, and Cian Cullinan, Toward Wearable Social Networking with iBand, CHI 2005 Extended Abstracts on Human Factors in Computing Systems, Portland, Oregon, 2 - 7 April 2005, ACM Press.

# References

45. Kautz, H., Selman, B., and Shah, M. 1997. Referral Web: combining social networks and collaborative filtering. *Commun. ACM* 40, 3 (Mar. 1997), 63-65.
46. David Kempe, Jon M. Kleinberg, Éva Tardos: Maximizing the spread of influence through a social network. *KDD 2003*: 137-146
47. Kermack, W. O. and McKendrick, A. G. "A Contribution to the Mathematical Theory of Epidemics." *Proc. Roy. Soc. Lond. A* 115, 700-721, 1927.
48. Young Ae Kim, Jaideep Srivastava: Impact of social influence in e-commerce decision making. *ICEC 2007*: 293-302
49. Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment" in *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 668-677, January 1998.
50. Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. *Cornell Computer Science Technical Report 99-1776*, October 1999
51. J.Kleinberg. Hubs, Authorities and Communities. *ACM Computing Surveys*, 31(4), December 1999.
52. Jon Kleinberg. Small-World Phenomena and the Dynamics of Information. *Advances in Neural Information Processing Systems (NIPS)* 14, 2001.
53. J. Kleinberg and P. Raghavan. Query Incentive Networks. In *FOCS '05: 46th Annual IEEE Symposium on Foundations of Computer Science*. Pittsburgh, PA, 132--141, 2005.
54. David Krackhardt, 1987. Qap partialling as a test of spuriousness. *Social Networks*, 9, 171-186
55. David Krackhardt and Jeffrey R. Hanson. Informal Networks: The Company Behind the Chart. *Harvard Business Review*, 1993.
56. Krebs, V. E. Mapping networks of terrorist cells. *Connections* 24, 3 (2001), 43–52.
57. Lam, C. 2004. SNACK: incorporating social network information in automated collaborative filtering. In *Proceedings of the 5th ACM Conference on Electronic Commerce (New York, NY, USA, May 17 - 20, 2004)*. EC '04.
58. H.Lauw, E-P.Lim, T.T.Tan and H.H. Pang. Mining social network from spatio-temporal events. In *Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining*, 2005.
59. Jure Leskovec, Lada Adamic, and Bernardo Huberman. The Dynamics of Viral Marketing. *EC'06*.
60. J.Leskovic and C.Faloutsos. Sampling from large graphs. *ACM SIGKDD 2006*: 631-636.
61. Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S. Glance, Matthew Hurst: Patterns of Cascading Behavior in Large Blog Graphs. *SDM 2007*
62. Juan-Zi Li, Jie Tang, Jing Zhang, Qiong Luo, Yunhao Liu, MingCai Hong. EOS: expertise oriented search using social networks. In *Proceedings of WWW'2007*. 1271-1272

# References

61. D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In CIKM'03: Proceedings of the 14th ACM international conference on Information and Knowledge Management, pp 556-559. ACM Press, 2003.
62. Hugo Liu and Pattie Maes. InterestMap: Harvesting Social Network Profiles for Recommendations. Workshop: Beyond Personalization, IUI'05, 2005.
63. Lonier, T. & Matthews, C., 2004. "Measuring the Impact of Social Networks on Entrepreneurial Success: The Master Mind Principle." Presented at the 2004 Babson Kauffman Entrepreneurship Research Conference, Glasgow, Scotland, June.
64. M. Makrehchi and M.S. Kamel. Building social networks from web documents: A text mining approach. In the 2nd LORNET Scientific Conference, 2005.
65. Makrehchi, M. and Kamel, M. S.: Learning Social Networks Using Multiple Resampling Method, IEEE International Conference on Systems, Man and Cybernetics, 2007
66. B. Malin. Unsupervised name disambiguation via social network similarity. In Proc. SIAM Wksp on Link Analysis, Counterterrorism, and Security, pages 93–102, Newport Beach, CA, 2005.
67. Murata, Tsuyoshi Moriyasu, Sakiko Link Prediction of Social Networks Based on Weighted Proximity Measures IEEE/WIC/ACM International Conference on Web Intelligence: 85-88: 2-5 Nov. 2007
68. P. Marsden 1990. "Network Data and Measurement." Annual Review of Sociology, Volume 16 (1990), pp. 435-463.
69. Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida and M. Ishizuka. Polyphnet: an advanced social network extraction system from the web. In WWW'06: Proceedings of 15th international conference on World Wide Web, pp 397-406, New York, NY, USA, 2006, ACM Press.
70. Matsuo, Y., Tomobe, H., and Nishimura, T.: Robust Estimation of Google Counts for Social Network Extraction. AAAI, 1395-1401, 2007
71. McDonald, D. W. 2003. Recommending collaboration with social networks: a comparative evaluation. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Ft. Lauderdale, Florida, USA, April 05 - 10, 2003). CHI '03.
72. Peter Mika. Flink: Using Semantic Web Technology for the Presentation and Analysis of Online Social Networks. Journal of Web Semantics 3(2), Elsevier, 2005.
73. S. Milgram, "The small world problem," Psychology Today 1, 61 (1967).
74. Martina Morris. Network Epidemiology: A Handbook for Survey Design and Data Collection (2004). London: Oxford University Press.

# References

75. N. Matsumura, D.E.Goldberg and X.Llor. Mining directed social network from message board. In WWW'05: Proceedings of 14th international conference on World Wide Web, pp 1092-1093, New York, NY, USA, 2005, ACM Press.
76. Nagaraja, S.: Anonymity in the Wild: Mixes on unstructured networks, Privacy Enhancing Technologies, 2007
77. M. E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. Phys.Rev. E64 (2001) 016131.
78. M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45(2):167-- 256, 2003.
79. M.E.J.Newman. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133, 2004.
80. M.E.J.Newman. Modularity and community structure in networks. Proc. of Natl Acad. of Sci. USA 103, 8577-8582, 2006.
81. J. Palau, M. Montaner, and B. Lopez. Collaboration analysis in recommender systems using social networks. In Eighth Intl. Workshop on Cooperative Info. Agents (CIA'04), 2004.
82. Palla, G., Barabási, A.-L., and Vicsek, T.: Quantifying social group evolution, Nature 446, 664-667, 2007
83. Paolillo, J.C. and Wright, E. Social Network Analysis on the Semantic Web: Techniques and challenges for Visualizing FOAF. Visualizing the Semantic Web (Draft Chapter), in press.,
84. Jialun Qin, Jennifer J. Xu, Daning Hu, Marc Sageman and Hsinchun Chen: Analyzing Terrorist Networks: A Case Study of the Global Salafi Jihad Network. Intelligence and Security Informatics, 2005.
85. Raghavan , U., Albert, R., and Kumara, S.: Near linear time algorithm to detect community structures in large scale networks, Phys. Rev. E 76, 036106, 2007
86. Resig, J., Teredesai, A., Dawara, S., & Homan, C., (2004) Extracting Social Networks from Instant Messaging Populations. KDD 2004 Link Discovery Workshop.
87. Matthew Richardson and Pedro Domingos. 2002. Mining knowledge-sharing sites for viral marketing. In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM Press, New York, NY, 61-70.
88. Kazumi Saito, Ryohei Nakano, Masahiro Kimura: Prediction of Link Attachments by Estimating Probabilities of Information Propagation. KES (3) 2007: 235-242
89. P. Sarkar and A. Moore. Dynamic social network analysis using latent space models. ACM SIGKDD Explorations Newsletter, 7(2), pp 31-40, 2005.
90. Schelling, T (1978). Micromotives and macrobehavior. New York: W. W. Norton.

# References

91. Shetty, J., & Adibi, J. The Enron email dataset database schema and brief statistical report (Technical Report). Information Sciences Institute, 2004.
92. J.Shetty and J.Adibi. Discovering important nodes through graph entropy the case of Enron email database. In Proceedings of 3rd international Workshop on Link Discovery in ACM SIGKDD'05, pp 74-81, 2005.
93. Smith, R.D. (2002). Instant Messaging as a Scale-Free Network. cond-mat/0206378.
94. Snijders, T.A.B. (2005). Models for Longitudinal Network Data. Chapter 11 in P. Carrington, J. Scott, & S. Wasserman (Eds.), Models and methods in social network analysis. New York: Cambridge University Press
95. Xiaodan Song, Belle L. Tseng, Ching-Yung Lin and Ming-Ting Sun, "ExpertiseNet: Relational and Evolutionary Expert Modeling", Intl. Conf. on User Modeling, Edinburgh, UK, July 2005.
96. E. Spertus, Mehran Sahami, Orkut Buyukkokten: Evaluating similarity measures: a large-scale study in the orkut social network. KDD 2005: 678-684
97. M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths. Probabilistic author-topic models for information discovery. In proceedings of 10th ACM SIGKDD, pp 306-315, Seattle, WA, USA, 2004.
98. Teredesai, A., Resig, J., (2004) A Framework for Mining Instant Messaging Services. SIAM DM 2004 Workshop on Link Analysis, Counter-Terrorism & Privacy.
99. J. Travers and S. Milgram, ``An experimental study of the small world problem,`` Sociometry 32, 425 (1969).
100. J.Tyler, D. Wilkinson and B. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. Communities and technologies, pp 81-69, 2003.
101. S. Wasserman, & K. Faust, Social Network Analysis: Methods and Applications. New York and Cambridge, ENG: Cambridge University Press (1994).
102. D. J. Watts, S. H. Strogatz, "Collective Dynamics of Small-World Networks." Nature 393, 440-442, 1998.
103. D. Watts, Network dynamics and the small world phenomenon. Americal Journal of Sociology 105 (2), 493-527. 1999
104. D. J. **Watts**, P. S. **Dodds**, and M. E. J. **Newman**. Identity and search in social networks. Science, 296, 1302-1305 (2002).
105. D. Watts. Small Worlds: The Dynamics of Networks between Order and Randomness, 2003.
106. Duncan J. Watts and Peter Sheridan Dodds, "Influentials, Networks, and Public Opinion Formation." Journal of Consumer Research: December 2007.
107. A.Y.Wu, M.Garland and J.Han. Mining scale free networks using geodesic clustering. ACM SIGKDD'04, pp 719-724, New York, NY, USA, 2004, ACM Press.
108. Xu, J. and Chen, H. 2005. Criminal network analysis and visualization. Commun. ACM 48, 6 (Jun. 2005), 100-107.

# References

107. Wan-Shiou Yang; Jia-Ben Dia; Hung-Chi Cheng; Hsing-Tzu Lin, "Mining Social Networks for Targeted Advertising," System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on , vol.6, no.pp. 137a- 137a, 04-07 Jan. 2006.
108. H Peyton Young, 2000. "The Diffusion of Innovations in Social Networks," Economics Working Paper Archive 437, The Johns Hopkins University, Department of Economics.
109. B. Yu and M. P. Singh. Searching social networks. In Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS). ACM Press, July 2003.
110. Jing Zhang, Jie Tang, Juan-Zi Li. Expert Finding in a Social Network. In Proceedings of DASFAA'2007. pp.1066~1069
111. Bin Zhao, Prithviraj Sen and Lise Getoor Entity and relationship labeling in affiliation networks, ICML workshop on Statistical Network Analysis, 2006.
112. Zhang, J. and Van Alstyne, M. 2004. SWIM: fostering social network based information search. In CHI '04 Extended Abstracts on Human Factors in Computing Systems (Vienna, Austria, April 24 - 29, 2004). CHI '04. ACM Press, New York, NY, 1568-1568.
113. Zheng, R., F. Provost and A. Ghose. Social Network Collaborative Filtering: Preliminary Results . Proceedings of the Sixth Workshop on eBusiness (WEB2007), December 2007.
114. D. Zhou, E.Manavoglu, J.Li, C.L.Giles and H.Zha. Probabilistic models for discovering e-communities. In WWW'06: Proceedings of 15th international conference on World Wide Web, pp 173-182, New York, NY, USA, 2006, ACM Press.