Privacy and Network Analysis: Examples and Questions

Ramayya Krishnan (rk2x@cmu.edu)
Director, iLab
Dean, Heinz College
School of Information Systems and Management
School of Public Policy and Management
Outline

• Introduction
  – The R-U framework
  – The traditional data privacy approaches

• Networks
  – Analysis using networks
    • Knowledge management example

• Privacy in Networks
  – Why is it complicated?
  – How does privacy protection affect analysis/inference?
  – Interesting open problem
The basic problem

• Micro data about individuals
  – Relational tuples with data about individual attributes. Each tuple assumed to be independent of the other.
  – Today: Network data from call data records, blogs, friendship networks etc.

• Publish micro-data
  – Maximize utility from the data
  – Subject to confidentiality constraints
The R-U Confidentiality Map (Duncan et al, 2001)

Utility – example 1: Inverse of the RMSE of the estimate of a statistic such as the sample Mean
example 2: sum of tuple information loss criterion
Risk – example 1: Width of the interval at a specified confidence level of value of a Confidential variable that will lead to re-identification; example 2: value of k in K-anonymity
## The Standard Privacy Problem

<table>
<thead>
<tr>
<th>Units</th>
<th>Variables</th>
</tr>
</thead>
</table>

### Solutions:
- Deleting cases
- Aggregating cases
- Deleting variables
- Adding noise
- Perturbations
  - K-anonymity
  - L-diversity
# Micro-data: an example

<table>
<thead>
<tr>
<th>Non-Sensitive</th>
<th>Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip Code</td>
<td>Age</td>
</tr>
<tr>
<td>1</td>
<td>13053</td>
</tr>
<tr>
<td>2</td>
<td>13068</td>
</tr>
<tr>
<td>3</td>
<td>13068</td>
</tr>
<tr>
<td>4</td>
<td>13053</td>
</tr>
<tr>
<td>5</td>
<td>14853</td>
</tr>
<tr>
<td>6</td>
<td>14853</td>
</tr>
<tr>
<td>7</td>
<td>14850</td>
</tr>
<tr>
<td>8</td>
<td>14850</td>
</tr>
<tr>
<td>9</td>
<td>13053</td>
</tr>
<tr>
<td>10</td>
<td>13053</td>
</tr>
<tr>
<td>11</td>
<td>13068</td>
</tr>
<tr>
<td>12</td>
<td>13068</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-Sensitive</th>
<th>Sensitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zip Code</td>
<td>Age</td>
</tr>
<tr>
<td>1</td>
<td>130**</td>
</tr>
<tr>
<td>2</td>
<td>130**</td>
</tr>
<tr>
<td>3</td>
<td>130**</td>
</tr>
<tr>
<td>4</td>
<td>130**</td>
</tr>
<tr>
<td>5</td>
<td>1485*</td>
</tr>
<tr>
<td>6</td>
<td>1485*</td>
</tr>
<tr>
<td>7</td>
<td>1485*</td>
</tr>
<tr>
<td>8</td>
<td>1485*</td>
</tr>
<tr>
<td>9</td>
<td>130**</td>
</tr>
<tr>
<td>10</td>
<td>130**</td>
</tr>
<tr>
<td>11</td>
<td>130**</td>
</tr>
<tr>
<td>12</td>
<td>130**</td>
</tr>
</tbody>
</table>

**Figure 1. Inpatient Microdata**

**Figure 2. 4-anonymous Inpatient Microdata**

Source: Machanavajjhala et al., 2008
The Canonical 3-D Problem

Table: OfficeVisit

<table>
<thead>
<tr>
<th>v#</th>
<th>Patient</th>
<th>Doctor</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>David</td>
<td>Christy</td>
<td>Compoz</td>
</tr>
<tr>
<td>123</td>
<td>John</td>
<td>Phillips</td>
<td>Fungicide</td>
</tr>
<tr>
<td>124</td>
<td>Israel</td>
<td>Christy</td>
<td>AZT</td>
</tr>
<tr>
<td>125</td>
<td>John</td>
<td>Hill</td>
<td>Compoz</td>
</tr>
</tbody>
</table>

\[ x_{ijk} = \text{count of visits over} \]

- Patient \( i \) \( i = 1, \ldots, I \)
- Doctor \( j \) \( j = 1, \ldots, J \)
- Treatment \( k \) \( k = 1, \ldots, K \)
The “Third Projection Problem”  
(Chowdhury, Duncan, Krishnan, Roehrig, Mukherjee)

• Given two 2-D projections, find bounds on cell values of the third 2-D projection

• Example: Given **Patient-Doctor** and **Doctor-Treatment**, find bounds on the sensitive table **Patient-Treatment**
The Decomposed Network

Arcs represent “flows” of treatments from doctor to patient.

The network splits into three smaller subgraphs.

Patient-Treatment maxima and minima are derived from flow algorithms.

Results correspond to MCA.
Results: Two-D Projection Bounds

Let \( A = [a_{ij}] \), \( B = [b_{jk}] \) and \( C = [c_{ik}] \) be the two-dimensional projections of the three-dimensional table \( T = [t_{ijk}] \).

**Proposition:** It is not possible in general to determine the entries of \( C \) given those of \( A \) and \( B \).

**Proposition (MCA):** Optimal upper bounds for the third projection \( C = [c_{ik}] \) are given by

\[
C^U_{ik} = A \otimes B = \sum_j \min(a_{ij}, b_{jk}).
\]

Optimal lower bounds for \( C \) are given by

\[
C^L_{ik} = A \otimes B = \sum_j \max(a_{ij} - \sum_{p \neq k} b_{jp}, 0).
\]
## The Network Privacy Problem

### Variables (Data for Units Corresponding to Nodes)

<table>
<thead>
<tr>
<th>Units</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Adjacency Matrix Linking Nodes (1=link; 0=no link)

<table>
<thead>
<tr>
<th>Adjacency Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

SAMSI, October 20, 2010
Society as a Graph

People are represented as nodes.

Source of next 3 slides: Rao, 2009
Society as a Graph

People are represented as *nodes.*

Relationships are represented as *edges.*

(Relationships may be acquaintanceship, friendship, co-authorship, etc.)
Society as a Graph

People are represented as *nodes*.

Relationships are represented as *edges*.

(Relationships may be acquaintanceship, friendship, co-authorship, etc.)

Allows analysis using tools of mathematical graph theory
The problem

• Publish network data
  – Maximize utility from the data
  – Subject to confidentiality constraints
  – Anonymize the network
    – Naïve approach of anonymizing node labels does not work (Hay, 2010) based on assumption of some prior background knowledge
      – Degree signature attack
      – Degree signature of node and that of neighbors
      – Leading to node re-identification and edge disclosure
      – But good from the standpoint of analysis since topology is not altered
Karate Club network- Anonymized

Zachary, 1977
# Network mappings

<table>
<thead>
<tr>
<th>Id</th>
<th>Gender</th>
<th>Age</th>
<th>Belt</th>
<th>Injury</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>F</td>
<td>18</td>
<td>yellow</td>
<td>ankle</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>21</td>
<td>white</td>
<td>ankle</td>
</tr>
<tr>
<td>31</td>
<td>F</td>
<td>42</td>
<td>white</td>
<td>elbow</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>16</td>
<td>white</td>
<td>none</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>19</td>
<td>white</td>
<td>clavicle</td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>21</td>
<td>white</td>
<td>knee</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>21</td>
<td>white</td>
<td>none</td>
</tr>
<tr>
<td>26</td>
<td>M</td>
<td>36</td>
<td>black</td>
<td>none</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>22</td>
<td>brown</td>
<td>none</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>45</td>
<td>black</td>
<td>ribs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Edge</th>
<th>YearsKnown</th>
</tr>
</thead>
<tbody>
<tr>
<td>{16, 8}</td>
<td>2</td>
</tr>
<tr>
<td>{16, 2}</td>
<td>3</td>
</tr>
<tr>
<td>{6, 8}</td>
<td>10</td>
</tr>
<tr>
<td>{6, 2}</td>
<td>11</td>
</tr>
<tr>
<td>{31, 2}</td>
<td>20</td>
</tr>
<tr>
<td>{31, 9}</td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>16</td>
</tr>
<tr>
<td>Bob</td>
<td>6</td>
</tr>
<tr>
<td>Carol</td>
<td>31</td>
</tr>
<tr>
<td>Dave</td>
<td>11</td>
</tr>
<tr>
<td>Ed</td>
<td>9</td>
</tr>
<tr>
<td>Fred</td>
<td>21</td>
</tr>
<tr>
<td>Gail</td>
<td>13</td>
</tr>
<tr>
<td>Mr. Hi</td>
<td>26</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
But first, a network analysis discussion
Visualization Software: Krackplot

Sources:  http://www.andrew.cmu.edu/user/krack/krackplot/mitch-circle.html
          http://www.andrew.cmu.edu/user/krack/krackplot/mitch-anneal.html
Connections

• Size
  – Number of nodes

• Density
  – Number of ties that are present
  – The amount of ties that could be present

• Out-degree
  – Sum of connections from an actor to others

• In-degree
  – Sum of connections to an actor
Distance

• Walk
  – A sequence of actors and relations that begins and ends with actors

• Geodesic distance
  – The number of relations in the shortest possible walk from one actor to another

• Maximum flow
  – The amount of different actors in the neighborhood of a source that lead to pathways to a target
Some Measures of Power & Prestige
(based on Hanneman, 2001)

• Degree
  – Sum of connections from or to an actor
    • Transitive weighted degree ➔ Authority, hub, pagerank

• Closeness centrality
  – Distance of one actor to all others in the network

• Betweenness centrality
  – Number that represents how frequently an actor is between other actors’ geodesic paths
Cliques and Social Roles
(based on Hanneman, 2001)

• Cliques
  – Sub-set of actors
    • More closely tied to each other than to actors who are not part of the sub-set
      – (A lot of work on “trawling” for communities in the web-graph)
      – Often, you first find the clique (or a densely connected subgraph) and then try to interpret what the clique is about

• Social roles
  – Defined by regularities in the patterns of relations among actors
Statistical approaches to network analysis

• Markov Graph-based models
  – Exponential random graph-based models

• Permutation test and regression-based approaches
  – E.g., QAP regression variants due to David Krackhardt at Heinz
Example 1: Product adoption – CRBT

Caller ringback tones
Groups

N-cliques
Exponential Random Graphs

• Very general families for modeling a single static network observation.

\[ P(N) = \exp \left\{ \theta \cdot u(N) - \ln Z(\theta) \right\} \]

• Can estimate the \( \theta \) parameters by MCMC MLE

• \( N \) is a network vector, \( u(N) \) are a set of sufficient statistics to estimate the parameter theta of the model
ERGM Example: CRBT-purchase in a cell phone network

- Classic example: (Frank & Strauss 1986)
- Once model is estimated, it can be used to predict the likelihood that a link will form between node I and node J
  - \( u_1(N) = \# \text{ edges in } N \)
  - \( u_2(N) = \# \text{ 2-stars in } N \)
  - \( u_3(N) = \# \text{ triangles in } N \)

\[
P(N) \propto \exp\{\theta_1 u_1(N) + \theta_2 u_2(N) + \theta_3 u_3(N)\}\]
Example 2: Analyzing an Intra-organizational blogosphere
Background

• Study conducted on an employee-only technical forum in a “top 5” Indian IT service provider

• Web-based Forum intended to serve two purposes:
  – Transfer knowledge across employees in different ‘silos’ by allowing anyone to post responses to queries
  – Archive posted discussions or threads for subsequent retrieval
Sample Query

• Query on: Singleton class and threads in Java
• Responses:

1. Singleton class means that any given time only one instance of the class is present, in one JVM. So, it is present at JVM level.

2. The thing is if two users (on two different machines which has separate JVMs) are requesting for singleton class then both can get one-one-one instance of that class in their JVM.
Sample data posting of query and responses

<table>
<thead>
<tr>
<th>threadid</th>
<th>associateid</th>
<th>postedtime</th>
<th>messagetype</th>
<th>subject</th>
<th>message</th>
</tr>
</thead>
<tbody>
<tr>
<td>{20070110}</td>
<td>138242</td>
<td>2007-01-10 06:41:15</td>
<td>Query</td>
<td>Panel Creation in REXX</td>
<td>&lt;p&gt;Hi,&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>122971</td>
<td>2007-01-10 07:42:54</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt;For retaining the input panel&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>107246</td>
<td>2007-01-10 13:20:24</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt; You are not creating the&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>128623</td>
<td>2007-01-17 07:19:18</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt; No need to VPUT you can&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>129498</td>
<td>2007-03-01 12:31:42</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt;it's simple .. if var1 var2 are the&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>107246</td>
<td>2007-03-01 13:49:16</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt;TYPE(INPUT) is to define the&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>125034</td>
<td>2007-04-14 07:17:32</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt;You can use the command&lt;/p&gt;</td>
</tr>
<tr>
<td>{20070110}</td>
<td>107246</td>
<td>2007-04-14 23:43:30</td>
<td>Response</td>
<td>Re: Panel Creation in REXX</td>
<td>&lt;p&gt; &lt;em&gt;&lt;strong&gt;ADDRESS&lt;/p&gt;</td>
</tr>
</tbody>
</table>
Data description

• Message level and thread-level data from forum

• Message characteristics
  – Posting time, EmployeeID, Thread, Type of message (query or response), content of message etc.

• User characteristics
  – EmployeeID, Tenure at firm, Age, Gender, Location, Division, Job Title

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of users participating</td>
<td>2974</td>
</tr>
<tr>
<td>Total number of queries</td>
<td>20090</td>
</tr>
<tr>
<td>Total number of responses</td>
<td>59038</td>
</tr>
<tr>
<td>Average responses per query</td>
<td>2.9</td>
</tr>
<tr>
<td>Average messages per day</td>
<td>162</td>
</tr>
<tr>
<td>Average time to first response</td>
<td>58 min</td>
</tr>
<tr>
<td>Number of users only posting queries</td>
<td>343</td>
</tr>
<tr>
<td>Number of users only posting responses</td>
<td>1377</td>
</tr>
<tr>
<td>Number of users posting queries and responses</td>
<td>1004</td>
</tr>
</tbody>
</table>
Network structure evolution

Sequence of Actions:

- User 301 posts a query Q1000
- Users 502, 641 post responses
- User 900 posts a query Q1001
- Users 301, 641 post responses

Directed Response Graph
Simmelian Ties

Why should this matter? Theory

1908: Simmel’s argument that Triads are different from Dyads, but adding more does not matter

Triads form Groups, with Norms, Rules, Values, Common Understandings, Pressure towards Compliance, Conformity and Cooperation
**Simmelian Decomposition:** Each network tie can be characterized as one of three mutually exclusive and exhaustive types:

- **Asymmetric:** \((a \rightarrow b)\)  

- **Sole-Symmetric:** \((a \rightarrow b) \land (b \rightarrow a)\)  

  ... but not Simmelian

- **Simmelian:**  
  \[ (a \rightarrow b) \land (b \rightarrow a) \land \exists c \text{ s.t. } (c \rightarrow a) \land (a \rightarrow c) \land (c \rightarrow b) \land (b \rightarrow c) \]

Each tie here is Simmelian.
Research Question

• Is the probability of response to a question posed by a node I contingent on the network structure that the node is embedded in?
  – Simmelian tie
  – Symmetric tie
  – Asymmetric tie

• Does the nature of the question (popular or not) which determines the context within which the tie was established make a difference?
Construction of Variables

Response Matrix

\[
\text{RESPONSES} = \begin{pmatrix}
0 & 0 & 1 & 1 & 0 \\
8 & 1 & 0 & 1 & 1 \\
7 & 0 & 1 & 4 & 0 \\
0 & 1 & 0 & 1 & 1 \\
0 & 2 & 0 & 1 & 1
\end{pmatrix},
\]

\( \text{RESPONSES}_{i,j} = \) number of times 'i' responds to 'j'

Directed Response Graph

1 → 2
1 → 3
1 → 4
1 → 5
2 → 1
2 → 3
2 → 4
2 → 5
3 → 1
3 → 2
3 → 4
3 → 5
4 → 1
4 → 2
4 → 3
4 → 5
5 → 1
5 → 2
5 → 4
5 → 5
Construction of Variables

\[
SIMMELIAN = \begin{pmatrix}
X & 0 & 0 & 0 & 0 \\
0 & X & 0 & 1 & 1 \\
0 & 0 & X & 0 & 0 \\
0 & 1 & 0 & X & 1 \\
0 & 1 & 0 & 1 & X
\end{pmatrix},
\]

\[
SIMMELIAN_{i,j} = 1 \text{ if 'i' and 'j' have a Simmelian tie}
\]
Construction of Variables

\[
\text{NON - SIMMELIAN} = \begin{pmatrix}
X & 0 & 1 & 1 & 0 \\
1 & X & 0 & 0 & 0 \\
1 & 0 & X & 1 & 0 \\
0 & 0 & 0 & X & 0 \\
0 & 0 & 0 & 0 & X
\end{pmatrix},
\]

\[
\text{NON - SIMMELIAN}_{i,j} = 1 \text{ if 'i' and 'j' have a non-Simmelian tie}
\]
Construction of Variables

Age difference

\[ \text{ABS}_\text{AGEDIFF} = \begin{pmatrix}
    0 & 18 & 26 & 28 & 40 \\
    18 & 0 & 8 & 10 & 22 \\
    26 & 8 & 0 & 2 & 14 \\
    28 & 10 & 2 & 0 & 12 \\
    40 & 22 & 14 & 12 & 0 \\
\end{pmatrix}, \]

\[ \text{ABS}_\text{AGEDIFF}_{i,j} = \text{absolute value of age difference between 'i' and 'j' (months)} \]
Construction of Variables

Locations Color Coded

\[
\text{SAMELOCATION} = \begin{pmatrix}
X & 0 & 1 & 0 & 1 \\
0 & X & 0 & 0 & 0 \\
1 & 0 & X & 0 & 1 \\
0 & 0 & 0 & X & 0 \\
1 & 0 & 1 & 0 & X
\end{pmatrix},
\]

\[
\text{SAMELOCATION}_{i,j} = 1 \text{ if 'i' and 'j' are collocated}
\]
Construction of Variables

**Verticals Color Coded**

\[
\text{SAMEVERTICAL} = \begin{pmatrix}
X & 1 & 0 & 0 & 0 \\
1 & X & 0 & 0 & 0 \\
0 & 0 & X & 0 & 1 \\
0 & 0 & 0 & X & 0 \\
0 & 0 & 1 & 0 & X
\end{pmatrix},
\]

\[
\text{SAMEVERTICAL}_{i,j} = 1 \text{ if 'i' and 'j' are part of the same vertical}
\]
Empirical Methodology

Want to characterize response behavior due to:

- Homophily
- Content
- Prior Network structure

Cannot use ordinary least squares regression

- Autocorrelation induced because of structural factors
  - Some users may respond more to all others

- Unbiased, but significance tests will be incorrect

- Use QAP (Quadratic Assignment Procedure) to test for significance
  - Krackhardt (1987) - reference
QAP - Regression

• Variant of QAP (Double Semi-Partialing)
  – Dekker et al (Psychometrika, 2007)
• Divide the data into two periods
  – P1 – Aug 2006 – Feb 2007
• Dependent variable is
  – Number of responses by A to B in period two
• Explanatory Variables:
  – Structural Properties in period one
  – Dyadic Homophily Measures
QAP Regression Specification

**Dependent Variable**
- \( Y_t \) = Number of responses from A to B in period ‘t’

**Independent variables**
- Abs(Difference between age)
- Abs(difference between tenure),
- Same location city dummy,
- Same vertical dummy,
- Number of queries posted,
- Structural Factors:
  - Simmelian and Non-simmelian of responses to:
    - (a) Low SP (Non-instrumental) threads
    - (b) High SP (Instrumental) threads
Dyadic QAP Regression Results

**Dependent variable:**
Number of response by A to B in period two
Dyadic QAP Regression Results

**Dependent variable:**
Number of response by A to B in period two

**Explanatory Variables:**
Dyadic Homophily Measures, *Structural Properties in period one*
Dyadic QAP Regression Results

*Dependent variable:* 
Number of response by A to B in period two

*Explanatory Variables:* 
Dyadic Homophily Measures, *Structural Properties in period one*

---

**Graph 1:** Instrumental Historical Context
- **Non-Simmelian** and **Simmelian** categories

**Graph 2:** Non Instrumental Historical Context
- **Non-Simmelian** and **Simmelian** categories

---
Dyadic QAP Regression Results

**Dependent variable:**
Number of response by A to B in period two

**Explanatory Variables:**
Dyadic Homophily Measures, *Structural Properties in period one*

![Graphs showing instrumental and non-instrumental historical context for Simmelian and non-Simmelian pairs.](image-url)
Dyadic QAP Regression Results

**Dependent variable:**
Number of response by A to B in period two

**Explanatory Variables:**
Dyadic Homophily Measures, **Structural Properties in period one**

![Graphs showing the comparison between instrumental and non-instrumental contexts for Simmelian and Non-Simmelian interactions in both current and historical data.](image)
Dyadic QAP Regression Results

*Dependent variable:*
Number of response by A to B in period two

*Explanatory Variables:*
Dyadic Homophily Measures, *Structural Properties in period one*

![Graphs showing Instrumental Historical Context and Non Instrumental Historical Context with bars for Non-Simmelian and Simmelian categories.](image-url)
Other iLab network data sets

• Reliance Telecom
  – 2009; 6 months, 3 million customers, Call Data Records and Caller Ring Back Tones Purchases; Mumbai
  – 2010; 4 months, 1 million customers, call data records and caller ring back tones purchase behavior; Pune

• Vodafone Portugal
  – 1 years worth of data from Portugal
  – Call data records, churn behavior and telecom service purchase behavior

• All data have individual level attributes and data about network behaviors like with the blog data set
Back to privacy and networks

• Recent interest in the literature
  – Hay et al., 2010 is a good review paper

• Focus on attacks on network data and attempts to anonymize the network through addition of “noise”
  – Directed alteration
  – Random alteration
  – generalization
k-degree anonymity

• The kind of attack
  – Vertex Refinement Queries

• Objective
  – The published graph
    • For every node \( v \), there exist at least \( k-1 \) other nodes in the graph with the same degree as \( v \)
    • Choose the number of edges that are added to achieve k-degree anonymity subject to minimally affecting the graph’s topology (more about this later)

• Approach
  – Add edges into the original anonymized graph to meet k-degree constraint
K-neighbor anonymity to prevent sub-graph attacks

(a) the social network

(b) the network with anonymous nodes

(c) the 1-neighborhood graph of Ada

(d) privacy-preserved anonymous network
Figure 8: Examples of random alteration applied to the anonymized karate club network of Figure 2. The original edges are shown in light gray. With random alteration, edges are randomly chosen and rewired until $m - 10$ edges have been altered (deleted edges in dashed black; inserted edges in solid black).

Hay, 2010
Generalization

Hay, 2010
**Exponential Random Graphs**

- Very general families for modeling a single static network observation.

\[
P(N) = \exp \left\{ \theta \cdot u(N) - \ln Z(\theta) \right\}
\]

- Can estimate the \( \theta \) parameters by MCMC MLE

- \( N \) is a network vector, \( u(N) \) are a set of sufficient statistics to estimate the parameter theta of the model
ERGM Example: CRBT-purchase in a cell phone network

• Classic example: (Frank & Strauss 1986)
• Once model is estimated, it can be used to predict the likelihood that a link will form between node I and node J
  – $u_1(N) = \# \text{ edges in } N$
  – $u_2(N) = \# \text{ 2-stars in } N$
  – $u_3(N) = \# \text{ triangles in } N$

\[
P(N) \propto \exp\{\theta_1 u_1(N) + \theta_2 u_2(N) + \theta_3 u_3(N)\} \]
Addition of noise affects inference

• Note that in ERGM models sufficient statistics that are inputs to parameters estimation are number of edges, number of 2-stars and number of triangles
  – All of these would be affected when edges are added to anonymize the network

• Similar problem with QAP regression since all the dyadic variables will be affected
Open problem

• Design a scalable anonymization technique that can be used to publish social network data such that it minimally affects the sufficient statistics used for parameter estimation of network statistics models