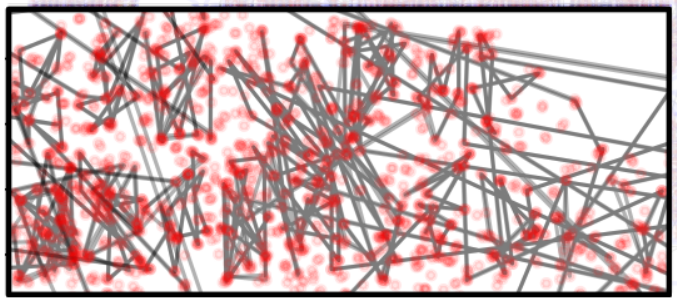
A Brief introduction to Social Network Analysis

Descriptives to Statistical Inference

Zack W Almquist*

*Department of Sociology and School of Statistics
University of Minnesota

MPC Inequality and Methods Workshop



Peter M. Blau, Exchange and Power in Social Life, 1964

"To speak of social life is to speak of the association between people – their associating in work and in play, in love and in war, to trade or to worship, to help or to hinder. It is in the social relations men establish that their interests find expression and their desires become realized."

J.L. Moreno, New York Times, April 13, 1933

"If we ever get to the point of charting a whole city or a whole nation, we would have . . . a picture of a vast solar system of intangible structures, powerfully influencing conduct, as gravitation does in space. Such an invisible structure underlies society and has its influence in determining the conduct of society as a whole."

Overview

- Motivating Examples
- Definitions
- Network Data Collection
- Descriptive Statistics
- Statistical Inference

- Tools
  - R/RStudio
  - github
  - Other resources

- Freeman, Linton (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, BC, Canada: Empirical Press.

Required Software

- github (`http://www.github.com`)
- R (`http://cran.r-project.org/`)
- RStudio (`http://www.rstudio.com/`)

Required Software

- R Packages for SNA
  - R packages:

```
install.packages("statnet")
install.packages("ergm")
install.packages("sna")
install.packages("network")
```

- R packages needed for installing from github
  - R packages:

```
install.packages("devtools")
```

- R packages from github
  - R packages:

```
install.packages("networkMethods")
install.packages("networkdata")
```

Examples: Some Types of Relationships

- Conceptual: shared or antithetic properties
  - E.g., similarity/difference in individual attributes, correlation among variables, inclusion/exclusion, surface matchings on proteins
- Co-categorical: shared membership
  - E.g., organizational co-membership, event co- participation, co-occurrence of words within texts
- Nominational: resulting from the behavior of ego
  - E.g., attributions of friendship/enmity, kinship (fictive or otherwise), causal narratives

Examples of Social Networks

For example:

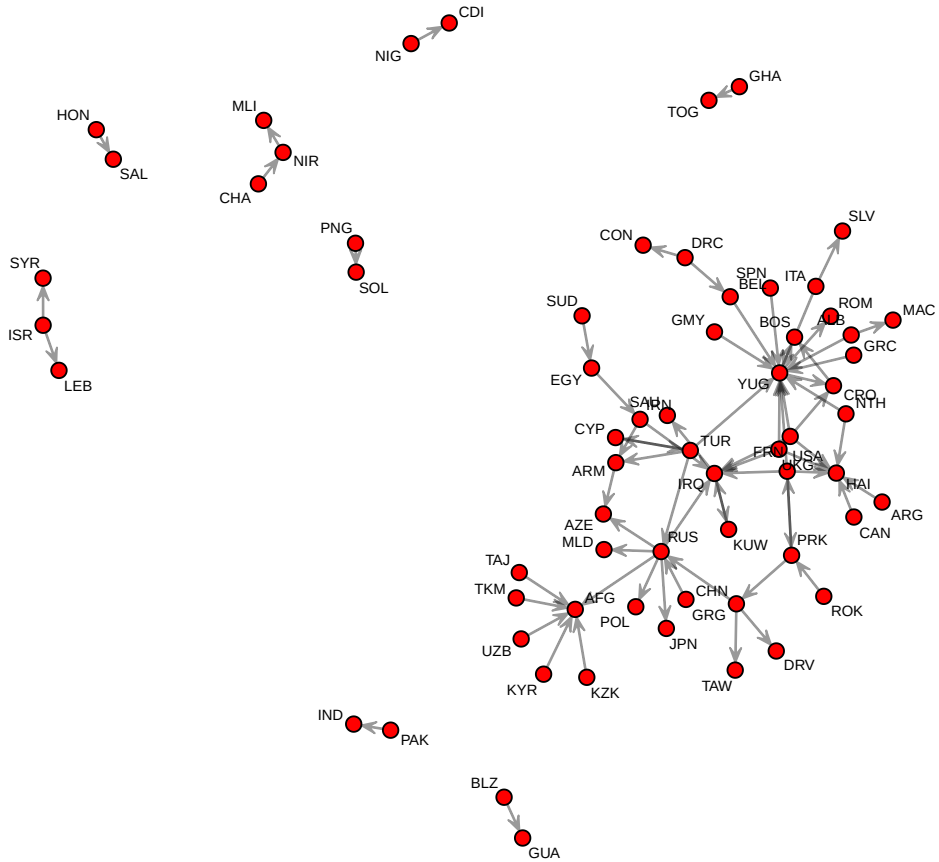
- Macro-Level Networks
- Interpersonal Networks

Macro-Level Networks

- International Relations
  - Militarized interstate disputes
- Migration Networks
  - County-to-county migration in the US
- Interorganizational Relations
  - Search and rescue (SAR) activities: Emergent multi-organizational networks (EMONs)

Militarized Interstate Disputes

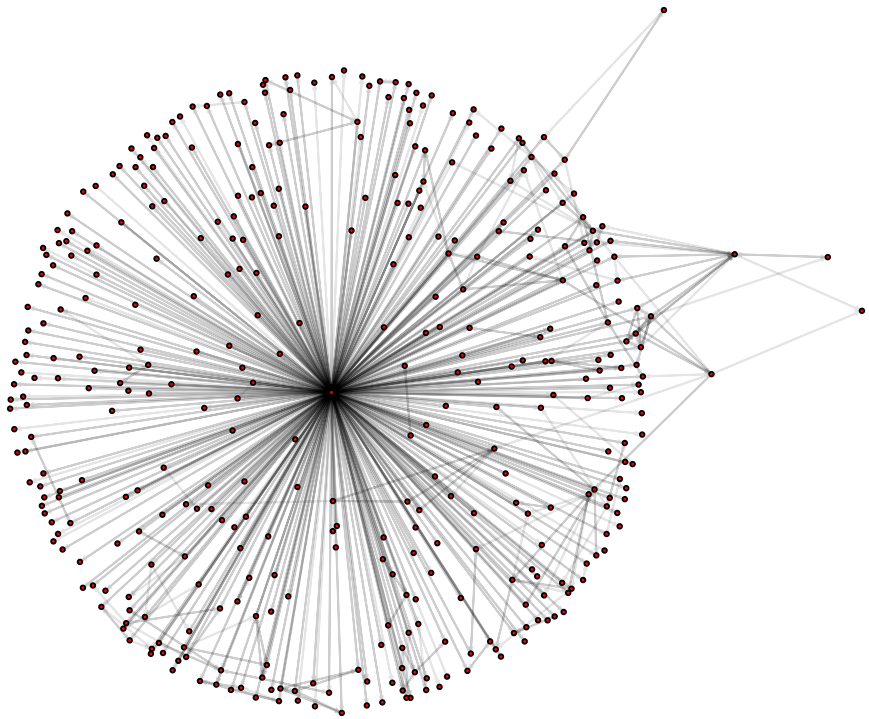
1993 militarized interstate disputes (MIDs)



Correlates of War Project

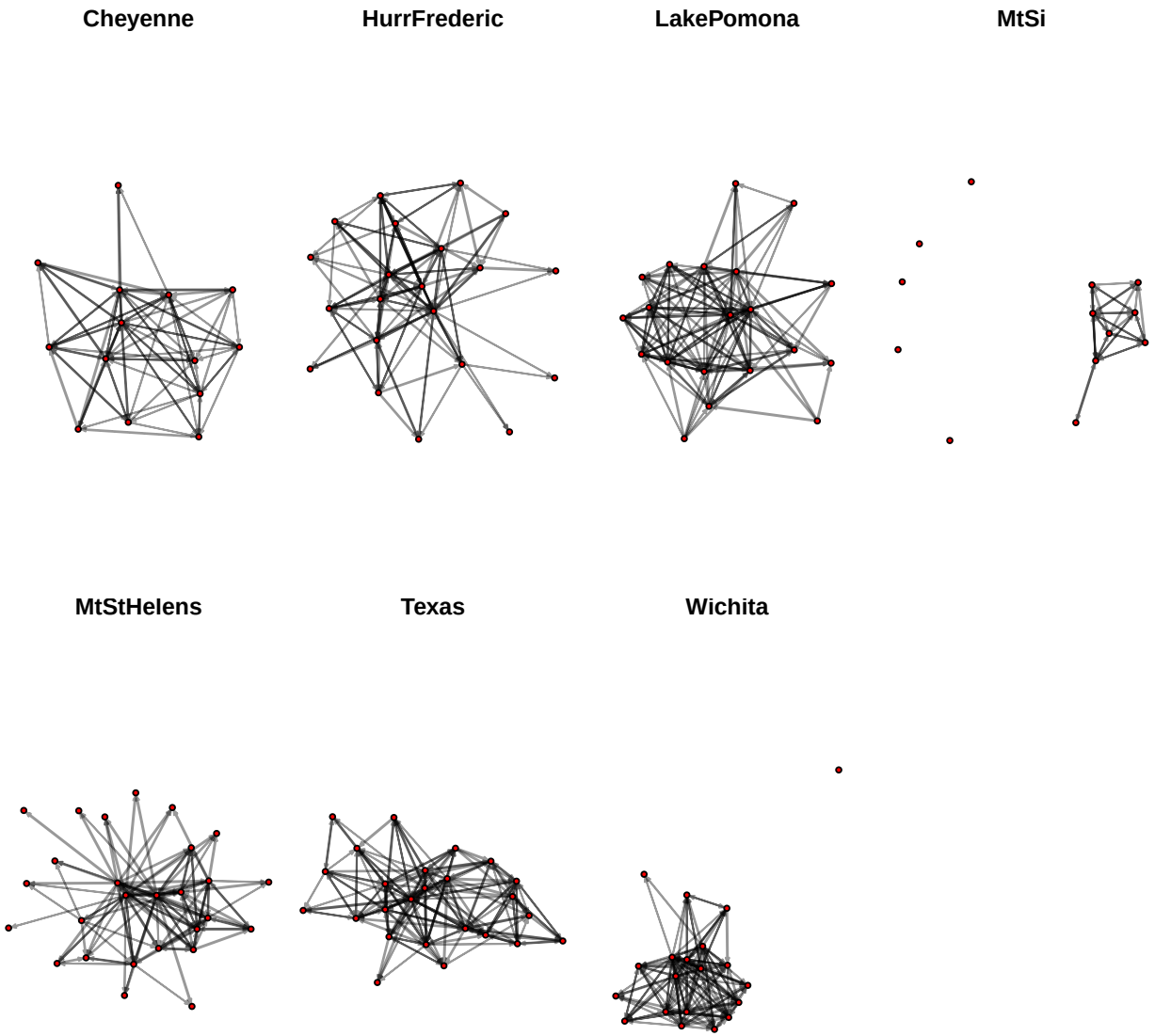
County-to-County Migration in the US

IRS Migration Data, 2000−2001



Threshold at 99%

SAR EMONs, All Reported Ties, from Drabek et al 1981

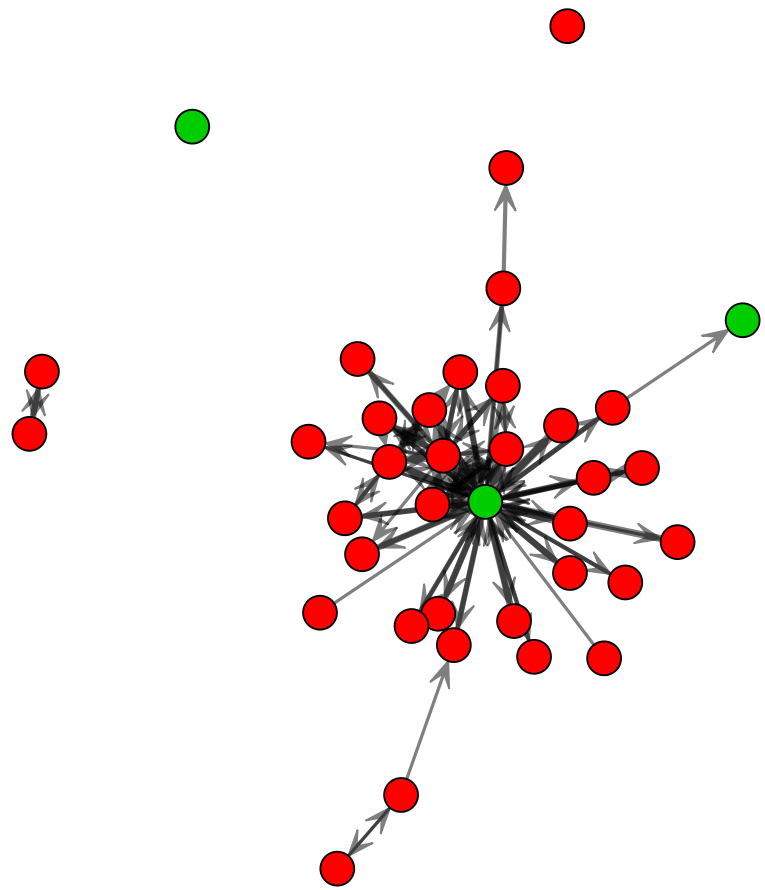


Interpersonal Networks

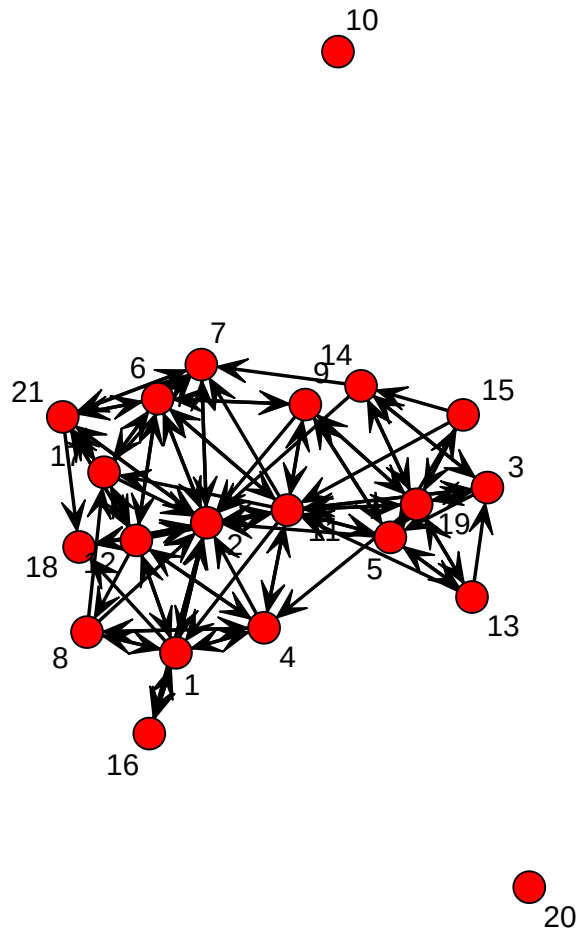
- Communication
  - WTC responder radio communications (aggregate and dynamic)
- Friendship
  - Perceived and estimated friendships among managers
- Interaction in Task Performance
  - WTC police reports
- Co-Participation
  - Research group co-participation in a large research project

World Trade Center Radio Communication

Data set coded by Butts et al. (2007)



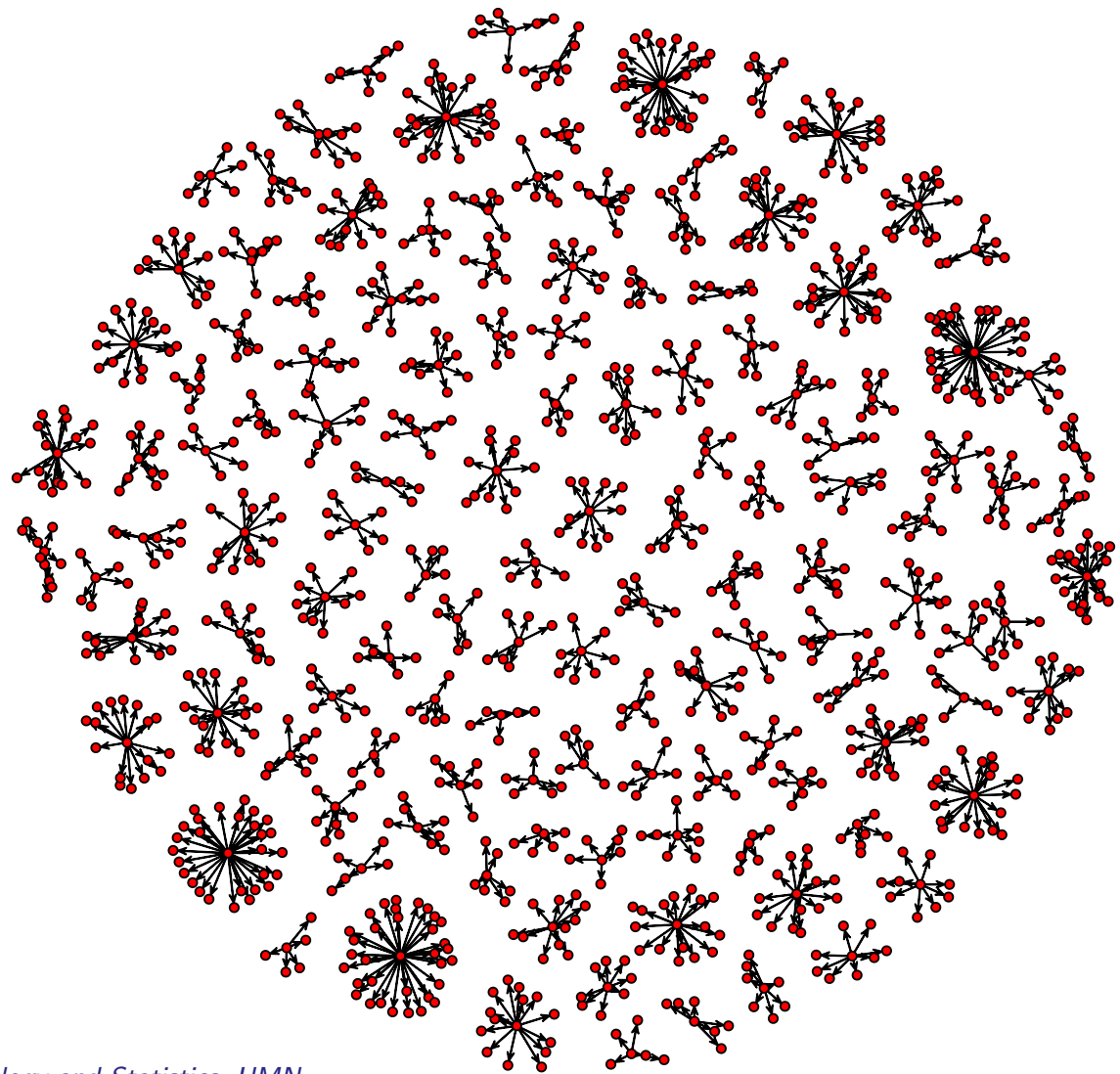
David Krackhardt (1987) Perceived Friendships



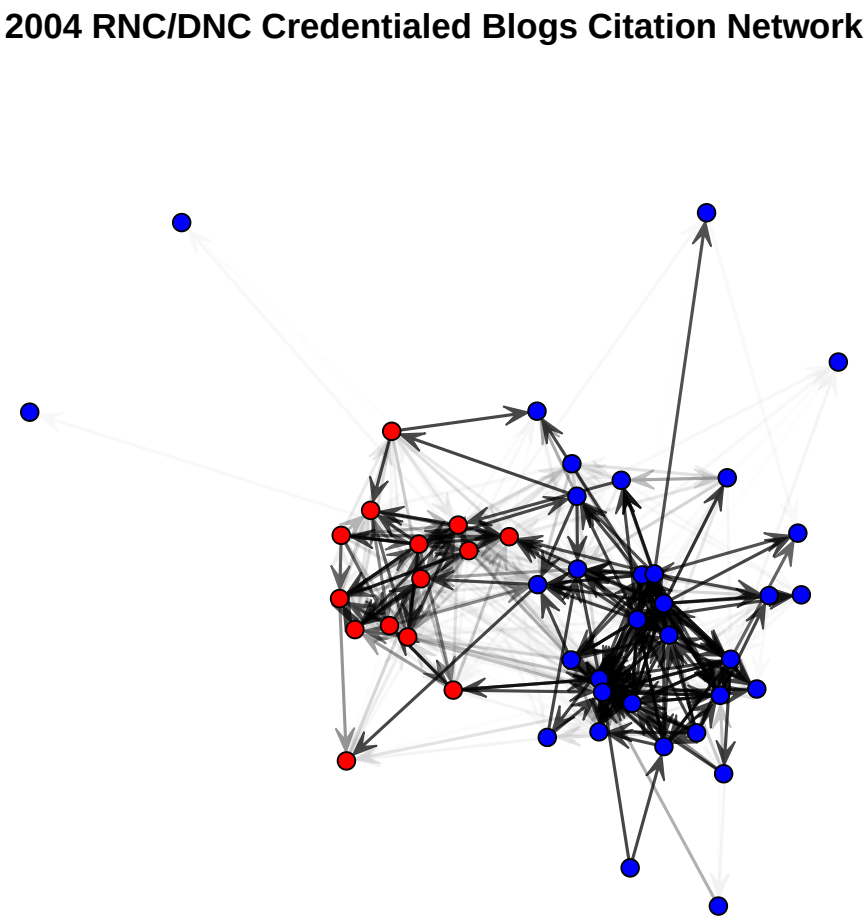
Citation Networks

- Authorship
  - Authorship of articles in a large research project
- Hyperreferences
  - Citations among weblogs
- Functional Citations
  - Function calls in the Linux kernel

Co-Author Network from J. Yang and J. Leskovec (2012).



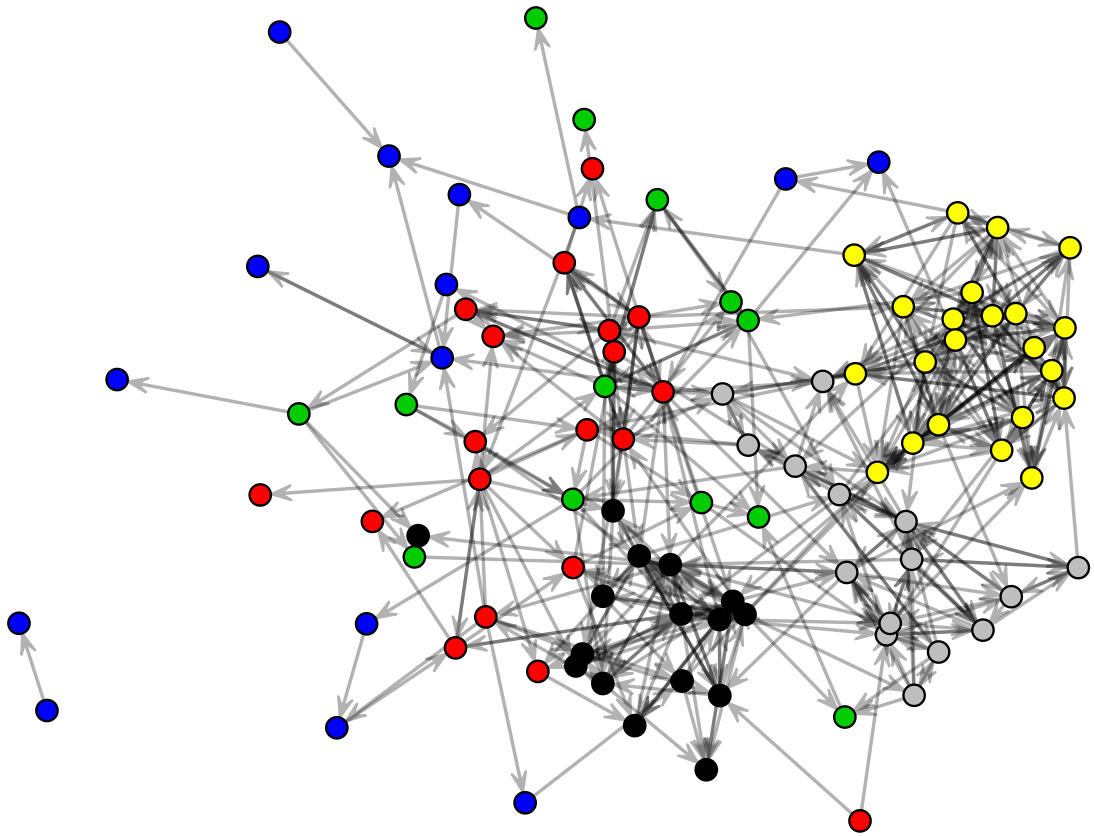
Weblog Citation Network from Butts and Cross (2005)



Even More Networks You Can Find In Everyday Circumstances

- Acknowledgments in articles, CD liners
- Citations among scholars
- Commodity chains (whence did your printer come?) Email messages
- Entailment in ownership (I only own A if I also own B...) Flow of inter-office mail envelopes
- Hyperlinks (actually a kind of citation network)
- Item co-ownership (do we read the same books?) Speed-dial settings and address books

Introduction to Network Concepts

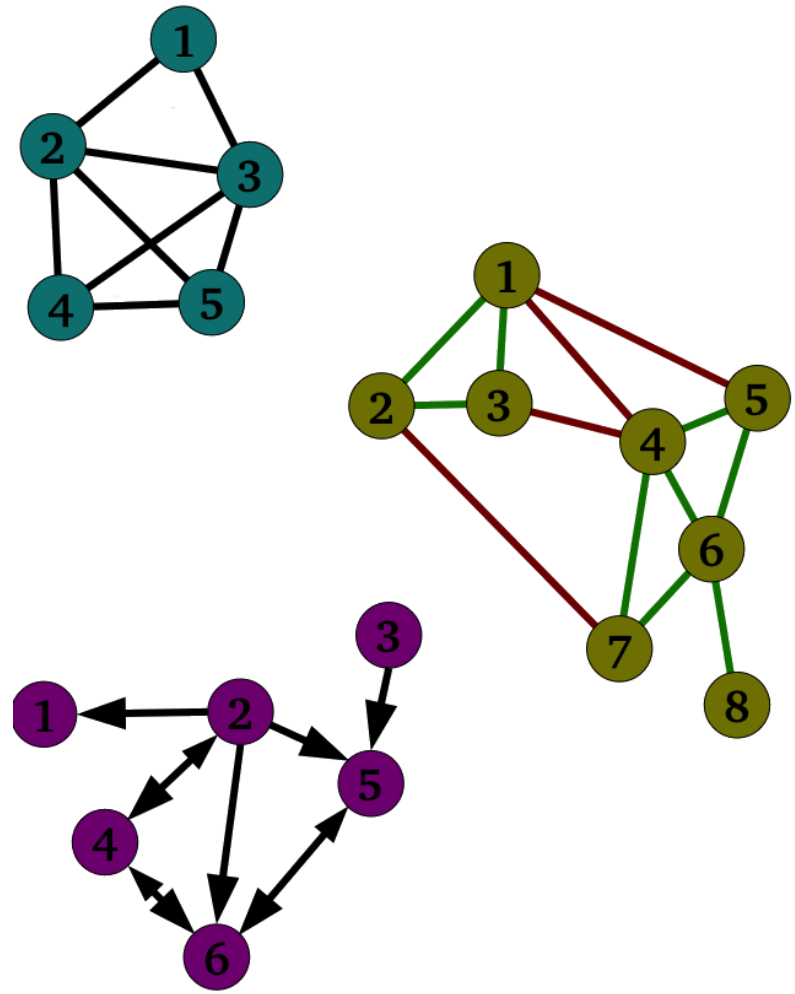


Focus on Relations

- Relationship: an irreducible property of two or more entities
  - Compare with properties of entities alone ("attributes")
- Focus: the properties and consequences of relations (rather than individual properties)
  - Entities can be persons, non-human animals, groups, locations, organizations, regions, etc.
  - Relationships can be communication, acquaintanceship, sexual contact, proximity, migration rate, alliance/conflict, etc.
  - Social network analysis: the study of relational data arising from social systems

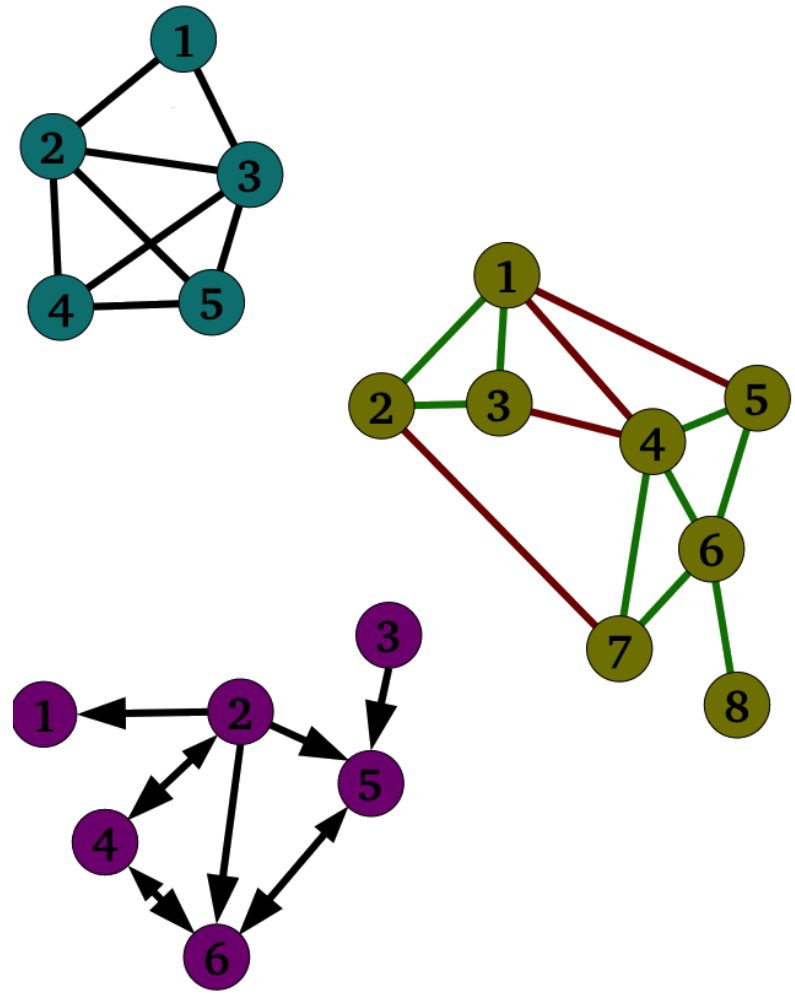
Some Vocabulary

- **Network**: a collection of entities, together with a set of relations on those entities
  - Entities: nodes, or vertices
  - Relations: edges, or ties
    - Focus on dyadic relations
    - Directed vs. undirected edges
    - May be signed or valued
- Graph: a set of vertices together with a set of edges
  - Entities can be persons, non-human animals, groups, locations, organizations, regions, etc.
  - Relationships can be communication, acquaintanceship, sexual contact, proximity, migration rate, alliance/conflict, etc.
  - Mathematical representation of social structure



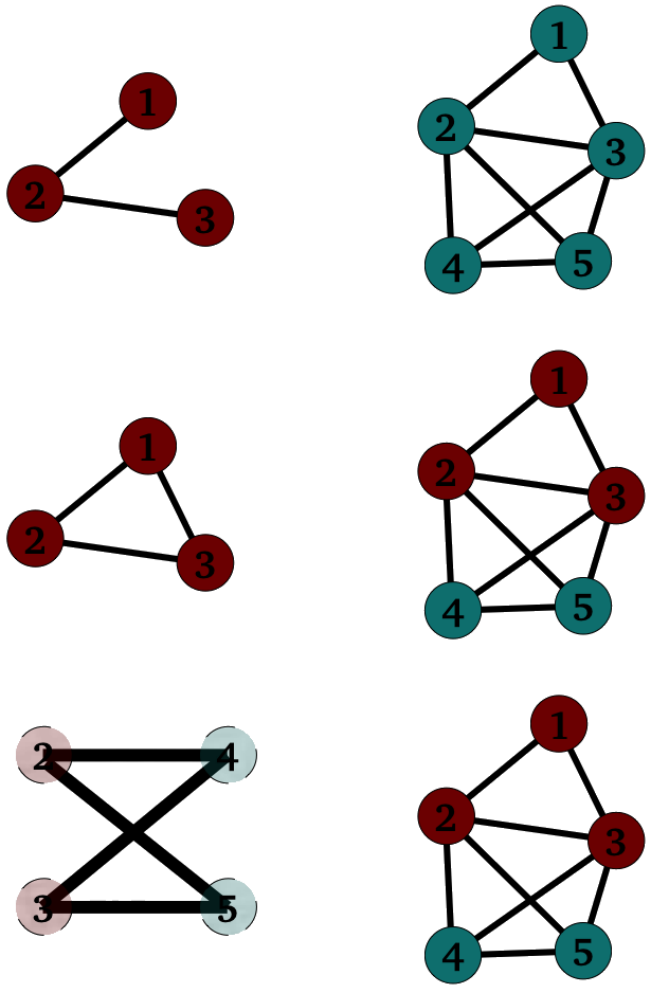
Sociometric Notation

- Based on *graph theory*
- Graph defined as $G = (V, E)$
  - Vertex set: $V = \{v_1, \ldots, v_N\}$
  - Edge set: E
    - Undirected: $E \subseteq \{\{v_i, v_j\} \ : \ v_i, v_j \in V\}$
    - Directed: $E \subseteq \{(v_i, v_j) \ : \ v_i, v_j \in V\}$
    - Simple iff $\{v_i, v_i\}, (v_i, v_i) \notin E$
- Simple operations
  - $C_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$
  - $G_1 \cap G_2 = (V_1 \cap V_2, E_1 \cap E_2)$



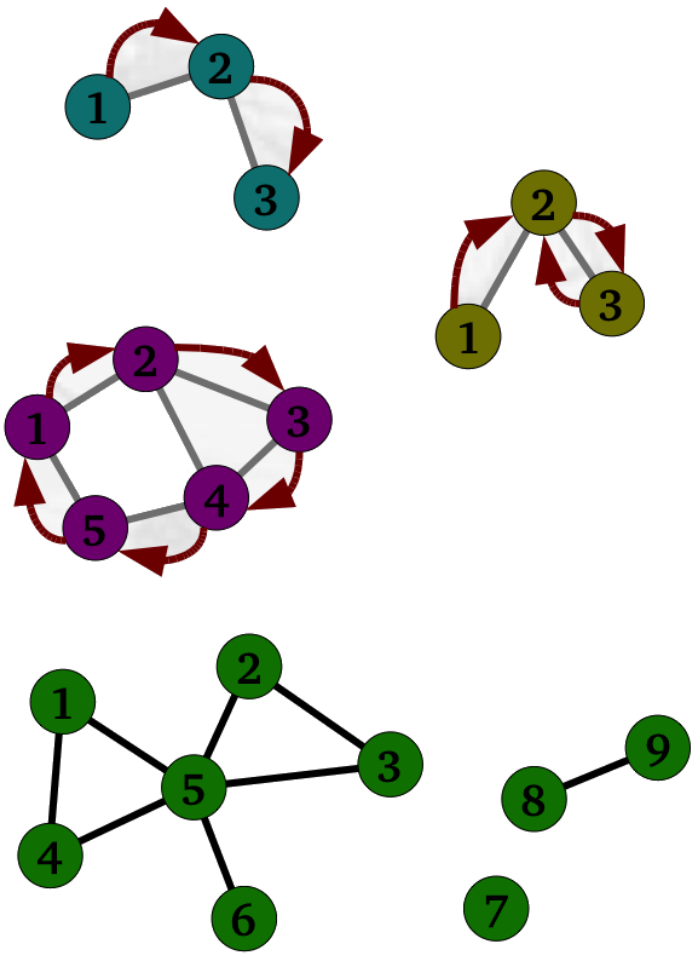
Subgraphs and Cuts

- Subgraph
  - Selection of vertices and edges from G
  - $G_1 \subseteq G_2$ iff $V_1 \subseteq V_2$, $E_1 \subseteq E_2$
- Induced Subgraph
  - Selection of vertices, w/all associated edges
  - $G[S] = (S, \{e \in E \ : \ e \subseteq S \times S\})$
- Edge cut
  - $G[S_1, S_2] = \{\{s_1, s_2\} \in E \ : \ s_1 \in S_1, s_2 \in S_2\}$
  - $G[S_1, S_2] = \{(s_1, s_2) \in E \ : \ s_1 \in S_1, s_2 \in S_2\}$



A Bit More Vocabulary

- Adjacency
  - i is adjacent to j iff i sends a tie to j
- Walks, paths, and cycles
  - *Path*: a sequence of adjacent vertices (and connecting edges) with no repetitions
  - *Walk*: like a path, but repetition is allowed
  - *Cycle*: like a path, but start/endpoints are the same
- Connectedness and components
  - i and j are connected iff an i, j path exists
  - *Component*: a maximal set of connected vertices
  - *Isolate*: a component of size 1

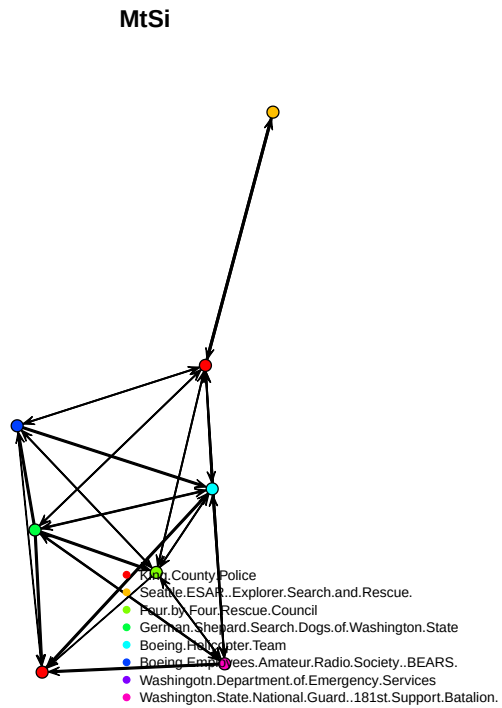


Relational Data

- Relational (network) data concerns connections among entities, rather than attributes of entities
  - Entities can be persons, organizations, concepts, etc.
  - Relations can be interaction, proximity, membership, etc.
- Two basic types (for now, at least!)
  - One-mode data: connections among one sort of entity
  - Two-mode data: connections among two sorts of entities

One-Mode Data

- Networks with one vertex class
- Represented by adjacency matrices
  - Vertices on rows and columns
  - $A_{ij} = 1$ if i sends a tie to j, else $A_{ij} = 0$
    - Can contain edge values, where applicable (A_{ij} is value of i, j edge)
  - Symmetric in undirected case
  - Diagonals represent self-ties
    - Often treated as undefined

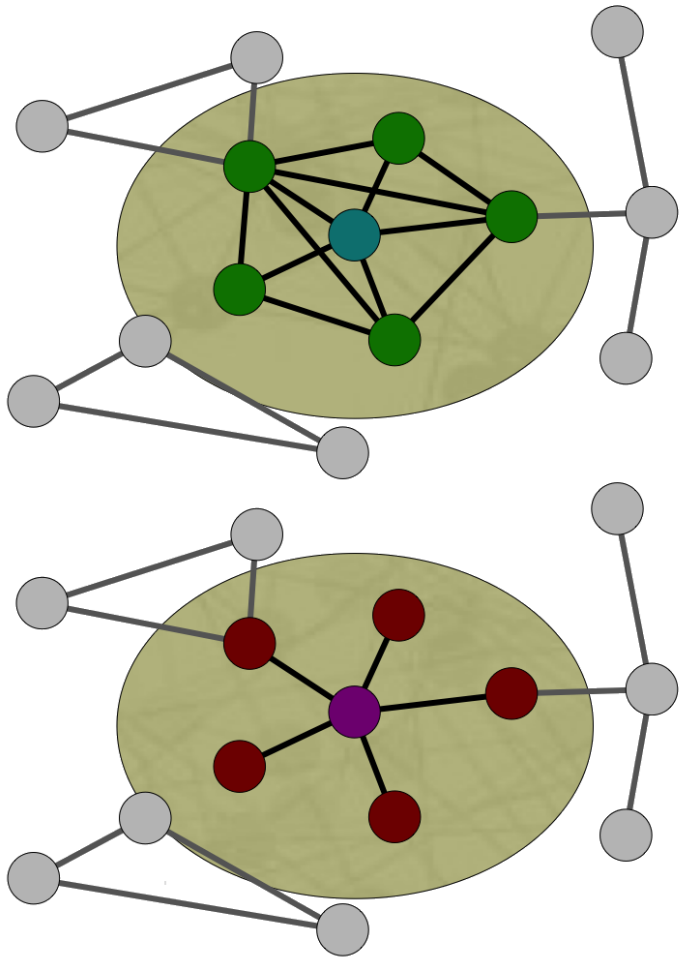


	1	2	3	4	5	6	7	8
1	0.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
2	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00
4	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
5	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
6	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
7	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00

Table: Mt. Si SAR EMON, Confirmed Ties

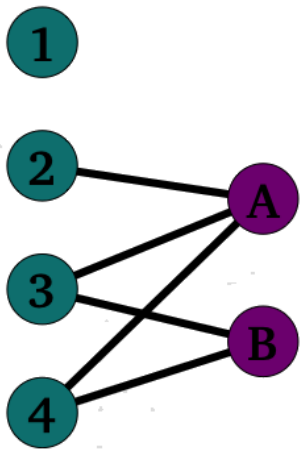
Special Case: Ego Nets

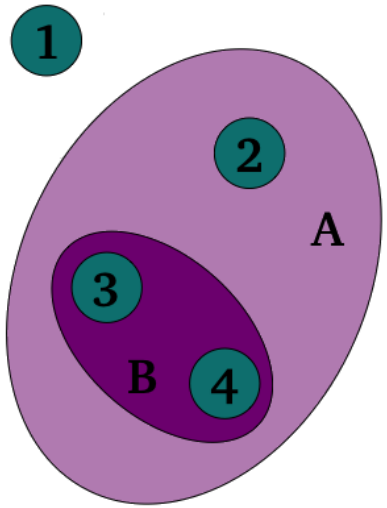
- Egocentric network: focal actor ("ego") + neighbors ("alters") + ties among alters
- What does it tell us?
  - Number of ties ego has (neighborhood size)
  - Triangles (3-cliques) containing ego
  - Connections among alters
  - Neighborhood composition (if asked)
- **Note:** Sometimes called *personal networks*



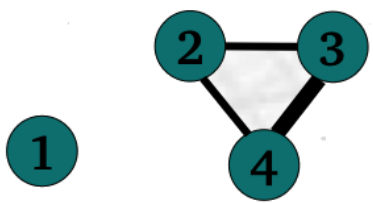
Two-Mode Data

- Networks with two vertex c
  - Different entity types
  - Membership
  - Matching/containment
- Represented by incidence m
  - "Senders" on rows, "rec on columns
- Can be used to obtain "dua representations

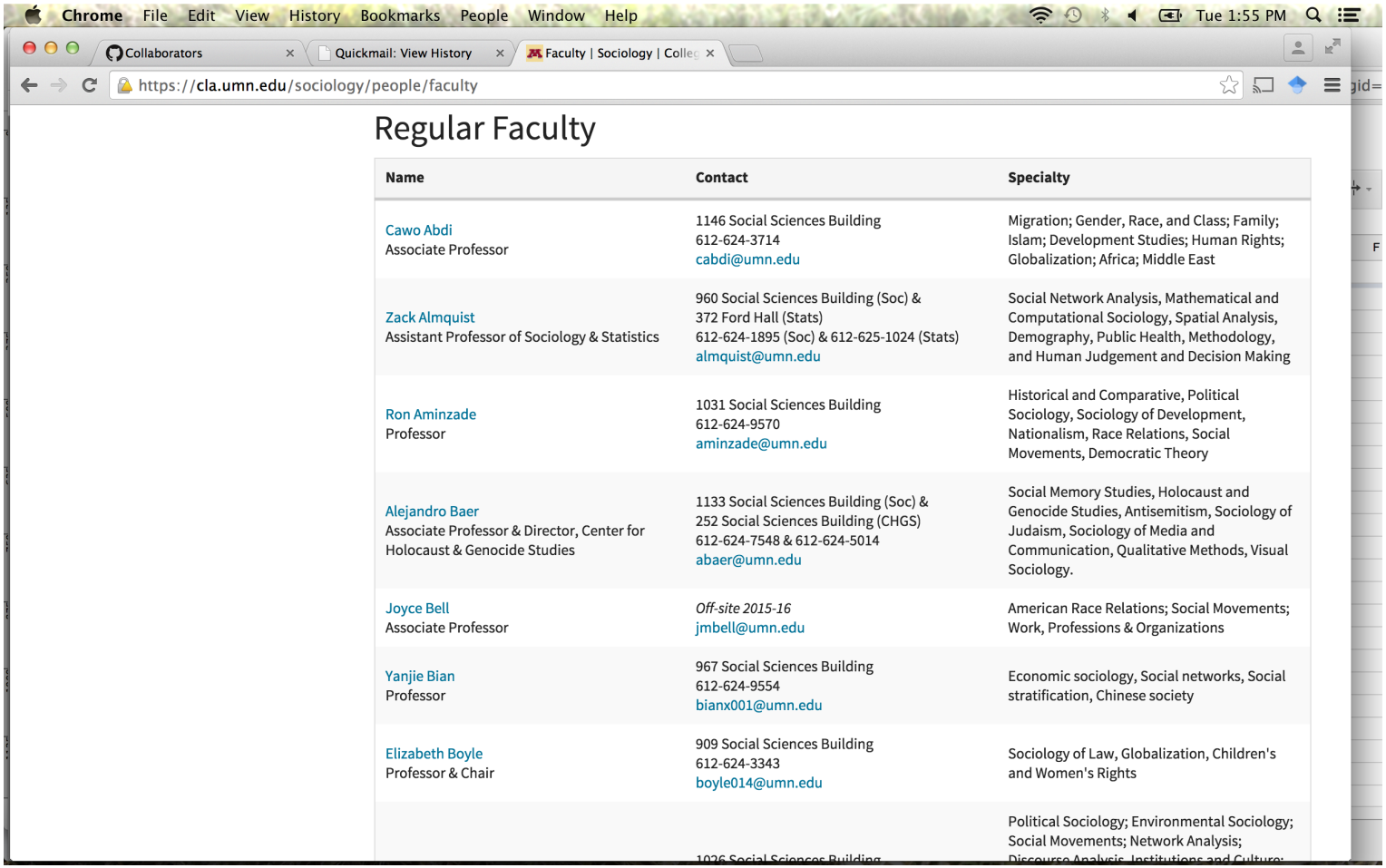




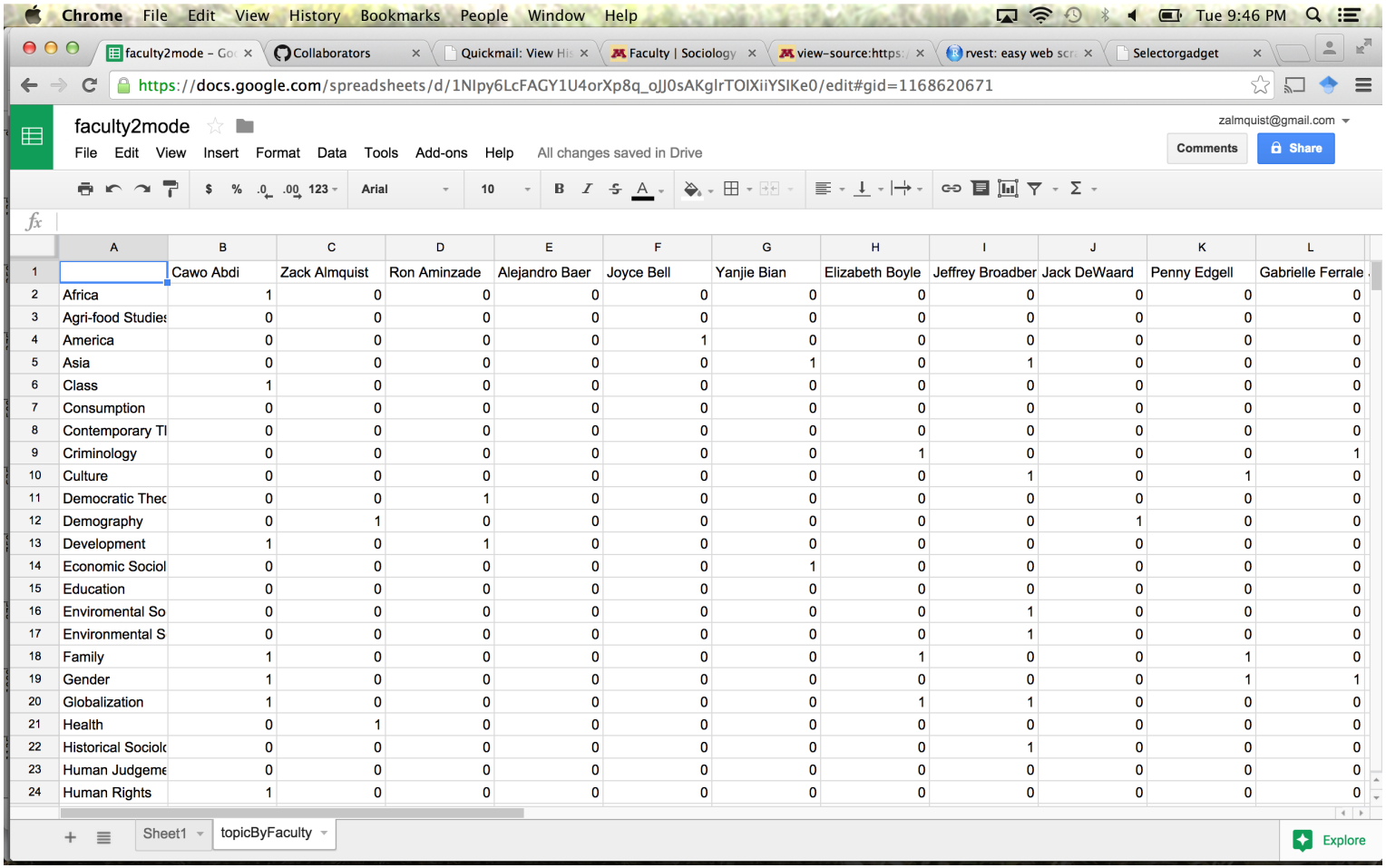
	A	B
1		
2	1	
3	1	1
4	1	1



Two-Mode Data



Two-Mode Data

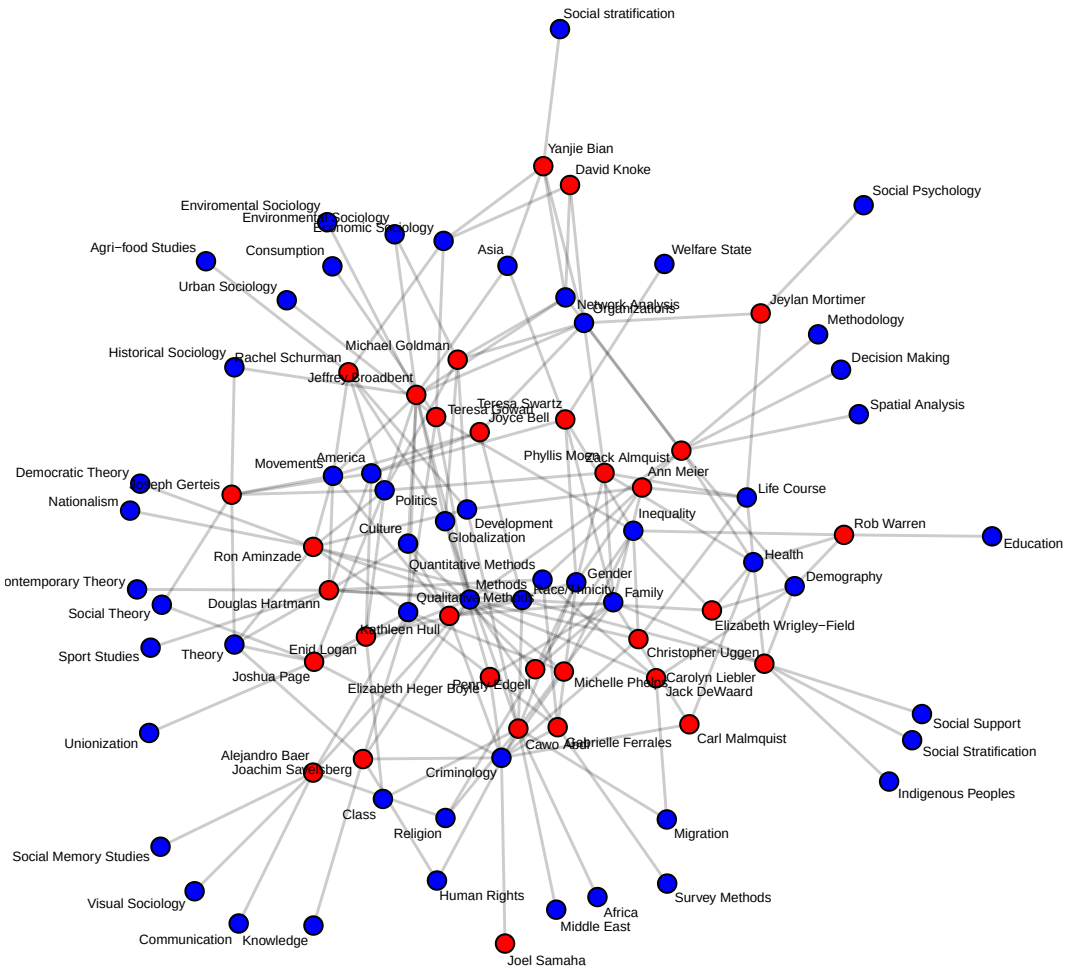


https://docs.google.com/spreadsheets/d/1NIpy6LcFAGY1U4orXp8q_oJJ0sAKglrTOIXiiYSIKe0/edit#gid=1168620671

faculty2mode

	A	B	C	D	E	F	G	H	I	J	K	L
1		Cawo Abdi	Zack Almquist	Ron Aminzade	Alejandro Baer	Joyce Bell	Yanjie Bian	Elizabeth Boyle	Jeffrey Broadber	Jack DeWaard	Penny Edgell	Gabrielle Ferrale
2	Africa	1	0	0	0	0	0	0	0	0	0	0
3	Agri-food Studies	0	0	0	0	0	0	0	0	0	0	0
4	America	0	0	0	0	1	0	0	0	0	0	0
5	Asia	0	0	0	0	0	1	0	1	0	0	0
6	Class	1	0	0	0	0	0	0	0	0	0	0
7	Consumption	0	0	0	0	0	0	0	0	0	0	0
8	Contemporary Tl	0	0	0	0	0	0	0	0	0	0	0
9	Criminology	0	0	0	0	0	0	1	0	0	0	1
10	Culture	0	0	0	0	0	0	0	1	0	1	0
11	Democratic Thec	0	0	1	0	0	0	0	0	0	0	0
12	Demography	0	1	0	0	0	0	0	0	1	0	0
13	Development	1	0	1	0	0	0	0	0	0	0	0
14	Economic Sociol	0	0	0	0	0	1	0	0	0	0	0
15	Education	0	0	0	0	0	0	0	0	0	0	0
16	Enviromental So	0	0	0	0	0	0	0	1	0	0	0
17	Environmental S	0	0	0	0	0	0	0	1	0	0	0
18	Family	1	0	0	0	0	0	1	0	0	1	0
19	Gender	1	0	0	0	0	0	0	0	0	1	1
20	Globalization	1	0	0	0	0	0	1	1	0	0	0
21	Health	0	1	0	0	0	0	0	0	0	0	0
22	Historical Sociol	0	0	0	0	0	0	0	1	0	0	0
23	Human Judgeme	0	0	0	0	0	0	0	0	0	0	0
24	Human Rights	1	0	0	0	0	0	0	0	0	0	0

Two-Mode Data



Two-Mode Data

One-Mode Projections

- Let A be an $N \times M$ incidence matrix; the row-projection of A is the $N \times N$ matrix B such that

$$B_{ij} = \sum_{k=1}^{M} A_{ik} A_{jk}$$

- Likewise, the column projection of A is the $M \times M$ matrix C such that

$$C_{ij} = \sum_{k=1}^{N} A_{kj} A_{kj}$$

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

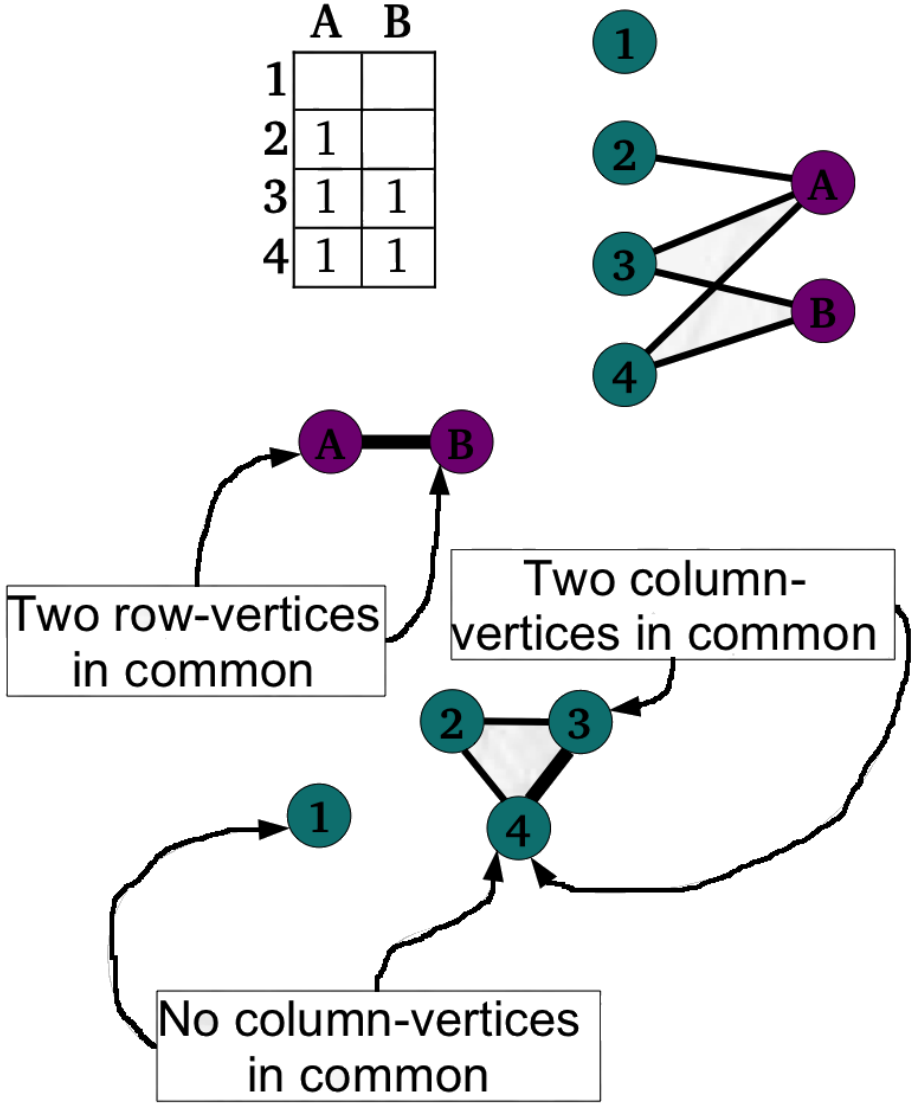
$$B = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}$$

$$C = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

What the Projections Mean

- Projections have simple meaning
  - Row: B_{ij} is the number of column elements shared by row elements i and j
  - Column: C_{ij} is the number of row elements shared by column elements i and j
  - Ex: Number of shared interests between two faculty; number of faculty having a given interest area in common

	A	B
1		
2	1	
3	1	1
4	1	1



Network Data Collection

- To analyze network data, we must first collect it!
  - Many approaches exist ? some better than others for particular purposes
  - Complex topic overall, but we will at least skim the surface. . .
- Two important concepts (not always separable):
  - Instruments: tools used to elicit information from respondents, assess presence/absence of ties from sensors or archival materials, etc.
  - Designs: protocols for determining how information should be elicited, who should be sampled, etc.

Relational Data - Designs

- Own-tie reports
  - Personal ties elicited from each ego
  - Standard instruments: roster and name generator
  - Pros: Easily implemented, most common design
  - Cons: Vulnerable to reporting error
- Egocentric network sampling
  - Personal ties elicited from ego, followed by induced ties
  - Standard instrument: name generator followed by roster
  - Pros: Well-suited to large-scale survey sampling; provides information on ego's neighborhood
  - Cons: Vulnerable to reporting error; false positives/negatives on own ties contaminate sampling of neighbors' ties

Instruments: Name Generators and Rosters

- Name generator: asks respondents to list names
  - E.g. "Think about the persons with whom you have talked in the past week. Please list all such persons in the following space." (followed by space to enter names)
- Pros:
  - Don't have to know name list; can use with large groups or organizations
- Cons:
  - High rate of forgetting; unclear boundary
- Roster: asks respondents to choose names from fixed list
  - E.g. "For each of the following persons, place a check in the associated blank if you have talked with him her in the past week." (followed by check list)
- Pros:
  - More accurate, clear boundary
- Cons:
  - List may be prohibitively long, can be imposing; alters must be known in advance

Instruments: Complete Ego Net

- Common way to elicit ego nets: complete instrument followed by roster
  - Asked to name those with whom you discussed important matters
  - Then, asked to fill in same question for all pairs of persons named initially
- Pros:
  - Relatively easy to administer; don't need entire list of possible alters; don't have to ask about all group members
- Cons:
  - Step 1, step 2 questions have different error rates; may need large roster if many alters; hard to use with paper-based surveys

Back to Designs

- Link-tracing
  - Personal ties elicited from ego; new ego(s) chosen from alters; process is iterated (possibly many times)
  - Standard instruments: multiwave own-report, RDS
  - Pros: Allows estimation of network properties for large and/or hard to reach populations; highly scalable; can be robust to poor seed sampling
  - Cons: Vulnerable to reporting error; reporting errors can contaminate design (but may be less damaging than ego net case); often difficult to execute
- Arc Sampling
  - Reports on third-party ties elicited from ego; multiple egos may be sampled for each third-party tie
    - Archival/observer data is a special case Standard instrument: CSS
  - Pros: Very robust to reporting error (via modeling); can be very robust to missing data
  - Cons: Can impose large burden on respondents; can be difficult to execute

Designs: CSS and RDS

- Cognitive Social Structure (CSS)
  - Ask each group member to report on all members‘ ties
  - Ex: “Which of the following persons does Steve go to for help or advice?”
- Pros:
  - Gets information on perception; can be used to get high-accurate estimates
- Cons:
  - Hard to use; requires roster; doesn’t scale well

- Respondent Driven Sampling (RDS)
  - Combine standard network instrument with recruitment “tickets”
  - Respondents given tickets to give to others; if they volunteer, both get paid
- Pros:
  - Can use with hidden, vulnerable populations
- Cons:
  - Difficult; expensive; complex to analyze; poorly understood

Coding Schemes as "Instruments"

- Can also think of coding schemes for archival materials as "instruments"
- Transcripts
  - Tag each line by sender/receiver
  - (i, j) tie if i sends to j
- Descriptive lists/tables
  - Common for two-mode data
  - Build entity/property table; fill in (i, j) as 1 if ith row entity has property j
- Video/Audio
  - Determine criterion for interaction
  - Find all interactions, code by sender/receiver
  - (i, j) tie if i sends to j
- Narrative documents
  - Determine criterion for interactions
  - As before, code by sender/receiver (or just by dyad, if not directed)
  - (i, j) tie if i sends to j, or $\{i, j\}$ tie if i and j interact

Coding Schemes as "Instruments"

- Can also think of coding schemes for archival materials as "instruments"
- Transcripts
  - Tag each line by sender/receiver
  - (i, j) tie if i sends to j
- Video/Audio
  - Determine criterion for interaction
  - Find all interactions, code by sender/receiver
  - (i, j) tie if i sends to j
- Narrative documents

$\{i, j\}$ tie if i and j interact

A fun example

Adams, jimi (2015). Glee's McKinley High: Following Middle America's sexual taboos. Network Science, 3(2): 293 – 295.

http://journals.cambridge.org/abstract_S2050124215000168

Related Issue: Network Boundary Problem

- Important problem: whence the vertex set?
  - If misspecified, theoretically relevant ties may be missed
- Typical cases
  - Exogenously defined
    - Physical region, group/cohort membership
  - Relationally defined
    - Isolated social unit (component), locally dense
  - Design defined
    - Alters named in ego net,link trace; sampled egos
- Major distinction: local versus global properties
  - Different sampling methods needed for each

Graph-Level Properties

- Earlier, we discussed the notion of node-level indices (mainly centrality)
  - Dealt with position of the individual within the network
- We will focus on properties at the graph level
  - Graph-level index: $f(v, G) \rightarrow \Re$
  - Describes aggregate features of structure as a whole
- Provide complementary insight into social structure
  - Node-level properties tell you who's where, but graph- level properties provide the broader context

Density

- Density: fraction of possible edges which are present
  - Probability that a given graph edge is in the graph
- Formulas:
  - Undirected: $\delta = \frac{2\sum_{i=1}^{N}\sum_{j=i}^{N} Y_{ij}}{N(N-1)}$
  - Directed: $\delta = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N} Y_{ij}}{N(N-1)}$

```
library(sna)
undirected <- rgraph(10,
    mode = "graph")
directed <- rgraph(10,
    mode = "digraph")
gden(undirected, mode = "graph")
```

```
 R >  [1] 0.5555556
```

```
2 * sum(undirected[upper.tri(undirected)])/(NROW(directed) *
    (NROW(directed) -
        1))
```

```
 R >  [1] 0.5555556
```

```
gden(directed, mode = "digraph")
```

```
 R >  [1] 0.5
```

```
sum(directed)/(NROW(directed) *
    (NROW(directed) -
        1))
```

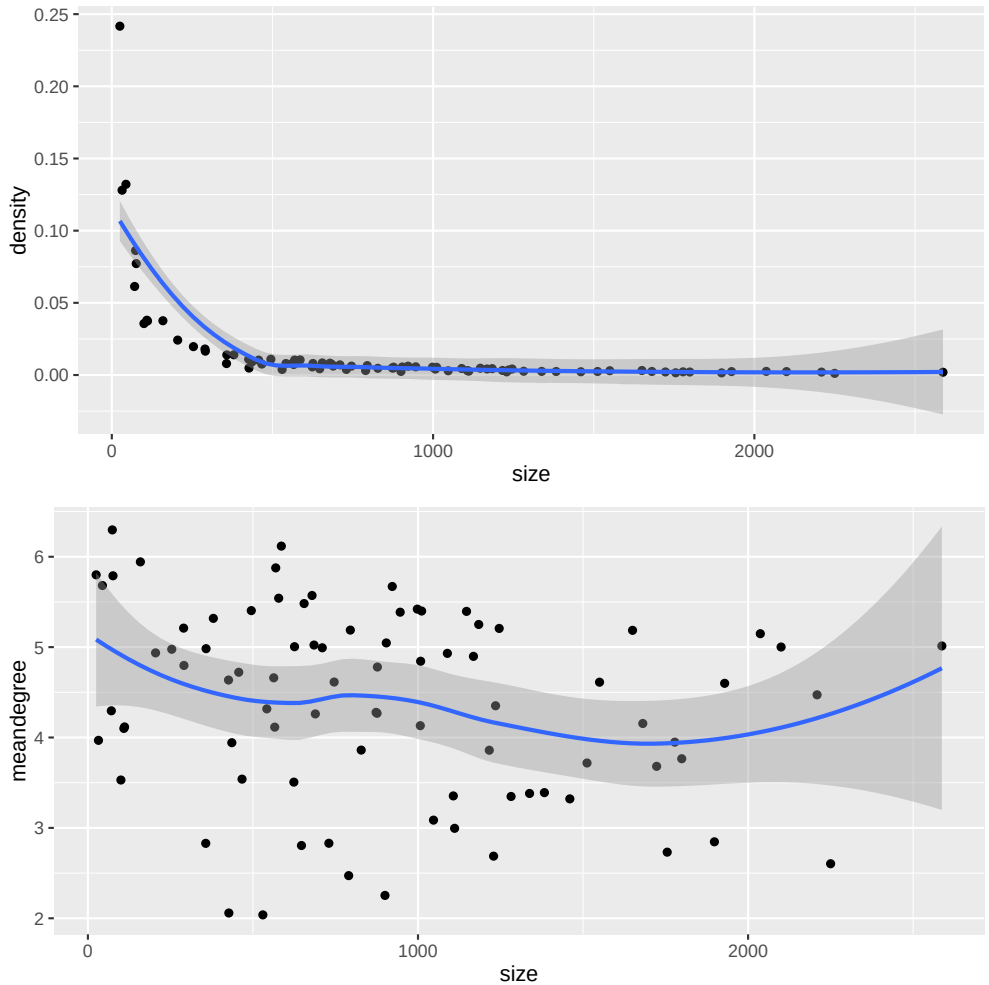
```
 R >  [1] 0.5
```

Size, Density, and Mean Degree

- Important fact: size, density, and mean degree are intrinsically related
  - Formally, $d_m = \delta(N-1)$ [I.e., mean degree = density times size-1]
  - Also, $\delta = d_m/(N-1)$ [I.e., density = mean degree over size-1]
- Simple fact, with non-obvious implications
  - If mean degree fixed, density falls with 1/group size
  - To maintain density, have to increase degree linearly, but actors can only support so many ties!
  - Thus, growing networks become increasingly sparse over time
    - Durkheim, Parsons, etc: modern social order depends on/produces norms of generalized exchange, since only tiny fraction of person can be directly related

Add Health Example

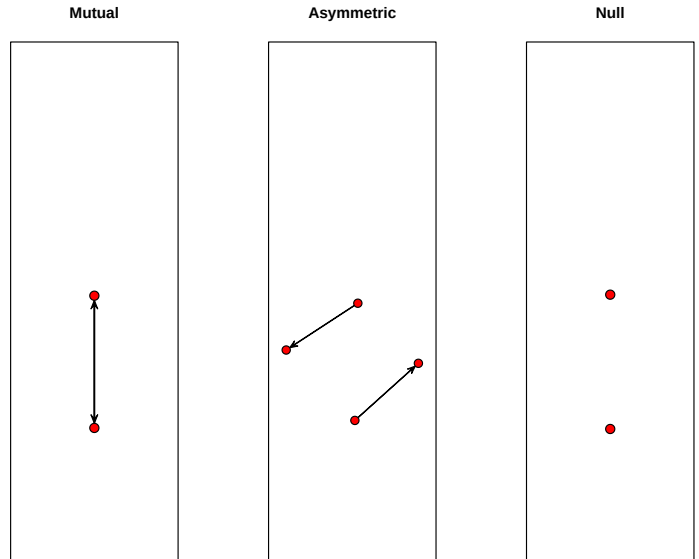
```
library(ggplot2)
library(gridExtra)
library(networkdata)
data(addhealth)
data(addhealth)
data <- data.frame(size = sapply(addhealth,
    network.size), density = sapply(addhealth,
    gden))
data$meandegree <- data$density *
    (data$size - 1)

p1 <- ggplot(data, aes(size,
    density)) + geom_point() +
    geom_smooth()
p2 <- ggplot(data, aes(size,
    meandegree)) + geom_point() +
    geom_smooth()
grid.arrange(p1, p2,
    ncol = 1)
```



Beyond Density: the Dyad Census

- Dyad census: a count of the number of mutual, asymmetric and null dyads in a network
  - Mutual: (i, j) and (j, i)
  - Asymmetric: (i, j) or (j, i), but not both
  - Null: neither (i, j) nor (j, i)
  - Traditionally written as (M, A, N)
- Used as "building block"
  - $M + A + N =$ Number of dyads
  - $2M + A =$ Number of edges
  - $(M + A/2)/(M + A + N) =$ Density

```
asym <- matrix(0, nc = 4,
    nr = 4)
asym[1, 4] <- 1
asym[3, 2] <- 1
mut <- matrix(c(0, 1,
    1, 0), nc = 2)
null <- matrix(c(0, 0,
    0, 0), nc = 2)
par(mfrow = c(1, 3))
gplot(mut, mode = "circle",
    main = "Mutual")
box()
gplot(asym, mode = "circle",
    main = "Asymmetric")
box()
gplot(null, mode = "circle",
    main = "Null")
box()
```



Reciprocity

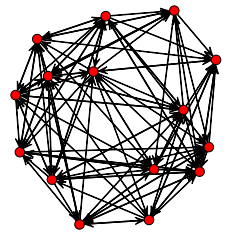
- Reciprocity: tendency for relations to be symmetric
- Several notions:
  - Dyadic: probability that any given dyad is symmetric (mutual or null)

$$\frac{M+N}{M+A+N}$$

  - Edgewise: probability that any given edge is reciprocated

$$\frac{2M}{2M+A}$$

```
graph <- rgraph(15)
gplot(graph)
```



```
dyad.census(graph)
```

```
R >       Mut Asym Null
R >  [1,]  31   49   25
```

```
grecip(graph, measure = "edgewise")
```

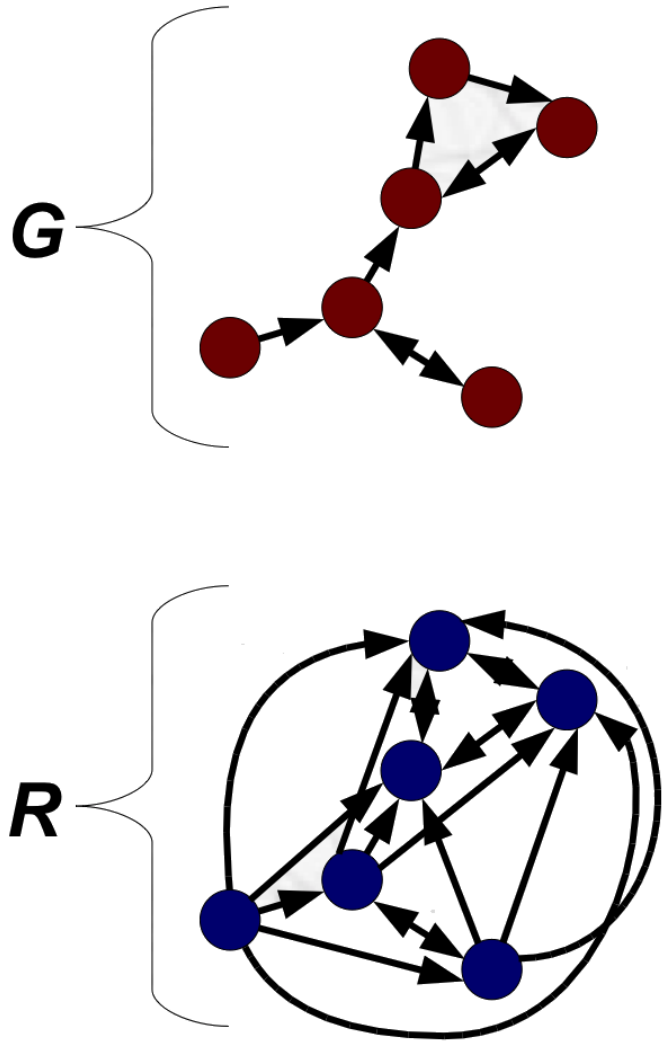
```
R >        Mut
R >  0.5585586
```

```
grecip(graph, measure = "dyadic")
```

```
R >        Mut
R >  0.5333333
```

Reachability

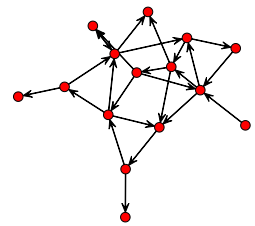
- Reachability graph
  - Digraph, R, based on G such that (i, j) is an edge in R iff there exists an i, j path in G
    - If G is undirected or fully reciprocal, R will also be fully reciprocal
  - Intuitively, an edge in R connects vertices which are connected in G
  - Strong components of G (including cycles) form cliques in R



Hierarchy

- Hierarchy: tendency for structures to be asymmetric
- As with reciprocity, many notions; for instance...
  - Dyadic Hierarchy: 1- (Dyadic Reciprocity)
    - Intuition: extent to which dyads are asymmetric
  - Krackhard Hierarchy: $1 - M/(M + A)$ in Reachability Graph
    - Intuition: for pairs which are in a contact, what fraction are asymmetric?

```
graph <- rgraph(15, tprob = 0.1)
gplot(graph)
```



```
hierarchy(graph, measure = "reciprocity")
```

```
R >        Mut
R >  0.2190476
```

```
hierarchy(graph, measure = "krackhardt")
```

```
R >  [1] 0.4607843
```

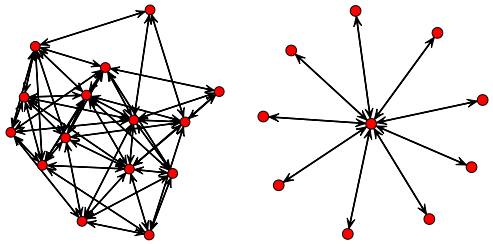
Centralization

- Centralization: extent to which centrality is concentrated on a single vertex
- Definition dut to Freeman (1979):

$$C(G) = \sum_{i=1}^{N} \left(\max_{v} c(v, G) - c(i, G) \right)$$

  - Defined for any centrality measure
  - Often used with degree, betweenness, closeness, etc.
- Most centralized structure usually star network
  - True for most centrality measures

```
graph <- rgraph(15, tprob = 0.5,
    mode = "graph")
star <- matrix(0, nc = 10,
par(mfrow = c(1, 2),
    mar = c(0, 0, 0,
        0))
gplot(graph)
gplot(star)
```



```
cbind(centralization(graph,
    g = 1, degree), centralization(graph,
    g = 1, closeness))
```

```
R >             [,1]
R >  [1,] 0.2527473
R >             [,2]
R >  [1,] 0.1234553
```

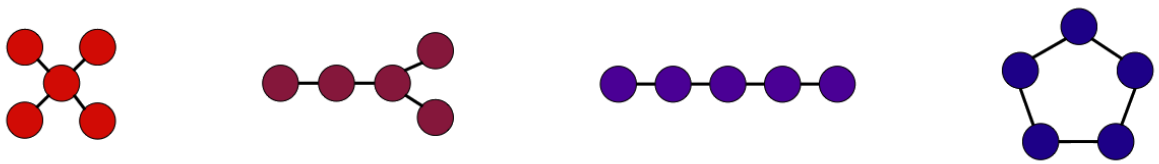
```
cbind(centralization(star,
    g = 1, degree), centralization(star,
    g = 1, closeness))
```

Centralization Versus Hierarchy

- Aren't centralization and hierarchy the same thing?
- No! Two very different ideas:
  - Hierarchy: asymmetry in interaction
  - Centralization: inequality in centrality
- Can have centralized mutual structures, hierarchical decentralized structures

Centralization and Team Performance

- Bavelas, Leavitt and others studied work teams with four structural forms:



- Performance generally highest in centralized groups
  - Star, "Y" took least time, made fewest errors, used fewest messages
- Satisfaction generally highest in decentralized groups
  - Circle>Chain>"Y">Star (but central persons had fun!)
- A lesson: optimal performance $\neq$ optimal satisfaction...

A First Look at Random Graphs

- So far, we have been measuring properties of graphs (and modeling their effects)
- Next step: modeling graphs themselves
  - Long-running and ongoing research area
  - Will crop up repeatedly in the coming weeks
- A quick introduction to some basic families
  - We'll see some uses of these model families in the next lecture...

The Notion of Random Graphs

- Let $G = (V, E)$ be a graph. If E (and perhaps V) is a random set, then G is a random graph
  - Can consider G to be a random variable on some set $\mathcal{G}$ of possible graphs ("multinomial" representation)
  - Write probability mass function (pmf) as $\Pr(G = g)$
- Let Y be the adjacency matrix of random graph G. Then Y is a random matrix
  - Write graph pmf as $\Pr(Y = y)$
  - Y_{ij} is a binary random variable which indicates the state of the (random) i, j edge
  - $\Pr(Y_{ij} = y_{ij})$ is the (marginal) proability of the Y_{ij} edge state

"Classical" Random Graphs

- Two families from the early (mathematical) literature:
  - The "N,M" family (Erdös-Rényi, size/density CUG)
  - Let M_m be the maximum number of edges in G. Then:

$$\Pr(G = g \mid N, M) = \binom{M_m}{M}^{-1}$$

  - The "N,p" family (homogeneous Bernoulli graphs)

$$\Pr(G = g \mid N, p) = p^{M}(1-p)^{M_m - M}$$

  - Both used as baseline models, but very limited
    - No heterogeneity, (almost) no dependence
    - Starting point for more complex models

Relaxation 1: Intra-dyadic Independence

- First way to build richer models: relax independence
- Intra-dyadic dependence models
  - Allow for size, density, reciprocity effects
- Two parallel models
  - Dyad census conditioned CUG $(U|MAN)$
  - Let G have fixed dyad census M, A, N. Then

$$\Pr(G = g \mid M, A, N) = \frac{M!A!N!}{(M + A + N)!}$$

  - Homogeneous dyadic multinomial family $(u|man)$

$$\Pr(G = g \mid m, a, n) = m^M a^A n^N$$

Relaxation 2: Homogeneity

- Second way to build richer models: relax homogeneity in edge probabilities
- Development for Bernoulli, multinomial cases:
- Inhomogeneous Bernoulli graph
  - Let $\Phi \in [0,1]^{N \times N}$ be a parameter matrix, and $B(X = x \mid p)$ the Bernoulli pmf.

$$\Pr(G = g \mid \Phi) = \prod_{(i,j)} B(Y_{ij} = y_{ij} \mid \Phi_{ij})$$

  - Inhomogeneous independent dyad graph
  - Let $\Phi, \Psi \in [0,1]^{N \times N}$ be parameter matrices w/ $\Phi_{ij} + \Psi_{ij} \leq 1$. Then

$$\Pr(G = g \mid \Phi, \Psi) = \prod_{(i,j)} \left[\Phi_{ij} y_{ij} y_{ji} + \Psi_{ij} (y_{ij}(1 - y_{ij}) + (1 - y_{ij} y_{ij}) + (1 - \Phi_{ij} - \Psi_{ij})(1 - y_{ij})(1 - y_{ij}) \right]$$

  - Intuitively, Φ sets the probability of mutuals, and Ψ sets the probability of asymmetrics

Simple Random Graph Models in Practice

- Models without trivial cross-dyadic dependence still have many uses
  - Baseline models for null hypothesis testing
  - Mathematical tools for exploring the space of graphs
  - Serious data models (in the inhomogeneous case)
    - Start with inhomogenous family, model parameter matrix using greression-like model (see, e.g., `sna::netlogit`) and/or with latent variables (e.g. package `latentnet`)
    - Can be extremely effective, if sufficiently strong covariates are available
  - Dyad dependent models are much more complex, but we'll see them later...

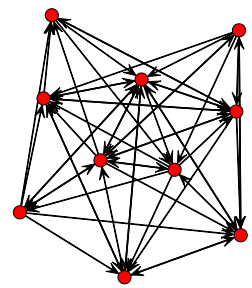
Graph-Level Properties

- Dyad census, etc.
  - Dyad census – most "foundational" structural element
- Triad and transitivity
  - Smallest nontrivial structural units
- Null hypothesis testing for GLIs
  - Application of random graphs

The Subgraph Census

- The dyad census counted al copies of the three dyadic isomorphism classes (M,A,N)
  - $G \cong H$ if there exists some relabeling f of $V(G)$: (u, v) in $E(G)$ iff $(f(u), f(v))$ in $E(H)$
- Generalization: subgraph census
  - Count of all subgraphs of a particular form (e.g., 2-stars, 3-cycles, etc)
  - Induced subgraph census: count of all induced subgraphs of a particular form (e.g., incomplete 2-paths – nulls count!)
  - Way of measuring network "composition"

```
g <- rgraph(10)
gplot(g)
```



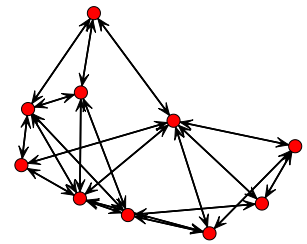
```
dyad.census(g)
```

```
R >       Mut Asym Null
R >  [1,]  15   23    7
```

From Dyads to Triads

- Famous observation of Simmel: triads differ fundamentally from dyads
  - Possibliity for exclusion, alliance formation, etc.
- Structurally, triads are the first inseparable order
  - Dyads are separable: the edges of distinct $\{i, j\}$ and $\{k, l\}$ do not overlap
  - Triads are inseparable: every $\{i, j, k\}$ triad intersects $3(N-3)$ other triads
- Triads are thus a "bridge" from local to global structure

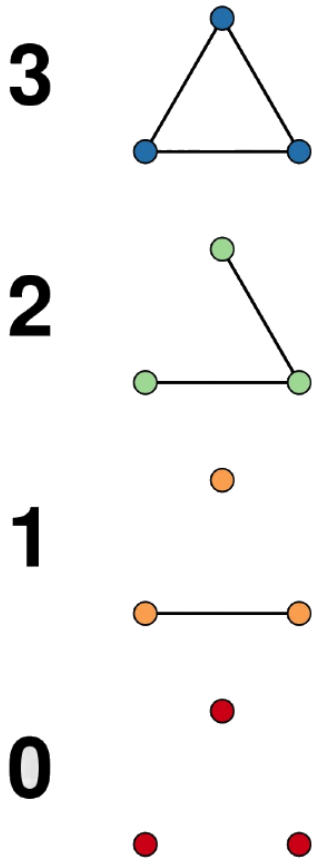
```
g <- rgraph(10, mode = "graph")
gplot(g)
```

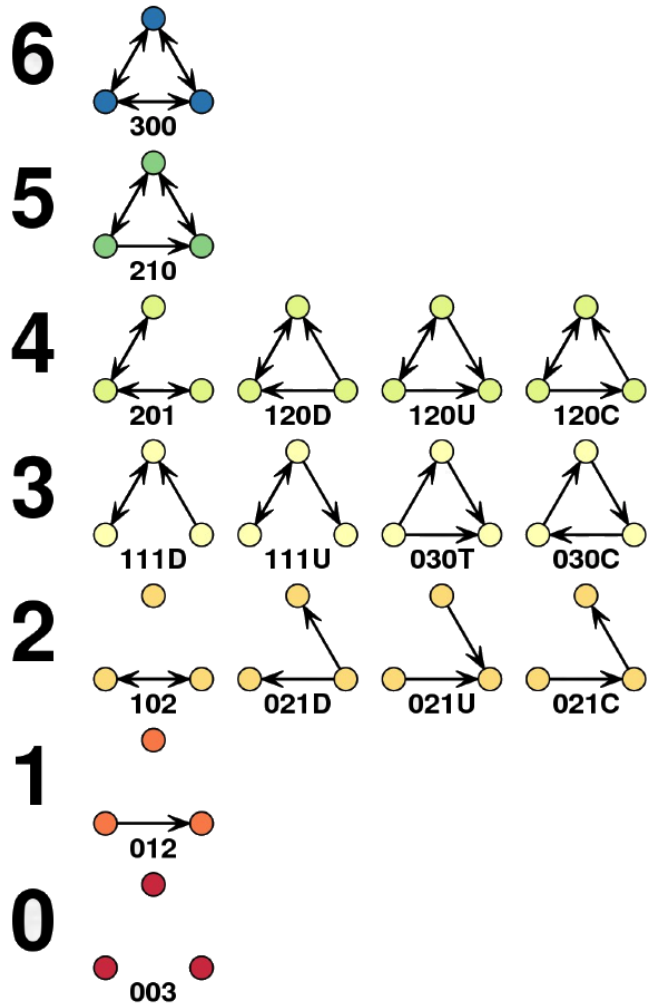


```
triad.census(g)
```

```
R >       003 012 102
R >  [1,]  15   0  53
R >       021D 021U 021C
R >  [1,]    0    0    0
R >       111D 111U 030T
R >  [1,]    0    0    0
R >       030C 201 120D
R >  [1,]    0  41    0
R >       120U 120C 210
R >  [1,]    0    0   0
R >       300
R >  [1,]  11
```

The Triad Census





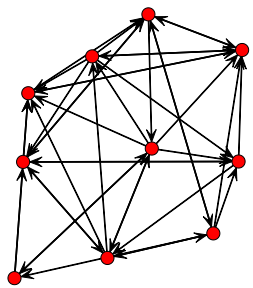
Undirected Case

Directed Case

Transitivity

- Transitivity: condition such that (i, j) and (j, k) implies (i, k)
  - Usually assessed as fraction of completed two-paths
  - Can be directed or undirected
  - In directed, case, same as triangle/3-cycle/3-clique
- Many Uses
  - Measure of clustering
  - Also used to indicate dominance, hierarchy

```
g <- rgraph(10)
gplot(g)
```



```
gtrans(g, measure = "weak")
```

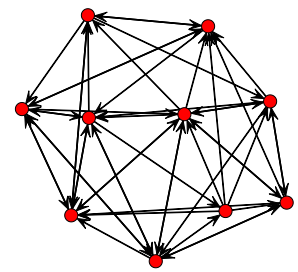
```
R >  [1] 0.3435115
```

`gtrans` with `measure="weak"`: The fraction of potentially intransitive triads obeying the weak condition $(a \to b \to c \implies a \to c)$.

The "Forbidden Triad" and Weak Ties

- Granovetter (following Heider): strong ties must be transitive
  - Weak ties, not so much
- Implications
  - No incomplete strong two-paths ("forbidden triad")
  - Strong ties form cliques
  - All bridging ties are weak ties
  - Thus, weak ties can play an important role in bridging social systems

```
g <- rgraph(10)
gplot(g)
```



```
triad.census(g)
```

```
R >       003 012 102
R >  [1,]   0   9   5
R >       021D 021U 021C
R >  [1,]    6    2    6
R >       111D 111U 030T
R >  [1,]   20   12    8
R >       030C 201 120D
R >  [1,]    6   6    6
R >       120U 120C 210
R >  [1,]    6    9  15
R >       300
R >  [1,]   4
```

Transitivity and the Triad Census

- Transitivity bias implies that some triads should be over/underrepresented
- Weighting vector approach (H+L)
  - Let t be the triad census vector, and w be a "weighting vector" counting number of configurations satisfying a given hypothesis
  - $\tau = t^T w$ then counts the number of configurations present
  - Comparing τ to a null distribution allows one to test various hypotheses about structure
    - $H + L$ use Gaussian approximation, but we can use direct computation (which is much safer)

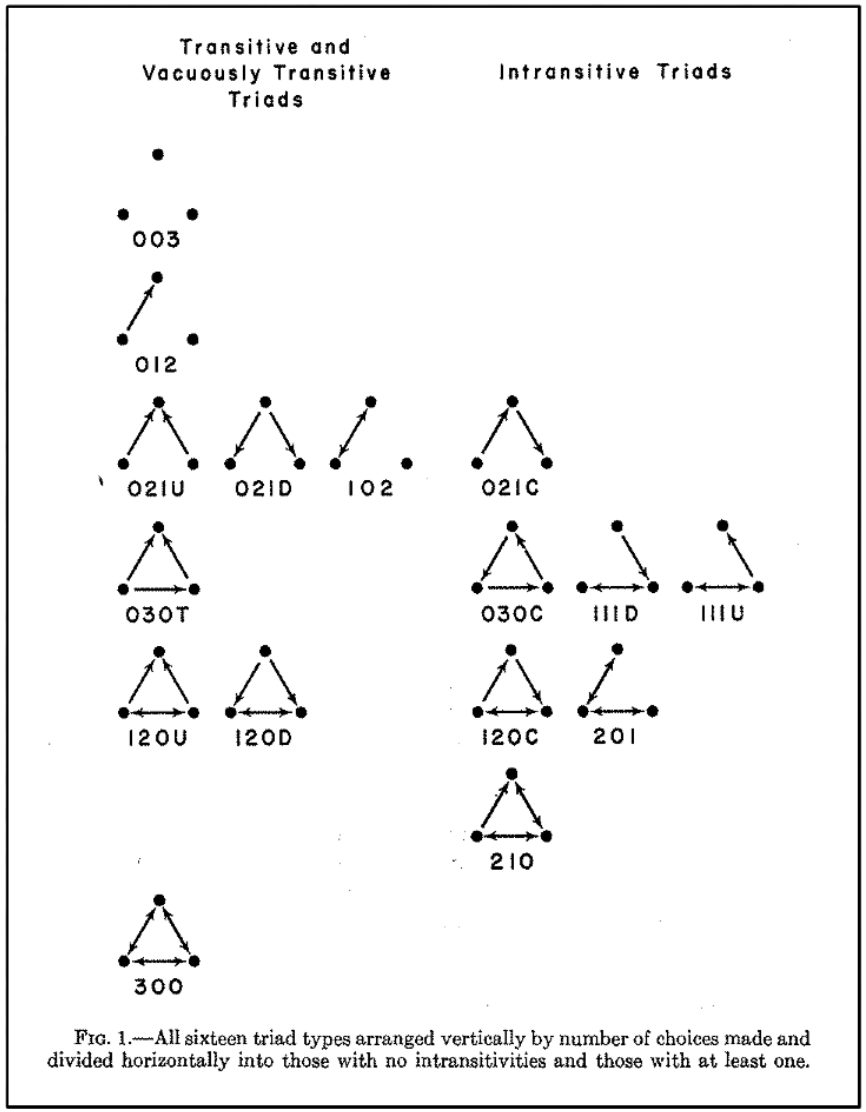


FIG. 1.—All sixteen triad types arranged vertically by number of choices made and divided horizontally into those with no intransitivities and those with at least one.

Null Hypothesis Testing for GLIs

Null Models

- Last time, we introduced random graph models
  - Focused on simple cases with little or no heterogeneity or dependence among edges
- These models can be used as null models when assessing network structure
  - Goal is the detection of structural biases (departures from purely random structure)
    - Confirmatory use: specific biases predicted by prior theory
    - Exploratory use: identifying which biases are present, before attempting to build more complex models
  - Compare to the role of standard null hypothesis testing in the conventional statistics literature
    - Same basic idea, although some philosophical differences

Initial Concept: Baseline Models

- Baseline model: model which treats social structure as random, given some fixed constraints
  - Constraints could include size, density, etc.
- Method of baseline models (from Mayhew)
  - Identify potentially constraining factors
  - Compare observed properties to baseline model
    - Interpret deviations from baseline
    - May repeat with additional constraints
  - Note similarity to classical null hypothesis testing
    - Baseline model acts as null hypothesis
  - Useful even when baseline model is not "realistic"
    - Emphasis on triangulation to identify nature of biases; multiple baselines may be used to "pin down" complex behavior

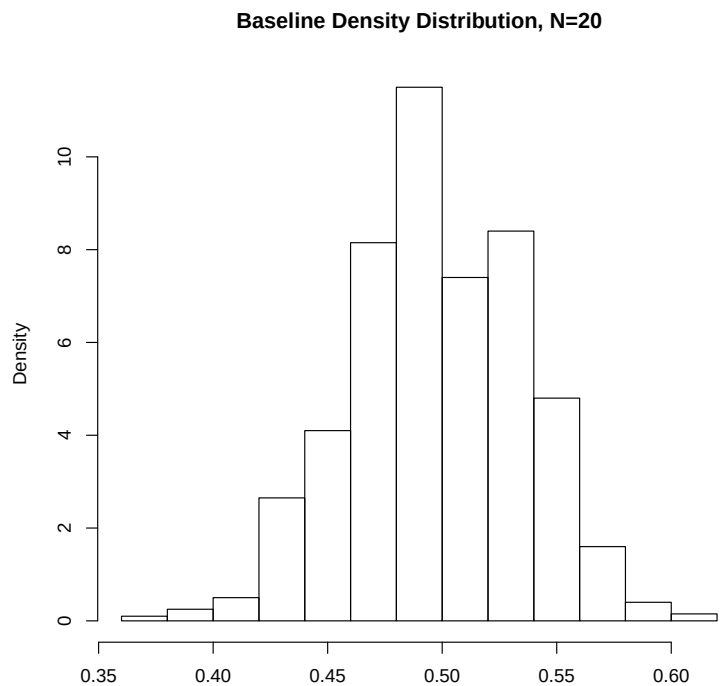
A Few Baseline Models

- Uniform conditional on size
  - Given number of individuals, all structures taken to be equally likely
- Uniform conditional on number of edges
  - Given number of individuals and interactions, structure is random (fixed density)
  - Valued version: condition on edge values (randomize over who gets edges)
- Uniform conditional on dyad census
  - Given number of individuals, mutual, asymmetric, and null relationships, structure is random (fixes density and reciprocity)
  - Valued verison: condition on dyad pair values (randomize over who belongs to each dyad)
- Uniform conditional on all unlabeled properties
  - This the permutation distribution.

Comparing Observations to Baseline

- To compare observations w/baseline behavior, must choose a statistic to evaluate
  - Should choose a statistic which reflects the type of property being examined
  - Obviously, cannot use a statistic on which one is conditioning
- Next, generate distribution of statistic under baseline model
  - Simulate networks from baseline model, then calculate statistic
- Finally, compare observed statistic to baseline distribution

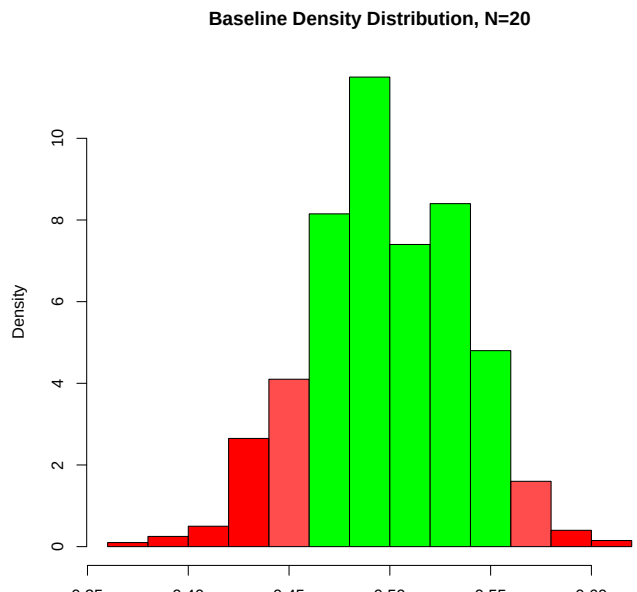
```
library(sna)
baselineDensity <- sapply(1:1000,
    function(x) {
        gden(rgraph(20,
            mode = "graph"))
    })
save <- hist(baselineDensity,
    main = "Baseline Density Distribution, N=20",
    xlab = "Density",
    probability = TRUE,
    breaks = 10)
```



Looking High and Low

- Primary quantities of interest: probabilities of obtaining values greater than/equal to or less than/equal to observed value under baseline model
  - $\Pr(s(G) \leq s(G_{obs})) \approx 0$ implies that $s(G_{obs})$ is large compared to baseline
    - Small chance of observing a value that large under baseline
  - $\Pr(s(G) \geq s(G_{obs})) \approx 0$ implies that $s(G_{obs})$ is small compared to baseline
    - Small chance of observing a value that small under baseling

```
qt <- quantile(baselineDensity,
    prob = seq(0, 1,
        0.025))[c("2.5%",
    "10%", "90%", "97.5%")]
col <- rep(rgb(0, 1,
    0, 1), length(save$breaks))
col[save$breaks <= qt[1]] <- col[save$breaks >=
    qt[4]] <- rgb(1,
    0, 0, 1)
col[save$breaks <= qt[2] &
    save$breaks > qt[1]] <- col[save$breaks >=
    qt[3] & save$breaks <
    qt[4]] <- rgb(1,
    0, 0, 0.7)
hist(baselineDensity,
    main = "Baseline Density Distribution, N=20",
    xlab = "Density",
    probability = TRUE,
    breaks = 10, col = col)
```



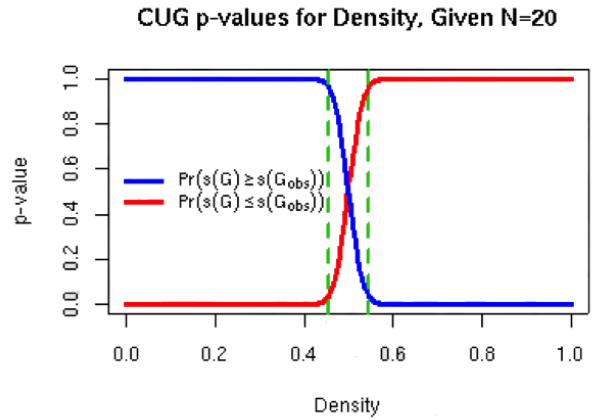
Conditional Uniform Graph Tests

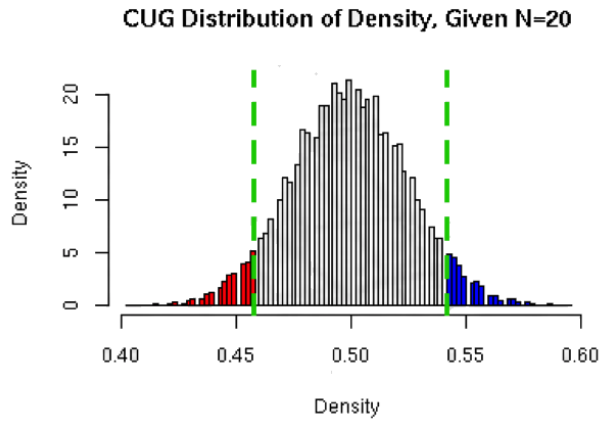
- CUG test: use of conditional uniform baseline as a null hypothesis test
  - Propose that $s(G_{obs})$ is drawn from a baseline model
  - Reject at significance level p if

    $\Pr(s(G) \leq s(G_{obs})) < p$ (upper tail) or

    $\Pr(s(G) \geq s(G_{obs})) < p$ (lower tail) or

  - Convential significance levels 0.05, 0.01, 0.001
- Interpretation
  - Rejection: Data shows noteworthy departure from model
  - Non-rejection: Data consistent with baseline model
  - Direction indicates nature of deviation





CUG Tests in R

- Can perform CUG tests the "hard way"
  - Use tools like rgraph, rguman, rgnm to simulate baseline draws
- Easier way: use the cug.test helper function
  - Can perform tests using several CUG baseline hypotheses

    - `cug.test` syntax

```
args(cug.test)
```

```
 R >  function (dat, FUN, mode = c("digraph", "graph"), cmode = c("size",
 R >      "edges", "dyad.census"), diag = FALSE, reps = 1000, ignore.eval = TRUE,
 R >      FUN.args = list())
 R >  NULL
```

Example: International Trade in Complex Manufactured goods

- First, test density against size-conditioned model

```
library(networkdata)
data(trade)
cden <- cug.test(trade$MANUFACTURED_GOODS,
    gden)
cden
```

```
 R >
 R >  Univariate Conditional Uniform Graph Test
 R >
 R >  Conditioning Method: size
 R >  Graph Type: digraph
 R >  Diagonal Used: FALSE
 R >  Replications: 1000
 R >
 R >  Observed Value: 0.5615942
 R >  Pr(X>=Obs): 0.001
 R >  Pr(X<=Obs): 0.999
```

- Next, test reciprocity against density-conditioned model

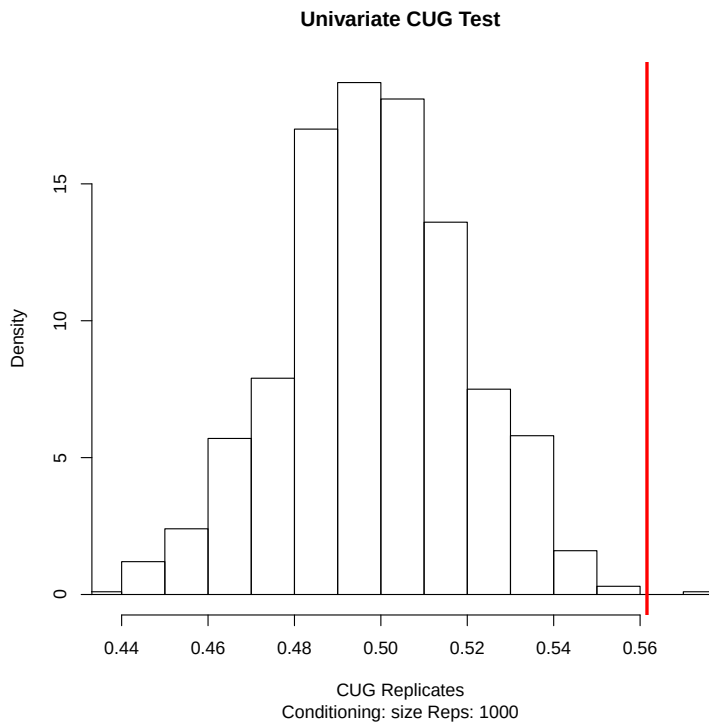
```
crecip <- cug.test(trade$MANUFACTURED_GOODS,
    grecip, cmode = "edges")
crecip
```

```
 R >
 R >  Univariate Conditional Uniform Graph Test
 R >
 R >  Conditioning Method: edges
 R >  Graph Type: digraph
 R >  Diagonal Used: FALSE
 R >  Replications: 1000
 R >
 R >  Observed Value: 0.7101449
 R >  Pr(X>=Obs): 0
 R >  Pr(X<=Obs): 1
```

Example: International Trade in Complex Manufactured goods

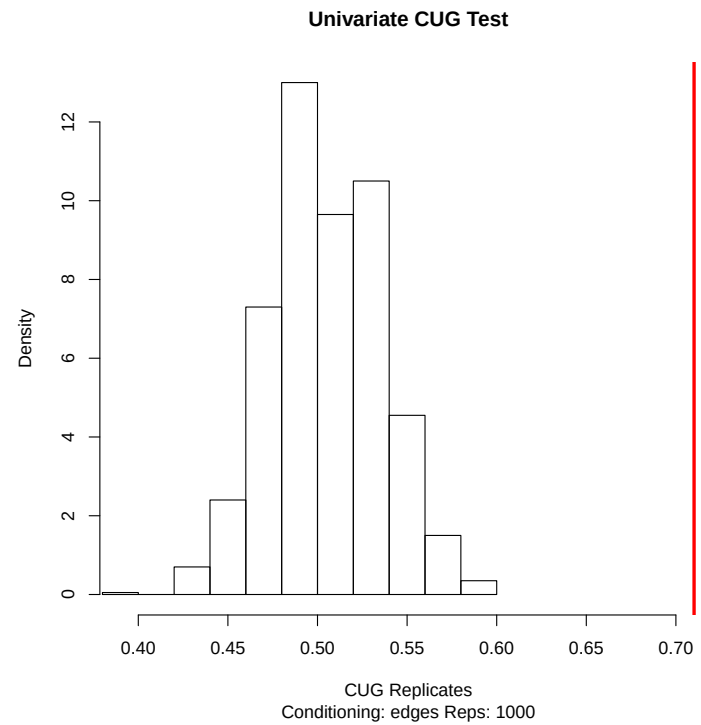
- First, test density against size-conditioned model

```
plot(cden)
```



- Next, test reciprocity against density-conditioned model

```
plot(crecip)
```



Example: International Trade in Complex Manufactured goods

- Now, test transitivity given the dyad census

```
library(networkdata)
data(trade)
ctrans <- cug.test(trade$MANUFACTURED_GOODS,
    gtrans, cmode = "dyad")
ctrans
```

```
 R >
 R >  Univariate Conditional Uniform Graph Test
 R >
 R >  Conditioning Method: dyad.census
 R >  Graph Type: digraph
 R >  Diagonal Used: FALSE
 R >  Replications: 1000
 R >
 R >  Observed Value: 0.7860323
 R >  Pr(X>=Obs): 0
 R >  Pr(X<=Obs): 1
```

- Repeat for indegree centralization

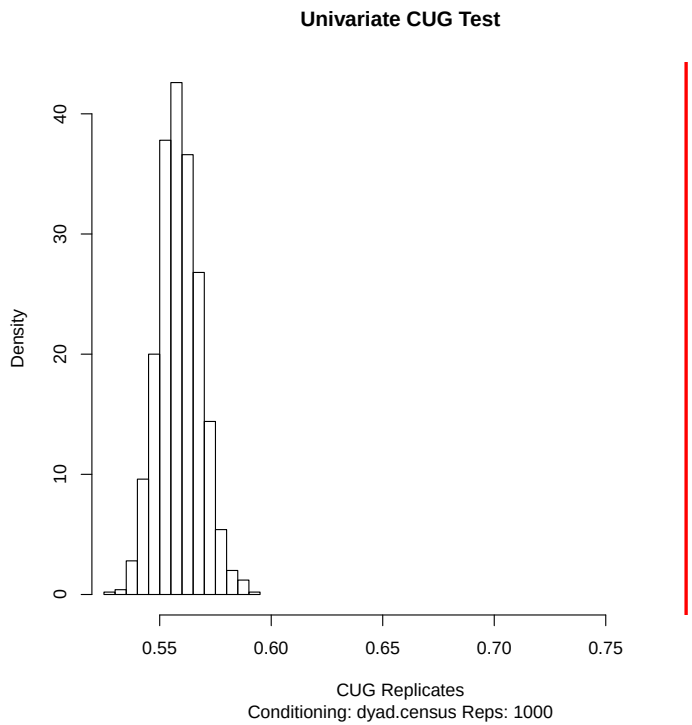
```
c_cent_ind <- cug.test(trade$MANUFACTURED_GOODS,
    centralization, cmode = "dyad",
    FUN.arg = list(FUN = degree,
        cmode = "indegree"))
c_cent_ind
```

```
 R >
 R >  Univariate Conditional Uniform Graph Test
 R >
 R >  Conditioning Method: dyad.census
 R >  Graph Type: digraph
 R >  Diagonal Used: FALSE
 R >  Replications: 1000
 R >
 R >  Observed Value: 0.2306238
 R >  Pr(X>=Obs): 0.452
 R >  Pr(X<=Obs): 0.872
```

Example: International Trade in Complex Manufactured goods

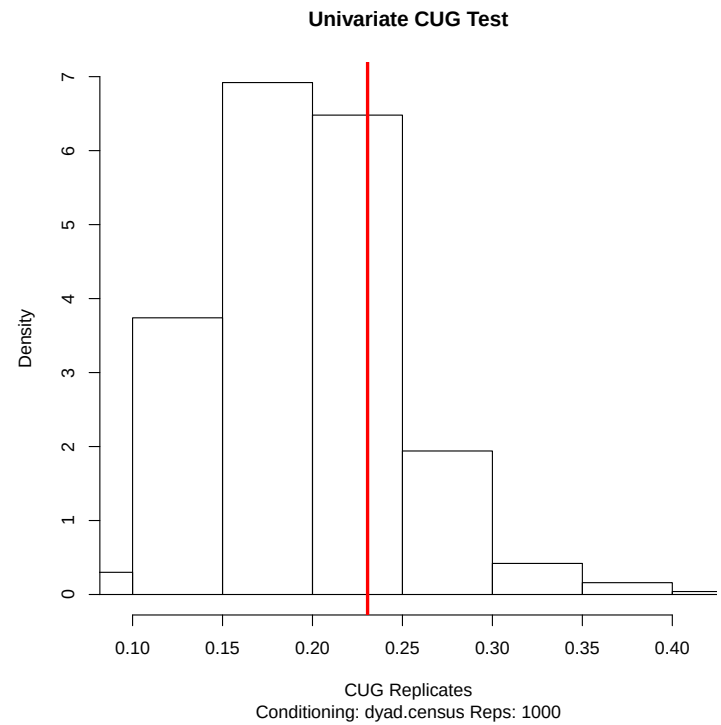
- Now, test transitivity given the dyad census

```
plot(ctrans)
```



- Repeat for indegree centralization

```
plot(c_cent_ind)
```



Advanced Topics

- Exponential Random Graph Models
- Temporal Network Models
- Spatial Network Models
- Cluster Analysis
- Latent Space Models
- Hierarchical Models (BERGM, HERGM, multilevel models)