Statistical Analysis of Networks

Lecture 6 – Basic concepts

NETWORK *CENTRALIZATION*

- combining centrality measures at node level to obtain an aggregate measure of *network centralization*

- the <u>larger</u> the meausure (index): more likely a single node is 'central' with the others considerably less central (in the periphery of a centralized system)

- index of centralization: how variable (heterogenous) the node centralities are

General formula for a centralization index:

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

C_A (i) = highest centrality index in the observed network*
C_A (i) = centrality index of node i

*C_A = **0** : all nodes have the same centrality index (circle graph)*
*C_A = **1** : only one node has the maximun centrality index (star graph)*

(NETWORK) CENTRALIZATION INDECES

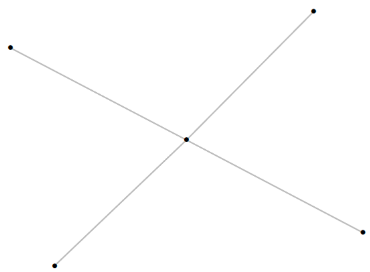
Degree: $$C_D = \frac{\sum_{i=1}^{n}[C_D(i^*) - C_D(i)]}{[(n-1)(n-2)]}$$

Closeness: $$C_C = \frac{\sum_{i=1}^{n}[C_C'(i^*) - C_C'(i)]}{[(n-1)(n-2)]/(2n-3)}$$ [$C_C'(i)$ = normalized closness index]

Betweeness: $$C_B = \frac{2\sum_{i=1}^{n}[C_B(i^*) - C_B(i)]}{[(n-1)^2(n-2)]} = \frac{2\sum_{i=1}^{n}[C_B'(i^*) - C_B'(i)]}{(n-1)}$$ [$C_B'(i)$ = normalized betweeness index]

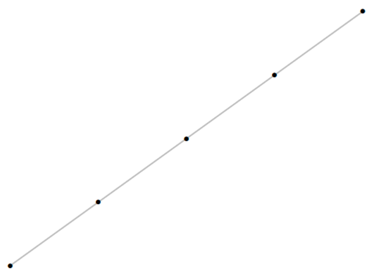
For proofs and computational details, see Freeman (1979)

(Network) centralization comparison



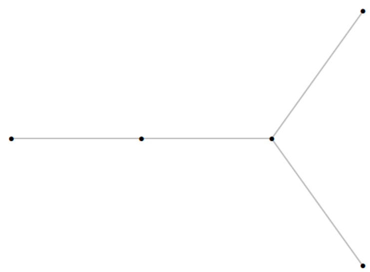
star graph

C_D = 1
C_C = 1
C_B = 1



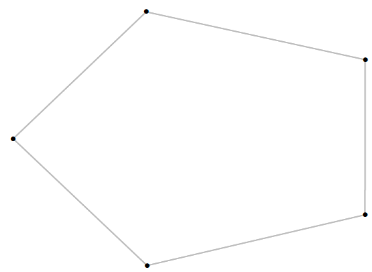
line graph

C_D = .1666667
C_C = .4222222
C_B = .4166667



Y-graph

C_D = .5833333
C_C = .6351852
C_B = .7083333



circle graph

C_D = 0
C_C = 0
C_B = 0

FLORENTINE FAMILIES: CENTRALITY INDECES

Medici is the most central wrt $C_D(i)$ and $C_B(i)$ *larger value wrt other next most central families* (*Guadagni* and Strozzi)

	n = 16		*n = 15*		
	$C'_D(n_i)$	$C'_B(n_i)$	$C'_D(n_i)^*$	$C'_C(n_i)^*$	$C'_B(n_i)^*$
Acciaiuoli	0.067	0.000	0.071	0.368	0.000
Albizzi	0.200	0.184	0.214	0.483	0.212
Barbadori	0.133	0.081	0.143	0.438	0.093
Bischeri	0.200	0.090	0.214	0.400	0.104
Castellani	0.200	0.048	0.214	0.389	0.055
Ginori	0.067	0.000	0.071	0.333	0.000
Guadagni	0.267	0.221	0.286	0.467	0.255
Lamberteschi	0.067	0.000	0.071	0.326	0.000
Medici	0.400	0.452	0.429	0.560	0.522
Pazzi	0.067	0.000	0.071	0.286	0.000
Peruzzi	0.200	0.019	0.214	0.368	0.022
Pucci	0.000	0.000	—	—	—
Ridolfi	0.200	0.098	0.214	0.500	0.114
Salvati	0.133	0.124	0.143	0.389	0.143
Strozzi	0.267	0.089	0.286	0.438	0.103
Tornabuoni	0.200	0.079	0.214	0.483	0.092
Centralization	0.267	0.383	0.257	0.322	0.437

Medici is the most central (with several other families almost as central: Albizzi, Guadagni, Ridolfi,…)

What about Strozzi (or Guadagni)?

What about comparison of the three indeces ?

rather small: not very great difference between the largest and smallest values (little variability)

Less variation in closeness centrality values than degree centrality values: more uniform spread of closeness

EIGENVECTOR CENTRALITY (BONANICH CENTRALITY, 1972)

It is an improvement on the concept of degree centrality

Main idea:
In degree centrality, each neighbor contributes equally to centrality

Bonacich centrality: *important* nodes contribute more

A node is central if it is connected to other central nodes.
More precisely, centrality of a node is proportional to the sum of scores of its neighbors.

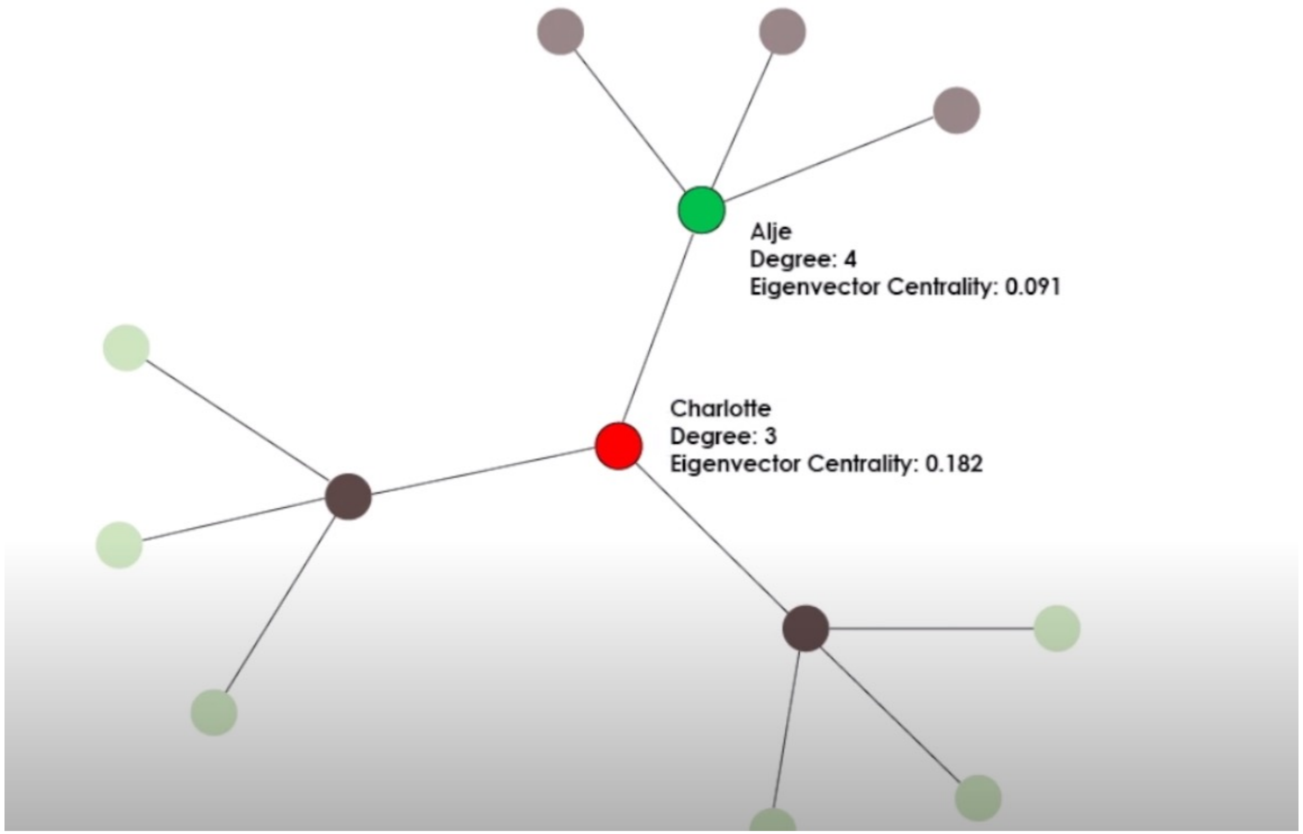
$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}x_j \quad \text{where } \lambda \text{ is a constant}$$

It can be determined by finding the principal eingvector in the adjancecy matrix with eigenvalue λ: ($\boldsymbol{Ax} = \boldsymbol{\lambda x}$)

If centralities have to be non-negative, it can be shown (using the Perron–Frobenius theorem) that λ must be the largest eigenvalue of the adjacency matrix and **x** the corresponding eigenvector.

This notion of centrality is closely related to ways in which scientific journals are ranked based on citations.

EIGENVECTOR CENTRALITY



many other variations on the above definitions: such as PageRank in Google search (*more 'important' websites are likely to receive more links from other websites)* which is related to Katz-Bonacich (centrality based on the number of walks emanating from a node *i,* each exponentially discounted based on their length) and eigenvector centralities (see the PageRank algorithm in igraph)

NODE CENTRALITY AND EDGES CENTRALITY

- centrality indeces: usually referred to nodes
- in some contexts, centrality of edges can be of major concern (often related to edge weight/strenght)

- *Betweeness centrality* is also defined for edges:
  - number of the shortest paths that go through an edge in a graph or network (Girvan and Newman, 2002)

- Not so straightforward for the other meausures:
  - specific definitions and solutions (community detection issues)

DEGREE CORRELATION

- degree distribution f_d summarizes node degree variation in a network
- networks can have the same degree distribution but differ in the way the nodes are associated
- *degree correlation:* basic structural metric that calculates the likelihood that nodes link to nodes of similar or dissimilar nodal degree

  - in many network, hubs - high degree nodes - tend to have ties to other hubs (e.g.: network of Celebrities, CEOs of major corporations)
  - in other networks, hubs tend to link to many small-degree nodes, generating a hub-and-spoke (star) pattern

DEGREE CORRELATION MEASURES

$f(k_1, k_2)$ Joint Degree Distribution

(frequency with which the 2 vertices at the end of an arbitrarily selected edge have a given pairs of degrees)

probability that an edge connects k_1- and k_2-degree nodes

$f(k_1, k_2) = L(k_1, k_2)/L$ if $k_1 = k_2$

$f(k_1, k_2) = L(k_1, k_2)/2L$ if $k_1 \neq k_2$

with $L(k_1, k_2)$ = # of edges connecting nodes of degrees k_1 and k_2

On JDD and its marginal distributions (Kolaczyk, 2009, pp. 86-88):

Pearson correlation coefficient $r(x, y)$ with $X = k_i$ and $Y = k_j$

ASSORTATIVE MIXING BY DEGREE (A VARIATION ON THE CONCEPT OF CORR. COEFFICIENT)

Assortative mixing (or *homophily*) is the tendency of vertices to connect to others that are like them in some way (e.g: with respect to a specific node attributes as gender, race, age, income, type of node, …)

Assortative mixing by degree: the high-degree nodes will be preferentially connected to other high-degree vertices, and the low to low (positive degree correlation)

(in a social network, for example, we have assortative mixing by degree if people with many friends (gregarious) are friends of others with many friends while the hermits have links with other hermits.

Disassortative mixing by degree: the gregarious people were hanging out with hermits and vice versa.

Mixing by degree is itself a property of the network structure not involving exogenous node attributes/characteristics.

(DIS)ASSORTATIVE MIXING BY DEGREE

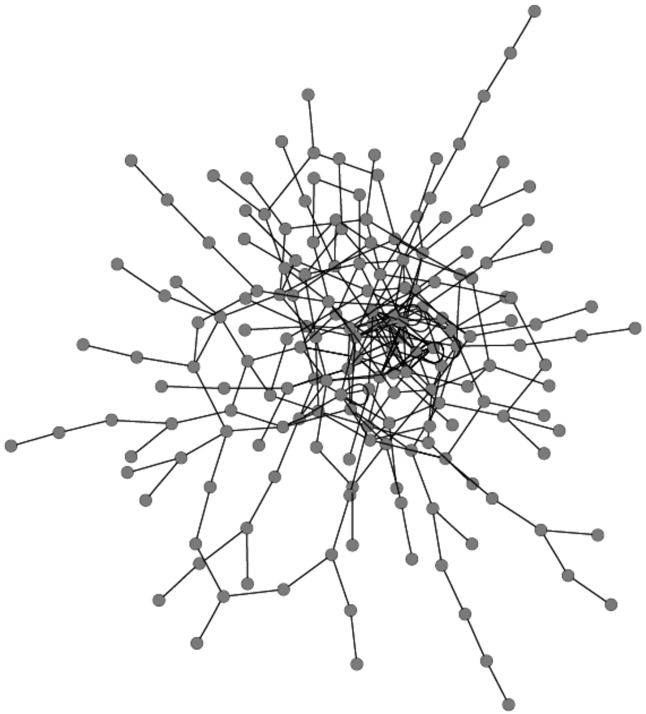
This structural property gives rise to some interesting features in networks:

- in an assortative network by degree (high-degree nodes tend to stick together) one expects to get a clump or core of such high-degree nodes in the network surrounded by a less dense periphery of nodes with lower-degree.

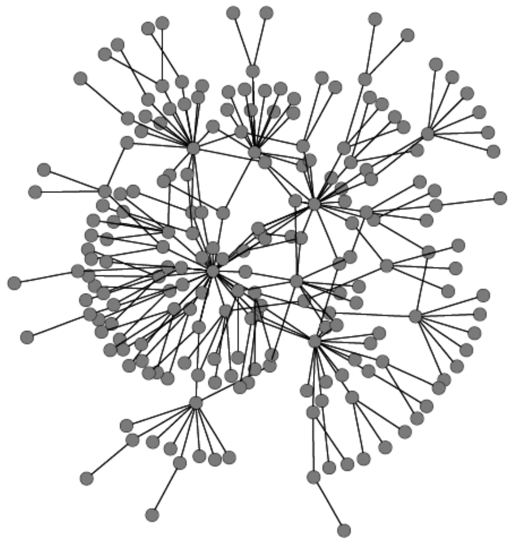
  - core/periphery structure is a common feature of social networks, many of which are found to be assortatively mixed by degree

- in a disassortative network by degree (high-degree nodes tend to connect to low-degree ones) star-like features are often readily visible.
  - disassortatively networks do not usually have a core/periphery split but are instead more uniform.

ASSORTATIVE AND DISASSORTATIVE NETWORKS BY DEGREE



A network that is assortative by degree, displaying the characteristic dense core of high-degree vertices surrounded by a periphery of lower-degree ones



A disassortative network, displaying the star-like structures characteristic of this case

Newman and Girvan (2003)

E-I INDEX

Given a partition of a network into a number of mutually exclusive groups (also defined by some attribute)

the E-I index is the number of ties external to the groups minus the number of ties that are internal to the group divided by the total number of ties:

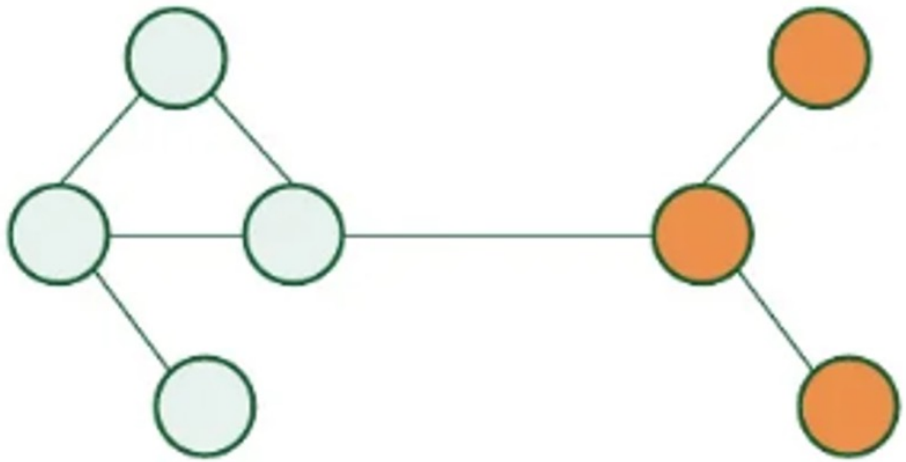
$$EI = \frac{External - Internal}{External + Internal}$$

EI can range from 1 to -1.
EI = -1: complete homophily - the node only has relationships with nodes of the same “type” as they themselves are.
EI = 1: complete heterophily - all the alters are of a different “type” than they themselves are.
EI = 0: an equal number of alters are of both the same “type” as the node, and different types.

(EI is also calculated for each group and for each individual node)

E-I INDEX

number of ties external to the groups minus the number of ties that are internal to the group divided by the total number of ties



(1-4)/7 = -3/7

= -0.43

(1-2)/7 = -1/7

= -0.14

whole network: EI = (1-6)/7 = -0.71

DESCRIPTIVE ANALYSIS OF NETWORK GRAPH CHARACTERISTICS (NETWORK *STATISTICS*/*METRICS*)

Structural analysis of network graphs

two broad categories can be distingished:

1. characterization of individual nodes and edges

2. characterization of network cohesion (involving more than just individual nodes and edges)