Freie Universität Berlin

Class 6:
Measures of Centrality

Course: Computational Network Analysis

Prof. Dr. Claudia Müller-Birn

Institute of Computer Science, «Human-Centered Computing»

Feb 29, 2016

Today's outline

- Motivation
- Degree
- Betweenness
- Closeness
- All together
- Eigenvector

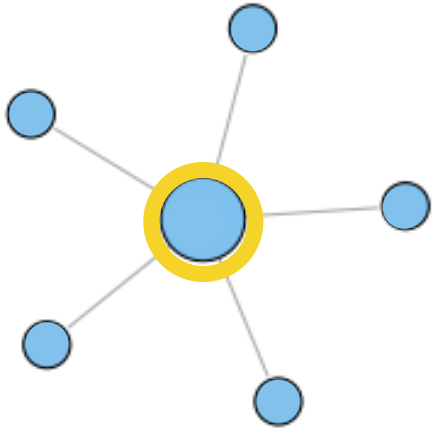
The term “Centrality”

- Introduced by Bavelas in 1948 to analyze communication in small groups by considering a relationship between structural centrality and influence in group processes

- Determining the most central node is important
  - to disseminate information in the network faster
  - to stop epidemics
  - to protect the network from breaking

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

What is the most central position in a network?



- That position has
  - the maximum possible **degree**
  - it falls on the geodesics **between** the largest possible number of other nodes
  - since it is located at the **minimum distance** from all other nodes, it is maximally close to them.

Dimensions of Analyzing Networks

Semantic Dimension of Network Analysis

Presentation of qualitative data (e.g. tabular with frequency, bar chart)	Presentation of quantitative data (e.g. histogram)	Measures of central tendency and variability (e.g. mean, range)

Syntactical Dimension of Network Analysis

Local structure	Global structure	*Partitions*
• Degree • **Degree Centrality** • **Closeness Centrality** • **Betweenness Centrality** • Local Clustering Coefficient	• Mean degree • Degree distribution • Density • **Network Centralization** • Global Clustering Coefficient • Components	• Local definition, such as clique, k-core, k-plex • Global definition with null model (modularity with modularity optimization and edge betweenness)

Scopes of measuring centrality

- The challenge is that the definition of ‘central’ varies by context/purpose
- We already discussed the degree as a **local measure** of centrality, for example the degree
- We also need to measure the degree relative to rest of network as a **normalized measure** - centrality
- We need to answer how evenly centrality is distributed among all vertices in the network as a **global measure** - centralization

Degree Centrality

If not indicated differently, slides has been adapted from Lada Adamic (2008).

Recap: Degree

- The **degree** of a vertex v in an undirected graph G = (V,E), denoted by d(v), is the number of edges in E that have v as an endvertex.

- If G is a multigraph, parallel edges are counted according to their multiplicity in E.

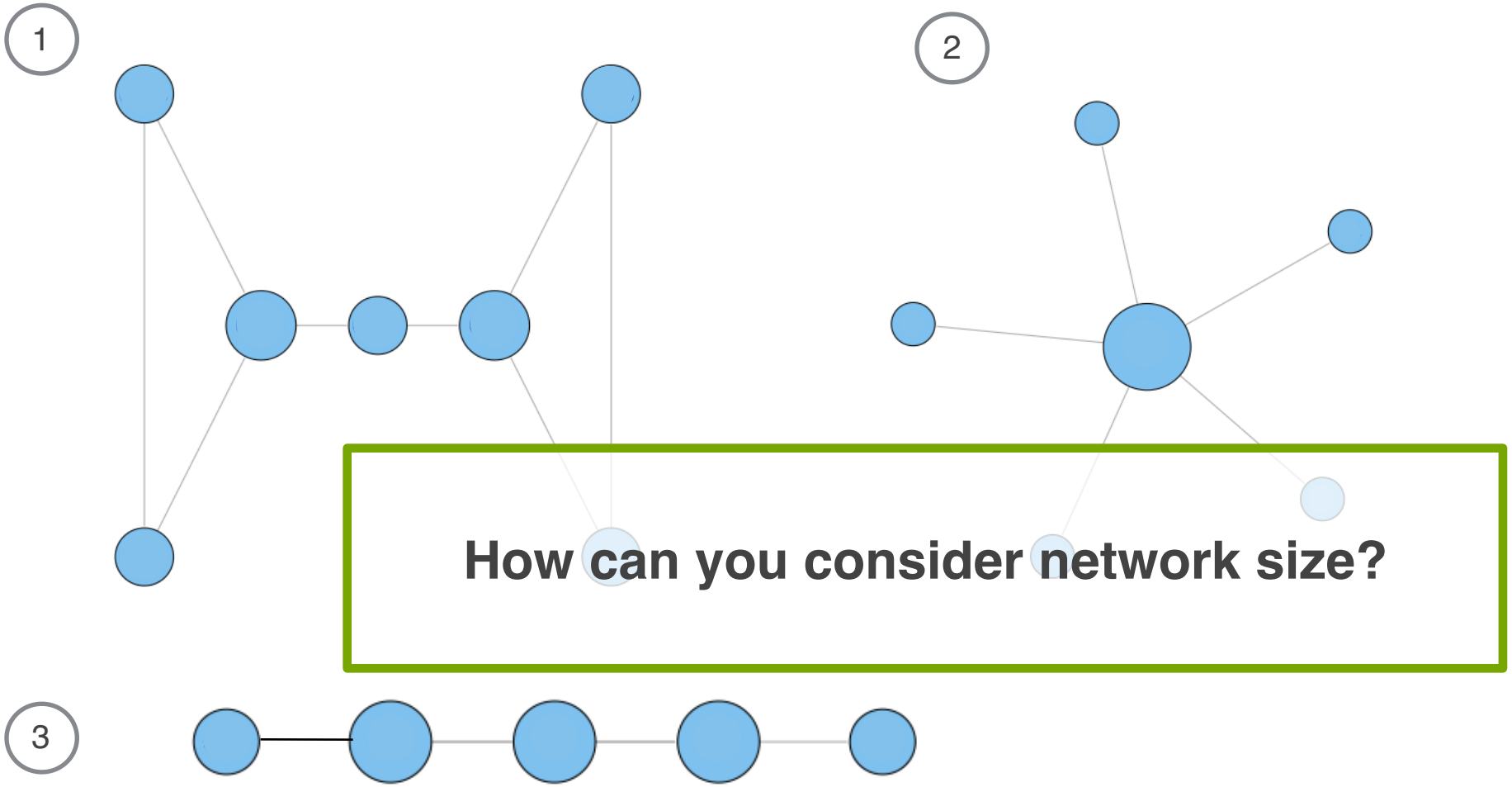
- In directed networks each vertex has two degrees
  - In-degree: number of ingoing edges connected to the vertex
  - Out-degree: number of outgoing edges connected to the vertex

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Degree Centrality

- Index of exposure to what is flowing through the network
- For example, in a gossip network, a central actor is more likely to hear a given bit of gossip
- Interpreted as opportunity to influence & be influenced directly
- Predicts variety of outcomes from virus resistance to power & leadership to job satisfaction to knowledge
- According to Freeman, we define: $C_{D_v} = d(v)$

Which vertex has the highest degree?



Normalized Degree centrality

- Evaluation of the position of a single node in a directed or undirected network
- A high value implies that the node is more independent from other nodes by acquisition or sharing of nodes

- Undirected network (normalized)

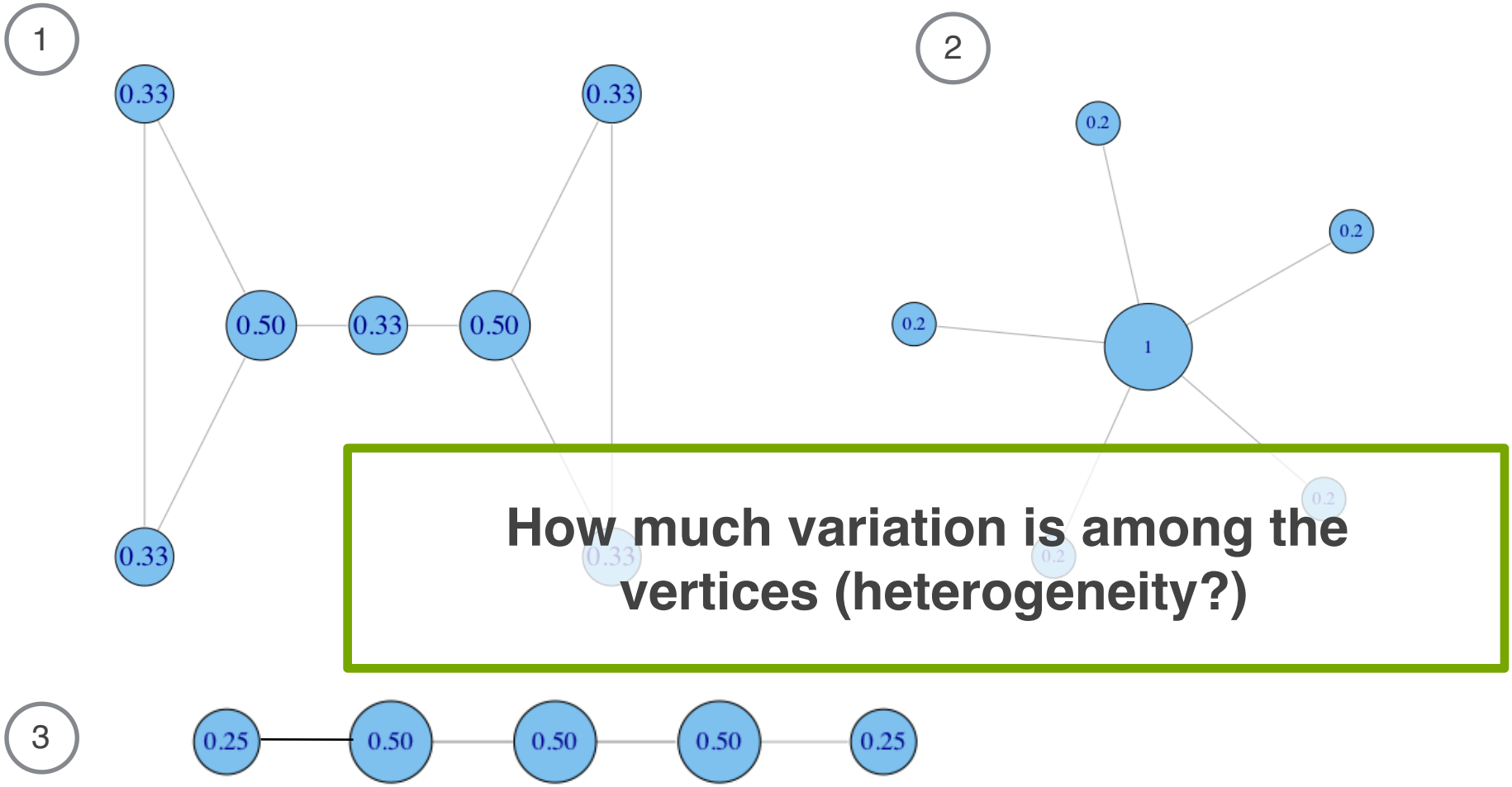
$$C'_{D_v} = \frac{d(v)}{N-1}$$

- Directed network (normalized)

$$C'_{D_v} = \frac{d_{out}(v)}{N-1}$$

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Which vertex has the highest relative degree centrality?



How much variation is among the vertices (heterogeneity?)

How can we define a global measure of centrality?

- Required features:
  - it should index the degree to which the centrality of the most central vertex exceeds the centrality of all other vertices, and
  - it should be expressed as a ratio of that excess to its maximum possible value for a graph containing the observed number of vertices.

- Given
  - N = number of vertices
  - C(v) = centrality of a vertex
  - C(v*) = largest value of centrality for any vertex in the network

- and

$$max \sum_{v=1}^{N} [C(v^*) - C(v)]$$

= the maximum possible sum of differences vertex centrality for a graph of N vertices

- Then $C = \frac{\sum_{v=1}^{N} [C(v^*) - C(v)]}{max \sum_{v=1}^{N} [C(v^*) - C(v)]}$

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

What does centralization mean?

- Centralization can be understood as an index that determine the degree to which C(v*) exceeds the centrality of all of the other vertices
- C is a ratio of an observed sum of differences to its maximum value, it will vary between 0 and 1
- C = 0 if and only if all C(v) are equal
- C = 1 if and only if one vertex v* completely dominates the network with respect to centrality

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Degree Centralization

- The question is: How equal are the nodes?

- Freeman's general formula for centralization

$$C_D = \frac{\sum_{v=1}^{N}[C_D(v^*) - C_D(v)]}{max\sum_{v=1}^{N}[C_D(v^*) - C_D(v)]}$$

$$= \frac{\sum_{v=1}^{N}[C_D(v^*) - C_D(v)]}{(N-1)(N-2)}$$

where $C_D(v^*)$ is the maximal degree in the network and N corresponds to the number of vertices in the network.

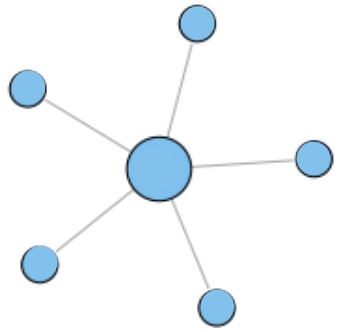
- Other metrics such as the Gini coefficient or standard deviation are possible as well

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Simplifying the denominator

- Goal: Determine the maximum sum of differences
- We have already established that the maximum value of $C_D(v^*)$ is N - 1 for a vertex that is adjacent to all of its neighbors
- If the graph is a star each of the other vertices will have $C_D(v) = 1$ and the differences will be

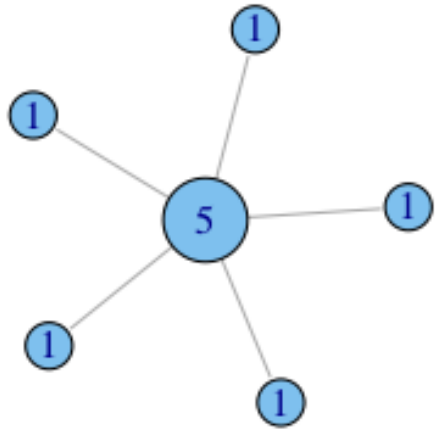
  (n-1)-1 = n-2



- for each of the N - 1 comparisons.
- Thus, the difference sum will be (n - 2)(n - 1)

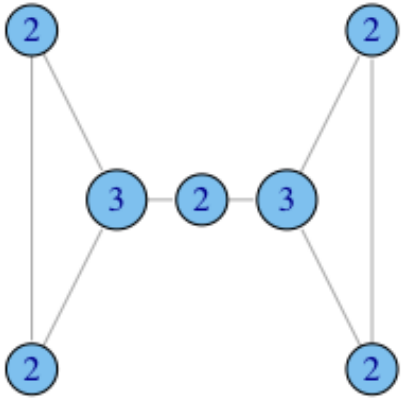
Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Result: Degree Centralization



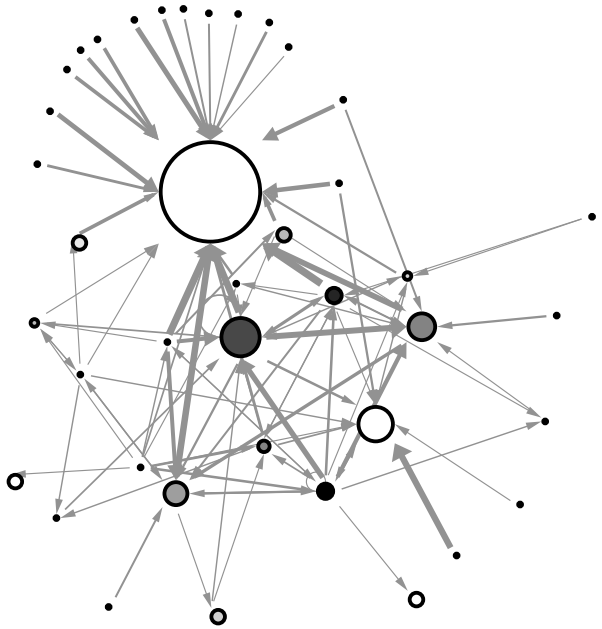
C_D = 1.0

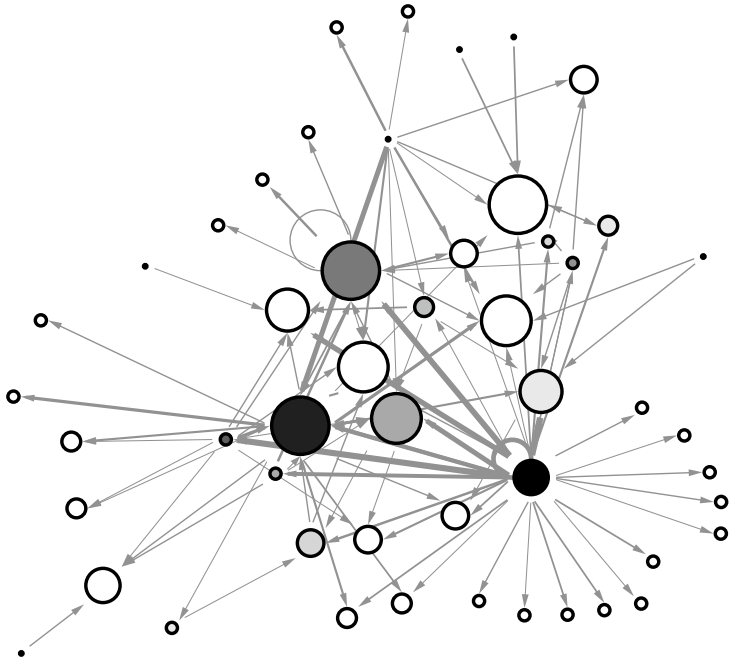
C_D = 0.167



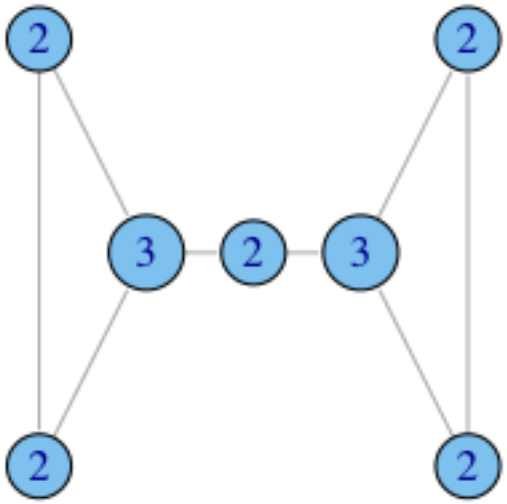
C_D = 0.167

Degree Centralization in financial trading networks





When degree isn't everything...



In what way does degree fail to capture centrality in the following networks?

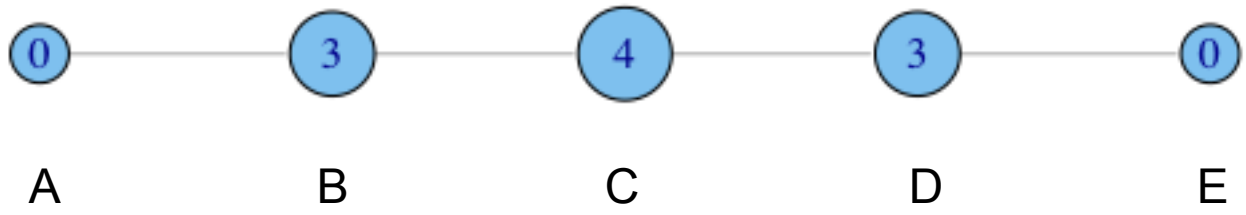
Betweenness Centrality

Motivation

- Persons are central if they are strategically located on the communication paths linking pairs of others
- Persons in such a position can influence the group by withholding or distorting information in transmission
- Additionally, persons occupying such positions for the maintenance of communications and they have a potential to be coordinators of group processes

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Determine the number of shortest paths

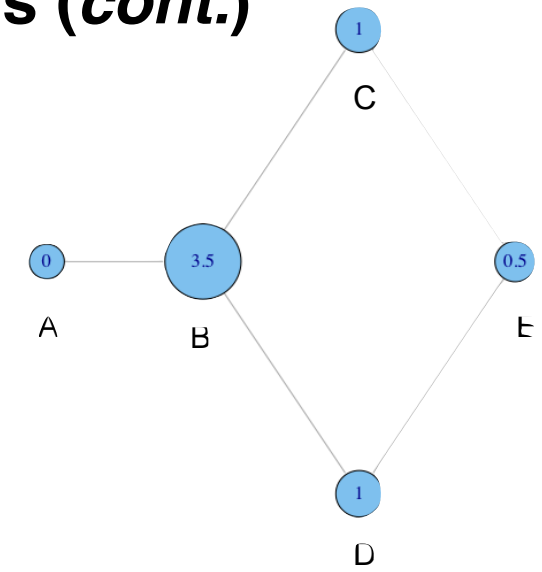


- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

- Note that there are no alternate paths for these pairs to take, so C gets full credit

Determine the number of shortest paths (*cont.*)

- There are two shortest paths between A und E.



- Partial betweenness: Two vertices i and j are indifferent with respect to the number of alternative geodesics, thus, the probability of using any one is $1/g_{ij}$ with g_{ij} is the number of geodesics linking of j and j

Betweeness

- The potential of point v for control of information passing between i and j then may be defined as the probability that v falls on a randomly selected geodesic connecting i and j.

- If $g_{ij}(v)$ = the number of geodesics linking i and j that contain v

  then the betweenness of v is $b_{ij}(v) = \frac{1}{g_{ij}} \times g_{ij}(v)$

- To determine the overall centrality of a vertex v, we sum its partial betweenness values for all unordered pairs of points where i # j # v:

$$C_B(v) = \sum_{i<j}^{N} \frac{g_{ij}(v)}{g_{ij}}$$

- High betweenness vertices can be thought of as a “cutpoint” in the shortest path connecting two other nodes, therefore being a critical link between them.

Normalized Betweenness Centrality

- Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not

- Undirected and directed network (normalized)

$$C'_B(v) = \frac{\sum_{i<j}^{N} = \frac{g_{ij}(v)}{g_{ij}}}{(n-1)(n-2)/2}$$

where
g_{ij} is the number of geodesic paths between two vertices i and j and
$g_{ij}(v)$ is the number of geodesic paths of the two vertices that contain v

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Betweeness Centralization

The central question:
How much variation is there in the centrality scores among the nodes?

$$C_B = \frac{\sum_{i=1}^{n}[C_B(i*) - C_B(i)]}{(n-1)}$$

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Closeness Centrality

What if it's not so important to have many direct friends?

Or be "between" others.

But one still wants to be in the "middle" of things, not too far from the center.

Motivation

- A vertex is viewed as central to the extent that it can avoid the control potential of others
- The independence of a vertex is determined by its closeness to all other vertices in a network (Leavitt, 1951)

- Bavelas (1948) suggested a message originating in the most central position in a network would spread throughout the entire network in minimum time
- Hakimi (1965) and Sabidussi (1966) defined With respect to time or cost efficiency, a vertex is central to the degree that the distances associated with all its geodesics are minimum.

- Short distances mean fewer message transmissions, shorter times and lower costs.

Freeman, Linton C., Douglas Roeder, and Robert R. Mulholland. "Centrality in social networks: II. Experimental results." Social networks 2.2 (1979): 119-141.

Normalized closeness centrality

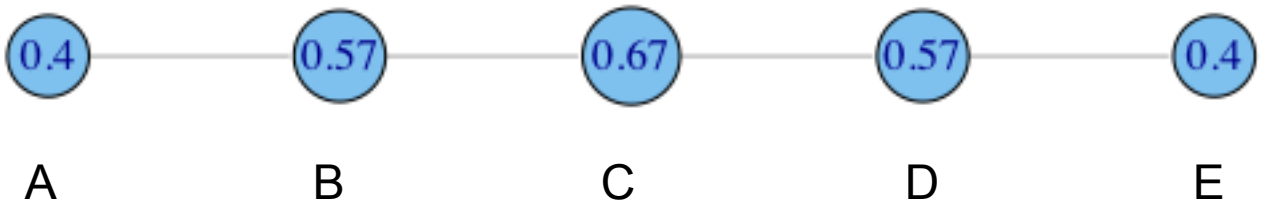
- A vertex is considered important if it is relatively close to all other vertices.
- If we let d(v, u) = the number of edges in the geodesic linking v and u
- Closeness is based on the inverse of the distance of each vertex to every other vertex in the network:

$$C_C(v) = \sum_{v \neq u} \frac{1}{d_{uv}}$$

- Normalized closeness centrality $C'_C(v) = \frac{1}{N-1} \sum_{v \neq u} \frac{1}{d_{uv}}$

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Another example



$$C'_C(v_A) = \frac{N-1}{\sum_{v=1}^{N} d(v_A, v_i)} = \frac{4}{1+2+3+4} = \frac{4}{10} = 0.4$$

Problems of using closeness centrality

- Geodesic distance d_{uv} in networks tends to be small, typically the distance increases logarithmically with the size of the whole network
- The ratio between the smallest distance (1) and the largest (order log n) is itself only log n and rather small
- But the smallest and the largest distance provide lower and upper bounds for the distances
- In practice, it might be difficult to distinguish between central and less central vertices

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Using closeness in disconnected networks

- The problem is if the network consists of components, thus, the geodesic distance between two vertices is infinite

- Strategy 1: Average over only those vertices in the same component as i, but then there is the problem of different component sizes

- Strategy 2: redefine closeness in terms of the harmonic mean distance between vertices (i.e., the average of the inverse distances)

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Closeness centralization

- The central question is:

  How much variation is there in the centrality scores among the nodes?

$$C_B = \frac{\sum_v^N [C'_C(v*) - C'_C(v)]}{(n-2)(n-1)/(2n-3)}$$

Wasserman, S, Faust, K. Social Network Analysis. Cambridge University Press. 1997.

Comparing the three most popular centrality measures

	Low degree	Low closeness	Low Betweenness
High degree		Embedded in cluster that is far from the rest of the network	Ego's connections are redundant communication bypasses him/her
High closeness	Key player tied to important/active alters		Probably multiple paths in the network, ego is near many people, but so are many others
High betweenness	Ego's few ties are crucial for network flow	Ego monopolizes the ties from a small number of people to many others	

Eigenvector Centrality

Recap: World Wide Web

- The fact that your page might point to other important pages is neither here nor there, for example anyone can set up a page that points to a thousand others
- Number and stature of web pages that point to your page can give a reasonable indication of how important or useful your page is
- Similar considerations apply also to citation networks and other directed networks

Eigenvector Centrality

- "Eigenvector centrality can be thought of as an extended form of degree centrality, in which we take into account not only how many neighbors a vertex has but also how central those neighbors themselves are." (Newman, 2010)

- In plain English: The eigenvector centrality of vertex v is the sum of its connections to other vertices, weighted by their centrality

- It can be calculated for either undirected or directed networks but it works best for the undirected case

- Can be calculated in different ways, but commonly it is a matrix-based calculation in which node j's centrality is determined by the jth entry in the eigenvector corresponding to the largest positive eigenvector

Eigenvector centrality

- Let's donate the centrality of vertex i by x_i , then we can allow for this effect by making x_i proportional to the average of the centralities of i's network neighbors

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$$

where λ is a constant.

- Defining the vector of centralities x = (x_1,x_2,...), we can rewrite this equation in matrix form as

$$\lambda x = A \cdot x$$

and hence we see that x is an eigenvector of the adjacency matrix with eigenvalue λ

M. E. J. Newman: The mathematics of networks.

Basic algorithm

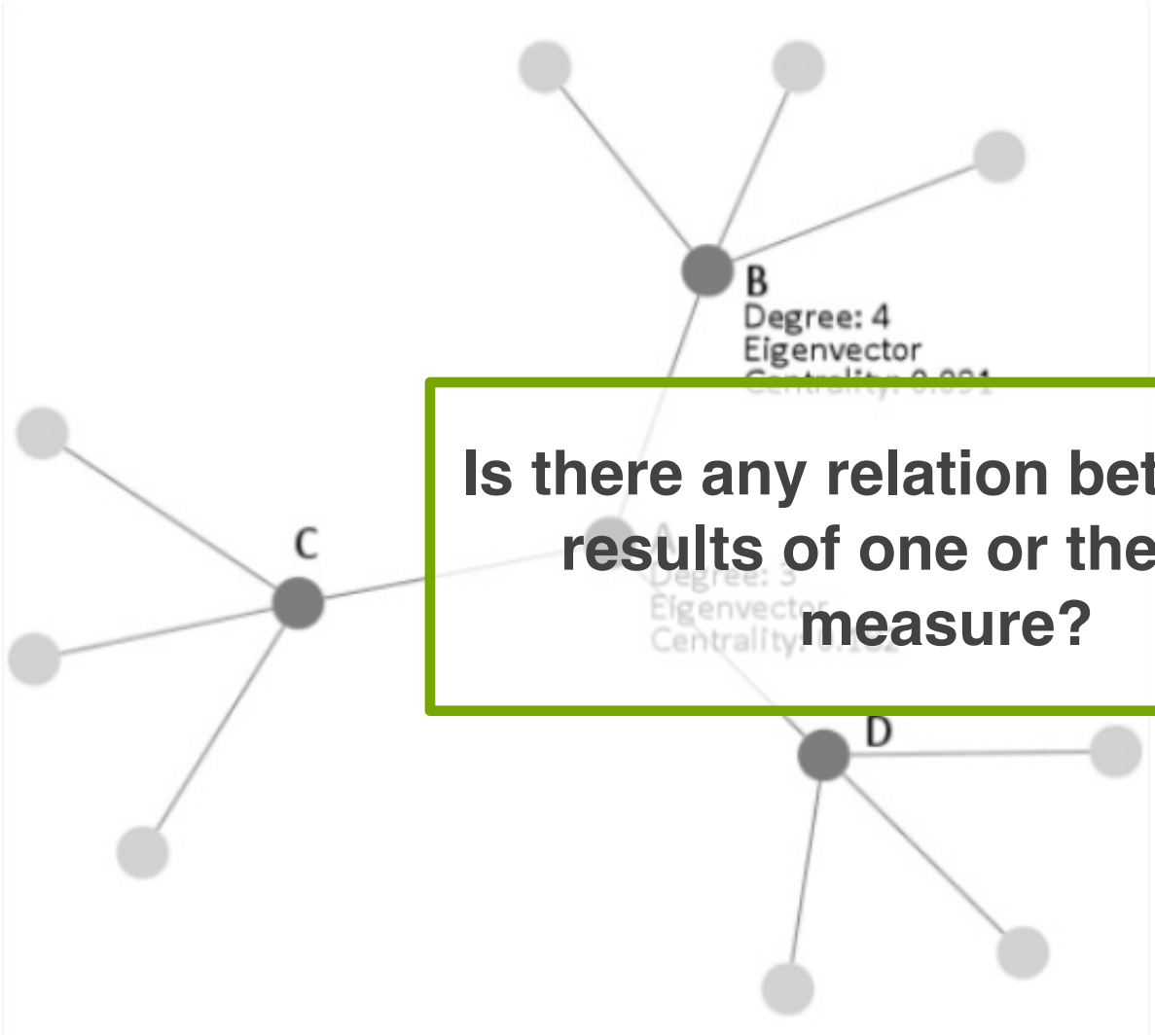
1. Start by assigning centrality score of 1 to all nodes ($v_i = 1$ for all i in the network)

2. Recompute scores of each node as weighted sum of centralities of all nodes in a node's neighborhood: $v_i = \sum_{j \in N} x_{ij} * v_j$

3. Normalize v by dividing each value by the largest value

4. Repeat steps 2 and 3 until values of v stop changing.

Example



Is there any relation between the results of one or the other measure?

Metrics comparison

- Association between two phenomena is usually measured by correlation coefficients

- Correlation coefficients range from 1 to –1
  - A positive coefficient indicates that a high score on one feature is associated with a high score on the other
  - A negative coefficient points toward a negative or inverse relation: a high score on one characteristic combines with a low score on the other

- Rule of thumb
  - No association, if the absolute value of the co- efficient is less than .05.
  - Weak association, if the absolute value of a coefficient is between .05 and .25
  - Moderate association: if the coefficient ranges from .25 to .60 (and from –.25 to –.60)
  - Strong association: if the coefficient ranges from.60 to 1.00 (or –.60 to –1.00)

WOUTER DE NOOY, A. M. A. V. B. (2005). Exploratory Social Network Analysis With Pajek, 1–362.

Measuring correlation

- Pearson's correlation coefficient uses the exact numerical scores on both characteristics

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  Shows linear dependence between variables, $-1 \leq r \leq 1$ (perfect when related by linear function)

- Spearman's rank correlation determines whether the ranking of vertices on one characteristic (e.g., indegree) matches the ranking on another characteristic (e.g., outdegree).
- It is a robust measure of association

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}$$

  Shows strength of monotonic association
  (perfect for monotone increasing/decreasing relationship)

Ranking comparison

- The Kendall tau rank distance is a metric that counts the number of pairwise disagreements between two ranking lists

- Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

$-1 \leq \tau \leq 1$, perfect agreement $\tau = 1$, reversed $\tau = -1$

- Example
  - Rank 1: A B C D E
  - Rank 2: C D A B E

$$\tau = \frac{6-4}{5(5-1)/2} = 0.2$$

Questions?