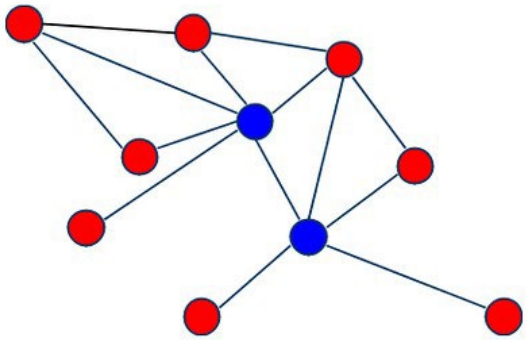


Social Network Analysis (SNA)

Frank Dignum

(slides based on slides from)

Dr. Giorgos Cheliotis (NUS) and Helle Hansen (TUD)

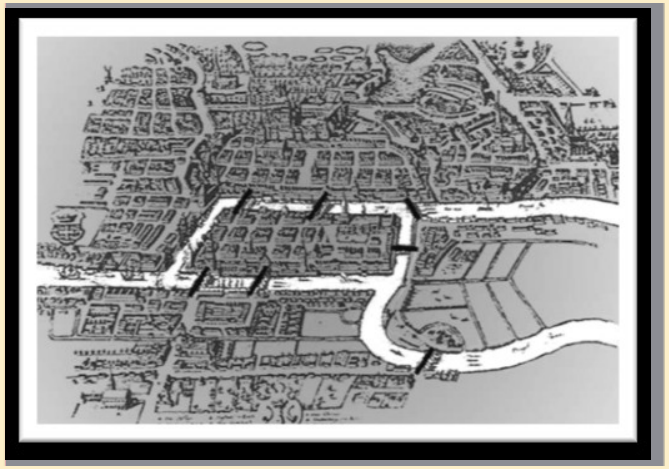
Background: Network Analysis

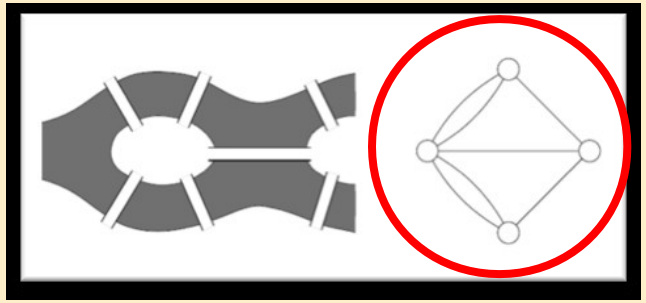
SNA has its origins in both social science and in the broader fields of *network analysis* and *graph theory*

Network analysis concerns itself with the formulation and solution of problems that have a network structure; such structure is usually captured in a *graph* (see the circled structure to the right)

Graph theory provides a set of abstract concepts and methods for the analysis of graphs. These, in combination with other analytical tools and with methods developed specifically for the visualization and analysis of social (and other) networks, form the basis of what we call SNA methods.

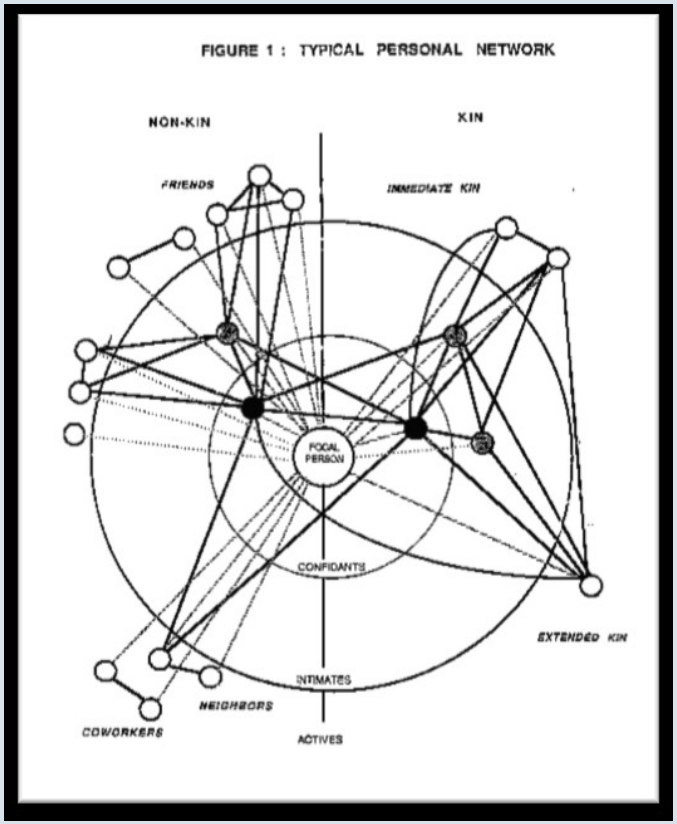
But SNA is not just a methodology; it is a unique perspective on how society functions. Instead of focusing on individuals and their attributes, or on macroscopic social structures, it centers on *relations* between individuals, groups, or social institutions





A very early example of network analysis comes from the city of Königsberg (now Kaliningrad). Famous mathematician Leonard Euler used a graph to prove that there is no path that crosses each of the city's bridges only once (Newman et al, 2006).

Background: Social Science



Wellman, 1998

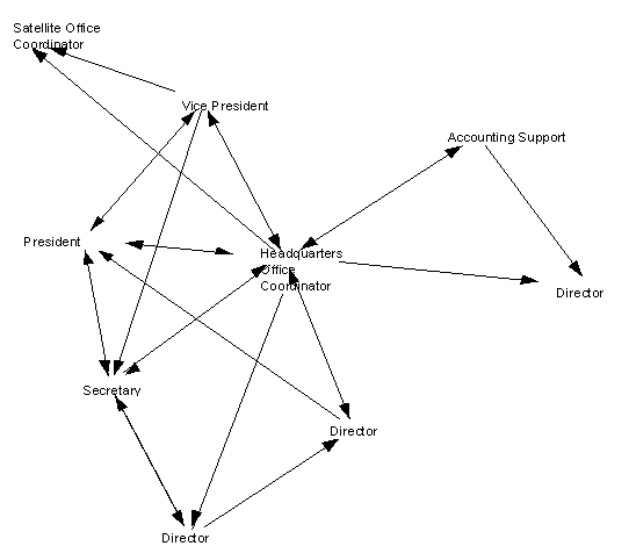
This is an early depiction of what we call an 'ego' network, i.e. a personal network. The graphic depicts varying tie strengths via concentric circles (Wellman, 1998)

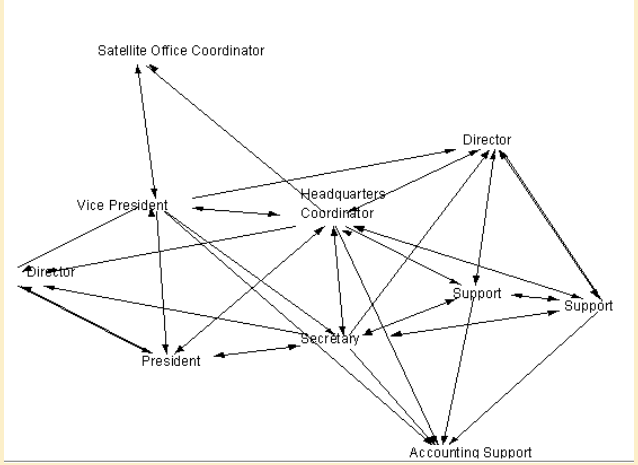
Studying society from a network perspective is to study individuals as embedded in a network of relations and seek explanations for social behavior in the structure of these networks rather than in the individuals alone. This 'network perspective' becomes increasingly relevant in a society that Manuel Castells has dubbed the network society.

SNA has a long history in social science, although much of the work in advancing its methods has also come from mathematicians, physicists, biologists and computer scientists (because they too study networks of different types)

The idea that networks of relations are important in social science is not new, but widespread availability of data and advances in computing and methodology have made it much easier now to apply SNA to a range of problems

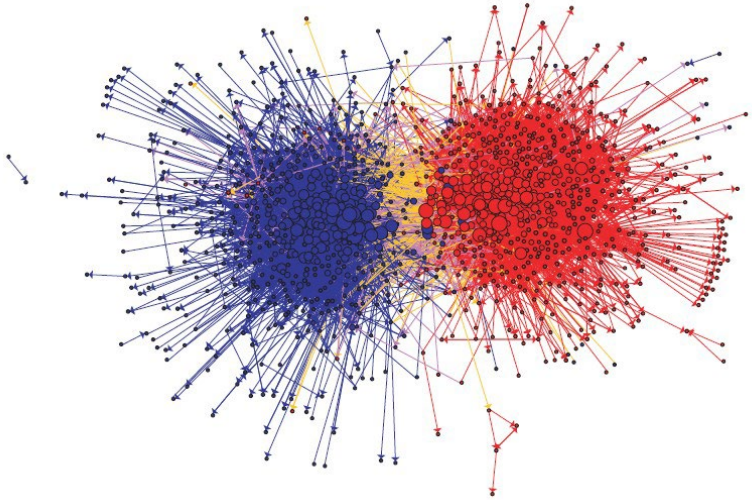
More examples from social science





These visualizations depict the flow of communications in an organization before and after the introduction of a content management system (Garton et al, 1997)

A visualization of US bloggers shows clearly how they tend to link predominantly to blogs supporting the same party, forming two distinct clusters (Adamic and Glance, 2005)



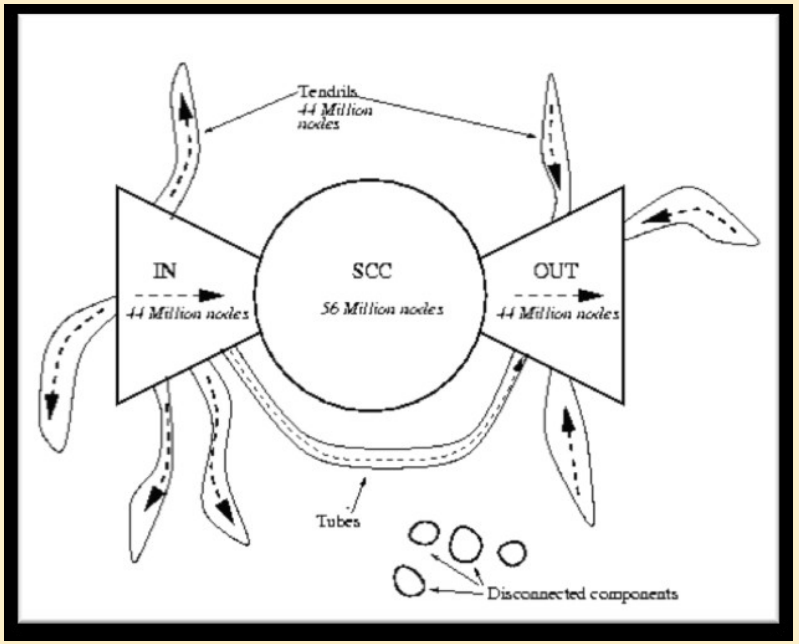
Background: Other Domains

(Social) Network Analysis has found applications in many domains beyond social science, although the greatest advances have generally been in relation to the study of structures generated by humans

Computer scientists for example have used (and even developed new) network analysis methods to study webpages, Internet traffic, information dissemination, etc.

One example in life sciences is the use of network analysis to study food chains in different ecosystems

Mathematicians and (theoretical) physicists usually focus on producing new and complex methods for the analysis of networks, that can be used by anyone, in any domain where networks are relevant



Broder et al, 2000

In this example researchers collected a very large amount of data on the links between web pages and found out that the Web consists of a core of densely inter-linked pages, while most other web pages either link to or are linked to from that core. It was one of the first such insights into very large scale human-generated structures (Broder et al, 2000).

Practical applications

Businesses use SNA to analyze and improve communication flow in their organization, or with their networks of partners and customers

Law enforcement agencies (and the army) use SNA to identify criminal and terrorist networks from traces of communication that they collect; and then identify key players in these networks

Social Network Sites like Facebook use basic elements of SNA to identify and recommend potential friends based on friends-of-friends

Civil society organizations use SNA to uncover conflicts of interest in hidden connections between government bodies, lobbies and businesses

Network operators (telephony, cable, mobile) use SNA-like methods to optimize the structure and capacity of their networks

Why and when to use SNA

Whenever you are studying a social network, either offline or online, or when you wish to understand how to improve the effectiveness of the network

When you want to visualize your data so as to uncover patterns in relationships or interactions

When you want to follow the paths that information (or basically anything) follows in social networks

When you do quantitative research, although for qualitative research a network perspective is also valuable

(a) The range of actions and opportunities afforded to individuals are often a function of their positions in social networks; uncovering these positions (instead of relying on common assumptions based on their roles and functions, say as fathers, mothers, teachers, workers) can yield more interesting and sometimes surprising results

(b) A quantitative analysis of a social network can help you identify different types of actors in the network or key players, whom you can focus on for your qualitative research

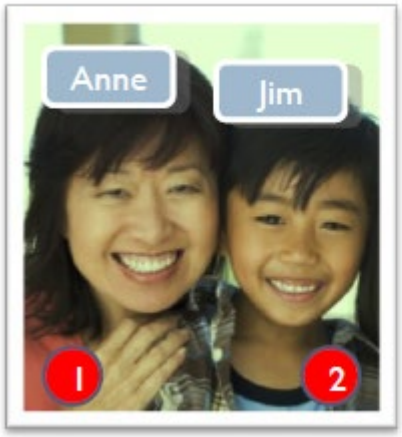
SNA is clearly also useful in analyzing SNS's, OC's and social media in general, to test hypotheses on online behavior and CMC, to identify the causes for dysfunctional communities or networks, and to promote social cohesion and growth in an online community

Basic Concepts

Networks
Tie Strength
Key Players
Cohesion

How to represent various social networks
How to identify strong/weak ties in the network
How to identify key/central nodes in network
Measures of overall network structure

Representing relations as networks



Can we study their interactions as a network?

Communication

Anne: Jim, tell the Murrays they're invited

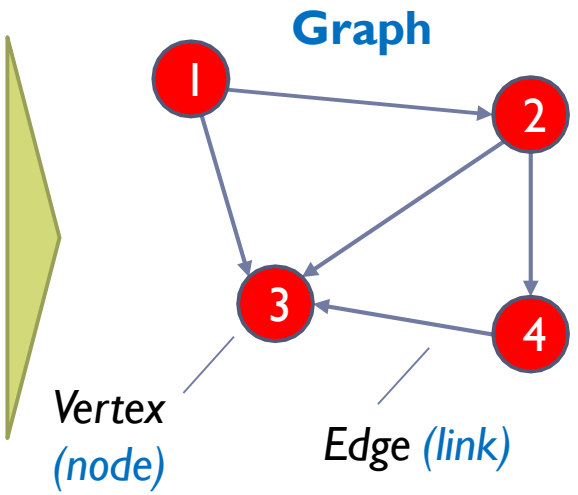
Jim: Mary, you and your dad should come for dinner!

Jim: Mr. Murray, you should both come for dinner

Anne: Mary, did Jim tell you about the dinner? You must come.

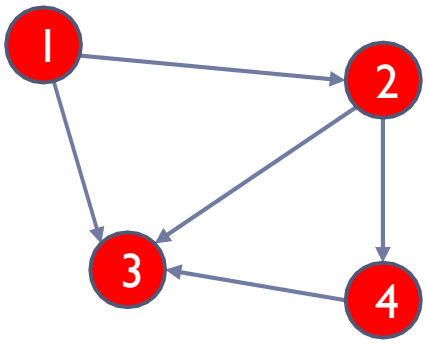
John: Mary, are you hungry?

…



Entering data on a directed graph

Graph (directed)



Edge list

Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

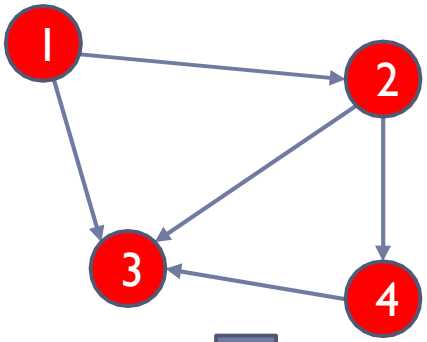
Adjacency matrix

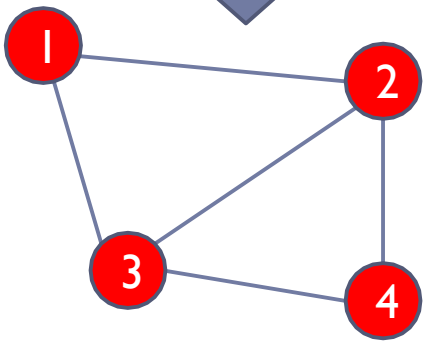
Vertex	1	2	3	4
1	-	1	1	0
2	0	-	1	1
3	0	0	-	0
4	0	0	1	-

Representing an undirected graph

Directed

(who contacts whom)





Undirected

(who knows whom)

Edge list remains the same

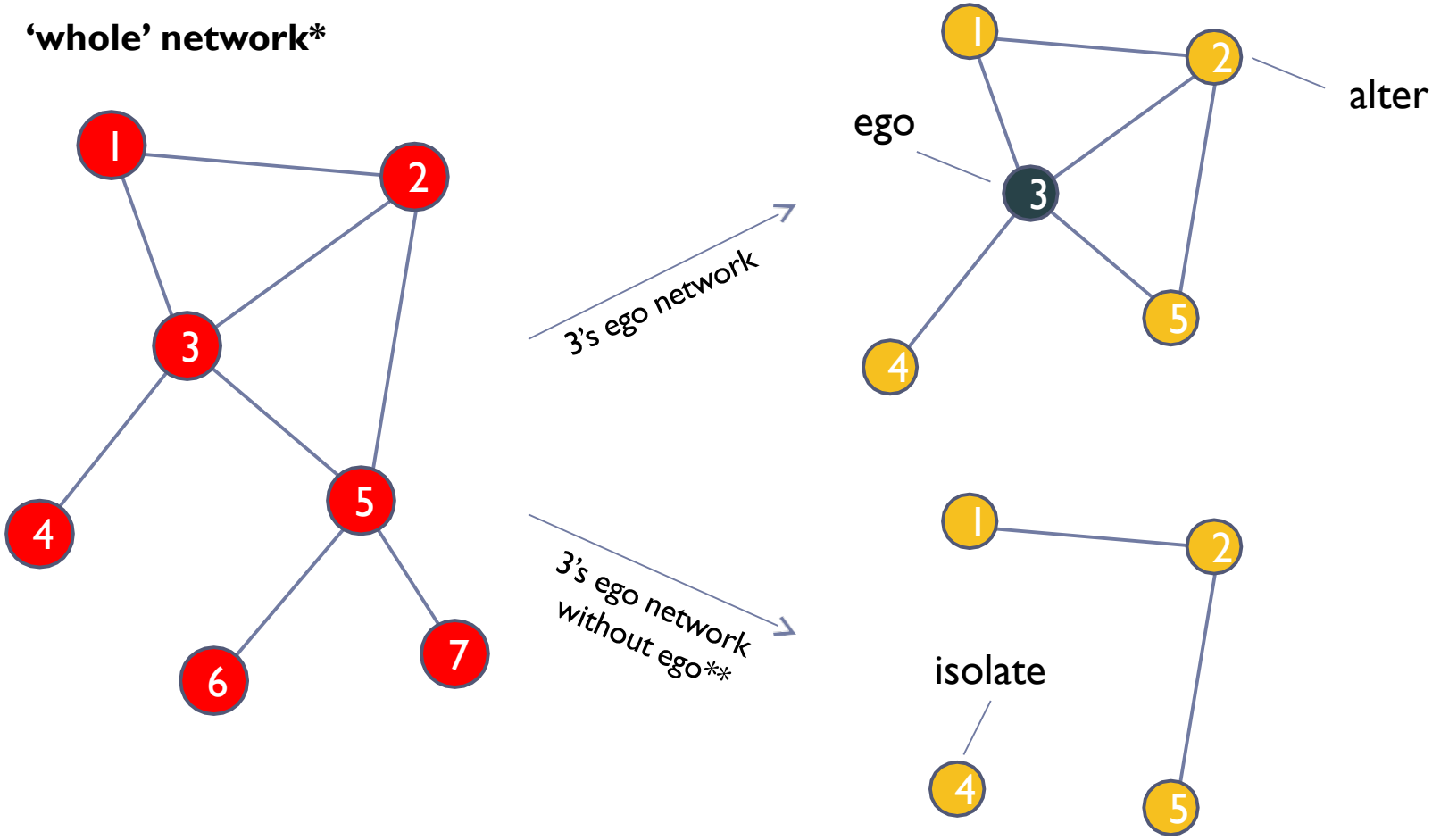
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

But interpretation is different now

Adjacency matrix becomes symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

Ego networks and 'whole' networks

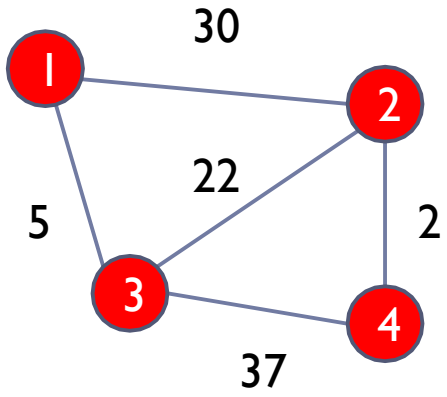


* no studied network is 'whole' in practice; it's usually a partial picture of one's real life networks (*boundary specification problem*)
** ego not needed for analysis as all alters are by definition connected to ego

Basic Concepts

Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
Key Players	How to identify key/central nodes in network
Cohesion	Measures of overall network structure

Adding weights to edges *(directed or undirected)*



Weights could be:

- *Frequency of interaction in period of observation*
- *Number of items exchanged in period*
- *Individual perceptions of strength of relationship*
- *Costs in communication or exchange, e.g. distance*
- *Combinations of these*

Edge list: add column of weights

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
3	4	37

Adjacency matrix: add weights instead of 1

Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

Edge weights as relationship strength

Edges can represent interactions, flows of information or goods, similarities/affiliations, or social relations

Specifically for social relations, a 'proxy' for the strength of a tie can be:

- the *frequency* of interaction (communication) or the amount of flow (exchange)
- *reciprocity* in interaction or flow
- the *type* of interaction or flow between the two parties (e.g., intimate or not)
- other *attributes* of the nodes or ties (e.g., kin relationships)
- The *structure* of the nodes' neighborhood (e.g. many mutual 'friends')

Surveys and interviews allows us to establish the existence of mutual or one-sided strength/affection with greater certainty, but proxies above are also useful

Homophily, transitivity, and bridging

Homophily is the tendency to relate to people with similar characteristics (status, beliefs, etc.)

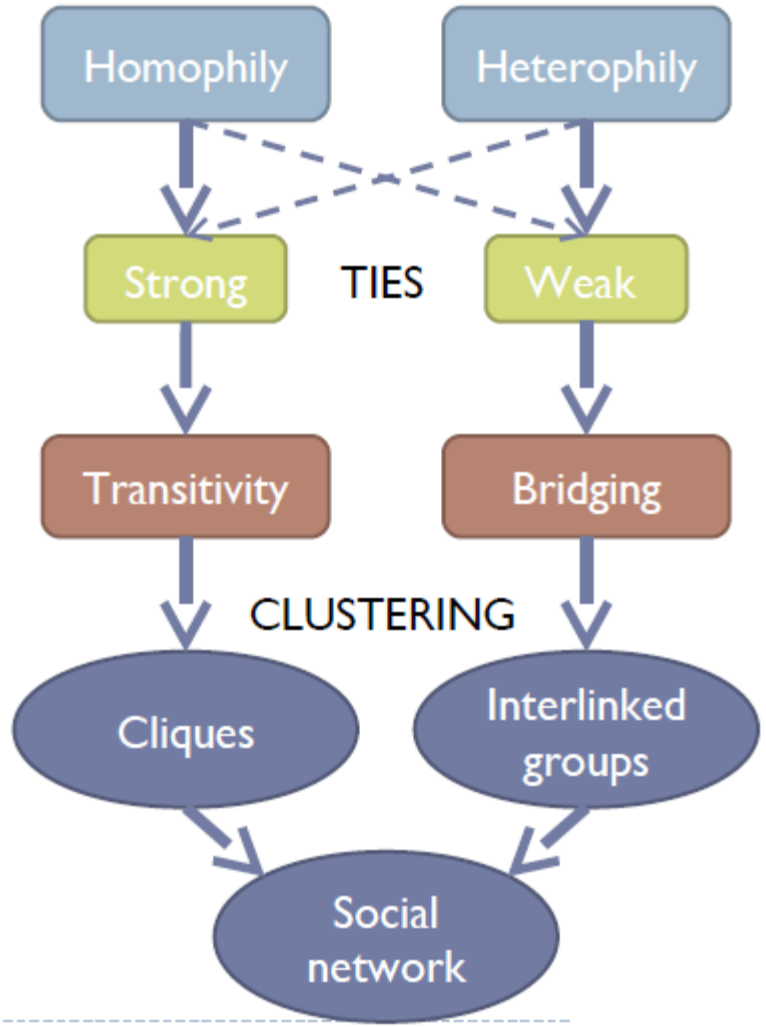
- It leads to the formation of homogeneous groups (*clusters*) where forming relations is easier
- Extreme homogenization can act counter to innovation and idea generation (*heterophily* is thus desirable in some contexts)
- Homophilous ties can be strong or weak

Transitivity in SNA is a property of ties: if there is a tie between A and B and one between B and C, then in a transitive network A and C will also be connected

- Strong ties are more often transitive than weak ties; transitivity is therefore evidence for the existence of strong ties (but not a necessary or sufficient condition)
- Transitivity and homophily together lead to the formation of *cliques* (fully connected clusters)

Bridges are nodes and edges that connect across groups

- Facilitate inter-group communication, increase social cohesion, and help spur innovation
- They are usually weak ties, but not every weak tie is a bridge



Basic Concepts

Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
Key Players	How to identify key/central nodes in network
Cohesion	Measures of overall network structure

Degree centrality

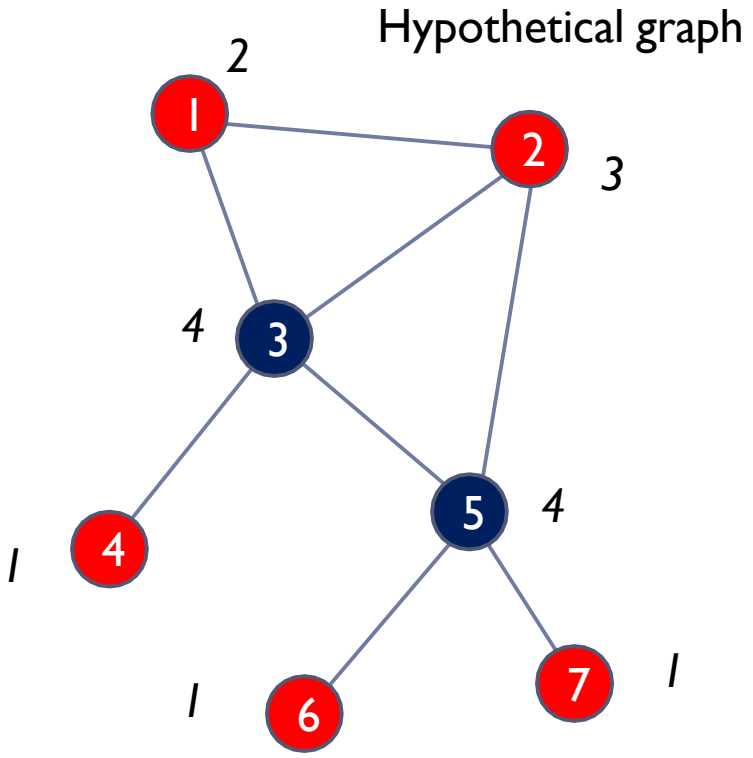
A node's (in-) or (out-)degree is the number of links that lead into or out of the node

In an undirected graph they are of course identical

Often used as measure of a node's degree of connectedness and hence also influence and/or popularity

Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate 'neighborhood'

Nodes 3 and 5 have the highest degree (4)



Values computed with the sna package in the R programming environment. Definitions of centrality measures may vary slightly in other software.

15th Century Florentine Marriages

(Padgett & Ansell, 1993), cf. (Jackson, 2010)

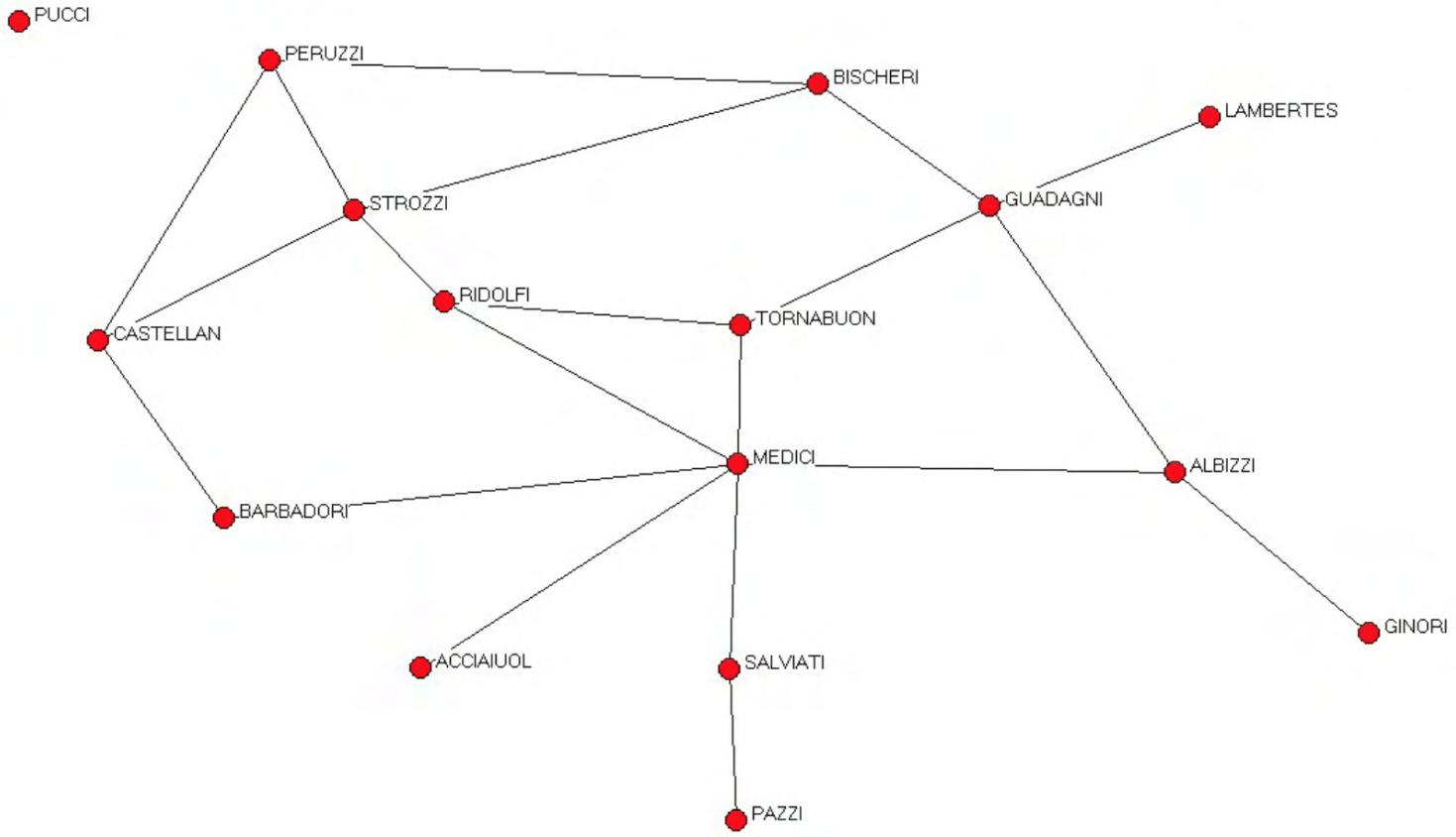


Figure 1.2.1 15th Century Florentine Marriges Data from Padgett and Ansell [491] (drawn using UCINET)

Network of Thrones

(Beveridge and Shan, 2016)

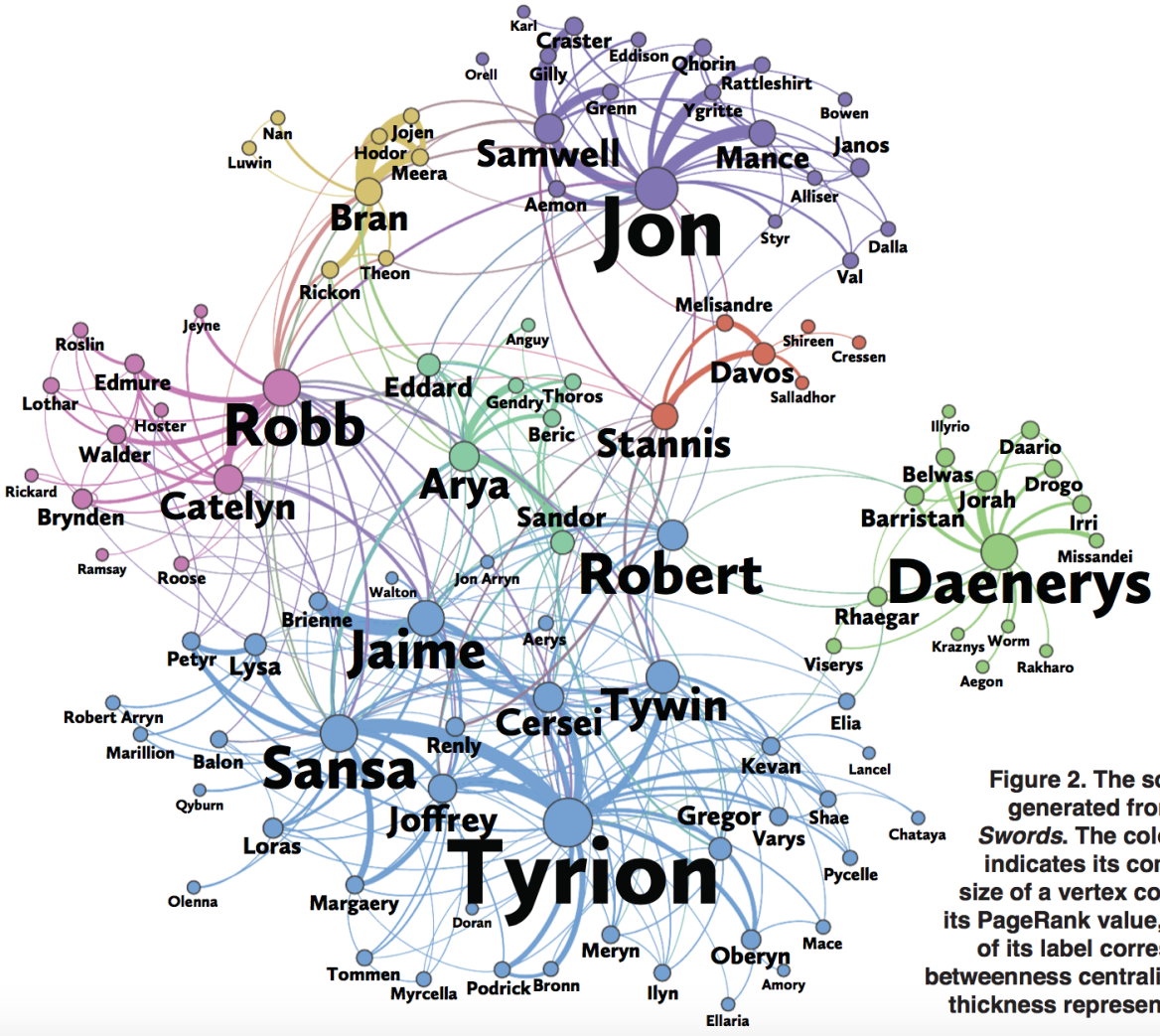


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Paths and shortest paths

Shortest path(s)

A *path* between two nodes is any sequence of non-repeating nodes that connects the two nodes

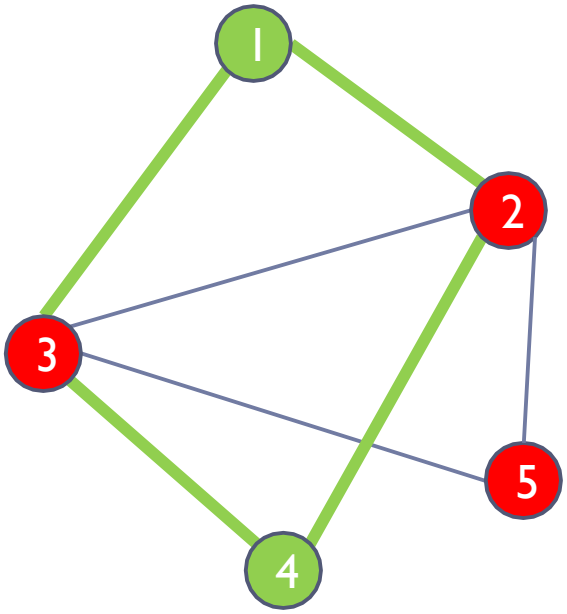
The *shortest path* between two nodes is the path that connects the two nodes with the shortest number of edges (also called the *distance* between the nodes)

In the example to the right, between nodes 1 and 4 there are two shortest paths of length 2: {1,2,4} and {1,3,4}

Other, longer paths between the two nodes are {1,2,3,4}, {1,3,2,4}, {1,2,5,3,4} and {1,3,5,2,4} (the longest paths)

Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

Hypothetical graph



Betweenness centrality

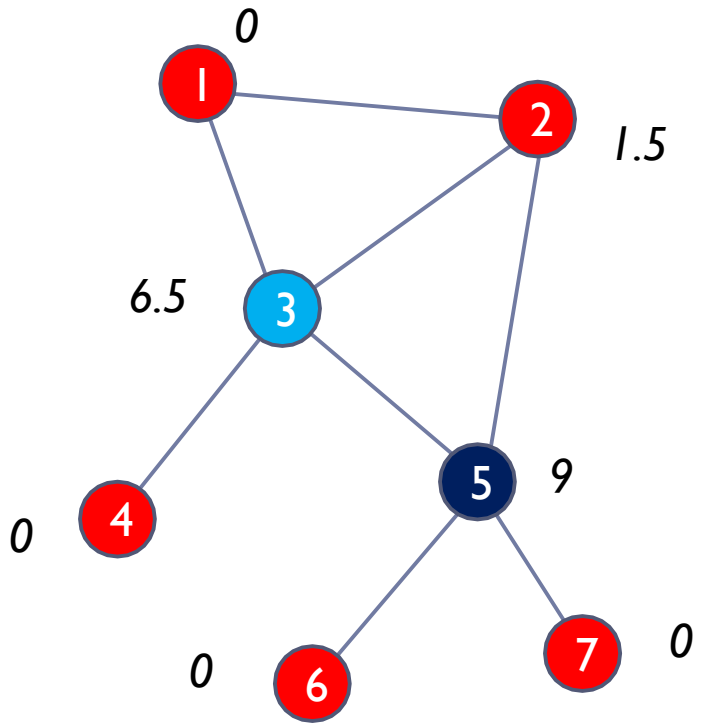
For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j

Sum the above values for all node pairs i,j

Sometimes normalized such that the highest value is 1 or that the sum of all betweenness centralities in the network is 1

Shows which nodes are more likely to be in communication paths between other nodes

Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)



Node 5 has higher betweenness centrality than 3

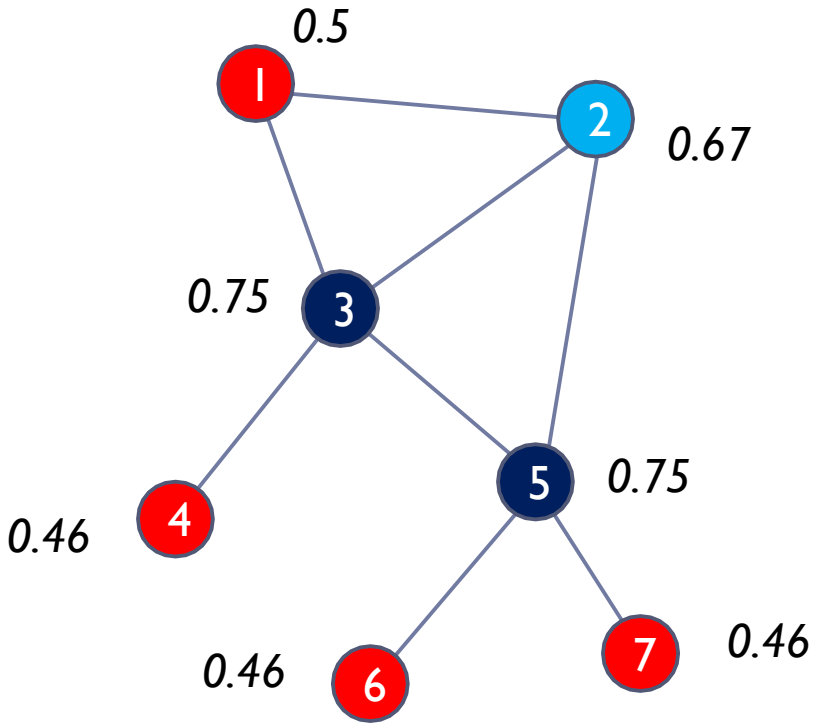
Values computed with the sna package in the R programming environment. Definitions of centrality measures may vary slightly in other software.

Closeness centrality

Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)

Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality

It is a measure of *reach*, i.e. the speed with which information can reach other nodes from a given starting node



Nodes 3 and 5 have the highest (i.e. best) closeness, while node 2 fares almost as well

Note: Sometimes closeness is calculated without taking the reciprocal of the mean shortest path length. Then lower values are 'better'.

Values computed with the sna package in the R programming environment. Definitions of centrality measures may vary slightly in other software.

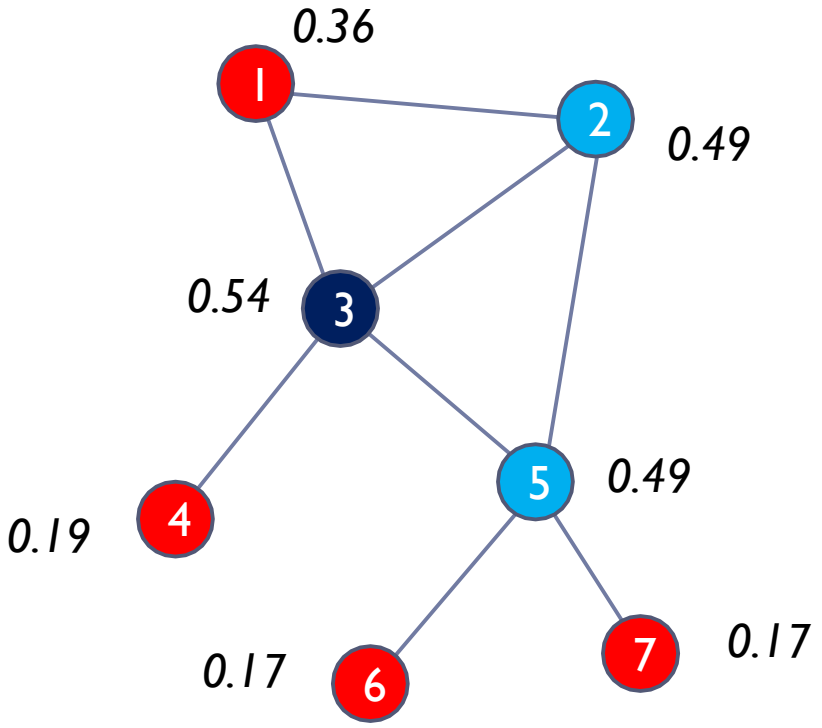
Eigenvector centrality

A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it

In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality

This is similar to how Google ranks web pages: links from highly linked-to pages count more

Useful in determining who is connected to the most connected nodes



Node 3 has the highest eigenvector centrality, closely followed by 2 and 5

Values computed with the sna package in the R programming environment. Definitions of centrality measures may vary slightly in other software.

Eigenvector Centrality

Idea: You are important if you have important friends.

The importance c_i of node i is proportional to the importance of its friends. I.e. for some non-zero constant $\lambda \in \mathbb{R}$,

$$\lambda \cdot c_i \quad = \quad \sum_{j \in \mathrm{Nbh}(i)} c_j \quad = \quad \sum_{j \in N} M_{ij} c_j \qquad (*)$$

where M is the adjacency matrix associated with G.
Let c be the vector of c_i's. The equation $(*)$ says

$$\lambda \cdot \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} M_{11} \cdots M_{1n} \\ \vdots \quad \vdots \\ M_{n1} \cdots M_{nn} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

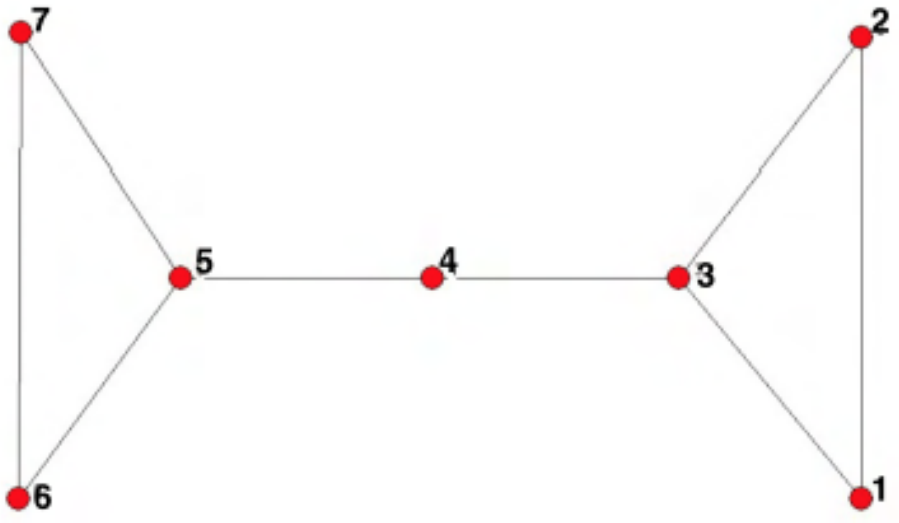
That is, c is a (right-side) eigenvector of M:

$$Mc = \lambda c \qquad \text{for some } \lambda \in \mathbb{R}$$

Eigenvector centrality

- An $n \times n$ matrix can have n different eigenvalues.
  By the Perron-Frobenius Theorem, a maximal eigenvalue always exists (if G is connected), it is positive, and all entries of the corresponding eigenvector are positive.
- For fixed eigenvalue, there can be many eigenvectors: If $Mc = \lambda c$ then for any $r \in \mathrm{R}$, $M(rc) = \lambda rc$ (i.e. if c is an eigenvector, then so is any scalar multiple of c.)
- The eigenvector c only gives a relative centrality score C-eigen(i) = c_i to each node i .
- Intuition: C-eigen(i) is the relative frequency with which i is visited during a random walk in the graph.
- Remark: Google PageRank is a similar to eigenvector centrality.

Example: Overview



nodes	1, 2, 6, 7	3, 5	4
C-deg(i)	0.33	0.50	0.33
C-clo(i)	0.40	0.55	0.60
C-betw(i)	0.00	0.53	0.60
C-eigen(i)	0.47	0.63	0.54

Network of Thrones

(Beveridge and Shan, 2016)

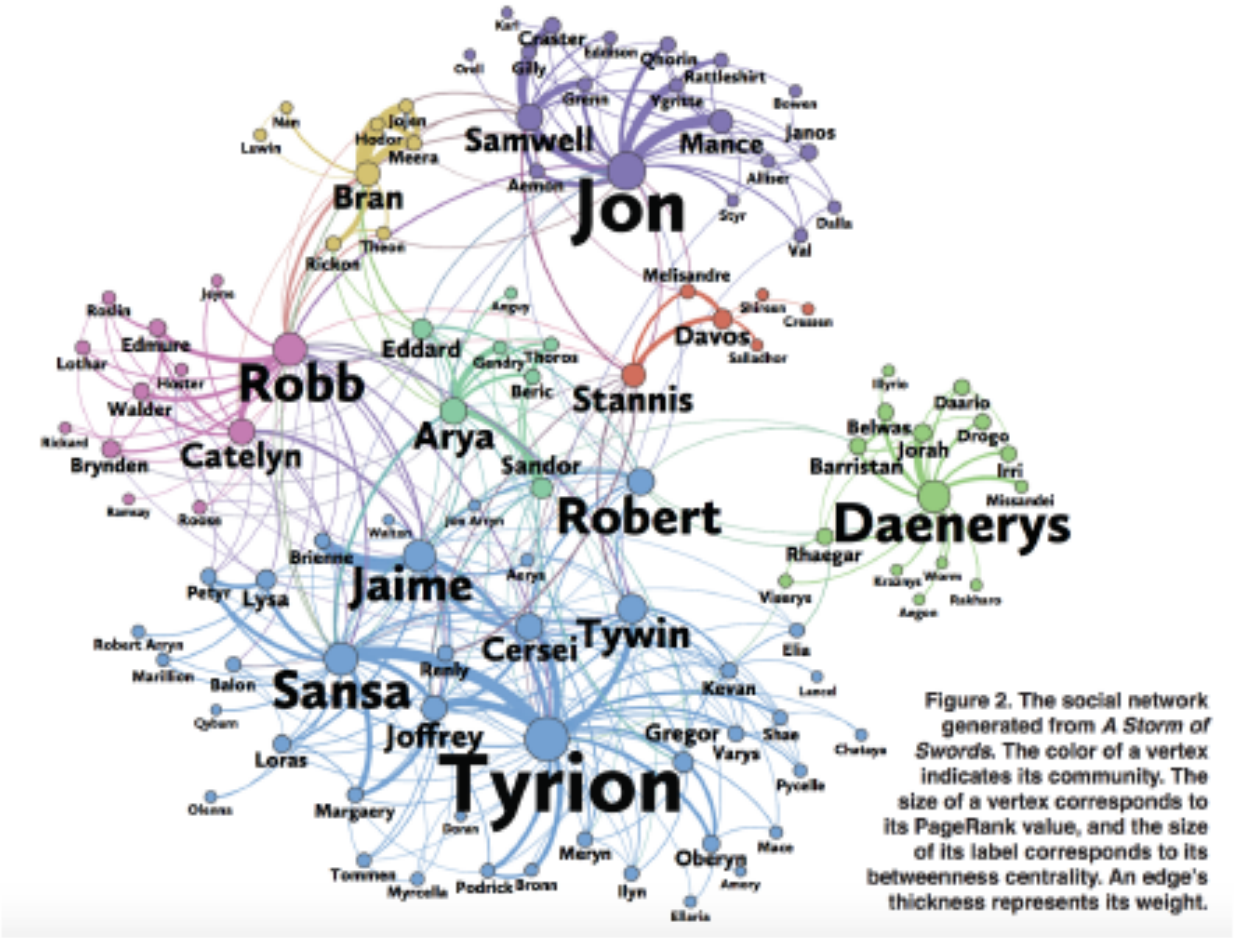


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

Network of Thrones: Centrality Measures

(Beveridge and Shan, 2016)

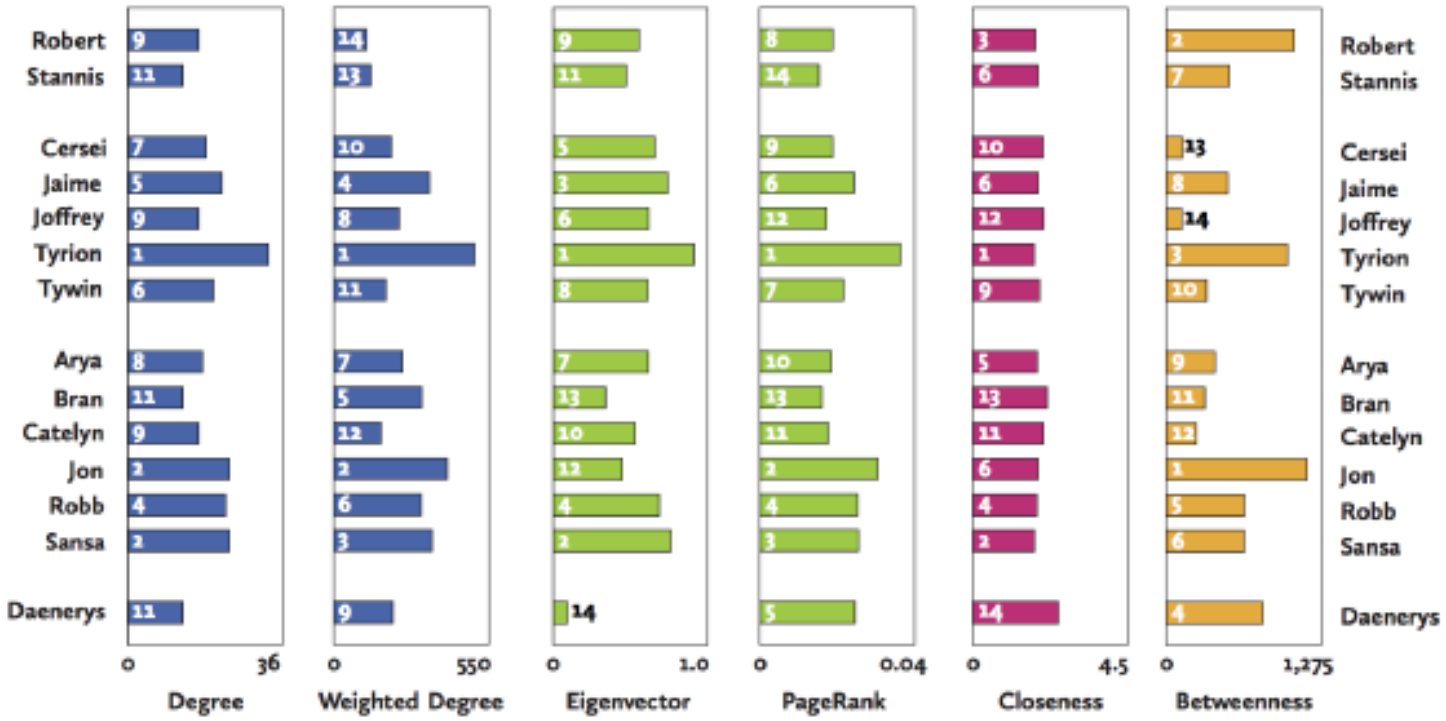


Figure 3. Centrality measures for the network. Larger values correspond to greater importance, except for closeness centrality, where smaller values are better. Numbers in the bars give the rankings of these characters.

Interpretation of measures (1)

Centrality measure	**Interpretation in social networks**
Degree	How many people can this person reach directly?
Betweenness	How likely is this person to be the most direct route between two people in the network?
Closeness	How fast can this person reach everyone in the network?
Eigenvector	How well is this person connected to other well-connected people?

Interpretation of measures (2)

Centrality measure	**Other possible interpretations…**
Degree	In network of music collaborations: how many people has this person collaborated with?
Betweenness	In network of spies: who is the spy though whom most of the confidential information is likely to flow?
Closeness	In network of sexual relations: how fast will an STD spread from this person to the rest of the network?
Eigenvector	In network of paper citations: who is the author that is most cited by other well-cited authors?

Centrality Measures Conclusion

- There is no single "right" centrality measure.
- Each centrality measure gives a different perspective.
- Edge centrality measures exist (e.g. edge betweenness)
- "Enhanced metrics" exist for graphs with more "features" (e.g. directed, weighted edges)

Identifying sets of key players

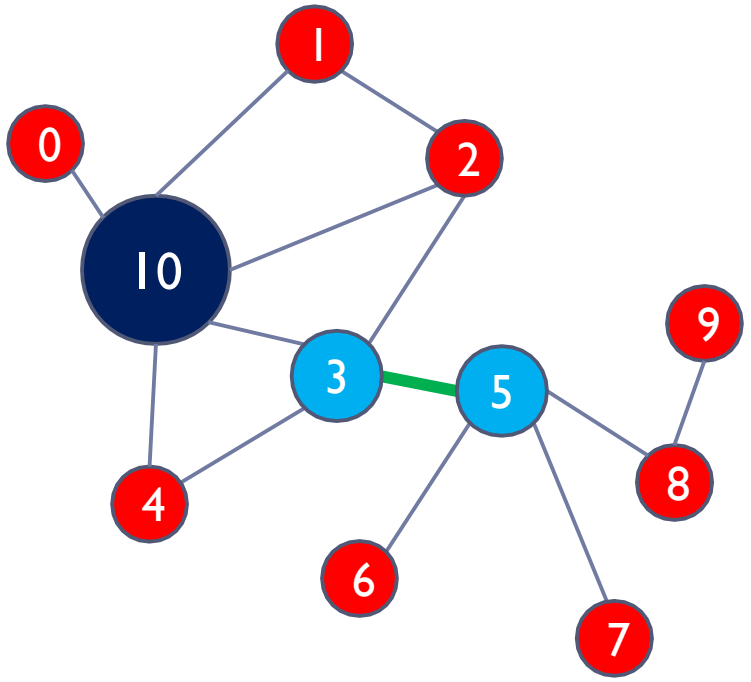
In the network to the right, node 10 is the most central according to degree centrality

But nodes 3 and 5 together will reach more nodes

Moreover the tie between them is critical; if severed, the network will break into two isolated sub-networks

It follows that other things being equal, players 3 and 5 together are more 'key' to this network than 10

Thinking about sets of key players is helpful!



Basic Concepts

Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
Key Players	How to identify key/central nodes in network
Cohesion	How to characterize a network's structure

Reciprocity (degree of)

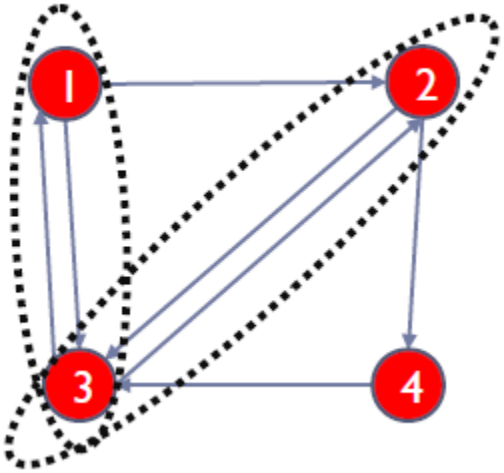
The ratio of the number of relations which are reciprocated (i.e. there is an edge in both directions) over the total number of relations in the network

…where two vertices are said to be related if there is at least one edge between them

In the example to the right this would be 2/5=0.4 (whether this is considered high or low depends on the context)

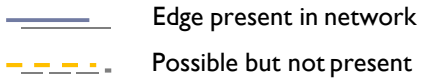
A useful indicator of the degree of mutuality and reciprocal exchange in a network, which relate to social cohesion

Only makes sense in directed graphs



Reciprocity for network = 0.4

Density



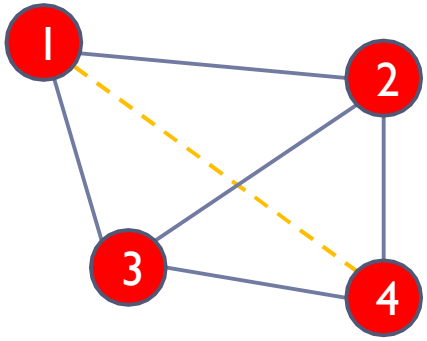
A network's *density* is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is $n(n-1)/2$, where n is the number of vertices, for an undirected graph)

In the example network to the right density=5/6=0.83 (i.e. it is a fairly *dense* network; opposite would be a *sparse* network)

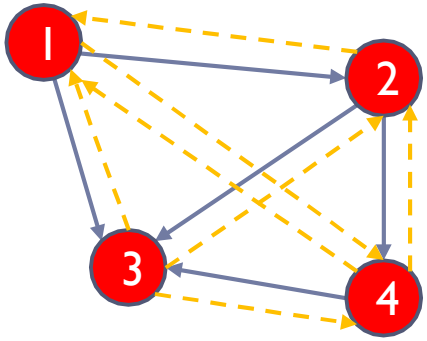
It is a common measure of how well connected a network is (in other words, how closely knit it is) – a perfectly connected network is called a *clique* and has density=1

A directed graph will have half the density of its undirected equivalent, because there are twice as many possible edges, i.e. $n(n-1)$

Density is useful in comparing networks against each other, or in doing the same for different regions within a single network



density = 5/6 = 0.83



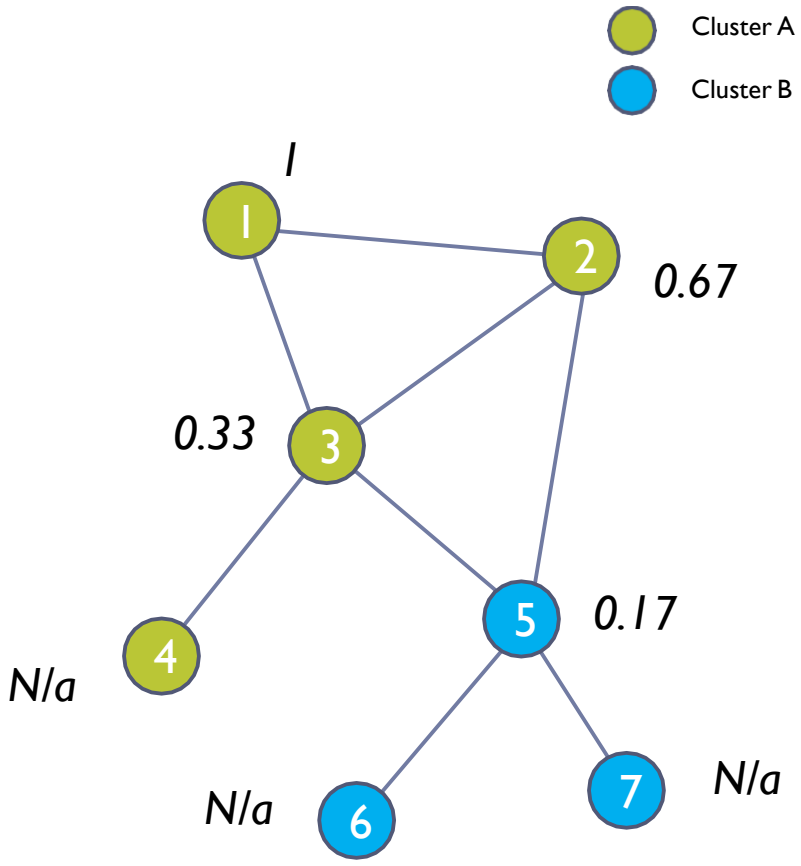
density = 5/12 = 0.42

Clustering

A node's *clustering coefficient* is the number of closed triplets in the node's neighborhood over the total number of triplets in the neighborhood. It is also known as *transitivity*.

E.g., node 1 to the right has a value of 1 because it is only connected to 2 and 3, and these nodes are also connected to one another (i.e. the only triplet in the neighborhood of 1 is closed). We say that nodes 1, 2, and 3 form a *clique*.

Clustering algorithms identify clusters or 'communities' within networks based on network structure and specific clustering criteria (example shown to the right with two clusters is based on *edge betweenness*, an equivalent for edges of the betweenness centrality presented earlier for nodes)



Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

Values computed with the igraph package in the R programming environment. Definitions of centrality measures may vary slightly in other software.

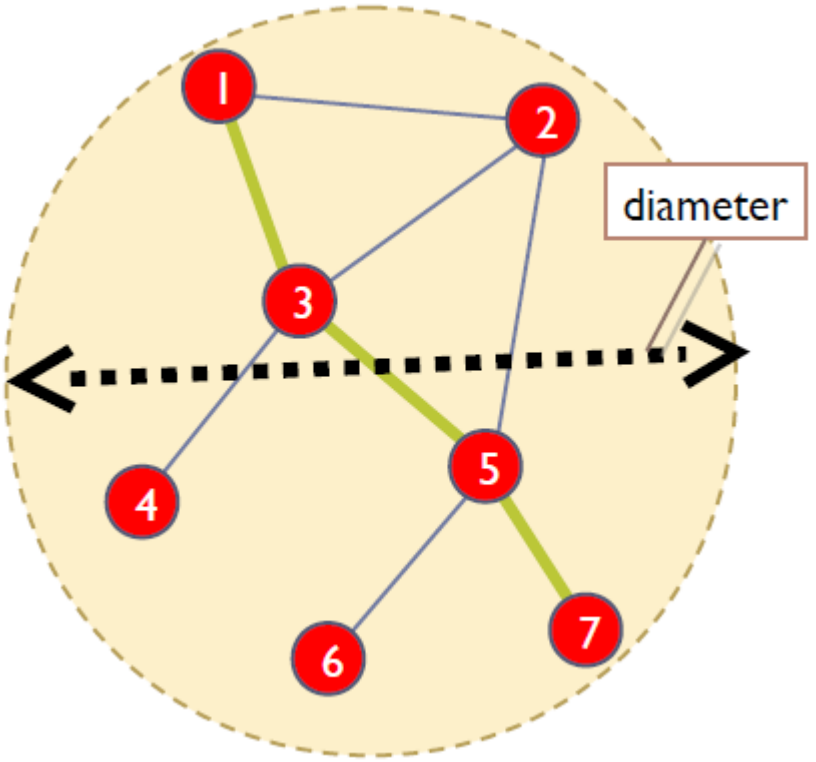
Average and longest distance

The longest shortest path (distance) between any two nodes in a network is called the network's *diameter*

The diameter of the network on the right is 3; it is a useful measure of the *reach* of the network (as opposed to looking only at the total number of vertices or edges)

It also indicates how long it will take at most to reach any node in the network (sparser networks will generally have greater diameters)

The average of all shortest paths in a network is also interesting because it indicates how far apart any two nodes will be on average (average *distance*)

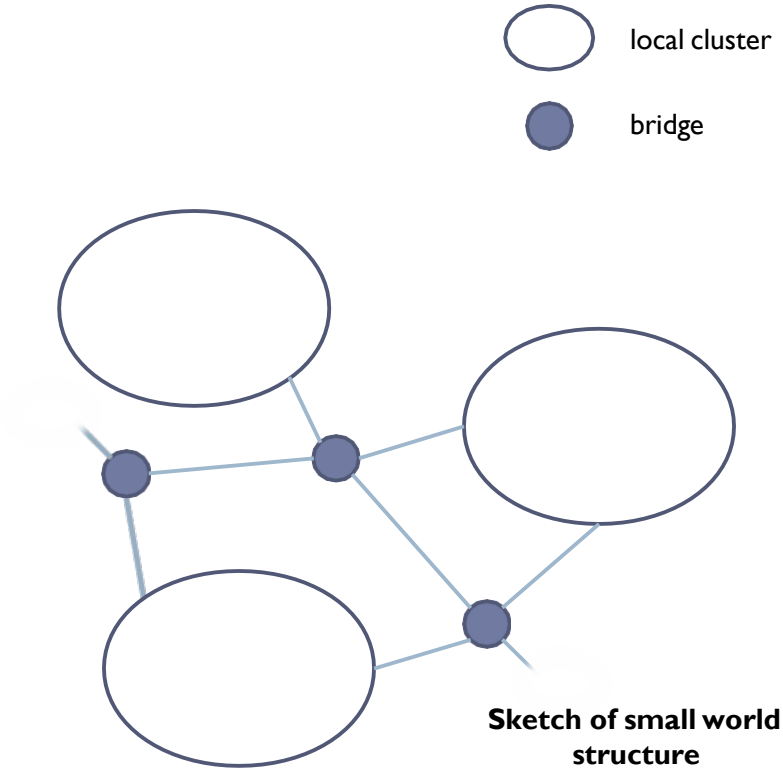


Small Worlds

A small world is a network that looks almost random but exhibits a significantly *high clustering coefficient* (nodes tend to cluster locally) and a relatively *short average path length* (nodes can be reached in a few steps)

It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters (see also next page…)

Such a network will have many clusters but also many bridges between clusters that help shorten the average distance between nodes



Sketch of small world structure

You may have heard of the famous "6 degrees" of separation

Preferential Attachment

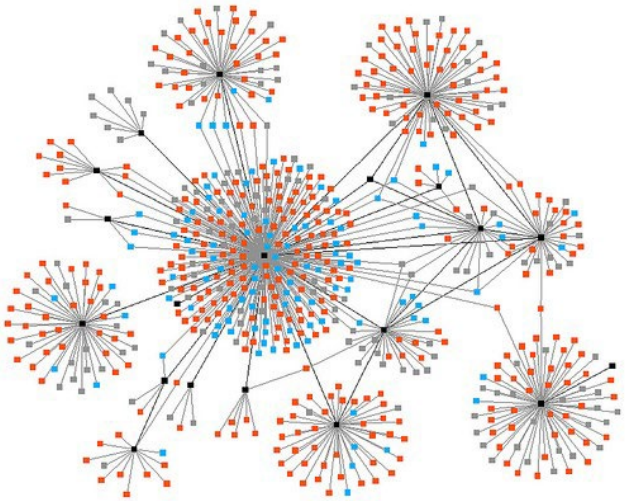
A property of some networks, where, during their evolution and growth in time, a the great majority of new edges are to nodes with an already high degree; the degree of these nodes thus increases disproportionately, compared to most other nodes in the network

The result is a network with few very highly connected nodes and many nodes with a low degree

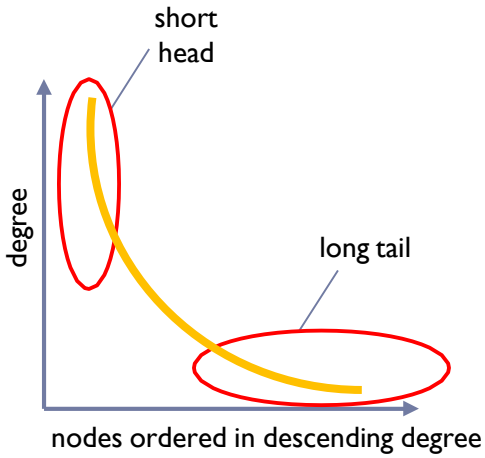
Such networks are said to exhibit a *long-tailed* degree distribution

And they tend to have a small-world structure!

(so, as it turns out, transitivity and strong/weak tie characteristics are not necessary to explain small world structures, but they are common and can also lead to such structures)



Example of network with preferential attachment



Sketch of long-tailed degree distribution

Reasons for preferential attachment

Popularity

We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective, measurable characteristics

Also known as 'the rich get richer'

Quality

We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention, faster

Also known as 'the good get better'

Mixed model

Among nodes of similar attributes, those that reach critical mass first will become 'stars' with many friends and followers ('halo effect')

May be impossible to predict who will become a star, even if quality matters

Core-Periphery Structures

A useful and relatively simple metric of the degree to which a social network is centralized or decentralized, is the *centralization* measure

(usually normalized such that it takes values between 0 and 1)

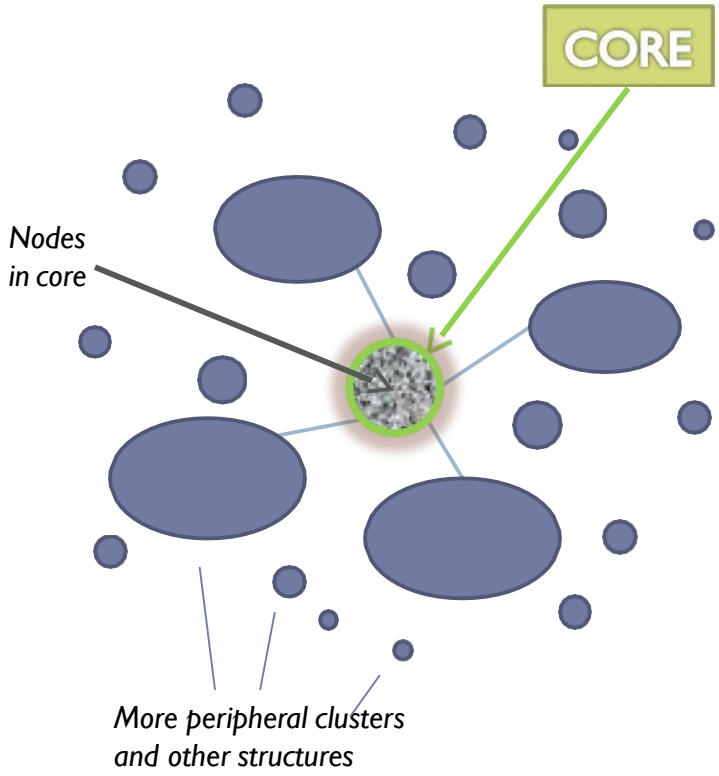
It is based on calculating the differences in degrees between nodes; a network that greatly depends on 1-2 highly connected nodes (as a result for example of preferential attachment) will exhibit greater differences in degree centrality between nodes

Centralized structures can perform better at some tasks (like team-based problem-solving requiring coordination), but are more prone to failure if key players disconnect

In addition to centralization, many large groups and online communities have a *core* of densely connected users that are critical for connecting a much larger periphery

Cores can be identified visually, or by examining the location of high-degree nodes and their joint degree distributions (do high-degree nodes tend to connect to other high-degree nodes?)

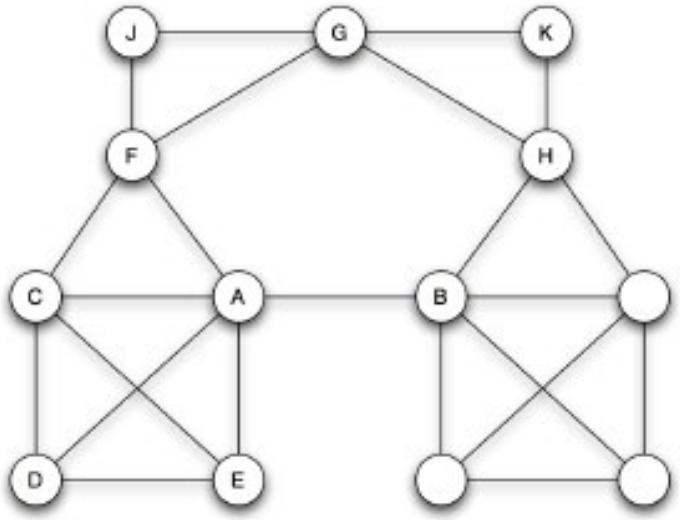
Bow-tie analysis, famously used to analyze the structure of the Web, can also be used to distinguish between the core and other, more peripheral elements in a network (see earlier example here)



Local Bridges

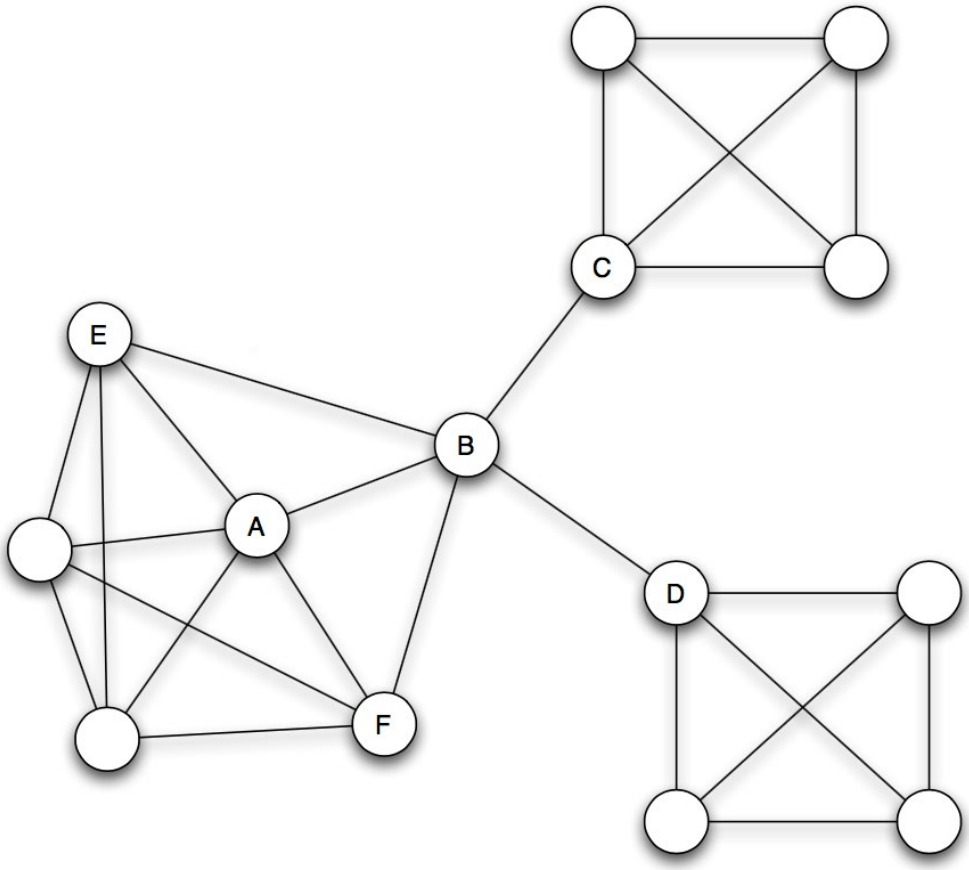
- **Definition.** An edge (i, j) is a local bridge if i and j have no friends in common.

  Example: (A,B) is a local bridge in this graph:



- Why are local bridges important?

Embeddedness



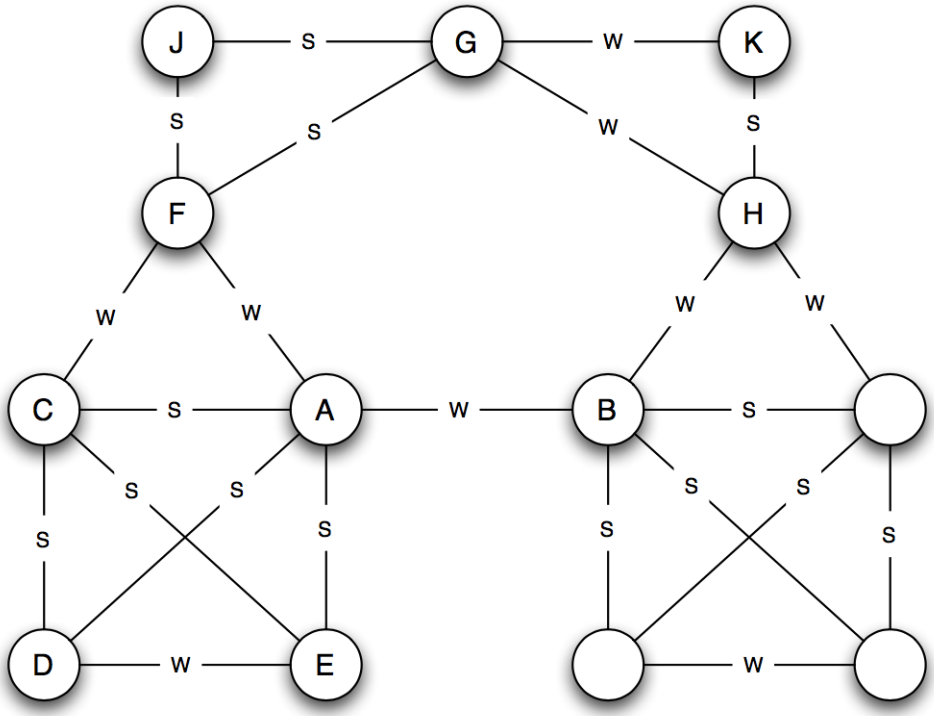
Definition. The embeddedness of an edge (i, j) is the number of common neighbours of i and j.

Example. Embeddedness of (A, F) is 3. Embeddedness of (B, D) is 0.

Higher embeddedness leads to more trust (due to potential social sanctions).

Strong and Weak Ties

- Assume we have labelled the edges as being 'strong' or 'weak' ties (e.g. based on data about frequency and duration of phone calls, reciprocity etc.)



Definition. A labelled graph satisfies the Strong Triadic Closure Property (STCP) if whenever a node *i* has strong ties to *j* and *k* (with $j \neq k$), then there is a (weak or strong) tie between *j* and *k*.

The Strength of Weak Ties

(Granovetter, 1973):
Novel/Useful information often comes from an acquaintance rather than from a close friend.

Why is STCP a useful concept?

- It links the local property of strong vs weak ties with the global property of being a local bridge.
- It is a mathematical definition that can be checked in empirical networks.
- It provides a concrete basis for studying the relationship between network properties and "social events" (e.g. finding a new job, life transitions)

Thoughts on Design

How can an online social media platform (and its administrators) leverage the methods and insights of social network analysis?

How can it encourage a network perspective among its users, such that they are aware of their 'neighborhood' and can learn how to work with it and/or expand it?

What measures can an online community take to optimize its network structure?

Example: cliques can be undesirable because they shun newcomers

What would be desirable structures for different types of online platforms? (*not* easy to answer)

How can online communities identify and utilize key players for the benefit of the community?

SNA inspired some of the first SNS's (e.g. SixDegrees), but still not used so often in conjunction with design decisions – much untapped potential here