E6885 Network Science Lecture 6:
Network Topology Inference

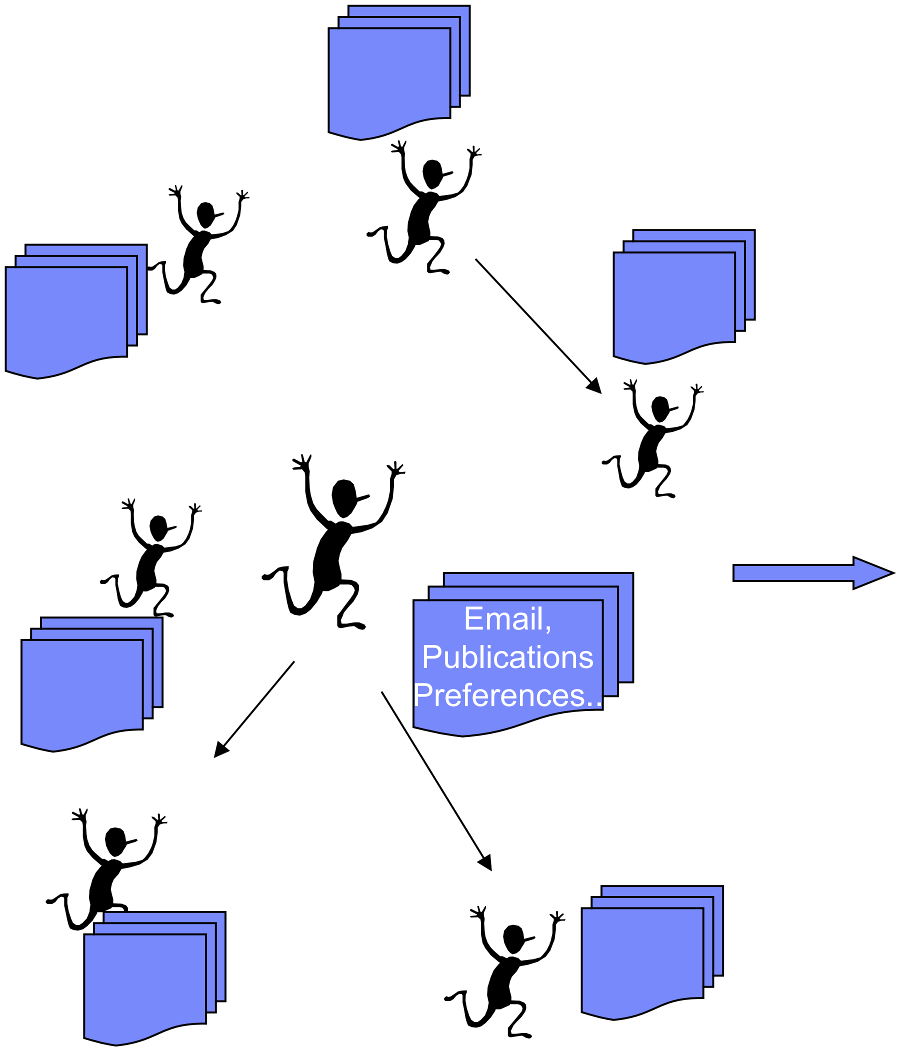
Ching-Yung Lin, Dept. of Electrical Engineering, Columbia University

October 15th, 2012

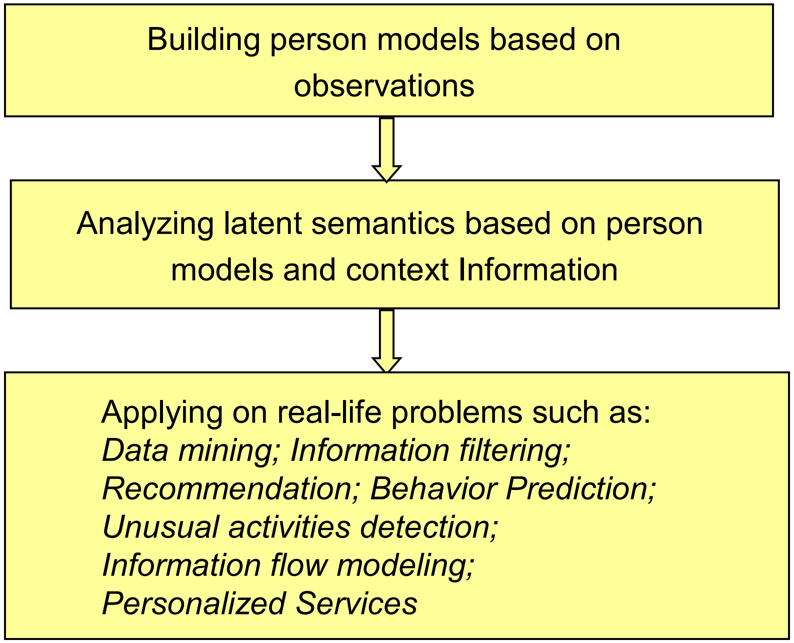
Course Structure

Class Date	Class Number	Topics Covered
09/10/12	1	Overview – Social, Information, and Cognitive Network Analysis
09/17/12	2	Network Representations and Characteristics
09/24/12	3	Network Partitioning, Clustering, and Visualization
10/01/12	4	Network Estimation and Modeling
10/09/12	5	Network Analysis Use Case
10/15/12	6	Network Topology Inference
10/22/12	7	Dynamic Networks
10/29/12	8	Info Diffusion in Networks
11/12/12	9	**Final Project Proposal Presentation**
11/19/12	10	Graphical Models
11/26/12	11	Privacy, Security, and Economy Issues in Networks
12/03/12	12	Behavior Understanding and Cognitive Networks
12/10/12	13	Large-Scale Network Processing System
12/17/12	14	**Final Project Presentation**

Motivation – Latent Person Behavior Modeling, Analysis and Applications



What are a person's
- roles in events?
- behavior evolutions?
-interests, tastes?
- ….

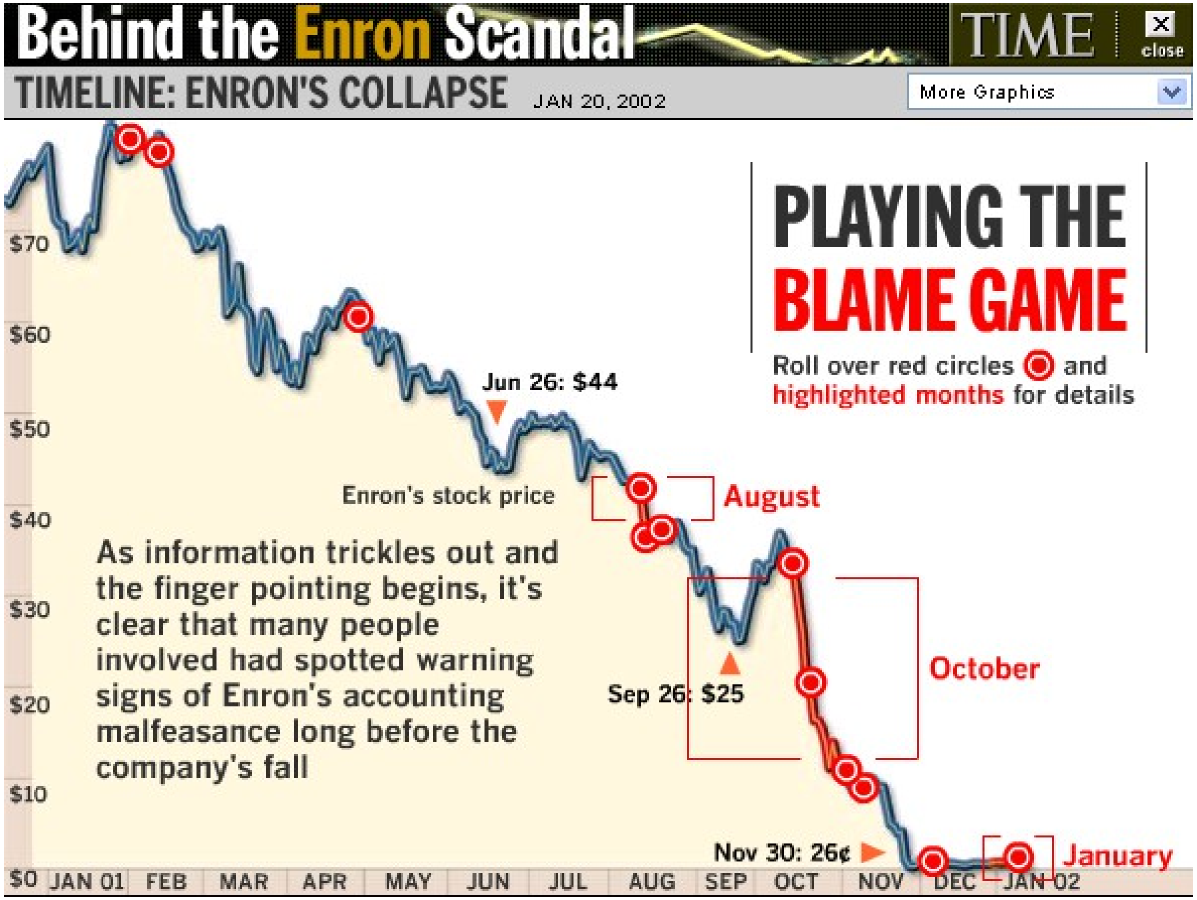


Building person models based on observations

Analyzing latent semantics based on person models and context Information

Applying on real-life problems such as:
Data mining; Information filtering;
Recommendation; Behavior Prediction;
Unusual activities detection;
Information flow modeling;
Personalized Services

Enron Corpus



Behind the Enron Scandal

TIMELINE: ENRON'S COLLAPSE JAN 20, 2002

PLAYING THE BLAME GAME

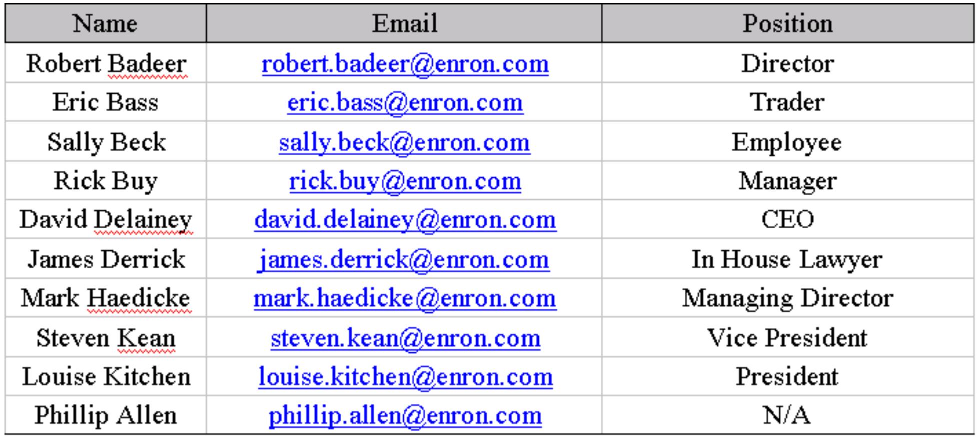
Roll over red circles ⊙ and highlighted months for details

As information trickles out and the finger pointing begins, it's clear that many people involved had spotted warning signs of Enron's accounting malfeasance long before the company's fall

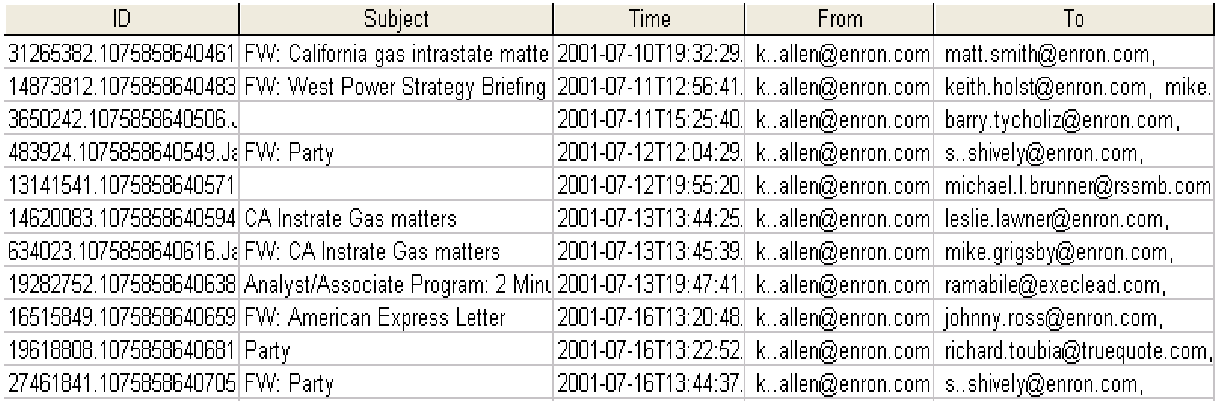
Enron Corpus

- Preprocessing
  - Original messages – 517,431
  - Remove empty messages – 493,391 remain
    - 1999 – 11196
    - 2000 – 196157
    - 2001 – 272875
    - 2002 – 35922
  - Remove repeated messages – 166,653 remain
  - Only keep intra-communications among 149 users within Enron – 25,428 remain
  - Number of terms: 84649
  - Number of users: 149

Collected information for each person



Name	Email	Position
Robert Badeer	robert.badeer@enron.com	Director
Eric Bass	eric.bass@enron.com	Trader
Sally Beck	sally.beck@enron.com	Employee
Rick Buy	rick.buy@enron.com	Manager
David Delainey	david.delainey@enron.com	CEO
James Derrick	james.derrick@enron.com	In House Lawyer
Mark Haedicke	mark.haedicke@enron.com	Managing Director
Steven Kean	steven.kean@enron.com	Vice President
Louise Kitchen	louise.kitchen@enron.com	President
Phillip Allen	phillip.allen@enron.com	N/A



ID	Subject	Time	From	To
31265382.1075858640461	FW: California gas intrastate matte	2001-07-10T19:32:29.	k..allen@enron.com	matt.smith@enron.com,
14873812.1075858640483	FW: West Power Strategy Briefing	2001-07-11T12:56:41.	k..allen@enron.com	keith.holst@enron.com, mike.
3650242.1075858640506.J		2001-07-11T15:25:40.	k..allen@enron.com	barry.tycholiz@enron.com,
483924.1075858640549.Ja	FW: Party	2001-07-12T12:04:29.	k..allen@enron.com	s..shively@enron.com,
13141541.1075858640571		2001-07-12T19:55:20.	k..allen@enron.com	michael.l.brunner@rssmb.com
14620083.1075858640594	CA Instrate Gas matters	2001-07-13T13:44:25.	k..allen@enron.com	leslie.lawner@enron.com,
634023.1075858640616.Ja	FW: CA Instrate Gas matters	2001-07-13T13:45:39.	k..allen@enron.com	mike.grigsby@enron.com,
19282752.1075858640638	Analyst/Associate Program: 2 Minu	2001-07-13T19:47:41.	k..allen@enron.com	ramabile@execlead.com,
16515849.1075858640659	FW: American Express Letter	2001-07-16T13:20:48.	k..allen@enron.com	johnny.ross@enron.com,
19618808.1075858640681	Party	2001-07-16T13:22:52.	k..allen@enron.com	richard.toubia@truequote.com,
27461841.1075858640705	FW: Party	2001-07-16T13:44:37.	k..allen@enron.com	s..shively@enron.com,

What happened?– collect the ground truth

- Summarize important events from different timelines
- The events with most occurrences from multiple media's timelines
  - 14 August 2001 -- Jeffrey Skilling resigns after just six months; Mr Lay returns to day-to-day management of the company.
  - 20 August 2001 -- Mr Lay exercises Enron share options worth $519,000.
  - 12 October 2001 -- Accounting firm Andersen begins destroying documents relating to the Enron audits.The destruction continues until November when the company receives a subpoena from the Securities and Exchange Commission.
  - 16 October 2001-- Enron reports losses of $638m run up between July and September and announces a $1.2 billion reduction in shareholder equity. The reduction in company value relates to partnerships set up and run by chief financial officer Andrew Fastow.
  - …

Dynamic Network Analysis

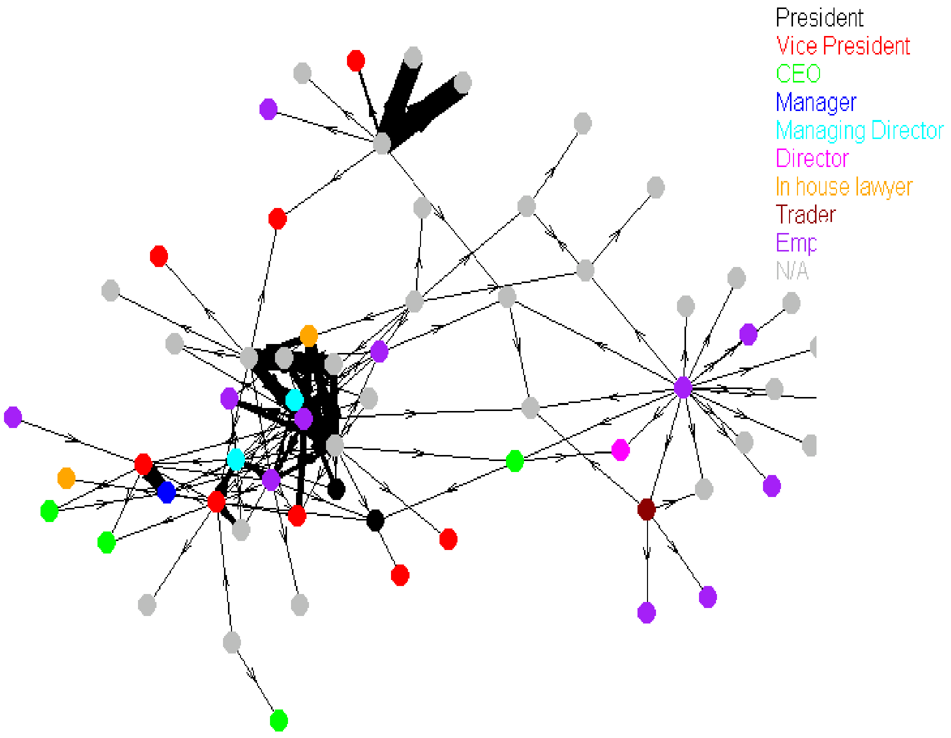
- Dynamical social networks
  - In social network analysis: Dynamic actor-oriented social network [Snijder 2002]
    - Changes in the network are modeled as the stochastic result of network effects (density, reciprocity, etc.)
    - Network evolution is modeled by continuous time Markov chain models
  - In information mining area:
    - link prediction problem - Infer which new interactions among its members are likely to occur in the near future [Liben-Nowell 2003]
    - Track changes in large-scale data by periodically creating an agglomerative clustering and examining the evolution of clusters over time [Kubica *et al. 2002*]

Dynamic Network Analysis

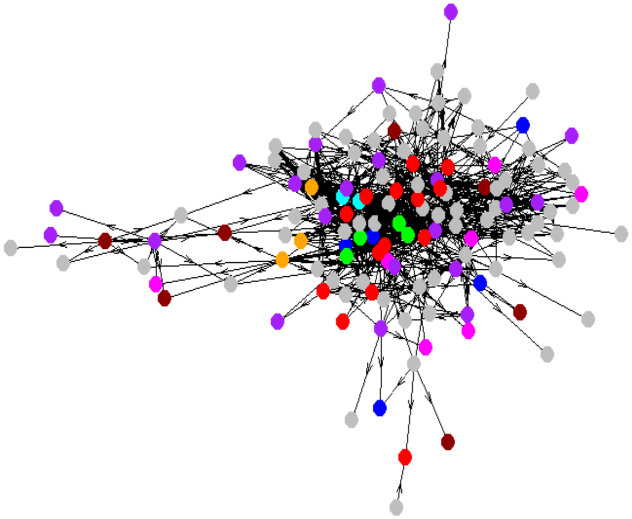
- Dynamical social networks
  - In social network analysis: Dynamic actor-oriented social network [Snijder 2002]
    - Changes in the network are modeled as the stochastic result of network effects (density, reciprocity, etc.)
    - Network evolution is modeled by continuous time Markov chain models
  - In information mining area:
    - link prediction problem - Infer which new interactions among its members are likely to occur in the near future [Liben-Nowell 2003]
    - Track changes in large-scale data by periodically creating an agglomerative clustering and examining the evolution of clusters over time [Kubica *et al. 2002*]

Dynamic social networks

Email contacts within Enron: 1999



Email contacts within Enron: 2000



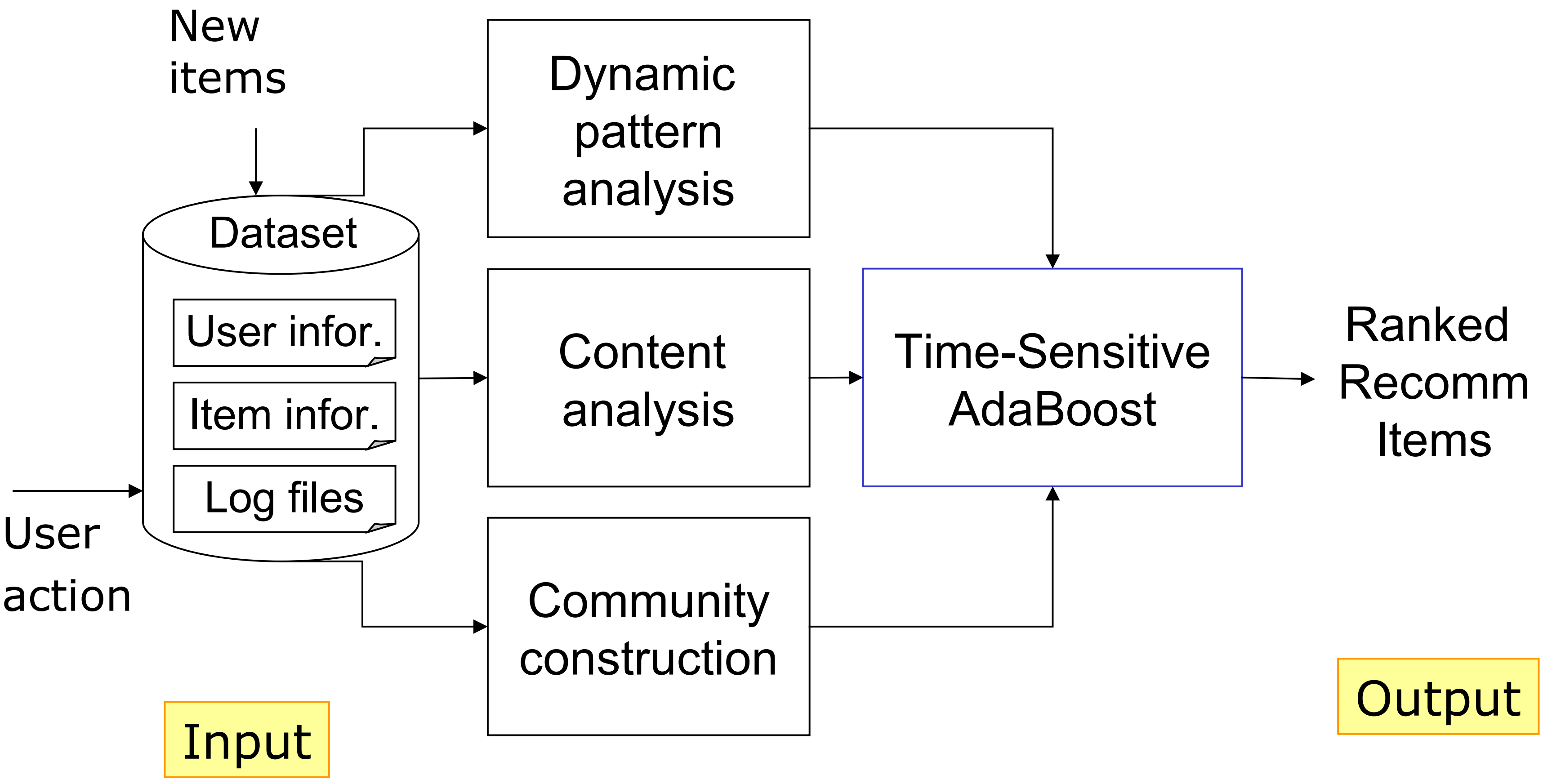
People with top 10 centralities in Enron

Centrality	1999		2000		2001		2002	
	Name	Position	Name	Position	Name	Position	Name	Position
1	Mark_Taylor	Employee	David_Delainey	CEO	Steven_Kean	Vice_President	Kevin_Presto	Vice_President
2	Tana_Jones	N/A	Steven_Kean	Vice_President	John_Lavorato	CEO	Louise_Kitchen	President
3	Sara_Shackleton	N/A	John_Lavorato	CEO	Jeff_Dasovich	Employee	John_Lavorato	CEO
4	Richard_Sanders	Vice_President	Vince_Kaminski	Manager	Vince_Kaminski	Manager	Hunter_Shively	Vice_President
5	Elizabeth_Sager	Employee	Jeff_Skilling	CEO	Louise_Kitchen	President	James_Steffes	Vice_President
6	Mark_Haedicke	Managing_Director	Mike_McConnell	N/A	David_Delainey	CEO	Greg_Whalley	President
7	John_Hodge	Managing_Director	Greg_Whalley	President	Greg_Whalley	President	Fletcher_Sturm	Vice_President
8	Steven_Kean	Vice_President	Sally_Beck	Employee	Mark_Haedicke	Managing_Director	Doug_Gilbert-Smith	N/A
9	Dan_Hyvl	Employee	Jeffrey_A_Shankman	N/A	Phillip_Allen	N/A	Dana_Davis	N/A
10	Carol_Clair	In_house_lawyer	John_Arnold	Vice_President	Mary_Hain	In_house_lawyer	Mark_Haedicke	Managing_Director

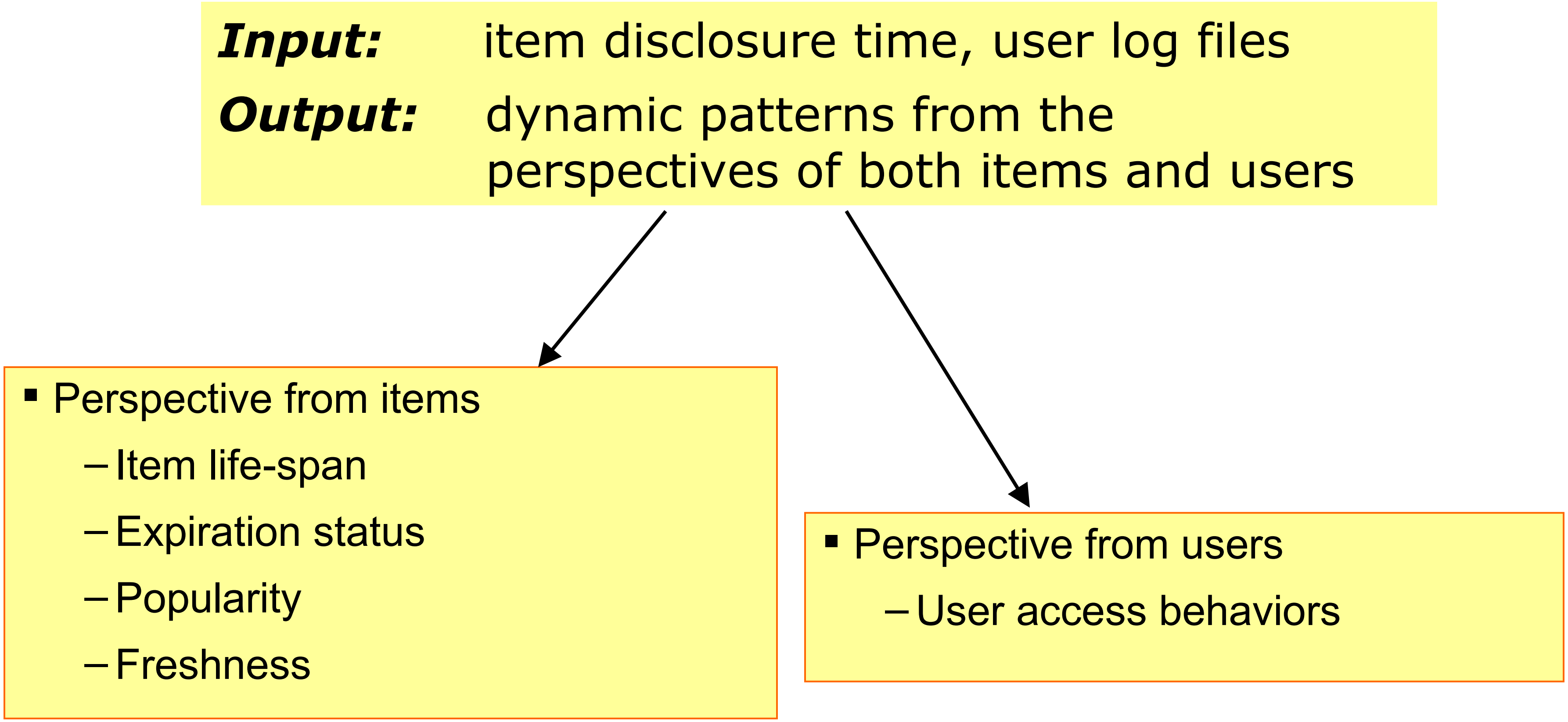
People with top-10 prestige in Enron

Prestige	1999		2000		2001		2002	
	Name	Position	Name	Position	Name	Position	Name	Position
1	John_Hodge	Managing_Director	John_Lavorato	CEO	John_Lavorato	CEO	Darron_Giron	N/A
2	Steven_Kean	Vice_President	Greg_Whalley	President	Louise_Kitchen	President	Phillip_Love	N/A
3	Vince_Kaminski	Manager	David_Delainey	CEO	Phillip_Allen	N/A	Kam_Keiser	Employee
4	Mark_Haedicke	Managing_Director	Steven_Kean	Vice_President	Greg_Whalley	President	Errol_McLaughlin	N/A
5	Elizabeth_Sager	Employee	Vince_Kaminski	Manager	Kevin_Presto	Vice_President	Stacey_White	N/A
6	Richard_Sanders	Vice_President	Rick_Buy	Manager	Barry_Tycholiz	Vice_President	Fletcher_Sturm	Vice_President
7	Kevin_Presto	Vice_President	Kevin_Presto	Vice_President	Steven_Kean	Vice_President	NA	NA
8	Mark_Taylor	Employee	Jeffrey_A_Shankman	N/A	Mike_Grigsby	N/A	NA	NA
9	Michelle_Cash	N/A	Phillip_Allen	N/A	David_Delainey	CEO	NA	NA
10	Stacy_Dickson	Employee	Jeff_Skilling	CEO	Hunter_Shively	Vice_President	NA	NA

Dynamic Network Analysis needs both Pattern and Content Analysis



Dynamic Pattern Analysis



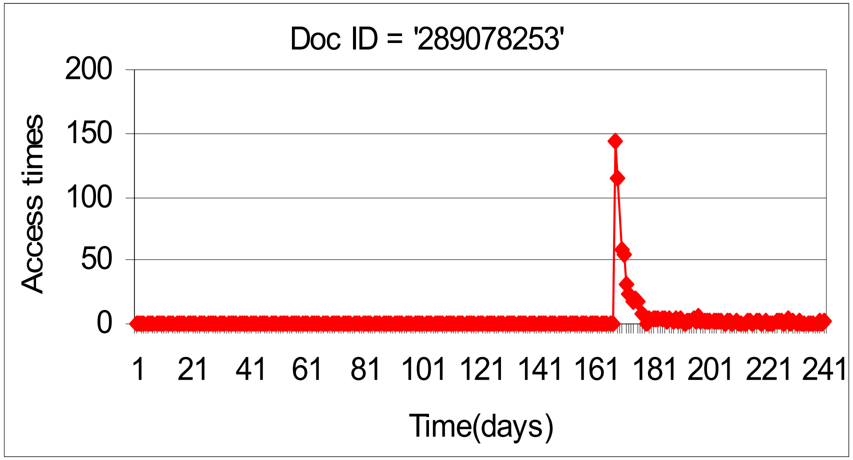
Input: item disclosure time, user log files

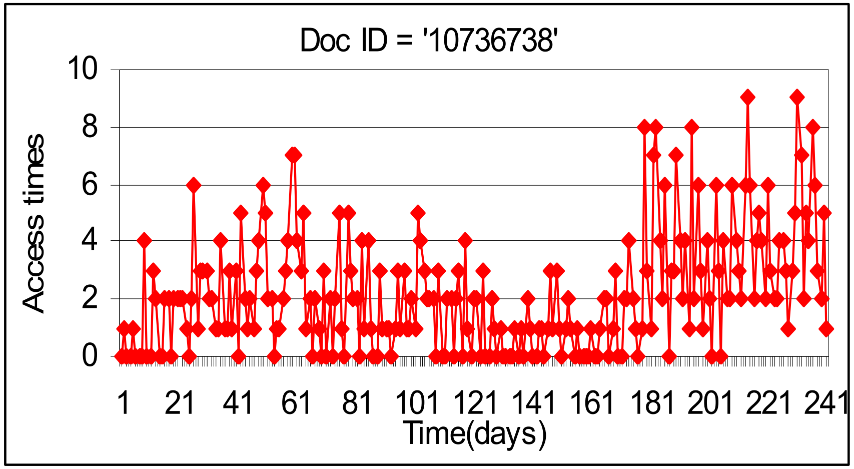
Output: dynamic patterns from the perspectives of both items and users

- Perspective from items
  - Item life-span
  - Expiration status
  - Popularity
  - Freshness

- Perspective from users
  - User access behaviors

Item Access Types





Peak on one day followed by a decay

announcements, conference notes, and executive summaries

Short-term items

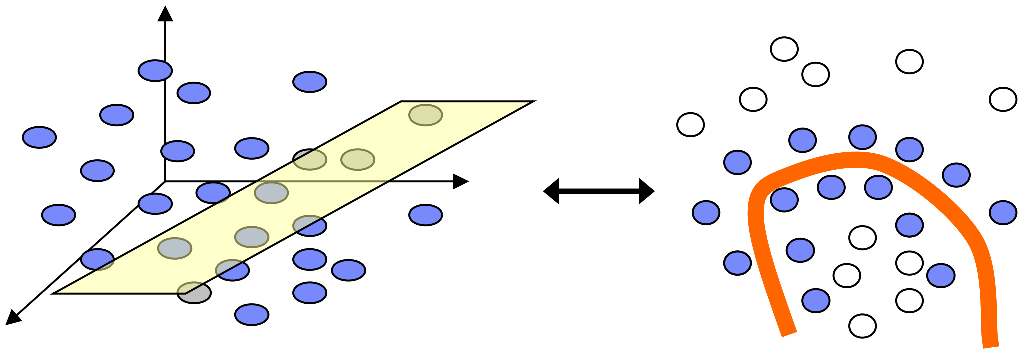
Sustained interest

handouts, specifications, and reference documents

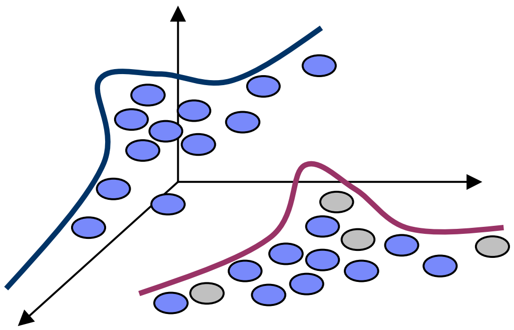
Long-term items

Statistical Modeling Approaches

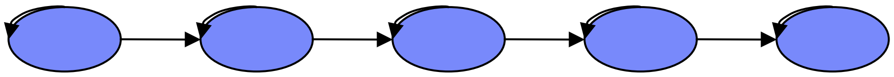
SVM



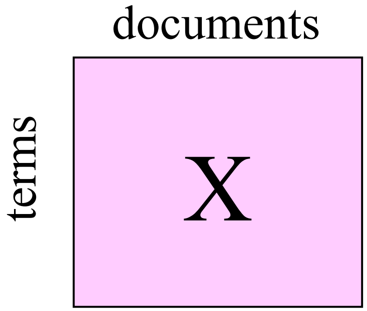
GMM



HMM



LSA



documents

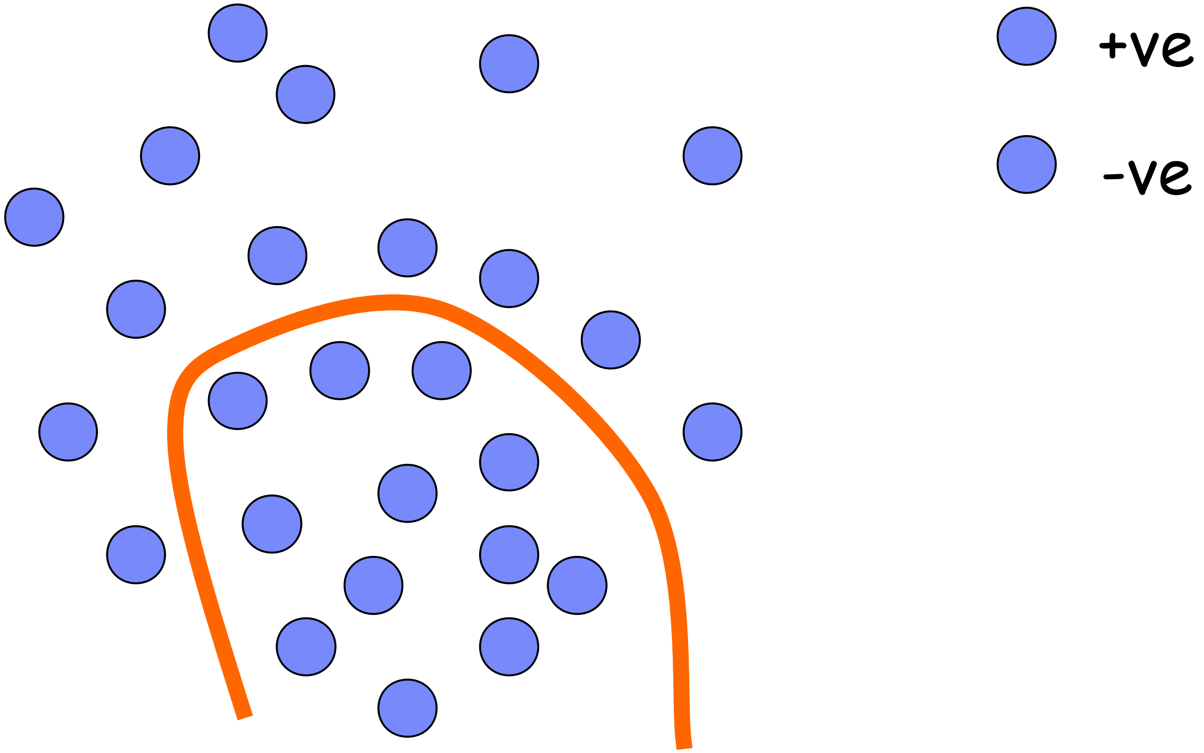
terms

X

Discriminant Modeling

Support Vector Machines

- **Objective**: How to separate observations of distinct classes
- **Solution**: Project to higher dimensionality to determine linear separability

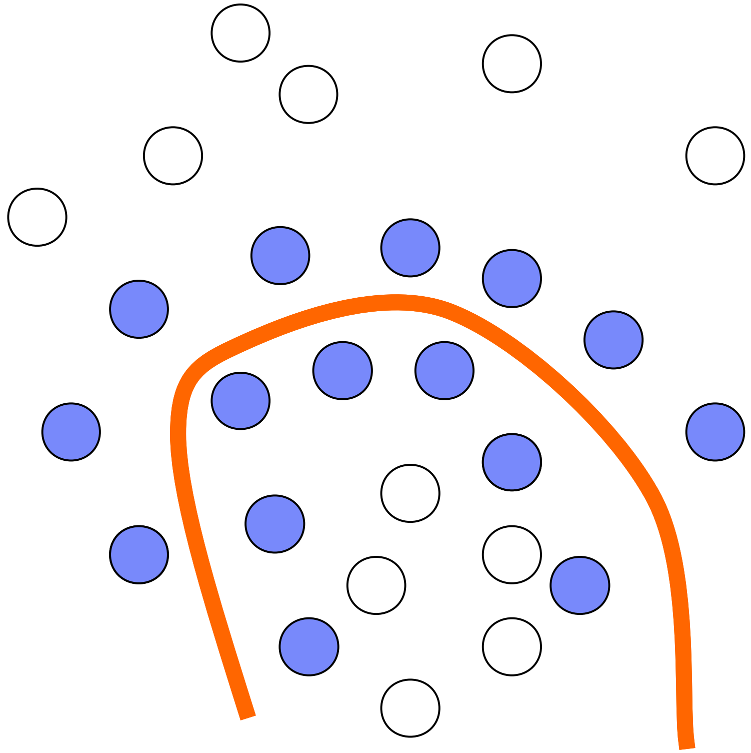


Slide Source: M. Naphade

Discriminant Modeling

Support Vector Machines

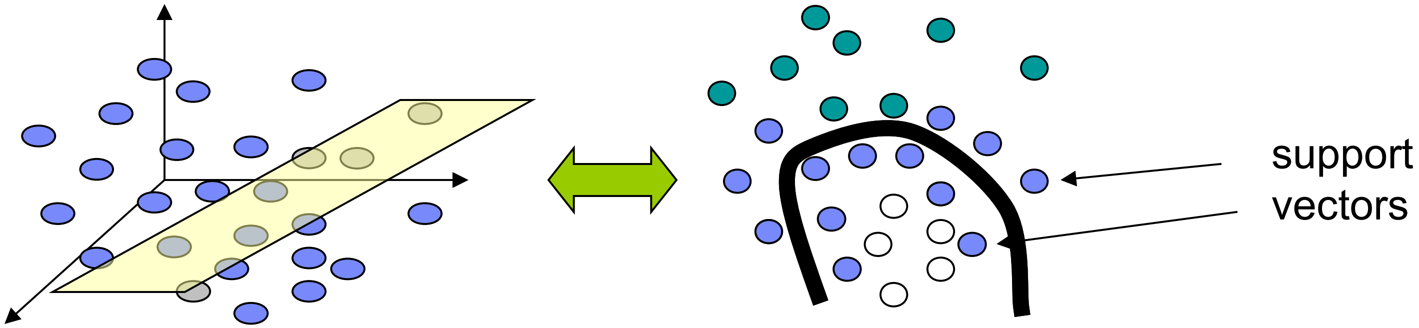
- **Objective**: How to separate observations of distinct classes
- **Solution**: Project to higher dimensionality to determine linear separability



- ● + support vectors
- ● - support vectors
- ○ useless examples

Slide Source: M. Naphade

SVM Classifiers



- Support Vector Machine
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space

- SVM Classifier:

$$f(x) = \sum_{i=1}^{S} a_i \cdot k(x, x_i) + b$$

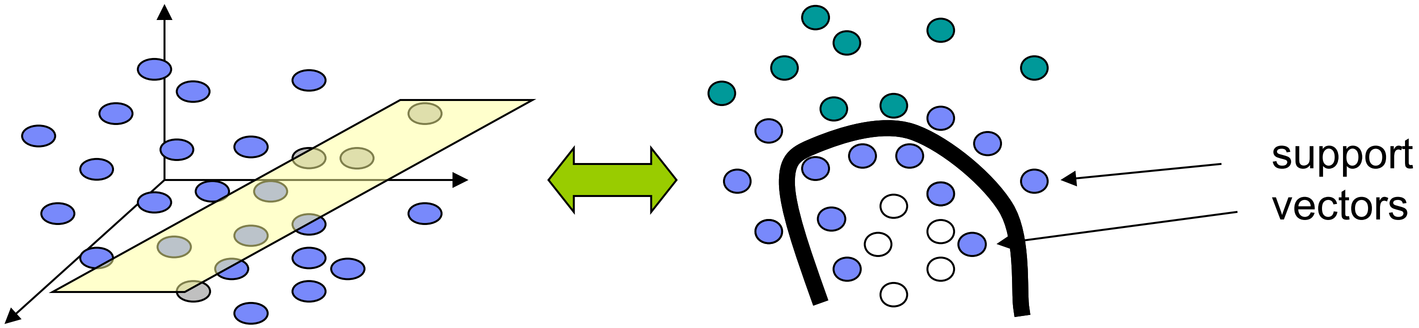
where *S* is the number of support vectors, k(.,.) is a kernel function. E.g., $k(x, x_i) = e^{-\frac{x - x}{r}}$

- Complexity *c*: operation (multiplication, addition) required for classification

$$c \propto S \cdot D$$

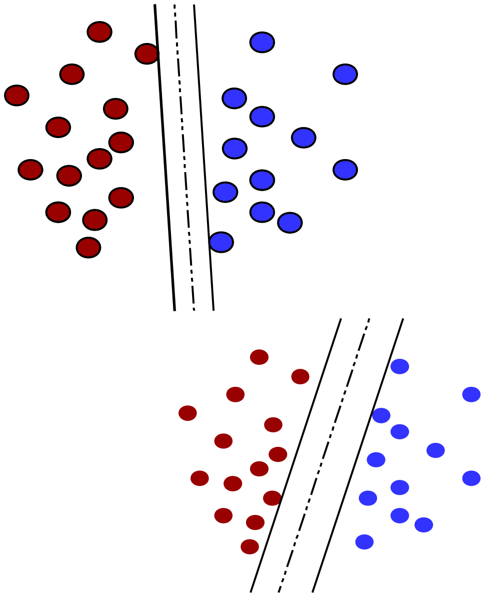
where *D* is the dimensionality of the feature vector

Support Vector Machine



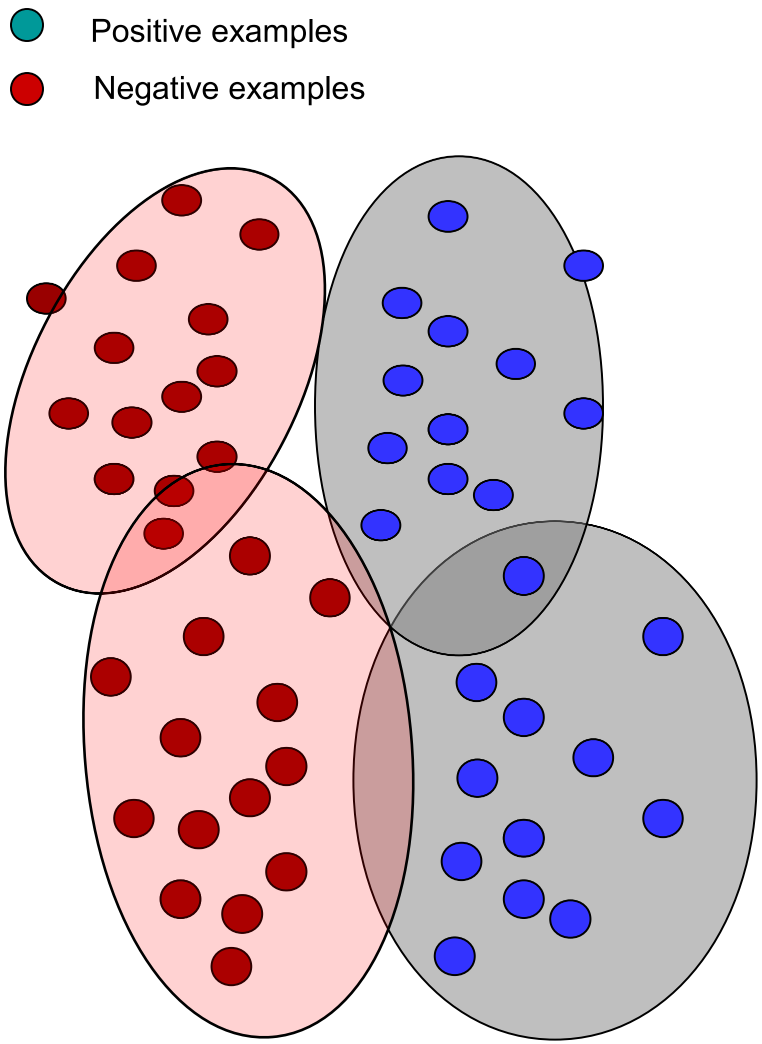
$$f(x) = \sum_{i=0}^{N_s} \alpha_i y_i K(x, s_i) + b$$

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}$$



- Support Vector Machine
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space
  - We use Radial Basis Functions as our kernels
  - Sequential Minimal Optimization = currently, fastest known algorithm for finding hyperplane

Gaussian Mixture Models



- The probability of examples belonging a cluster (e.g., positive set) is modeled as a weighted summation of multiple Gaussians distributed in the feature space

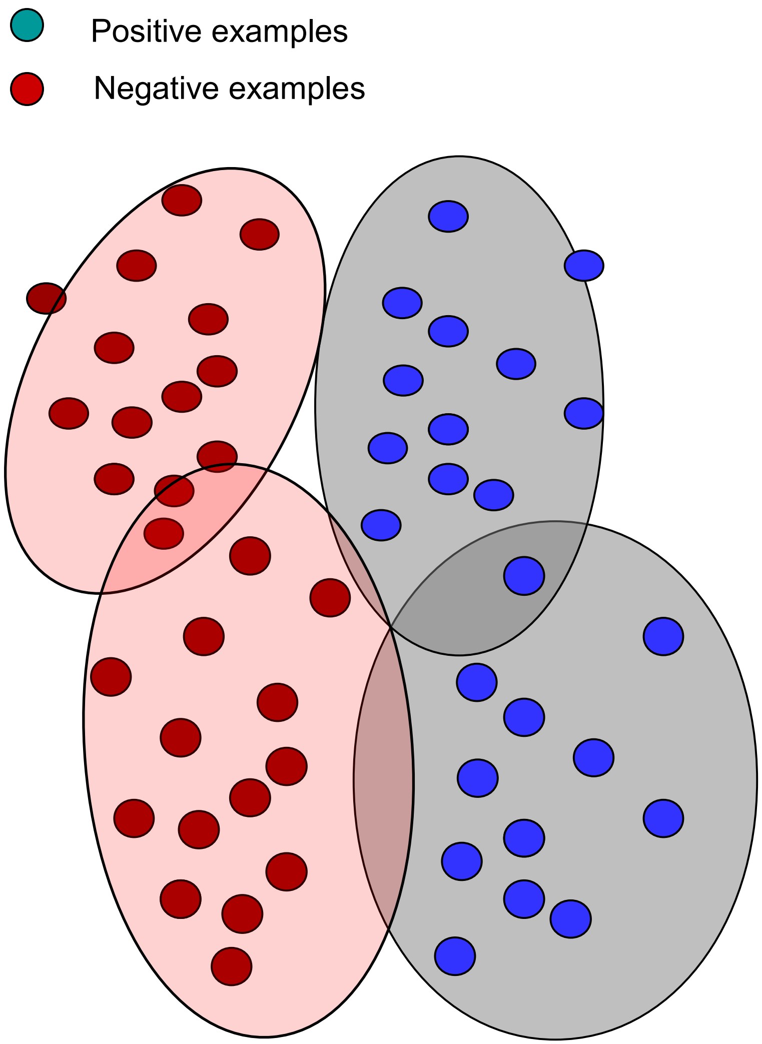
- Given a vector **x** at N-dimensional feature space, the probability that it belongs to a model C is:

$$p(\mathbf{x} \mid C) = \sum_{j=1}^{M} p(\mathbf{x} \mid c_j) p(c_j)$$

where M is the number of Gaussian components. Each component is an N-dimensional Gaussian function which is determined by its mean vector μ and covariance matrix Σ.

$$p(\mathbf{x} \mid c_j) = \frac{1}{(2\pi)^{N/2} |\Sigma_j|^{1/2}} \exp^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}-\boldsymbol{\mu}_j)}$$

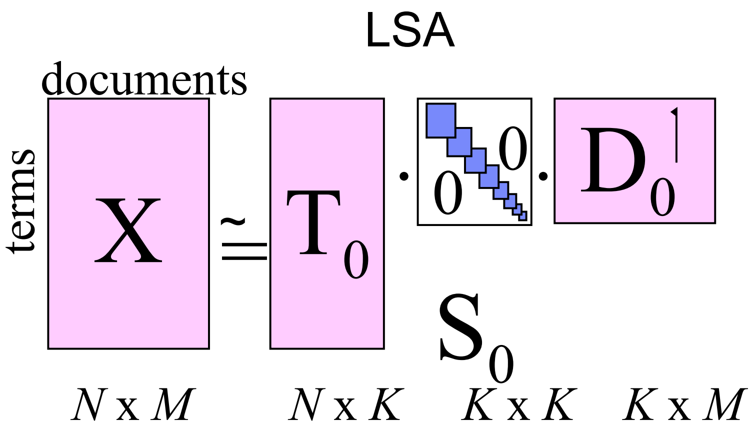
Gaussian Mixture Models



- The probability of examples belonging a cluster (e.g., positive set) is modeled as a weighted summation of multiple Gaussians distributed in the feature space.
- Pros:
  - The size of model parameters can be very small. In a lot of applications, we may assume the Gaussian components at different dimensions are independent. Thus (mean, std) at each feature dim.
- Cons:
  - Usually Estimation-Maximization (EM) Models are used to estimate models. The results may vary depending on the initial condition

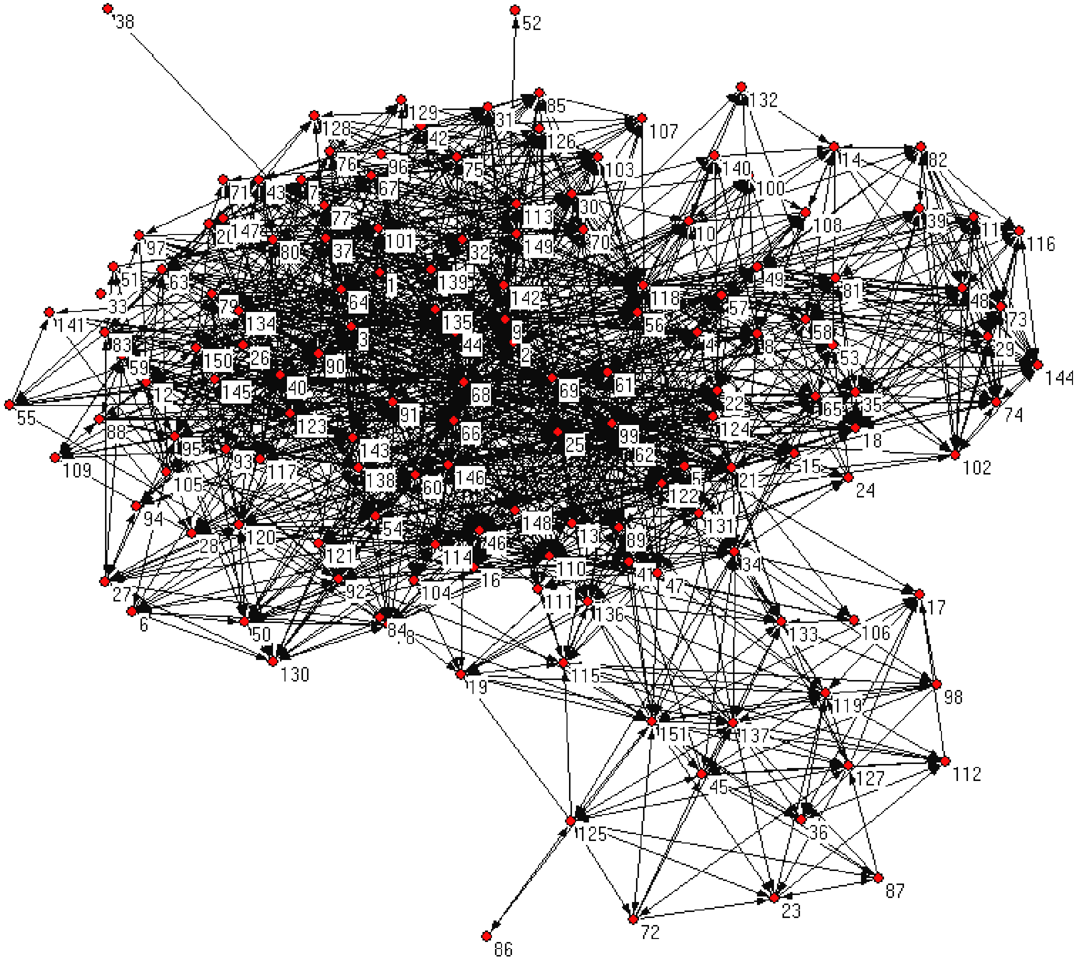
Content Analysis Prior Art I -- Latent Semantic Analysis

- Latent Semantic Analysis (LSA) [Landauer, Dumais 1997]
  - Descriptions:
    - Capture the semantic concepts of documents by mapping words into the latent semantic space which captures the possible synonym and polysemy of words
    - Training based on different level of documents. Experiments show the synergy of the # of training documents and the psychological studies of students at 4th, 10th, and college level. Used as an alternative to TOEFL test.
  - Based on truncated SVD of document-term matrix: optimal least-square projection to reduce dimensionality
  - Capture the concepts instead of words
    - Synonym
    - Polysemy



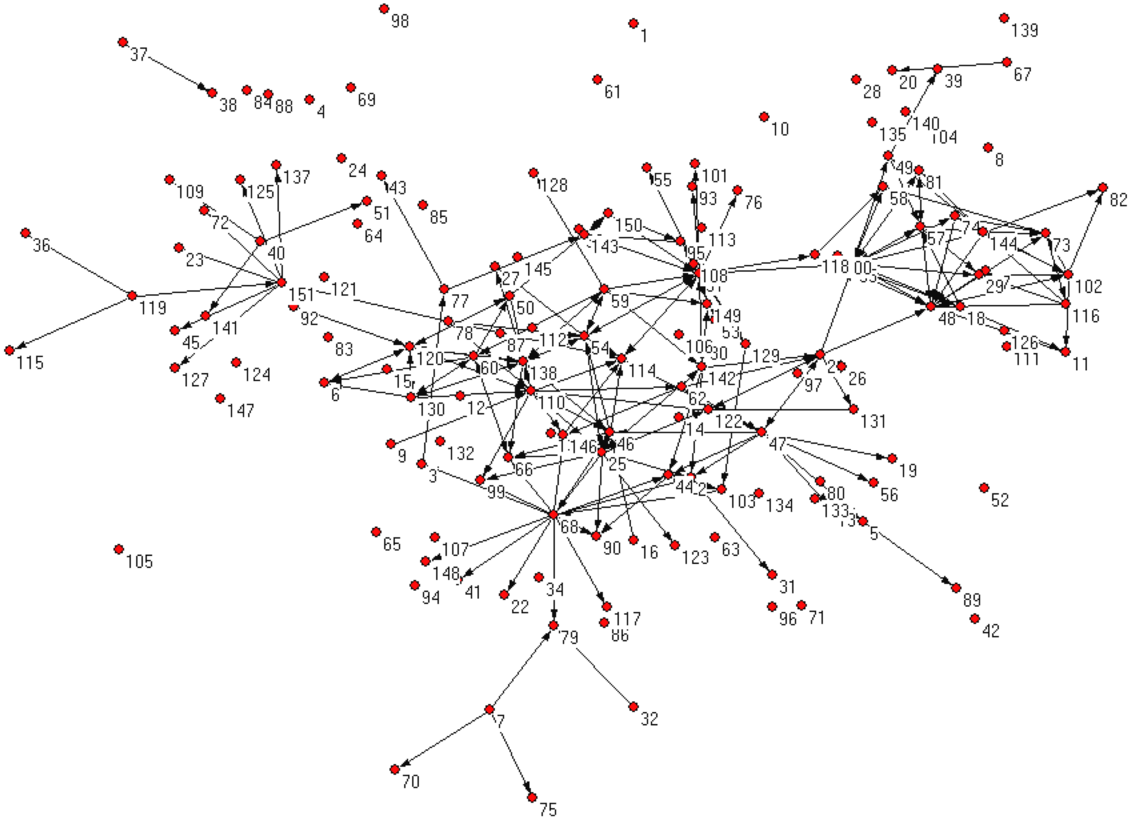
Example: Social Network of Enron Managers

- If we try to build social networks based on communications regardless expertise or topics, it is difficult.



We can first find the experts and then see how this community works (II)

- Rosalee Fleming played an important role at "Market Opportunities." She received info from Actor 119 (Mike Carson) and Actor 23 (James Steffes – VP of Gov. Affairs of Enron.)
- Actor 68 (Rod Havslett -- CFO) is also a major information spreader.



Network Topology Inference

- Link Inference – how to decide a link?
  - User-specified decisions and rules
  - Domain / expert knowledge

- Validation:
  - Whether the network graph representation is accurate?
  - What accuracy ideally can be expected, given the available measurable info?
  - Whether there are other similar representations that are equally or almost as accurate?
  - How robust the representation is to small changes in the measurements?
  - How useful the representation is as a basis for other purposes?

Statistical Inference on Graph

- Take our goal to select an appropriate number of graph that “best” captures the underlying state of the system.
- Based on the information in the data, as well as:
  - Any other prior information
  - Using techniques of statistical modeling and inference.

- Unfortunately, there is now no single coherent body of formal results on inference problems of this type…

➔ different forms, different context, different goals….

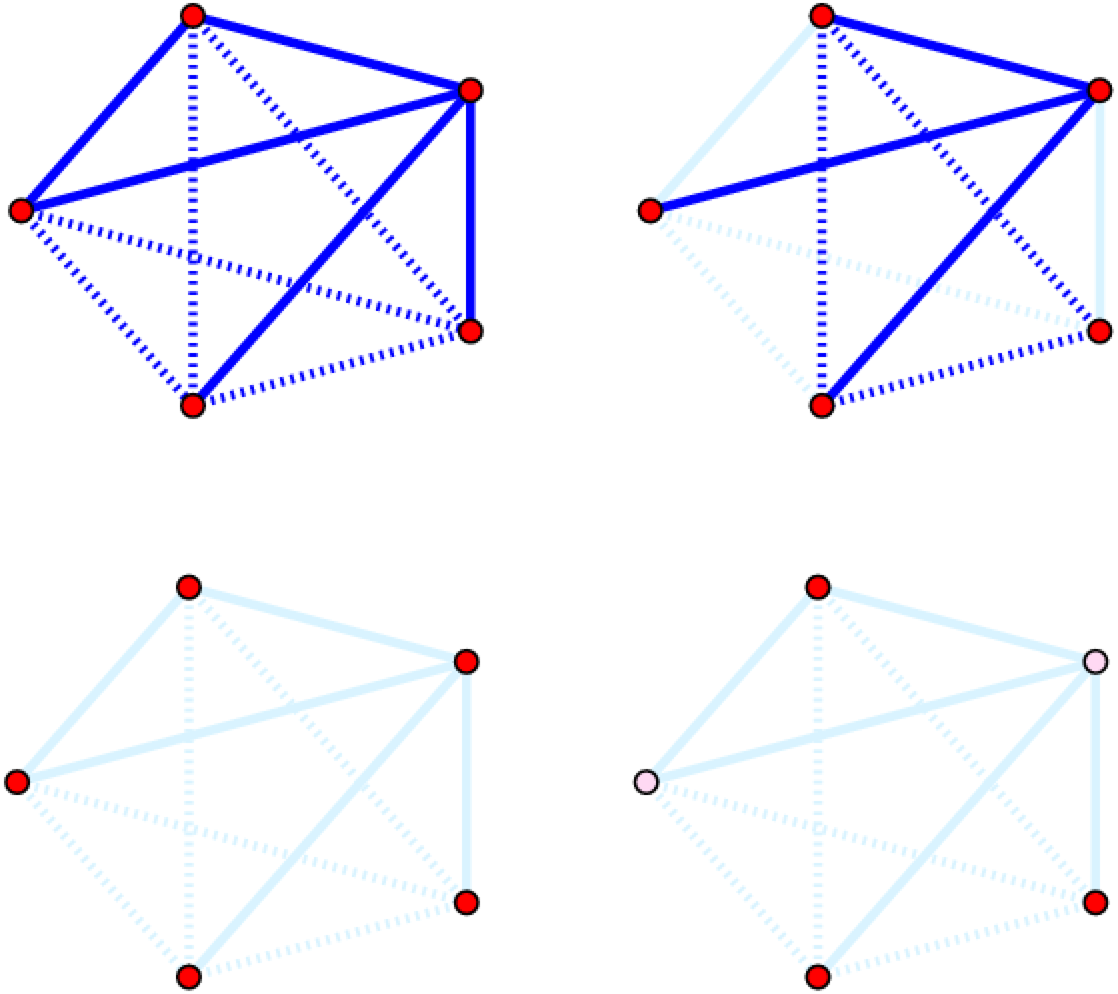
Canonical problems

- Link prediction:
  - Whether a pair of vertices does or does not have an edge between them.
  - Assume knowledge of all the vertices and the status of some of the edges/non-edges of the graph, and seeks to infer the rest of the edges.

- Inference of Association Networks:
  - Edges are defined by a "nontrivial" level of association between certain characteristics of the vertices.
  - No knowledge of edge status anywhere in the network graph.

- Tomographic Network Inference:
  - Measurements are available only at vertices that are somehow at the 'perimeter' of the network.
  - Infer the presence or absence of both edges and vertices in the 'interior'.
  - Involves measurements at only a particular subset of vertices

Visual characterization of three types of network topology inference problems



Link Prediction

- Let $G = (V, E)$ be an undirected random network graph.

- $V^{(2)}$ is the set of distinct unordered pairs of vertices.

- It is the union of:
  - The set E of edges in G,
  - The set $V^{(2)} \setminus E$ of non-edges in G.

- The presence or absence of an edge is observed only for some subset of pairs, say $V_{obs}^{(2)}$. For the remaining pairs, say $V_{miss}^{(2)} = V^{(2)} \setminus V_{obs}^{(2)}$, this information is missing.

Link Prediction (cont'd)

- Let $\mathbf{Y}$ be the random $N_v \times N_v$ binary adjacency matrix.
- Denote $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{miss}$ the entries of $\mathbf{Y}$.
- The problem of link prediction is to predict the entries in $\mathbf{Y}^{miss}$, given the values $\mathbf{Y}^{obs} = \mathbf{y}^{obs}$ and possibly various vertex attribute variables

  $\mathbf{X} = \mathbf{x} \in R^{N_v}$

- Sometimes edges are missing, because of the difficulty in observation in a certain point in time.
- Sometimes edges are missing because of sampling.

A common assumption – missing at random

- Hoff (2007); Taskar et. al. (2004):
  - Missing information on edge presence/absence is *missing at random*.
    - The probability that an edge variable is observed depends only on the values of those other edge variables observed and not on its own value.
  - If edge variables are missing with probability depending upon themselves, this is called *informative missingness*.

- An example of information missingness in biological networks:
  - Network derived from databases (summarizing known results)
  - Bias toward 'positive' results in literatures.

The prediction model

- Given an appropriate model for $\mathbf{X}$ and $(\mathbf{Y}^{obs}, \mathbf{Y}^{miss})$, we might aim to jointly predict the elements of $\mathbf{Y}^{miss}$ based on a model for

$$P(\mathbf{Y}^{miss} \mid \mathbf{Y}^{obs} = \mathbf{y}^{obs}, \mathbf{X} = \mathbf{x})$$

➔Use the network models, e.g., as discussed in the previous 2 lectures.

➔ This approach simplify matters considerably in many ways

Informal Scoring Methods

- Scoring methods:
  - Somewhat less formal than the model-based methods
  - Popular and can be effective.
  - A score function can be incorporated into the model-based methods as explanatory variables.

- With scoring methods, for each pair of vertices i and j whose edge status is unknown, a score $s(i,j)$ is computed.

- A set of predicted edges may then be returned either by applying a threshold s^* to those scores, or by ordering them and keeping those pairs with the top n^* values.

Informal Scoring Methods (cont'd)

- A simple score, inspired by the small-world principle, is negative the shortest-path distance between i and j,

$$s(i,j) = -\,dist_{G^{obs}}(i,j)$$

Jaccard coefficient and other coefficients

- Scores based on comparison of the observed neighborhoods N_i^{obs} and N_j^{obs} of i and j in G^{obs} including the number of common neighbors:

$$s(i,j) = | N_i^{obs} \cap N_j^{obs} |$$

- A standardized version of this value, called the Jaccard coefficient,

$$s(i,j) = \frac{| N_i^{obs} \cap N_j^{obs} |}{| N_i^{obs} \cup N_j^{obs} |}$$

- A variation of this idea due to Liben-Nowell and Kleinberg (2003), weighting more heavily those common neighbors that are themselves not highly connected

$$s(i,j) = \sum_{k \in N_i^{obs} \cap N_j^{obs}} \frac{1}{\log | N_k^{obs} |}$$

Probabilistic Classification Methods

- Learning approaches.

- Logistic Regression Classifiers:

$$\log\left[\frac{P_\beta(Y_{ij}=1\,|\,\mathbf{Z}_{ij}=\mathbf{z})}{P_\beta(Y_{ij}=0\,|\,\mathbf{Z}_{ij}=\mathbf{z})}\right]=\beta^T\mathbf{z}$$

where β is a vector of regression coefficients, and $\boldsymbol{Z}_{ij}$ is a vector of explanatory variables indexed in the unordered pairs $\{i,j\}$.

➔ Standard techniques, based on an efficient iteratively re-weighted least-squares algorithm, may be used to produce maximum likelihood estimates.

Probabilistic Classification Methods (cont'd)

- Potential edges are then classified as being present or absent according to whether or not the estimated classification probabilities:

$$P_{\hat{\beta}}(Y_{ij}^{miss} = 1 \mid \mathbf{Z}_{ij} = \mathbf{z}) = \frac{\exp(\hat{\beta}^T \mathbf{z})}{1 + \exp(\hat{\beta}^T \mathbf{z})}$$

- Two important issues:
  - The standard logistic regression framework assumes that the observations in the training data are independent, conditional on the explanatory variables ➔ no formal work to explore the implications on prediction accuracy of ignoring possible dependencies
  - The nature of underlying mechanism of missingness is relevant. If the underlying missing edges is not at random, the accuracy will suffer.

Logistic Regression with latent Variables

- Hoff (2007, 2008):

- Let $\mathbf{M}$ be an unknown random, symmertric $N_v \times N_v$ matrix of the form

$$\mathbf{M} = \mathbf{U}^T \mathbf{\Lambda} \mathbf{U} + \mathbf{E}$$

- Where **U** is a random orthonormal matrix, $\mathbf{\Lambda}$ is a random diagonal matrix, and **E** is a matrix of i.i.d. noise variables.

$$\log\left[\frac{P_\beta(Y_{ij} = 1 \mid \mathbf{Z}_{ij} = \mathbf{z}, M_{ij} = m)}{P_\beta(Y_{ij} = 0 \mid \mathbf{Z}_{ij} = \mathbf{z}, M_{ij} = m)}\right] = \beta^T \mathbf{z} + m$$

➔ Although Y_{ij} are still conditionally independent, given the Z_{ij} and M_{ij}, they are now conditionally dependent given only the Z_{ij}.

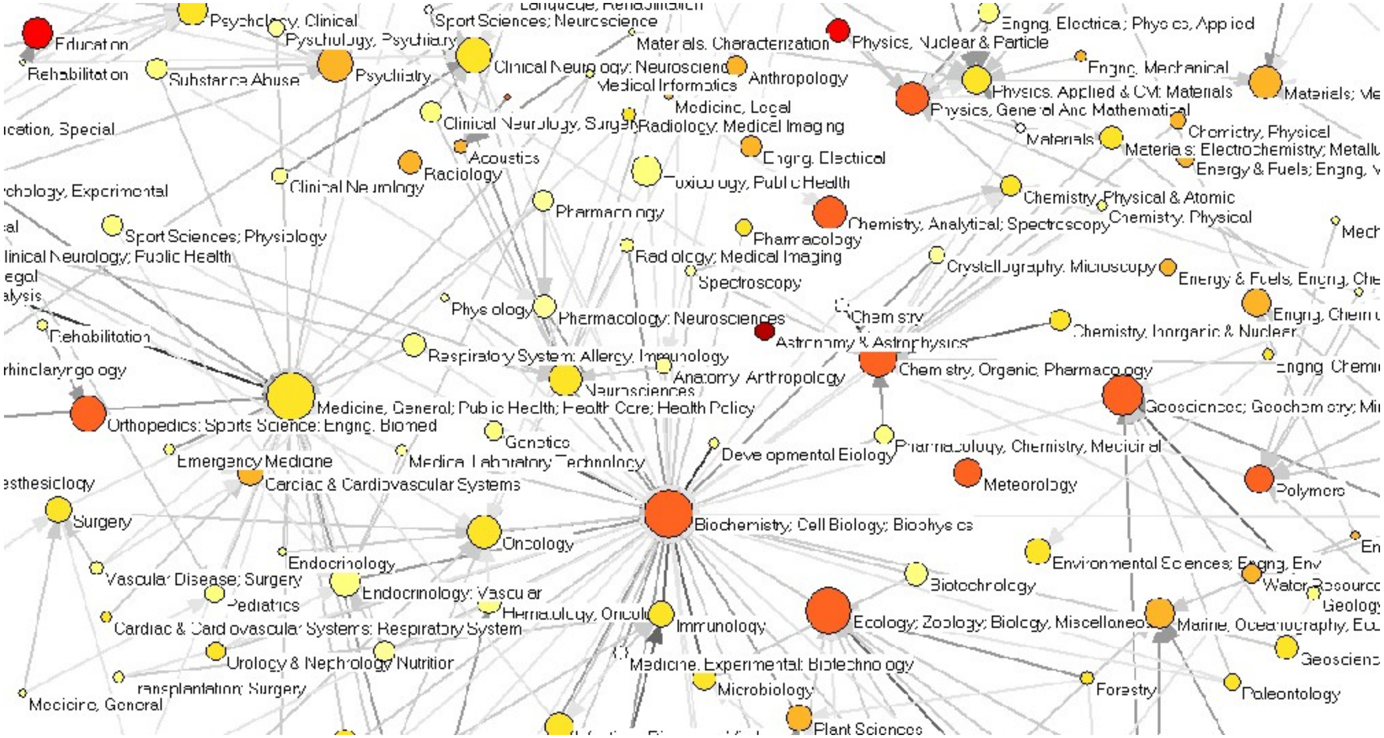
Logistic Regression with latent Variables (cont’d)

- If a Bayesian approach to intference is to be used, a prior distribution for ✔ is also needed.
- An intuitive choice for modeling U is the uniform distribution on the space of all matrices.
- For other variables, multivariate Gaussian distributions are convenient, due to the manner in which they facilitate MCMC sampling of the relevant posterior distribution through conjugacy properties.

$$E\left(\frac{\exp(\hat{\beta}^T \mathbf{Z}_{ij} + M_{ij})}{1 + \exp(\hat{\beta}^T \mathbf{Z}_{ij} + M_{ij})} \,\middle|\, \mathbf{Y}^{obs} = \mathbf{y}^{obs}, \mathbf{Z}_{ij} = \mathbf{z} \right)$$

Inference of Association Networks

- For instance, this network is derived from use of the Jaccard measure of similarity and a combination of ad hoc thresholding and expert-guided clustering rules.


Language, Rehabilitation
Psychology, Clinical
Sport Sciences; Neuroscience
Engng, Electrical; Physics, Applied
Education
Psychology, Psychiatry
Materials, Characterization
Physics, Nuclear & Particle
Rehabilitation
Substance Abuse
Psychiatry
Clinical Neurology; Neuroscience
Medical Informatics
Anthropology
Engng, Mechanical
Physics, Applied & CM: Materials
Materials; Me
Medicine, Legal
Physics, General And Mathematical
Education, Special
Clinical Neurology, Surgery
Radiology; Medical Imaging
Materials
Chemistry, Physical
Materials; Electrochemistry; Metallu
Acoustics
Radiology
Engng, Electrical
Energy & Fuels; Engng, V
Psychology, Experimental
Clinical Neurology
Toxicology, Public Health
Chemistry, Physical & Atomic
Pharmacology
Chemistry, Analytical; Spectroscopy
Chemistry, Physical
Sport Sciences; Physiology
Pharmacology
Mech
Clinical Neurology; Public Health
Radiology; Medical Imaging
Legal
Crystallography; Microscopy
Energy & Fuels, Engng, Che
Analysis
Spectroscopy
Physiology
Pharmacology; Neurosciences
Chemistry
Engng, Chem
Rehabilitation
Astronomy & Astrophysics
Chemistry, Inorganic & Nuclear
Respiratory System; Allergy; Immunology
Otorhinolaryngology
Anatomy, Anthropology
Engng, Chemi
Neurosciences
Chemistry, Organic, Pharmacology
Medicine, General; Public Health; Health Care; Health Policy
Geosciences; Geochemistry; Mi
Orthopedics: Sports Science: Engng, Biomed
Genetics
Developmental Biology
Pharmacology, Chemistry, Medicinal
Emergency Medicine
Medical Laboratory Technology
Anesthesiology
Cardiac & Cardiovascular Systems
Meteorology
Polymers
Surgery
Biochemistry; Cell Biology; Biophysics
Oncology
En
Endocrinology
Environmental Sciences; Engng, Env
Vascular Disease; Surgery
Biotechnology
Water Resources
Endocrinology; Vascular
Pediatrics
Geology
Hematology, Oncology
Immunology
Ecology; Zoology; Biology, Miscellaneous
Marine, Oceanography, Ec
Cardiac & Cardiovascular Systems; Respiratory System
Urology & Nephrology
Nutrition
Geoscienc
Medicine, Experimental; Biotechnology
Transplantation; Surgery
Microbiology
Forestry
Paleontology
Medicine, General
Plant Sciences

Correlation Networks

- Pearson product-moment correlation between two nodes.

$$\rho_{ij} = corr(X_i, X_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

- Empirical correlations

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$$

- If the pair of variables (Xi,Xj) has a bivariate Gaussian distribution, the density $\hat{\rho}_{ij}$ under H0: $\rho_{ij} = 0$ has a concise closed-form expression, but it is somewhat complicated and requires numerical integration or tables to produce p-values.

Partial Correlation Networks

- Important -- ‘Correlation does not imply causation’

- For instance:
  - Two vertices may have highly correlated attributes because the vertices somehow strongly ‘influence’ each other in a direct fashion.
  - Alternatively, their correlation may be high primarily because they each are strongly influenced by a third vertex.

➔ Need more considerations!!

Partial Correlation Networks (cont'd)

- If it is felt desirable to construct a graph G where the inferred edges are more reflective of direct influence among vertices, rather than indirect influence, the notion of partial correlation becomes relevant.

- The partial correlation of attributes X_i and X_j of vertices i, j, defined with respect to the attributes $X_{k1}, \ldots X_{km}$ of vertices $k_1, \ldots k_m$ in $k_1, \ldots, k_m \in V \setminus \{i, j\}$, is the correlation between X_i and X_j left over after adjusting for those effects of $X_{k1}, \ldots X_{km}$ common to both.

- Let $Sm=\{k_1, \ldots k_m\}$, we define the partial correlation of X_i and X_j , adjusting for $\mathbf{X}_{S_m} = (X_{k_1}, \ldots, X_{km})^T$, as

$$\rho_{ij|S_m} = \frac{\sigma_{ij|S_m}}{\sqrt{\sigma_{ii|S_m}\sigma_{jj|S_m}}}$$

Partial Correlation Networks (cont'd)

- Here

$$Cov\begin{pmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{pmatrix} = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

for $\mathbf{W}_1 = (X_i, X_j)^T$ and $\mathbf{W}_2 = \mathbf{X}_{S_m}$.

and then $\sigma_{ii|S_m}$, $\sigma_{jj|S_m}$, and $\sigma_{ij|S_m} = \sigma_{ji|S_m}$ are the diagonal and off-diagonal elements of this 2x2 partial covariance matrix.

$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

- For example, for a given choice of m, we may dictate that an edge be present only when there is correlation between X_i and X_j regardless of which m other vertices are conditioned upon.

$$E = \left\{ \{i,j\} \in V^{(2)} : \rho_{ij|S_m} \neq 0, \forall\, S_m \in V_{\setminus\{i,j\}}^{(m)} \right\}$$

Gaussian Graphical Model Networks

- A special and popular case of the use of partial correlation coefficients is when $m=N_v$-2 and the attributes are assumed to have a multivariate Gaussian joint distribution.

- Here the partial correlation between attributes of two vertices is defined conditional upon the attribute information at all other vertices.

- The graph with edge set

$$E = \left\{ \{i, j\} \in V^{(2)} : \rho_{ij|V\setminus\{i.j\}} \neq 0 \right\}$$

is called a conditional independence graph. The overall model, combing the multivariate Gaussian distribution with the graph G, is called a Gaussian graphical model.

- The partial correlation coefficients may be expressed in the form:

$$\rho_{ij|V\setminus\{i,j\}} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

where ω_{ij} is the (I,j)-th entry of $\Omega = \Sigma^{-1}$, the inverse of the covariance matrix of vertex attributes.

Tomographic Network Topology Inference

- Inference of ‘interior’ components of a network – both vertices and edges – from data obtained at some subset of ‘exterior’ vertices.
- E.g., in computer networks, desktop and laptop computers are typical instances of ‘exterior’ vertices, while Internet routers to which we do not have access are effecitvely ‘interior’ vertices.
- Network tomography: describe prblems in the context of computer network monitoring, in which aspects of the ‘internal’ workings of the network, suchas intensity of traffic flowing between vertices, are inferred from ‘external’ measurements.
- This is an’ill-posed inverse proble’ in mathematics. Mapping being many-to-one.
- For the tomographic inference of network topologies, a key structural simplification has been the restriction to inference of networks in the form of trees.

Tomographic Inference of Tree Topologies

- A tree structure.

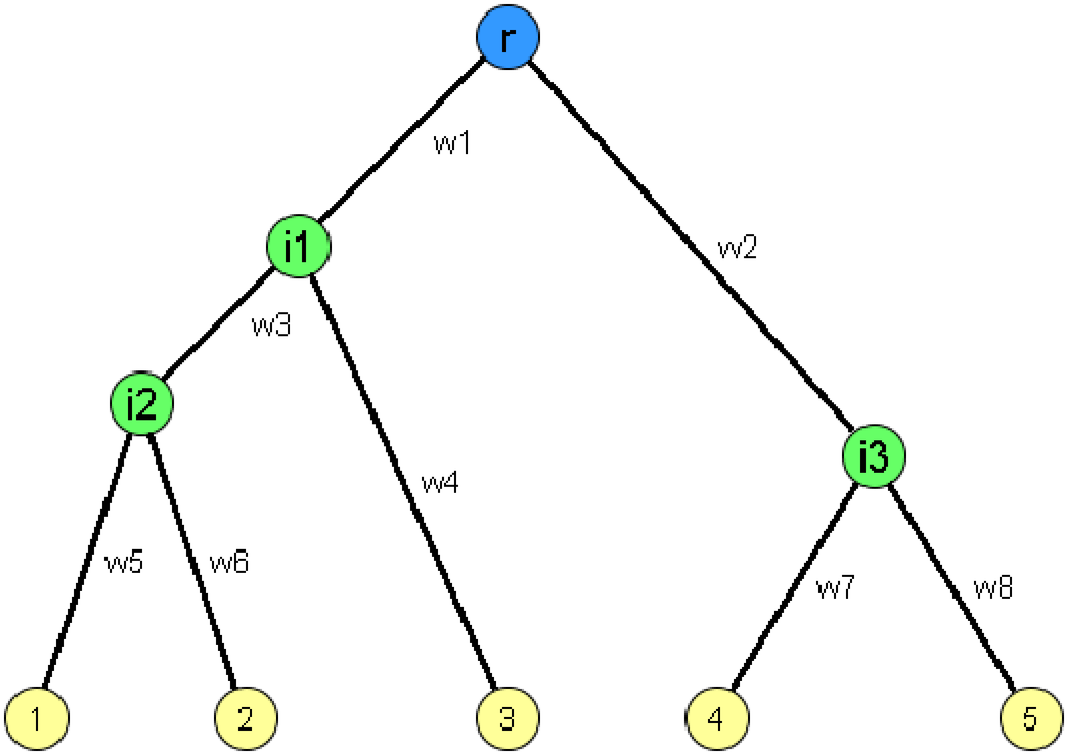


Fig. 7.8 Schematic representation of a binary tree in association with the tomographic network inference problem. Measurements are available at the leaves 1, 2, 3, 4, and 5, in yellow. The internal vertices, i_1, i_2, and i_3, in green, and possibly the root r, in blue, are unknown, as are the branches joining the various vertices.

Tomographic network inference problem

- Suppose a set of N_l vertices, we have n iid observations of some random variables $\{X_1, \ldots X_{Nl}\}$. We aim to find that three of all binary trees with N_l labeled leaves that best explains the data.

- Example: Multicast probes

Two most popular classes of methods

- Hierarchical clustering and related ides

- Likelihood-based methods

Hierarchical Clustering- based methods

- We treat the N_l leaves as the ‘objects’ to be clustered and the tree corresponding to the resulting clustering as our inferred tree.

- The tree corresponding to the entire set of parititions is our focus.

- Example: Rantnasamy and McCanne (1999): the observed rate of shared losses of packets should be fairly indicative of how close two leaf vertices (i.e., destination addresses) are on a multicast tree T.
  - Two different types of shared loss between a pair of leaf vertices – ‘true’ and ‘false’ shared loss.
  - The true shared losses are due to loss of packets on the path common to the vertices i and j.
  - The valse shared losses would refer to cases where packets were lost separately on the two paths from i1 to the vertices 1 and 3.

Likelihood-based Methods

- If we are willing to specify probability models, then we have the potential for likelihood-based methods of inference.

- In general, maximum likelihood inference of tree topologies may also be pursued through the use of MCMC. MCMC is critical to the use of Bayesian methods to tomographic inference of trees.

Final Project

- Team work: 1 – 3 people per team
- Project Idea proposal, no later then 11/01/2012 ➔ please send me email (cylin at ee dot columbia dot edu) of your rough idea / plan, discussing with me or TA.
- Project initial preparation & plan presentation: 11/12/2012 ➔ 5 mins per person. Team will present together.

- Potential Topics:
  - Classification & Prediction of People Behavior in Networks
  - Predicting Network Characteristics Evolution
  - Visualizing Networks
  - Verifying fitness of various network models to practical data
  - Anything related to network…..

Anything on Networks

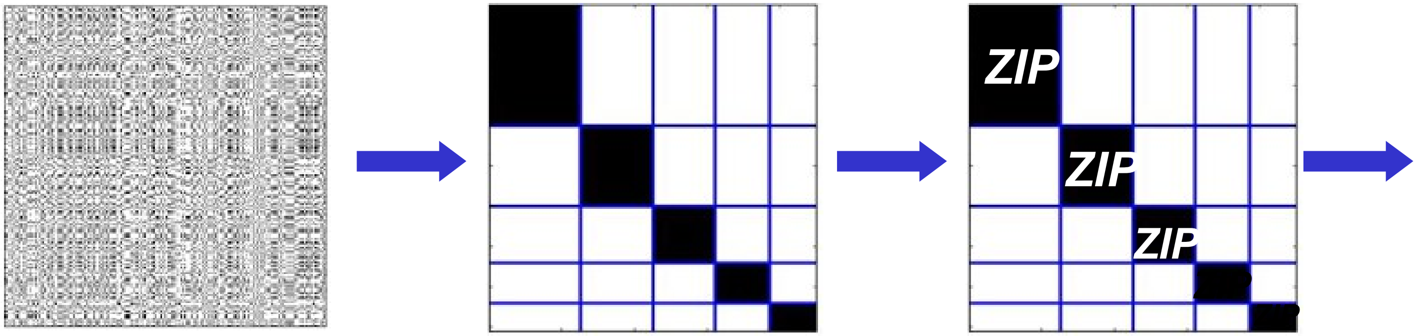
- Formation of Network
  - Communications ← *Electrical Engineering*
  - Information ← *Computer Science*
  - People ← *Sociology, Public Health*
  - Companies / Organizations ← *Economics, Management, Politics*
  - Nations ← *International Relationships, History*
- Network Data Collection ← *Law*
- Network Science Infrastructure
- Network Applications
- Network Visualization ← *Arts, Math*
- Network Sampling, Indexing and Compression ← *Math*
- Network Flow ← *Physics*
- Network Evolution and Dynamics
- Network Impact
- Cognitive Networks ← *Bio, Cognition, Behavior Science*

Other possible topics not discussed in the textbook

- Network compression
- Cognitive networks
- Mobile applications

Large-Scale Graph Indexing and Network Management

- Build efficient indexing to support efficient storing, retrieving, and querying large graphs.



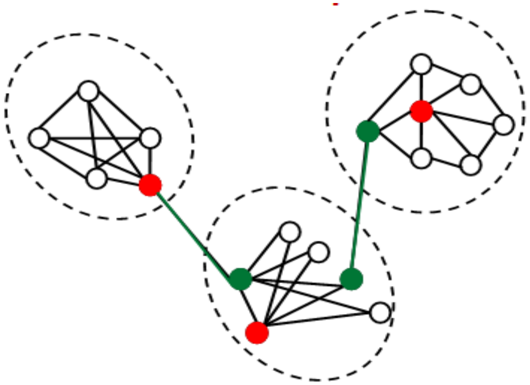
Neighbor
Community
Ego-net
Shortest path
Diameter
…

Raw Graph

Partition

Store, Index, Retrieval

Querying



❖Find Communities
❖Find a few nodes/edges to describe
  - each community
  relationship between 2 communities

Algorithms that make indexing & queries possible for large graphs

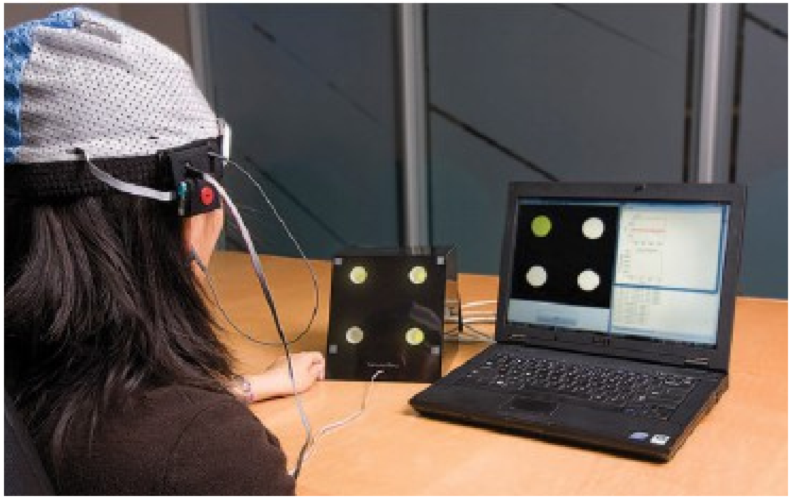
Traditional Cognitive (EEG) Sensor Signal Extraction

- Challenges:
  - Gel
  - Needles
  - Wires

New Type of EEG Detector and Signal Analysis

- Fundamental Research on EEG Signal Processing
- New Dry Sensing
  - Classifying Attention, Relaxation, etc.
  - Classifying Target – P300 signals
  - Classifying Visual Cortex Signals
- Breakthrough Non-Contact Sensing – suitable for everyday/normal use
- Cognitive Wireless Sensor becomes possible

EEG Wireless Sensor developed by an IBM partner



Network Analysis Application on Mobile

- SmallBlue Applications on Mobile Phones (Nov 2009)

Show Expertise of ‘SNA’ inside:

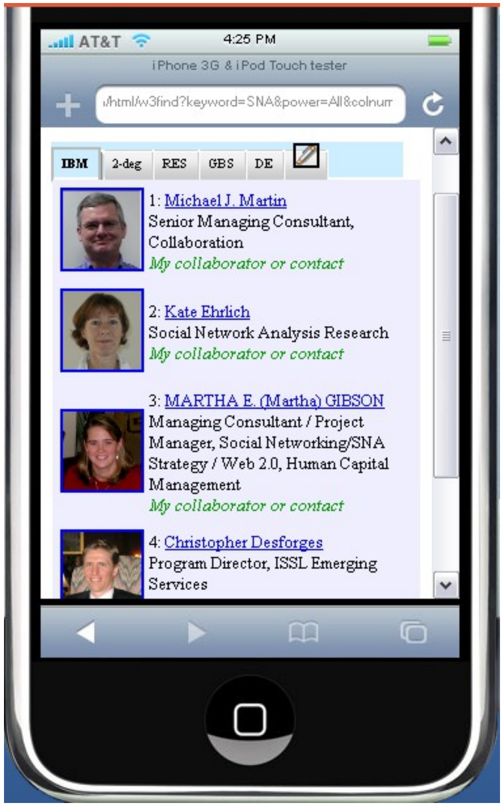
(1) IBM

(2) My 2-degree network

(3) Research division

(4) Global Business Services

(5) Any group – e.g., Distinguished Engineers



AT&T
4:25 PM
iPhone 3G & iPod Touch tester
/html/w3find?keyword=SNA&power=All&colnum
IBM
2-deg
RES
GBS
DE

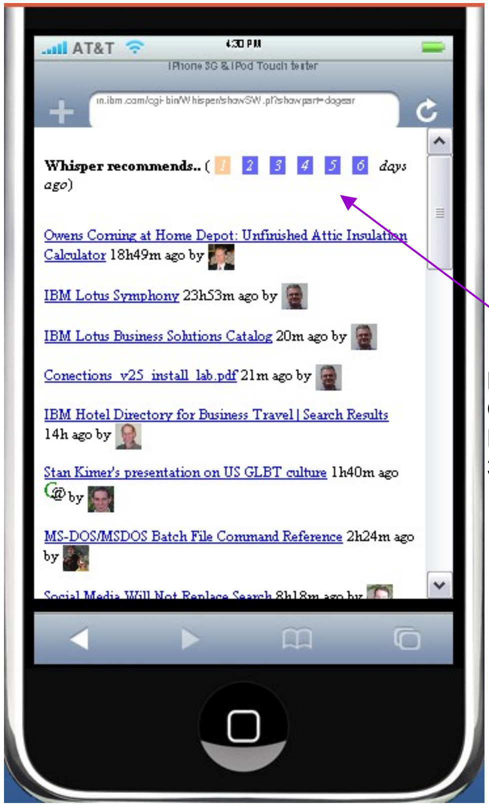
1: Michael J. Martin
Senior Managing Consultant, Collaboration
My collaborator or contact

2: Kate Ehrlich
Social Network Analysis Research
My collaborator or contact

3: MARTHA E. (Martha) GIBSON
Managing Consultant / Project Manager, Social Networking/SNA Strategy / Web 2.0, Human Capital Management
My collaborator or contact

4: Christopher Desforges
Program Director, ISSL Emerging Services

SmallBlue Find Widget in Mobile



AT&T
4:30 PM
iPhone 3G & iPod Touch tester
m.ibm.com/cgi-bin/Whisper/showSW.pl?showpart=dogear

Whisper recommends.. (1 2 3 4 5 6 *days ago*)

Owens Corning at Home Depot: Unfinished Attic Insulation Calculator 18h49m ago by

IBM Lotus Symphony 23h53m ago by

IBM Lotus Business Solutions Catalog 20m ago by

Conections_v25_install_lab.pdf 21m ago by

IBM Hotel Directory for Business Travel | Search Results 14h ago by

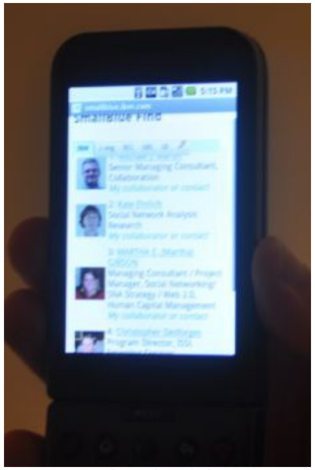
Stan Kimer's presentation on US GLBT culture 1h40m ago by

MS-DOS/MSDOS Batch File Command Reference 2h24m ago by

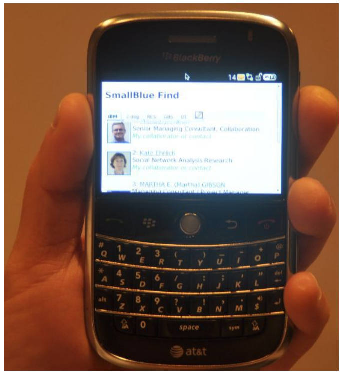
Social Media Will Not Replace Search 8h18m ago by

SmallBlue Whisper Widget in Mobile

Recommend Contents from Friends within 3-degrees



Android



BlackBerry

Nokia