Social Network Analysis: Data Collection and Descriptives Part 1

EPIC - SNA, Columbia University

Zack W Almquist

June 12th, 2018

University of Minnesota

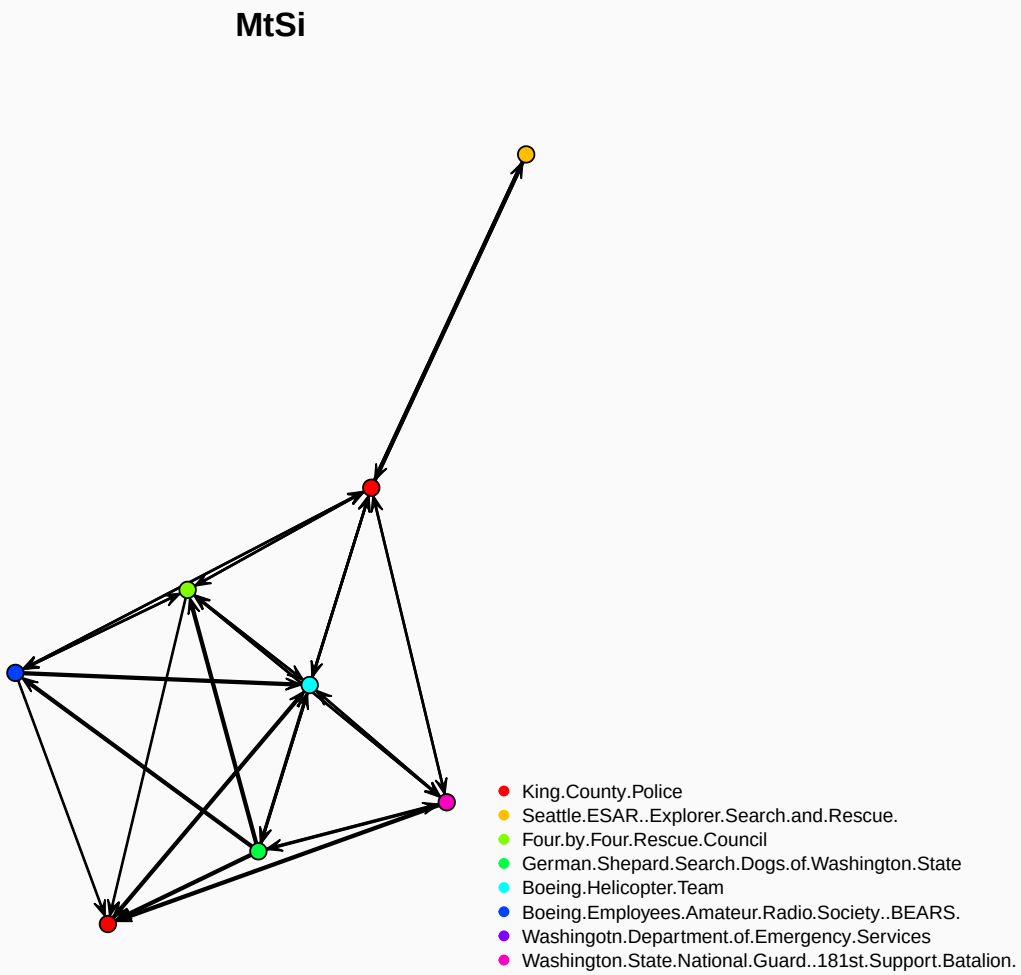
Network Data Collection

Descriptives: What are Descriptive Statistics for Networks?

R for Descriptive Analysis: An Introduction

Network Data Collection

One-mode data



One-mde data

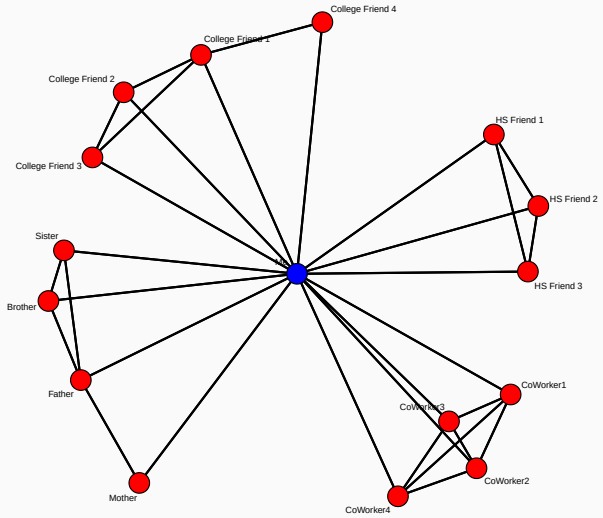
	1	2	3	4	5	6	7	8
1	0.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
2	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00
4	1.00	1.00	1.00	0.00	0.00	1.00	0.00	0.00
5	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00
6	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
7	1.00	0.00	0.00	0.00	0.00	0.00	0.00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

Table 1: Mt. Si SAR EMON, Confirmed Ties

Special Case: Egocentric network

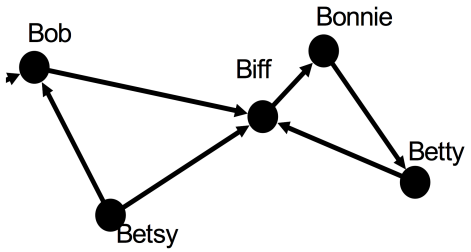
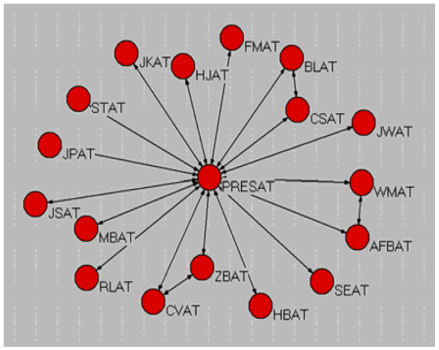
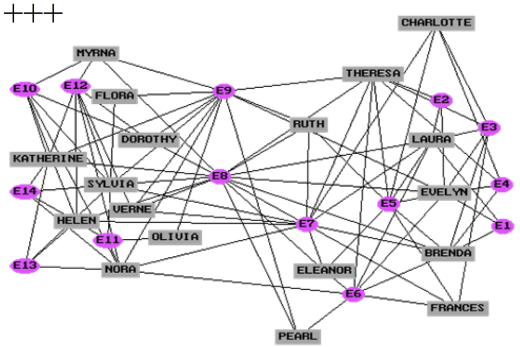
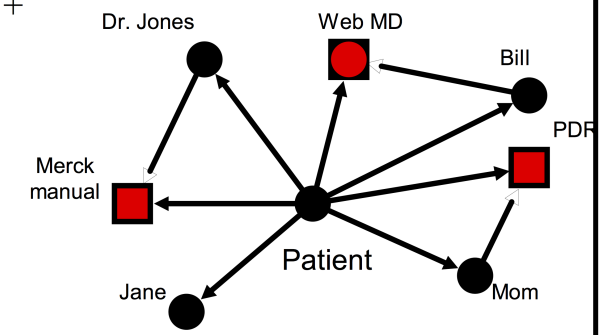
Egocentric network: focal actor ("ego") + neighbors ("alters") + ties among alters

- **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*



Modes of Data: Review

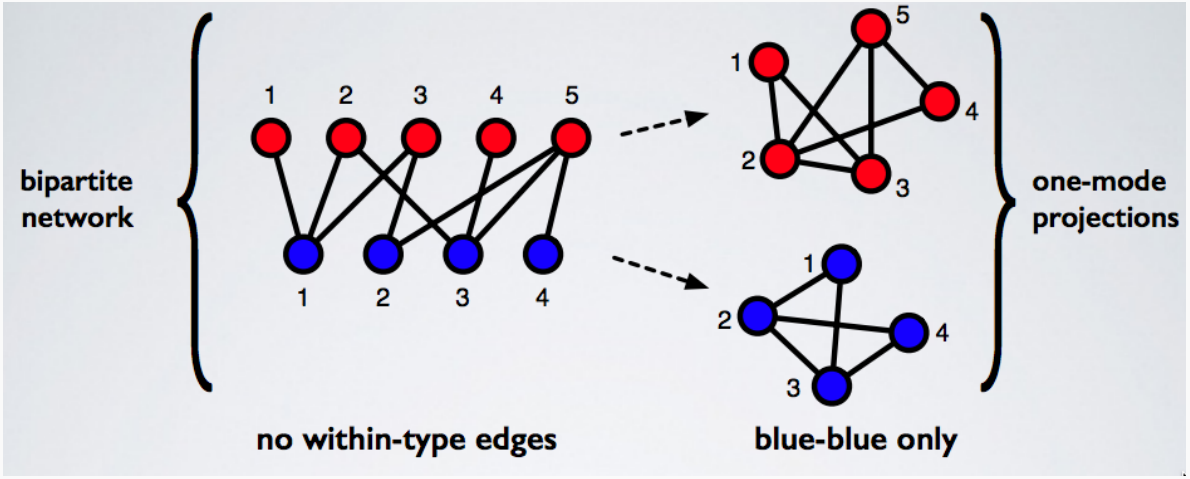
Kinds of Network Data

	Complete	Ego
1-mode	+++++ 	+++++ 
2-mode	+++ 	+ 

Credit: ??

Two mode data

- Networks with two vertex class
  - Different entity types
  - Membership
  - Matching/containment
- Represented by incidence matrices
  - "Senders" on rows, "receivers" on columns
- Can be used to obtain "dual" representations



Aaron Clauset. 2015. A short introduction to networks. Santa Fe Institute.

Two mode data: Projecting into one mode data

- 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

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

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, B = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}, \text{ and } C = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$$

Two mode data: Projecting into one mode data

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

Two mode data: Projecting into one mode data

What the Projections Mean

- 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.

Levels of Analysis

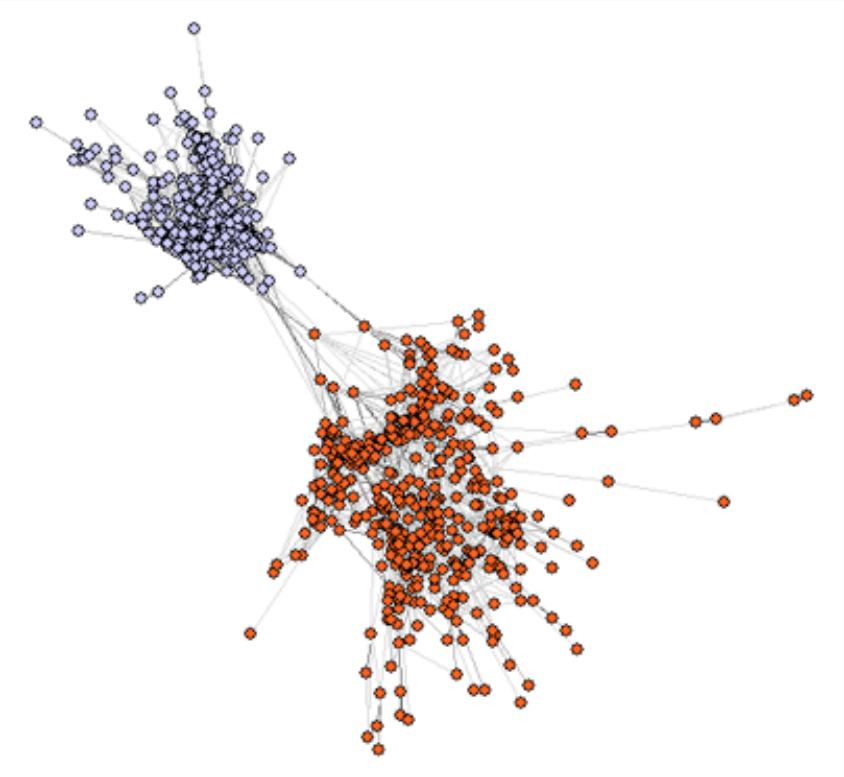
What scope of information do you want?

- Boundary Specification: key is what constitutes the "edge" of the network

	LOCAL	GLOBAL
"Realist" (Boundary from actors' Point of view)	Everyone connected to ego in the relevant manner (all friends, all (past?) sex partners)	All relations relevant to social action ("adolescent peers network" or "Ruling Elite")
Nominalist (Boundary from researchers' point of view)	Relations defined by a name-generator, typically limited in number ("5 closest friends")	Relations within a particular setting ("friends in school" or "votes on the supreme court")

Levels of Analysis

Boundary Specification Probblem



While students were given the option to name friends in the other school, they rarely do. As such, the school likely serves as a strong substantive boundary

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 Egonet

- 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

Carter Butts. Social Network Methodology. University of California, Irvine.

GSS: Important Matters

- Famous Example: General Social Survey

  From time to time, most people discuss important matters with other people. Looking back over the last six months—who are the people with whom you discussed matters important to you? Just tell me their first names or initials.

- IF LESS THAN 5 NAMES MENTIONED, PROBE: Anyone else?

GSS: Important Matters

- Famous Example: General Social Survey

  From time to time, most people discuss important matters with other people. Looking back over the last six months—who are the people with whom you discussed matters important to you? Just tell me their first names or initials.

- Survey

GSS: Important Matters

Basic findings

McPherson, Smith-Lovin, Brashears, "Social Isolation in America: Changes in Core Discussion Networks over Two Decade" ASR 2006

- The number of people saying there is no one with whom they discuss important matters nearly tripled. The mean network size decreases by about a third (one confidant), from 2.94 in 1985 to 2.08 in 2004.
- The modal respondent now reports having no confidant; the modal respondent in 1985 had three confidants

GSS: Important Matters

Basic findings

Small et al. "How stable is the core discussion network." Social Networks 2015.

1. "We found that when actors enter new institutional environments, their core discussion network changes rather quickly"
2. "Our findings are consistent with the idea that the core discussion network may include people who are not close associates or intimates."

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

Carter Butts. Social Network Methodology. University of California, Irvine.

Designs: 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: RDS

- 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

Carter Butts. Social Network Methodology. University of California, Irvine.

Example: Link-tracing

The Data

The data was aggregated by Martina Morris (University of Washington) and Richard Rothenberg (Emory University) and put online at ICPSR. The original data can be found here. In this exercise we are goint to investigate four networks derived from the Rural Arizona risk networks in Flagstaff, AZ. These networks were collected from May 1996 to Jan 1998 and originally had 95 respondents interviewed 5 times each. All participants are over 18 years old.

Example: Link-tracing

Instrument

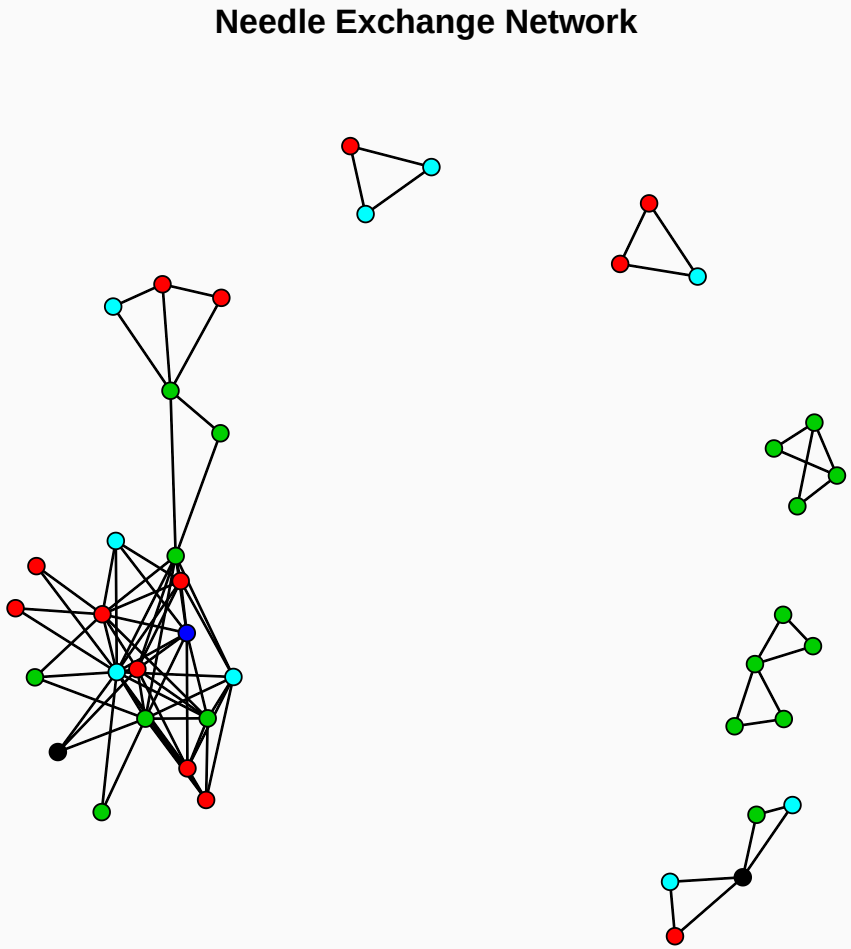
- Name generator
  - Sex, needle, other (illicit) drug contact, social contact in last 6 months
  - Sampling strategy
    - Six seeds chosen at random within same geographic area (Flagstaff) from persons presumed to be at elevated risk for HIV acquisition (through sex and/or drug behaviors)

Designs: Link-tracing

```
addr<-"https://github.com/zalmquist/ERGM_Lab/raw/master/data/flagstaff_rural.rda"
load(url(addr))
```

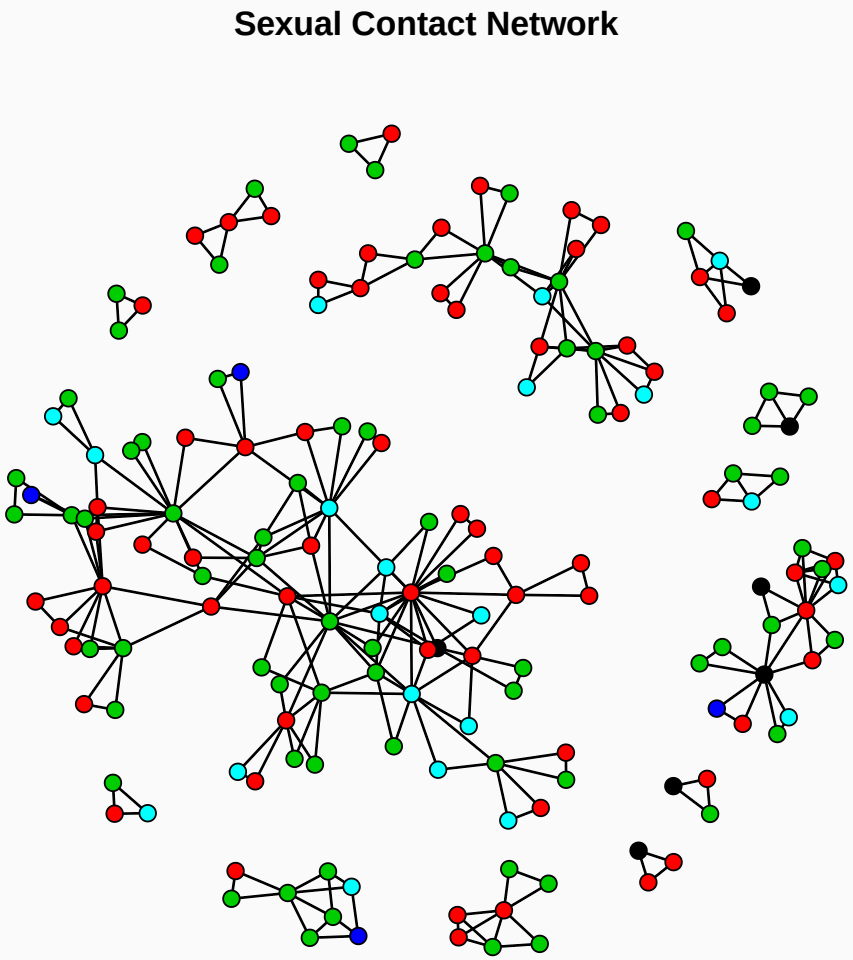
Designs: Link-tracing

```
plot(flag_needle_net, vertex.col = "race",
     main = "Needle Exchange Network")
```



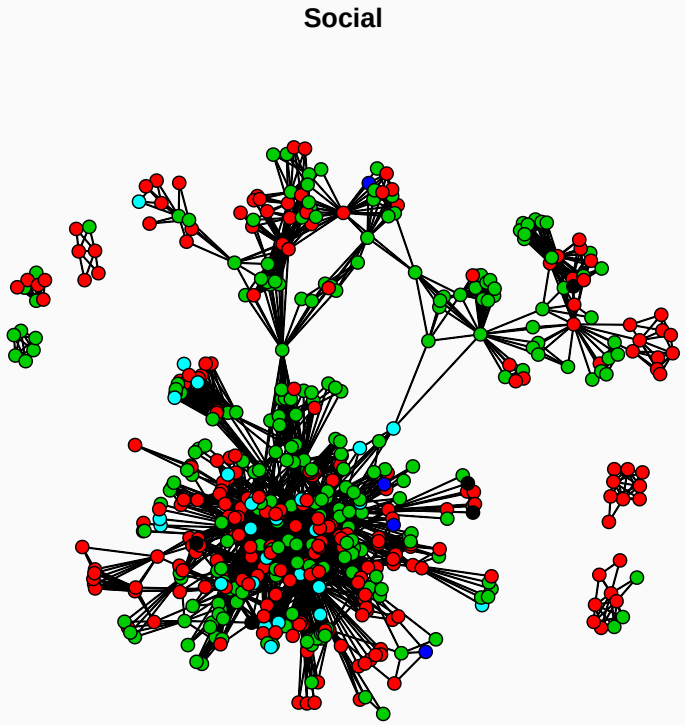
Designs: Link-tracing

```
main <- "Sexual Contact Network"
plot(flag_sex_net, vertex.col = "race", main = main)
```



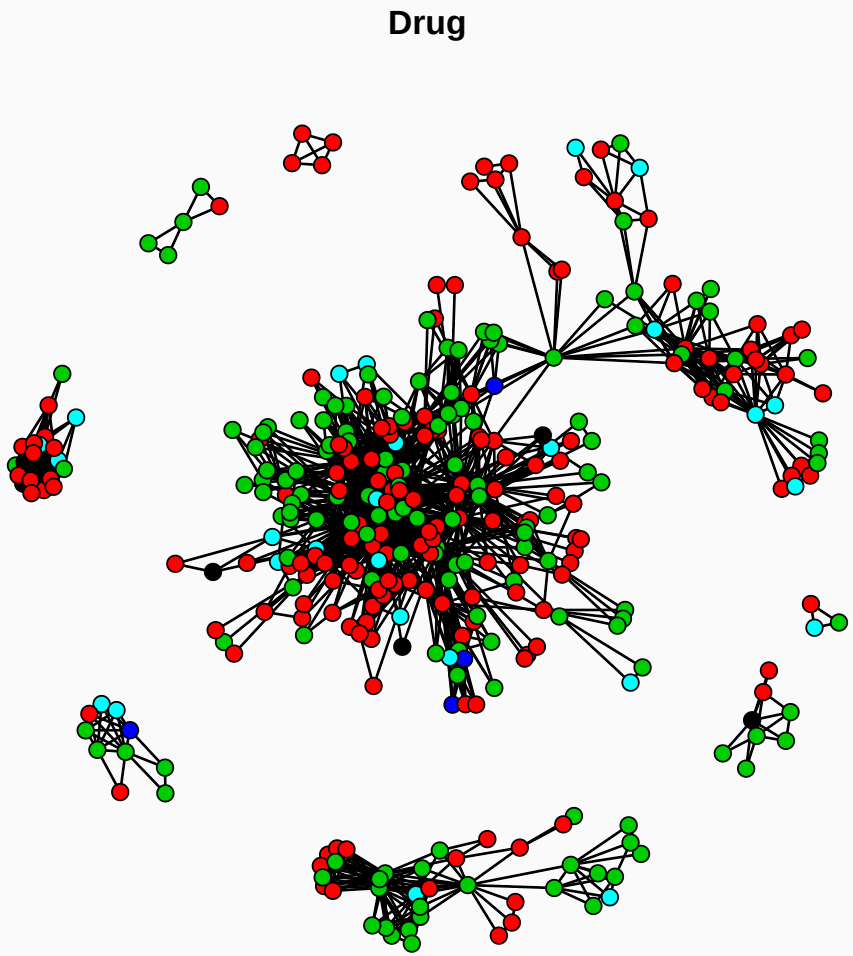
Designs: Link-tracing

```
plot(flag_social_net, vertex.col = "race",
     main = "Social")
```



Designs: Link-tracing

```
plot(flag_drug_net, vertex.col = "race",
     main = "Drug")
```



Designs: CSS

- 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

Carter Butts. Social Network Methodology. University of California, Irvine.

Designs: CSS

The Data David Krackhardt collected cognitive social structure data from 21 management personnel in a high-tech, machine manufacturing firm to assess the effects of a recent management intervention program. The relation queried was

- "Who does X go to for advice and help with work?" (`krackad`)
- "Who is a friend of X?" (`krackfr`).

Each person indicated not only his or her own advice and friendship relationships, but also the relations he or she perceived among all other managers, generating a full 21 by 21 matrix of adjacency ratings from each person in the group.

Designs: CSS

```
library(networkdata)
data(krack)
length(krack[[1]])
```

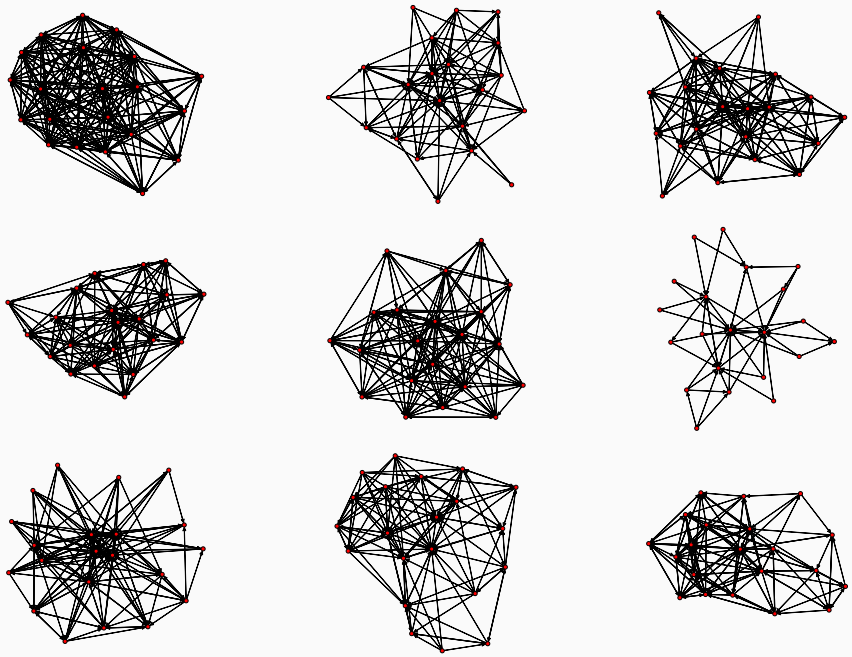
```
[1] 21
```

```
length(krack[[2]])
```

```
[1] 21
```

Designs: CSS

```
par(mfrow = c(3, 3), mar = c(0, 0, 0, 0) +
    0.1)
for (i in 1:9) plot(krack[[1]][[i]])
```



Designs: CSS Analysis

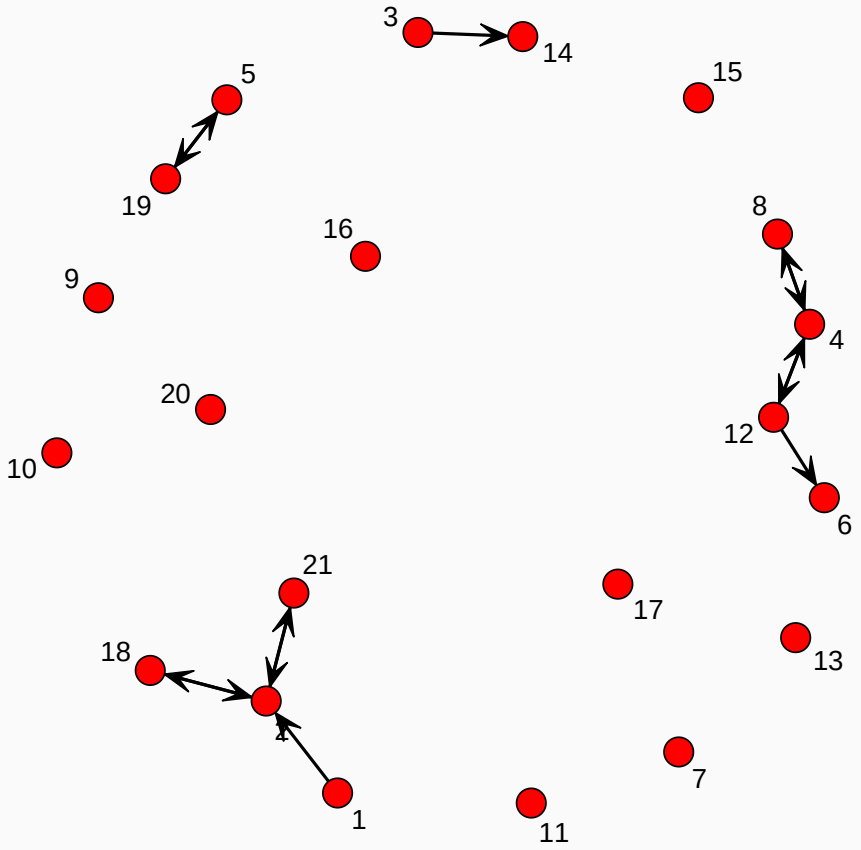
```
kfr <- as.sociomatrix.sna((krack$krackfr))
np <- matrix(0.5, 21, 21)  # 21 x 21 matrix of Bernoulli p
emp <- sapply(c(3, 11), rep, 21)  # Beta(3,11) priors for
epp <- sapply(c(3, 11), rep, 21)  # Beta(3,11) priors for
```

```
kfr.post.fixed <- bbnam.fixed(kfr, nprior = np,
    em = 3/(3 + 11), ep = 3/(3 + 11))
kfr.post.pooled <- bbnam.pooled(kfr, nprior = np,
    em = emp[1, ], ep = epp[1, ])
kfr.post.actor <- bbnam.actor(kfr, nprior = np,
    em = emp, ep = epp)
```

Butts, C. T. (2003). Network inference, error, and informant (in) accuracy: a Bayesian approach. social networks, 25(2), 103-140.

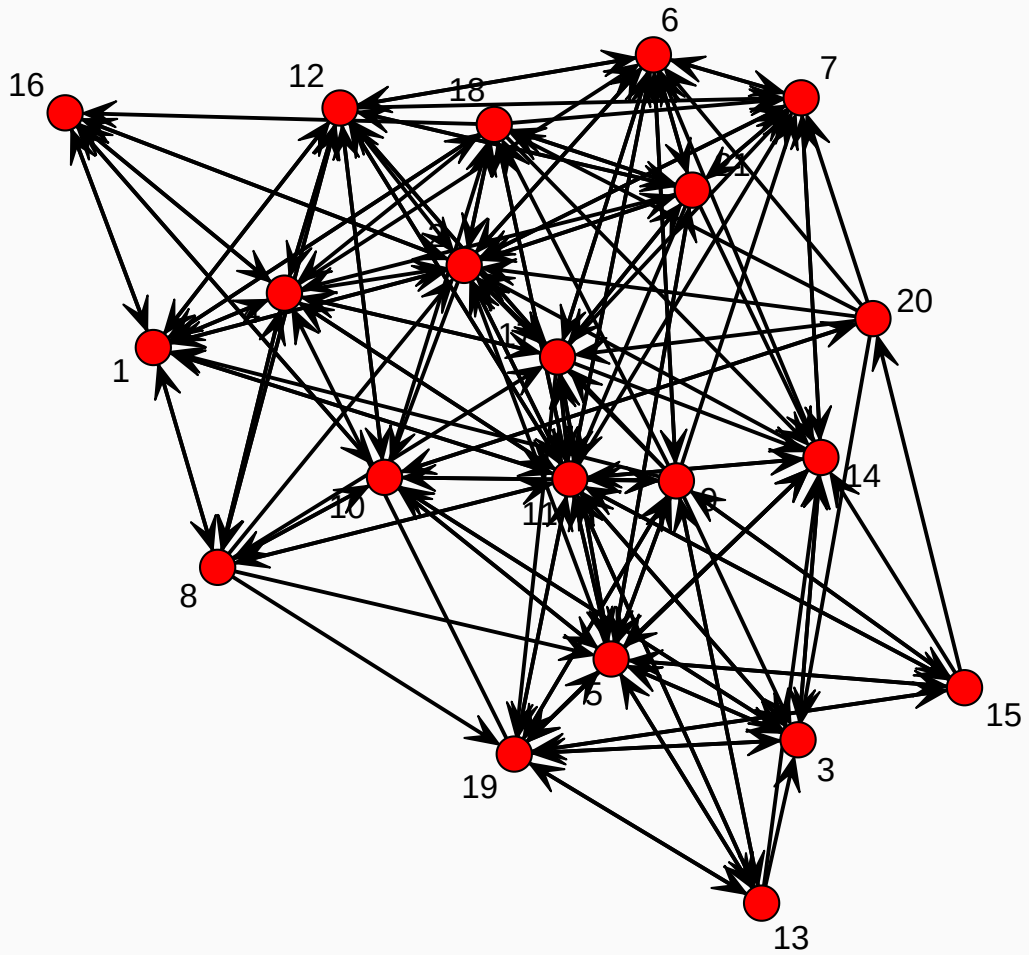
Designs: CSS

```
gplot(apply(kfr.post.fixed$net, c(2, 3),
    median), displaylabels = TRUE, boxed.lab = FALSE)
```



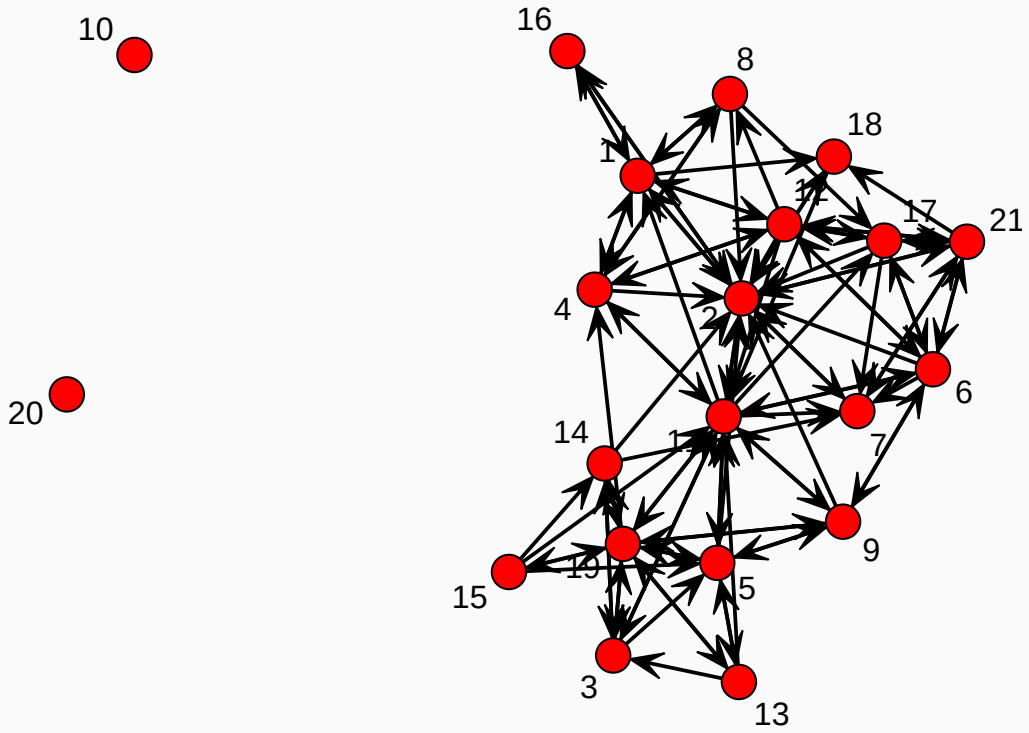
Designs: CSS

```
gplot(apply(kfr.post.pooled$net, c(2, 3),
    median), displaylabels = TRUE, boxed.lab = FALSE)
```



Designs: CSS

```
gplot(apply(kfr.post.actor$net, c(2, 3),
    median), displaylabels = TRUE, boxed.lab = FALSE)
```



Designs: 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

Carter Butts. Social Network Methodology. University of California, Irvine.

Designs: Coding Schemes as "Instruments"

- 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

Carter Butts. Social Network Methodology. University of California, Irvine.

Designs: Coding Schemes as "Instruments"

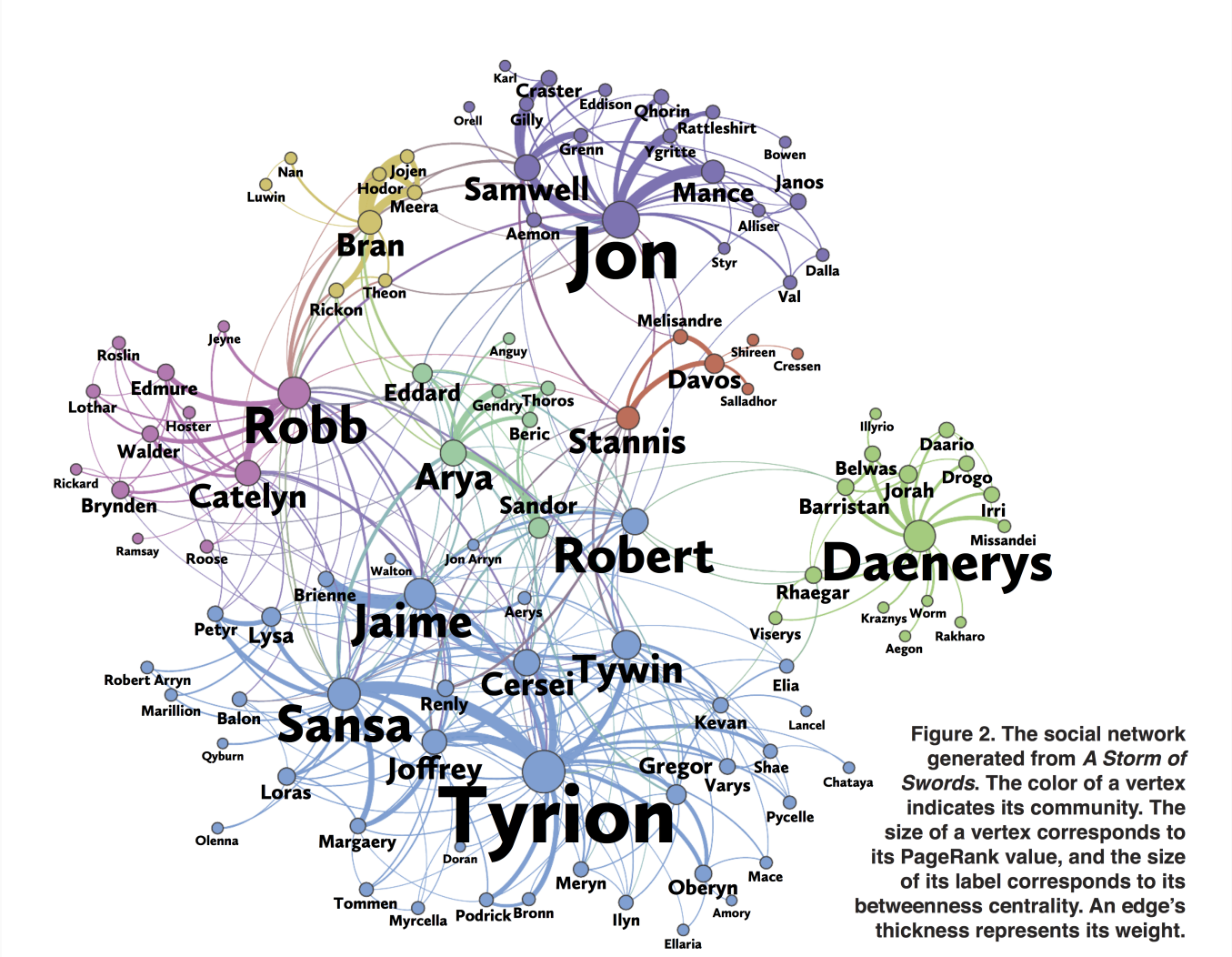


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Andrew Beveridge and Jie Shan's Network of Thrones

Descriptives: What are Descriptive Statistics for Networks?

Network Summary Statistics

Relational data can be complicated!

$$Y = \begin{bmatrix} NA & 1 & 0 & 0 & \dots \\ 1 & NA & 0 & 1 & \dots \\ 0 & 0 & NA & 1 & \dots \\ 0 & 1 & 0 & NA & \dots \\ \vdots & & & & \end{bmatrix}$$

Network Summary Statistics

Relational data can be complicated!

$$Y = \begin{bmatrix} & & & & \\ NA & 1 & 0 & 0 & \dots \\ 1 & NA & 0 & 1 & \dots \\ 0 & 0 & NA & 1 & \dots \\ 0 & 1 & 0 & NA & \dots \\ \vdots & & & & \end{bmatrix}$$

- **Statistic:** A *statistic* $t(Y)$ is any function of the data
- **Descriptive data analysis:** A representation of the main features of a dataset via a set of statistics $t_1(Y), \dots, t_k(Y)$

Network Summary Statistics

Relational data can be complicated!

$$Y = \begin{bmatrix} NA & 1 & 0 & 0 & \dots \\ 1 & NA & 0 & 1 & \dots \\ 0 & 0 & NA & 1 & \dots \\ 0 & 1 & 0 & NA & \dots \\ \vdots & & & & \end{bmatrix}$$

- **Statistic:** A *statistic* $t(Y)$ is any function of the data
- **Descriptive data analysis:** A representation of the main features of a dataset via a set of statistics $t_1(Y), \dots, t_k(Y)$

Network Summary Statistics

Many important statistics can be computed from the sociomatrix using basic matrix calculations!

$$Y = \begin{bmatrix} NA & 1 & 0 & 0 & \dots \\ 1 & NA & 0 & 1 & \dots \\ 0 & 0 & NA & 1 & \dots \\ 0 & 1 & 0 & NA & \dots \\ \vdots & & & & \end{bmatrix}$$

Network Summary Statistics

The most basic statistic of a relational dataset is the *mean* or *average*

- **Mean:** The sum of the relational measurements divided by the number of relational measurements.

For a *fully* observed directed relation with n nodes:

- The sum of the relational measurement is $\sum_{i \neq j} y_{ij}$
- The number of relational measurments is $n \times (n-1)$
- The mean is

$$\bar{y} = \frac{\sum_{i \neq j} y_{ij}}{n(n-1)}$$

Network Summary Statistics

- The most basic statistic of a relational dataset is the *mean* or *average*
- **Mean:** The sum of the relational measurements divided by the number of relational measurements.

For a *fully* observed undirected relation with n nodes: * The sum of the relational measurement is $\sum_{i<j} y_{ij}$ * The number of relational measurments is $n \times (n-1)/2$ * The mean is

$$\bar{y} = \frac{\sum_{i<j} y_{ij}}{n(n-1)/2}$$

Peter Hoff. Statistical Networks. University of Washington

Network Summary Statistics

Means can be computed from adjacency matrices:

$$Y = \begin{bmatrix} NA & 0 & 2 & 6 \\ 1 & NA & 0 & 0 \\ 0 & 0 & NA & 0 \\ 3 & 0 & 2 & NA \end{bmatrix}$$

Let's do this on the board

$$\sum_{i \neq j} y_{ij} = 14$$

$$n(n-1) = 12$$

So the mean is $14/12 = 1.67$

Network Summary Statistics: Now in R

```
Y <- matrix(c(NA, 0, 2, 6, 1, NA, 0, 0, 0,
    0, NA, 0, 3, 0, 2, NA), nc = 4, byrow = TRUE)
Y
```

```
     [,1] [,2] [,3] [,4]
[1,]   NA    0    2    6
[2,]    1   NA    0    0
[3,]    0    0   NA    0
[4,]    3    0    2   NA
```

Network Summary Statistics: Now in R

```
sum(Y)
```

```
[1] NA
```

```
sum(Y, na.rm = TRUE)
```

```
[1] 14
```

```
length(Y)
```

```
[1] 16
```

```
sum(Y, na.rm = TRUE)/length(Y)  ## This is wrong!
```

```
[1] 0.875
```

Network Summary Statistics: Now in R

```
Y <- matrix(c(NA, 0, 2, 6, 1, NA, 0, 0, 0,
    0, NA, 0, 3, 0, 2, NA), nc = 4, byrow = TRUE)
Y
```

```
     [,1] [,2] [,3] [,4]
[1,]   NA    0    2    6
[2,]    1   NA    0    0
[3,]    0    0   NA    0
[4,]    3    0    2   NA
```

Network Summary Statistics: Now in R

```
sum(Y, na.rm = TRUE)
```

```
[1] 14
```

```
sum(!is.na(Y))
```

```
[1] 12
```

```
length(Y[!is.na(Y)])
```

```
[1] 12
```

```
sum(Y, na.rm = TRUE)/sum(!is.na(Y))
```

```
[1] 1.166667
```

Network Summary Statistics: Now in R

```
Y <- matrix(c(NA, 0, 2, 6, 1, NA, 0, 0, 0,
    0, NA, 0, 3, 0, 2, NA), nc = 4, byrow = TRUE)
Y
```

```
     [,1] [,2] [,3] [,4]
[1,]   NA    0    2    6
[2,]    1   NA    0    0
[3,]    0    0   NA    0
[4,]    3    0    2   NA
```

Network Summary Statistics: Now in R

```
mean(Y)
```

```
[1] NA
```

```
mean(Y, na.rm = TRUE)
```

```
[1] 1.166667
```

Network Summary Statistics

Means from adjacency matrices: The undirected case

$$Y = \begin{bmatrix} NA & 0 & 2 & 4 \\ 0 & NA & 0 & 0 \\ 2 & 0 & NA & 3 \\ 4 & 0 & 3 & NA \end{bmatrix}$$

Let's do this on the board

$$\sum_{i<j} y_{ij} = 9$$

$$n(n-1)/2 = 6$$

So the mean is $9/6 = 1.5$

Network Summary Statistics: Now in R

```
Y <- matrix(c(NA, 0, 2, 4, 0, NA, 0, 0, 2,
    0, NA, 3, 4, 0, 3, NA), nc = 4, byrow = TRUE)
Y
```

```
     [,1] [,2] [,3] [,4]
[1,]   NA    0    2    4
[2,]    0   NA    0    0
[3,]    2    0   NA    3
[4,]    4    0    3   NA
```

```
mean(Y, na.rm = TRUE)
```

```
[1] 1.5
```

Network Summary Statistics: Means overview

For an undirected relation:

- mean of the relation = mean of the "upper triangle" of the sociomatrix = mean of the "lower triangle" of the sociomatrix = mean of the sociomatrix
- So for either directed or undirected relations,

$\bar{y}$ = average of the non-missing values of the sociomatrix

Network Summary Statistics: Means via edgelists

Directed

```
Y <- matrix(c(NA, 0, 2, 6, 1, NA, 0, 0, 0,
    0, NA, 0, 3, 0, 2, NA), nc = 4, byrow = TRUE)
Ed <- sna::as.edgelist.sna(Y)
Ed <- Ed[!is.na(Ed[, 3]), ]
Ed
```

```
     snd rec val
[1,]   2   1   1
[2,]   4   1   3
[3,]   1   3   2
[4,]   4   3   2
[5,]   1   4   6
```

Network Summary Statistics: Means via edgelists

How can we compute the mean?

```
sum(Ed[, 3])/(4 * 3)
```

```
[1] 1.166667
```

Network Summary Statistics: Means via edgelists

Undirected

```
Y <- matrix(c(NA, 0, 2, 4, 0, NA, 0, 0, 2,
    0, NA, 3, 4, 0, 3, NA), nc = 4, byrow = TRUE)
Y[lower.tri(Y)] <- NA
Eu <- sna::as.edgelist.sna(Y)
Eu <- Eu[!is.na(Eu[, 3]), ]
Eu
```

```
     snd rec val
[1,]   1   3   2
[2,]   1   4   4
[3,]   3   4   3
```

Network Summary Statistics: Means via edgelists

How can we compute the mean?

```
sum(Eu[, 3])/(4 * 3/2)
```

```
[1] 1.5
```

- When using an edgelist, you need to use the formula and be aware if the relation is directed or undirected
- Additionally, it is more difficult to account for missing data with edgelists

Network Summary Statistics: Density

Density: The proportion of edges present in a graph $>$
$= \frac{\text{the number of edges}}{\text{the maximum possible number of edges}}$

The number of observed is $|E|$ The number of possible edges is

- $n(n-1)$ in a directed graph
- $n(n-1)/2$ in an undirected graph
  - **Derivation** A n by n adjacency matrix (minus its diagonals) has $2 * \binom{n}{2} = \frac{n!}{2!(n-2)!} = n(n-1)$ cells

$$\delta_d = \frac{|E|}{n(n-1)} \text{ , } \delta_u = \frac{|E|}{n(n-1)/2}$$

Peter Hoff. Statistical Networks. University of Washington

Network Summary Statistics: Density

Let y_{ij} be the binary indicator of an eged from i to j

Then,

- $|E| = \sum_{i<j} y_{ij}$ or an undirected graph
- $|E| = \sum_{i \neq j} y_{ij}$ for an undirected graph

Thus, the density of a graph (undirected or directed) is the mean of the corresponding adjecancy matrix

Peter Hoff. Statistical Networks. University of Washington

Network Summary Statistics: Density in R

Directed

```
sum(Y, na.rm = TRUE)
```

```
[1] 6
```

```
nrow(Y)
```

```
[1] 5
```

Network Summary Statistics: Density in R

Directed

```
sum(Y, na.rm = TRUE)/(nrow(Y) * (nrow(Y) -
    1))
```

```
[1] 0.3
```

```
mean(Y, na.rm = TRUE)
```

```
[1] 0.3
```

Network Summary Statistics: Examples!

Densities can be view as,

- Probabilities of the existence of a tie between randomly sampled nodes
- Estimates of these probabilities

Let,

- i and j be two randomly sampled individuals
- Let θ be the probability that $y_{ij} = 1$

$$\Pr(Y_{ij} = 1) = \theta$$

Then

$\bar{y} = \theta$ if your nodeset is the entire population of nodes $\bar{y} = \hat{\theta}$ if your nodeset is arandom sample of nodes

R for Descriptive Analysis: An Introduction

References and Places for More Information i

Network Data Collection

Descriptives: What are Descriptive Statistics for Networks?

R for Descriptive Analysis: An Introduction