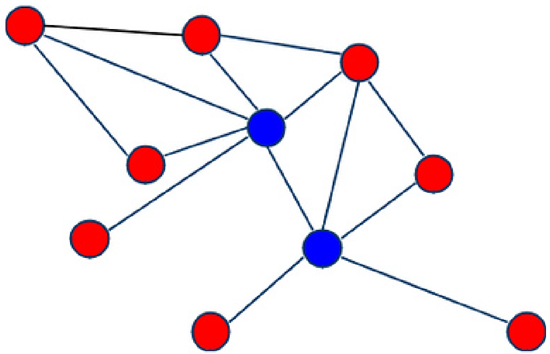


Social Network Analysis (SNA)

including a tutorial on concepts and methods

Social Media – Dr. Giorgos Cheliotis (gcheliotis@nus.edu.sg)
Communications and New Media, National University of Singapore

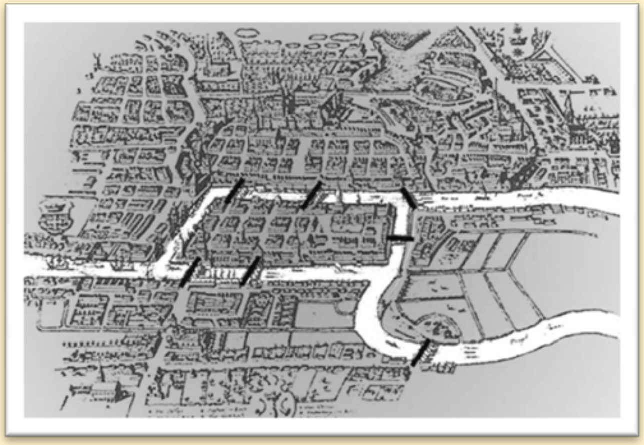
Background: Network Analysis

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

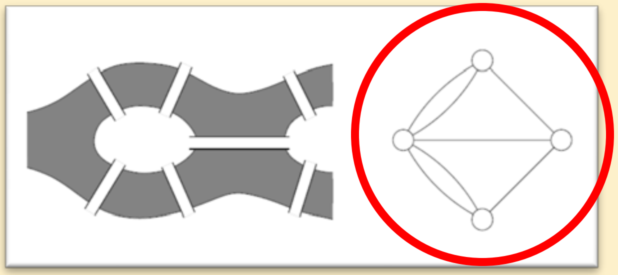
Network analysis concerns itself with the formulation and solution of problems that have a network structure; such structure is usually captured in a *graph* (see the circled structure to the right)

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



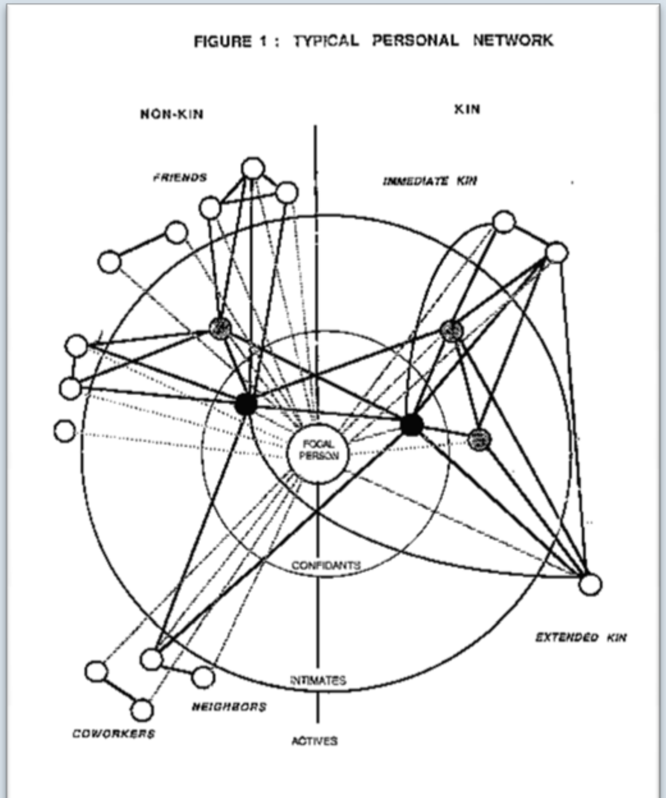
Newman et al, 2006



Newman et al, 2006

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

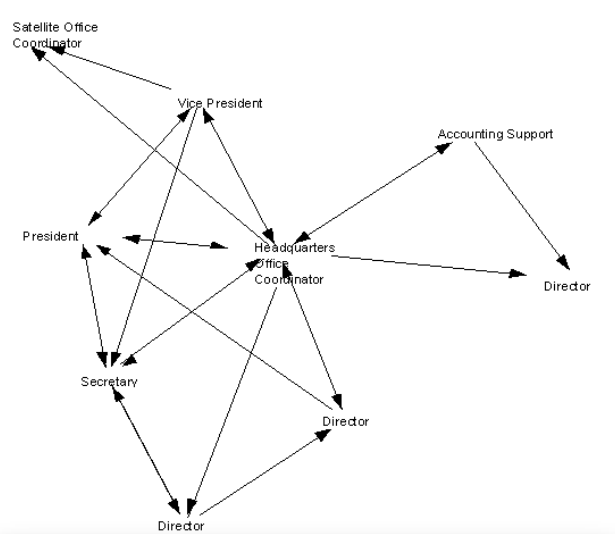
This is an early depiction of what we call an 'ego' network, i.e. a personal network. The graphic depicts varying tie strengths via concentric circles (Wellman, 1998)

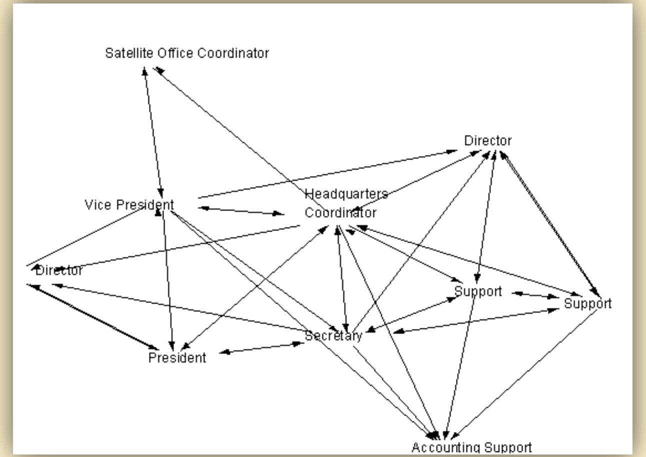
Studying society from a network perspective is to study individuals as embedded in a network of relations and seek explanations for social behavior in the structure of these networks rather than in the individuals alone. This 'network perspective' becomes increasingly relevant in a society that Manuel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

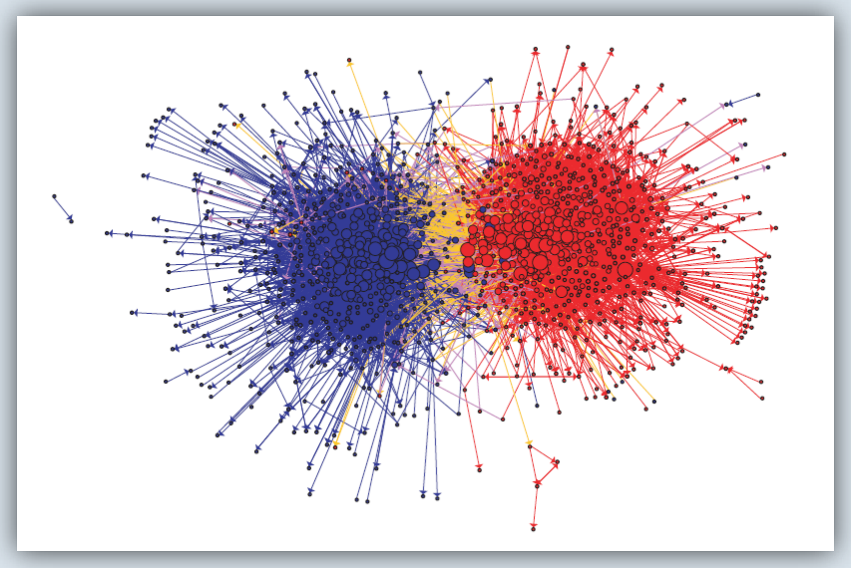
More examples from social science





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)



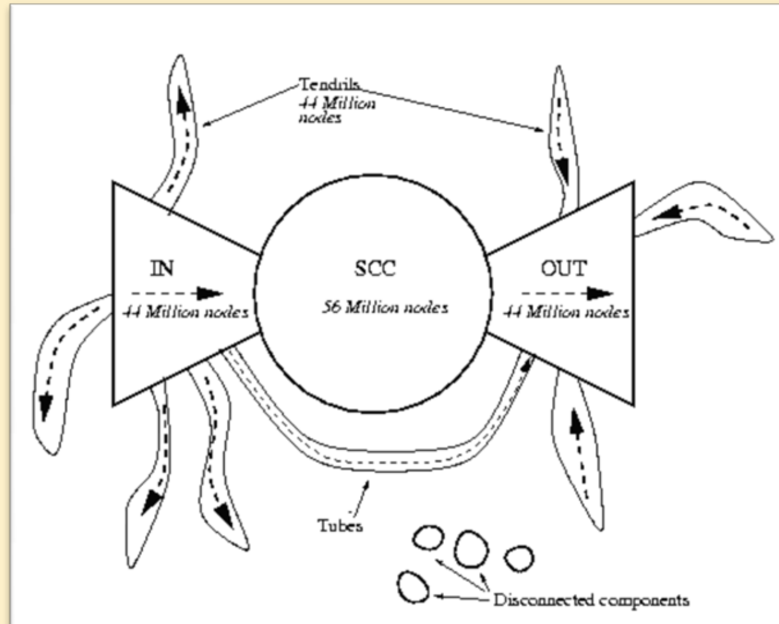
Background: Other Domains

(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

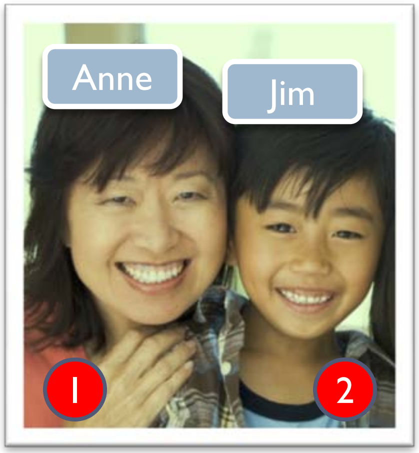
How to represent various social networks

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

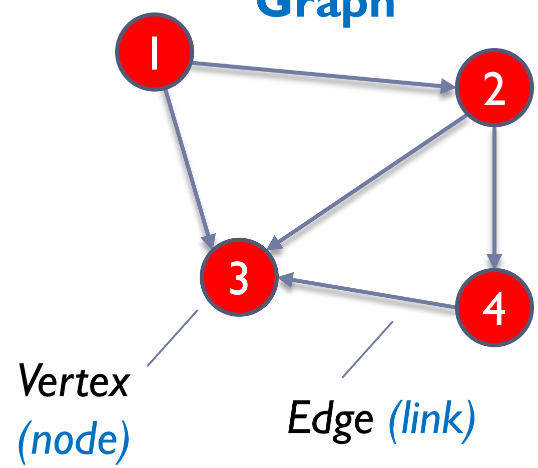


Can we study their interactions as a network?

Communication

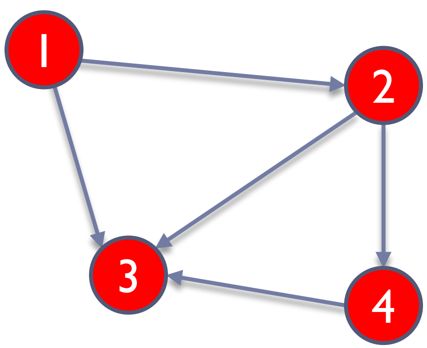
Anne: Jim, tell the Murrays they're invited
Jim: Mary, you and your dad should come for dinner!
Jim: Mr. Murray, you should both come for dinner
Anne: Mary, did Jim tell you about the dinner? You must come.
Mary: Dad, we are invited for dinner tonight
John: *(to Anne)* Ok, we're going, it's settled!

Graph



Entering data on a directed graph

Graph (directed)



Edge list

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

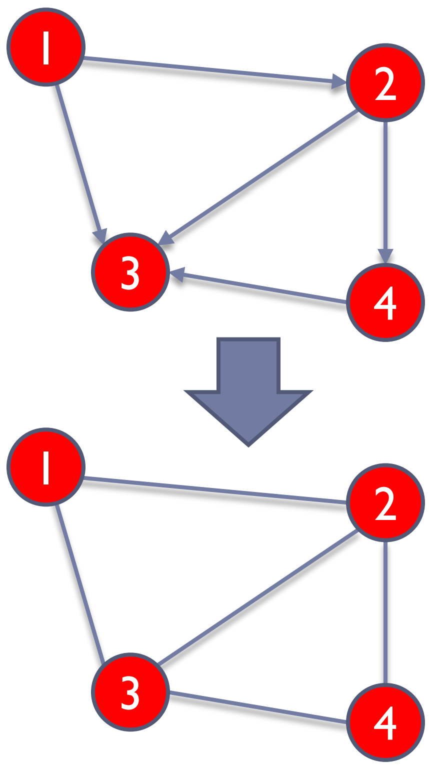
Adjacency matrix

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

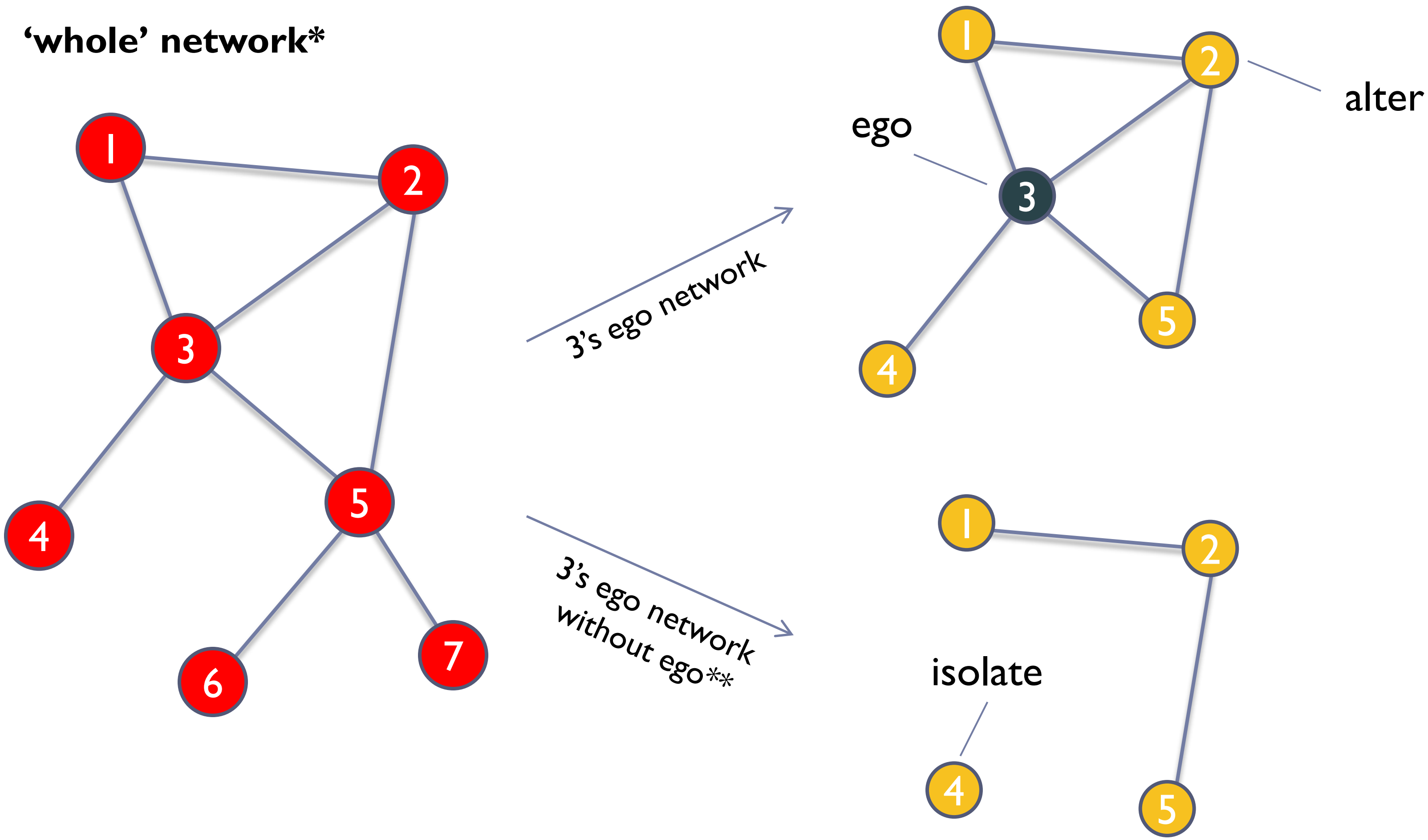
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

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

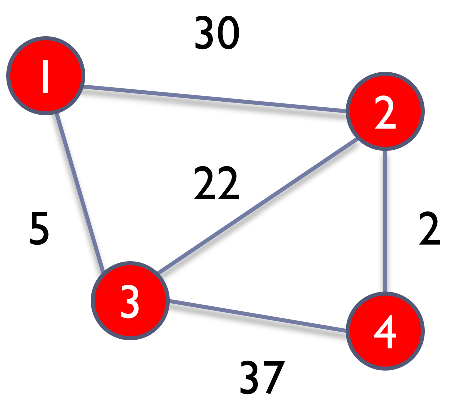


* no studied network is 'whole' in practice; it's usually a partial picture of one's real life networks (*boundary specification problem*)
** ego not needed for analysis as all alters are by definition connected to ego

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)*



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	-

Weights could be:

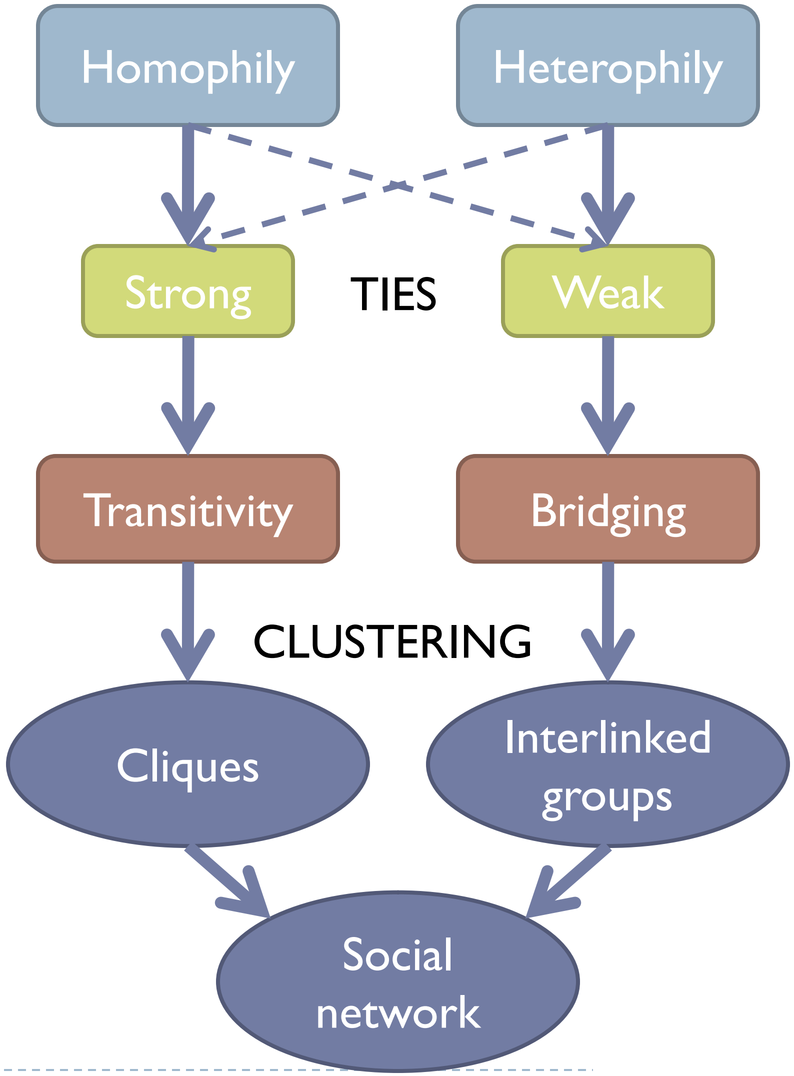
- *Frequency of interaction in period of observation*
- *Number of items exchanged in period*
- *Individual perceptions of strength of relationship*
- *Costs in communication or exchange, e.g. distance*
- *Combinations of these*

Edge 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:
  (a) the *frequency* of interaction (communication) or the amount of flow (exchange)
  (b) *reciprocity* in interaction or flow
  (c) the *type* of interaction or flow between the two parties (e.g., intimate or not)
  (d) other *attributes* of the nodes or ties (e.g., kin relationships)
  (e) 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

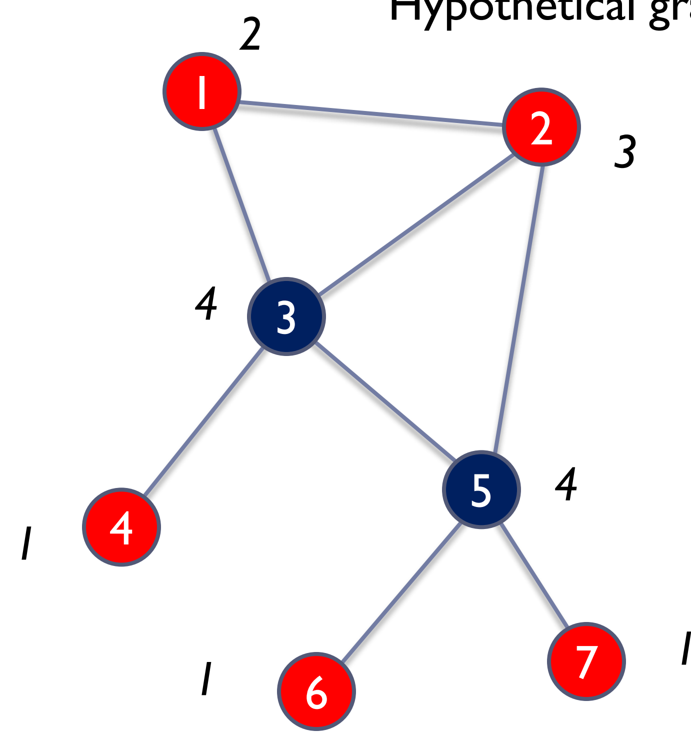
Networks	How to represent various social networks
Tie Strength	How to identify strong/weak ties in the network
Key Players	How to identify key/central nodes in network
Cohesion	Measures of overall network structure

Degree centrality

NodeXL output values

- 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'

Hypothetical graph

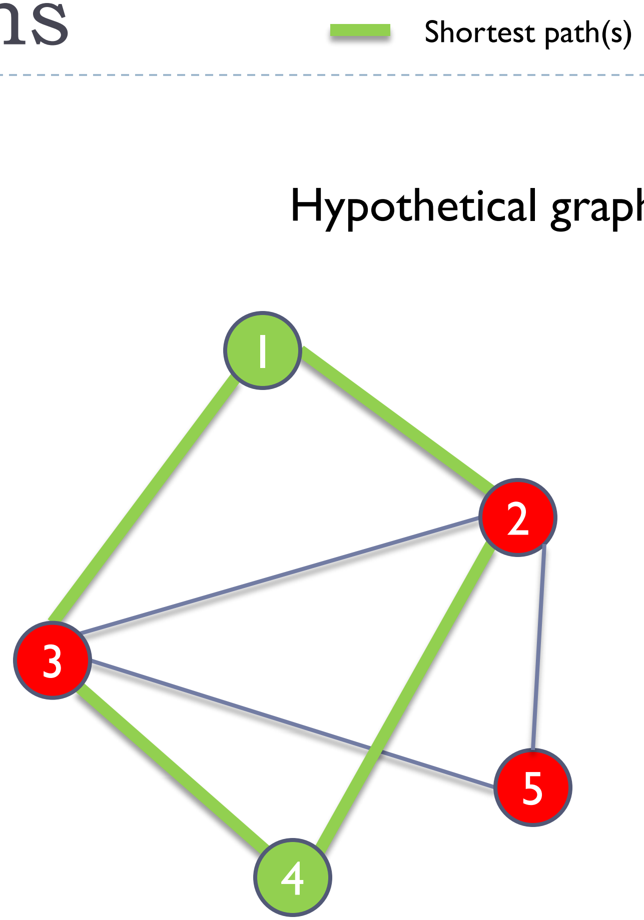


Nodes 3 and 5 have the highest degree (4)

Paths and shortest paths

- 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)

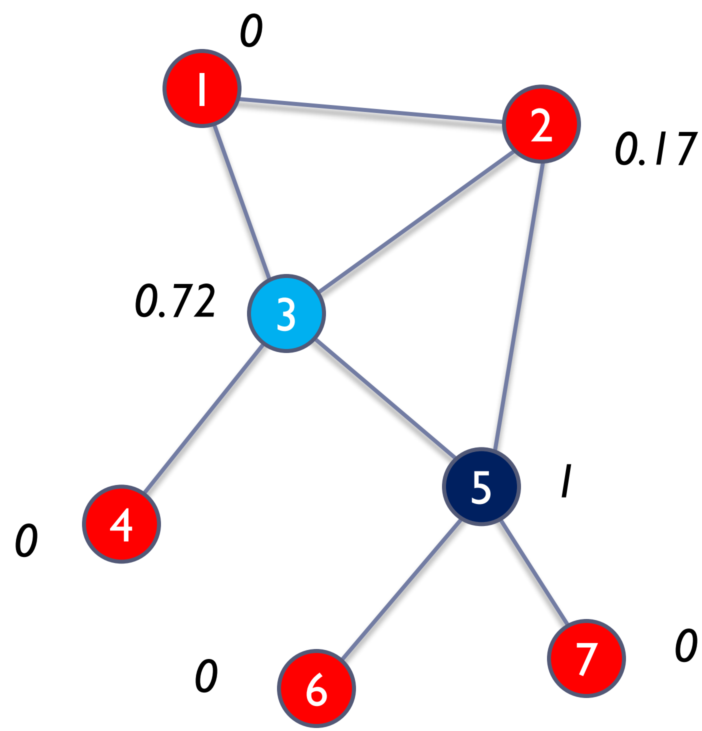
Hypothetical graph



Betweeness centrality

NodeXL output values

- The number of shortest paths that pass through a node divided by all shortest paths in the network
- Sometimes normalized such that the highest value 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)

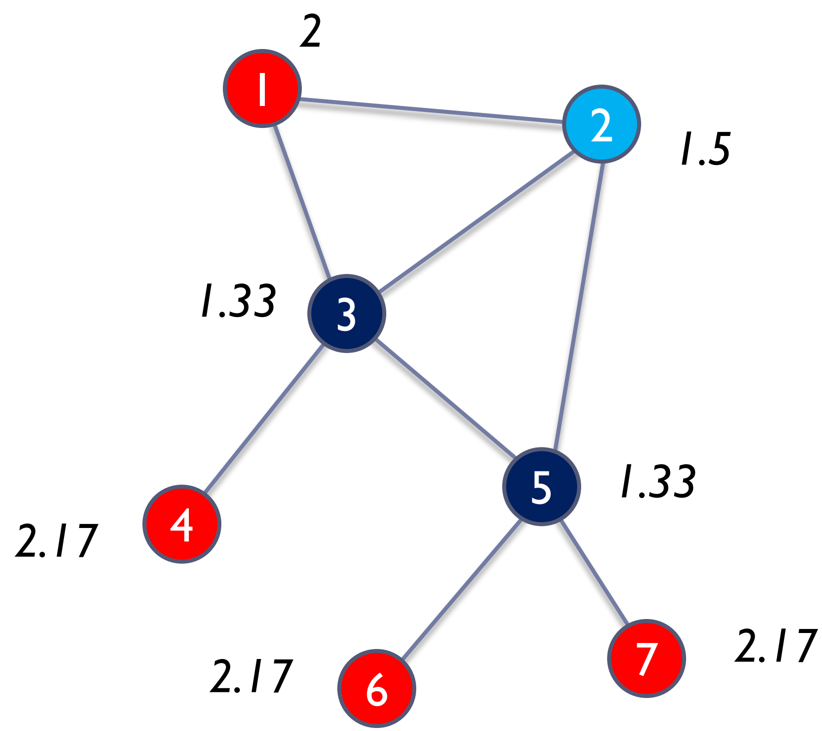


Node 5 has higher betweenness centrality than 3

Closeness centrality

NodeXL output values

- 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)
- It is a measure of *reach*, i.e. how long it will take to reach other nodes from a given starting node
- Useful in cases where speed of information dissemination is main concern
- Lower values are better when higher speed is desirable



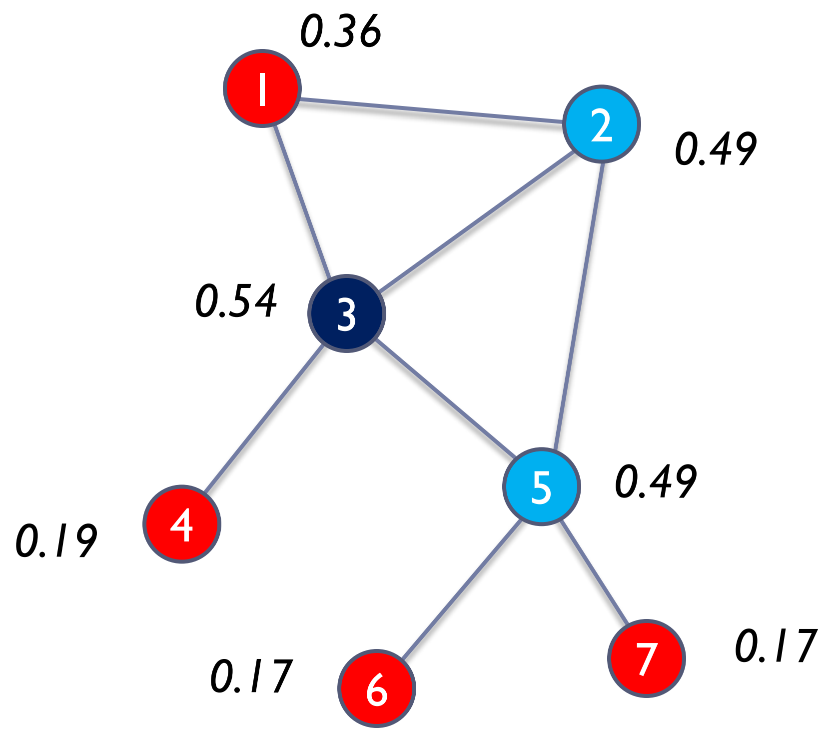
Nodes 3 and 5 have the lowest (i.e. best) closeness, while node 2 fares almost as well

Note: Sometimes closeness is defined as the reciprocal of this value, i.e. 1/x, such that higher values would indicate faster reach

Eigenvector centrality

NodeXL output values

- 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

Note: The term 'eigenvector' comes from mathematics (matrix algebra), but it is not necessary for understanding how to interpret this measure

Interpretation of measures (1)

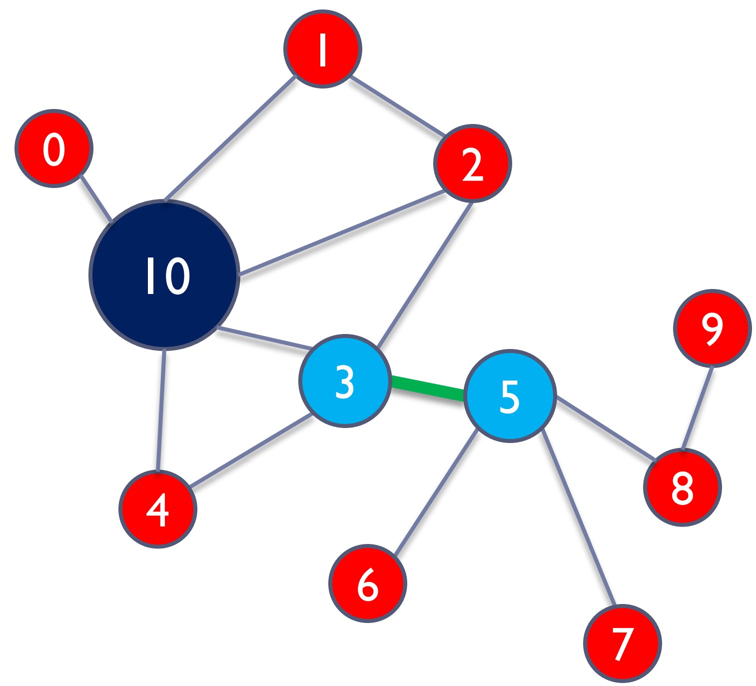
Centrality measure	**Interpretation in social networks**
Degree	How many people can this person reach directly?
Betweenness	How likely is this person to be the most direct route between two people in the network?
Closeness	How fast can this person reach everyone in the network?
Eigenvector	How well is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

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!

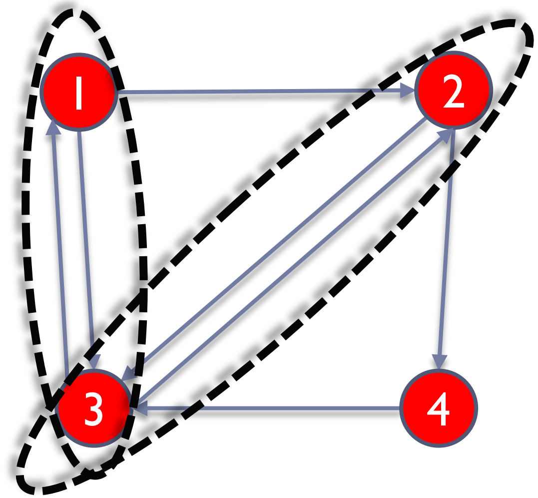


Basic Concepts

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



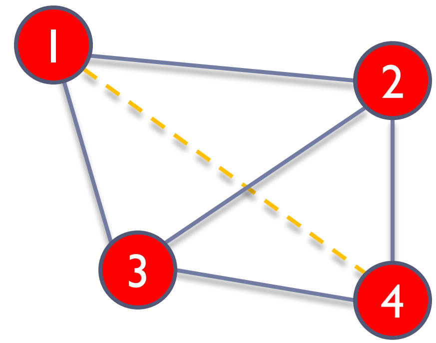
Reciprocity for network = 0.4

Density

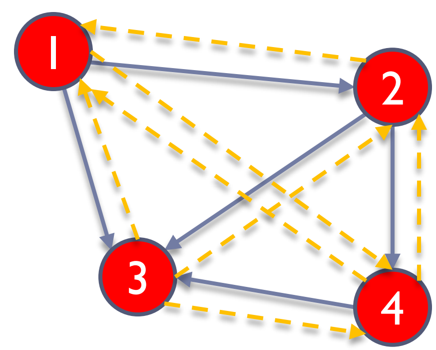
Edge present in network

Possible but not present

- 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

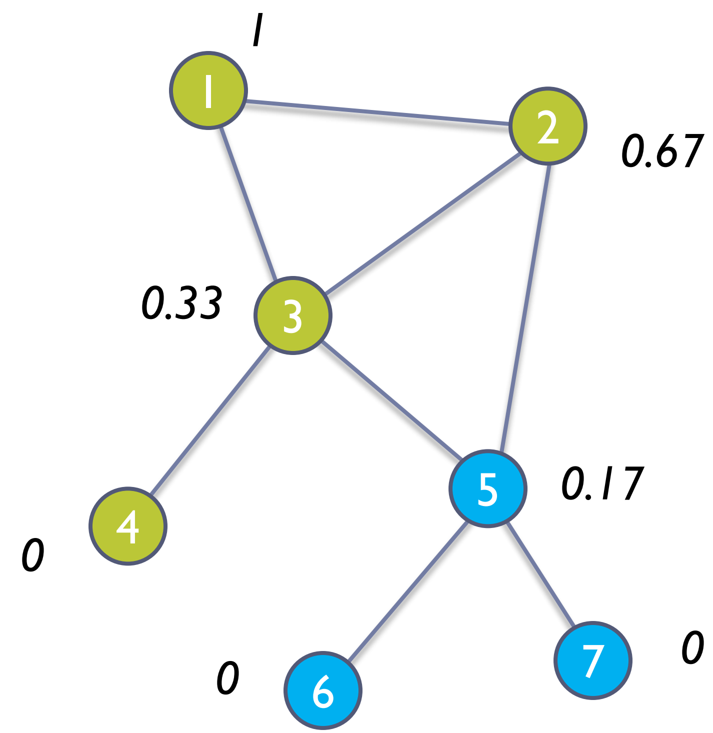
Clustering

Cluster A

Cluster B

NodeXL output values

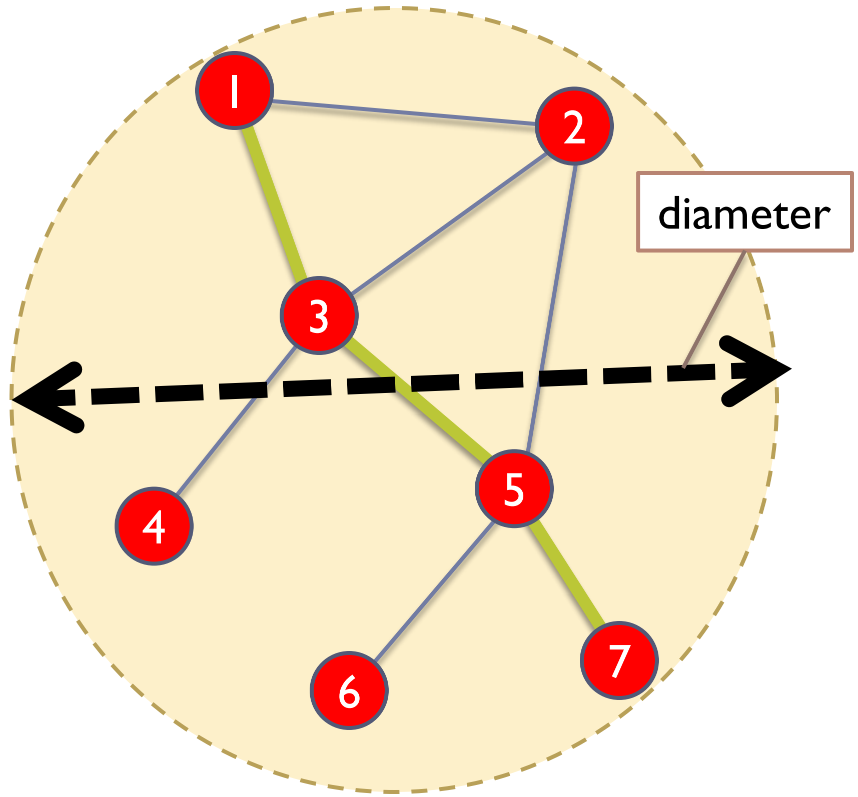
- A node's *clustering coefficient* is the density of its *neighborhood* (i.e. the network consisting only of this node and all other nodes *directly* connected to it)
- E.g., node 1 to the right has a value of 1 because its neighbors are 2 and 3 and the neighborhood of nodes 1, 2 and 3 is perfectly connected (i.e. it is a 'clique')
- The clustering coefficient for an entire network is the average of all coefficients for its nodes
- Clustering algorithms try to maximize the number of edges that fall within the same cluster (example shown to the right with two clusters identified)
- Clustering indicative of the presence of different (sub-)communities in a network



(Network clustering coefficient = 0.31)

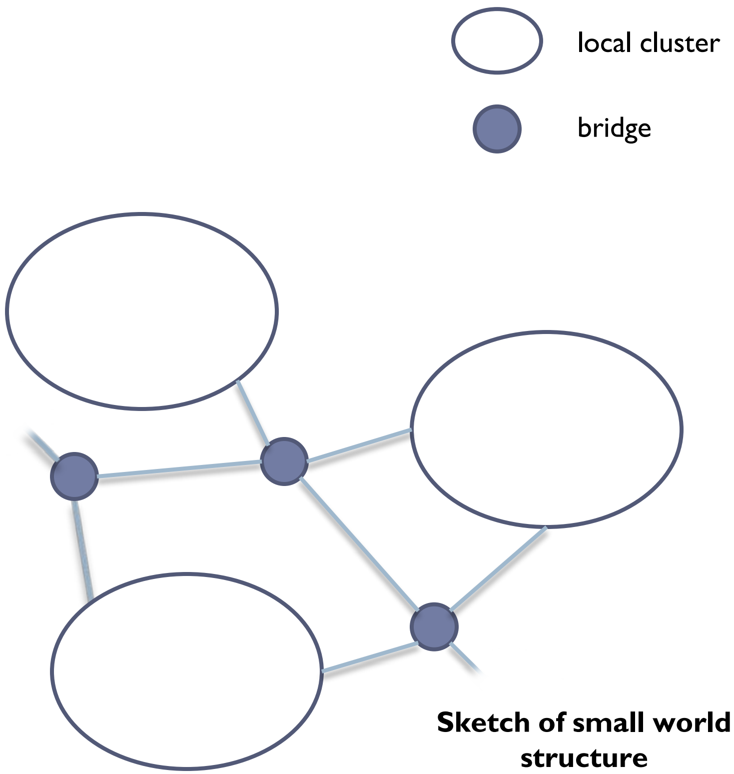
Average and longest distance

- 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

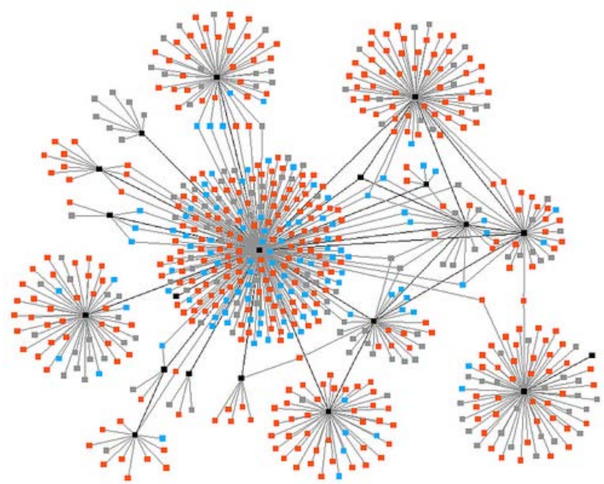
You may have heard of the famous "6 degrees" of separation

Preferential Attachment

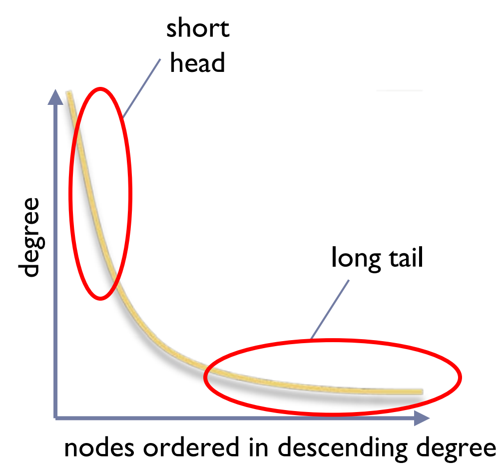
A property of some networks, where, during their evolution and growth in time, a the great majority of new edges are to nodes with an already high degree; the degree of these nodes thus increases disproportionately, compared to most other nodes in the network

- 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

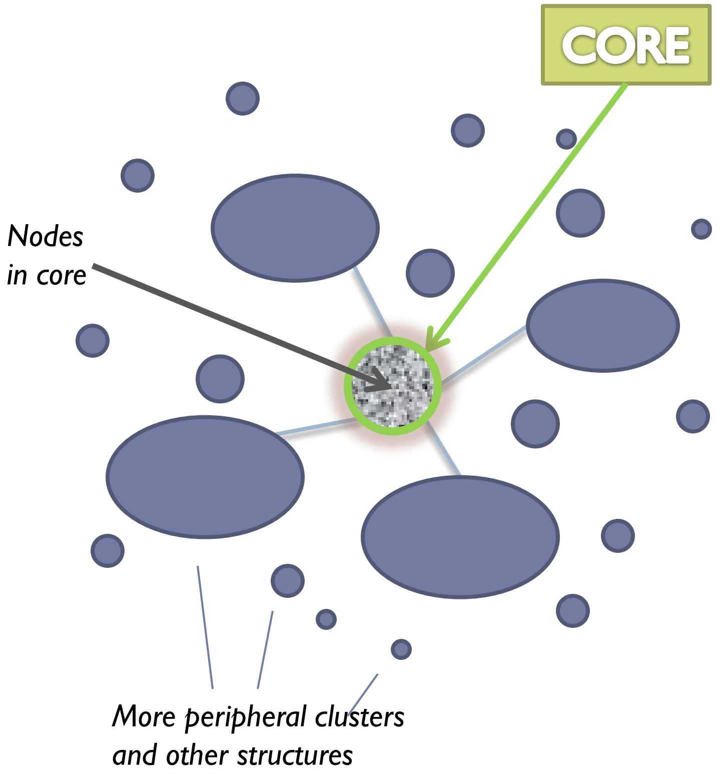
Among nodes of similar attributes, those that reach critical mass first will become 'stars' with many friends and followers ('halo effect')

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)



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

Analyzing your own ego-network

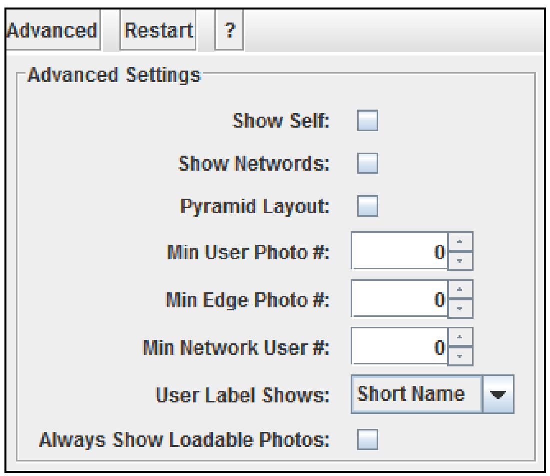
Now you will learn how to quickly visualize and analyze your own network on Facebook or Twitter, using freely available tools!

- Use the steps outlined in the following pages to visualize and analyze your own network
- Think about the key players in your network, the types of ties that you maintain with them, identify any clusters or communities within your network, etc.
- Objective: practice SNA with real data!
- **Present your findings in class next week!**

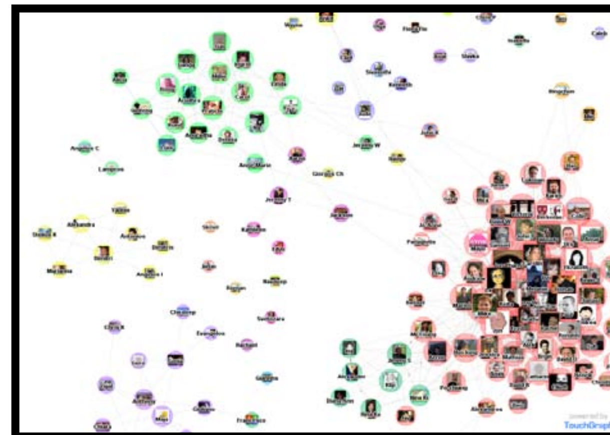
Visualizing Facebook ego-network online

- Launch the [TouchGraph Facebook Browser]
  - You should see a visualization of your network like the one to the right
  - Make sure to set 

    to a value that will allow you to see your entire network (friends ranked according to highest *betweeness centrality* according to TouchGraph Help)
  - Go to "Advanced" and remove all filters on the data so that settings look like below:



- Note: ***not* possible to export your data for further analysis**
- You may also want to try [TouchGraph Google Browser] (it's fun!)



Example TouchGraph Facebook Layout

Navigate the graph, examine friend 'ranks', friend positions in the network, clusters and what they have in common, try to identify weak and strong ties of yours and assess overall structure of your ego-network

Exporting data for offline analysis

- Data is more useful when you can extract it from an online platform and analyze with a variety of more powerful tools
  - Facebook, Twitter and other platforms have public Application Programming Interfaces (API's) which allow computer programs to extract data among other things
  - Web crawlers can also be used to read and extract the data directly from the web pages which contain them
  - Doing either of the above on your own will usually require some programming skills
  - Thankfully, if you have no such experience, you can use free tools built by others :)
- Bernie Hogan (Oxford Internet Institute) has developed a Facebook application that extracts a list of all edges in your ego-network (see instructions on next page)
- Also, NodeXL (Windows only, see later slide) currently imports data from: ***Twitter, YouTube, Flickr, and your email client***!

- Let's start with installing and learning to use NodeXL…

Using NodeXL for visualization & analysis

- Download and install NodeXL

  Windows 7/Vista/XP, requires Excel 2007

  (installation may take a while if additional software is needed for NodeXL to work)

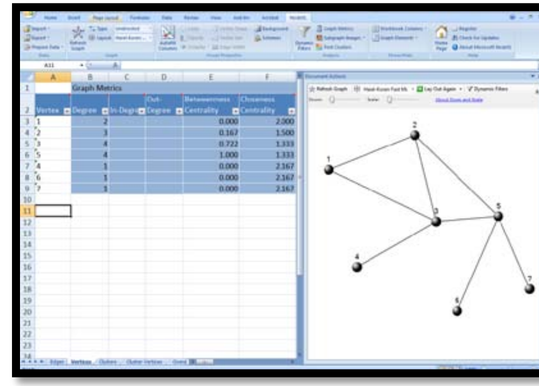
- Launch Excel and select

  New -> My Templates ->NodeXLGraph.xltx

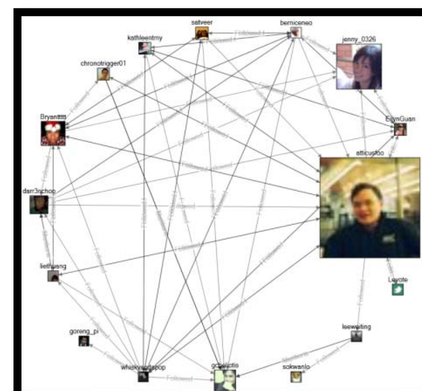
- Go to "Import" and select the appropriate option for the data you wish to import (for Facebook import see **next slide first**!)

- Click on [Graph Metrics] to ask NodeXL to compute centralities, network density, clustering coefficients, etc.

- Select [Refresh Graph | Harel-Koren Fast Mu | Lay Out Again] to display network graph. You can customize this using [Dynamic Filters] and [Autofill Columns] as well as [Options]



NodeXL sample screenshot



NM4881A Tweep Network (weekly data)

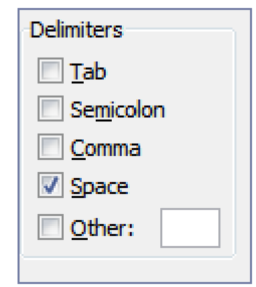
For more info read this **NodeXL tutorial**

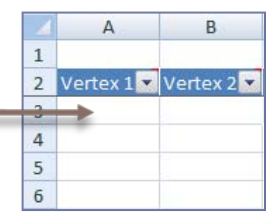
Exporting Facebook ego-network data

This will explain how to export your Facebook data for analysis with a tool like NodeXL

- To use Bernie Hogan's tool on Facebook, click here. From the two options presented, select "UCInet". This is a format specific to another tool, not NodeXL, but we will import this data into NodeXL because it's easier to use.
- After selecting "UCInet", *right-click* on the link given to you and select to save the generated file to a folder on your computer.
- Launch Excel and open the file that you just saved to your computer.
  1. Excel will launch the Text Import Wizard. Select "Delimited". Click "Next >"
  2. Select "Space" as the delimiter, as shown here
  3. Select "Next >" and then "Finish".
  4. A new file will be created in Excel. It contains a list of all the nodes in your ego-network, followed by a list of all edges. Scroll down until you find the edges, select all of them and copy them (Ctrl-C)
  5. You can now open a new NodeXL file in Excel as explained in the previous slide. Instead of using NodeXL's import function, paste the list of edges to the NodeXL worksheet, right here
  6. In NodeXL select 

and "Get Vertices from Edge List"
  7. Now you can compute graph metrics and visualize your data like explained in the previous slide!





More options

- Many more tools are used for SNA, although they generally require more expert knowledge. Some of these are:
  - Pajek (Windows, free)
  - UCInet (Windows, shareware)
  - Netdraw (Windows, free)
  - Mage (Windows, free)
  - GUESS (all platforms, free and open source)
  - R packages for SNA (all platforms, free and open source)
- The field is continuously growing, so we can expect to see more user-friendly applications coming out in the next years…

Credits and licensing

- Front page network graph by ilamont (license: CC BY)
- Bridge routing illustrations in Newman et al, *The Structure and Dynamics of Networks*, Princeton University Press
- Personal social network diagram in Mark Wellman (ed.), *Networks in the Global Village*, Westview Press
- Visualization of interactions in organization in Garton et al, *Studying Online Social Networks*, JCMC
- Visualization of US political bloggers in Lada Adamic and Natalie Glance, *The Political Blogosphere and the 2004 US election: Divided They Blog*, Proceedings of the International Conference on Knowledge Discovery and Data Mining, ACM Press, 2005
- Web bow tie diagram by Broder et al, *Graph Structure in the Web*, Computer Networks, Elsevier
- Bond/tie photo by ChrisK4u (license: CC BY-ND)
- Visualization of small world network by AJC1 (license: CC BY-NC)