E6885 Network Science Lecture 6: 

*Network Topology Inference*

Ching-Yung Lin, Dept. of Electrical Engineering, Columbia University
October 15th, 2012
## Course Structure

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

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

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Subject</th>
<th>Time</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>31265362.1075886640461</td>
<td>FW, California gas intrastate matters</td>
<td>2001-07-10T19:32:23</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:matt.smith@enron.com">matt.smith@enron.com</a>, <a href="mailto:keith.hols@enron.com">keith.hols@enron.com</a>, <a href="mailto:mike.hanks@enron.com">mike.hanks@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>14973812.1075886640483</td>
<td>FW, West Power Strategy Briefing</td>
<td>2001-07-11T12:56:41</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:keith.hols@enron.com">keith.hols@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>3650242.1075886640506</td>
<td></td>
<td>2001-07-11T15:25:40</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:bary.tycholiz@enron.com">bary.tycholiz@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>493902.1075886640564</td>
<td>Je, FW, Party</td>
<td>2001-07-12T12:04:29</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:s.shively@enron.com">s.shively@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>13141541.1075886640571</td>
<td></td>
<td>2001-07-12T19:55:20</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:michael.bruner@essmo.com">michael.bruner@essmo.com</a>,</td>
<td></td>
</tr>
<tr>
<td>14620063.1075886640594</td>
<td>CA, Instrate Gas matters</td>
<td>2001-07-13T13:44:25</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:leslie.lawne@enron.com">leslie.lawne@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>634023.1075886640616</td>
<td>Je, FW, CA, Instrate Gas matters</td>
<td>2001-07-13T13:45:30</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:mike.grigsby@enron.com">mike.grigsby@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>19292752.1075886640638</td>
<td>Analysis/Associate Program, 2 Min</td>
<td>2001-07-13T19:47:41</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:ramsbillo@exxonmobil.com">ramsbillo@exxonmobil.com</a>,</td>
<td></td>
</tr>
<tr>
<td>16515849.1075886640569</td>
<td>FW, American Express Letter</td>
<td>2001-07-16T13:20:48</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:johnny.ross@enron.com">johnny.ross@enron.com</a>,</td>
<td></td>
</tr>
<tr>
<td>13618088.1075886640681</td>
<td>Party</td>
<td>2001-07-16T13:32:52</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:richard.touba@truequote.com">richard.touba@truequote.com</a>,</td>
<td></td>
</tr>
<tr>
<td>27461841.1075886640706</td>
<td>FW, Party</td>
<td>2001-07-16T13:44:37</td>
<td><a href="mailto:k.allen@enron.com">k.allen@enron.com</a>, <a href="mailto:s.shively@enron.com">s.shively@enron.com</a>,</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Centrality</th>
<th>1999</th>
<th>Position</th>
<th>Name</th>
<th>2000</th>
<th>Position</th>
<th>Name</th>
<th>2001</th>
<th>Position</th>
<th>Name</th>
<th>2002</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>Mark_Taylor</td>
<td></td>
<td></td>
<td>David_Delainey</td>
<td></td>
<td>Steven_Kean</td>
<td>Vice_President</td>
<td></td>
<td>Kevin_Presto</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>Tana_Jones</td>
<td></td>
<td></td>
<td>Steven_Kean</td>
<td></td>
<td>John_Lavorato</td>
<td>CEO</td>
<td></td>
<td>Louise_Kitchen</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>Sara_Shackleton</td>
<td></td>
<td></td>
<td>John_Lavorato</td>
<td></td>
<td>Jeff_Dasovich</td>
<td>Employee</td>
<td></td>
<td>John_Lavorato</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>Richard_Sanders</td>
<td>Vice_President</td>
<td>Vince_Kaminski</td>
<td>Manager</td>
<td></td>
<td>Vince_Kaminski</td>
<td>Manager</td>
<td></td>
<td>Hunter_Shively</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>Elizabeth_Sager</td>
<td>Employee</td>
<td>Jeff_Skilling</td>
<td>CEO</td>
<td></td>
<td>Louise_Kitchen</td>
<td>President</td>
<td></td>
<td>James_Steffes</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Managing_Director</td>
<td>Mark_Haedicke</td>
<td>Mike_McConnell</td>
<td>N/A</td>
<td>David_Delainey</td>
<td></td>
<td>Greg_Whalley</td>
<td>President</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Managing_Director</td>
<td>John_Hodge</td>
<td>Greg_Whalley</td>
<td>President</td>
<td>Greg_Whalley</td>
<td></td>
<td></td>
<td>Fletcher_Sturm</td>
<td></td>
<td>Vice_President</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Vice_President</td>
<td>Steven_Kean</td>
<td>Sally_Beck</td>
<td>Employee</td>
<td>Mark_Haedicke</td>
<td>Managing_Director</td>
<td>Doug_Gilbert-Smith</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Employee</td>
<td>Dan_Hyvl</td>
<td>Jeffrey_A_Shank</td>
<td>N/A</td>
<td>Phillip_A llen</td>
<td>N/A</td>
<td>Dana_Davis</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>In_house_lawyer</td>
<td>Carol_Clair</td>
<td>John_Arnold</td>
<td>Vice_President</td>
<td>Mary_Hain</td>
<td>In_house_lawyer</td>
<td>Mark_Haedicke</td>
<td>Managing_Director</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
People with top-10 prestige in Enron

<table>
<thead>
<tr>
<th>Prestige</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
<td>Position</td>
<td>Name</td>
<td>Position</td>
</tr>
<tr>
<td>1</td>
<td>John_Hodge</td>
<td>Managing_Director</td>
<td>John_Lavorato</td>
<td>CEO</td>
</tr>
<tr>
<td>2</td>
<td>Steven_Kean</td>
<td>Vice_President</td>
<td>Greg_Whalley</td>
<td>President</td>
</tr>
<tr>
<td>3</td>
<td>Vince_Kaminski</td>
<td>Manager</td>
<td>David_Delainey</td>
<td>CEO</td>
</tr>
<tr>
<td>4</td>
<td>Mark_Haedicke</td>
<td>Managing_Director</td>
<td>Steven_Kean</td>
<td>Vice_President</td>
</tr>
<tr>
<td>5</td>
<td>Elizabeth_Sager</td>
<td>Employee</td>
<td>Vince_Kaminski</td>
<td>Manager</td>
</tr>
<tr>
<td>6</td>
<td>Richard_Sanders</td>
<td>Vice_President</td>
<td>Rick_Buy</td>
<td>Manager</td>
</tr>
<tr>
<td>7</td>
<td>Kevin_Presto</td>
<td>Vice_President</td>
<td>Kevin_Presto</td>
<td>Vice_President</td>
</tr>
<tr>
<td>8</td>
<td>Mark_Taylor</td>
<td>Employee</td>
<td>Jeffrey_A_Shan_kman</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>Michelle_Cash</td>
<td>N/A</td>
<td>Phillip_Allen</td>
<td>N/A</td>
</tr>
<tr>
<td>10</td>
<td>Stacy_Dickson</td>
<td>Employee</td>
<td>Jeff_Skilling</td>
<td>CEO</td>
</tr>
</tbody>
</table>

Most of them have relatively high position in Enron, which reveal the roles in the social network actually are almost corresponding to the roles in the real life.
Dynamic Network Analysis needs both Pattern and Content Analysis

Input

User action

Dataset

User infor.

Item infor.

Log files

New items

Dynamic pattern analysis

Content analysis

Community construction

Time-Sensitive AdaBoost

Output

Ranked Recomm. Items
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

Sustained interest

- handouts, specifications, and reference documents

Short-term items

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
Discriminant Modeling
Support Vector Machines

• **Objective**: How to separate observations of distinct classes
• **Solution**: Project to higher dimensionality to determine linear separability
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., \( k(x, x_i) = e^{-r} \)

- **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

- 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

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

\[ K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2) \]
Gaussian Mixture Models

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

- Given a vector $\mathbf{x}$ at N-dimensional feature space, the probability that it belongs to a model $C$ is:

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

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

\[
p(x \mid c_j) = \frac{1}{\sqrt{2\pi} \sqrt{|\Sigma_j|}} \exp\left(-\frac{(x-\mu_j)^T \Sigma_j^{-1} (x-\mu_j)}{2}\right)
\]
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 Hayslett -- 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 $V^{(2)}_{obs}$. For the remaining pairs, say $V^{(2)}_{miss} = V^{(2)} \setminus V^{(2)}_{obs}$, this information is missing.
Link Prediction (cont’d)

- Let $Y$ be the random $N_v \times N_v$ binary adjacency matrix.

- Denote $Y^{obs}$ and $Y^{miss}$ the entries of $Y$.

- The problem of link prediction is to predict the entries in $Y^{miss}$, given the values $Y^{obs} = y^{obs}$ and possibly various vertex attribute variables $X = x \in \mathbb{R}^{N_v}$.

- 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 $X$ and $(Y^{obs}, Y^{miss})$, we might aim to jointly predict the elements of $Y^{miss}$ based on a model for $P(Y^{miss} \mid Y^{obs} = y^{obs}, X = x)$

- 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 $j$,

$$ s(i, j) = - \text{dist}_{G^{obs}} (i, j) $$
Jaccard coefficient and other coefficients

- Scores based on comparison of the observed neighborhoods $\mathcal{N}_i^{obs}$ and $\mathcal{N}_j^{obs}$ of $i$ and $j$ in $G^{obs}$ including the number of common neighbors:

$$s(i, j) = | \mathcal{N}_i^{obs} \cap \mathcal{N}_j^{obs} |$$

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

$$s(i, j) = \frac{| \mathcal{N}_i^{obs} \cap \mathcal{N}_j^{obs} |}{| \mathcal{N}_i^{obs} \cup \mathcal{N}_j^{obs} |}$$

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

$$s(i, j) = \sum_{k \in \mathcal{N}_i^{obs} \cap \mathcal{N}_j^{obs}} \frac{1}{\log | \mathcal{N}_k^{obs} |}$$
Probabilistic Classification Methods

- Learning approaches.

- Logistic Regression Classifiers:

\[
\log \left[ \frac{P_\beta(Y_{ij} = 1 \mid Z_{ij} = z)}{P_\beta(Y_{ij} = 0 \mid Z_{ij} = z)} \right] = \beta^T z
\]

where \( \beta \) is a vector of regression coefficients, and \( Z_{ij} \) 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:

\[
P_{\hat{\beta}}(Y_{ij}^{\text{miss}} = 1 | Z_{ij} = z) = \frac{\exp(\hat{\beta}^T z)}{1 + \exp(\hat{\beta}^T z)}
\]

- Two important issues:
  - The standard logistic regression framework assumes that the observations in the training data are independent, conditional on the explanatory variables \( \Rightarrow \) 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


- Let $M$ be an unknown random, symmetric $N_v \times N_v$ matrix of the form

  \[ M = U^\top \Lambda U + E \]

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

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

- 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 inference is to be used, a prior distribution for $\beta$ 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.

$$E \left( \frac{\exp(\hat{\beta}^T Z_{ij} + M_{ij})}{1 + \exp(\beta^T Z_{ij} + M_{ij})} \middle| Y^{obs} = y^{obs}, Z_{ij} = z \right)$$
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.
Correlation Networks

• Pearson product-moment correlation between two nodes.

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

• Empirical correlations

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

• If the pair of variables \((X_i, X_j)\) has a bivariate Gaussian distribution, the density \(\rho_{ij}\) 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 $X_{k_1}, \ldots, X_{k_m}$ 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 $X_{k_1}, \ldots, X_{k_m}$ common to both.

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

\[
\text{Cov}
\begin{pmatrix}
W_1 \\
W_2
\end{pmatrix} = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
\]

for \( W_1 = (X_i, X_j)^T \) and \( W_2 = X_{S_m} \).

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.

\[
E = \left\{ \{i, j\} \in V^{(2)} : \rho_{ij|S_m} \neq 0, \forall S_m \in V^{(m)} \setminus \{i, j\} \right\}
\]
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 \( \omega_{ij} \) 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 effectively ‘interior’ vertices.

- Network tomography: describe problems in the context of computer network monitoring, in which aspects of the ‘internal’ workings of the network, such as intensity of traffic flowing between vertices, are inferred from ‘external’ measurements.

- This is an ‘ill-posed inverse problem’ 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 $\{X_1, \ldots, X_{N_l}\}$. 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 ideas
- 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 partitions 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 false shared losses would refer to cases where packets were lost separately on the two paths from $i_1$ 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
  - Information
  - People
  - Companies / Organizations
  - Nations

- Network Data Collection

- Network Science Infrastructure

- Network Applications

- Network Visualization

- Network Sampling, Indexing and Compression

- Network Flow

- Network Evolution and Dynamics

- Network Impact

- Cognitive Networks

Electrical Engineering

Computer Science

Sociology, Public Health

Economics, Management, Politics

International Relationships, History

Law

Arts, Math

Math

Physics

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.

- 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

Recommend Contents from Friends within 3-degrees