Social Network Analysis

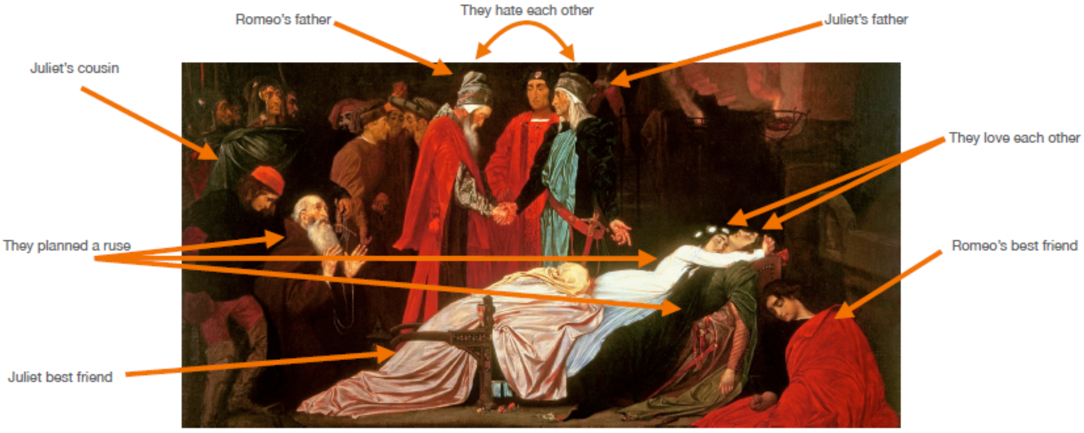
Matteo Donati

October 12, 2022

Contents

1 Introduction

A network is a simplified representation that reduces a system to an abstract structure or topology, capturing only the basics of the connection patterns and little else.



In particular, networks capture the pattern of interactions between the parts of a system. In turn, the pattern of interactions have a sensible effect on the behaviour of a system. Knowing the structure of a network is essential to fully understand how its corresponding system works. Moreover:

- The systems studied can have interesting features not represented by the network. For example, the detailed behaviours of individual nodes, such as people and the precise nature of the interactions between them.

- One can capture additional information by labelling the nodes and/or edges of the network, such as with names or strengths of interactions. However, every time one defines a representation of a full system, one decides to filter out some information.

1.1 Network Analysis

Network analysis is about studying the nature and function of the system a network describes, given such network. A first step in analysing the structure of a network is often to make a picture of it. Automatic tools help in managing, visualising, and exploring networks. Indeed, the human eye is enormously gifted at discerning patterns, and visualisations allow us to put this gift to work on our network problems. However, the visualisation process can help study the system only when the network is sparse (i.e. the number of edges is quite small). To address the issue of large and dense networks, network theory has developed a large tool-chest of measures and metrics that mimic some specific abilities of our eyes. Examples of such measures are **centrality** (namely, quantifying how important a node is), and the **small-world effect** (namely, the shortest distance between a given pair of nodes).

2 Research Design

The elements of a research paper are:

- Context. This includes the general idea of the paper and the specific application.
- Problem/motivation. This describes what are the problems the author want to address, why these problems are important and what the main contributions of the paper are.
- Data. This section usually describes how the authors gather data, what tools they used, and what measures they applied.
- Results. This describes the connection among the gathered data, the applied measures and the properties found.

Papers are usually peer-reviewed.

2.1 Elements of Network Analysis Research

The all network analysis research uses a tool from discrete mathematics called **graph theory**. This theory can be used to reduce and draw conclusions from naturally-occurring phenomena. In particular, there exist two fundamental kinds of network research designs:

- **Whole network**, in which one studies the set of ties among all pairs of nodes in a given set. Whole network designs enable researchers to employ the full set of network concepts and techniques, which often assume that the entire network is available. However, the cost of assembling and managing the network can quickly rise due to the whole-network scope.
- **Personal network**, in which one studies a set of nodes called *egos* and their ties to others, called *alters*. Personal network designs have the advantage of simplifying the gathering and management of the network.

The main sources of network data are:

- **Primary sources**, where the researcher collects the data first-hand.
- **Secondary sources**, where the researcher gathers data that already exists somewhere.

In order to determine when enough data is enough, one must consider the nature of the research questions. Indeed, the amount of data varies with respect to such type of question. Moreover, when considering sources of data, one should be aware of reliability and validity issues:

- **Validity** is about measuring what one intends to measure. In a network study this entails understanding how closely one's model represents reality (e.g. a map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness). Validity errors include:

- **Omission errors**: missing edges and nodes have huge impacts on errors in network variables, by making the network appear more/less disconnected than it really is, or make nodes and edges in the network appear to be more important than they really are.
  - **Commission errors**: dual to omission errors. Namely, the erroneous inclusion of nodes and edges can effect the ultimate determination of node-level measures and the identification of key nodes.
  - **Data collection and retrospective errors**: these errors are related to data collected from individuals where the network-elicitation question deals with reports of behaviour, in particular when one has to do with social interactions of a temporally discrete nature.
  - **Edge/node attribution errors**: mis-assignment of a behaviour to a node can yield linkages that in reality do not exist.

- **Reliability** is about finding out if, given the same conditions, the results of an experiment would turn out to be exactly the same. Usually, if one relies on objective data and one applies objective measures, then one would increase the reliability of the study. Threats to reliability include:

  - **Data fusion/aggregation**: when aggregating data on different temporal, relational or spatial scales, it is possible to exclude important nodes and edges because these have lost their importance in the network. For this reason, there should exist aggregation decisions.
  - **Errors in secondary sources and data mining**: secondary-source data may have inherent biases, which should be considered in any analysis.
  - **Formatting errors**: when mining data, errors can derive from unexpected differences in document formatting. These errors can lead to the over- or under-representation of terms, actors, attributes, etc. in the data retrieval process.

3 Mathematics of Networks

A network, also called a graph in the mathematical literature, is a collection of nodes (or vertices) joined by edges. Nodes and edges are also called *sites* and *bonds*, or *actors* and *ties*. Generally:

- The common notation for the number of nodes in a network is n and the number of edges is m.
- The simplest networks have at most a single edge between any pair of nodes. When there are more than one edge between the same nodes, the set of edges between two nodes are called **multi-edge**. When nodes have edges to themselves, these nodes are called **self-edges**.
- Networks that neither have self-edges nor multi-edges are called **simple networks**.
- A network with multi-edges is called a **multigraph**.

The fundamental mathematical representation of a network is the **adjacency matrix**. Such matrix A is defined to be the $n \times n$ matrix with elements A_{ij} such that:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between nodes } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Edges in a graph can be given a weight. In this case, the network (or graph) is called a **weighted network**. These weights can be stored in the usual adjacency matrix. Moreover:

- **Directed networks** or **directed graphs** are networks in which each edge has a direction, pointing from one node to another. In this case:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge from node } j \text{ to node } i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- A cycle in a directed network is a closed loop of edges with the arrows on each of the edges pointing the same way around the loop. A self-edge counts as a cycle.
- Directed networks that have no cycles are called **acyclic networks** (**DAG**).
- A **bimodal network** is a network with two kinds of nodes and edges that run only between nodes of different kinds. The equivalent of the adjacency matrix for a bipartite graph is a rectangular matrix called the **incidence matrix**. This matrix B a $n \times g$ matrix, where n is the number of nodes in the network and g is the number of groups:

$$B_{ij} = \begin{cases} 1 & \text{if item } j \text{ belongs to group } i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

- A **walk** is a network is a sequence of nodes such that every consecutive pair of nodes in the sequence is connected by an edge. The length of a walk in a network is the number of edges traversed along the walk, counted separately as they are traversed. Walks that do not intersect themselves are called **paths**.

- To calculate the number of walks of a given length r on a network one can apply the following rule:

$$N_{ij}^{(r)} = [A^r]_{ij} \tag{4}$$

  For example, the number of walks of length two is computed as:

$$N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik} A_{kj} = [A^2]_{ij}$$

- The **shortest path** in a network, is the shortest walk between a given pair of nodes. Mathematically, the shortest distance between nodes i and j is the smallest value of r such that $[A^r]_{ij} > 0$.

- The **diameter** of a network is the length of the longest among all existing shortest paths between every pair od nodes in the network. This can be useful to understand the connectedness of networks.

- A network does not necessarily consist of a single connected set of nodes. Indeed, frequently networks have separate parts that are disconnected from one another. Such parts are called **components**. Technically, a component is a subset of the nodes of a network with the following properties:

  - There exists at least one path from each member to each other member of that subset.
  - No other node in the network can be added to the subset.

  A network in which all nodes belong to the same single component is said to be **connected**.

4 Data Collection and Data Management

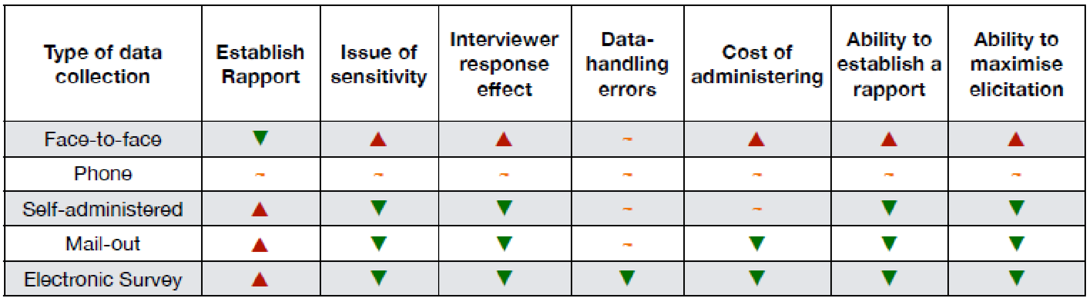
Networks questions are an example of data collection. Indeed, these are all those questions which help build a network. The proper selection of the network questions and formats is critical to the success of any network study. A fundamental issue in the design of network questions is whether to use an open- or closed-ended format:

- **Close-ended** questions require the definition of the set of nodes of the network beforehand and respondents respond to answers on their relations with those actors.
- **Open-ended** questions require no prior decisions on who to obtain information about.

On the other hand, the **respondent burden** represents the commitment required from the respondent to participate to the study, including time, attention, and emotions. A guiding principle to relieve respondent burden is to minimize the respondent's anger and frustration.

4.1 Data Collection and Reliability

Network approaches are usually sensitive to missing/wrong data. Moreover, the smaller the network, the larger the effect of omissions or commissions of actors and/or ties. The process of collection of network data has a profound impact on actor participation and on the reliability and validity of the collected data. In general:



Type of data collection	Establish Rapport	Issue of sensitivity	Interviewer response effect	Data-handling errors	Cost of administering	Ability to establish a rapport	Ability to maximise elicitation
Face-to-face	▼	▲	▲	~	▲	▲	▲
Phone	~	~	~	~	~	~	~
Self-administered	▲	▼	▼	~	~	▼	▼
Mail-out	▲	▼	▼	~	▼	▼	▼
Electronic Survey	▲	▼	▼	▼	▼	▼	▼

Moreover:

- To perform **archival data-collection**, the archival sources must contain information of social relations that are amenable to either a one-mode or two-mode network format. Also, less structured archival data can be a source for relational studies.
- The collection of data from **electronic sources** is similar to the collection of network data in archival or historical research. Many of these sources already have information available in a one-mode or two-mode network format, while others require using/writing data-mining software to put it into data formats that can be more readily analysed.

Lastly, when storing network data digitally, one needs to decide a form of representation in the memory of the computer. The first step in representing a network in a computer is to label the nodes so that each can be uniquely identified. The most common way of doing this is to give each node a numerical label, usually an integer. Then, one could either use an adjacency matrix or an adjacency list to store the edges between nodes. Often, the nodes in a network have annotations or values attached to them, in addition to labels. All of these other notations and values can be stored in the memory by defining an array of a suitable type, one for each node.

4.2 Data Transformation

There are many transformations applicable to data in the course of an analysis to make some evidence emerge:

- **Transposing** a matrix means interchanging its rows with its columns. Transposition, applied to a non-symmetric adjacency matrix, reverses the direction or arcs and can be helpful in maintaining a consistent interpretation of the ties in the network.

- Many graph-theoretic measures are sensible to **missing values**. A naive solution is to eliminate the nodes of which we miss the data. However, node removals should be performed with a reason, accounted for in the analysis of the validity and reliability of the study. In the case of symmetric or undirected relations, a direct solution is to fill-in the missing rows with the data found in the corresponding column. For non-symmetric relations, however, this technique would make no sense. However, if one has two non-symmetric relations that can be used to fill each other's missing data, one can use the transpose of the second matrix to fill-in the missing rows in the first matrix, and vice versa.

- **Symmetrising** means creating a new dataset in which all ties are reciprocated. In general, there are many reasons to symmetrise data:

  - Some analytical techniques assume symmetric data.
  - Some data-cleaning processing has a symmetrisation step.
  - When studying inherently symmetric relations.

  From the point of view of a matrix A representing a network, when one symmetrises, one is comparing an entry A_{ij} with corresponding entry A_{ji} and, if needed, one makes them the same. More in general, the **union policy** corresponds to taking the larger of the two entries while the **intersection policy** takes the smaller of the two. Besides the above two policies, others are possible.

- **Dichotomising** refers to converting valued data to binary data. The main reason for this is that some measures are only applicable to binary data. Dichotomising is the first example of a more general concept of edge cut-off, useful to reduce the density of a large network and to make it more efficient/feasible to handle.

$$
\begin{array}{c|cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 0 & 0 & 0 & 4 & 0 & 0 \\
2 & 0 & 0 & 4 & 0 & 0 & 0 \\
3 & 1 & 0 & 0 & 0 & 2 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 1 \\
5 & 0 & 0 & 0 & 3 & 0 & 2 \\
6 & 0 & 2 & 0 & 0 & 0 & 0
\end{array}
\xrightarrow{\text{cut-off set to three}}
\begin{array}{c|cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
2 & 0 & 0 & 1 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 1 & 0 & 0 \\
6 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

- Most network studies collect, on the same set of nodes, multiple relations which are then useful to be **combined** into one. Mathematically, to combine relations, one can sum the separate matrices:

$$
\begin{aligned}
A + B &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} \\
&= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}
\end{aligned} \tag{5}
$$

- Is some studies it is useful to re-express, **standardize** or **normalize** network data to ensure one is making fair comparisons across rows, columns or entire matrice.

4.3 Running Time and Computational Complexity

Computational complexity is a measure of the running time of a computer algorithm, as a function of the size of the problem it is tackling. Let n be the number of steps needed to solve a given problem in the worst possible case. Then, the time complexity of such algorithm is of order n, or just $O(n)$ for short.

5 Measures and Metrics, Nodes

A way to simplify and represent a studied network is to define mathematical measures that capture interesting features of the network structure quantitatively, boiling down large volumes of complex structural data into numbers that are an indication of the studied phenomena. These measures can be of different type:

- **Binary scale**. Conventionally, 1 indicates the presence of a relationship and 0 indicates its absence. Being the ground floor of the information, it can always be obtained starting from another metric, defining a threshold value (cut-off point) below which all values are reported to 0 and above which all values are reported to 1.
- **Multi-category nominal scales**. This metric indicates for each relation the type that it assumes, with respect to a multiple-choice list (e.g. lover, friend, colleague, enemy, etc.).
- **Ordinal scales**. The simplest ordinal metric refers to a three-value scale of the type -1, 0, $+1$, where -1 implies the presence of a negative relationship, 0 indicates indifference, and $+1$ implies the symmetrical situation to the negative one.
- **Scalar scales**. Scala metrics are useful when handling values representing either physical quantities, or information units of account.

Examples of metrics are:

- **Centrality**. This metric measures which nodes are the most important in the network.
  - One of the simplest centrality measure for a node is its **degree** (namely, the number of arcs entering and exiting the specific node). In directed networks, nodes have both an in-degree and an out-degree.
  - In many circumstances a node's importance in a network is increased by having connections to other nodes that are themselves important. **Eigenvector centrality** is an extension of degree centrality that takes this factor into account. Instead of just awarding one point for every network neighbour a node has, eigenvector centrality awards a number of points proportional to the centrality scores of the neighbours. Considering an undirected network of n nodes, the eigenvector centrality x_i of node i can be computed as:

    $$x_i = \alpha \sum_{j \in \text{neighbours}(i)} x_j \tag{6}$$

    which can be also expressed as:

    $$x_i = \alpha \sum_{j=1}^{n} A_{ij} x_j \quad \Rightarrow \quad \alpha^{-1} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{7}$$

Still, the values for $x_1, \ldots, x_n$ are unknown. However, this last transformation let one understand that the vector of centralities is one of the possible eigenvectors of the matrix A ($kx = Ax$, where k is the eigenvalue and x is the eigenvector):

$$k \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = k^{-1} A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{8}$$

Assuming one wants the centrality values to be all positive, then one can use the Perron-Frobenius theorem, by which for a square matrix with all elements non-negative there exists a unique largest eigenvalue k and the corresponding eigenvector x, called leading, that have strictly positive components. The eigenvector centrality x_i of node i is the i-th element of the leading eigenvector of the adjacency matrix and the value of the constant k is the leading eigenvalue.

For the case of directed networks, the eigenvector centrality poses some complications due to the asymmetricity of adjacency matrices. This translates into two sets of eigenvectors, left and right. Which to choose among the two depends on the reason of the calculation of the centrality measure. The right eigenvector measures centrality as bestowed by others to the node. The left eigenvector measure centrality as connections of the node to the others.

Eigenvector centrality also has the problem that, whenever some node A has only outgoing edges and no ingoing ones, its eigenvector centrality will be zero. Moreover, all other nodes B which have only one ingoing edge, and such edge comes from A, then also its centrality will be zero.

– To solve the problem of zero-trailing in eigenvector centrality for directed networks, **Katz centrality** was proposed. This metric gives each node a small amount of centrality regardless of its position in the network or the centrality of its neighbours:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta \tag{9}$$

One problem with Katz centrality is that, if a node with high Katz centrality has edges pointing to many others, than all of those others also get high centrality. To solve this problem, one can define a variant of the Katz centrality where one derives the centrality of the neighbours as proportional to their centrality divided by their out-degree. Then nodes that point to many others pass only a small amount of centrality on to each of those others, even if their own centrality is high:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{\mathrm{od}(j)} + \beta \tag{10}$$

where $\mathrm{od}(j)$ represents the out-degree of node j. If j has an out-degree equal to zero, then $\mathrm{od}(j) = 1$.

- The centrality measures for directed networks described so far all follow the same basic principle: high centrality goes with being pointed by others with high centrality. In some cases, nodes are highly central when they point to other highly central ones. In this kind of networks there are two types of important nodes, **hubs and authorities**: authorities are nodes that hold useful resources, while hubs are nodes that are gateways toward the most resourceful authorities.

  Hyperlink-induced topic search is a centrality measure that gives each node i in a directed network two different centrality scores: the authority centrality x_i and the hub centrality y_i, defined using the constants α and β, and by swapping the indices of the matrix element (since the hub centrality of a node i is defined by the nodes it points to):

  $$x_i = \alpha \sum_j A_{ij} y_i \qquad y_i = \beta \sum_j A_{ji} x_j \tag{11}$$

  Hub- and authorities-centrality circumvent the problems that ordinary eigenvector centrality has with directed networks: in the hub-and-authority approach, nodes not pointed by any others have authority centrality zero but they can still have non-zero hub centrality, and the nodes that they point to can then have non-zero authority centrality by virtue of being pointed.

- **Closeness centrality**, instead, uses the shortest paths in networks, measuring the mean distance from a node to other nodes. Let d_{ij} be the shortest distance from node i to node j. Then, the mean shortest distance from i to every node in the network is:

  $$\ell_i = \frac{1}{n} \sum_j d_{ij} \tag{12}$$

  Thus, ℓ_i is not a centrality measure per-se, since it gives low values to more central nodes and high values to less central ones. To be used as a centrality, one can use the inverse of ℓ_i:

  $$C_i = \ell_i^{-1} = \frac{n}{\sum_j d_{ij}} \tag{13}$$

  Closeness centrality has a problem with networks with more than one component and non-existing paths set to infinite. There, when nodes belong to different components, ℓ_i is infinite and C_i is zero. To solve this problem, it is possible to average over only those nodes in the same component as node i. An alternative solution is to redefine closeness in terms of the harmonic mean distance:

  $$\ell_i' = \frac{n-1}{\sum_{j(\neq i)} \frac{1}{d_{ij}}} \qquad C_i' = \frac{1}{n-1} \sum_{j(\neq i)} \frac{1}{d_{ij}} \tag{14}$$

  When $d_{ij} = \infty$ because i and j are in different components, the measure zeroes the term and drops it. Moreover, it gives more weight to nodes that are close to i than to those far away.

- **Betweenness centrality**, also based on shortest paths, measures the extent to which a node lies on paths between other nodes. The assumption is that paths lying on trafficked shortest paths have a more central role in the network, as gateways favoured by their closeness ro reach the other nodes. Supposing one has an undirected network in which there is at most one shortest path between any pair of nodes, and let n^i_{sd} be 1 if node i lies on the shortest path from the source s to the destination d, and 0 if it does not or if there is no such path. Then, the betweenness centrality x_i is computed as:

  $$x_i = \sum_{sd} n^i_{sd} \tag{15}$$

  where the sum is over all shortest paths (sd).

  Since it is possible for two shortest paths between the same pair of nodes to overlap, one can refine n^i_{sd} to be the number of shortest paths from s to d that pass through i, and define g_{sd} as the total number of shortest paths from s to d, obtaining:

  $$x_i = \sum_{sd} \frac{n^i_{sd}}{g_{sd}} \tag{16}$$

  Assuming as convention $n^i_{sd}/g_{sd} = 0$ if both terms are zero, the newly-defined value of x_i corresponds to the average rate of the traffic that passes through node i.

  Sometimes is convenient to normalize betweenness. One natural choice is to normalize the path count by dividing it by the total number of node pairs, which is n^2, so that betweenness becomes the fraction of paths that run through a given node:

  $$x_i = \frac{1}{n^2} \sum_{sd} \frac{n^i_{sd}}{g_{sd}} \tag{17}$$

  This new measure has the additional benefit of limiting the values of centrality between 0 and 1.

- **Groups of nodes**. Many networks divide naturally into groups or communities. Besides calculating their centrality, it is possible to apply measures to nodes to detect their membership to one or more constituent groups.

  - A **clique** is a set of nodes within an undirected network such that every member of the set if connected by an edge to every other. The occurrence of a clique in an otherwise sparsely connected network is normally an indication of a highly cohesive subgroup.
  - A ***k*-core** is a connected set of nodes where each on is joined to at least k of the others. The k-core is not the only possible relaxation of a clique, but it is a particularly useful one for the very practical reason that k-cores are easy to find.
  - A **component** in an undirected network is a maximal set of nodes, each with a path to each of the others. A useful generalization of this concept is the ***k*-component**. A k-component is a set of nodes such that each is reachable from each of the others by at least

k node-independent paths (i.e. paths that do not share any node but the source and the target ones). In particular, a 1-component is an ordinary component.

One disadvantage of k-components is that for $k \geq 3$ they can be non-contiguous. Sometimes, non-contiguous components are inappropriate to identify groups of nodes. For this reason, researchers introduced alternative grouping definitions: n-cliques, n-clans, k-plexes, and k-groups.

- An ***n*-clique** is a generalization of cliques that replaces the strong constraint of the complete and maximum sub-graph with the existence of a relationship between all the actors through a path of maximum length n.
- An ***n*-clan** is a restriction of an n-clique through the constraint that the longest path in the group is less than or equal to n. This corrects a defect in n-cliques that can form spurious groups by including neighbouring members that are literally closer to other groups.
- A ***k*-plex** is another generalization of cliques that accepts as a member of the group any node that has at least $n-k$ links with the other nodes, where n is the total number of nodes that make up the group. k-plexes generate many more smaller groups than the previous methods.

Regarding groups of nodes, one could also study transitivity and clustering coefficients:

- In mathematics, a relation $\mathscr{R}$ is said to be **transitive** is $a \mathscr{R} b$ and $b \mathscr{R} c$ together imply $a \mathscr{R} c$. In networks, if $\mathscr{R}$ is "connected by an edge" and $\mathscr{R}$ is transitive, one would have that "if a and b are connected and b and c are connected, then a and c are connected".

  Perfect transitivity holds in a network when network is a clique (and its graph is complete). **Partial transitivity** instead can indicate the tendency to extend that mission relation. For example, if a and b are friends and b and c are friends, that does not guarantee that a and c are friends, however it makes it likely.

  Transitivity if a property of triads that characterize different network structural configurations: **siolation** (when the triad is disconnected), **dyad** (when only two out of three nodes are in $\mathscr{R}$), **structural hole** (when the three nodes are in $\mathscr{R}$ except one dyad), **cluster** (when the triad enjoys perfect transitivity). Clusters are also called **closed triads** as they form two-edge long paths among the members of the triad, closed by a third edge.
- The **clustering coefficient** is the fraction of paths of length two in the network that are closed. That is, one counts all paths of length two, one also counts how many of them are closed, and then one divides the second number by the first to get a clustering coefficient C that lies in the range from zero to one:

$$C = \frac{\text{number of closed paths of length two}}{\text{number of paths of length two}} \tag{18}$$

  In particular, $C = 1$ implies perfect transitivity, while $C = 0$ implies no closed triads.
- While the clustering coefficient is a property of an entire network, it is also useful to define a local clustering coefficient C_i for a single node i:

$$C_i = \frac{\text{number of pairs of neighbours of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i} \tag{19}$$

The **local clustering coefficient** represents the average probability that a pair of nodes related to it by $\mathscr{R}$ are also in $\mathscr{R}$ with each other. Since for nodes with degree zero or one the number of pairs of neighbours is zero and C_i would not be well defined, by convention $C_i = 0$ for those cases.

Local clustering can be used as an indicator of structural holes in a network. Thus, local clustering can be seen as a type of centrality measure, where the smaller the values the more powerful the node.

- Local clustering is strictly linked to the concept of **redundancy**, whose definition R_i of a node i corresponds to the average number of connections from a neighbour of i to the other neighbours of i. The minimum possible value of the redundancy of a node i is zero, and the maximum value is $d_i - 1$, where d_i is the degree of the node.

- **Reciprocity**. This measure focuses on more closed groups than triads. For example, in a directed network one can have loops of length two and it would be interesting to ask about the frequency of occurrence of these loops. The frequency of loops of length two is measured by the reciprocity, which estimates how likely it is that two nodes point at each other. If there is a directed edge from node i to node j, and there is also an edge from j to i, then the two edges are reciprocated. Let m be the total number of edges in the given network, reciprocity is:

$$r = \frac{1}{m} \sum_{ij} A_{ij} A_{ji} \tag{20}$$

- **Similarity**. There are two fundamental measures of network similarity: structural equivalence and regular equivalence:

  - **Structural equivalence** is a count of the number of common neighbours two nodes have. In an undirected network, the number n_{ij} of common neighbours of nodes i and j is given by:

$$n_{ij} = \sum_k A_{ik} A_{kj} \tag{21}$$

However, focussing on the total number of nodes penalises nodes with low degree. The cosine similarity is a similarity that compounds varying degrees of nodes:

$$\frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{d_i}\sqrt{d_j}} = \frac{\overbrace{n_{ij}}^{\text{number of common neighbours}}}{\underbrace{\sqrt{d_i d_j}}_{\text{geometric mean of their degree}}} \tag{22}$$

There are alternative measures to cosine similarity. For example, the Jaccard coefficient is computed as:

$$J_{ij} = \frac{n_{ij}}{d_i + d_j - n_{ij}} \tag{23}$$

The Pearson correlation, which expresses the degree of linear association between two variables:

$$r_{ij} = \frac{\sum_k (A_{ik} - \langle A_i \rangle) \sum_k (A_{jk} - \langle A_j \rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i \rangle)^2} \sqrt{\sum_k (A_{jk} - \langle A_j \rangle)^2}} \tag{24}$$

where $\langle A_x \rangle$ is the average of the x-th row of matrix A. Lastly, the Hamming distance calculates the number of neighbours two nodes do not have in common:

$$h_{ij} = \sum_k (A_{ik} - A_{jk})^2 \tag{25}$$

– **Regular equivalence** of two nodes is the count of neighbours that are themselves similar. The basic idea is to define a similarity score σ_{ij} such that i and j have high similarity if they have neighbours k and l that themselves have high similarity. For an undirected network:

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} \tag{26}$$

with α constant and σ the leading eigenvector, as in the case of Katz centrality.

- **Homophily**, or **assortative mizing**. This measure is related to similarity and equivalence. It reports the tendency of nodes in the network to draw ties with other nodes that are similar/equivalent to them. In general, a network is assortative is a significant fraction of the edges in the network run between nodes of the same type. To measure the level of assortativity, one can calculate a) the fraction of edges that run between nodes of the same type and subtract from that b) the fraction of such edges one would expect to find if edges were positioned at random without regard for node type. Hence, this measure quantifies the level of non-randomness in the placement of edges in the network:

$$a) = \frac{1}{2} \sum_{ij} A_{ij} \delta_{g_i g_j} \tag{27}$$

where δ is the Kronecker delta function: $\delta_{kl} = 1$ if $k = l$ and 0 otherwise, and $g_i, g_j \in [1, N]$ represent the group/class/type of nodes i and j respectively. Then:

$$b) = \frac{1}{2} \sum_{ij} \frac{d_j}{2m} d_i \delta_{g_i g_j} \tag{28}$$

where $2m$ is the maximum number of ends of edges in an entire network, with m being the number of edges. Thus, given an edge with an end at i, the chance that the other end belongs to j is $d_j/2m$. Lastly, putting all together:

$$\text{a)} - \text{b)} = \frac{1}{2}\sum_{ij}\left(A_{ij} - \frac{d_j d_i}{2m}\right)\delta_{g_i g_j} \tag{29}$$

One can calculate assortative mixing also on networks with ordered characteristics, like age or income, which support the calculation of approximation of assortativity based on the distance between those characteristics. If network nodes with similar values of a scalar characteristic tend to be connected together more likely than those with different values, then the network is considered assortatively mixed according to the characteristic. To measure this, one relies on the covariance of the network. Let us have x_i being the value of attribute x for node i, then:

$$\text{COV}(x_i, x_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu_i^x)(x_j - \mu_j^x)}{\sum_{ij} A_{ij}} \tag{30}$$

Being:

$$\mu_i^x = \frac{\sum_{ij} A_{ij} x_i}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_i d_i x_i \tag{31}$$

then:

$$\text{COV}(x_i, x_j) = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right) x_i x_j \tag{32}$$

Assortativity with respect to the total number of edges is called **modularity**, denoted with Q, and it measures the extent to which similar nodes are likely to connect to each other:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta_{g_i g_j} \tag{33}$$

Lastly, it is sometimes convenient to normalize the covariance so that it takes the value 1 in a network with perfect assortative mixing (namely, a network in which all edges fall between nodes with precisely equal values of x_i):

$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right) x_i x_j}{\sum_{ij}\left(d_i \delta_{ij} - \frac{d_i d_j}{2m}\right) x_i x_j} \tag{34}$$

$r \in [-1, 1]$ is called **assortativity coefficient**. The higher this coefficient is, the more assortative the network is.

6 Measures and Metrics, Networks

Examples of measures and metrics on networks are:

- **The small-world effect**. A renowned, and measurable, network phenomenon is the small-world effect. Informally, one has a small-world effect when one can find shorter-than-expected distances between pairs of nodes. Mathematically, let d_{ij} be the length of the shortest path through a network between nodes i and j; then, the mean distance ℓ_i for a node i corresponds to:

$$\ell_i = \frac{\sum_j d_{ij}}{n} \tag{35}$$

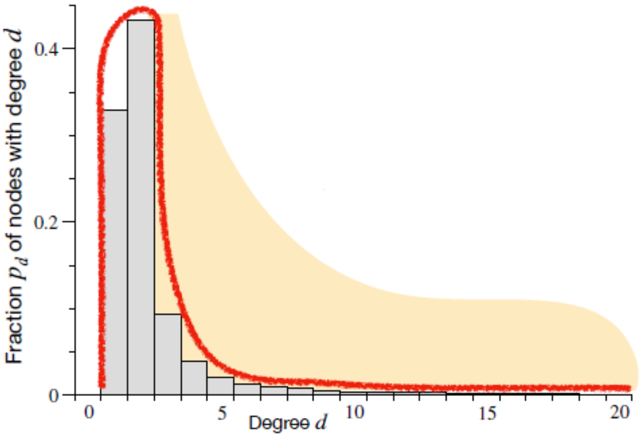
  and the mean distance for the whole network corresponds to:

$$\ell = \frac{\sum_i \ell_i}{n} = \frac{\sum_{ij} d_{ij}}{n^2} \tag{36}$$

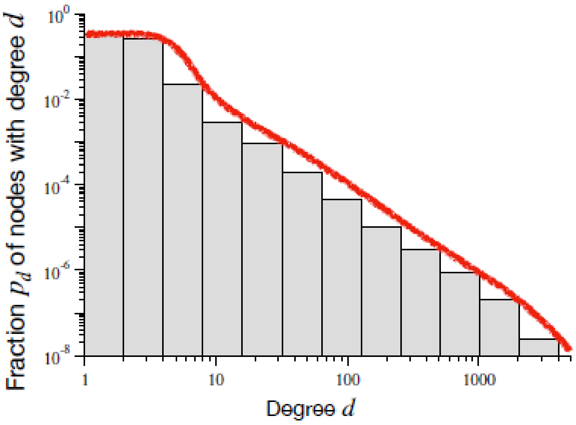
  A family of networks shows small-world effect when $\ell \propto \log n$. Small-world networks have: many highly-clustered groups (e.g. cliques) where all nodes are densely connected, hubs that serve as mediators to shorten the lengths between other edges, robustness to random perturbations thanks to the low hub-to-leat ratio.

- **Degree distribution**. Considering a network and letting p_d be the fraction of nodes that have degree d, then that ratio is essentially the probability of a given node to have that degree.

- **Power laws and scale-free networks**. Considering the degrees of a portion of the Internet and plotting the degree distribution of it, one can obtain the following bar-plot:



  From this plot one can see how most of the nodes in the network have a low degree. However, there exists a significant tail of nodes with substantially higher degree. When plotted in a log-log scale, power-law distributions tend to follow a straight-line behaviour:



Distributions of this kind are described by the formula:

$$\ln p_d = -\alpha \ln d + c \tag{37}$$

where α and c are constants that respectively modify the slope and normalize the curve of the distribution. By taking the exponential of both sides of the formula:

$$p_d = Cd^{-\alpha} \tag{38}$$

with $C = e^c$. Since the distribution is dependent on a power (with exponent α) of the degree d, it is called a **power law** distribution.

To detect power-law behaviours, one can use the cumulative distribution function, which is defined by:

$$p_d = \sum_{d'=d}^{\infty} p_{d'} \tag{39}$$

so that p_d is the fraction of nodes that have degree d or greater. One can get a precise measure of how close the given distribution approximates a power-law by calculating the value of α. Indeed, if $p_d = Cd^{-\alpha}$, then:

$$p_d = C \sum_{d'=d}^{\infty} d'^{-\alpha} \underbrace{\approx}_{\text{assuming } \alpha > 1} C \int_d^{\infty} d'^{-\alpha} \partial d' = \frac{C}{\alpha - 1} d^{-(\alpha-1)} \tag{40}$$

so that α becomes the exponent determining the distribution (on d) as:

$$\alpha = 1 + n \left(\sum_i \ln \frac{d_i}{d_{min} - 1/2} \right)^{-1} \tag{41}$$

Empirically, in power-law distributions $2 \leq \alpha \leq 3$.

Networks whose degree distribution follows a power-law behaviour are usually called **scale-free** networks. Scale-free networks are highly robust networks that can survive the failure of a sensible number of their nodes.

- **Distribution of other centrality measure**. One could also study the distribution of other centrality measures, such as the eigenvector centrality, the betweenness centrality, and the closeness centrality.

- **Cohesion and Connectedness**. This represents the likelihood of nodes being connected to each other. Notably, cohesion does not indicate social aggregation. The simplest measure of cohesion is **density** (i.e. the ratio between the number of ties in the network with respect to the total number of possible ties $n(n-1)/2$). To avoid the issue of comparing sensibly different networks over density alone, one can resort to a cohesion measure on the average degree of the network. This is obtained by calculating the average of the degrees of each node (i.e. the row sums of the adjacency matrix).

  Connectedness is a more sensitive measure of cohesion defined as the proportion of pairs of nodes that can reach each other by a path of any length. Or, alternatively, the proportion of pairs of nodes that are located in the same component. For directed, non-reflexive networks:

$$\frac{\sum_{i \neq j} r_{ij}}{n(n-1)} \tag{42}$$

  where $r_{ij} = 1$ when i and j are in the same component, 0 otherwise. Connectedness can be used in evaluating changes to a network either in reality or as part of a what-if simulation.

  A variation of connectedness is **compactness**. This weights the paths connecting nodes inversely by their length:

$$\frac{\sum_{i \neq j} d_{ij}^{-1}}{n(n-1)} \tag{43}$$

  with $d_{ij}^{-1} = 0$ when no path exists between i and j. Intuitively, compactness considers network cohesion as a measure of how easily things can flow through it, accounting also for disconnected components. For example, with compactness ≈ 1, nodes tend to be all connected and close.

- **Centralization and core-periphery indices**. Centralization refers to the extent a network is dominated by a single node. A maximally centralized network looks like a star: the node at the centre of the network has ties to all other nodes, and no other ties exist. A simple method for finding the core-periphery structure assumes that the nodes in the core have higher degree than the nodes in the periphery, and divide the nodes according to degree. Another method is to find the k-cores of the network, slicing the network into different, nested layers. A more refined method for detecting dichotomic core-periphery structures relies on finding the division into core and periphery, defined by a value g_i, such that it minimize a measure ρ that calculates the difference between the number of edges in the periphery and the expected number of such edges if placed at random:

$$\rho = \frac{\sum_{ij}(A_{ij} - p_{ij})g_i g_j}{2} \tag{44}$$

with:

$$g_k = \begin{cases} 0 & \text{if } k \in \text{core} \\ 1 & \text{otherwise} \end{cases} \tag{45}$$

and p_{ij} equal to the average probability of the same of edges, it placed at random.

- **Random graphs**. A random graph is a model network in which the values of certain properties of the network are fixed, but the network is, in other respects, random. One of the simplest examples of a random graph is the one where one fixes only the number of nodes n and the number of edges m. This model is often referred to by its mathematical name $G(n, m)$. More specifically, one can define a random graph model as a family of networks defined by a probability distribution:

  $$P(G) = \frac{1}{\binom{\binom{n}{2}}{m}} \tag{46}$$

  where $\binom{n}{2}$ represents the pairs of nodes between which one could place an edge, and the entire denominator represents the ways of placing the m edges.

  A special family of random graphs is that of $G(n, p)$, where one does not fix the number of edges but the probability of edges between nodes. $G(n, p)$ is the ensemble of simple networks with n nodes in which each network G appears with probability defined by the distribution:

  $$P(G) = p^m(1 - p)^{\binom{n}{2} - m} \tag{47}$$

  This family of graphs is important since it is capable to simulate the tendency of real-world networks to become sparser and sparser the larger they grow. Formally, their average degree grows slower than their size. This is also captured by the degree distribution of $G(n, p)$, which for smaller values of n follows a binomial Bernoullian distribution, while for larger values approximates a Poissonian distribution.

7 Testing Hypotheses

In quantitative studies, hypotheses are statistical hypotheses (i.e. they are testable by considering the observed data as values taken from a collection of random variables). The difference between the two models (the actual and the random one) is deemed statistically significant if, according to a threshold probability, the actual data is unlikely to have occurred randomly. The hypothesis that the two models are comparable is called **null hypothesis**. If one can reject (through testing) the null hypothesis, one has ground for believing that there is a dependence relationship between two or more phenomena represented by the data. This second hypothesis is called **alternative**.

The term **correlation** refers to the degree to which a pair of variables are linearly related. Correlations are useful because they can indicate a predictive relationship. To test the hypothesis of correlation, one measures a set of variables (e.g. two) on a set of cases drawn via a probability sample from a population and then calculate a correlation coefficient between the variables. If the correlation measure found in the sample appears in rare cases with respect to all possible random configurations, then one has a high degree of confidence in rejecting the null hypothesis (i.e. the hypothesis that the two variables are independent) and accept the alternative (i.e. that the two variables are correlated). The most common of correlation coefficients is **Pearson's coefficient**.

7.1 Testing Hypotheses in Network Analysis

Applying correlation tests over network data has some peculiarities, mainly because standard (i.e. statistical) tests make assumptions about the data which are violated by network data. An assumption such tests usually make is that observations are statistically independent, which, in the case of adjacency matrices, they are not. Another typical assumption of standard tests is that the variables are drawn from a population with a particular distribution, such as a normal distribution. However, many network population distributions are not normal. The main solutions to solve the problem of statistical testing of network hypotheses are two:

- Using statistical models specifically designed for studying the distribution of ties in a network.
- Using a generic methodology of randomization/permutation tests with modified standard methods, like regression.

In general, there are four main levels of analysis for network hypotheses:

- **Monadic** hypotheses, are the closest to non-network ones and the correlation stands between a characteristic of the node and another.
- **Dyadic** hypotheses consider relations among pairs (dyads) of nodes, and are normally organized as $N \times N$ matrices one correlates.
- **Mixed** hypotheses consider the relation between monadic and dyadic properties. Often-times these kinds of hypotheses are tested as dyadic by rephrasing the monadic property.
- **Network-level** hypotheses consider the structure of the network with respect to some variable, with the cases being whole networks instead of nodes.