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

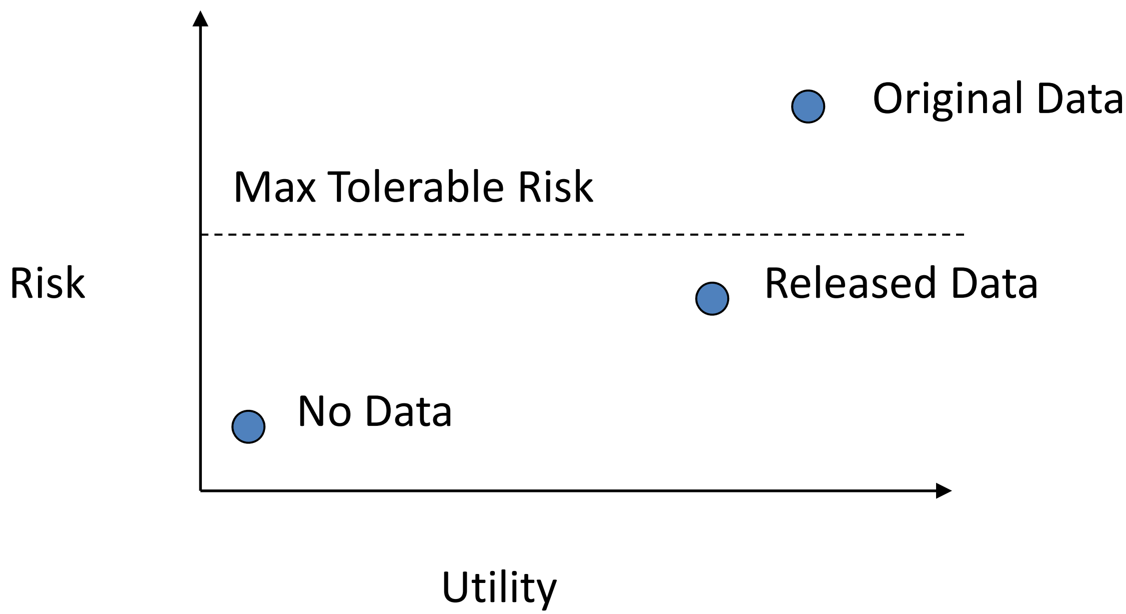
- Introduction
  - The R-U framework
  - The traditional data privacy approaches
- Networks
  - Analysis using networks
    - Knowledge management example
- Privacy in Networks
  - Why is it complicated?
  - How does privacy protection affect analysis/inference?
  - Interesting open problem

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)

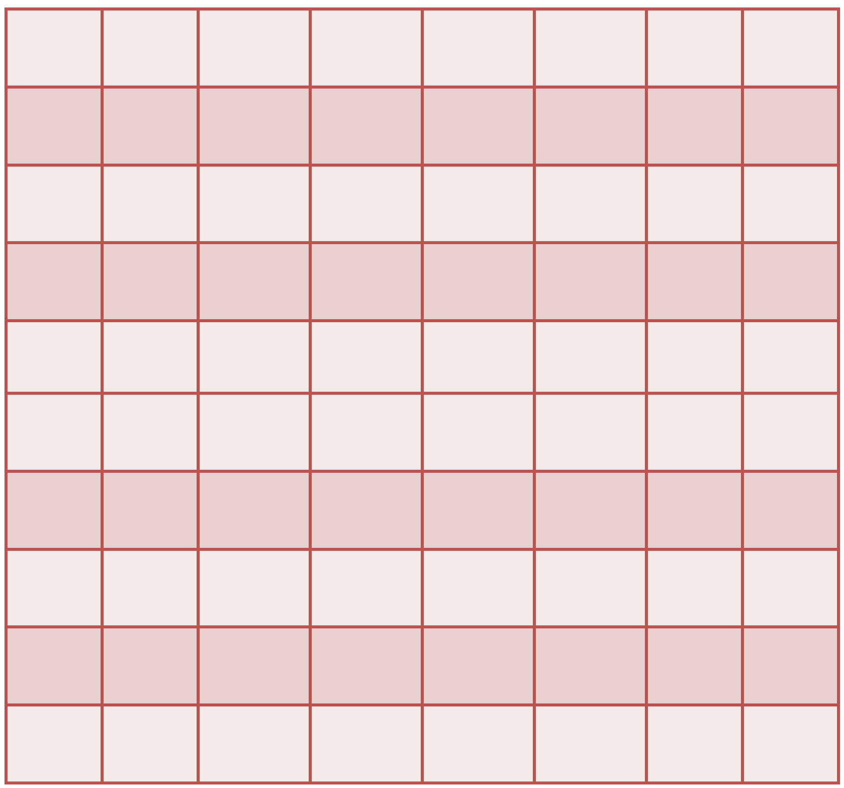


Utility –example 1: Inverse of the RMSE of the estimate of a statistic such as the sample Mean
example 2: sum of tuple information loss criterion
Risk – example 1: Width of the interval at a specified confidence level of value of a Confidential variable that will lead to re-identification; example 2: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

The Standard Privacy Problem

Variables

Units



“Solutions”:

- Deleting cases
- Aggregating cases
- Deleting variables
- Adding noise
- perturbations
    - K-anonymity
    - L-diversity

Micro-data: an example

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

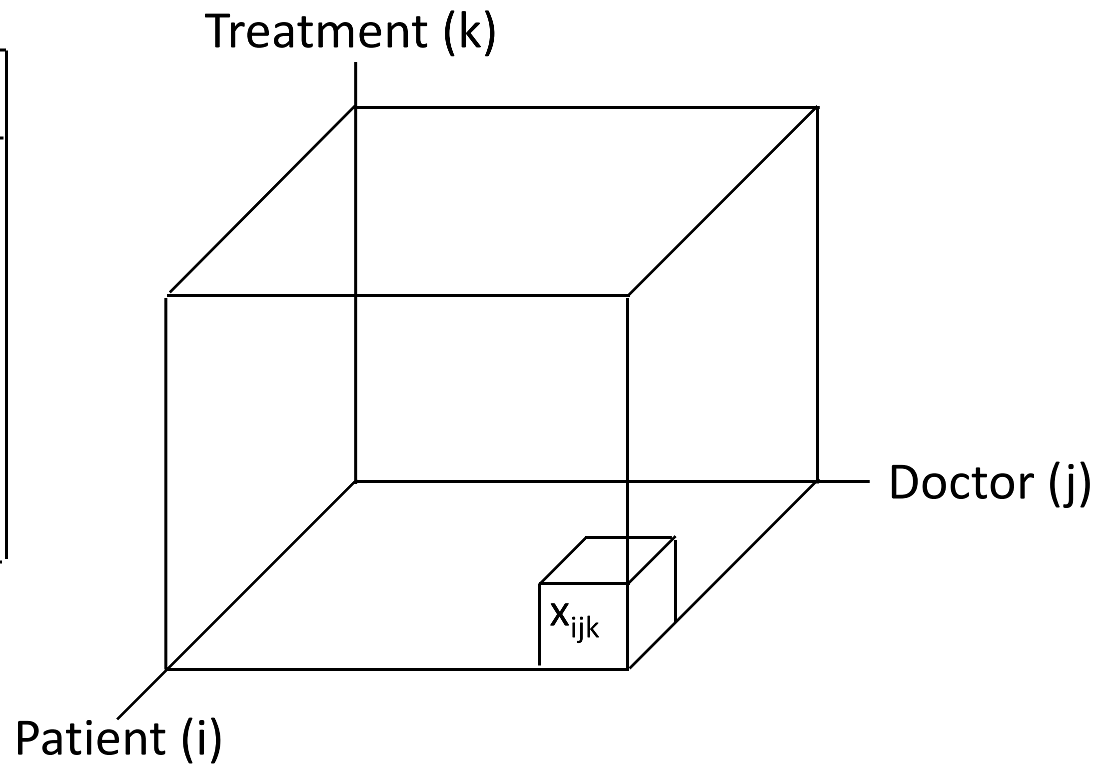
Figure 2. 4-anonymous Inpatient Microdata

Source: Machanavajjhala et al., 2008

The Canonical 3-D Problem

Table: OfficeVisit

v#	Patient	Doctor	Treatment
122	David	Christy	Compoz
123	John	Phillips	Fungicide
124	Israel	Christy	AZT
125	John	Hill	Compoz
:	:	:	:



x_{ijk} = count of visits over

Patient i $\quad i = 1,\ldots,I$

Doctor j $\quad j = 1,\ldots,J$

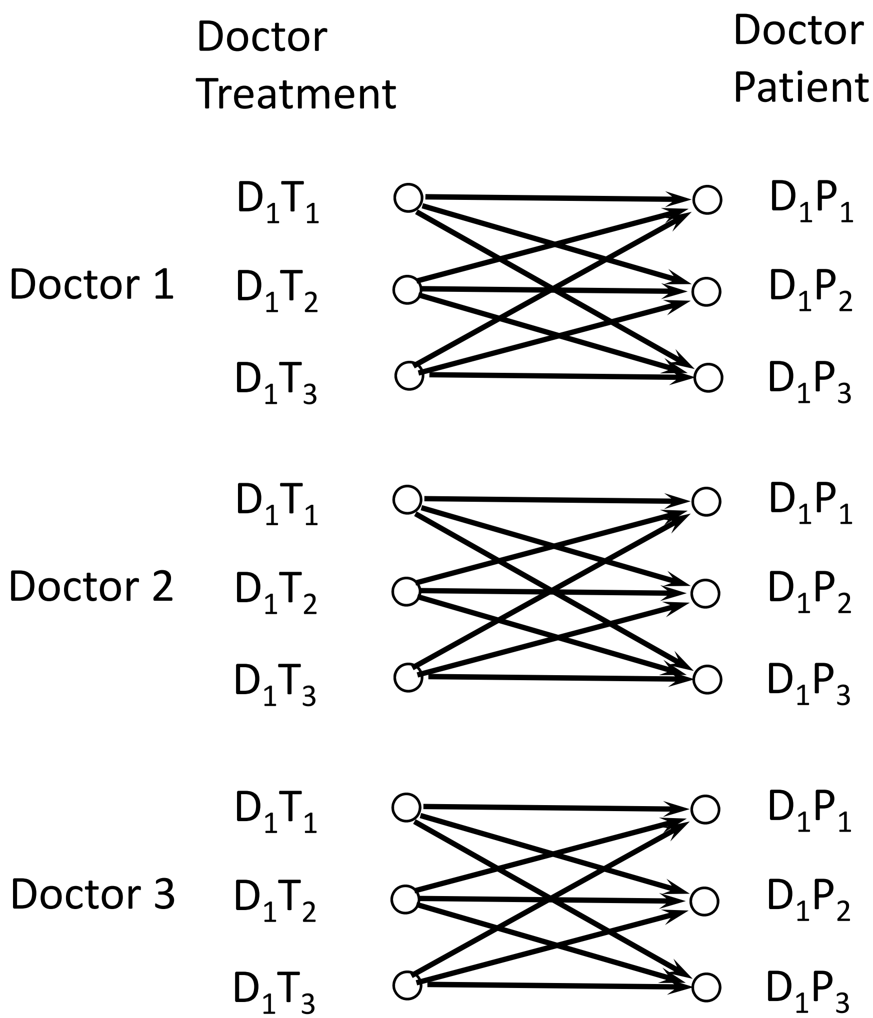
Treatment k $\quad k = 1,\ldots,K$

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

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

Optimal lower bounds for C are given by

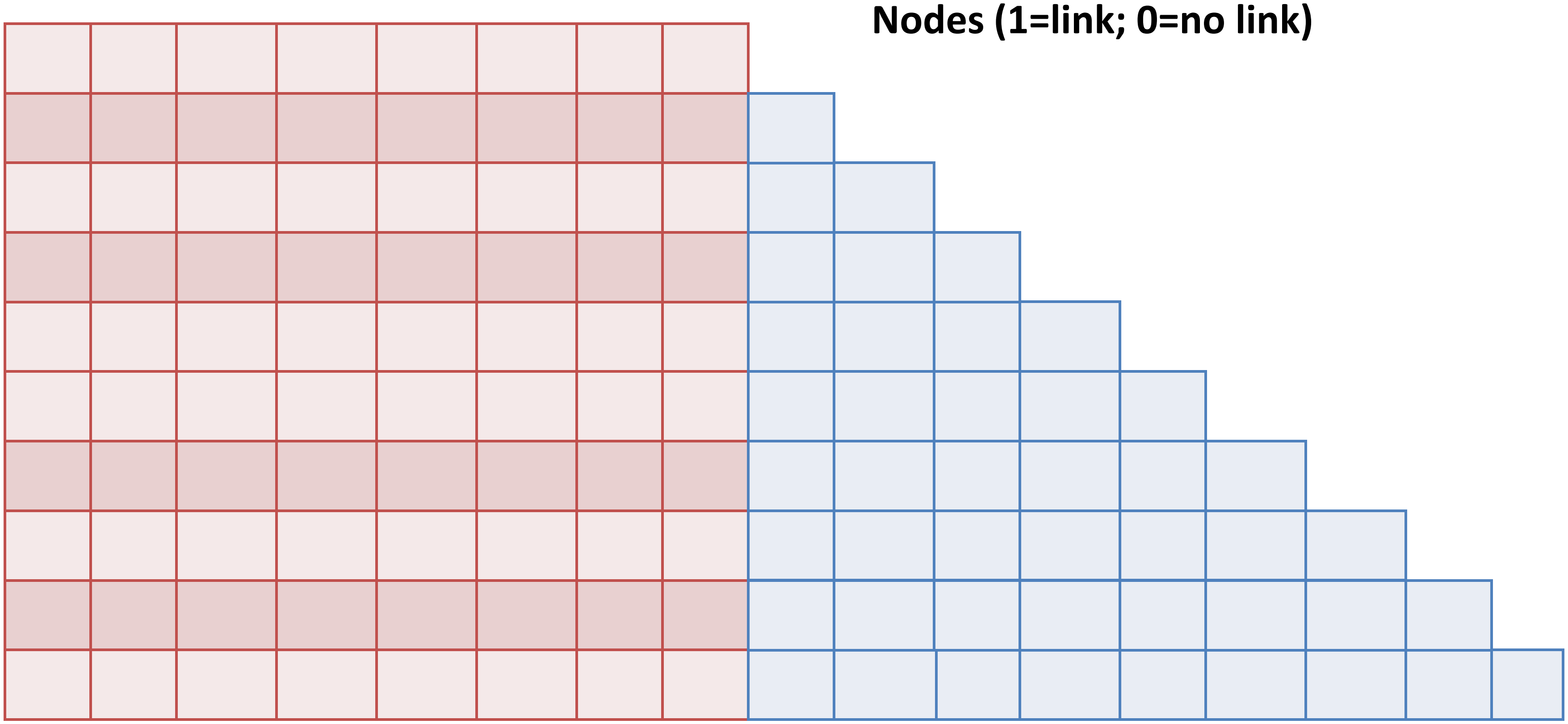
$$C^L_{ik} = A \underline{\otimes} B = \sum_j \max(a_{ij} - \sum_{p \neq k} b_{jp}, 0).$$

The Network Privacy Problem

Variables (Data for Units Corresponding to Nodes)

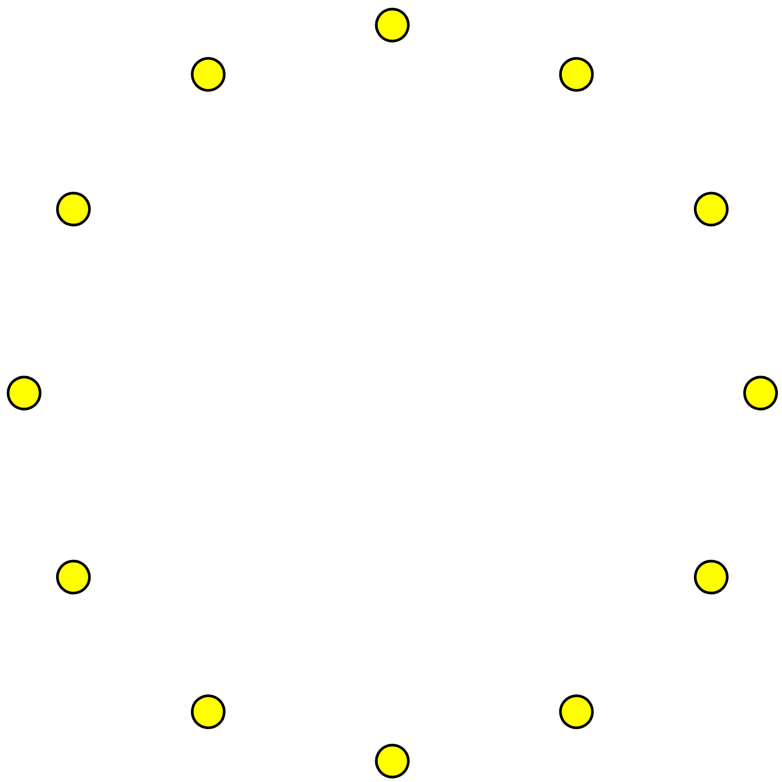
Adjacency Matrix Linking Nodes (1=link; 0=no link)

Units



Society as a Graph

People are represented as *nodes.*



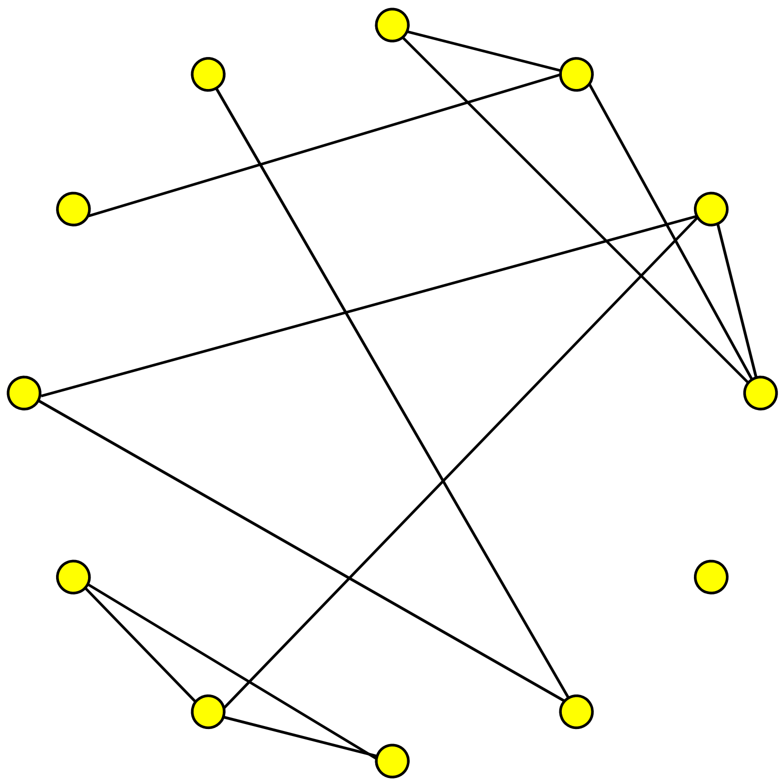
Source of next 3 slides: Rao, 2009

Society as a Graph

People are represented as *nodes.*

Relationships are represented as *edges.*

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



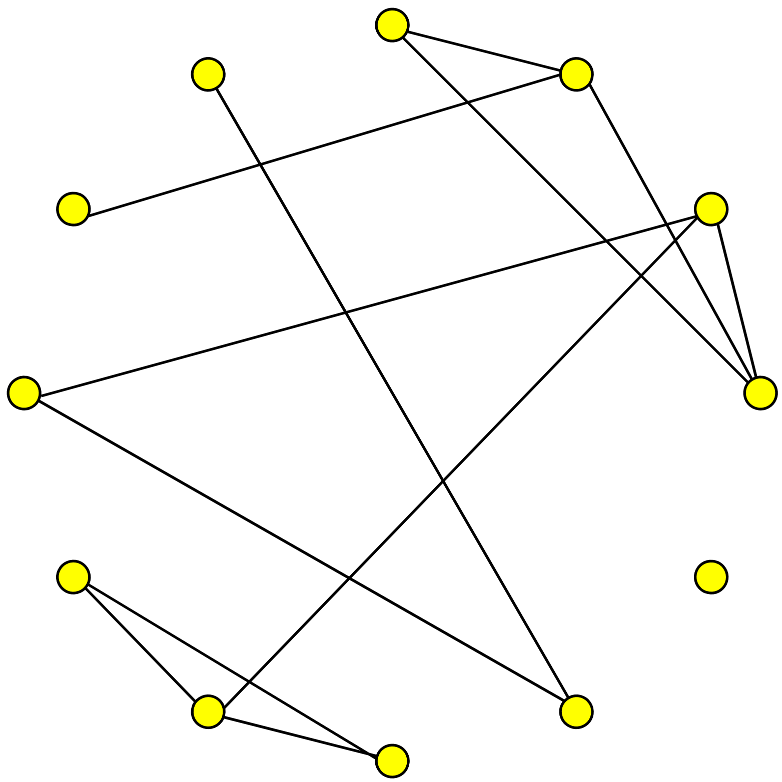
Society as a Graph

People are represented as *nodes.*

Relationships are represented as *edges.*

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

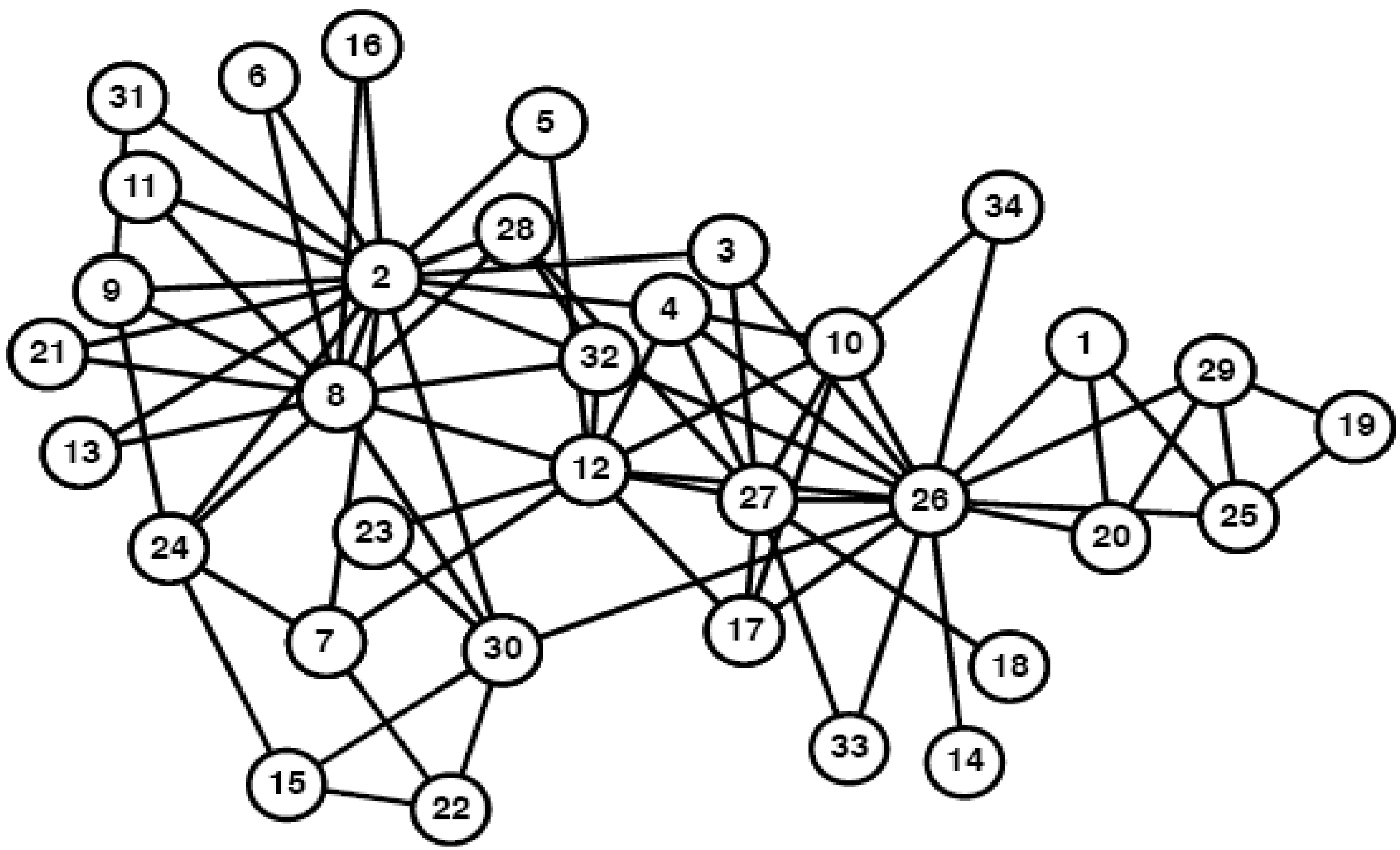
Allows analysis using tools of mathematical graph theory



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

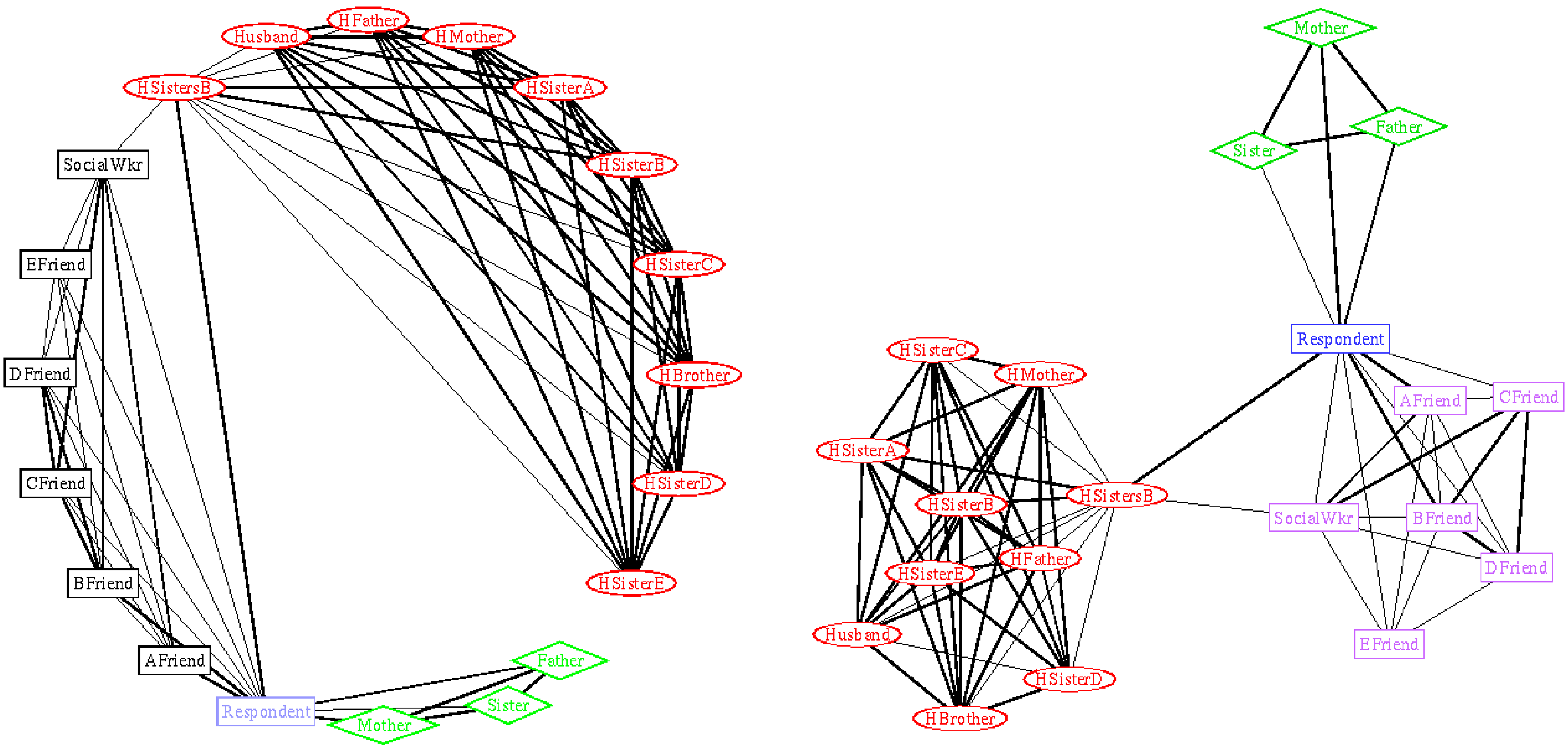
Id	Gender	Age	Belt	Injury
16	F	18	yellow	ankle
6	M	21	white	ankle
31	F	42	white	elbow
11	M	16	white	none
9	M	19	white	clavicle
21	M	21	white	knee
13	F	21	white	none
26	M	36	black	none
8	F	22	brown	none
2	M	45	black	ribs
		...		

Edge	YearsKnown
{ 16, 8 }	2
{ 16, 2 }	3
{ 6, 8 }	10
{ 6, 2 }	11
{ 31, 2 }	20
{ 31, 9 }	1
...	

Alice	**16**
Bob	**6**
Carol	**31**
Dave	**11**
Ed	**9**
Fred	**21**
Gail	**13**
Mr. Hi	**26**
...	**...**

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)

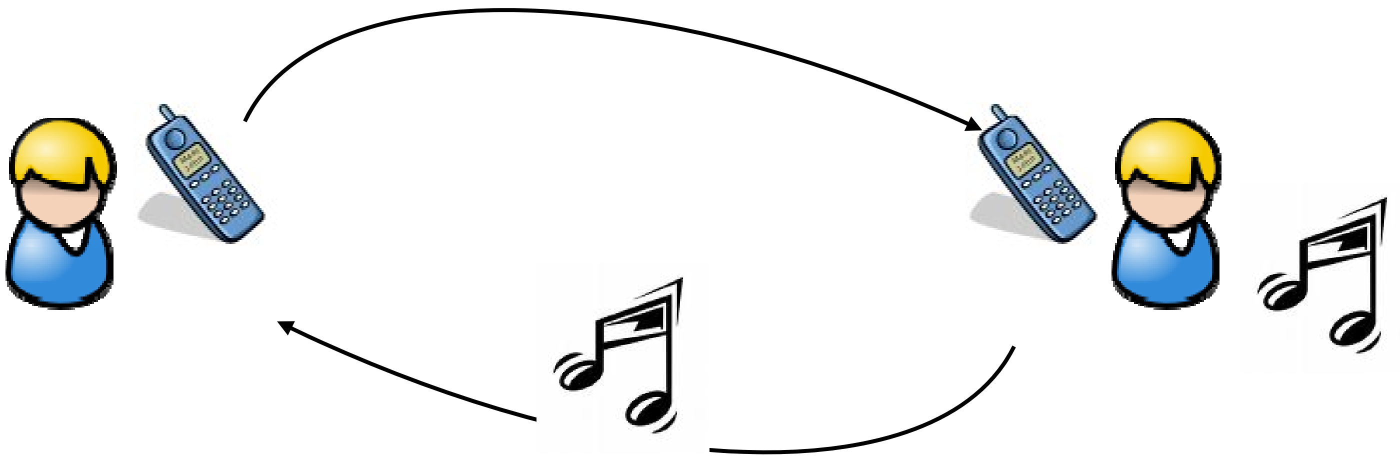
- Cliques
  - Sub-set of actors
    - More closely tied to each other than to actors who are not part of the sub-set
      - (A lot of work on “trawling” for communities in the web-graph)
      - Often, you first find the clique (or a densely connected subgraph) and then try to interpret what the clique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

- Social roles
  - Defined by regularities in the patterns of relations among actors

Statistical approaches to network analysis

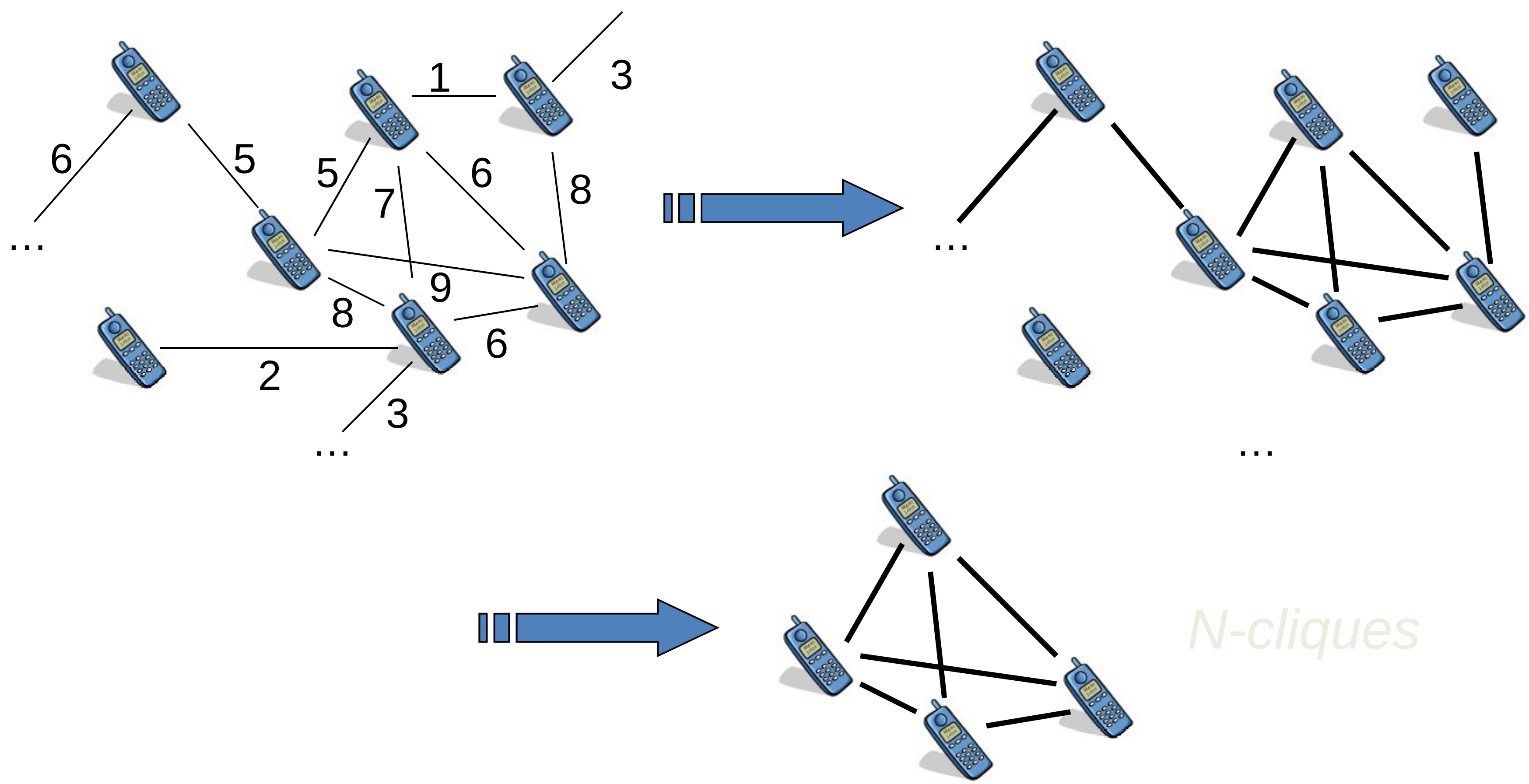
- Markov Graph-based models
  - Exponential random graph-based models
- Permutation test and regression-based approaches
  - E..g, QAP regression variants due to David Krackhardt at Heinz

Example 1: Product adoption – CRBT



Caller ringback tones

Groups



Exponential Random Graphs

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

$$P(N) = \exp\{\theta \cdot u(N) - \ln Z(\theta)\}$$

- Can estimate the θ 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
  - $u_1(N)$ = # edges in N
  - $u_2(N)$ = # 2-stars in N
  - $u_3(N)$ = # 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 instance of that class in their JVM.

Sample data posting of query and responses

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

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

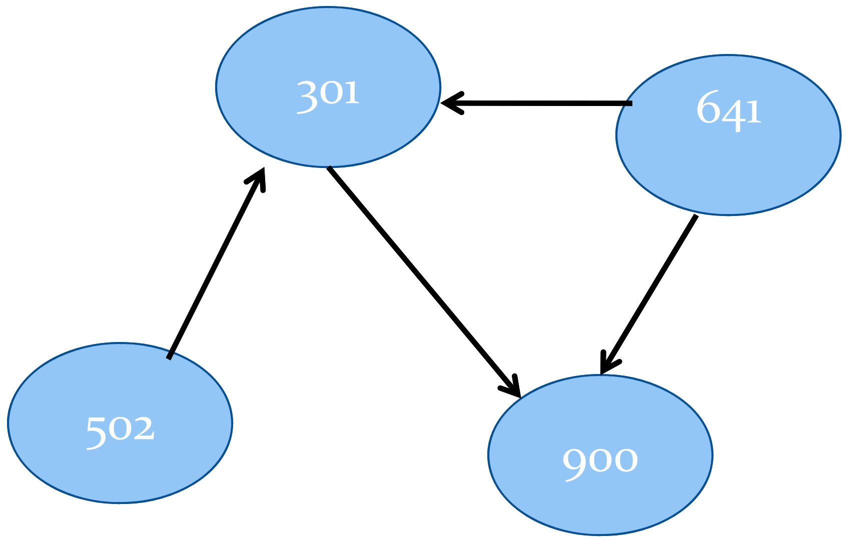
Summary statistics of forum (8/2006-8/2007)

Statistic	Value
Total number of users participating	2974
Total number of queries	20090
Total number of responses	59038
Average responses per query	2.9
Average messages per day	162
Average time to first response	58 min
Number of users only posting queries	343
Number of users only posting responses	1377
Number of users posting queries and responses	1004

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*

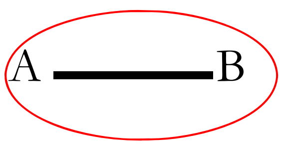


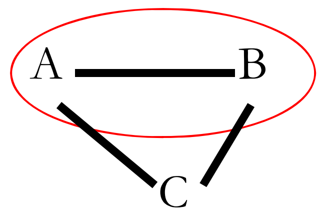
Directed Response Graph

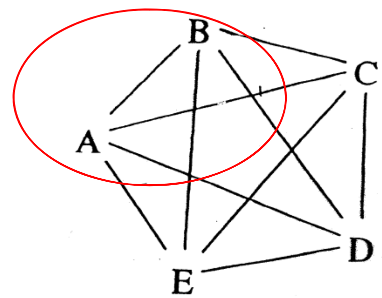
Simmelian Ties

Why should this matter? Theory

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



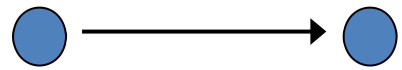




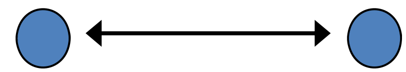
Triads form *Groups*, with Norms, Rules, Values, Common Understandings, Pressure towards Compliance, Conformity and Cooperation

Simmelian Decomposition: Each network tie can be characterized as one of three mutually exclusive and exhaustive types:

Asymmetric: $(a \rightarrow b)$



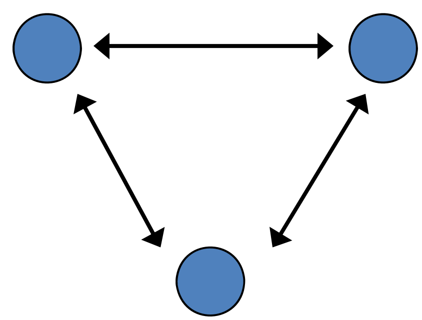
Sole-Symmetric: $(a \rightarrow b) \wedge (b \rightarrow a)$



… but not Simmelian

Simmelian: $(a \rightarrow b) \wedge (b \rightarrow a)$ *and*

$$\exists\, c \;\; s.t.\; (c \rightarrow a) \wedge (a \rightarrow c) \wedge (c \rightarrow b) \wedge (b \rightarrow c)$$



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?

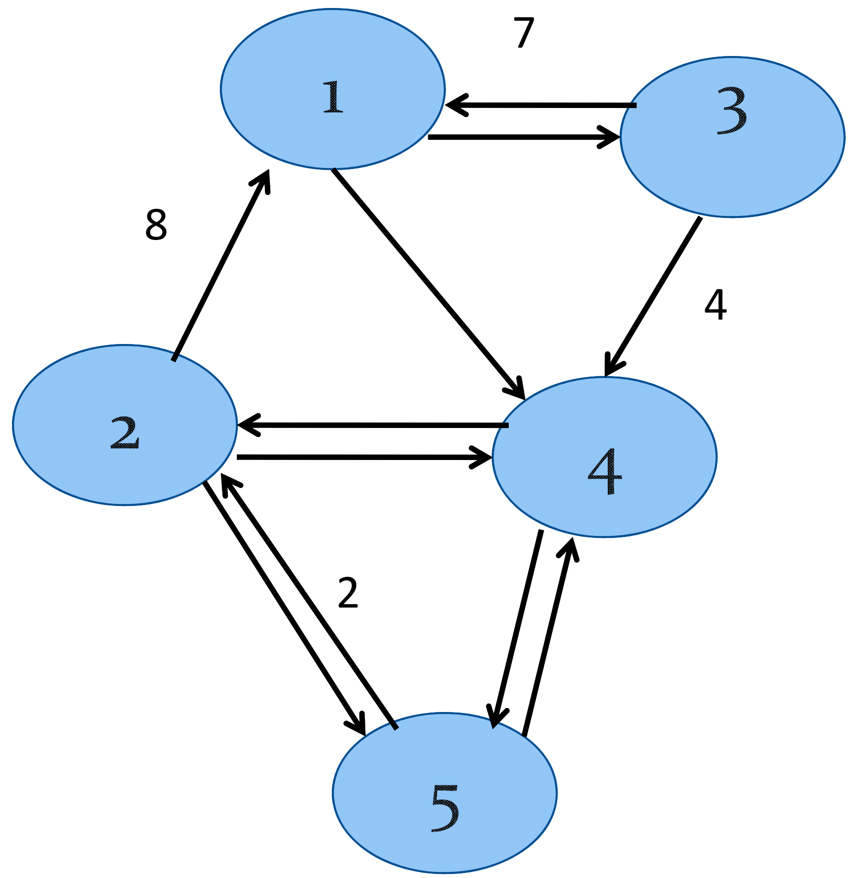
Construction of Variables

Response Matrix

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

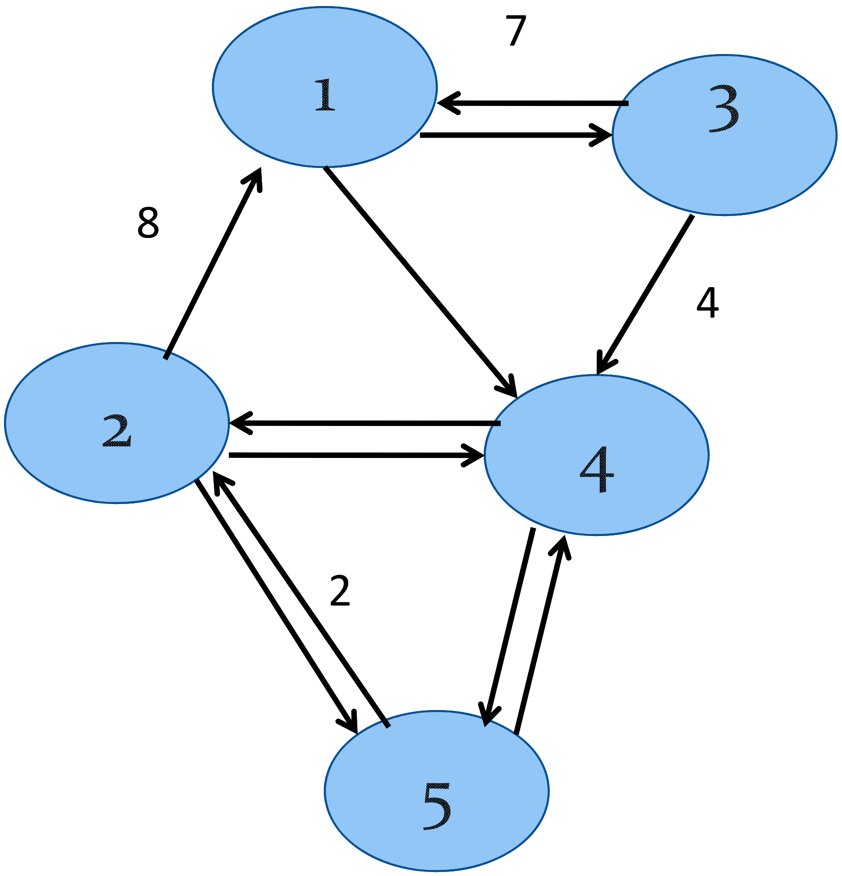
$$RESPONSES_{i,j} = \text{number of times 'i' responds to 'j'}$$

Directed Response Graph



Construction of Variables

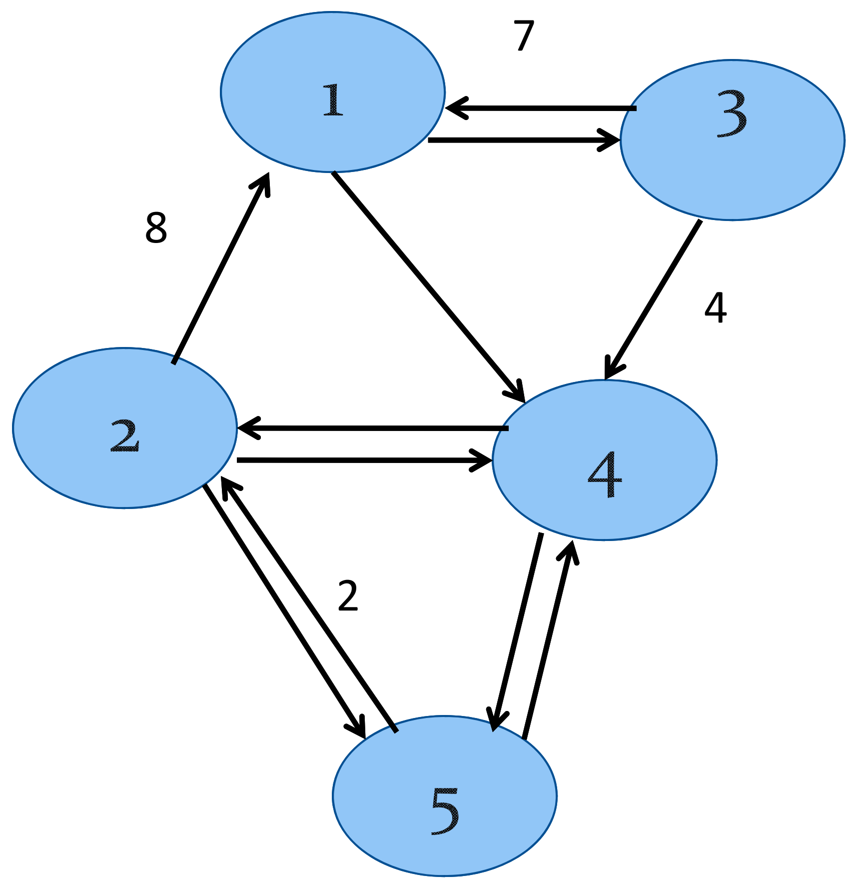
$$SIMMELIAN = \begin{pmatrix} X & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 1 & 1 \\ 0 & 0 & X & 0 & 0 \\ 0 & 1 & 0 & X & 1 \\ 0 & 1 & 0 & 1 & X \end{pmatrix},$$



$$SIMMELIAN_{i,j} = \text{ 1 if 'i' and 'j' have a Simmelian tie}$$

Construction of Variables

$$NON-SIMMELIAN = \begin{pmatrix} X & 0 & 1 & 1 & 0 \\ 1 & X & 0 & 0 & 0 \\ 1 & 0 & X & 1 & 0 \\ 0 & 0 & 0 & X & 0 \\ 0 & 0 & 0 & 0 & X \end{pmatrix},$$

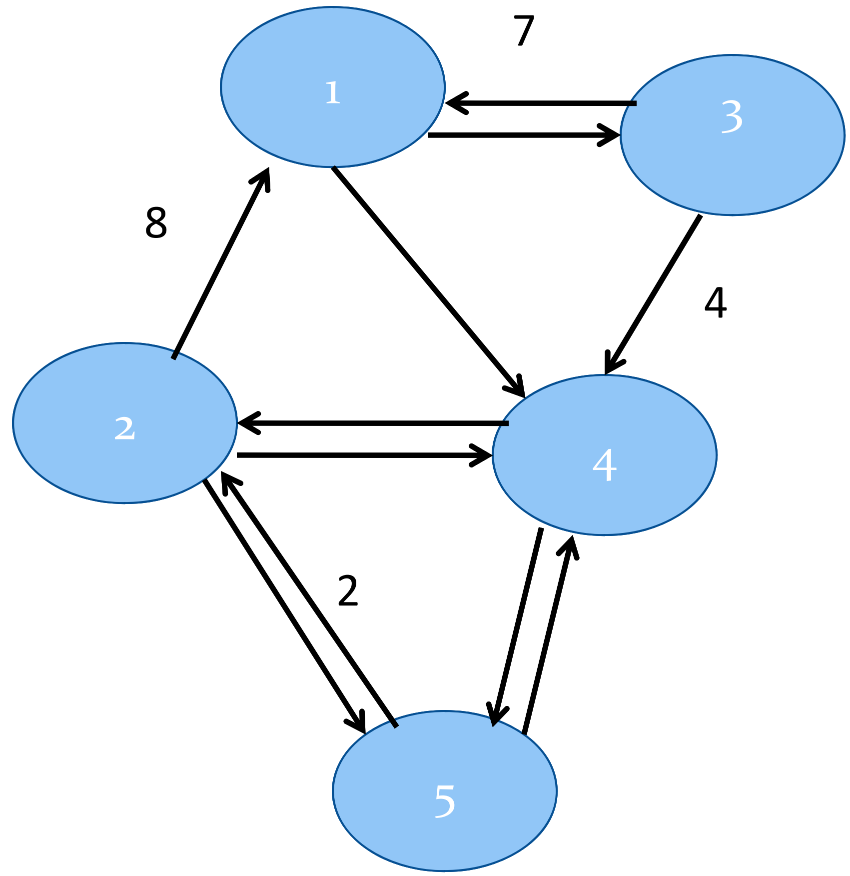


$$NON-SIMMELIAN_{i,j} = \text{ 1 if 'i' and 'j' have a non-Simmelian tie}$$

Construction of Variables

Age difference

$$ABS\_AGEDIFF = \begin{pmatrix} 0 & 18 & 26 & 28 & 40 \\ 18 & 0 & 8 & 10 & 22 \\ 26 & 8 & 0 & 2 & 14 \\ 28 & 10 & 2 & 0 & 12 \\ 40 & 22 & 14 & 12 & 0 \end{pmatrix},$$



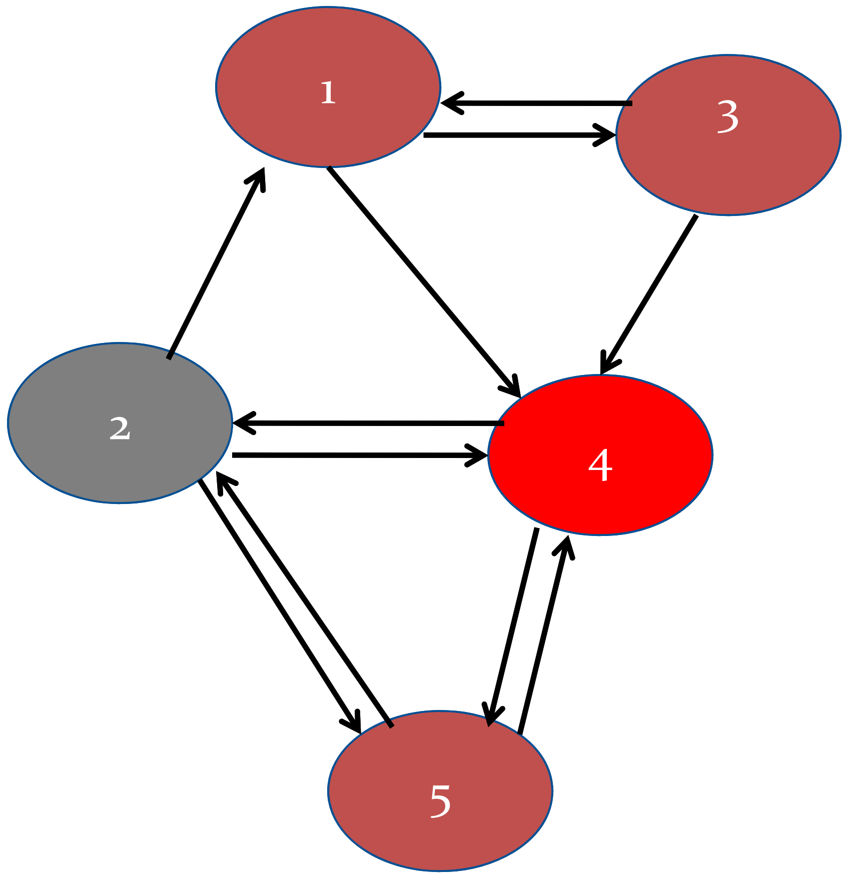
$$ABS\_AGEDIFF_{i,j}$$

$$= \text{ absolute value of age difference between 'i' and 'j' (months)}$$

Construction of Variables

Locations Color Coded

$$SAMELOCATION = \begin{pmatrix} X & 0 & 1 & 0 & 1 \\ 0 & X & 0 & 0 & 0 \\ 1 & 0 & X & 0 & 1 \\ 0 & 0 & 0 & X & 0 \\ 1 & 0 & 1 & 0 & X \end{pmatrix},$$

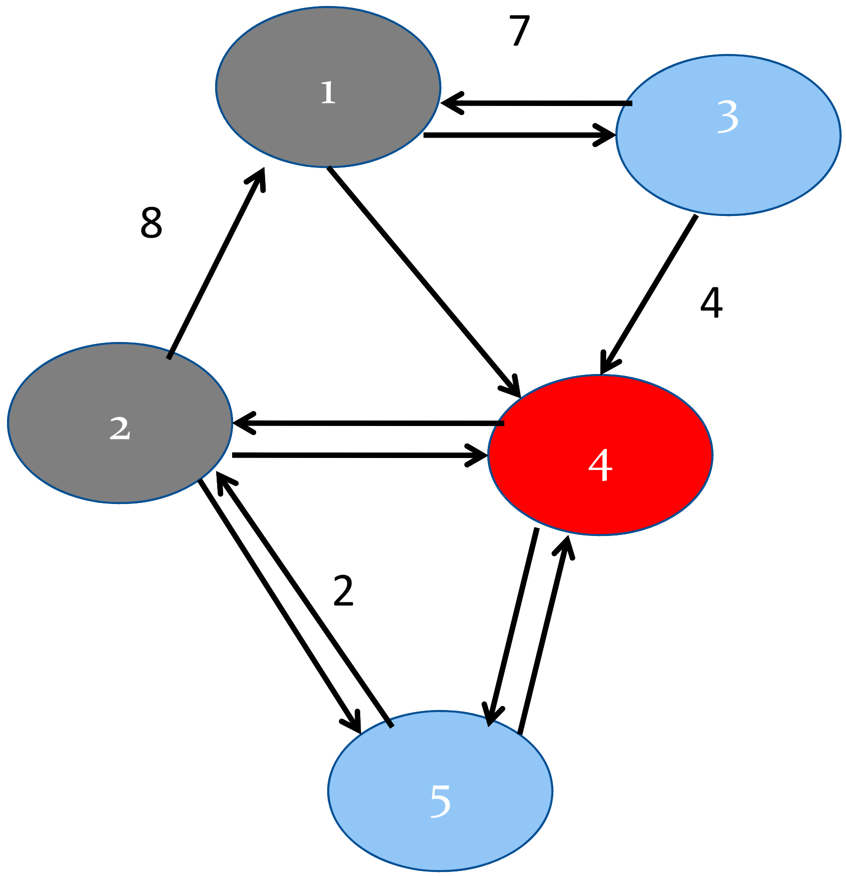


$$SAMELOCATION_{i,j} = 1 \text{ if 'i' and 'j' are collocated}$$

Construction of Variables

Verticals Color Coded

$$SAMEVERTICAL = \begin{pmatrix} X & 1 & 0 & 0 & 0 \\ 1 & X & 0 & 0 & 0 \\ 0 & 0 & X & 0 & 1 \\ 0 & 0 & 0 & X & 0 \\ 0 & 0 & 1 & 0 & X \end{pmatrix},$$



$$SAMEVERTICAL_{i,j} = 1 \text{ if 'i' and 'j' are part of the same vertical}$$

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
  - P2 – Feb 2007 – Aug 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:

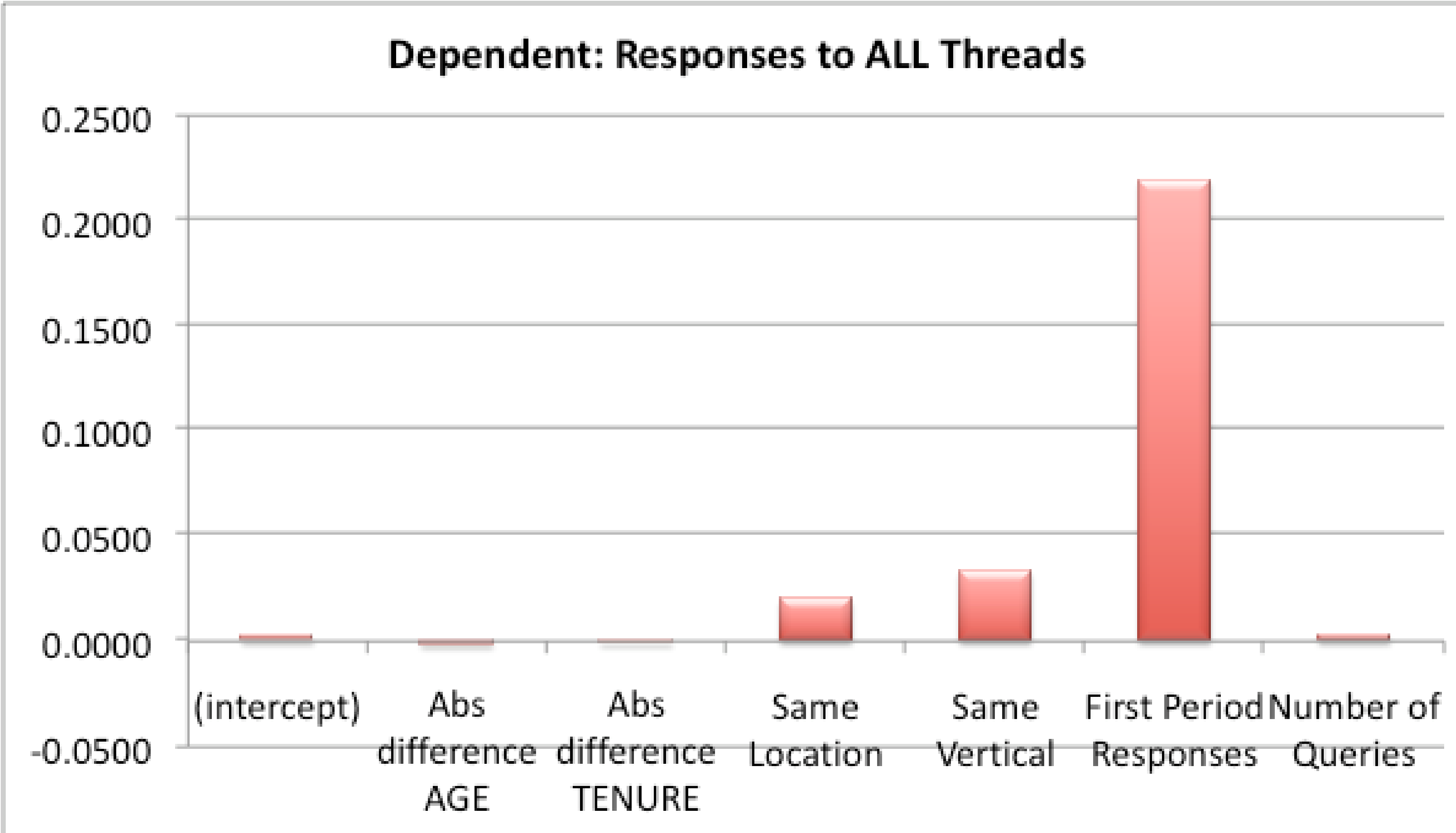
  (a)Low SP (Non-instrumental) threads

  (b)High SP (Instrumental) threads

Dyadic QAP Regression Results

Dependent variable:

Number of response by A to B in period two



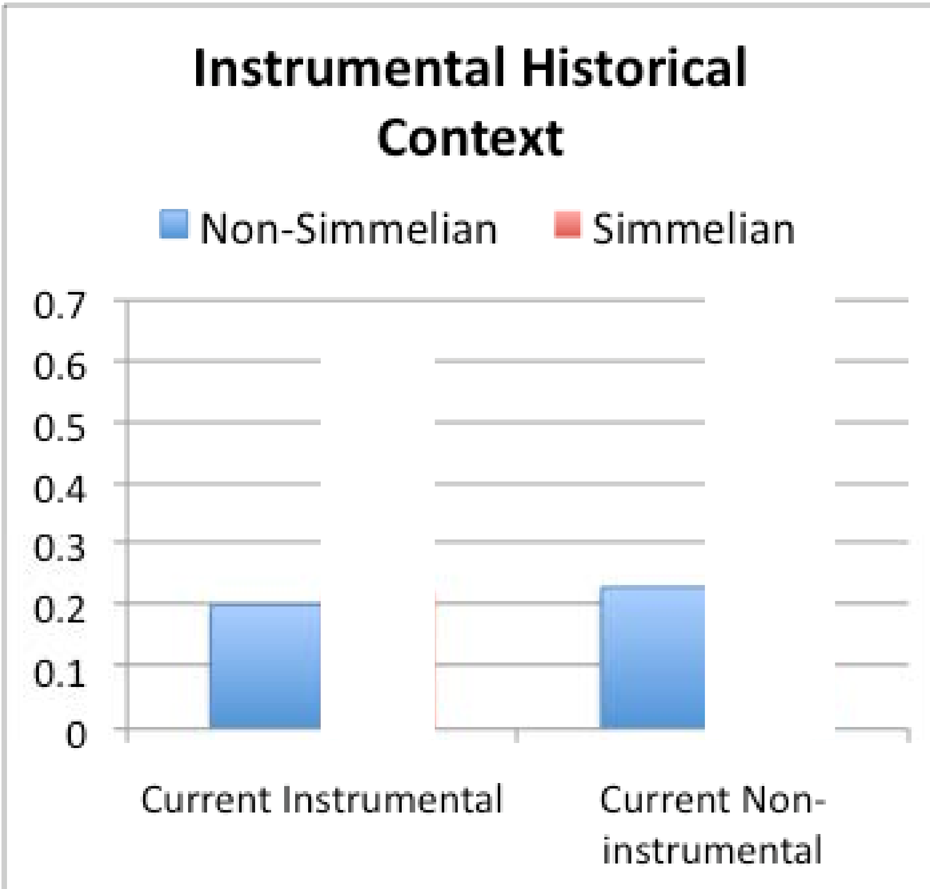
Dyadic QAP Regression Results

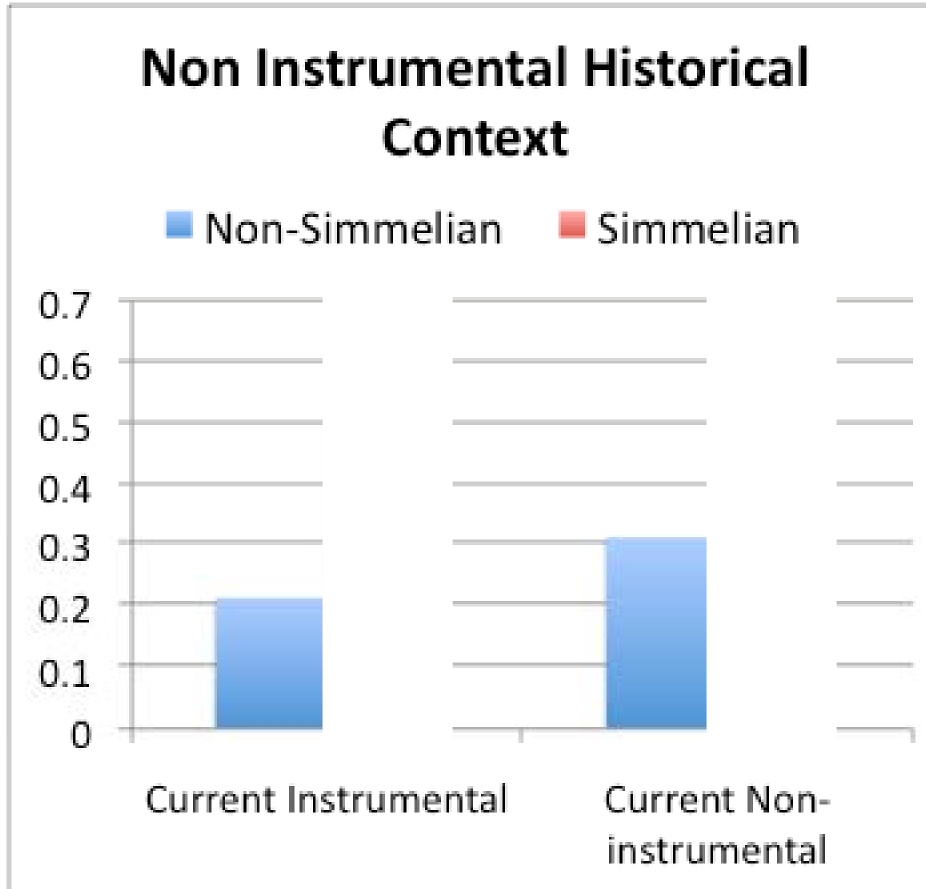
Dependent variable:

Number of response by A to B in period two

Explanatory Variables:

Dyadic Homophily Measures, **Structural Properties in period one**





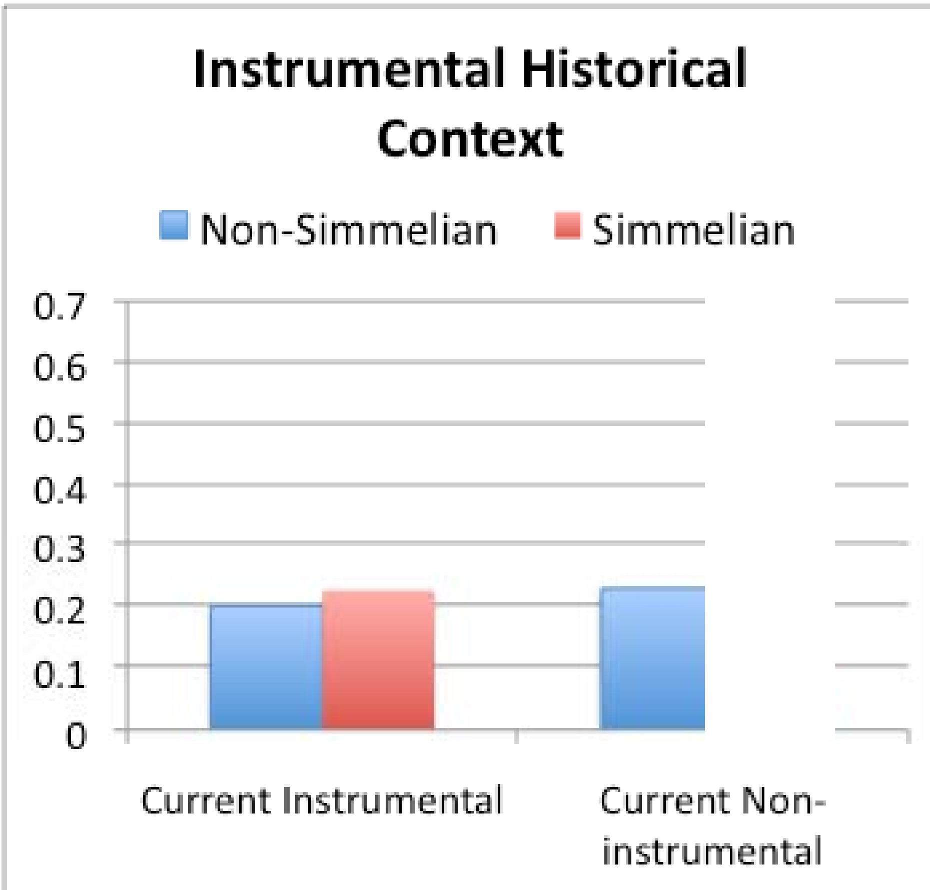
Dyadic QAP Regression Results

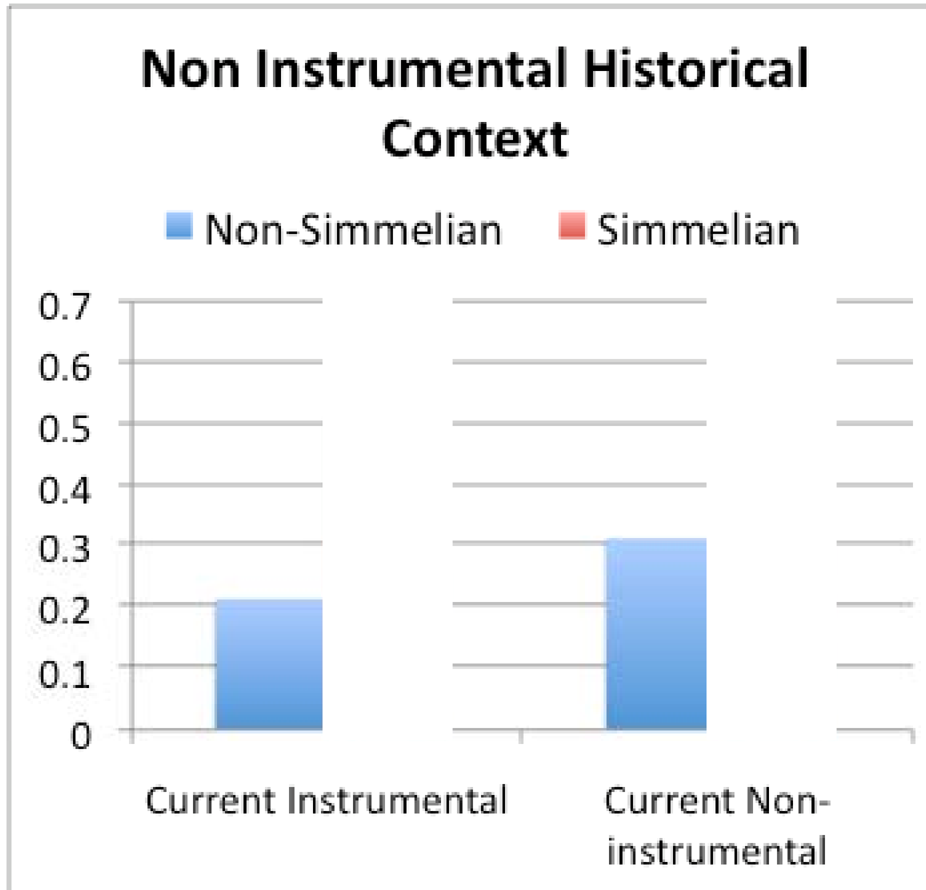
Dependent variable:

Number of response by A to B in period two

Explanatory Variables:

Dyadic Homophily Measures, **Structural Properties in period one**





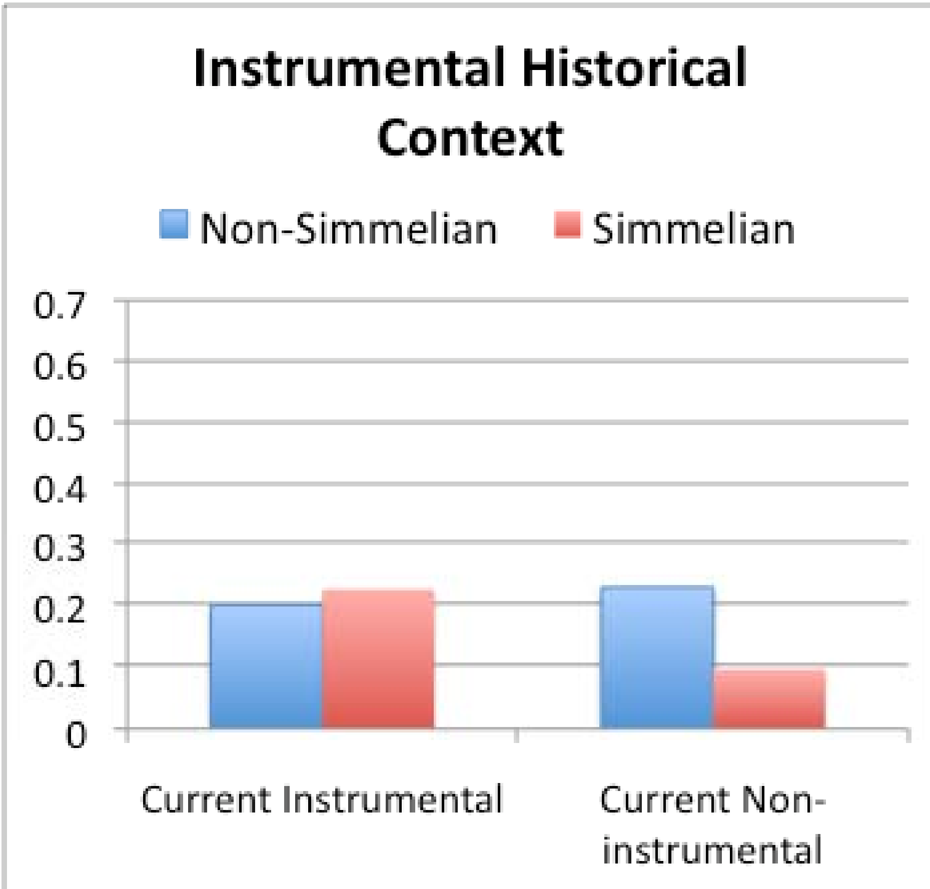
Dyadic QAP Regression Results

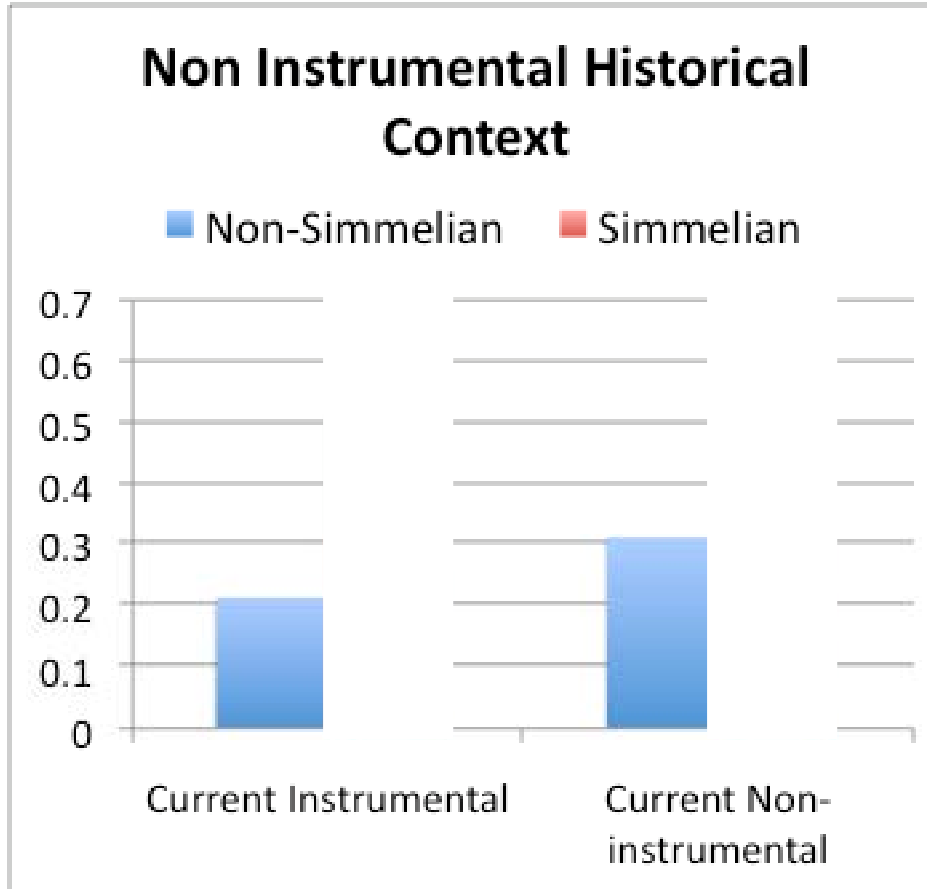
Dependent variable:

Number of response by A to B in period two

Explanatory Variables:

Dyadic Homophily Measures, **Structural Properties in period one**





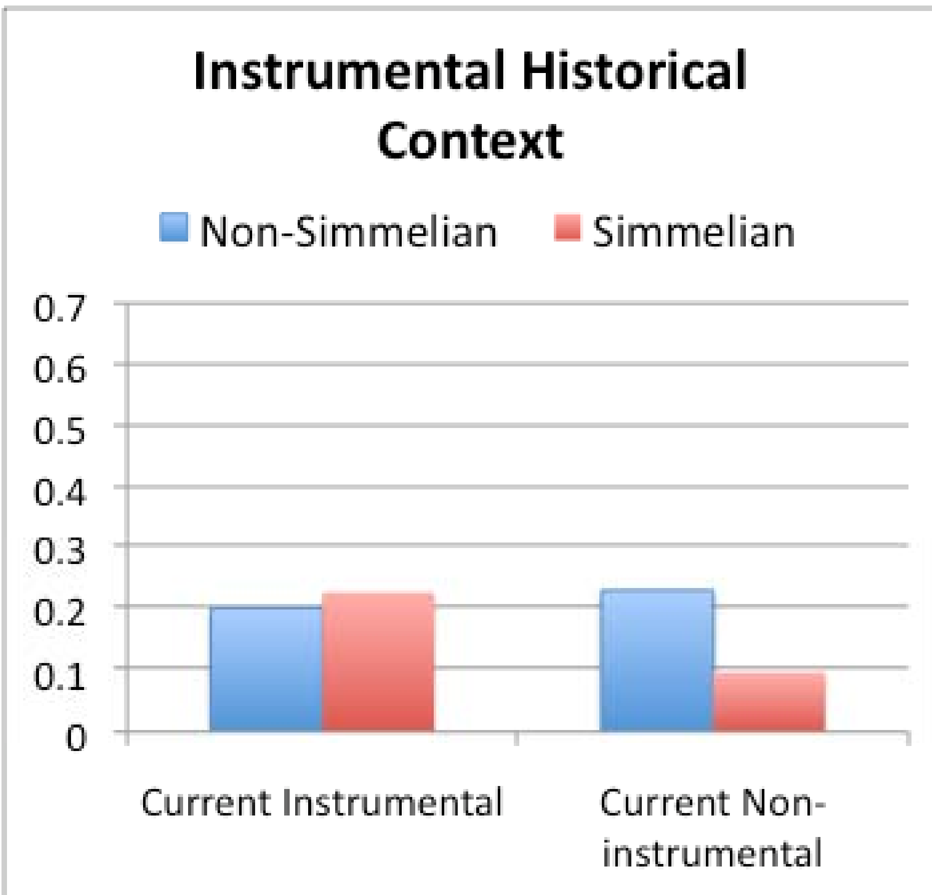
Dyadic QAP Regression Results

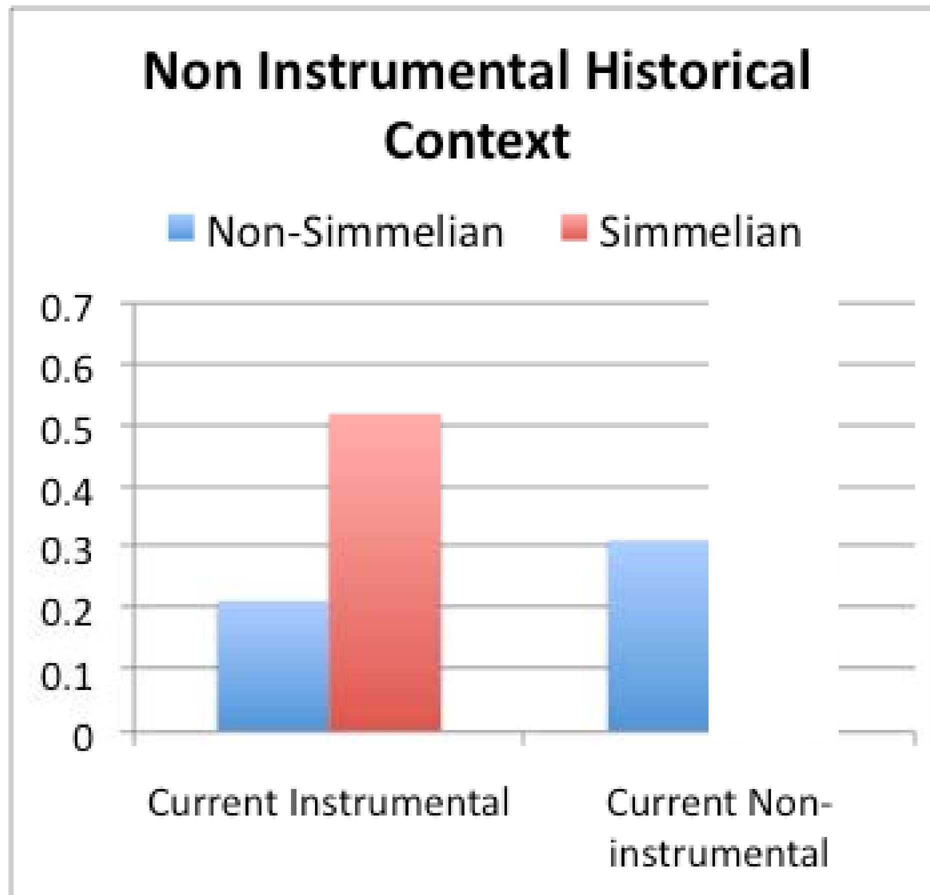
Dependent variable:

Number of response by A to B in period two

Explanatory Variables:

Dyadic Homophily Measures, **Structural Properties in period one**





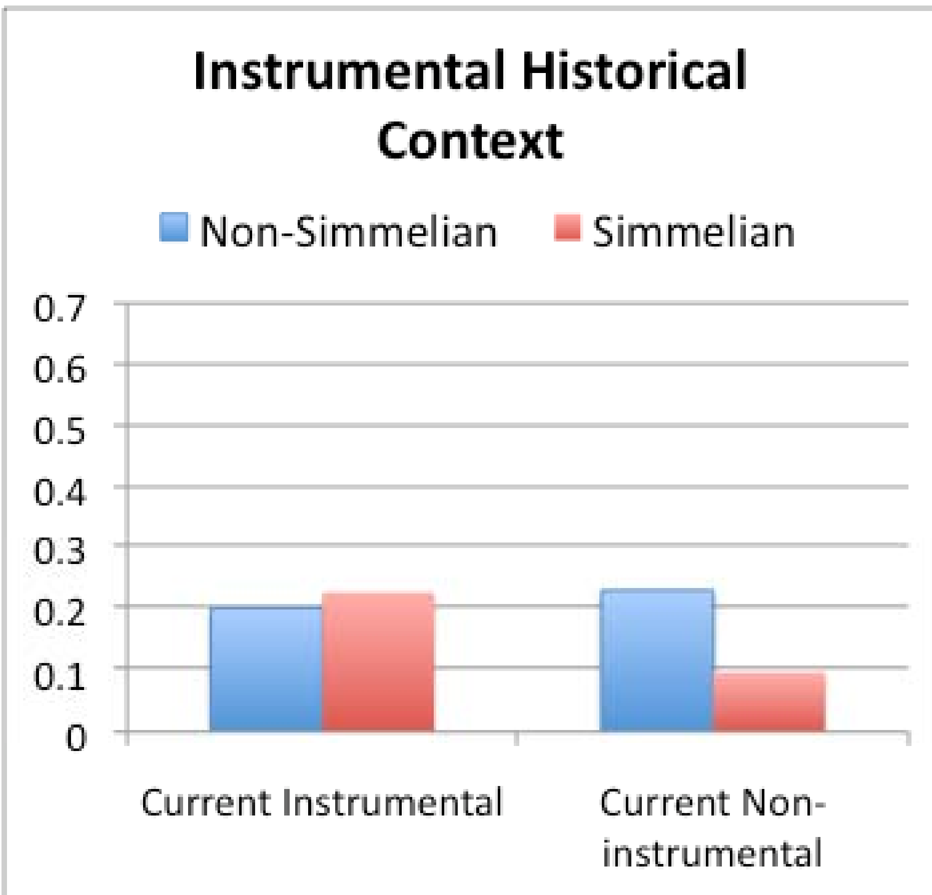
Dyadic QAP Regression Results

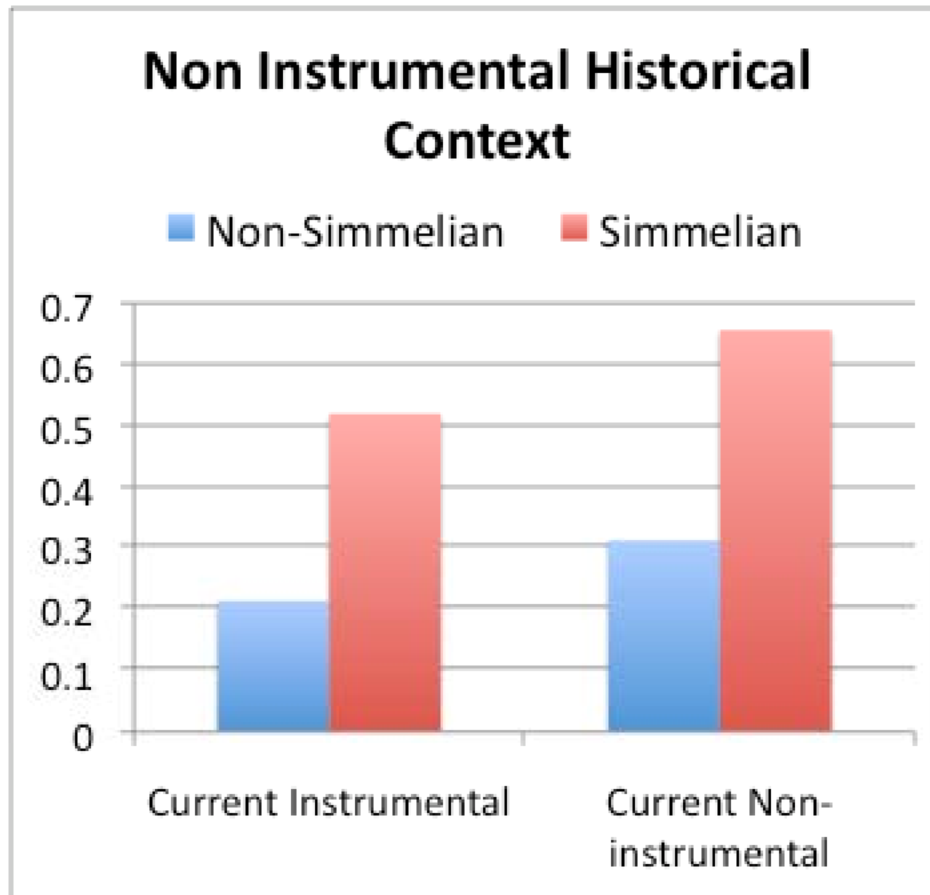
Dependent variable:

Number of response by A to B in period two

Explanatory Variables:

Dyadic Homophily Measures, **Structural Properties in period one**





Other iLab network data sets

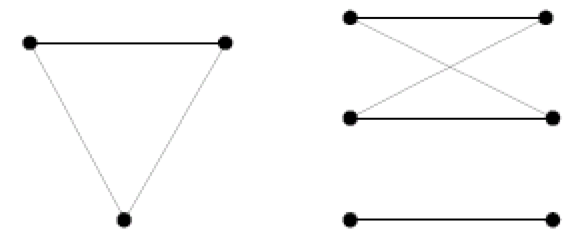
- Reliance Telecom
  - 2009; 6 months, 3 million customers, Call Data Records and Caller Ring Back Tones Purchases; Mumbai
  - 2010; 4 months, 1 million customers, call data records and caller ring back tones purchase behavior; Pune
- Vodafone Portugal
  - 1 years worth of data from Portugal
  - Call 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

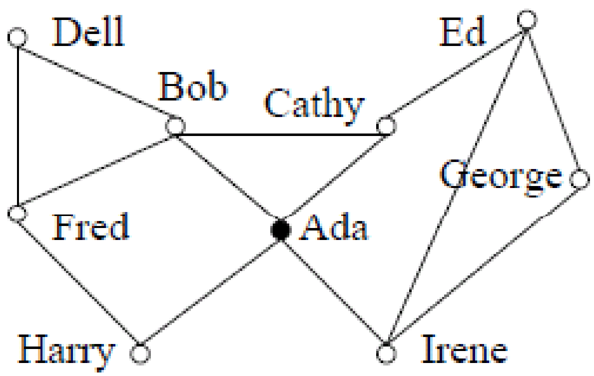
- Focus on attacks on network data and attempts to anonymize the network through addition of "noise"
  - Directed alteration
  - Random alteration
  - generalization

k-degree anonymity

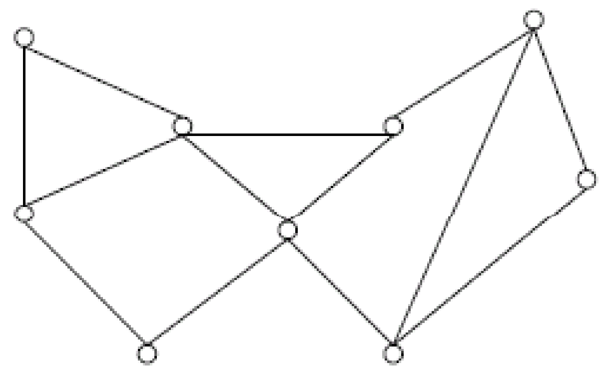


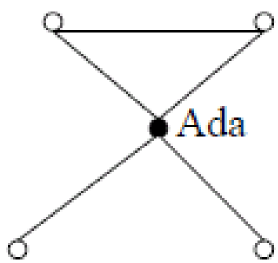
- The kind of attack
  - Vertex Refinement Queries
- Objective
  - The published graph
    - For every node v, there exist at least k-1 other nodes in the graph with the same degree as v
    - Choose the number of edges that are added to achieve k-degree anonymity subject to minimally affecting the graph’s 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

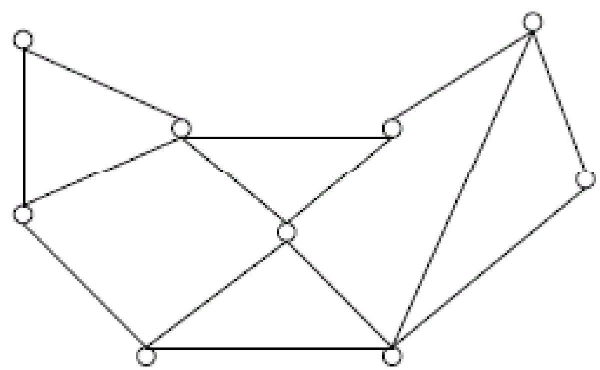


(a) the social network


(b) the network with anonymous nodes



(c) the 1−neighborhood graph of Ada


(d) privacy−preserved anonymous network

Random alteration

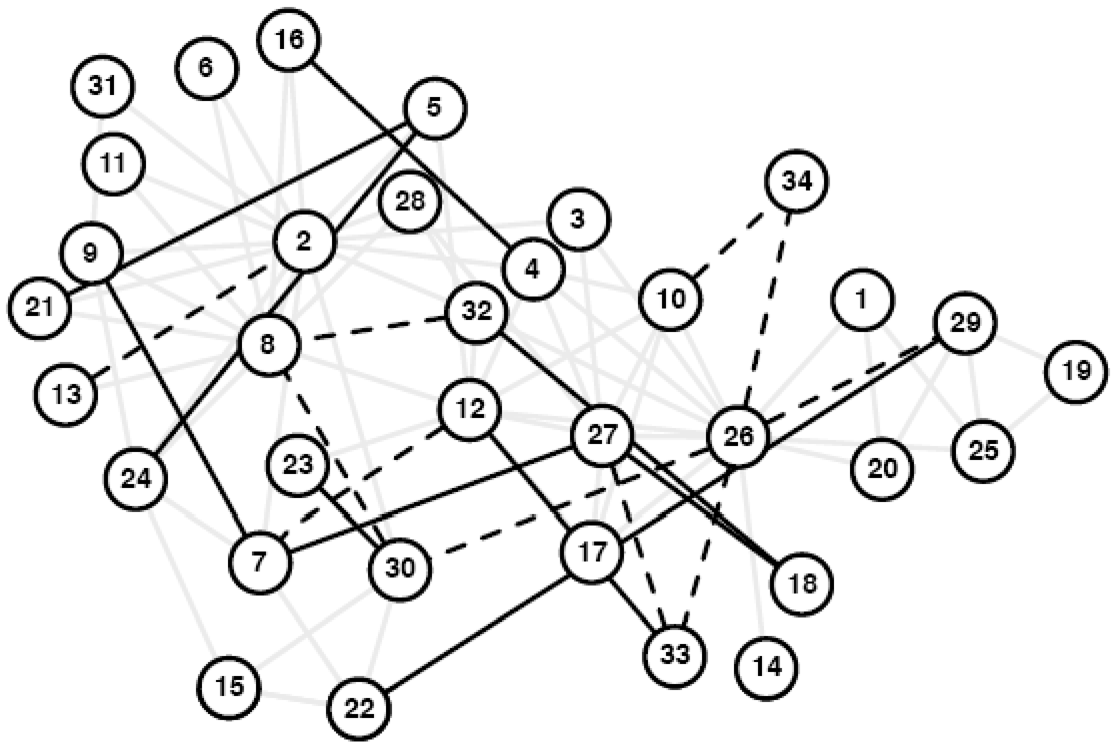
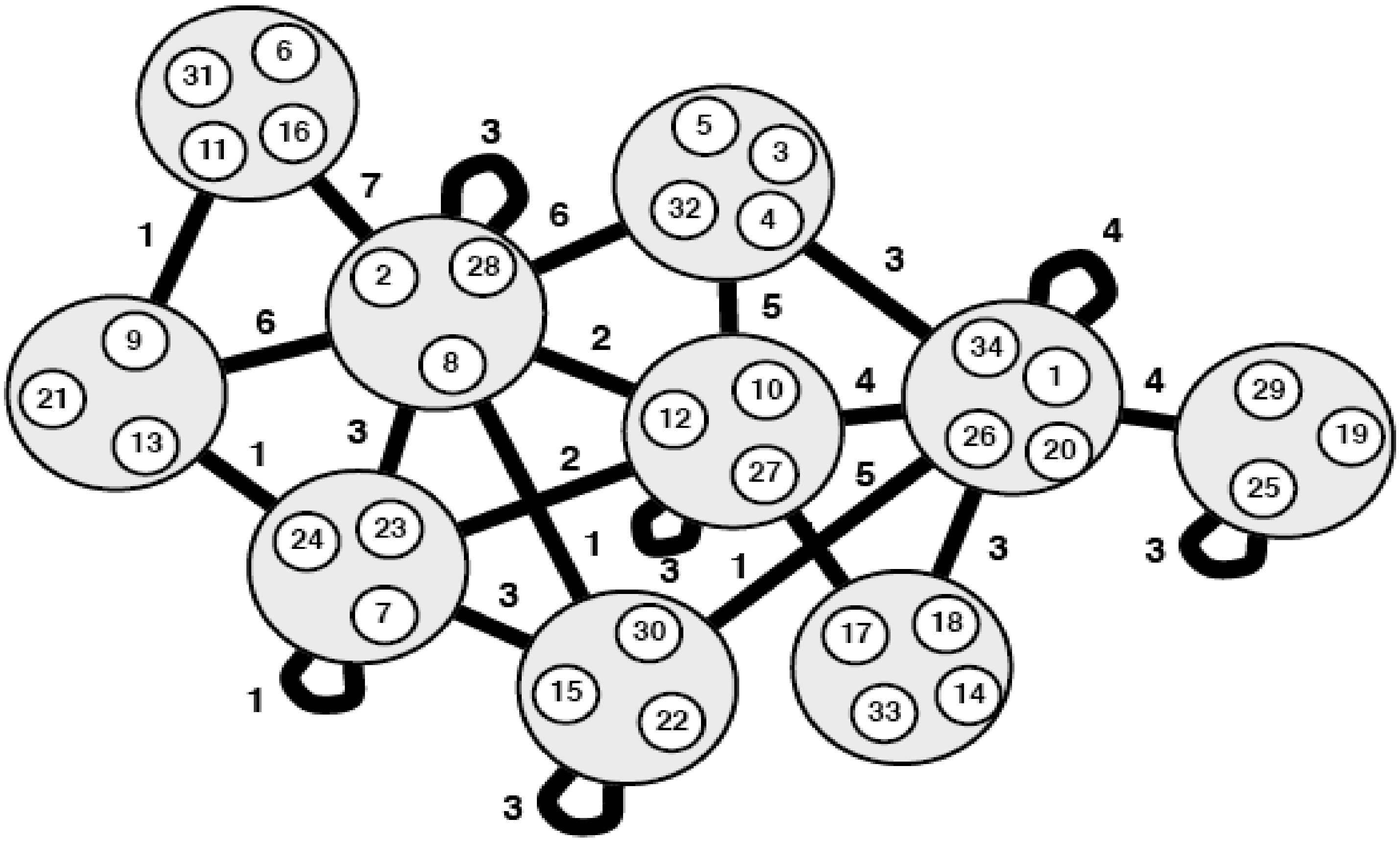


Figure 8: Examples of random alteration applied to the anonymized karate club network of Figure 2. The original edges are shown in light gray. With random alteration, edges are randomly chosen and rewired until $m = 10$ edges have been altered (deleted edges in dashed black; inserted edges in solid black).

Hay, 2010

Generalization



Hay, 2010

Exponential Random Graphs

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

$$P(N) = \exp\{\theta \cdot u(N) - \ln Z(\theta)\}$$

- Can estimate the θ 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
  - $u_1(N)$ = # edges in N
  - $u_2(N)$ = # 2-stars in N
  - $u_3(N)$ = # triangles in N

$$P(N) \propto \exp\left\{\theta_1 u_1(N) + \theta_2 u_2(N) + \theta_3 u_3(N)\right\}$$

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