Social Network Analysis (SNA)

Nik Bear Brown

In this lesson we will disucss Social network analysis and graph theory. There are no data sets, or libraires to be installed.

Social network analysis (SNA)

Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories.

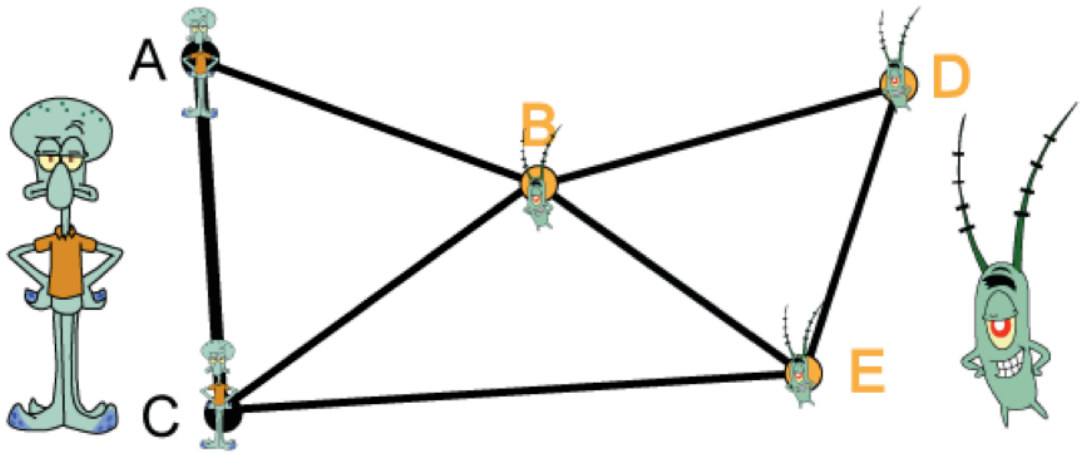
A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency, such as values, visions, ideas, financial exchange, friends, kinship, dislike, conflict, trade, web, sexual relations, disease transmission.

Social Networks

What is a network?

A network is a graph. That is:

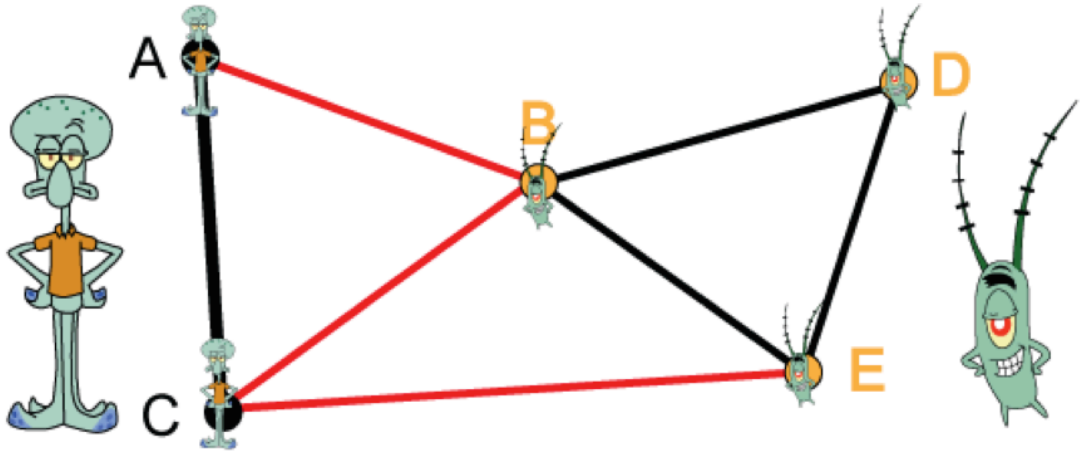
- A set of vertices (or nodes) joined by edges.
- Vertices and edges can have properties (feature values).
- Edges can have weights and be directed.



-
Graph

Cut Sets

In graph theory, a graph cut is a partition of the vertices of a graph into two disjoint subsets. Any cut determines a cut-set, the set of edges that have one endpoint in each subset of the partition.

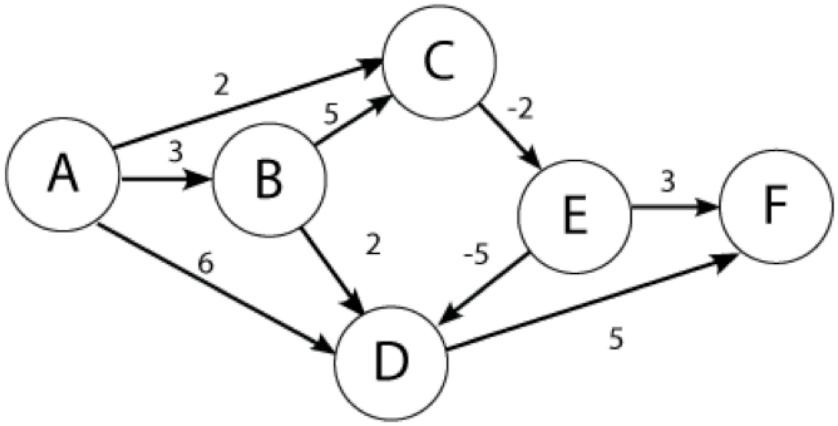


- Cut Set

Directed graph (digraph)

Each edge is an ordered pair of vertices, to indicate direction Lines become arrows.

- The indegree of a vertex is the number of incoming edges
- The outdegree of a vertex is the number of outgoing edges



- Directed graph (digraph)

Paths and Connectivity

A path in an undirected graph G = (V, E) is a sequence P of nodes $v_1, v_2, \ldots, v_{k-1}, v_k$ with the property that each consecutive pair v_i, v_{i+1} is joined by an edge in E.

A path is simple if all nodes are distinct.

An undirected graph is connected if for every pair of nodes u and v, there is a path between u and v.

Cycles

A cycle is a path $v_1, v_2, \ldots, v_{k-1}, v_k$ in which $v_1 = v_k, k > 2$, and the first k-1 nodes are all distinct.

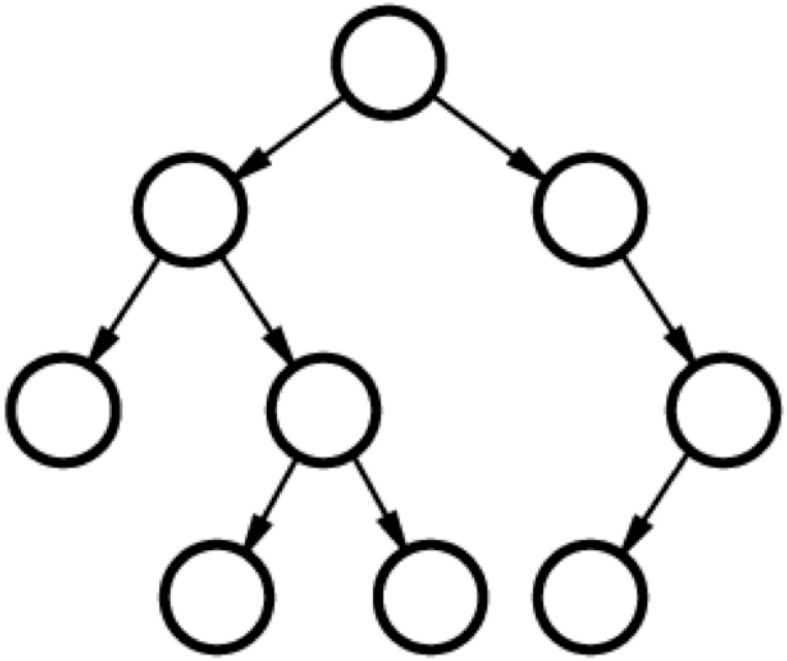
Trees

An undirected graph is a tree if it is connected and does not contain a cycle.

Theorem. Let G be an undirected graph on n nodes. Any two of the following statements imply the third. * G is connected.
* G does not contain a cycle.
* G has n-1 edges.

Rooted Trees

Rooted tree. Given a tree T, choose a root node r and orient each edge away from r.



- Rooted tree

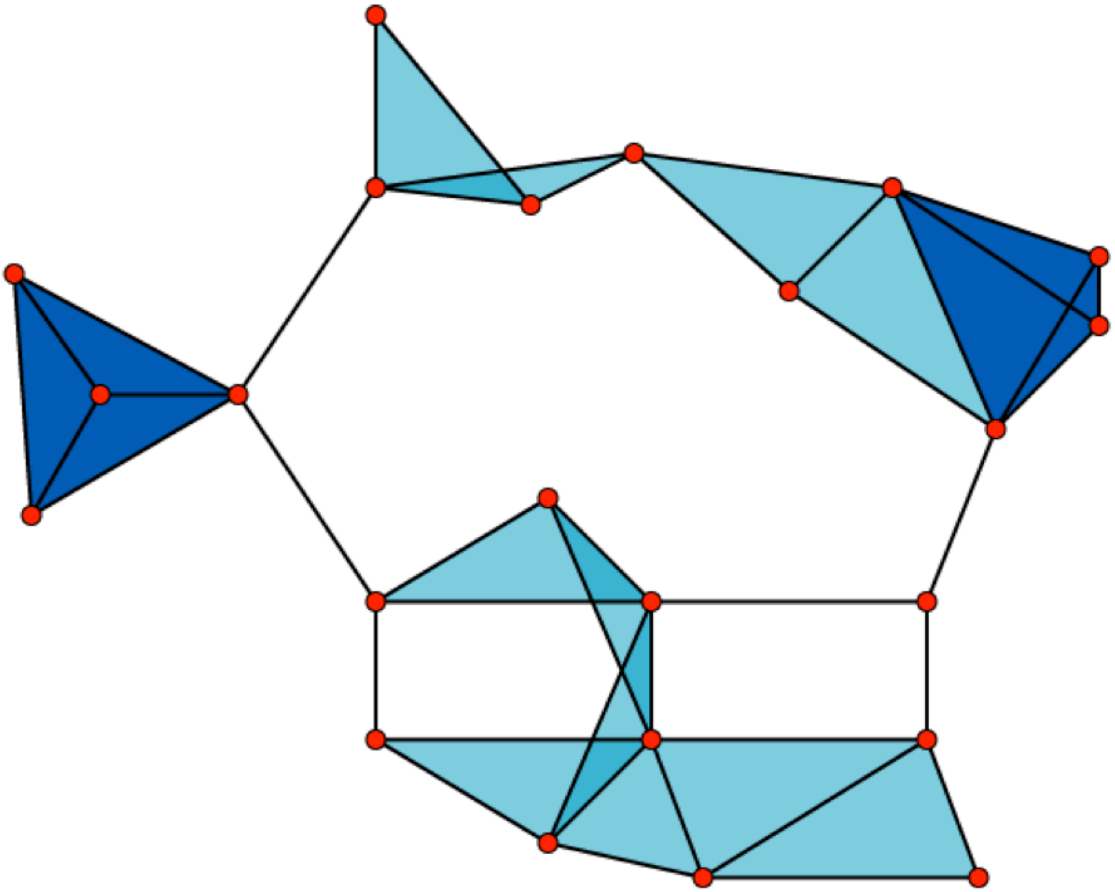
Importance. Models hierarchical structure.

Diameter

The longest shortest path in the graph. In other words, a graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another when paths which backtrack, detour, or loop are excluded from consideration.

Clique

A clique is a fully connected graph. A graph that has all possible n(n-1)/2 edges.



Clique (graph theory)

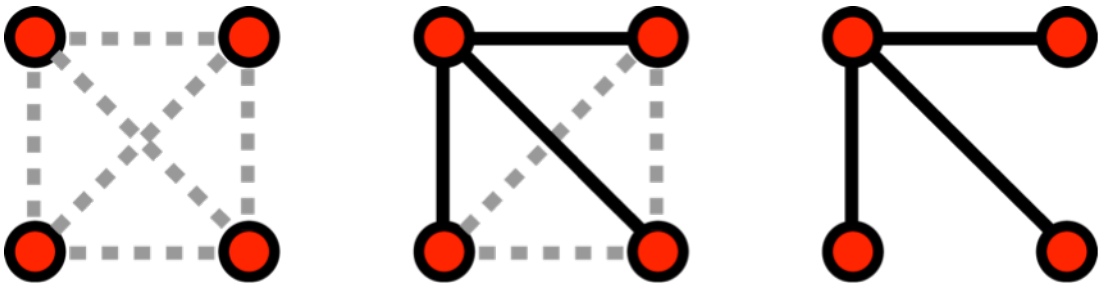
A graph with * 23 x 1-vertex cliques (the vertices), * 42 x 2-vertex cliques (the edges), * 19 x 3-vertex cliques (the light and dark blue triangles), and * 2 x 4-vertex cliques (just the dark blue areas).

The 11 light blue triangles form maximal cliques. The two dark blue 4-cliques are both maximum and maximal, and the clique number of the graph is 4.

from Clique (graph theory)

Erdos-Renyi Random Graphs

In graph theory, the Erdos-Renyi model is either of two closely related models for generating random graphs. They are named after Paul Erdős and Alfréd Rényi.



- Erdos-Renyi model is generated with N = 4 nodes. from en.wikipedia.org/wiki/Network_science#/media/File:ER_model.png

The G(n, M) model, * n is the the number of vertices or nodes
* M is the the number of edges * $0 \leq p \leq 1$ * for each pair (i,j), generate the edge (i,j) independently with probability p

Equivalently, all graphs with n nodes and M edges have equal probability of

$$p^M(1-p)^{\binom{n}{2}-M}.$$

The parameter p in this model can be thought of as a weighting function; as p increases from 0 to 1, the model becomes more and more likely to include graphs with more edges and less and less likely to include graphs with fewer edges.

Random graphs degree distribution follows a binomial.

$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

where n is the total number of vertices in the graph. Since

$$P(\deg(v) = k) \to \frac{(np)^k e^{-np}}{k!} \quad \text{as } n \to \infty \text{ and } np = \text{const},$$

this distribution is Poisson for large n and np = const.

The Strength of Weak Ties

Granovetter's "The Strength of Weak Ties" (considered one of the most important sociology papers written in recent decades with about 30,000 citations according to Google Scholar). Granovetter argued that "weak ties" could actually be more advantageous in politics or in seeking employment than "strong ties", because weak ties allowed an individual to reach a higher number of other individuals.

SNA Metrics

Structural (quantitative) * Size - Number of nodes.
* Density - Number of ties that are present vs the amount of ties that could be present.
* Diversity/Homophily - Homophily (i.e., "love of the same") is the tendency of

individuals to associate and bond with similar others.
* Out-degree- Sum of connections from a node to others.
* In-degree - Sum of connections to a node.
* Structural Holes - In the context of social networks, social capital exists where people have an advantage because of their location in a network. Most social structures tend to be characterized by dense clusters of strong connections. When two separate clusters possess non-redundant information, there is said to be a structural hole between them.
* Isolates/Cliques - Highly connected subnodes.
* Centrality - Nodes who have more ties to other nodes may be advantaged positions in social networks. Many ties provide redundancy and reachability. Centrak nodes have alternative ways to satisfy needs, and hence are less dependent on other individuals.
* Betweeness - a hub, other nodes need to connect through the between node to connect to a desired node. Also called betweenness centrality (i.e. A number that represents how frequently a node is between other nodes geodesic paths.
* Closeness - Average Path distance * Closeness: Reach (i.e. number of hops) * Closeness: Geodesic distance - shortest path. * Closeness centrality - Distance of one node to all others in the network.
* Diameter - Maximum greatest least distance between any node and another.
* Maximum flow - The amount of different nodes in the neighborhood of a source that lead to pathways to a target.

Relational (qualitative) * Strength of ties * Accessibility * Likeability/"fun" * Reputation * Expected reciprocity? * Competing unit? * Dependence * Trust

The Tipping Point

The Tipping Point: How Little Things Can Make a Big Difference is the debut book by Malcolm Gladwell, first published by Little Brown in 2000.

Malcolm Gladwell describes the "three rules of social epidemics", that is, how epidemics spread through social networks. The three rules:

1 The Law of the Few

Gladwell theorizes that a few specialized nodes are needed for virality of a message.

Connectors are the people in a community who know large numbers of people and who are in the habit of making introductions. A connector is essentially the social equivalent of a computer network hub.

Mavens are "information specialists", or "people we rely upon to connect us with new trusted information." The represents authority and trust in new information.

Salesmen are "persuaders", charismatic people with powerful negotiation skills.

2 The Stickiness Fnode

The specific content of a message that renders its impact memorable.

3 The Power of Context

Malcolm Gladwell says, "Epidemics are sensitive to the conditions and circumstances of the times and places in which they occur".

Centrality Measures

Local Centrality (Degree): The number of links an node has with other nodes.

- A potential sign of power.
- High in-degree can be a sign of prominence or prestige.
- High out-degree can be a sign of influence.
- Betweeness: The degree to which an node is situated between two groups, and is a necessary route between those groups.
  nodes with high betweeness have the potential to have a major influence
- A node with high betweenness has great influence over what flows in the network indicating important links and single point of failure.
- They can be mediators/brokers, gatekeepers, bottlenecks, or obstacles to communication.
- They are especially valuable when the link two diverse groups.

Global Centrality (Closeness): the average distance between an node and all other nodes in a network.

- Most likely to be “in the know” about what is happening

Small world phenomena

The small-world experiment comprised several experiments conducted by Stanley Milgram and other researchers examining the average path length for social networks of people in the United States. Letters were handed out to people in Nebraska to be sent to a target in Boston. People were instructed to pass on the letters to someone they knew on first-name basis The letters that reached the destination followed paths of length around six. The experiments are often associated with the phrase "six degrees of separation", although Milgram did not use this term himself.

See Small world project

Measuring the Small World Phenomenon

Diameter - The longest shortest path in the graph. In other words, once the shortest path length from every node to all other nodes is calculated, the diameter is the longest of all the calculated path lengths.

Characteristic (Average) path length - Average path length is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. This shows us, on average, the number of steps it takes to get from one member of the network to another.

Harmonic mean - In mathematics, the harmonic mean is one of several kinds of average, and in particular one of the Pythagorean means. The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n$ is defined to be

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n \cdot \prod_{j=1}^{n} x_j}{\sum_{i=1}^{n} \frac{\prod_{j=1}^{n} x_j}{x_i}}.$$

From the third formula in the above equation, it is more apparent that the harmonic mean is related to the arithmetic and geometric means.

Clustering Coefficient

In graph theory, a clustering coefficient is a measure of degree to which nodes in a graph tend to cluster together.

Clustering Coefficients were introduced by Watts & Strogatz in 1998, as a way to measure how close a node (or vertex) and its neighbors are from being a clique, or a complete graph within a larger graph or network.

Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterized by a relatively high density of ties.

Mixing patterns

Assume that we have various types of nodes. What is the probability that two nodes of different type are linked? Social networks tend to be "Birds of a Feather." Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience.

Individuals in homophilic relationships share common characteristics (beliefs, values, education, etc.) that make communication and relationship formation easier. The opposite of homophily is heterophily.

Resources

- Social Network Analysis using R and Gephis | R-bloggers
- Social Network Analysis - University of Michigan | Coursera
- Social Network Analysis in R
- Social Network Analysis | R-bloggers
- Social Networks in R - Shizuka Lab
- SOCIAL NETWORK ANALYSIS WITH R