Statistical Analysis of Networks

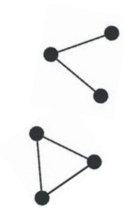
Lecture 5 – Basic concepts

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

Questions of interest phrased as questions on aspects of the structure (characteristics) of the network graph:

e.g.:

- underlying structural (social) processes: particular patterns of triplets of nodes (triad configurations: 2-stars or cyclic or transitive closed triangles)



- flows of information (knowledge) or commodities: presence of paths in the network
- importance of individual units: how *central* is the correspondent node in the network (according some suitable definition/meausurement of centrality)
- presence of 'groups' of nodes characterised by specific network behaviour (community detection): detection of sub-graphs by means of network partition

→ structural analysis of network graphs: traditionally a descriptive task

(mainly performed with graph -mathematical and computer sciences- tools than statistical ones)

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

<u>Structural analysis of network graphs</u>

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)

DESCRIPTIVE ANALYSIS OF NETWORK GRAPH CHARACTERISTICS
DEGREE OF A NODE (OR NODAL DEGREE)

$d(n_i)$ $[d_i]$ = # of lines that are incident with i
or, equivalentely,
= # of nodes adjacent to i

$d(n_i) = 0$ *(min value)* if no nodes are adjacent to i (isolated node)
$d(n_i) = n - 1$ *(max value) if* i *is adjacent to all other nodes in the network*

undirected

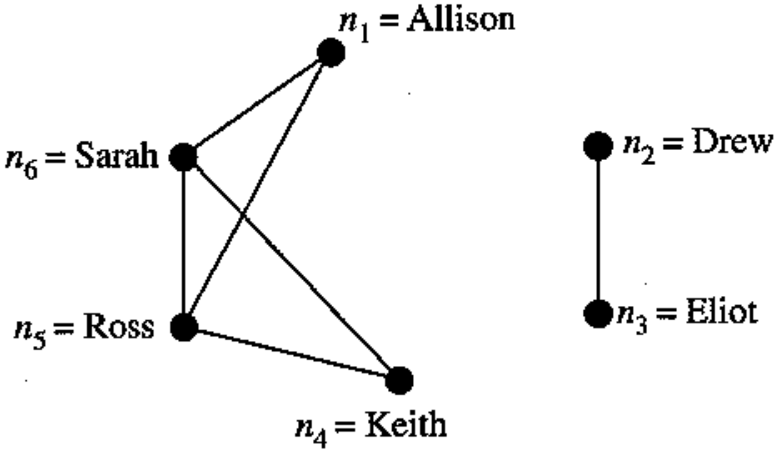
Mean degree $M(d) = \sum_n d_i/n = 2L/n$ $(2 * L$ (# of lines) $= \sum$ degrees)

Degree: basic quantification of the extent to which a node is connected to other nodes in the graph (in SNA: measure of the 'activity' of a node)

Important notion: *degree distribution*

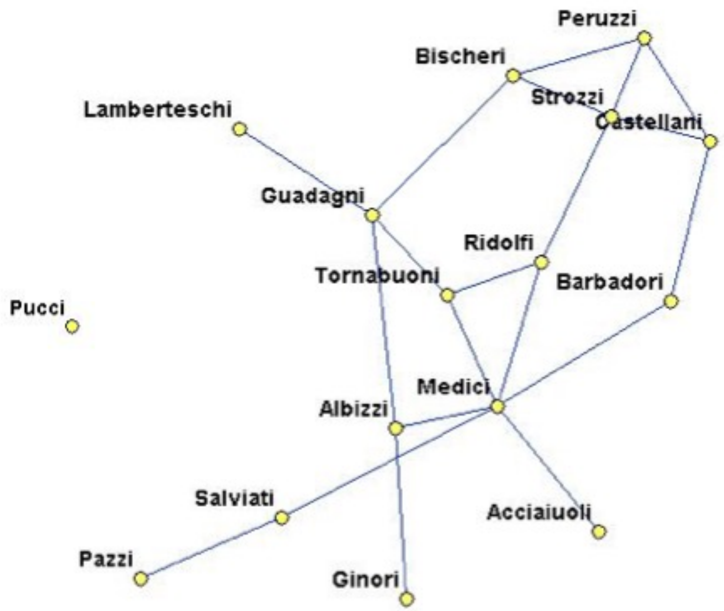
DEGREE – UNDIRECTED NETWORK

Graph of "lives near" relation for six children



$$d(n_1) = 2, d(n_2) = 1, d(n_3) = 1, d(n_4) = 2, d(n_5) = 3, d(n_6) = 3.$$

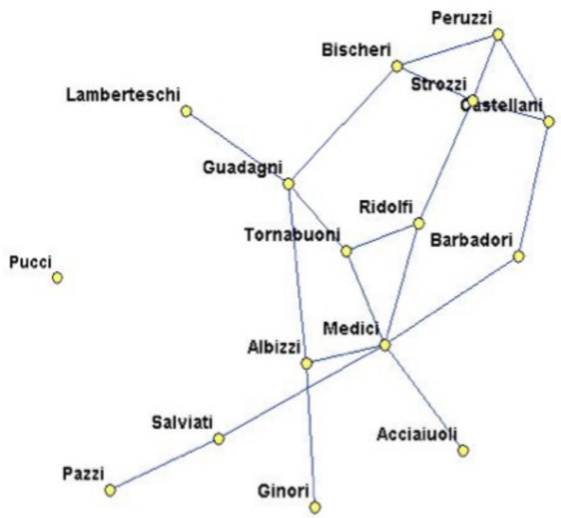
Florentine Families (Padgett, 1994)
marriage relations between pairs of families in the early 15 century



16 families

One mode
undirected network

FLORENTINE FAMILIES – DEGREE

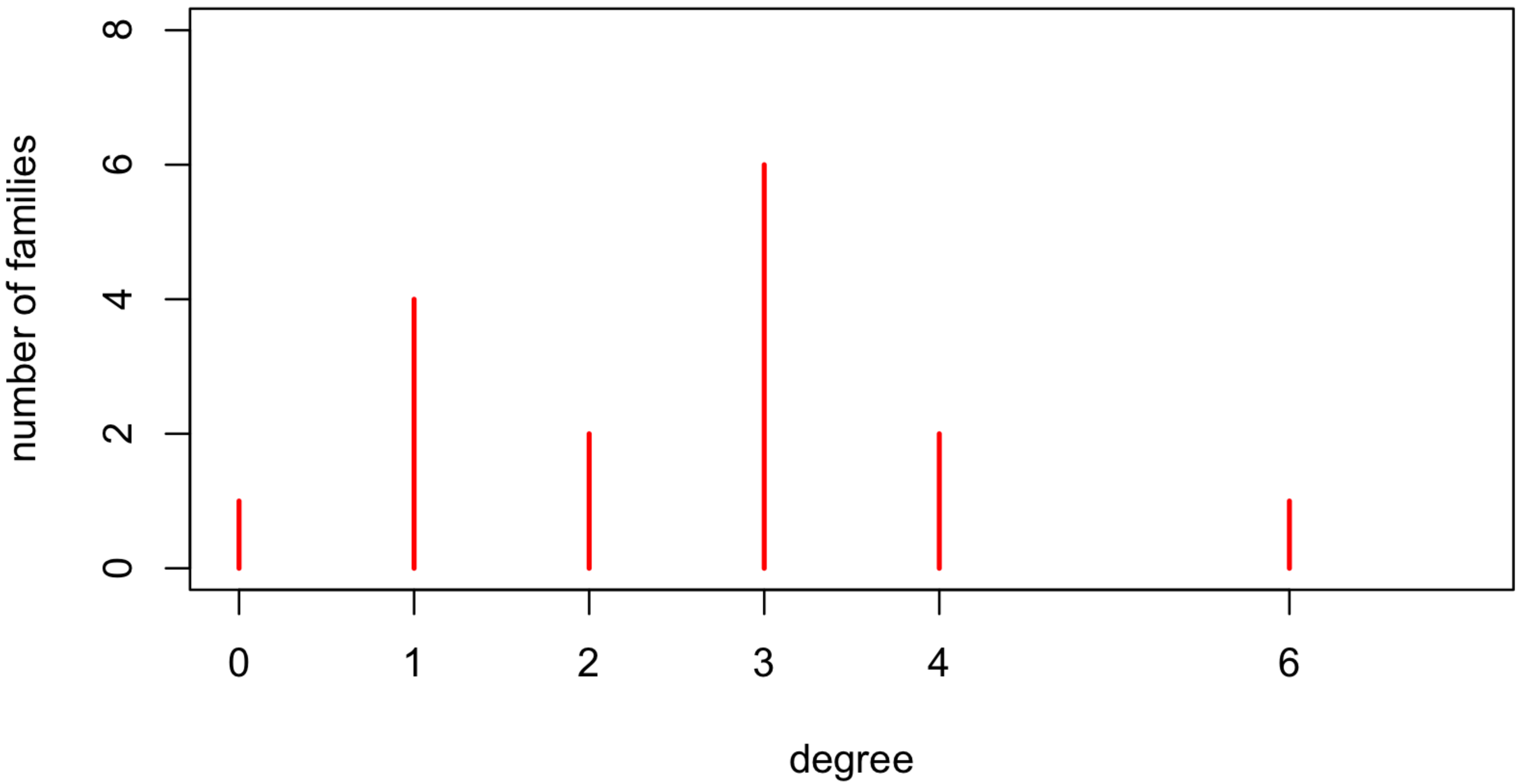


	Degree
Acciaiuoli	1
Albizzi	3
Barbadori	2
Bischeri	3
Castellani	3
Ginori	1
Guadagni	4
Lamberteschi	1
Medici	6
Pazzi	1
Peruzzi	3
Pucci	0
Ridolfi	3
Salviati	2
Strozzi	4
Tornabuoni	3

M(d) = 2.5

Medici, Strozzi, Guadagni, Peruzzi, Bischeri, ..., Tornabuoni are the families with higher degree (especially Medici family)

Florentine Families – degree distribution (absolute values)



FLORENTINE FAMILIES – ADJACENCY MATRIX

Family		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Acciaiuoli	**1**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2 Albizzi	**2**	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
3 Barbadori	**3**	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
4 Bischeri	**4**	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
5 Castellani	**5**	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
6 Ginori	**6**	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7 Guadagni	**7**	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
8 Lambertesc	**8**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9 Medici	**9**	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1
10 Pazzi	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
11 Peruzzi	**11**	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
12 Pucci	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13 Ridolfi	**13**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
14 Salviati	**14**	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
15 Strozzi	**15**	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0
16 Tornabuoni	**16**	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0

$$A_{i+} = \sum_j A_{ij} = d_i$$

DEGREE – DIRECTED NETWORK

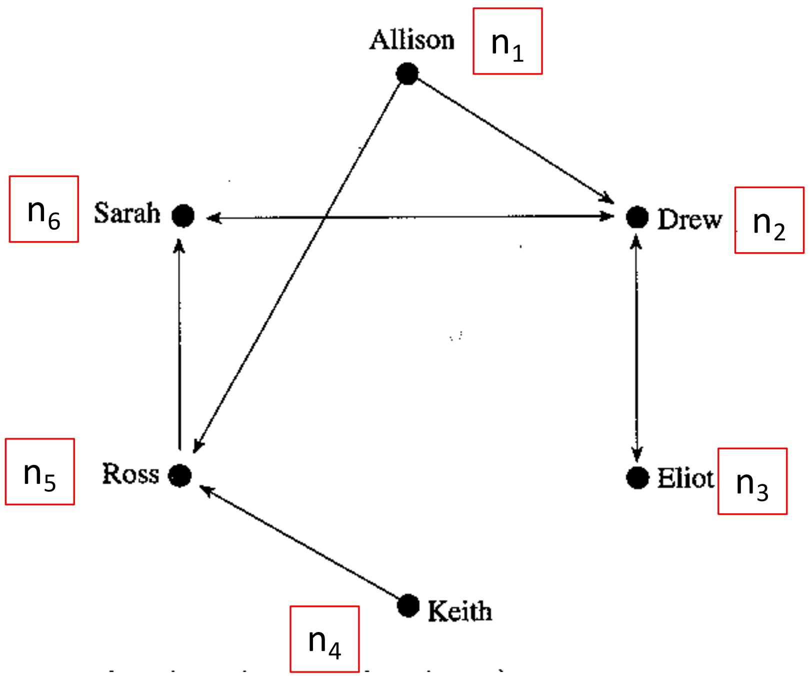
A node can be *adjacent to* and *adjacent from* an other node, depending of the *direction* of the arc

$d_{in}(i)$ = # of nodes adjacent to i *(#* of arcs terminating at i*)*
$d_{out}(i)$ = # of nodes adjacent from i *(#* of arcs originating with i*)*

L (# of lines) $= \sum d_{in} = \sum d_{out}$

Mean degree $M(d) = \sum d_{in}/n = \sum d_{out}/n = L/n$
in SNA:
In degree: 'popularity' measure of a node
Out degree: 'expansiveness' measure of a node

FRIENDSHIP AT THE BEGINNING OF THE COURSE



In degree

- $d_I(n_1) = 0$
- $d_I(n_2) = 3$
- $d_I(n_3) = 1$
- $d_I(n_4) = 0$
- $d_I(n_5) = 2$
- $d_I(n_6) = 2$

Out degree

- $d_O(n_1) = 2$
- $d_O(n_2) = 2$
- $d_O(n_3) = 1$
- $d_O(n_4) = 1$
- $d_O(n_5) = 1$
- $d_O(n_6) = 1$

NETWORK DENSITY

Density = proportion of *possible* edges actually present in the graph (determined by # of nodes *n*)

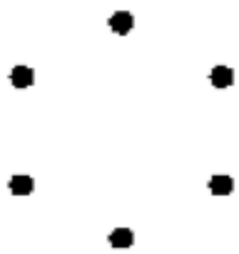
- Undirected graph:
  Max # of lines: $n(n-1)/\,2$ (no loops), *L # of observed edges*
  $D = L/[n\,(n-1)/2] \; = \; 2L/[n\,(n-1)]$

- Directed graph:
  Max # of lines: $n(n-1)$ (no loops)

  $D = L/[n\,(n-1)]$

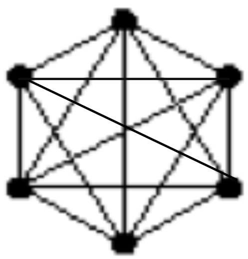
NETWORK DENSITY

Empty graph



Density $D = 0$

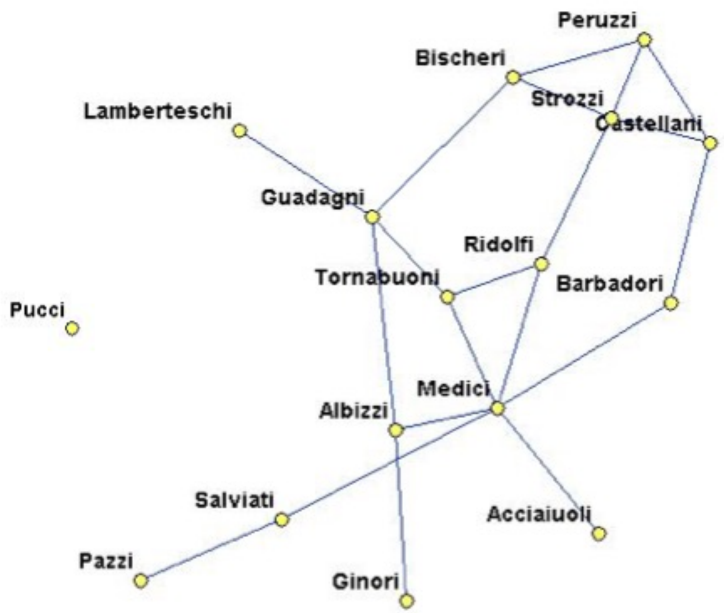
Complete graph



Density $D = 1$

- Sparse network: the number of lines is of the same order as the number of vertices ($m \approx kn$). In real life, many networks are very large but sparse.
- Dense networks: in general, the number of lines can be much higher than the number of vertices.

FLORENTINE FAMILIES - DENSITY



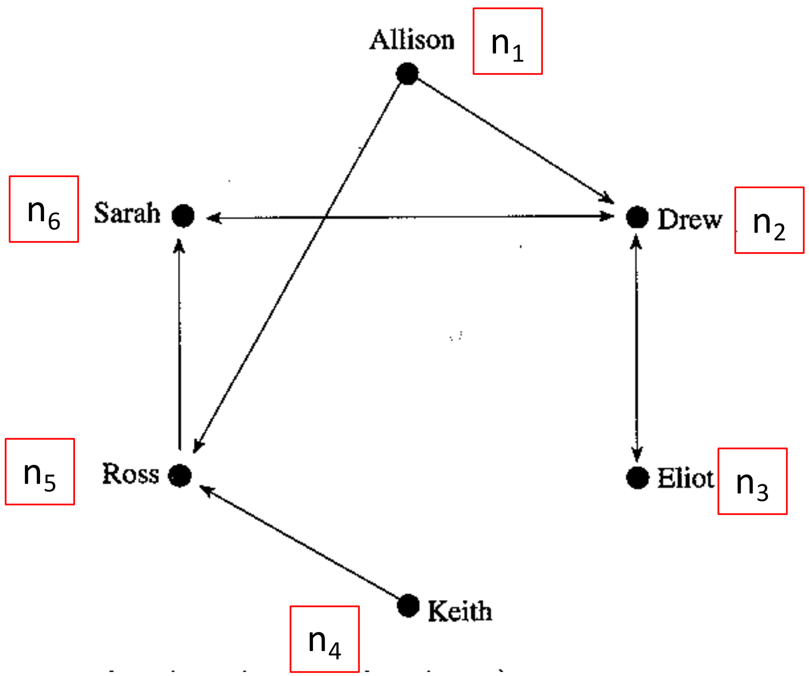
16 families

One mode
undirected network

20 marriage relations

$$D = 2 * 20/(16 * 15) = .167$$

FRIENDSHIP AT THE BEGINNING OF THE COURSE - DENSITY



6 students

One mode
directed network

8 friendship relations

$$D = 8/(6 * 5) = .267$$

NETWORKS AND DEGREE (BARABASI, 2016)

Network	Nodes (N)	Links (L)	Type	N	L	Av. Degree ($\langle d \rangle$)
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

or $\langle k \rangle$ with k_i = degree of node i

What is ?

DEGREE DISTRIBUTION

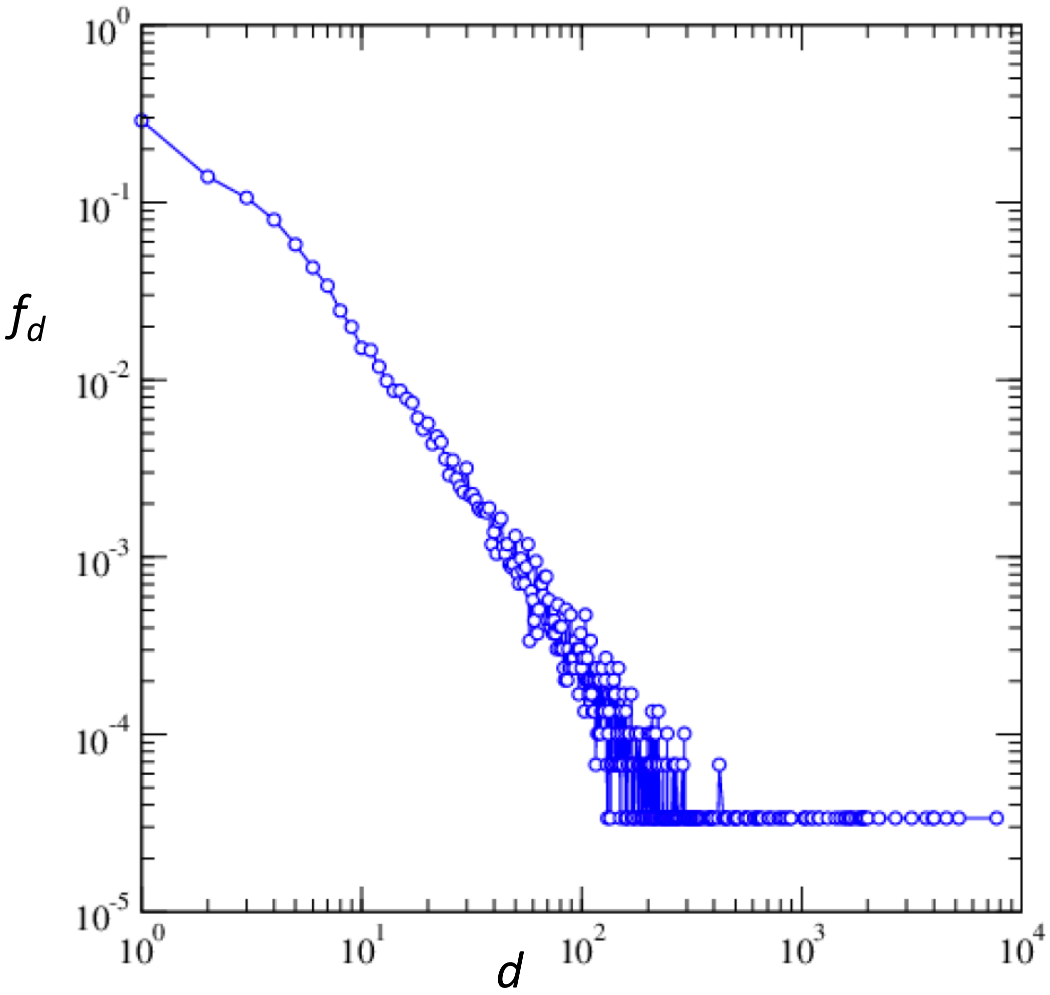
Degree distribution (network of *n* nodes) is the *relative frequency distribution of the node degrees*

$f_d = n_d / n$ (n_d = # of degree-*d* nodes)

f_d = probability that a randomly selected node has degree *d*

Degree distribution: central role in network theory since the discovery of specific functional forms (i.e. power-law) of the distribution in several real (sometimes large) networks is related to many network phenomena and network properties

DEGREE DISTRIBUTION

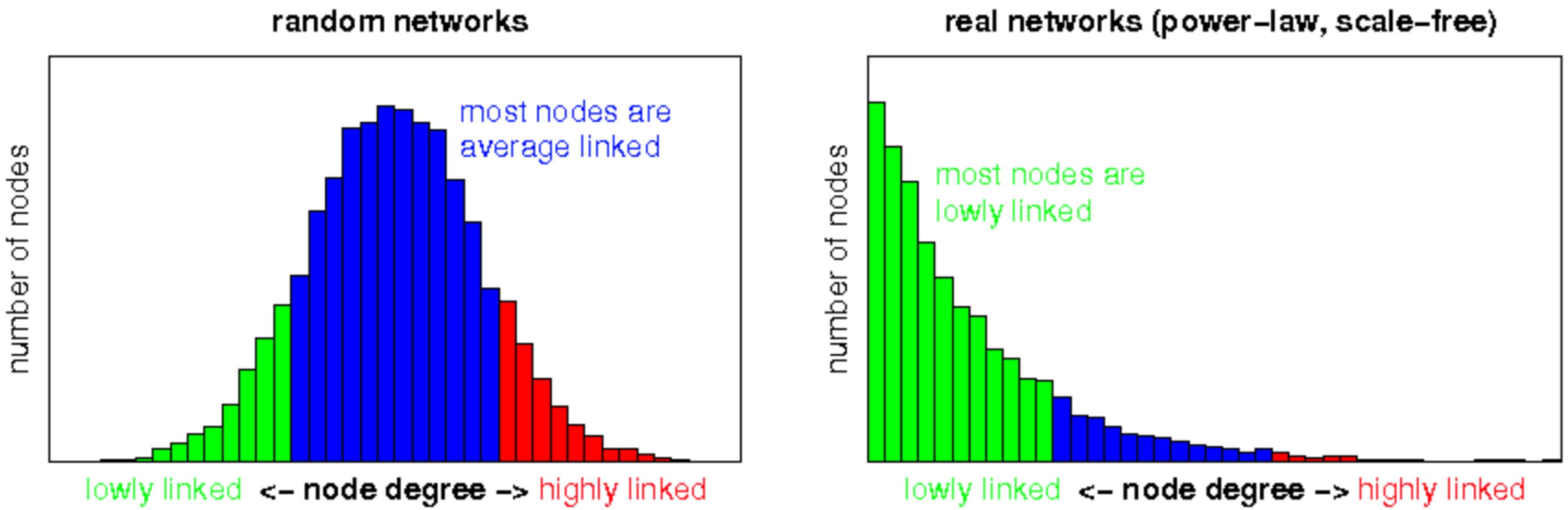


- Global syntactic dependency network (English)
- Nodes: words
- Links: syntactic dependencies
- Many degrees occurring just once!

a kind of dependency in Dependency Grammar in which linguistic units, e.g. words, are connected to each other by directed links

See also Kolaczyk (2009), pp. 81-82

DEGREE DISTRIBUTION



What is the functional form of f_d ?

NODE'S CENTRALITY

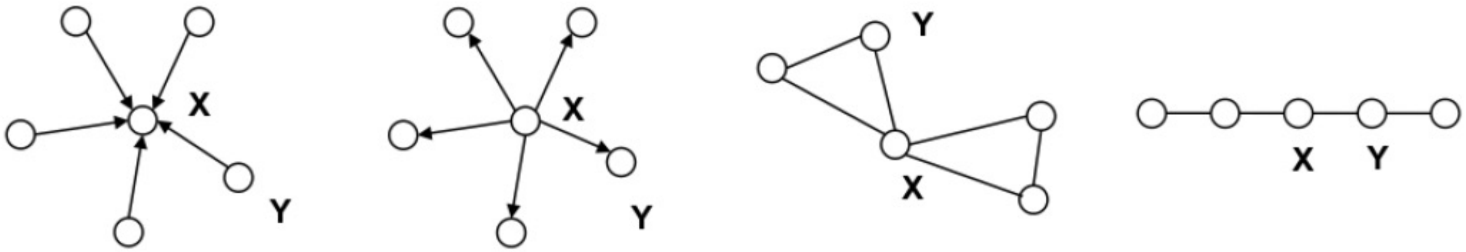
Centrality is a node's meausure (index) *w.r.t.* other nodes:

- identification (quantification) of the *importance* (prominence) of a node in the network structure
- several definition of *importance:* a *variety* of centrality measures
  - related to describe 'node location' in the network
  - most important nodes are usually located in strategic 'positions' within the network
  - general idea in SNA: actors occupying central positions have more opportunities with respect to those having more peripheral positions in the network

majority of the concepts underlying centrality meausures: designed for undirected dichotomous relations (with extensions to directed dichotomous relations)

WHAT DOES CENTRALITY MEAN?

- A central node is *important* and/or *powerful*
- A central node has an *influential position in the network*
- A central node has an *advantageous position in the network*



CENTRALITY MEASURES

- Graph-theoretical centrality
  - Degree centrality
  - Closeness centrality
  - Betweenness centrality

- Eigenvector-based centrality
  - Eigenvector centrality

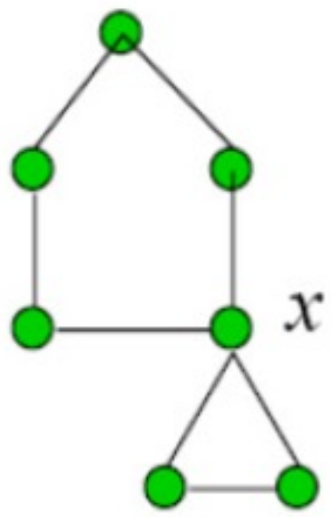
DEGREE CENTRALITY

Importance (power) through *connections*

- Degree Centrality $C_D(i)$ is based on node degree
  - $C_D(i) = d_i$, where d_i = is the degree of i

Normalized index (in the range [0,1]): degree centrality is divided by the maximum possible degree centrality value (= $n - 1$)

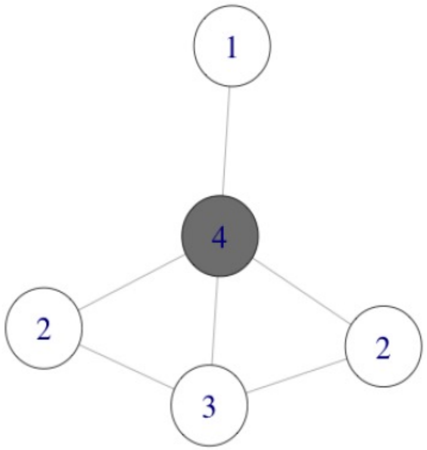
- It measures the "*communication potential*" of i



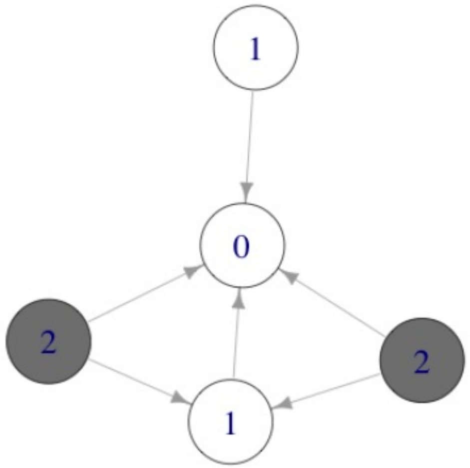
$C_D(x) = 4$, normalized: 0.67

IN/OUT DEGREE CENTRALITY

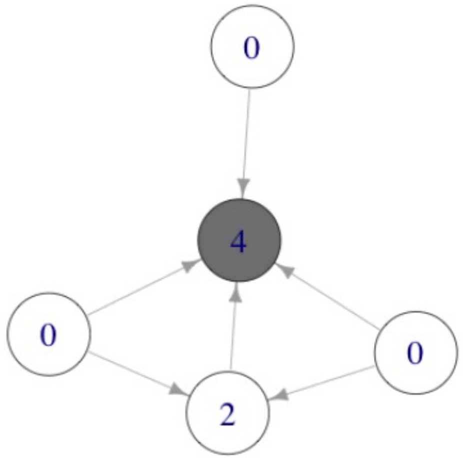
Undirected



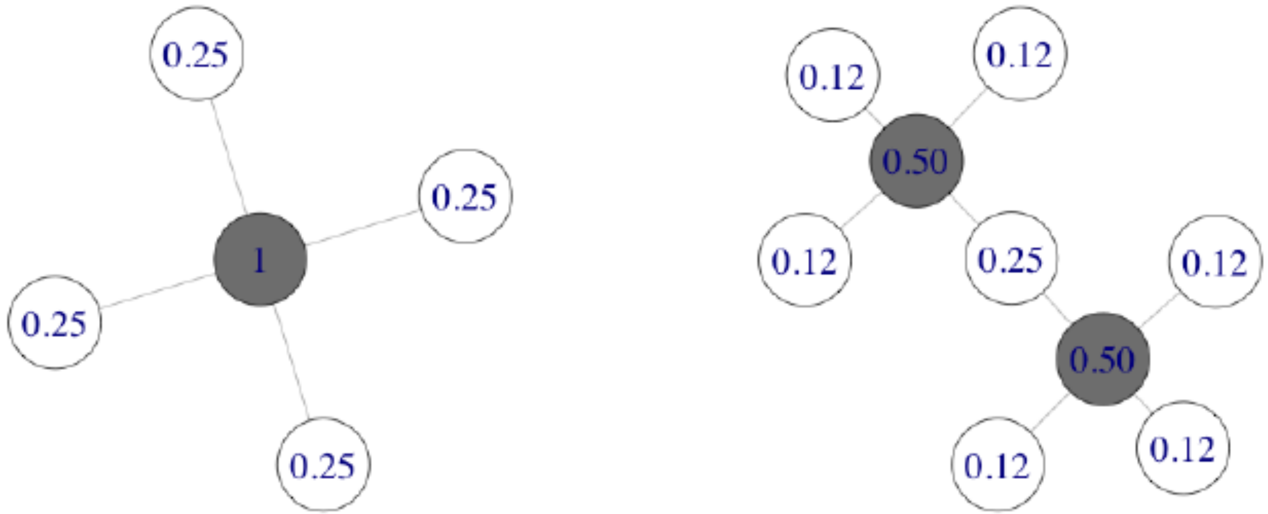
Directed: outdegree centrality (most used)



Directed: indegree centrality



DEGREE CENTRALITY



What can we say about degree centrality here?

Closeness centrality

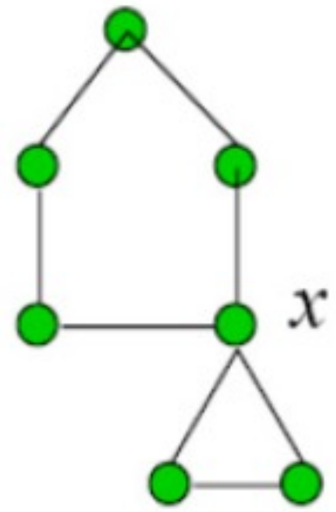
Importance (power) through *proximity to others*

- Closeness $C_c(i)$ is based on geodesic distances

$$\left(\sum_{i \neq j} g_{ij}\right)^{-1} = \frac{1}{\sum_{i \neq j} g_{ij}}$$

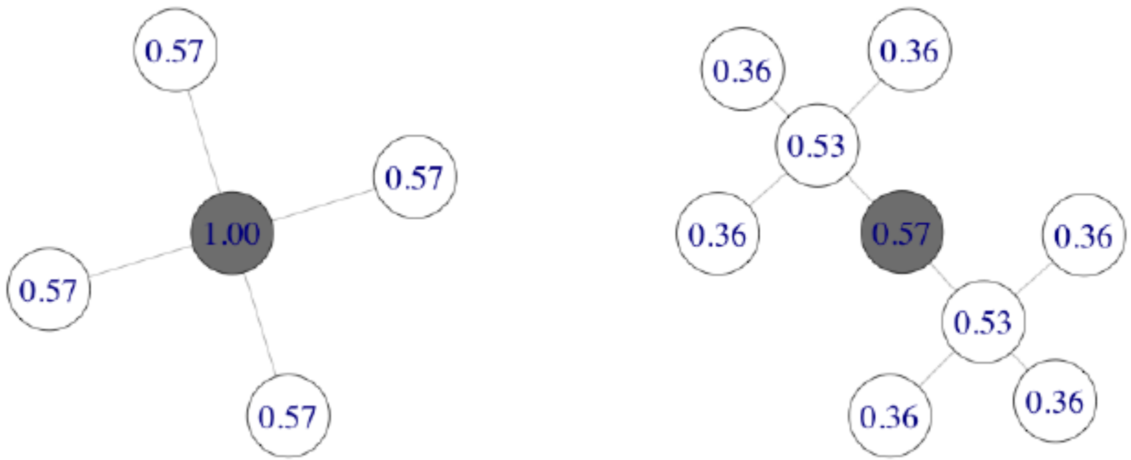
Normalized index (in the range [0,1]): closeness centrality is divided by the maximum possible closeness centrality value (= $1/(n-1)$)
the node is <u>adjacent</u> to <u>all</u> other nodes

- It measures the proximity of node i with respect to the others in the network ("*independent communication potential*")



$C_c(x) = 1/[1+1+1+1+2+2] = 1/8 = .125$, normalized: 0.75

CLOSENESS CENTRALITY



What matters is to be close to everybody else, i.e., to be easily reachable or have the power to quickly reach others.

Closeness centrality in directed network (digraph)

- Undirected network
  - closeness index is only meaningful in a connected graph
  - in presence of disconnected graphs, it is usually computed on the *giant component*
    - i.e. the largest subgraph in terms of number of reachable connected nodes included in it

- Directed network
  - closeness index is only meaningful in a *strongly* connected digraph: every node is reachable by a path (the path from n_i to n_j *can be different from the* path from n_j to n_i *)*

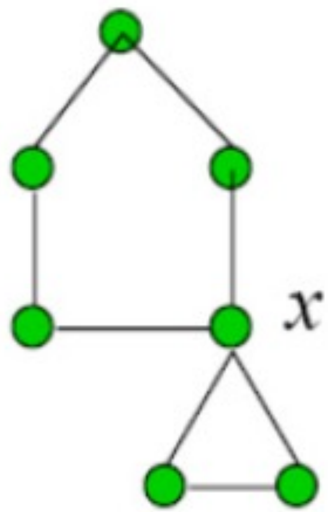
BETWEENESS CENTRALITY

Importance (power) through *brokerage*

- Betweenness $C_B(i)$ is based on # shortest paths passing through a node
  - $$C_B(i) = \sum_{j,k \in V, j \neq k} \frac{g_{jk}(i)}{g_{jk}}$$

g_{jk} is the number of shortest-paths between j and k, and $g_{jk}(i)$ is the number of shortest-paths through i

Normalized index (in the range [0,1]): betweeness centrality is divided by the maximum possible betweeness centrality value (= [(n – 1) (n – 2)] /2)

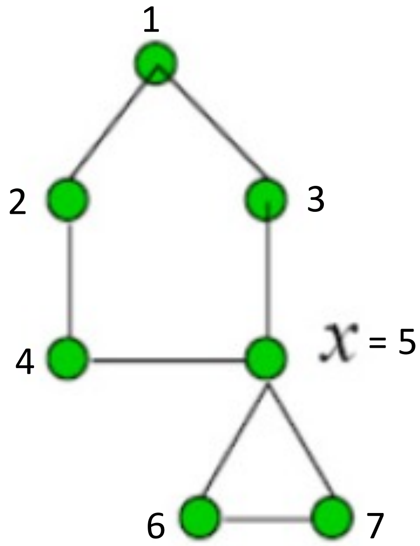


C_B (x) = 9

- It measures the potential of a node in "controlling" the communication flows (*broker* or *bridging* role) (based on the assumption that shortest paths are important to propagate information)

Betweeness centrality

(not normalized)



$C_B(x) = 9$

$x = 5$

$g_{1,2}(5)/ g_{1,2} = 0/1$
$g_{1,3}(5)/ g_{1,3} = 0/1$
$g_{1,4}(5)/ g_{1,4} = 0/1$
$g_{1,6}(5)/ g_{1,6} = 1/1$
$g_{1,7}(5)/ g_{1,7} = 1/1$
$g_{2,3}(5)/ g_{2,3} = 0/1$
$g_{2,4}(5)/ g_{2,4} = 0/1$
$g_{2,6}(5)/ g_{2,6} = 1/1$
$g_{2,7}(5)/ g_{2,7} = 1/1$
$g_{3,4}(5)/ g_{3,4} = 1/1$
$g_{3,6}(5)/ g_{3,6} = 1/1$
$g_{3,7}(5)/ g_{3,7} = 1/1$
$g_{4,6}(5)/ g_{4,6} = 1/1$
$g_{4,7}(5)/ g_{4,7} = 1/1$

Betweeness centrality

(not normalized)