Social Network Analysis

Discussion Paper

Claudia Zehnpfund

January 2005

Contents

1 Social Networks Analysis (SNA)

1.1 Basic Definitions

Over several decades researchers have either used the social networks study to investigate the patterns that emerge from the interaction among people belonging to various kinds of networks, e.g. kinship, work, friendships etc; or to identify the major actors or vertices in the network which influence the structure of the network in a number of ways. It is important to note that despite the fact that social networks research is described in terms of the repertoire of the measures, it has a long history of qualitative research. Social networks, may be regarded as important sub-discipline in the social sciences in broader terms. Hence studies in social networks range from pure qualitative research to development of sophisticated quantitative metrics that capture the macro-properties of the underlying network.

In order to grasp a firm understanding of mathematical network theory we will formulate some basic definitions in the first section. In the section after that we will discuss various network measures and their applications. Network measures are meant to be a tool through which we are able to analyze networks efficiently. Most importantly they enable us to compare different networks at a certain point of time or the development of the same network over time.

Definition 1.1 *A* NONDIRECTED GRAPH OR A GRAPH G *is an ordered pair* $G = (V, E)$*, which fulfils the following conditions:*

1. V *is a set of vertices or nodes*
2. E *is a set of unordered pairs of different vertices, called edges. The vertices* u *and* v *that belong to an edge* $e = \{u, v\}$ *are called vertices of the edge or start vertex and end vertex.*

V (and therefore E) are normally considered as finite sets and many well known results are not true for infinite graphs. Therefore we define V to be a finite set unless we declare it otherwise.

Definition 1.2 *A* DIRECTED GRAPH *is an ordered pair $G = (V, A)$ with*

1. *a set V of vertices*
2. *a set A of ordered pairs of vertices, called arrows. An arrow $a = (x, y)$ is considered to be directed from x to y; y is called the head and x is called the tail of the arrow.*

Definition 1.3 *A* WEIGHTED GRAPH *G is a graph in which each edge is associated with a specific value.*

Definition 1.4 *A* SUBGRAPH *$G = (V, E)$ of a graph $H = (V', E')$ is a graph with $V \subseteq V'$ and $E \subseteq E'$.*

Definition 1.5 *Let $G = (V, E)$ be a graph. A* PATH P *is a sequence of vertices $v_1, \ldots v_k$, such that for all $i \in \{1, 2 \ldots, k-1\}$ the condition holds that $\{v_i, v_{i+1}\} \in E$ and $v_i \neq v_j$ für $1 \leq i < j \leq k$.*
If at that the condition $\{v_k, v_1\} \in E$ holds, this path is called a CYCLE.
A DIRECTED PATH *is defined analogously to a path with the exception that a directed path includes at least one arrow. An* ORIENTED PATH *consists only of arrows.*
A MINIMAL CYCLE *is a cycle which chooses the shortest path in order to arrive again at the start vertex.*
The LENGTH *of a path is defined as the number of edges that the path has to traverse in order to go from the start vertex to the end vertex.*
A weighted graph associates a weight to each edge of the graph. The WEIGHT OF A PATH *in a weigthed graph is therefore the sum of all the weights of the traversed edges.*

Definition 1.6 *A graph $G = (V, E)$ is called* CONNECTED *if there is a path from u to v, $\forall\ u, v \in V$. A* CONNECTED COMPONENT *of a graph is a subgraph in which all vertices are connected.*
If the graph G is connected even after the removal of $k-1$ arbitrary vertices, G is called K-CONNECTED. *For a graph G to be k-connected there has to be k disjoint path between any two vertices in the graph.*

A DIRECTED GRAPH *is called* CONNECTED *if there is a directed path from any vertex to every other vertex. A directed graph is called* WEAKLY CONNECTED *if every vertex in the graph is connected with all the other vertices by directed paths where the arrows are not necessarily directed in the right direction.*

Definition 1.7 *Two vertices u and v are called* NEIGHBOURS *if $\{u, v\} \in E$. A* CLIQUE *of a graph $G = (V, E)$ is a set of pairwise neighbouring vertices.*

Definition 1.8 *Let $G = (V, E)$ be a graph. The* DISTANCE *$dG(u, v)$ between two (not necessarily different) vertices u and v is the length of the shortest path between them.*

1.2 Summery of selected network measures

There are numerous network measure and indices that can be used to analyse the efficiency of a network.
Among other applications they can be used to compare different networks at a certain point of time or to analyse the evolution of a network to different points of time.
Below we will present some network measures.

1. DEGREE: The degree of a vertex v is the number of edges that are incident with v. A vertex of degree 0 is called ISOLATED.

2. CLUSTERING COEFFIZIENT

   The CLUSTERING COEFFICIENT is a measure for the magnitude of the clustering in a graph. One distinguishes the local clustering coefficient for a certain vertex of the graph and the global clustering coefficient of the whole graph.
   The local clustering coefficient of a vertex v in a graph G denotes the ratio of the number of edges that are present between it's direct neighbours and the number of edges that could be present between it's neighbours.
   The global clustering coefficient is the average of the local clustering coefficients of all the vertices in the graph. Small-World networks have a very high global clustering coefficient compared with the clustering coefficient of a random graph, [1]. Scale-free networks have a small

characteristic pathlength and a high clustering as well. But the clustering in small-world networks is still more extreme than the clustering in scale-free networks. At that the clustering coefficient of small-world networks is independent of the size of the network, [3]. This feature is found in totally ordered lattices as well. In contrast to small-world networks, the clustering coefficient of scale-free networks is dependent on the size of the network. If the size of the network grows, the clustering coefficient of scale-free networks converges to 0, [3].

3. DENSITY: The density of a graph is the proportion of present edges to possible edges in the graph. A graph in which every vertex is connected with every other vertex is called complete. This graph has the maximal possible density 1.

   Definition 1.9 *Formally that means for a graph $G = (V, E)$ that the global clustering coefficient can be calculated via $\frac{|E|}{M}$ mit $M = \begin{pmatrix} |V| \\ 2 \end{pmatrix}$*

4. CHARACTERISTIC PATHLENGTH The characteristic pathlength of a graph is the average of the distance (the smallest pathlength) between two arbitrary vertices in the graph.

5. NUMBER OF COMPONENTS A graph $G = (V, E)$ ist divided in connected components. If G consists of only one connected component the graph G is called connected.

6. NUMBER OF CUTPOINTS Cutpoints are central points in the sense that they are the vertices that hold parts of the graph together that would not be connected if it were not for these certain vertices. Therefore they are the only vertices that have to be traversed by the paths from one part of the graph to the other.
   The concept of a cutpoint can be extended from one vertex to a set of vertices. If a set M of vertices is necessary to hold a graph together, then M is called a cutset. Analogously: Cutedge.

   Theorem 1.1 *A vertex w of a connected graph $G = (V, E)$ with $|V| \geq 3$ is a cutpoint of G if and only if there exist vertices u and v, distinct of w, such that w is on every path that goes from u to v.*

Theorem 1.2 *An edge $e = \{u, v\}$ of a connected graph G is a cutedge if and only if e does not belong to a cycle of G.*

7. DIAMETER AND RADIUS: The EXCENTRITY $EG(v)$ of a vertex v in a graph $G = (V, E)$ is the maximal distance of v to another vertex in the graph. (That means the maximum of the shortest pathlengths over all vertices in the graph.) The DIAMETER $diam(G)$ of a graph G is the maximal excentrity over all vertices of G and the RADIUS of G, $rad(G)$, is the minimal excentrity. If there are two components in G one defines $diam(G)$ and $rad(G)$ to be infinite. Vertices with maximal excentrity are called peripheral vertices and vertices with minimal excentrity form the "'center"' of the graph.

   Theorem 1.3 *A tree has at most two vertices that lie in the center.*

   The diameter enables us to measure the evolution of a network over time. The bigger the diameter the less connected the network tends to be. In the case of a complex graph one can calculate the diameter with the help of a topological distance matrix, which calculates for every pair of vertices the minimal topological distance.

   (E.g.: The definition of the Shimbel distance matrix can be seen at: http://people.hofstra.edu/geotrans/eng/ch1en/meth1en/shimbelmatrix.html)

8. REACHABILITY/CONNECTIVITY A measure for the connectedness of two vertices u and v in a network is the so called reachability or connectivity. In this context u and v are called

   (a) weakly connected if there is a path connecting them

   (b) unilateral connected if there is a directed path from u to v OR a directed path from v to u.

   (c) strongly connected if there is a directed path from u to v AND a directed path from v to u.

   (d) recursively connected it they are strongly connected and the path from u to v uses the same edges and arrows as the directed path from v to u.

9. CENTRALITY OF CERTAIN VERTICES OR OF THE WHOLE NETWORK
An important application of social network theory is the identification of "'important"' vertices in the network. There are a lot of different ways to define an actor (a vertex) as important or central. More often than not, vertices who are important are distributed at strategic important places in the network.

   (a) A vertex v is called CENTRAL if the degree of v is great in comparison with the other vertices of the network. We don't make a difference between incoming and outgoing edges at the moment.

   (b) One can also consider the DEGREE DEPENDENT CENTRALITY. One problem with the consideration of degrees is that they are dependent of the magnitude $g = |V|$ of the network. The maximal degree a vertex can achieve is $g - 1$. To standardize this measure one can define a degree dependent centrality $C_A(v) = \frac{degree(v)}{g-1}$. Then it is possible to compare the value $C_A(v)$ and therefore different vertices over networks of different magnitudes.
   A similar measure is the EGO-DENSITY where the degree of a vertex is divided by the maximal possible number of edges this vertex could possess. This measure is independent of the magnitude of the network as well.

   (c) One can expand the same concepts to a directed graph. Here one talks of HIGH PRESTIGE if a vertex has a lot of incoming edges compared to the other vertices in the network. The prestige of a vertex gets bigger if more arrows are pointing to it but normaly it doesn't change if it points new arrows to other vertices.

   (d) One can consider the centrality of the whole network in the sense that one can compare the centrality of all vertices in the network and finds out if the centrality of the vertices is distributed evenly or not. A star graph for example is an example for a graph where one vertex is very central and all other vertices are not important with regard to centrality.
   Let $C_A(n_i)$ be the degree dependent centrality (definition see under (b) above) of the vertex n_i. Furthermore let $C_A(n^*)$ be the maximum of the degrees over all vertices of the network, that means, $C_A(n^*) = max_i C_A(n_i)$, $i \in \{1, \ldots, |V|\}$.
   The sum of the differences between $C_A(n^*)$ and the centrality in-

dices of the other vertices, $\sum_{i=1}^{|V|} [C_A(n^*) - C_A(n_i)]$ is a network measure for the distribution of centrality of the vertices in the whole network.

(e) A further point of view to define the centrality of a vertex is via DISTANCE OR CLOSENESS. The idea is that a vertex is central if it can be in contact with all or many of the other vertices very quickly.
For that purpose let $d(n_i, n_j)$ be the number of edges of the shortest path from the vertex n_i to the vertex n_j. The absolute distance that the vertex n_i has from all other vertices is defined by $\sum_{j=1}^{|V|} d(n_i, n_j)$ with $j \neq i$.
As the centrality index of a vertex has to get smaller the greater the pathlengths to all other vertices, one defines the closeness index as:

$$C(n_i) = \left[\sum_{j=1}^{|V|} d(n_i, n_j)\right]^{-1}$$

If the vertex n_i is incident with all other vertices, that means if n_i thus has the greatest possible closeness, the value of $C(n_i)$ is $(g-1)^{-1}$. The minimal value approaches 0 asymptotically. The value 0 is assigned to a network if a vertex is not reachable by a path from n_i To make these values indepentend of the magnitude of the network, one multiplies $C(n_i)$ with $(g-1)$. Then the maximal value is 1 and the minimal still 0:

$$SC(n_i) = C(n_i) * (g-1)$$

Summarized that means that a vertex n_i with a standardized closeness value, $SC(n_i)$ of 1 is a vertex that is only one edge away from any other vertex in the network, i.e. the middle vertex of a stargraph. The nearer the standardized closeness value comes to 0 the farther away is the examined vertex from most of the other vertices.

(f) One can also define a closeness measure for the whole network. For that purpose one considers the standardised centrality index $SC(n_i)$ for all vertices n_i, $i \in \{1, \ldots, |V|\}$
The group centrality index is defined as

$$\sum_{i=1}^{|V|} [SC(n^*) - SC(n_i)],$$

where $SC(n^*)$ corresponds to the vertex with the greatest standardised centrality index.
It was shown in [5] that the maximal value of this formula is $[(|V|-2)(|V|-1)](2|V|-3)$ so that the group centrality index that is independent of network size is defined as

$$C_C = \frac{\sum_{i=1}^{|V|}[SC(n^*)-SC(n_i)]}{[(|V|-2)(|V|-1)](2|V|-3)}$$

This index is 1 if a vertex is incident to all other vertices and all other vertices have pathlength 2 to every other vertex. This is exactly the case of a star graph.

(g) Interactions between vertices are often dependent of other vertices that are *between* them. In certain cirumstances the vertices in-between can have influence on the vertices which would like to get in contact with one another. See for example the definition of cutpoints and cutedges.

Definition 1.10 *Let $G = (V, E)$ be a graph. A vertex $v \in V$ is called a* CUTPOINT *if the number of connected components in the graph that contain v is fewer than the number of components in the subgraph that results from deleting v from the graph. The set of all cutpoints is called* CUTSET.
A BRIDGE *or* CUTEDGE *similary is an edge, such that the graph containing this edge has fewer components than the subgraph that is obtained after the edge is removed.*

But cutpoints are not the only way one can look at a vertex as being central in regard to inbetweenness. Vertices are regarded as central if the probability that they are on the chosen path from one vertex to another vertex is high.
If one would like to calculate the probability that a certain vertex lies on the path between two other vertices one calculates at first the probability that a certain path is chosen. Let v and w be two vertices in the network. We consider all of the shortest paths between them. If there are several shortest paths we assume that any one of them is chosen with the same probability, e.g., to transport information from v to w. Let p_{vw} be the number of paths between v and w with the shortest length. The probability that

one of these paths will be used to get from v to w is then $\frac{1}{p_{vw}}$.
One calculates then the probability that a certain vertex j lies on the chosen path. Let $p_{vw(j)}$ be the number of shortest paths between v and w that lead over j. The probability that a path between v and w is chosen that leads over j is therefore $\frac{p_{vw}(j)}{p_{vw}}$. (In these calculations we always assumed that paths of equal lengths were chosen with the same probability.)
The index of the measures of betweenness centrality of a vertex v is then the sum over all calculated probabilities

$$C_B(v) = \sum_{j<k} p_{jk}(v)/p_{jk}$$

for $v \neq i, j$.

THIS MEASURE MEASURES THEREFORE HOW MUCH A VERTEX LIES *between* OTHER VERTICES. The minimum value is 0. This is the case if the vertex v lies on none of the shortest paths between any two vertices. The maximum value is $(|V| - 1)(|V| - 2)/2$, which coincides with the number of pairs of vertices which do not include v. The index reaches it's maximum if the vertex v lies on all shortest paths between any two vertices. The example for that is again a star graph with v as vertex in the center.
As the index is again dependent of the magnitude of the network we standardize as ususal:

$$SC_B(n_i) = C_B(n_i)/\left[(|V| - 1)(|V| - 2)/2\right].$$

Standardized in this way the indes SC_B now takes values between 0 and 1 and one can compare vertices from different networks.
The idea here is to define a vertex as central if he lies on the shortest paths between many other vertices.
In [10] it is describe why edges that connect two highly intraconnected clusters that are not connected otherwise are so especially interesting. They are the only bridge between those clusters (see definiton of cutedge above). For further details see Granovetters "'The Strength of Weak Ties, A Network Theory Revisited"', [6].
The definition and inspection of betweeness centrality still goes further, compare [9].

OTHER THAN SEVERAL INDICES LIKE E.G. THE CLOSENESS INDEX ONE CAN UTILIZE THIS INDEX EVEN IF THE NETWORK IS NOT CONNECTED. This is of course a big advantage and it leads to the fact that this index is the index that is used most frequently in social network analysis. Algorithms to find shortest paths and count how many vertices are in the respective paths are available and are for example implemented in UCINET.

Analogously one can define the GROUP INDEX OF BETWEENNESS CENTRALITY. This index allows to compare different networks in relation to their respective heterogenity of betweenness. It is defined as

$$\sum_{i=1}^{|V|} [C_B(n^*) - C_B(n_i)],$$

with $C_B(n^*)$ being the vertex with the biggest betweenness centrality index for the set of vertices of one social network.
The maximal value of the measure is $(|V|-1)^2(|V|-2)/2$, so that the standardized index for betweenness centrality of a group is defined as (compare [5])

$$C_B = \frac{2\sum_{i=1}^{|V|}[C_B(n^*)-C_B(n_i)]}{[(|V|-1)^2(|V|-2)]}.$$

In the above cited reference is shown that the index has it's minimal value 0 if all vertices have excactly the same betweenness value.

The betweeness theory can still be extended. Different assumptions can lead to different results. One does not, for example, have to consider each path to be chosen with the same probabilty and one could also consider other paths than just the shortest.
At that we did not go into much detail in the theorie of directed networks. Many of the above mentioned measures can be used by directed graph or can be defined analogously.

10. The STRUCTURAL BALANCE THEORY is based on the studies of Fritz Heider concerning the perception of social situations. It was extended with the analysis how the opinion and attitude of a single person fits or does not fit together with the opinions and attitudes of persons

in a group. In doing so edges between persons get the weight + or - according to the consensus between them. A group is defined to be structural balanced if for every two vertices whose edges are both weighted with + agree in the evalution of all other vertices. If e.g. two persons like each other (+) then they should evaluate all other actors (vertices) in the same way. It can be shown that the vertices in a structural balanced graph can be divided into two subgroups in such a way that all edges between them are either all positive or all negative. Structural balance was used in many application for example in the examination of international connections. For further details see [9] chapter 6.

11. Especially interesting in network theory is the identification of COHESIVE SUBGROUPS in a network. That means subgroups of vertices/actors between whom there are relatively strong, direct or positve ties/edges. As an first example one could remember the definition of a clique, but one easily understands that this definition is of low impact for real world applications as there are nearly no cliques in any real networks.
    A possibility to define a cohesive subgroup is the n-clique. A n-clique is a subgroup of vertices between there is a path of length $\leq n$.
    But n-cliques have two big disadvantages: First of all, the diameter of the subgraph of a n-clique can be greater than n and secondly the n-clique does not have to be connected. These problems arise because there was no requirement in the definition for the path between two vertices to use only vertices that are part of the n-clique. Therefore a path could be used that is partly outside the n-clique and so possibly raises the diameter of the n-clique. Under special circumstances it could then happen that the n-clique is not even connected. Therefore, n-cliques are not as cohesive as we would like them to be.
    This problem is solved with the definition of n-clans and n-clubs. An n-clan is a subgraph where all n-cliques of a graph are identified and then those are thrown out which have a diameter greater than n. The collection of the remaining subgraphs is called n-clan.
    An n-club is an n-clique in which the shortest pathlength between any two vertices has to be smaller than n for a path that is entirely inside the subgraph.

12. ALPHA INDEX α: Let $G = (V, E)$ be a connected graph. Another measure for the magnitude of connectivity inside the graph G is the α-index, who divides the number of minimal cycles in a graph through the number of maximal possible minimal cycles. The greater the alpha index the more interwoven is the network. Trees have an alpha index of 0. The value 1 stands for a completely connected network, a clique. This index is interesting because it measures the degree of connectivity independently of the number of vertices in the graph.
The following formular holds:

$$\alpha = \frac{C}{2|V|-5},$$

where C is defined as the number of minimal cycles in the graph. In the following figure one can see four networks with their respective α-indeces:
Network A has α index 0

Network B has α index $\frac{1}{3}$

Network C has α index $\frac{2}{3}$

Network D has α index 1

2 Possibilities of calculations with matrices

A possibility to visualize a network is the construction of a SOCIOMATRIX OR ADJACENCY MATRIX. This matrix indicates if two vertices are neighbours or not. There is a row and a column for each vertex of the network. The entry in the matrix is 1 if there is an edge between these two vertices and 0 otherwise. In a directed or weighted graph the entries are a bit different. In a weighted graph the entry is the weight of the respective edge and in a directed graph there is an entry 1 in the field x_{vw} if and only if an arrow is directed from v to w.
Sociamatrices in nondirected semantic relationships are symmetrical.
In order to analyze networks one can use some basic matrix operations: One

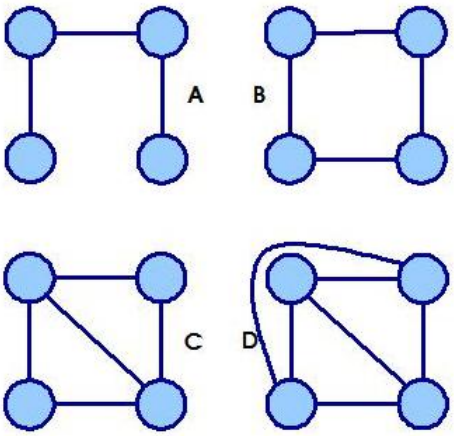

Figure 1: Four networks with their respective α-indices

can build the transpose, the inverse of a matrix, one can add, subtract and multiply...
E.g. matrixmultiplication can be used to calculate paths and reachability in networks, compare [9], chaper 4
The shortest pathlength between two vertices can be calculated with the multiplication of the sociomatrix with itself, compare [9] (chapter 4, pages 160-163).
The degree of a vertex can be calculated with the sociomatrix as well. The degree is equal to the sum of a column or the row of the respective vertex. That means:

$$\text{degree } (n_i) = \sum_{j=1}^{|V|} x_{ij}.$$

Analogously one can identify the degree of incoming and outgoing edges. [9] (chapter 4, pages 161-162).
The density of a graph can be calculated as the sum of all entries of the matrix, divided by the possible number of entries:

$$\text{Density}(G) = \frac{\sum_{i=1}^{|V|} * \sum_{j=1}^{|V|} x_{ij}}{|V|(|V|-1)}.$$

For further details see [7] and [9].

References

[1] Albert, R. and Barabasi, A. (2002) Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* 74, 47-97.

[2] Barabasi, A. (2003). *Linked*. Plume.

[3] Barthélémy, O., 2003, "'The impact of the model structure in social simulations"', in Proceedings of the 1st European Social Simulation Association (ESSA) Conference, Groningen; also available as Centre for Policy Modelling Report No. CPM-03-121, http://cfpm.org/cpmrep121.html.

[4] Freemann, L.C., et al (1979). Centrality in social networks: I. Conceptual clarification. *Social Networks*. 1, 215-239.

[5] Freemann, L.C., et al (1980). II. Centrality in social networks: 2.Experimental results. *Social Networks*. 2, 119-141.

[6] Granovetter, M., "'The Strength of Weak Ties, a Network Theory Revisited"' *American Journal of Sociology*, Volume 78 (1973), 1360-1380.

[7] de Nooy, W. et al (2004). *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

[8] Semple S., and Steel, M. (2003) *Phylogenetics*. Oxford University Press.

[9] Wassermann, S. and Faust K. (1994). *Social Network Analysis*. Cambridge University Press.

[10] Watts, D.J., 2003, Six Degrees: The Science of a Connected Age, W. W. Norton & Company.

[11] Watts, D.J., (1999). *Small Worlds*. Princeton University Press.

[12] Ziegler, G. (2003) *Lectures on Polytopes*. Springer Verlag Berlin.