Social Network Analysis: Node and Graph Level Statistics Part 1

EPIC - SNA, Columbia University

Zack W Almquist

June 13th, 2018

University of Minnesota

Node-Level Indices

Node-Level Indices

Introduction to classic Social Network Metrics (Positional or Node-level indices)

Node-level Indices

- Node-level index: a real- valued function of a graph and a vertex
  - Purely structural NLIs depend only on unlabeled graph properties
    - I.e., $f(v, G) \rightarrow \Re$
    - Invariant to node relabeling
  - Covariate-based NLIs use both structural and covariate properties
    - I.e., $f(v, G, X) \rightarrow \Re$
    - Not labeling invariant
- Primary uses:
  - Quantify properties of individual positions
  - Describe local neighborhood
- Several common families:
  - Centrality
  - Ego-net structure
  - Alter covariate indices
- Centrality is the most prominent, and our focus today/lecture

Centrality

- Returning to the core question: how do individual positions vary?
- One manner in which positions vary is the extent to which they are “central" in the network
  - Important concern of social scientists (and junior high school students)
- Many distinct concepts
  - No one way to be central in a network - many different kinds of centrality!
  - Different types of centrality aid/hinder different kinds of actions
  - Being highly central in one respect doesn’t always mean being central in other respects (although the measures generally correlate)

Types of Centrality Measures

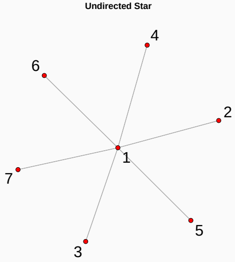
- One attempted classification by Koschutzki et al. (2005):
  - Reach: Centrality based on ability of ego to reach other vertices
    - Degree, closeness
  - Flow Mediation: Centrality based on quantity/weight of walks passing through ego
    - Stress, betweenness
  - Vitality: Centrality based on effect of removing ego from the network
    - Flow betweenness (oddly), cutpoint status
  - Feedback: Centrality of ego defined as a recursive function of alter centralities
    - Eigenvector centrality, Bonacich Power

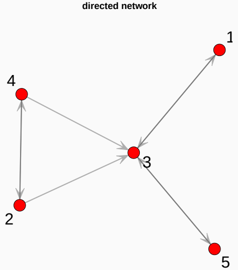
Degree

- Degree: number of direct ties
  - Overall activity or extent of involvement in relation
  - High degree positions are influential, but also may be subject to a great deal of influence from others
- Formulas:
  - Degree (undirected):

$$d(i, Y) = \sum_{j=1}^{N} Y_{ij}$$

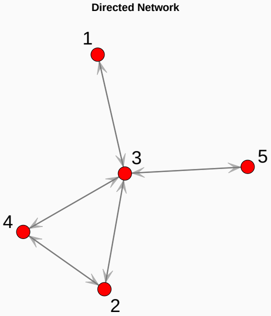
  - Indegree: $d_i(i, Y) = \sum_{j=1}^{N} Y_{ji}$
  - Outdegree: $d_o(i, Y) = \sum_{j=1}^{N} Y_{ij}$





Review: Shortest Paths

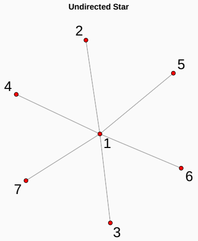
- A shortest path from i to j is called an i, j geodesic
  - Can have more than one (but all same length, obviously)
  - The length of an i, j geodesic is called the geodesic distance from i to j



	1	2	3	4	5
1	0.00	2.00	1.00	2.00	2.00
2	2.00	0.00	1.00	1.00	2.00
3	1.00	1.00	0.00	1.00	1.00
4	2.00	1.00	1.00	0.00	2.00
5	2.00	2.00	1.00	2.00	0.00

Betweenness

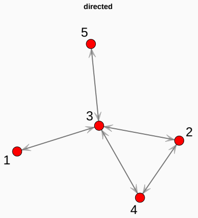
- Betweenness: tendency of ego to reside on shortest paths between third parties
  - Quantifies extent to which position serves as a bridge
  - High betweenness positions are associated with “broker" or “gatekeeper" roles; may be able to “firewall" information flow



- Formula

$$b(i, Y) = \sum_{j \neq i} \sum_{k \neq l} \frac{g^{'}(j, k, l)}{g(j, k)}$$

Where $g(j, k)$ is the number of j, k geodesics, $g^{'}(j, k, i)$ is the number of j, k geodesics including i

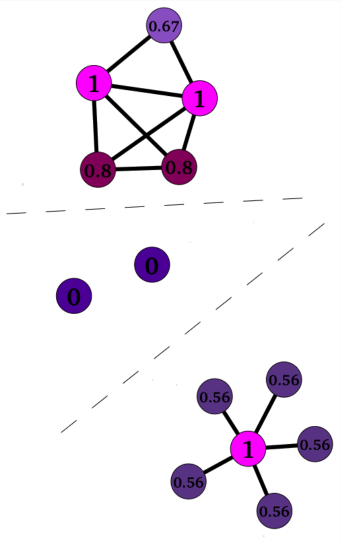


Closeness

- Closeness: ratio of minimum distance to other nodes to observed distance to other nodes
  - Extent to which position has short paths to other positions
  - High closeness positions can quickly distribute information, but may have limited direct influence
  - Limitation: not useful on disconnected graphs (may need to symmetrize directed graphs, too)
- Formula

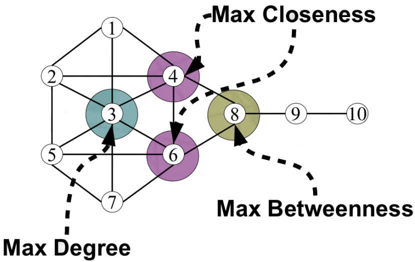
$$c(i, Y) = \frac{N-1}{\sum_{j=1}^{N} D(i,j)}$$

Where $D(I,j)$ is the distance from i to j



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

Classic Centrality Measures Compared



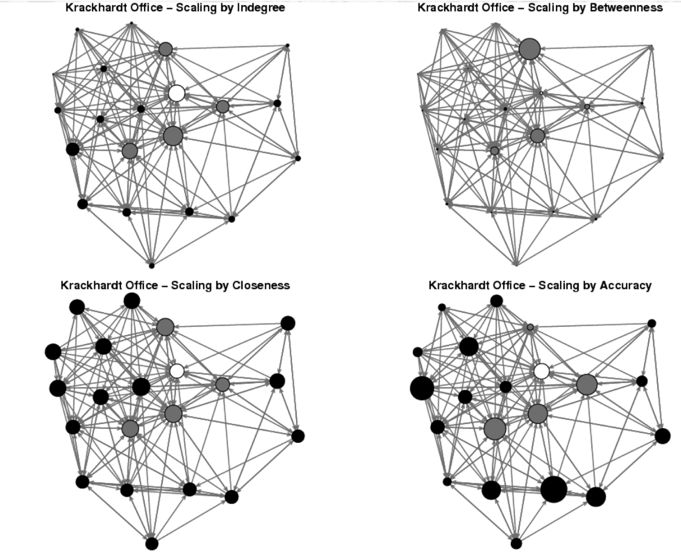
Top 3 by Degree
- 1: Node 3
- 2: Nodes 4 and 6
- 3: Nodes 2 and 5

Top 3 by Closeness
- 1: Nodes 4 and 6
- 2: Nodes 3 and 8
- 3: Nodes 2 and 5

Top 3 by Betweenness
- 1: Node 8
- 2: Nodes 4 and 6
- 3: Node 9

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

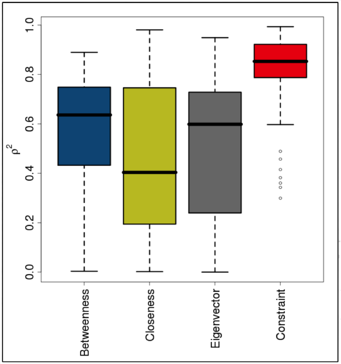
Classic Centrality Measures Compared



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

Relatedness of Centrality Indices

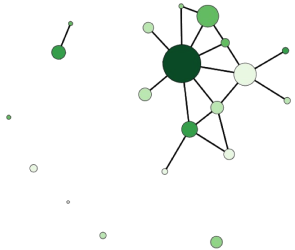
- Centrality indices are strongly correlated in practice
- Simple example: total degree versus “complex" NLIs
  - Squared correlations for sample UCINET data sets
  - Some diversity, but usually accounts for majority of variance
  - Theoretical insight: if you can capture degree, you can capture many other aspects of social position



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

Relating NLIs to Vertex Covariates

- Common question: are NLIs related to non-structural covariates?
  - Centrality to power or influence
  - Constraint to advancement
  - Diversity to attainment



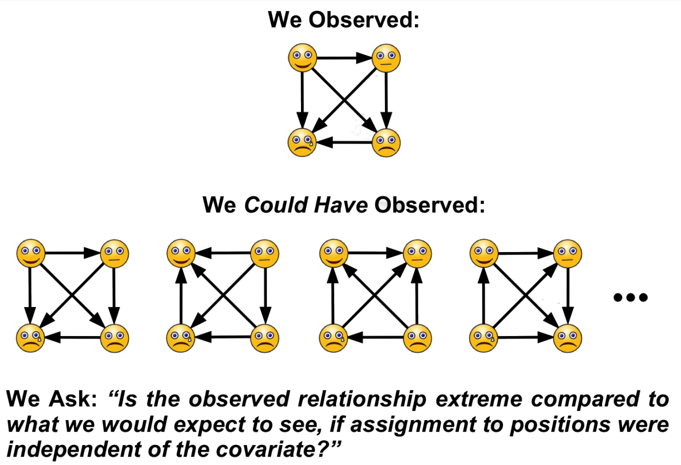
(Texas SAR EMON Decision Rank Score (scale) vs. Degree (color), Mutually Reported "Continuous" Communication" - ρ=0.86)

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

“Linear" Permutation Tests

- Simple, nonparametric test of association between vectors
  - Sometimes called “linear" or “vector" permutation test (or monte carlo test)
  - Tests marginal association against exchangeability null (independence conditional on marginal distributions)
- Null interpretation: “musical chairs" model
  - If we randomly switched the positions of people in the network (leaving structure as-is), what is the chance of observing a similar degree of association?
- Monte Carlo procedure:
  - Let $x_{obs} = (f(v_1, G), \ldots, f(v_N, G))$ be the observed NLI vector, w/covariate vector y
  - Let $t_{obs} = s(x_{obs}, y)$
  - For i in $1, \ldots, n$
    - Let $x^{(0)}$ be a random permutation of x_{obs}
    - Let $t^{(i)} = s(x^{(i),y})$
- Estimated p-values:
  - One-seided
    - $\Pr(t^{(i)} \leq t_{obs}) \approx \sum_i I(t^{(i)} \leq t_{obs})/n$
    - $\Pr(t^{(i)} \geq t_{obs}) \approx \sum_i I(t^{(i)} \geq t_{obs})/n$
  - Two-sided
    - $\Pr(|t^{(i)}| \geq |t_{obs}|) \approx \sum_i I(|t^{(i)}| \geq |t_{obs}|)/n$

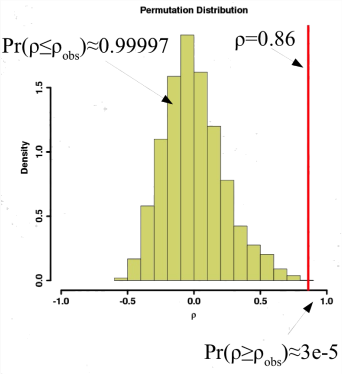
Understanding the Null Model



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

Texas SAR EMON Example

- Question: do organizations in constant communication w/many alters end up more/less prominent in the decision-making process?
    - Measure (s): correlation of decision rank score (y) with degree in confirmed "continuous communication" network (x_{obs})
    - Null: no relationship between degree and decision making
    - Alternative: decision making has linear marginal relationship w/degree
- Results
    - $t_{obs} = 0.86$; $\Pr(|t^{(i)} \geq |t_{obs}|) \approx 3e-5$



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

NLIs as Covariates

- NLIs can also be used as covariates (e.g., in regression analyses)
  - Modeling assumption: position properties predict properties of those who hold them
  - Conditioning on NLI values, so dependence doesn't matter (if no error in G)
  - NLIs as dependent variables are much more problematic; we'll revisit this problem when we discuss ERGs
- Things to keep in mind....
  - Make sure that your theory really posits a direct relationship w/the NLI
  - NLI distributions could be quite skewed or irregular; be sure this makes sense (e.g., via analysis of residuals)
  - Multiple NLIs may be strongly correlated; may not be able to distinguish among related measures in practice

Graph Level Indicies

Graph-Level Properties

- Earlier, we discussed the notion of node-level indices (mainly centrality)
  - Dealt with position of the individual within the network
- Today, 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

Review 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: $\delta = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N} Y_{ij}}{N(N-1)}$

R Code

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

```
[1] 0.4222222
```

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

```
[1] 0.5222222
```

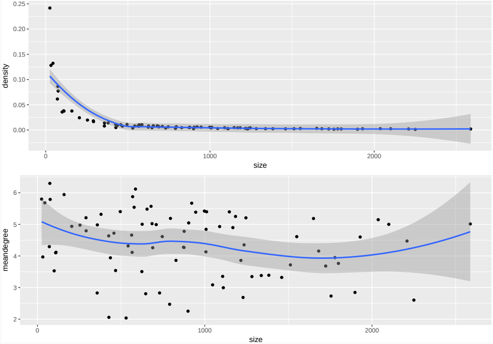
Size, Density, and Mean Degree

- 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

Illustration: Mean Degree Constancy and Density Decline

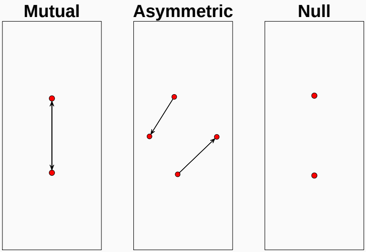
```
library(ggplot2)
library(gridExtra)
library(networkdata)
data(addhealth)
data <- data.frame(size = sapply(addhealth, network.size), density = sapply(addhealth,
    gden))
data$meandegree <- data$density * (data$size - 1)

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



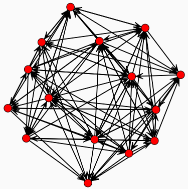
Reciprocity

- 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

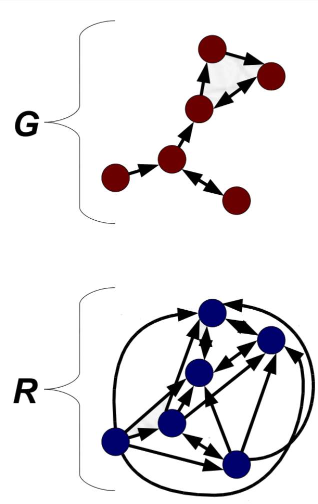
$$\frac{2M}{2M+A}$$



	Mut	Asym	Null
1	19.00	64.00	22.00

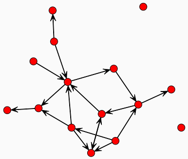
Reachability

- 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?



Reciprocity

0.15

Krackhardt

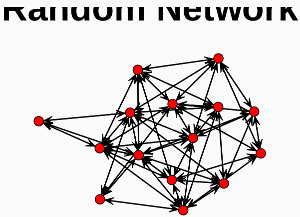
0.83

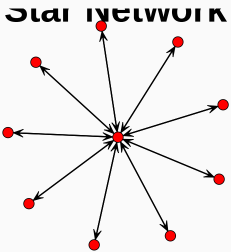
Centralization

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

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

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



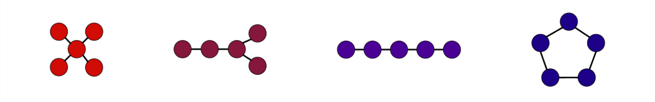


Centralization Versus Hierarchy

- 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 ...

References and Places for More Information

References and Places for More Information i

Node-Level Indices

Graph Level Indicies

References and Places for More Information