Centrality

567 Statistical analysis of social networks

Peter Hoff

Statistics, University of Washington

Centrality

A common goal in SNA is to identify the "central" nodes of a network.

What does "central" mean?

- active?
- important?
- non-redundant?

Koschutzki et al. (2005) attempted a classification of centrality measures

- Reach: ability of ego to reach other vertices
- Flow: quantity/weight of walks passing through ego
- Vitality: effect of removing ego from the network
- Feedback: a recursive function of alter centralities

Common centrality measures

We will define and compare four centrality measures:

- degree centrality (based on degree)
- closeness centrality (based on average distances)
- betweeness centrality (based on geodesics)
- eigenvector centrality (recursive: similar to page rank methods)

Standardized centrality measures

Node-level indices

Let $c_1, \ldots, c_n$ be node-level centrality measures:

$$c_i = \text{ centrality of node } i \text{ by some metric}$$

It is often useful to standardize the c_i's by their maximum possible value:

$$\tilde{c}_i = c_i / c_{\max}$$

Network centralization

Network-level indices

How centralized is the network?

To what extent is there a small number of highly central nodes?

- Let $c^* = \max\{c_1, \ldots, c_n\}$
- Let $S = \sum_i [c^* - c_i]$

Then

- $S = 0$ if all nodes are equally central;
- S is large if one node is most central.

Network centralization

Network level centralization index

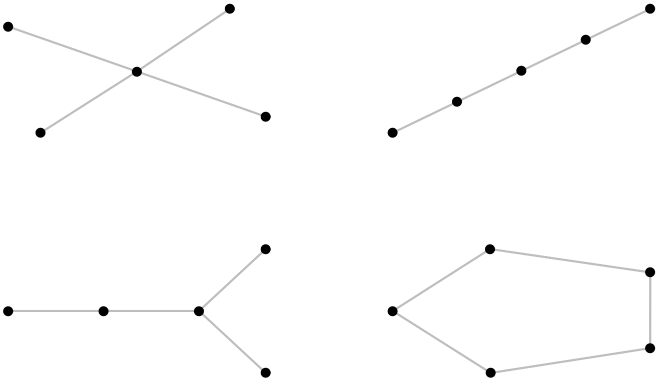
$$C = \frac{\sum_i [c^* - c_i]}{\max \sum_i [c^* - c_i]}$$

The "max" in the denominator is over all possible networks.

- $0 \leq C \leq 1$;
- $C = 0$ when all nodes have the same centrality;
- $C = 1$ if one actor has maximal centrality and all others have minimal.

Networks for comparison

We will compare the following graphs under different centrality measures:



These are the star graph, line graph, y-graph, the circle graph.

Which do you feel is most "centralized"? Which the least?

Degree centrality

Idea: A central actor is one with many connections.

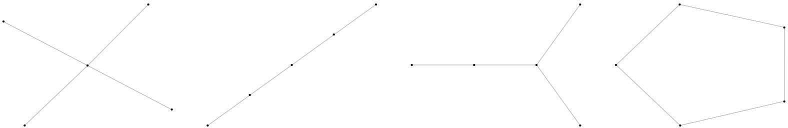
This motivates the measure of **degree centrality**

- undirected degree centrality: $c_i^d = \sum_{j:j\neq i} y_{i,j}$
- outdegree centrality: $c_i^o = \sum_{j:j\neq i} y_{i,j}$
- indegree centrality: $c_i^i = \sum_{j:j\neq i} y_{j,i}$

The standardized degree centrality is

$$\tilde{c}_i^d = c_i^d / c_{\max}^d = c_i^d / (n-1)$$

Degree centrality



```
apply(Ys,1,sum,na.rm=TRUE)

## 1 2 3 4 5
## 1 1 4 1 1

apply(Yl,1,sum,na.rm=TRUE)

## 1 2 3 4 5
## 1 2 2 2 1

apply(Yy,1,sum,na.rm=TRUE)

## 1 2 3 4 5
## 1 2 3 1 1

apply(Yc,1,sum,na.rm=TRUE)

## 1 2 3 4 5
## 2 2 2 2 2
```

Degree centralization

$$c_i^d : \text{ actor centrality}$$

$$c^{d*} : \text{ maximum actor centrality observed in the network}$$

$$\sum_i [c^{d*} - c_i^d] : \text{sum of differences between most central actor and others}$$

Centralization

$$C^d = \frac{\sum_i [c^{d*} - c_i^d]}{\max_Y \sum_i [c^{d*} - c_i^d]}$$

What is the maximum numerator that could be attained by an n-node graph?

Degree centralization

The maximum occurs when

- one node has the largest possible degree ($c^{d*} = n - 1$),
- the others have the smallest possible degree $c_i^d = 1$.

This is the star graph.

$$\begin{aligned}\max_Y \sum_i [c^{d*} - c_i^d] &= \sum_i [(n-1) - c_i^d] \\ &= 0 + (n-1-1) + \cdots + (n-1-1) \\ &= (n-1)(n-2)\end{aligned}$$

$$C^d(\mathbf{Y}) = \frac{\sum_i [c^{d*} - c_i^d]}{(n-1)(n-2)}$$

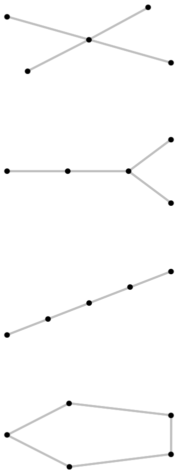
Degree centralization

Exercise: Compute the degree centralization for the four $n = 5$ graphs:

- the star graph;
- the line graph;
- the y-graph;
- the circle graph.

Degree centralization

```
Cd<-function(Y)
{
  n<-nrow(Y)
  d<-apply(Y,1,sum,na.rm=TRUE)
  sum(max(d)-d)/( (n-1)*(n-2) )
}

Cd(Ys)

## [1] 1

Cd(Yy)

## [1] 0.5833333

Cd(Yl)

## [1] 0.1666667

Cd(Yc)

## [1] 0
```



Closeness centrality

Idea: A central node is one that is close, on average, to other nodes.

This motivates the idea of **closeness centrality**

- (geodesic) distance: $d_{i,j}$ is the minimal path length from i to j;
- closeness centrality: $c_i^c = 1/\sum_{j:j\neq i} d_{i,j} = 1/[(n-1)\bar{d}_i]$;
- limitation: not useful for disconnected graphs.

Closeness centrality

$$c_i^c = 1/[(n-1)\bar{d}_i]$$

Recall,

$$d_a < d_b \Rightarrow \frac{1}{d_a} > \frac{1}{d_b}$$

and so a node i would be "maximally close" if $d_{i,j} = 1$ for all $j \neq i$.

$$c_{\max}^d = \frac{1}{n-1}$$

The standardized closeness centrality is therefore

$$\begin{aligned}\tilde{c}_i^c &= c_i^c / c_{\max}^d \\ &= (n-1)c_i^c = 1/\bar{d}_i.\end{aligned}$$

$$
\begin{aligned}
c_i^c &: \text{ actor centrality} \\
c^{c*} &: \text{ maximum actor centrality observed in the network} \\
\sum_i [c^{c*} - c_i^c] &: \text{sum of differences between most central actor and others}
\end{aligned}
$$

Centralization

$$C^c = \frac{\sum_i [c^{c*} - c_i^d]}{\max_Y \sum_i [c^{c*} - c_i^c]}$$

What is the maximum numerator that could be attained by an n-node graph?

Closeness centralization

The maximum occurs when

- one node has the largest possible closeness ($\bar{d}^* = 1, c^{c*} = 1/(n-1)$),
- the others have the smallest possible closeness, given that $c^{c*} = 1/(n-1)$.

(Freeman, 1979)

For what graph are these conditions satisfied?

- For $c^{*c} = 1/(n-1)$, one node must be connected to all others.
- To then maximize centralization, the centrality of the other nodes must be minimized.

This occurs when none of the non-central nodes are tied to each other, i.e. the star graph.

Closeness centralization

For a non-central node in the star graph,

$$\begin{aligned}\bar{d}_i &= \frac{1+2+\cdots+2}{n-1}\\ &= \frac{2(n-2)+1}{n-1}\\ &= \frac{2n-3}{n-1}\\ c_i^c &= 1/[(n-1)\bar{d}_i] = \frac{1}{2n-3}.\end{aligned}$$

Therefore, for the star graph

$$\begin{aligned}\sum_i [c^{c*} - c_i^c] &= 0 + (\frac{1}{n-1} - \frac{1}{2n-3}) + \cdots (\frac{1}{n-1} - \frac{1}{2n-3})\\ &= (n-1) \times \left(\frac{1}{n-1} - \frac{1}{2n-3}\right)\\ &= (n-1) \times \frac{n-2}{(2n-3)(n-1)}\\ &= \frac{n-2}{2n-3}\end{aligned}$$

Closeness centralization

To review, the maximum of $\sum_i [c^{c*} - c_i^c]$ occurs for the star graph, for which

$$\sum_i [c^{c*} - c_i^c] = \frac{n-2}{2n-3}$$

Therefore, the centralization of any graph $\mathbf{Y}$ is

$$\begin{aligned} C^c(\mathbf{Y}) &= \frac{\sum_i [c^{c*} - c_i^c]}{\max_Y \sum_i [c^{c*} - c_i^c]} \\ &= \frac{\sum_i [c^{c*} - c_i^c]}{(n-2)/(2n-3)} \end{aligned}$$

Alternatively, as $\tilde{c}_i^c = (n-1)c_i^c$,

$$\begin{aligned} C^c(\mathbf{Y}) &= \frac{\sum_i [c^{c*} - c_i^c]}{(n-2)/(2n-3)} \\ &= \frac{\sum_i [\tilde{c}^{c*} - \tilde{c}_i^c]}{[(n-1)(n-2)]/(2n-3)} \end{aligned}$$

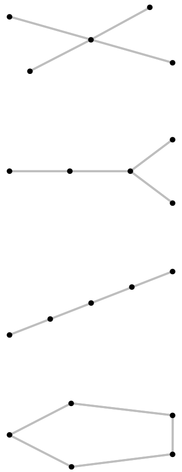
Closeness centralization

Exercise: Compute the closeness centralization for the four $n = 5$ graphs:

- the star graph;
- the line graph;
- the y-graph;
- the circle graph.

Closeness centralization

```
Cc<-function(Y)
{
n<-nrow(Y)
D<-netdist(Y)
c<-1/apply(D,1,sum,na.rm=TRUE)

sum(max(c)-c)/( (n-2)/(2*n-3) )
}

Cc(Ys)

## [1] 1

Cc(Yy)

## [1] 0.6351852

Cc(Yl)

## [1] 0.4222222

Cc(Yc)

## [1] 0
```



Betweeness centrality

Idea: A central actor is one that acts as a bridge, broker or gatekeeper.

- Interaction between unlinked nodes goes through the shortest path (geodesic);
- A "central" node is one that lies on many geodesics.

This motivates the idea of **betweenness centrality**

- $g_{j,k}$ = number of geodesics between nodes j and k;
- $g_{j,k}(i)$ = number of geodesics between nodes j and k going through i;
- $c_i^b = \sum_{j<k} g_{j,k}(i)/g_{j,k}$

Betweeness centrality

Interpretation: $g_{j,k}(i)/g_{j,k}$ is the probability that a "message" from j to k goes through i.

- j and k have $g_{j,k}$ routes of communication;
- i is on $g_{j,k}(i)$ of these routes;
- a randomly selected route contains i with probability $g_{j,k}(i)/g_{j,k}$.

Note: WF p.191
"(betweenness centrality) can be computed even if the graph is not connected" (WF)

- Careful: If j and k are not reachable, what is $g_{j,k}(i)/g_{j,k}$?
- By convention this is set to zero for unreachable pairs.

Betweeness centrality

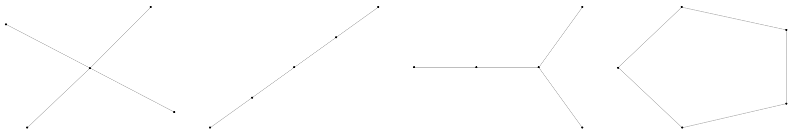
$$c_i^b = \sum_{j<k} g_{j,k}(i)/g_{j,k}$$

- $0 \leq c_i^b$, with equality when i lies on no geodesics (draw a picture)
- $c_i^b \leq \binom{n-1}{2} = \frac{(n-1)(n-2)}{2}$, with equality when i lies on all geodesics.

The standardized betweenness centrality is

$$\tilde{c}_i^b = 2c_i^b/[(n-1)(n-2)].$$

Betweenness centrality



Exercise: Compute the betweenness centrality for each node in each graph.

```
betweenness(Ys,gmode="graph")

## [1] 0 0 6 0 0

betweenness(Yl,gmode="graph")

## [1] 0 3 4 3 0

betweenness(Yy,gmode="graph")

## [1] 0 3 5 0 0

betweenness(Yc,gmode="graph")

## [1] 1 1 1 1 1
```

Betweenness centralization

$$c_i^b : \text{ actor centrality}$$

$$c^{b*} : \text{ maximum actor centrality observed in the network}$$

$$\sum_i [c^{b*} - c_i^b] : \text{sum of differences between most central actor and others}$$

Centralization

$$C^b = \frac{\sum_i [c^{b*} - c_i^b]}{\max_Y \sum_i [c^{b*} - c_i^b]}$$

What is the maximum numerator that could be attained by an n-node graph?

Betweenness centralization

The maximum occurs when

- one node has the largest possible betweeness ($c^{b*} = \binom{n-1}{2}$),
- the others have the smallest possible betweeness ($c_i^d = 0$).

Again, this is the star graph.

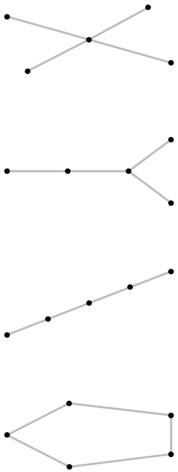
$$\begin{aligned}\max_Y \sum_i [c^{b*} - c_i^b] &= \sum_i [\binom{n-1}{2} - c_i^b] \\ &= 0 + (\binom{n-1}{2} - 0) + \cdots + (\binom{n-1}{2} - 0) \\ &= (n-1)\binom{n-1}{2}\end{aligned}$$

$\binom{n-1}{2} = (n-1)(n-2)/2$, so

$$\begin{aligned}C^b(\mathbf{Y}) &= \frac{\sum_i [c^{b*} - c_i^b]}{(n-1)\binom{n-1}{2}} \\ &= 2\frac{\sum_i [c^{b*} - c_i^b]}{(n-1)^2(n-2)}\end{aligned}$$

Betweenness centralization

```
Cb<-function(Y)
{
  require(sna)
  n<-nrow(Y)
  b<-betweenness(Y,gmode="graph")
  2*sum(max(b)-b)/(  (n-1)^2 * (n-2) )
}

Cb(Ys)

## [1] 1

Cb(Yy)

## [1] 0.7083333

Cb(Yl)

## [1] 0.4166667

Cb(Yc)

## [1] 0
```



Eigenvector centrality

Idea: A central actor is connected to other central actors.

This definition is recursive:
Eigenvector centrality: The centrality of each vertex is proportional to the sum of the centralities of its neighbors

- Formula: $c_i^e = \frac{1}{\lambda} \sum_{j:j \neq i} y_{i,j} c_j^e$
- Central vertices are those with many central neighbors
- A variant of eigenvector centrality is used by Google to rank Web pages

Google Describing PageRank: *PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."*

Eigenvector centrality

$$c_i^e = \frac{1}{\lambda} \sum_{j:j \neq i} y_{i,j} c_j^e$$

Using matrix algebra, such a vector of centralities satisfies

$$\mathbf{Y}\mathbf{c}^e = \lambda \mathbf{c}^e,$$

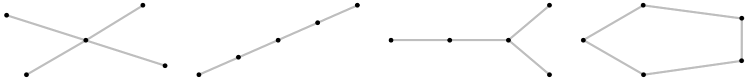
where the missing diagonal of $\mathbf{Y}$ has been replaced with zeros.

A vector $\mathbf{c}^e$ satisfying the above equation is an **eigenvector** of $\mathbf{Y}$.

There are generally multiple eigenvectors. The centrality is taken to be the one corresponding to the largest value of λ.

- this corresponds with the best rank-1 approximation to $\mathbf{Y}$;
- nodes with large c_i^e's have "strong activity" in the "primary dimension" of $\mathbf{Y}$.

Eigenvector centrality



```
evecc<-function(Y)
{
  diag(Y)<-0
  tmp<-eigen(Y)$vec[,1] ; tmp<-tmp*sign(tmp[1])
  tmp
}

evecc(Ys)

## [1] 0.3535534 0.3535534 0.7071068 0.3535534 0.3535534

evecc(Yl)

## [1] 0.2886751 0.5000000 0.5773503 0.5000000 0.2886751

evecc(Yy)

## [1] 0.2705981 0.5000000 0.6532815 0.3535534 0.3535534

evecc(Yc)

## [1] 0.4472136 0.4472136 0.4472136 0.4472136 0.4472136
```

Eigenvector centralization

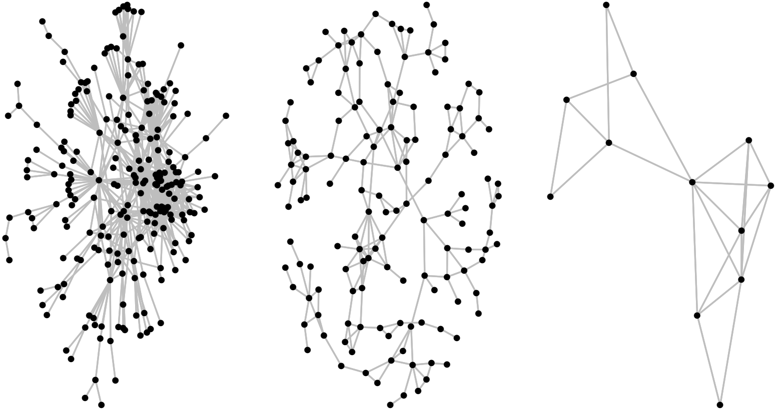
```
Ce<-function(Y)
{
  n<-nrow(Y)
  e<-evecc(Y)
  Y.sgn<-matrix(0,n,n) ;  Y.sgn[1,-1]<-1 ; Y.sgn<-Y.sgn+t(Y.sgn)
  e.sgn<-evecc(Y.sgn)
  sum(max(e)-e)/  sum(max(e.sgn)-e.sgn)
}

Ce(Ys)

## [1] 1

Ce(Yy)

## [1] 0.802864

Ce(Yl)

## [1] 0.5176381

Ce(Yc)

## [1] 9.420555e-16
```

Empirical study: Comparing centralization of different networks

Comparison of centralization metrics across four networks:

- `butland_ppi`: binding interactions among 716 yeast proteins
- `addhealth9`: friendships among 136 boys
- `tribes`: postive relations among 12 NZ tribes

Empirical study: Comparing centralization of different networks



Empirical study: Comparing centralization of different networks

	degree	closeness	betweenness	eigenvector
ppi	0.13	0.26	0.31	0.35
addhealth	0.04	0.14	0.42	0.61
tribes	0.35	0.5	0.51	0.47

Empirical study: Comparing centralization of different networks

Comments:

- The protein network looks visually centralized, but
  - most centralization is local;
  - globally, somewhat decentralized.
- The friendship network has small degree centrality (why?).
- The tribes network has one particularly central node.