4 Analysis

You will now learn about methods that help to make sense of network data. There are two essential steps here: first, the specific functioning of a network is represented in a dyad index that assesses the direct and indirect relations between pairs of actors; second, such dyad indices are used to identify influential positions (centrality) and groups of cohesive (clustering) or similar (roles) actors. In addition, we will take a brief look at models of networks that can be used to explain or predict structural features, also in relation to actor attributes.

Having collected and stored data on the phenomenon of interest, we are ready to make sense of it.

Generally speaking, the purpose of analysis is to extract from input data information that can be used to describe, explain, or predict situations or developments. In principle, there is nothing special about *network* data analysis as, like other statistical analyses, it merely involves the attempt to make distinctions and find associations between quantities. The difference, however, lies in the definition (and thus computation) of the quantities involved. Since networks yield relational data, concepts that differ from those used in the analysis of conventional data tables are available—and required.

This chapter provides a structured overview of the formal toolbox of network analysis (Brandes and Erlebach 2005). While details are provided for the most commonly applied methods, information is also provided on where to look for additional methods, should they be required for a particular study.

Our basic data format is a single network of actors who may have additional attributes. Many methods extend to more general situations, how-

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

ever. Although not entirely accurate, the following linkage between tasks and methods provides a reasonable initial approximation and may serve as a guide through this chapter.

Description: The structure of networks can be summarized using various indicators that frequently constitute quantitative assessments of variance in the structural positions of actors and their composition. Beyond their descriptive function, such indicators are often related to other quantities when the objective is to spot network effects or determinants.

Explanation: The structure of observed networks is commonly explained using network formation models. This is done, for example, by devising a generative model that yields networks with structural characteristics similar to the observed networks, or by fitting a parametric model to an observation.

Prediction: If only determining factors such as a preceding network configuration are known, one way to predict the formation or evolution of a network is simulation. Simulation, like fitting, is a specific use of network models.

Network data types are described in more detail in the following section; the subsequent sections are organized according to analytical interests.

It should be noted that we do not distinguish between complete and ego network analysis in this chapter because most methods are relevant to both cases. In fact, the most important differences do not show up during evaluation but during data collection and interpretation. Nevertheless, it should be noted that the methods most relevant for ego network analysis are network characteristics as described in Section 4.2.2.

4.1 Mathematical Representation

This chapter appears to be more formal than previous ones because of its more extensive use of mathematical symbols and formulas. If you are not familiar with symbolic notation, please observe that it is not only used for the purpose of brevity, it often offers a convenient way of attaining an appropriate level of precision. While many books attempt to provide less mathematical treatments, these often result in long-winded and more complicated verbal statements that ultimately generate greater confusion because it is difficult to avoid ambiguity using natural language. Note also

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Box 17: Some Mathematical Notation

Symbol $\in$ is shorthand notation for "contained in," i.e., $x \in S$ denotes that x is an element in set S. For two sets S and T, $S \subseteq T$ indicates that every element of S is also an element of T, i.e., S is a subset of T. Similarly, $S_1 \cup S_2$ and $S_1 \cap S_2$ denote the union and intersection of sets S_1, S_2. The number of elements in a set S, i.e., the set's cardinality, is denoted by $|S|$.

The sum over all elements x of a set S is abbreviated as $\sum_{x \in S} x$. The binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ gives the number of k-element subsets of a set with n elements, where $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$ denotes the nth factorial.

For a range of values R, $R^{n \times m}$ denotes the set of all matrices with n rows, m columns, and entries from R.

The type of a function f is declared as $f : D \to R$, meaning that f maps elements from a domain D to values in a range R.

that, although we use symbolic notation for clarity, we do not practice mathematical derivations.

In previous chapters, we used graphical and matrix notations to represent the structure of a network. For many of the methods introduced in this chapter, it is more convenient to use the combinatorial representation of a graph defined below.

The matrix representation may be most familiar to you as this is how network data is entered into spreadsheets. Each cell is indexed by a row index and a column index. If both rows and columns are labeled with actors of a (one-mode) network, the matrix is square and each cell can be referenced by specifying a row actor i and a column actor j. The corresponding cell content, a_{ij}, describes the relationship between i and j. For a non-valued relation, i.e., one that is either present or absent, the conventional choice is $a_{ij} = 1$ if the relation is present, and $a_{ij} = 0$ otherwise. Two-mode networks yield rectangular matrices, in which the rows and columns are labeled with the elements of the two distinct modes, however, the meaning of an entry is analogous.

For one-mode networks there is a second entry, a_{ji}, indexed by the same two actors but in the opposite order. If the relation coded in a matrix is symmetric, as in being friends, then all pairs of entries a_{ij} and a_{ji}

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

graphical	matrix	graph
i ● ● j	$a_{ij}=0, a_{ji}=0$	$(i,j)\notin E, (j,i)\notin E$ resp. $\{i,j\}\notin E$
i ●⟶● j	$a_{ij}=1, a_{ji}=0$	$(i,j)\in E, (j,i)\notin E$
i ●⟵● j	$a_{ij}=0, a_{ji}=1$	$(i,j)\notin E, (j,i)\in E$
i ●⟷● j	$a_{ij}=1, a_{ji}=1$	$(i,j)\in E, (j,i)\in E$
or i ●—● j		resp. $\{i,j\}\in E$

Figure 17: Equivalent representations of the four possible dyad configurations in an ordered (or unordered) boolean relation.

will be equal. The coding is thus redundant and the matrix is also referred to as symmetric. This need not be the case if the relation is potentially asymmetric, as in one actor considering the other to be a friend. For a non-valued relation this yields the four possible combinations, two symmetric and two asymmetric ones, for every pair of actors. We refer to a pair of actors such as i and j, irrespective of their relation, as a *dyad*. By convention, potentially asymmetric relations are coded in such a way that they extend from the row actor to the column actor. The four possible configurations thus suggest the graphical notation depicted in Figure 17.

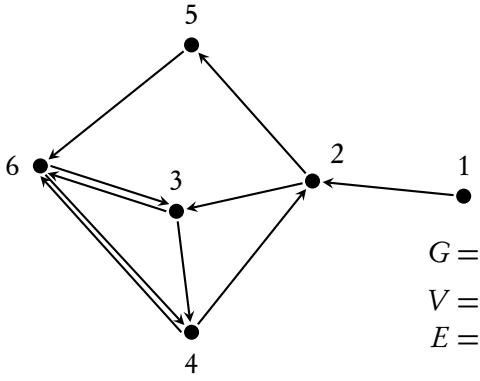
4.1.1 Graphs

Matrix representations are useful for data management and certain calculations, and graphical representations are a convenient means of communication. For most of the structural considerations that we will be dealing with in this chapter, however, the combinatorial representation of a graph is more appropriate.

A *graph* G is a pair $G=(V,E)$ that consists of a set V of *vertices*, and a set E of *edges*. While the elements of V represent the actors of the network, the ties between them are represented in E. Therefore, an edge is simply a pair of vertices, however, since the relation may be potentially asymmetric, the pair may have to be ordered. Hence, if the relation between a pair of vertices $i,j\in V$ is potentially asymmetric, the order of i and j in the pair is used to resolve potential ambiguities analogous to the matrix entries. In this case we say that the edges are *directed* and de-

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



$G = (V, E)$ where
$V = \{1, 2, 3, 4, 5, 6\}$ and
$E = \{(1,2), (2,3), (2,5), (3,4), (3,6), (4,2), (4,6), (5,6), (6,3), (6,4)\}$

Figure 18: A graph with $n = 6$ vertices, $m = 10$ directed edges, and $6 \cdot 5/2 = 15$ dyads.

note them using parentheses $(i, j) \in E$. Directed edges are thus a subset $E \subseteq V \times V = \{(i, j) : i, j \in V\}$ of all possible pairs of vertices, where pairs (i, j) and (j, i) are considered to be different edges corresponding to the two matrix entries a_{ij} and a_{ji} if $i \neq j$. Otherwise, edges are referred to as *undirected* and denoted using curly braces $\{i, j\} \in E$. The set notation indicates that $\{i, j\}$ and $\{j, i\}$ denote the same edge.

Since every undirected edge is represented equivalently by two directed edges (one for each ordering), directed graphs are more general. For most purposes in this chapter, an undirected graph can be thought of as a symmetric directed graph. If not stated otherwise, we assume in this chapter that edges are directed, although it should be noted that most definitions apply analoguously to undirected (symmetric) relations. Also, we assume that there are no *loops* (i, i) connecting a vertex $i \in V$ with itself, i.e., no diagonal elements in the matrix representation.

We will use $n = |V|$ to denote the number of vertices, and $m = |E|$ to denote the number of edges throughout.

The set of all dyads in a graph $G = (V, E)$ is $D(G) = \{\{i, j\} : i \neq j \in V\}$. Consequently, a graph with n vertices has $\binom{n}{2} = \frac{n(n-1)}{2}$ dyads. In an undirected graph without loops, there can be at most one edge per dyad, so that $E \subseteq D(G)$ and thus $m \leq \frac{n(n-1)}{2}$. In a directed graph without loops, there can be two edges per dyad (one in each direction), so that $m \leq n(n-1)$. In either case, the number of edges is at most quadratic in the number of vertices.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Just as numbers are abstracted from the objects being counted to allow for more general statements about quantities, vertices and edges are abstracted from actors and their ties to allow for more general statements about structures. It should be noted, however, that all other information specific to these entities is ignored deliberately. If vertices or edges represent qualitatively different actors or ties, vertex or edge attributes are needed for differentiation.

There is a straightforward correspondence between matrix and graph representations. The type of matrix we have been using so far has $n \times n$ entries and is called the *adjacency matrix* $A(G) = (a_{ij})_{i,j \in V}$ of G, where

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise .} \end{cases}$$

An alternative representation based on membership of actors in dyads is the $n \times m$ *incidence matrix* $B(G) = (b_{ie})_{i \in V, e \in E}$ with

$$b_{ie} = \begin{cases} 1 & \text{if } e = (i,j) \in E \text{ for some } j \in V \\ -1 & \text{if } e = (j,i) \in E \text{ for some } j \in V \\ 0 & \text{otherwise .} \end{cases}$$

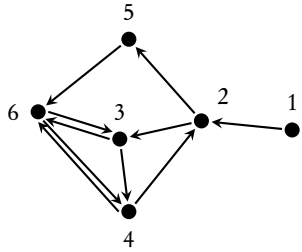
In an incidence matrix, each column has two entries, because every column corresponds to one edge, and the sign indicates its direction. Since symmetric relations would yield two columns per dyad that differ only in their signs, undirected graphs are frequently represented with one column per edge and no signs, i.e., entries are either 0 or 1, but not -1.

Please note that relations involving more than two actors can be represented in an incidence matrix but not in an adjacency matrix. This becomes more obvious when incidence matrices are interpreted as *two-mode* matrices, in which actors are the row-mode and their relations the column-mode. Two-mode networks are introduced in Section 4.1.3 below.

A graph $G' = (V', E')$ is a *subgraph* of a graph $G = (V, E)$, $G' \subseteq G$, if $V' \subseteq V$ and $E' \subseteq E$. Note that G' is required to be a graph, i.e., we can only have those edges in E' that connect vertices present in V'. By $G[V']$ we denote the unique *vertex-induced* subgraph of G that contains the vertices in V' and all edges that connect vertices of V' in G. Likewise, the *edge-induced* subgraph $G[E']$ contains the edges in E' and precisely those vertices that are incident to any edge of E' in G.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



(a) diagram

A	1	2	3	4	5	6
1	0	1	0	0	0	0
2	0	0	1	0	1	0
3	0	0	0	1	0	1
4	0	1	0	0	0	1
5	0	0	0	0	0	1
6	0	0	1	1	0	0

(b) adjacency matrix

B	(1,2)	(2,5)	(5,6)	(2,3)	(4,2)	(3,4)	(3,6)	(6,3)	(6,4)	(4,6)
1	1	0	0	0	0	0	0	0	0	0
2	−1	1	0	1	−1	0	0	0	0	0
3	0	0	0	−1	0	1	1	−1	0	0
4	0	0	0	0	1	−1	0	0	−1	1
5	0	−1	1	0	0	0	0	0	0	0
6	0	0	−1	0	0	0	−1	1	1	−1

(c) incidence matrix

Figure 19: Three alternative representations of the same graph.

4.1.2 Ego-Centered Networks

As discussed in Chapter 2, ego-centered networks comprise a mixture of network and population studies as a population of actors (the *egos*) is studied in terms of their (individual) social embeddings and not in terms of the relations between them. However, their embeddings are captured by relations with and between other actors (the *alteri*) on a per-ego basis.

Depending on the specific data available about the environment of each ego, ego networks require different representations. The following are typical scenarios (cf. also Section 3.1.3):

– The embedding of a person in his or her social environment is described by a collection of numerical or categorical values. While these values may characterize the network among alteri, it is possible that this network was never determined explictly because the values were surveyed directly. For example, the respondents may have been asked to rate the diversity of their personal networks without providing any details about it. Either way, there is no graph involved in the subsequent analyses.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

- Knowing the relationship of ego to some alteri, but not those among alteri, yields the intermediate case of ego networks consisting only of ego-alter dyads. Although these form a (star-like) graph, they are usually represented as a list of ego attributes.
- If data about the relationships between alteri is available, the resulting personal networks can be treated like any other complete network. In this case, the relationships between ego and each alter are best represented in an alter attribute.

Despite the fact that the first two scenarios are grounded in a structural perspective, in general, the data analyzed are not relational. From a technical point of view, ego network studies using such data are more akin to population statistics. In the third scenario, however, genuine relational data is observed, if only between ego and alteri as well as between alteri but not among egos. When a collection of networks based on the same type of relation consisting of one network per ego is obtained, they are sometimes referred to as a *personal network ensemble*.

4.1.3 Two-Mode Networks

Sometimes the relation between actors is, in fact, established indirectly via entities that are not considered as actors in the same sense. In studies of interlocking directorates, for instance, relations between directors are derived from joint board memberships, i.e., from relations involving two types of entities, directors and boards.

Networks consisting of two categorically different types of entities and a relation involving one member of each type are called *two-mode networks*. As discussed in Chapter 3, they can be represented in rectangular matrices $B \in \{0,1\}^{n \times m}$ in which one type of entity (e.g., actors) indexes the n rows, and the other type (e.g., social settings) indexes the m columns. Let an entry b_{ie} in that matrix be 1, if entity i of one type is tied to entity e of the other type, and 0 otherwise. Matrices of two-mode networks then generalize regular incidence matrices to the case of *undirected hypergraphs* because a column corresponds to a (hyper)edge, which can have any number of incident vertices rather than just two.

The analytical procedures available for two-mode networks (Freeman and White 1993; Faust 1997; Borgatti and Everett 1997; Doreian, Batagelj, and Ferligoj 2004) are less developed than those available for single-mode networks. Therefore, two-mode networks are often transformed into

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

weighted undirected graphs via an operation called *projection*. Projection to the row-mode or column-mode yields adjacency matrices

$$A^{row} = BB^T \quad \text{or} \quad A^{col} = B^T B\,,$$

where B^T is the matrix *transpose* of B in which row and column indices are swapped, i.e., the entries are related by $b^T_{ij} = b_{ji}$. Hence, an entry $a^{row}_{ij} = \sum_{e=1}^{m} b_{ie} b^T_{ej} = \sum_{e=1}^{m} b_{ie} b_{je}$ of the $n \times n$ row-mode projection corresponds to the number of entries that rows i and j of B have in common.

Projection to the row-mode or column-mode is generally irreversible because several different two-mode networks may yield the same projection. Whenever possible, you should try to operate on the two-mode data directly, also because projections tend to be dense and thus more computationally demanding in subsequent analyses. This is because every entity of degree d in one mode induces a complete subgraph of d vertices (and thus $d(d-1)/2$ edges) in the projection to the other mode.

An example of data with more than two modes are the *meta-networks* of Carley (2002). These are collections of one-mode and two-mode networks that result from combinations of several modes. For example, people P, organizations O, and locations L may form PP, PO, PL, OO, OL, and LL networks.

4.2 Indexing and Grouping

Numerous structural indicators have been devised to describe and analyze, for example, the particular structure of a given network, the relative position of actors in it, and the decomposition into meaningful groups of actors.

We will not attempt to enumerate them all. Instead, we will cover the most common methods for the most common analytic tasks in a systematic way. We will also provide pointers to more sophisticated and less frequently used methods, where appropriate. The crucial descriptive tasks addressed in this section are summarized in the following three guiding questions:

- What are the overall characteristics of the network?
- Who are the most influential actors?
- Which actors form meaningful groups in the network?

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

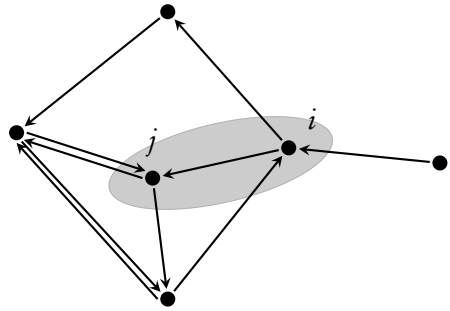


Figure 20: Dyad formed by vertices i and j. While being adjacent, the dyad is not mutual.

What distinguishes network analysis from other, more commonly applied techniques for empirical data analysis is the inherently relational nature of the data. Consequently, an analysis is often based on (possibly derived) relations between actors and the units of analysis are dyads rather than actors.

Hence, we begin this section with indicators that allow the quantification of direct and indirect pair-wise relations of various kinds. While more elaborate indicators are defined in the subsequent sections, in most cases, they only involve the aggregation of the elementary indices presented next.

4.2.1 Dyads as the Unit of Analysis

A dyad is a pair of actors that may or may not be linked by a tie. Dyads are thus represented by pairs of vertices, irrespective of whether these are connected by an edge or not. Dyads are the basic unit of analysis for social networks as other methods are essentially based on the aggregation of dyadic information.

Since most other indicators are derived from dyad indices, the choice of an appropriate dyad index is essential and should be guided by theoretical reasoning about the nature of the relationships between dyad members. For most of the indices discussed below, we will provide examples of instances in which they may be appropriate.

Since a dyad index is supposed to capture the nature of a relation it is closely tied to the type of network under examination. The method of aggregation, on the other hand, determines how positions and composi-

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

tions are constructed, and is therefore more closely tied to the interest we have in a network.

Elementary Dyad Indices

Simple examples of dyad indices are derived immediately from the presence or absence of edges in a graph $G = (V, E)$. The vertices i, j of a dyad are called *adjacent*, if $(i, j) \in E$ or $(j, i) \in E$, and a dyad is *mutual* or *symmetric*, if $(i, j) \in E$ and $(j, i) \in E$. In terms of elements of the adjacency matrix, these concepts can be expressed equivalently as

$$\begin{aligned} \text{adjacent:} \quad & a_{ij} + a_{ji} > 0 \\ \text{mutual:} \quad & a_{ij} + a_{ji} = 2\,. \end{aligned}$$

Please note that there is no difference between these two conditions in the case of symmetric relations, and that the vertices in a mutual dyad are necessarily adjacent.

Many other quantities can be derived directly from relations between the two actors of a dyad. If data on multiple relations are available, for instance, one can define a *degree of multiplexity* in terms of the number of ties in any of these relations for the same dyad. Likewise, *valued* relations, such as interaction *frequency*, immediately yield a dyad index. An example of a dyad index derived from non-network data is *dissimilarity* based on vertex attributes.

The following are indices that take into account the entire structure, in which a dyad is embedded.

Distance

To define a (structural) concept of distance between two actors, it is necessary to specify how actors can reach each other. Therefore, we define a (directed) *path* in a (directed) graph $G = (V, E)$ as a sequence of edges such that the endpoint of one is the starting point of the next.

Let $s, t \in V$ be any two vertices. Then a sequence of edges (i_0, i_1), $(i_1, i_2), \ldots, (i_{k-1}, i_k) \in E$ with $i_0 = s$ and $i_k = t$ is called a *(directed) st–path*. A sequence of edges that forms an *st*–path after reversing any number of them is called an *undirected st*–path. A path is called *simple*, if no vertex (and thus no edge) is contained twice.

If an *st*–path exists, then t is *reachable* from s. If t is reachable from s and vice versa, then s and t are said to be *(strongly) connected*. They are

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

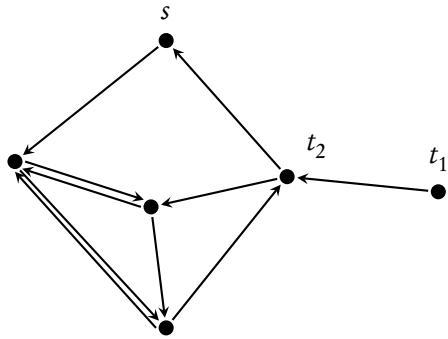


Figure 21: While vertex t_1 is not reachable from s, t_2 has (shortest-path) distance 3 from s. Hence, $d_G(s,t_1)=\infty$ and $d_G(s,t_2)=3$). The graph is weakly connected but has two strongly connected components, one of which consists only of t_1.

weakly connected, if at least one undirected *st*–path exists, otherwise s and t are *disconnected*. The inclusion-maximal subgraphs in which all pairs of vertices are strongly or weakly connected are called the strongly or weakly connected *components*. Note that there is no difference between strong and weak connectedness in undirected graphs.

Connectedness and reachability are thus relations (of indirect linkage) derived from the elementary relation of adjacency (i.e., direct linkage), where only connectedness is necessarily symmetric. They can be valued by considering the *length* of paths, i.e., the number of edges in the sequence. Alternatively, if the edge-relation is already valued, the edge values can be aggregated into path values. The most common and intuitive example is a network of locations in which edges are valued by spatial distance. The length of a path is then the sum of the edge lengths, i.e., the total distance along its edges. Graphs without edge values are often treated as a special case of valued graphs, in which all values are uniformly one. In this case, the sum of edge values and the number of edges coincide so that the former is a proper generalization of the latter.

The minimum length of an *st*–path is an essential dyad index, and called (*shortest-path* or *geodesic*) *distance*. It is the lowest number of edges needed to go from s to t, and denoted by $d_G(s,t)$ or simply $d(s,t)$ if it is clear which graph G is referred to. See Figure 21 for an example.

Note that higher-level indices based on shortest-path distances thus embody the assumption that connections between actors are established

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

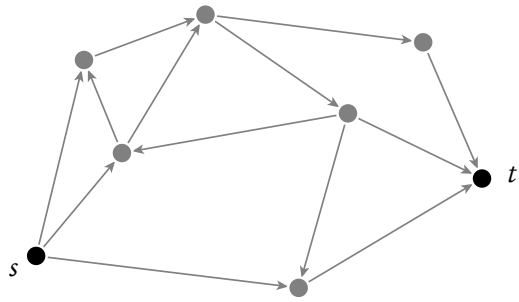


Figure 22: The dyad $\{s,t\}$ has lower vertex connectivity than edge connectivity, $\nu(s,t)=2<3=\lambda(s,t)$.

along the shortest possible routes. If the relation studied does not grant this assumption, alternative notions of distance may be applicable. For example, the current in an electrical network spreads along paths of any length, but in an energy-efficient way with more current flowing along paths of less resistance. Since this is determined by the resistance in each wire and the way that wires are joined, the effective resistance that needs to be overcome between two network nodes is also a measure of distance derived from aggregated edge values. It appears that current flow is a reasonable model, e.g., for the spreading of rumors.

It is thus crucial to be aware of the implicit assumptions that arise with a particular spreading model. We will return to this issue when discussing centrality indices.

Connectivity

While distances generally evaluate the difficulty actors have in reaching other actors, a *connectivity* index quantifies the ease or certainty of reaching them. If a dyad is connected via two independent paths, this can be interpreted as an indication that reachability is more reliable.

The two most common concepts of connectivity are *vertex connectivity* and *edge connectivity*. A dyad $s,t \in V$ is called *k-vertex-connected*, or simply *k-connected*, if there are at least k *st*–paths that do not share a vertex other than s and t. Equivalently, at least k vertices must be removed to cut all *st*–paths. The maximum k for which s and t are k-connected is the *(vertex) connectivity*, $\nu(s,t)$, of s and t.

If connectivity is not intercepted at vertices but at edges, we are inter-

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

ested in the *edge connectivity*, $\lambda(s,t)$, of s and t. It is defined analoguosly with $s,t \in V$ being *k-edge-connected*, if there are at least k *st*–paths that do not share an edge. Equivalently, at least k edges must be removed to cut all *st*–paths.

It is plausible (and can indeed be proven) that $\nu(s,t) \leq \lambda(s,t)$ for every dyad $s,t \in V$ of every graph $G = (V,E)$. An example in which strict inquality holds is provided in Figure 22.

Connectivity indices are employed, e.g., in applications such as communication network design where tolerance to network failure is important. However, they are also used to define concepts of cohesion, and will hence be revisited in the subsequent section on grouping.

As for distances, there are numerous ways of defining connectivity. The amount of current flowing between two nodes in an electrical network is inversely related to the resistance between them and thus a notion of connectivity. Similarly, there is the concept of network flow, that is not subject to resistance but capacity constraints on the edges.

Embeddedness

The final class of dyad indices considered here does not evaluate the nature of connections between two actors but compares the connections they maintain with others.

In the simplest of all cases, two vertices have exactly the same neighbors (possibly excluding themselves). Since their network positions are indistinguishable, any pair of such vertices is called *structurally equivalent*. Note that this new dyadic relation is reflexive (a vertex always has the same neighbors as itself), symmetric, and transitive (if a vertex has the same neighbors as another, and this second one has the same neighbors as a third vertex, then the first and third necessarily also have the same neighbors), and therefore induces a partition into classes of equivalent vertices. It is thus a method of grouping actors as discussed in Section 4.2.5 below.

An important weaker relation called a *Simmelian tie* is defined for undirected graphs by requiring that the two vertices are adjacent and have at least one neighbor in common (Krackhardt 1999). The resulting relation is a subset of the original relation, restricting adjacencies to pairs of vertices that are members of a triangle.

These non-valued relations can be valued by the *degree of overlap* in neighborhoods, i.e., the commonality of the two sets $N(i)$ and $N(j)$ of neighbors of vertices i and j. Note that in directed graphs, we may

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

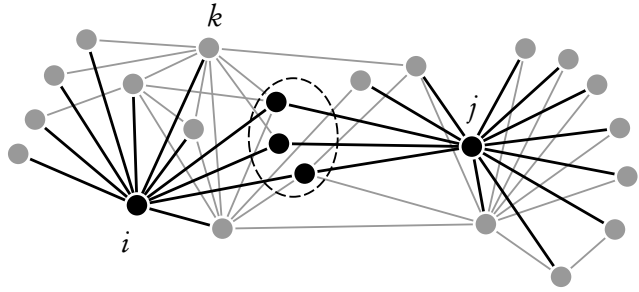


Figure 23: Barely overlapping neighborhoods with Jaccard index $J_G(i,j) = \frac{3}{21} \approx 0.14$ *and Euclidean neighborhood distance* $\sqrt{18} \approx 4.24$. *For comparison,* $J_G(i,k) = \frac{7}{13} \approx 0.54$.

also be interested in the overlap of in-neighbors and out-neighbors separately. Multiple measures for quantifying the overlap of any two sets exist (see, e.g., Sneath and Sokal 1973) and can thus be applied to neighborhoods. Two of the more common of these are the number of joint neighbors, $|N(i) \cap N(j)|$, and the relative number of joint neighbors, $J_G(i,j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$ (known as the *Jaccard index*). These are similarity indices, but since $J_G(i,j)$ is always in the range from 0 to 1, we can construct a *Jaccard distance* from $1 - J_G(i,j)$.

A common measure of dissimilarity that also takes the joint absence of ties into account is the *Euclidean neighborhood distance* of the characteristic vectors of neighborhoods, i.e., the Euclidean distance of the two rows or columns in the adjacency matrix. For out-neighborhoods, the Euclidean neighborhood distance is the Euclidean distance of the row vectors A_i and A_j in the adjacency matrix A,

$$||A_i - A_j|| = \sqrt{\sum_{k \in V} (a_{ik} - a_{jk})^2}\,.$$

Similar concepts of this kind include *Hamming distance* (counting the number of vertices that are adjacent to one of i and j but not the other) and *neighborhood correlation* (correlation of the row or column vectors of the adjacency matrix).

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

4.2.2 Network Characteristics

Before dealing with more fine-grained indicators, let us consider some coarse characteristics of an entire network. These are intended to be helpful in answering the first of our three guiding questions mentioned on p. 111.

Network characteristics are especially useful in comparative studies when the variance in network structure is assumed to be a cause or consequence of other variables. In fact, this is the main hypothesis underlying ego network analysis, in which characteristics of personal networks are used as features of individuals (i.e., of the egos). Please note that it may be easier to survey the characteristics directly than to collect the actual network first and then determine the quantity of interest from it.

Density

The simplest such indicator is a network's *density*, which is defined as the ratio of the number of edges to the number of dyads, i.e., the ratio of the number of actual to possible edges. While the number of dyads $D(G) = \{\{i, j\} : i \neq j \in V\}$ in a graph is always $\frac{n(n-1)}{2}$, the number of edges a dyad can have depends on whether the graph is directed or undirected. We define

$$\textit{density}(G) = \frac{\text{number of edges}}{\text{number of possible edges in a dyad} \times \text{number of dyads}}$$

and thus obtain $\textit{density}(G) = \frac{m}{2|D(G)|} = \frac{m}{n(n-1)}$ if G is a directed graph, and $\textit{density}(G) = \frac{m}{|D(G)|} = \frac{2m}{n(n-1)}$ if it is undirected.

It is instructive to view density from another angle too. Since each edge represents an adjacency, density can be understood as the network average of the dyad-level adjacency index.

Although intuitive as a measure, density has an undesirable scaling behavior. Assume that the network data were collected via questionnaries using a limited-choice design, i.e., each respondent was allowed to name at most k alteri, where k is a constant, say $k = 5$. For simplicity, let us assume that every respondent named exactly k alteri, and that all alteri are also respondents. If the resulting network has n vertices (respondents), it has $k \cdot n$ directed edges (nominations). Its density is therefore $\frac{kn}{n(n-1)} = \frac{k}{n-1}$. Since k is constant, density is tending to zero by design with increasing sample size n.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

This vanishing behavior is encountered quite generally as, even in free-choice questionnaire designs, the average number of nominations is usually bounded by a constant so that the total number of nominations behaves as in the above example.

For an alternative, observe that the average outdegree is at most k by design in the above example. The deviation from k is hence a reasonable indicator of how many nominations were made in comparison to how many were allowed. Since the total outdegree always equals the number of edges, the average outdegree equals density times $n-1$. For networks that are typically sparse, the average degree is generally a better measure of whether the number of edges is large or small, and therefore preferred to density.

In the same way that density summarizes adjacency, *reciprocity* summarizes mutuality. It is expressed as a fraction of the adjacent dyads,

$$\textit{reciprocity}(G) = \frac{\text{number of mutual dyads}}{\text{number of adjacent dyads}},$$

because counting in all non-adjacent dyads also results in an over-emphasis if these are considered as mutual and an under-emphasis if they are considered as non-mutual (compliant with the above definition of mutuality).

Degrees

A frequently used network characteristic is the distribution of degrees, i.e., the fractions of vertices having the same degree. Note that degree is itself a vertex-level aggregation of a dyad index, namely the number of adjacent dyads, in which a vertex is involved. It should also be recalled that we defined a variant concept of density by averaging over all vertex degrees.

Many graphs representing observed networks have been found to contain vertices of surprisingly high degree. More precisely, their *degree sequences* display a scaling behavior defined as follows. Let $d_1 \geq d_2 \geq \ldots \geq d_n$ be the size-ordered sequence of degrees in an undirected graph, then this sequence is said to (approximately) satisfy a *power law* with exponent α, if

$$i \approx c \cdot d_i^{-\alpha}$$

for any constant c and all $i = 1, \ldots, n$. The scatterplot with points (d_i, i) is called *size-rank* plot, and taking the logarithm on both axes yields points lying close to a line with slope $-\alpha$.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Graphs with a power-law degree sequence are called *scale free* for the following reason. Let us assume we are observing a different graph G' with n' vertices satisfying a power law with the same parameters c, α. This can be interpreted as changing the resolution on the y-axis of the size-rank plot; depending on whether n' is larger or smaller than n we add or remove points. Now fix a rank $1 \leq i \leq n$ and consider a corresponding rank $1 \leq i' \leq n'$ that is in approximately the same position after the resolution has been changed, i.e. $\frac{i'}{i} \approx \frac{n'}{n}$. Since the degree sequence $d'_1 \geq \ldots \geq d'_{n'}$ of G' satisfies $i' \approx c \cdot (d'_{i'})^{-\alpha}$ for all $i' = 1, \ldots, n'$, it follows that $\frac{n'}{n} \approx \frac{i'}{i} \approx \left(\frac{d'_{i'}}{d_i}\right)^{-\alpha}$ and therefore $\frac{d'_{i'}}{d_i} \approx \left(\frac{n'}{n}\right)^{-\frac{1}{\alpha}}$, so that the x-axis is essentially scaled by a constant as well. Hence, scaling the size of the graph only results in a corresponding scaling of degrees. For more details on this, see Cooper and Lu (2007).

Please note that if a degree sequence scales with exponent α, degree frequencies also scale with exponent $\gamma = 1 + \alpha$. The interpretation of frequencies as probabilities leads to the graphic interpretation that high-degree vertices are much more likely than they would be in a normal distribution. However, it is a common mistake to start from the seemingly more intuitive frequencies and infer a scaling behavior by fitting lines in doubly logarithmic size-frequency plots (Li, Alderson, Doyle, and Willinger 2005).

Connectivity

The notion of dyad connectivity discussed above can be extended to the entire network. We say that an undirected graph is *connected*, if each dyad is connected. If it is not connected, the inclusion-maximal subsets of vertices that are pair-wise connected are called the *connected components*.

Most commonly, dyad connectivity is extended to entire graphs not by averaging but by taking the minimum. For communication and infrastructure networks, the minimum edge or vertex connectivity between any pair of vertices in the corresponding graph is an indicator of the fault tolerance or attack resilience of the network.

The (weakly) connected components of a graph can often be treated separately because there are no ties creating dependencies between them. Some network-analytic measures require that a graph is connected, or strongly connected, to be well-defined. It may, therefore, be necessary to check for connectivity prior to carrying out an analysis.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Distance

A simple network characteristic based on distances is *Wiener's Index*, $W(G) = \sum_{s,t \in V} d_G(s,t)$ (Wiener 1947), where $d_G(s,t)$ is the shortest-path distance from s to t. It is a requirement that the graph be strongly connected since the index is infinite or undefined if there is a dyad with vertices that are not mutually reachable. Please remember that $d_G(s,t) = 0$ if $s = t$ and observe that each dyad is counted twice (once in each direction). To avoid double counting in undirected graphs, the above value can be divided by two.

The normalized version of Wiener's Index representing the average distance of each pair is referred to as the *characteristic path length*, $L(G) = \frac{1}{n(n-1)} W(G)$ and one of the criteria used to classify networks as *small worlds* (Watts and Strogatz 1998).

Configuration Counts

As in many variations of the above examples and other dyad-level or vertex-level indices, network characteristics can be obtained from their distribution and statistics of these distributions.

A different class of network characteristics is obtained from counting configurations rather than aggregating lower-level indices. A simple example is the *dyad census* of a directed graph. Since vertices of any dyad can be linked by at most two edges, it consists of three values for the number of mutual, asymmetric, and null dyads (Holland and Leinhardt 1970).

The next biggest example is the *triad census* (Holland and Leinhardt 1976), in which the frequency of all 16 possible configurations of the three dyads formed by three vertices as shown in Figure 24 are determined. More generally, one is often interested in particular, or particularly frequent, subgraphs which are then referred to as *motifs*. Counting statistics are pivotal in network modeling as discussed in Section 4.3.

Macro Shapes

Some aggregates of lower-level indices serve to characterize certain tendencies that relate to images of a global shape of the network. As one dimension of informal organization, Krackhardt (1994) quantifies the degree to which a graph is *hierarchic* as $1 - reciprocity(G_r)$, where G_r is the graph defined by the reachability relation and *reciprocity* is as defined on p. 119. A relation such as pecking is thus hierarchic, if the number of

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

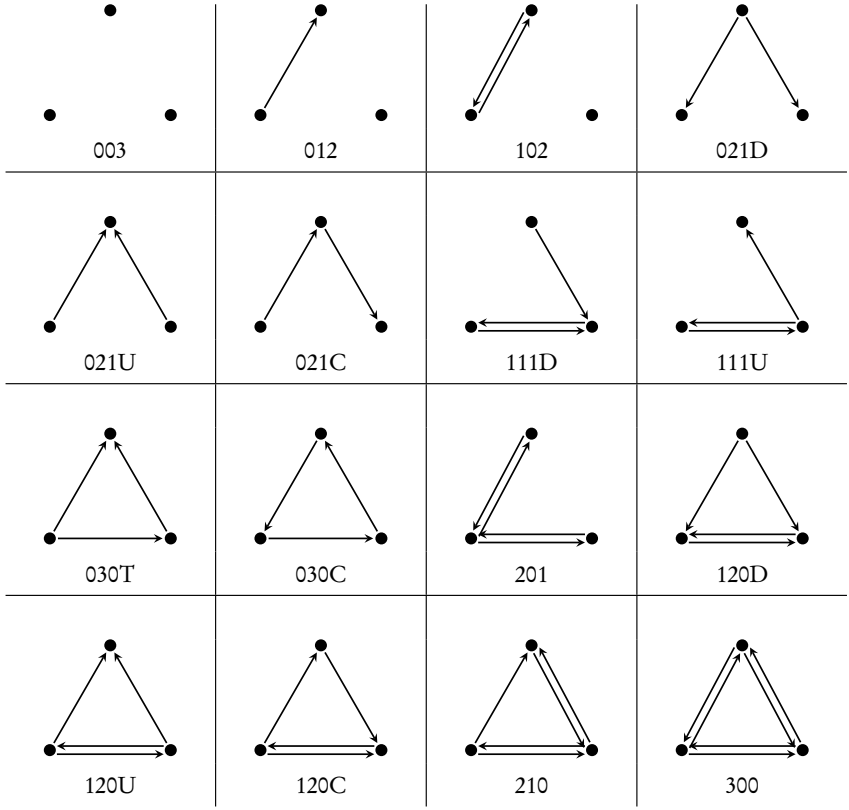


Figure 24: Triad types in directed networks. The conventional numbering scheme is based on the number of mutual, asymmetric, and null dyads (MAN for short), with trailing characters U: up, D: down, T: transitive, and C: cyclic.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

dyads in which one actor can reach the other but not vice versa is relatively high (see also Everett and Krackhardt 2012).

A summary statistic of vertex-level indices is *centralization* (Freeman 1979). A vertex centrality assigns values to vertices that quantify the structural importance of vertices, and we treat them extensively below. One such index is simply the degree of a vertex. Whether a graph is centralized is then assessed by the degree to which few vertices are notably more central than the others,

$$\frac{\sum_{i \in V} \hat{c} - c(i)}{C(n)},$$

where $c(i)$ is the centrality of vertex $i \in V$, $\hat{c} = \max_{j \in V} c(j)$ is highest centrality of any vertex in the graph, and $C(n)$ is the maximum value that the numerator can attain in any graph with the same number n of vertices.

Centralization is superficially related to the concept of having a *core-periphery structure*. The latter refers to the possibility of classifying the actors into a cohesive core and a non-cohesive periphery (Borgatti and Everett 1999). Core-periphery structures can also occur as artifacts of poor boundary specification or snowball sampling.

4.2.3 Centrality

We shall now address the second guiding question of this section by introducing methods developed for the identification of (structurally) important actors and, more generally, the relative importance of all actors. In fact, this is one of the primary tasks in network analysis.

The most common approach to this task is the assignment of values to actors such that larger values correspond to greater importance. In this scenario, centrality is an actor-level index, and we will present several indices that, together, can be considered representative.

Most actor-level indices are aggregates of dyad-level indices from precisely those dyads in which the actor is involved. The most elementary ones are *degrees* and can be considered as measures of an actor's activity or involvement. These are simple counts of the number of adjacencies of the corresponding vertex. We define

$$\begin{aligned} \text{indegree} \quad & d^-(i) = \textstyle\sum_{j \in V} a_{ji} \\ \text{outdegree} \quad & d^+(i) = \textstyle\sum_{j \in V} a_{ij} \\ \text{degree} \quad & d(i) = d^-(i) + d^+(i) \end{aligned}$$

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

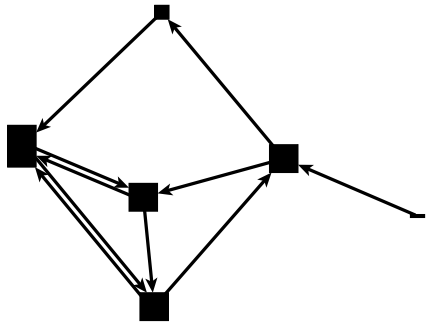


Figure 25: Degree visualized in terms of node size.

While, say, indegree corresponds to the number of in-neighbors in non-valued graphs, the above definition generalizes to valued and multigraphs. In undirected graphs, in-neighbors and out-neighbors are the same and degree is therefore defined as $d(i) = d^-(i) = d^+(i)$.

Obviously, we can define similar actor indices by replacing adjacencies with any other dyad-level index such as mutuality.

Centralities are probably the most commonly used actor-level indices. There is no agreement as to what exactly centrality is and which properties a centrality index should satisfy (Freeman 1979) but, in general, the idea is that centralities express a structural advantage, importance, or dominance.

In this sense, degree is a centrality index because activity may already be an indication of importance. The oldest example of this kind is *sociometric choice*, in which the popularity of an actor is operationalized as the indegree of the corresponding vertex.

Radial Centralities

A more elaborate way of defining an actor centrality is, again, to aggregate information over all dyads involving the actor of interest. An example of this is *closeness centrality* (Bavelas 1950; Sabidussi 1966),

$$c_C(i) = \frac{1}{\sum_{t \in V} d(i,t)} = \frac{1}{\text{total distance to all other vertices}},$$

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

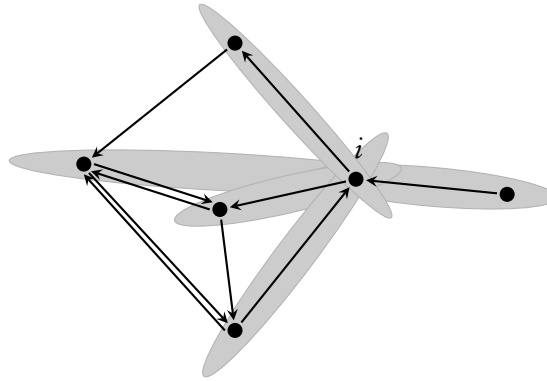


Figure 26: Radial dyads centered on i.

in which the inverse of the sum of distances to all other vertices is taken so that large distances correspond to low centralities. Closeness is thus an example of what is termed a *radial* notion of centrality because the underlying principle is that the centrality of an actor is defined in terms of the dyadic relations of this actor with everyone else in the network (Borgatti and Everett 2006). Obviously, the dyadic relation may be instantiated with any of the dyad indices discussed above (distance, connectivity, etc.), and we can also aggregate them in various ways (sum, harmonic sum, maximum, median, etc.).

The choice of dyad index should be guided by the kind of network relation considered (Borgatti 2005). When studying a network of personal communication links, for example, the importance of an actor as a source of rumors may be defined using a dyad-level index that models how likely it is that a rumor originating from one actor will arrive at the other within a given time span and without alteration. The aggregation of these indices with the same source vertex then evaluates its overall importance for placing rumors.

Burt (1992) also defines a radial centrality, *constraint*, by summing redundancy (as defined on p. 131) over all dyads, in which the actor participates.

Medial Centralities

Instead of assessing the position of an actor by aggregating over all dyads, in which the actor is involved, we can aggregate over all other dyads the importance of the actor's position for that dyad.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

One interpretation of such involvement is the dependency of a dyad on the actor of interest to serve as an intermediary. The standard example is *betweenness centrality* (Freeman 1977). Each dyad contributes the dependency of its members on the focal actor. For betweenness, this is called *pair-dependency* and defined as

$$\delta(s,t|i) = \frac{\sigma(s,t|i)}{\sigma(s,t)} = \frac{\text{number of shortest } st\text{–paths via } i}{\text{number of shortest } st\text{–paths}}$$

where $\sigma(s,t)$ is a dyad-level index defined as the number of different shortest paths from s to t, and $\sigma(s,t|i)$ is the number of those that involve i as an intermediate vertex. Since

$$\sigma(s,t|i) = \begin{cases} \sigma(s,i)\cdot\sigma(i,t) & \text{if } d_G(s,t) = d_G(s,i) + d_G(i,t) \\ 0 & \text{otherwise}, \end{cases}$$

$\sigma(s,t|i)$ and also $\delta(s,t|i)$ are easily recognized as a dyad index derived from another one (shortest-path distance). Betweenness centrality is obtained by aggregating dependencies over all dyads that do not contain the focal actor,

$$c_B(i) = \sum_{s\neq i\neq t\in V} \delta(s,t|i)\,.$$

Again, pair-dependency based on shortest paths can be replaced by dependencies based on other dyad-level indices, such as current flow or connectivity, and selecting the appropriate one is a substantive matter.

Feedback Centralities

Feedback refers to the dependence of an actor's centrality on the centrality of other actors. The classic example of this kind is *eigenvector centrality* (Bonacich 1972; 1987; Bonacich and Lloyd 2001), which is defined by

$$c_E(i) = \alpha \sum_{j\in N^-(i)} c_E(j)$$

where $\lambda = \frac{1}{\alpha}$ is an eigenvalue of the adjacency matrix. Please note that the resulting system of linear equations has one equation per vertex and it can be written as $Ac_E = \lambda c_E$, where A is the adjacency matrix of the graph. If the graph is undirected or strongly connected, and λ is selected as the largest eigenvalue of A, this system has a unique solution with only

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

positive entries. Note that we can replace the in-neighbors by any other neighborhood.

The rationale underlying eigenvector centrality is straightforward. An actor's centrality is proportional to the total centrality of its neighbors. Conversely, an actor contributes its own centrality to each of its neighbors. This implies that the contribution must be understood as something replicable that scales to any number of neighbors. If the contributions that can be made are limited, for example because it requires time to pass on importance, it may be better to divide the contribution by the number of recipients as in

$$c(i) = \alpha \sum_{j \in N^-(i)} \frac{c_e(j)}{d^+(j)} .$$

Interestingly, in undirected graphs this yields $\alpha = 1$ and is equivalent to degree. In directed graphs it is more informative but requires strong connectivity. A variant that circumvents this problem is the *PageRank index* used in Google's search engine (Brin and Page 1998),

$$c_P(i) = (1-\omega) \left(\sum_{j \in N^-(i)} \frac{c_P(j)}{d^+(j)} \right) + \omega \frac{1}{n(G)} ,$$

where the second term corresponds to artifical links that fully connect the graph but have low and uniform weight, and $0 < \omega < 1$ is a parameter trading off structure versus the uniform a-priori influence.

Interestingly, feedback centralities can also be expressed in terms of dyad-level indices, although the relation is intricate. For instance, a frequently used measure called *status* (Katz 1953) can be defined either in feedback terms as

$$c_S(i) = \sum_{(j,i) \in V} \alpha \cdot (1 + c_S(j))$$

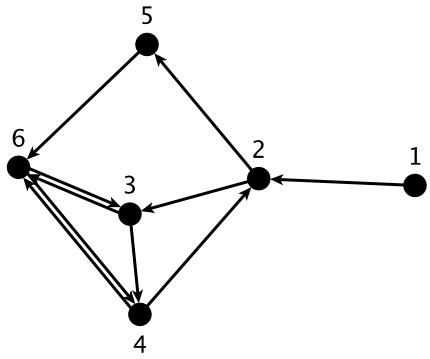
where $0 < \alpha < 1$ is a sufficiently small constant, or as a radial centrality based on the sum over dyad scores

$$c_S(i) = \sum_{s \in V} \chi_\alpha(s,i)$$

where $\chi_\alpha(s,i)$ is the sum over all (simple or not) *si*–paths of any length, where the contribution of paths of length k is weighted by α^k, i.e., longer walks contribute less. This dyad index thus considers direct and indirect choices, and therefore generalizes sociometric choice.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



vertex	raw	std.	pct.
1	0.00	0.0000	0.0%
2	3.75	0.1875	37.5%
3	1.50	0.0750	15.0%
4	2.50	0.1250	25.0%
5	0.50	0.0250	5.0%
6	1.75	0.0875	17.5%

Figure 27: Betweenness centrality scores: raw values, standardized by dividing by the maximum possible score of $5 \cdot 4 = 20$*, and normalized by their sum* 10*. Relative scores remain the same within the network, but need not across networks.*

Choosing a Centrality Index

The relations between radial, medial, and feedback centralities are not fully understood. Nevertheless, reasonably educated choices can be made by considering the following two dimensions.

Dyad index: In your study, you will investigate particular kinds of ties, and have specific theories and ideas about how actors relate directly or indirectly through these ties. Identifying a dyad-level index that reflects the nature of these relations is, therefore, the crucial first step.

Aggregation: In a second step, the way, in which these relations yield special positions, is derived from theories of, e.g., power attribution, and operationalized in (i) the choice of dyads and (ii) the way their evaluations are aggregated.

Comparing Actor Centralities

Centrality scores are usually considered as being ordinal, i.e., while the rank of an actor's score in the sorted list of all scores is important, the absolute value is not. To compare actor centrality across networks, however, scores must be mapped to a common scale. Two main approaches have been proposed and whereas in both cases scores are mapped to lie in the interval from zero to one, the reference quantity is different.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Standardization: Each vertex centrality score is divided by the maximum possible score for a vertex in any graph with the same number of vertices (Freeman 1979). Therefore, the standardized score is one exactly when a vertex cannot be more central in any configuration with the same number of others. Usually, but not always, this corresponds to comparing the position of a vertex with that of the center of a star.
The advantage is that the centrality of an actor is assessed in absolute terms, compared to what is possible. A disadvantage is that this theoretical maximum is not always well-defined (as is the case with eigenvector centrality, for instance) and that it is difficult to compare the centrality of actors in networks of different sizes. Even for networks with the same number of vertices, it could be argued that the maximum should really be determined only among graphs that also have the same number of edges.

Normalization: Instead of an absolute benchmark, you can use the scores of the other vertices in the graph for reference. This is achieved by dividing each centrality score by the sum of all scores, which yields the share of importance that a vertex received. Such a relative notion enables comparison across all kinds of networks, and even centrality measures. However, it is not possible to state that a score is high or low in absolute terms.

Neither transformation changes the rank order of actors in the same network. Since standardization is proposed in Freeman (1979), it is the default form for presenting centrality scores in many tools for network analysis, although sum-normalized scores are more generally applicable. We already pointed out at the end of Section 4.2.2 that summary statistics of the centrality distribution can be used to characterize the network as a whole.

Edge Centralities

In this section, we briefly discuss means of assessing the importance of individual edges. This is very similar to what we have just discussed for vertices.

In fact, concepts introduced for vertices carry over to edges as we can transform a graph into a new graph that has a vertex for every edge of the original one, and two of these new vertices are adjacent, if the original edges share a vertex. More precisely, a directed graph $G = (V, E)$ yields a *line graph* $\mathscr{L}(G) = (E, \{(e_1, e_2) : e_1 = (i, j), e_2 = (j, k) \in E\})$. Line

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

graphs thus reflect how edges are connected among themselves, and any vertex index can be applied to the line graph to define an edge index in the original graph.

However, genuine edge indices also exist. In his seminal paper, Granovetter (1973) distinguishes strong and weak ties based on a forbidden triad configuration. As this classification is not unique, however, he also proposes a valued assessment of the importance of an edge for shortcutting otherwise distant dyads. An edge (i,j) of a non-valued graph G is called a *k-bridge*, if $d_{G-(i,j)}(i,j) \geq k$, i.e., if the distance from i to j is at least k if (i,j) is removed. Edges that are k-bridges for some large k are considered indicative of weak ties because they link distant vertices.

A similar operationalization of weak ties is based on betweenness. A proper extension of the pair-dependencies on intermediate vertices defined above is

$$\delta(s,t|(i,j)) = \begin{cases} \sigma(s,i)\cdot\sigma(j,t) & \text{if } d(s,t) = d(s,i)+1+d(j,t) \\ 0 & \text{otherwise,} \end{cases}$$

so that an *edge betweenness centrality* can be defined as

$$c_B(e) = \sum_{s,t\in V} \frac{\sigma(s,t|e)}{\sigma(s,t)}\,.$$

Since, unfortunately, this index does not coincide with vertex betweenness centrality in the line graph, care must be taken in choosing an appropriate index.

4.2.4 Cohesion

Cohesion denotes the tendency towards dense, redundant connections. For example, we may want to assess the cohesion of a team because a group performance theory relates it with effectiveness. More generally, the variance in cohesion across a network can be used to identify cohesive groups and thus provide one type of answer to the third guiding question of this section.

We have already discussed an elementary index for cohesion, namely (network-level) density. Since density is defined as the ratio of existing to possible edges, it can be defined for any subnetwork by restricting the set of dyads taken into consideration. Some actor-level and group-level versions of density are listed below for comparison with more sophisticated methods.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

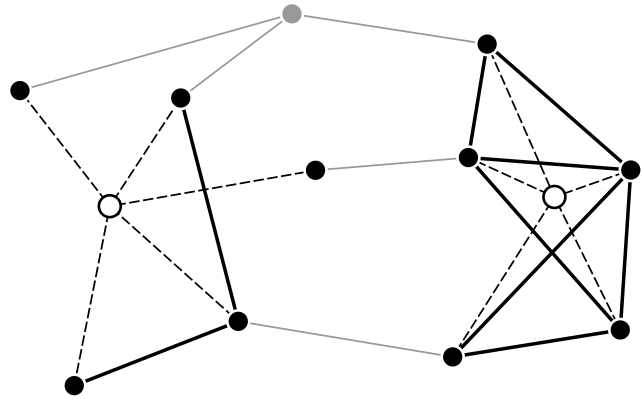


Figure 28: Two neighborhoods of size 5 with different density. The left one has density $\frac{2}{5 \cdot 4} = 0.1$, the right one $\frac{7}{20} = 0.35$. By definition, these values are the clustering coefficients of the two vertices defining these neighborhoods.

Actor Level

On the actor level, cohesion refers to the tight integration of an actor into the surrounding structure. The best-known measure of cohesion on the actor level is the *clustering coefficient* (Watts and Strogatz 1998), which is simply the density of the graph induced by the neighborhood of an actor. Whereas degree is defined in terms of the number of neighbors, the clustering coefficient is based on their interconnectedness.

We have already argued, however, that density has undesirable scaling properties, and thus recommended that average degree be used instead. In fact, Burt (1992) defines the *redundancy* in an actor's neighborhood as the average degree of its members in the induced subgraph.[1] In addition to redundancy, Burt (1992) defines three other cohesion-based indicators describing the neighborhood of an actors.

The opposite of cohesion among neighbors is their segregation. In particular, an actor may be a *broker* that is vital for dyad-level indices of connectivity in the sense that these deteriorate when its vertex is removed from the graph. In the extreme case, the graph is no longer connected without this vertex.

1 This is actually a simplified interpretation of redundancy in unweighted graphs pointed out in Borgatti (1997).

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Group Level

A cohesive group is a group of actors with higher internal than external cohesion. The most cohesive group imaginable is one in which all members are pair-wise adjacent. A set $C \subseteq V$ of vertices in a graph is called a *clique* (Luce and Perry 1949), if all dyads in the induced graph $G[C]$ are mutual. Remember that mutual and adjacent are the same in undirected graphs and note that any subset of a clique also forms a clique. A common form of analysis is, therefore, the enumeration of all inclusion-maximal cliques of at least three vertices.

In observed networks, there may be few inclusion-maximal cliques and their size is usually small. This is because the requirement for a set of vertices to form a clique is extreme, and a single missing edge rules out the enlargement of a clique. Of the many weaker concepts of strong internal cohesion of a group, we mention just one, k-cores.

The *k-core* of an undirected graph is a vertex-induced subgraph $G[C_k]$ induced by an inclusion-maximal set of vertices $C_k \subseteq V$ such that every vertex $i \in C_k$ has at least k neighbors in C_k, i.e., the induced degree $d_{G[C_k]}(i) \geq k$ (Seidman 1983). Note that C_k may be smaller than the set of all vertices of degree at least k in the whole graph G. In fact, the k-core of a graph is uniquely defined for all $k = 0, \ldots, n-1$. Since cores are nested, i.e., the $k+1$-core is a subgraph of the k-core, we have a successively stronger requirement.

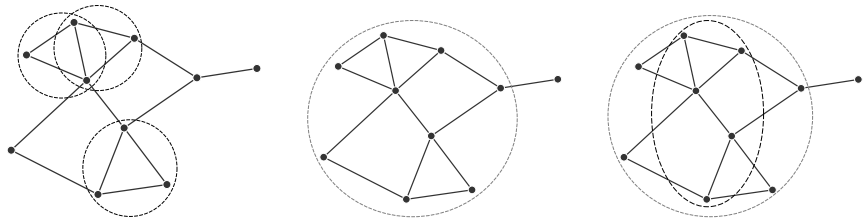
On the other hand, the following example shows that by focussing on degree, we may lose the connectivity aspect of cohesion. Consider a graph consisting of two separate cliques of $k+1$ vertices each. These two cliques together then form the k-core because each vertex has, indeed, at least k neighbors in this graph.

If connectivity is more important than degree, a corresponding concept of group cohesion can also be defined. A *λ-set* is a subset of vertices $C \subseteq V$, in which every internal dyad has a higher edge-connectivity than any dyad that involves a member and a non-member of that group (Borgatti, Everett, and Shirey 1990). Again, we obtain a nested structure of successively more cohesive groups.

Despite the differences between them, all of the above concepts share the property that each subset of vertices is either a group in the corresponding sense or not. This is because cohesion is defined as an absolute requirement and, consequently, a graph may not contain any cohesive group.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



(a) cliques of size 3 (b) 2-core (no 3-core) (c) λ-sets for $\lambda = 2,3$

Figure 29: Comparison of cliques, cores, and λ-sets.

Clustering

Without an absolute definition of cohesion, a graph can nevertheless be decomposed into relatively cohesive groups by trading-off internal cohesion against external separation. The strategy is to determine a set $\mathscr{C} = \{C_1, \ldots, C_r\}$ of *clusters*. Commonly, $\mathscr{C}$ is a *partition*, i.e., the clusters are disjoint, non-empty, and cover the set of vertices. Overlapping clusters are sometimes also considered, however.

Various objective functions have been proposed for assessing the quality of a partition with respect to the internal-external criterion (Schaeffer 2007; Fortunato 2010). The one currently used most frequently is called *modularity* and defined as

$$Q(\mathscr{C}) = \sum_{C \in \mathscr{C}} \frac{|m(C)|}{m} - \frac{\sum_{i \in C} d(i)^2}{(2m)^2},$$

where $m(C)$ is the number of edges with both vertices in C (Newman and Girvan 2004).

Modularity thus trades off two competing goals. The first term corresponds to the percentage of edges internal to the clusters and is, therefore, large if clusters are growing. It is maximum for a single cluster covering all edges. For the second term, observe first that the sum over all vertex degrees is twice the number of edges and that the square of a sum of positive values is never smaller than the sum of the squares of these values. The second term is thus small, if vertex degrees are scattered across clusters. It is minimum for the partition, in which every vertex forms a singleton cluster. A partition is thus considered a good clustering, if most edges are internal to clusters but clusters are small and balanced in the total degree of their vertices. An example is given in Figure 30.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

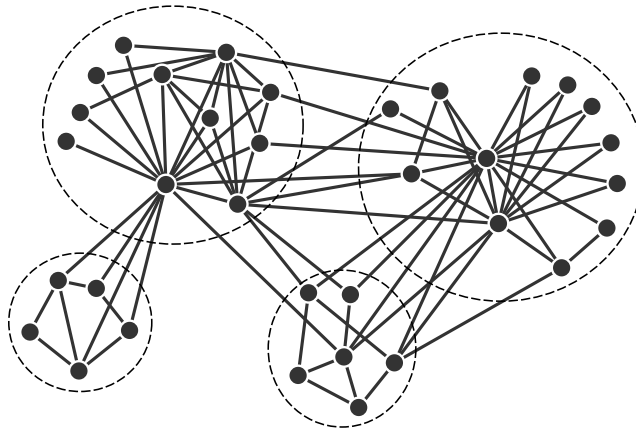


Figure 30: Partition with maximum modularity.

In recent years, extensive research has also unveiled quite a few disadvantages of the modularity objective. Like most other graph clustering methods, it is computationally intractable to optimize (Brandes, Delling, Gaertler, Görke, Hoefer, Nikoloski, and Wagner 2008), and partitions of similarly good quality can be very different (Good, de Montjoye, and Clauset 2010).

On the other hand, if there is a prominent clustering with a high variance of local density, any of the established methods will recover the corresponding partition. The precise relations with network formation theories and difference in outcome are even less well understood than for centrality indices. Hence, for now, at least, you need not worry too much about which graph clustering method to use. It is more important to be careful in the interpretation of results which are bound to be subject to artifacts or arbitrariness.

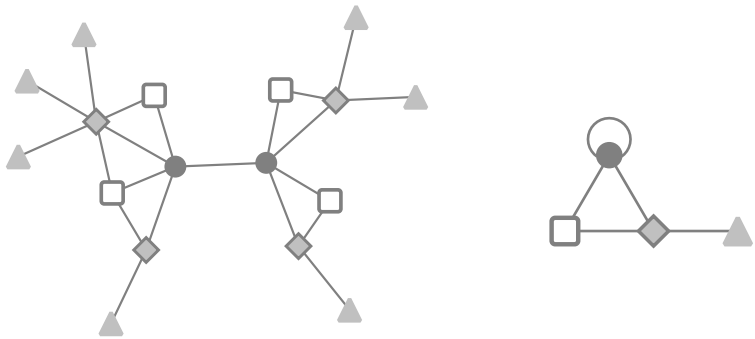
4.2.5 Roles

Cohesion appears to be a straightforward concept for grouping. Groups need not, however, be characterized by tighter intergration, actors can also be classified by similarity of relations to other (groups of) actors. Just like cohesion, this concept may be defined in various ways.

Nadel (1957), for instance, argued that the relations among actors in a social network are governed by their *roles*, where a role may be defined,

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



(a) graph with role assignment (b) corresponding role graph

Figure 31: In a regular role assignment, every pair of vertices having the same role is adjacent to the same other roles.

e.g., by social category, membership in a social group, or other actor characteristics.

Formally, a *role assignment* is a mapping $r : V \to R$ of vertices $v \in V$ to roles $r(v) \in R$. A role assignment reduces the complexity of a network by abstracting from the individual actors. Think of a food web, in which not individual animals and plants, but their types are represented. The predator-prey relationship of an animal of one type towards one of another type is generally the same for all members of these types (see, e.g., Luczkovich, Borgatti, Johnson, and Everett 2003).

Figure 31 shows how a role assignment can be used to reduce a graph to the corresponding *role graph*. Note that the edges in the role graph only make sense, if they are *compatible* with those in the original graph. Therefore, not every mapping of vertices to roles is a meaningful role assignment. In the following, we will discuss the two main classes of constraints that are used to ensure compatibility of a role assignment with the given structure.

Structural Equivalence

A straightforward and purely structural definition of role, which may or may not have a deeper foundation, is *structural equivalence* (Lorrain and White 1971). Two actors are considered structurally equivalent, only if they have exactly the same neighbors. Thus, structurally equivalent actors can not be distinguished by their relations to others.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

A role assignment is based on structural equivalence, if assuming the same role implies having the same neighbors,

$$r(i) = r(j) \quad \Longrightarrow \quad N(i) = N(j)\,.$$

A graph may admit several role assignments based on structural equivalence, some of which are refinements of others. In particular, assigning a different role to each vertex is always compatible with the definition of structural equivalence; this is because no two vertices have the same role, therefore the set of conditions is empty.

On the other hand, structural equivalence is a rather strict concept. A single added or omitted tie may render an otherwise meaningful role assignment infeasible. Relaxations of the concept have been proposed, therefore, in which actors are grouped not by equality but similarity of their neighborhoods. We have already discussed corresponding dyad indices at the end of Section 4.2.1, and in particular the Jaccard coefficient. Groups can be formed from such pair-wise similarities or dissimilarities by applying general data clustering methods. An unconventional approach to classification based on iterated neighborhood correlation is proposed in Breiger, Boorman, and Arabie (1975).

Regular Equivalence

Structural equivalences and their relaxations are inappropriate if the notion of role we are intersted in is not defined in terms of ties with the same actors. Two whales do not play the same role in a food web because they feed on the same krill but because they both feed on krill.

The prototypical example in which roles are defined in terms of relations to actors who are not necessarily the same but have the same role, is called *regular equivalence* (White and Reitz 1983). A role assignment based on regular equivalence must satisfy

$$r(i) = r(j) \quad \Longrightarrow \quad r(N(i)) = r(N(j))$$

where $r(N(i))$ is the set of roles in the neighborhood of $i \in V$. Note that this introduces a notion of feedback into the definition, similar to what we did with centralities. Note also that the number of occurrences of a role in a neighborhood does not matter. The roles in Figure 31 are indeed based on regular equivalence, and the two diamond-shaped vertices on the left have a different number of triangle-shaped neighbors. If we require

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

that the number of neighbors having the same role also be equal, i.e., if we interpret $r(N(i))$ as a multiset, the role assignment is said to be based on *exact equivalence* (Everett and Borgatti 1996).

Like structural equivalence, regular equivalence may allow for several different role assignments. Again, the trivial role assignment of one role per vertex is compatible with the definition, and, moreover, the assignment of the same role to all vertices does not violate any constraint either, unless there are both isolate and non-isolate vertices. In the majority of cases, we are interested, therefore, in the coarsest role assignment based on regular equivalence that refines a given partition based on other criteria. A number of results about role equivalence relations is compiled in Everett and Borgatti (1994).

4.2.6 Blockmodeling

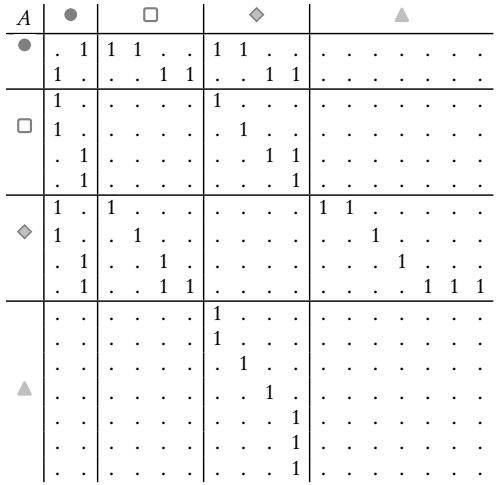
The general problem of partitioning the actors of a network into groups such that the relations within and between groups are close to some idealized pattern is called *blockmodeling*. The name stems from the adjacency matrix representation, in which rows and columns can be reordered to have groups form intervals. The relationships within groups then appear as quadratic submatrices along the diagonal, and relationships between groups as rectangular off-diagonal submatrices. Both types of submatrices are referred to as *blocks*. A partition is considered feasible, if specific requirements are met by each block.

As initial examples for such requirements, note that the above approaches to clustering and role equivalence are special cases of blockmodeling. In clustering, we seek to determine a partition, in which the diagonal blocks are as full as possible and the non-diagonal blocks as empty as possible. Each of the clustering approaches mentioned has its own specific concept of full and empty. Vertex equivalences, on the other hand, yield blocks that correspond to the defining equivalence relation. Regular equivalence, for instance, yields blocks that are either empty or have at least one non-zero entry in each row and column. These so-called regular blocks indicate that each member of the row group is adjacent to at least one member of the column group.

Similar to the role graph for equivalence classes, we can generate a more abstract view of a blockmodel by replacing each block with a single entry indicating its type as shown in Figure 32. This simplification is called the *image matrix* of a blockmodel.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH



(a) blockmodel

A/r	●	□	◇	▲
●	1	1	1	.
□	1	.	1	.
◇	1	1	.	1
▲	.	.	1	.

(b) image matrix

Figure 32: Blockmodel with groups corresponding to the regular equivalence classes of Figure 31. Consequently, all non-empty blocks are regular because they contain at least one 1 *in every row and column.*

It need not be the case, however, that groups are uniform in the sense that their internal and external relations follow the same pattern across all groups. A simple, yet important, example is given in Figure 33. The image matrix represents a prototypical *core-periphery structure*, in which there is a cohesive center with a periphery of near-isolates.

Several degrees of freedom can be leveraged to fit blockmodels to empirical data (see, e.g., Snijders and Nowicki 1997). Depending on the context, the presence of several groups with special inter-group relations may be anticipated. This corresponds to assuming the presence of certain types of blocks, while others, for which no expectations are made, can be of any type. Moreover, the evaluation of a blockmodel is highly dependent on the way in which the degree of non-conformance with an ideal block type is measured.

To illustrate this point, consider the ideal block types for cohesion-based clustering, namely full and empty blocks. These are almost never observed empirically because they would correspond to a collection of cliques. The relaxations proposed differ in their goodness-of-fit criteria

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

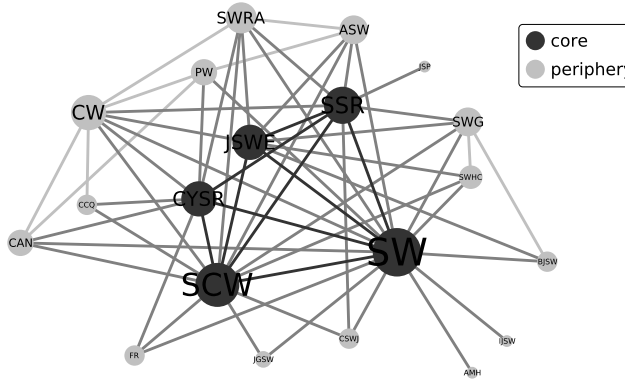
© Campus Verlag GmbH

	core	periphery
core	1	1
periphery	1	0

(a) pre-specified core-periphery image matrix

	core					periphery														
SSR	1	1	1	1	1	.	.	1	.	.	.	1	.	.	.	1	1	1	.	1
CYSR	1	1	.	1	1	1	.	.	1	.	.	.	.	1	1	1	.	.	.	1
JSWE	1	.	1	1	1	.	.	.	.	.	.	1	1	.	.	1	.	1	1	1
SCW	1	1	1	1	1	1	.	1	1	.	1	1	.	.	1	1	.	1	1	1
SW	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	1	.	1	1	1
CAN	.	1	.	1	1	1	.	.	.	.	.	.	.	1	.	1	.	.	.	.
AMH	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.
CSWJ	1	.	.	1	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
FR	.	1	.	1	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
IJSW	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.
JGSW	.	.	.	1	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.
ASW	1	.	1	1	1	.	.	.	.	.	.	1	.	1	.	.	.	.	.	1
BJSW	.	.	1	.	1	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.
PW	.	1	.	.	1	1	.	.	.	.	.	1	.	1	.	1	.	.	.	.
CCQ	.	1	.	1	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.
CW	1	1	1	1	1	1	.	.	.	.	.	.	.	1	1	1	.	.	.	1
JSP	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.
SWG	1	.	1	1	1	.	.	.	.	.	.	.	1	.	.	.	.	1	1	.
SWHC	.	.	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.
SWRA	1	1	1	1	1	.	.	.	.	.	.	1	.	.	.	1	.	.	.	1

(b) network fitted to core-periphery model



(c) network diagram (size represents degree)

Figure 33: Example of Borgatti and Everett (1999), in which citation data between social work journals (Baker 1992) is fitted to a core-periphery model.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

and, in many cases, these can be separated into goodness-of-fit statistics for the individual blocks.

Almost all criteria other than role equivalence lead to computationally intractable partitioning problems. Hence, the software tools available for blockmodeling resort to heuristic algorithms which generally yield suboptimal solutions. Interpretation should therefore be careful and generally avoid focusing on individuals. A comprehensive overview of block types, fitting methods, and interpretation of blockmodels is provided in Doreian, Batagelj, and Ferligoj (2005).

4.3 Modeling

Models express our – often simplified – assumptions about underlying mechanisms or possible outcomes. Statistical models, in particular, express our assumptions about associations between variables that are subject to noise, measurement error, or uncertainty.

Goldenberg, Zheng, Fienberg, and Airoldi (2009) and Snijders (2011) provide detailed and comprehensive surveys of network modeling approaches. We differentiate somewhat more informally between three different forms of modeling that feature prominently in network analysis:

- The first consists of statistical models of associations between (network or non-network) variables but is not special to network analysis. This happens when the statistics involved may be obtained from networks via methods such as those described in Section 4.2 but are treated like any other quantity from then on. As an example of an important method that requires adaptation for relational data, we mention permutation tests (Krackhardt 1987; Dekker, Krackhardt, and Snijders 2007).
- The second class of models is used to describe the state or evolution of a network, often with networks as the dependent variables. Since statistical network models require their own kind of reasoning and are, therefore, special compared to other distributions, the remainder of this section is devoted to them.
- The third approach is simulation, in which likely outcomes are predicted from observed or deliberately chosen boundary conditions and assumptions about their influence. A prominent example is agent-based modeling in social simulation. Since these approaches are only

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

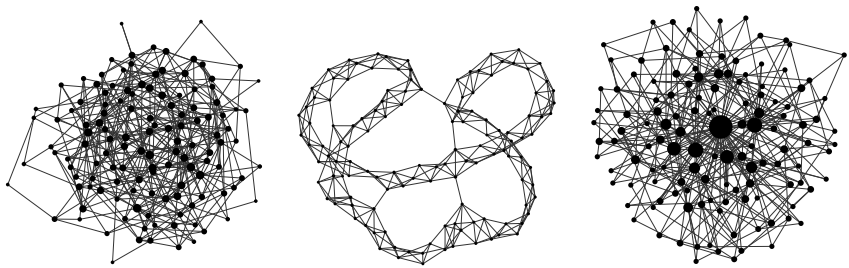


Figure 34: Typical samples from simple models are highly artificial. All three have the same number of vertices and edges, and vertex size represents degree.

rarely used in empirical research, they are omitted from this book. Should you be interested in finding out more about them, we recommend Axelrod (1997) for an introduction to this domain.

A statistical network model consists of a set of conceivable network configurations together with a family of probability distributions on them. These distributions describe whether certain features shall be considered typical or atypical for a network. Typically, a parameterized family of distributions is specified (*modeling*) and a particular member of that family is selected based on an observed network (*estimation*). Whether features of the observation are significant is then assessed relative to the selected distribution. The role of the distribution is to represent variability in a population, measurement error, or uncertainty, and may, therefore, also be selected on other grounds. These general statements are made more concrete in the following.

4.3.1 Idealized Models

Let us start with basic models for the structure of simple, undirected graphs; no attributes, no directed edges considered. Although it is sometimes the case, the models presented in this section are not intended for use in empirical studies. Instead, they rather serve as idealized baseline models that illustrate a particular concept. Figure 34 shows anecdotal evidence that samples from these models are easily recognized as such.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Random Graphs

The number of vertices and the number of edges are two most straightforward features of a graph. We can define a statistical model on the class of all simple undirected graphs by postulating that all graphs with the same number of vertices and edges, n and m, are equally likely.

Given a simple undirected graph G that represents the structure of an observed network, the only reasonable choice of distribution from this model is the unique one with matching numbers of vertices and edges. Since empirical networks are, however, the result of a formation process that is usually far from being uniformly random, the expectation for features of graphs drawn from this distribution will be significantly different from those of the observed network structure. Loosely speaking, this is because a typical random graph has little variation in the density of any subgraph, i.e., almost no clustering, almost no deviation from average degrees, etc.

The study of random graphs is really a branch of mathematics, and it was neither initiated nor developed for the assessment of empirical phenomena but for showing the existence of graphs with certain structural characteristics, and the conditions under which they are rare or abundant. See Bollobás (2001) for a comprehensive treatment.

The model defined above is the original *Erdős-Rényi model* which we refer to as $\mathcal{G}(n, m)$. The more common variant, denoted $\mathcal{G}(n, p)$, is defined on the same class of graphs but with probability distributions parameterized by the number n of vertices and a common probability p for the existence of each individual edge (independent and identically distributed). From a mathematical perspective, these models are largely exchangable; for empirical purposes, they can serve as a no-knowledge baseline at best.

Small Worlds

The model of Watts and Strogatz (1998) confirms that it is possible to design graphs with structural features that seem plausible for social networks, namely overall sparseness, high local density, and short average distance.

This is achieved by starting from a cycle of n vertices, each of which is adjacent to its k nearest neighbors on the cycle for some small k, say $k = 6$. This highly artifical graph is sparse because k is chosen much smaller than n, and locally dense because most neighbors of a vertex are

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

connected themselves. However, its diameter, the longest shortest path of any dyad, is large as it scales with $\frac{n}{k}$. The crucial observation is that even sparse random graphs have a low diameter because it is highly unlikely that none of the many possible shortcuts of a long shortest path is realized. Hence, when adding or rewiring a few edges at random, the average distance in this artificially constructed graph is reduced much faster than the other properties are lost.

Similar to random graphs, samples from this small-world model may have superficial similarity with graphs from observed networks, but, because of their sustained high degree of craftedness, they would yield a poor statistical model.

Preferential Attachment

Another frequent feature of observed network structures is the presence of vertices with high degree. Such vertices are very unlikely to feature in random graphs and small worlds, but they do appear in graphs sampled via a process proposed by Barabási and Albert (1999). In this process, a graph is constructed by starting from a small initial graph, to which more vertices are added one at a time. These are made adjacent with some k previously created vertices, where vertices with higher degree are more likely to be chosen as neighbors. This may be phrased as a tendency of new actors to attach themselves preferentially to already popular ones, increasing the popularity of the latter even further.

Even when the process appears plausible and the degree histogram of a preferential-attachment graph matches that of an observed structure fairly well, it would appear far-fetched to conclude that the actual formation process would have been similar. Like small worlds, preferential-attachment graphs only show that it is possible to construct graphs reproducing (insufficient) statistics.

Preferential attachment clearly yields graphs, in which vertex degrees generally decrease monotonically with creation time, and which are too clean and simple to reproduce more complex features generally found in empirical data.

4.3.2 Exponential-Family Random Graph Models

A family of models consisting of more general distributions is known as *exponential (family) random graph models* or simply the *ERGM* family.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

They are defined on the class of all directed graphs with a fixed number of n vertices and, therefore, easily specialized to undirected graphs.

The probability distributions in ERGMs are from the exponential family, hence the name. Therefore, a graph G has probability $P(G) = \frac{1}{Z} \cdot e^{H(G)}$, where Z is a normalizing constant ensuring that the probabilities sum to one and H is any function assigning a number to each graph. Note that probabilities are monotone in H and that the model solely depends on the choice of H. To model the inherent complexity of real data while limiting the complexity of the form of H, H is restricted to linear combinations of any number of statistics (network-level indices), s_i, weighted by corresponding coefficients θ_i.

An ERGM thus has probability distributions

$$P(G) = \frac{1}{Z(\theta)} \cdot \exp\left\{\sum_{i=1}^{k} \theta_i \cdot s_i(G)\right\}$$

for a fixed selection of statistics $s_1, \ldots, s_k$ and associated parameters $\theta = (\theta_1, \ldots, \theta_k)$. A straightforward example of a statistic is the edge count $m(G)$ and it can be checked that a $\mathcal{G}(n, p)$ random graph model is indeed a very simple ERGM with $H(G) = \left(\log \frac{p}{1-p}\right) \cdot m(G)$.

Typical ERGMs include $m(G)$ to control for density and then add several other substantively or otherwise meaningful statistics such as degrees, transitive triangles, or stars. In empirical research, model parameters θ are fitted to observed data. Parameter interpretation is not straightforward but this is a useful perspective: a parameter θ_i associated with statistic s_i measures the log-odds of forming an edge if that increases s_i by one, provided that the rest of the network remains constant.

As an example, let us consider the model used in Goodreau, Kitts, and Morris (2009) to investigate patterns of friendship formation among adolescents in the AddHealth data set (Resnick, Bearman, Blum, Bauman, Harris, Jones, Tabor, Beuhring, Sieving, Shew, Ireland, Bearinger, and Udry 1997). Three main factors are supposed to drive sociodemographic clustering, i.e., homophily in terms of sociodemographic attributes such as gender, race, or school grade.

The first factor is referred to as sociality, a personal characteristic, and included in the model as the total outdegree of all actors with the same attribute value for race, gender, and grade. The second factor is dyadic and called selective mixing. It is included by two types of statistics, the

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

number of ties within groups of equal attribute values (gender) and the number of ties for each combination of two different attribute values (race and grade). The third factor is called triad closure and modeled by the so-called geometrically weighted edge-wise shared partner statistic (Hunter and Handcock 2006),

$$s_{\mathrm{GWESP}}(G) = e^{\tau} \cdot \sum_{i=1}^{n-2} p_i(G) \left(1 - (1 - e^{\tau})^i\right) ,$$

where $p_i(G)$ counts the number of adjacent dyads with exactly i common neighbors and τ is a parameter modeling the decreasing marginal contribution of additional shared partners; the authors use $\tau = \frac{1}{4}$.

With total edge number and sociality complementing the sociodemographic variables as controls, Goodreau, Kitts, and Morris (2009) find evidence for both selective mixing and triadic closure, with hints at complex interaction effects. Disregarding the substantive results reported, the study illustrates nicely how relational models can be used to paint a more realistic picture of friendship formation. A more detailed introduction is provided in Robins, Pattison, Kalish, and Lusher (2007).

4.4 Summary

In principle, the analysis of networks is similar to the analysis of other, non-relational, data. However, for relations the unit of analysis is a dyad rather than a monad. Since dyads are heavily interdependent, and this interdependence is the actual substantive interest, other methods are required.

The link to an underlying network theory is best established through dyad indices because these evaluate the specific quality of relations between pairs of actors. Most methods essentially aggregate evaluations of particular dyads and they are generic in the sense that very often the dyad index can be chosen to account for a specific theoretical context.

While the methods of analysis discussed in Section 4.2 are often used for description, the quantities determined through them may also represent the core variables in inferential statistics. In addition, we have seen that modeling a population of networks, to which observations are contrasted, is a complex task and an area still under development.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

We have deliberately omitted longitudinal networks because the inclusion of time leads to a combinatorial explosion of cases to consider. This is because different kinds of dynamics can be present in attributes, structure, composition, and processes taking place in a network, externalities, and any combination thereof, and with respect to any of the types of analysis discussed. Let us simply note that the usual first steps are time series analyses of the variables determined on cross-sectional snapshots of a longitudinal network and the *stochastic actor-oriented modeling* approach (Snijders et al. 2010).

4.5 Exercises

1. Assume that friendship network data was collected using a fixed-choice design, in which a class of 27 school children were asked to list up to five best friends from within the class in rank order. What can we conclude about density and average degree in the asymmetric network of nominations? How would that be different if the boundary was specified as including all 234 children in the entire school, or in a free-choice questionnaire design?
2. Consider any undirected graph $G = (V, E)$. The so-called *Handshake Lemma* is a formal statement which asserts that the sum of all degrees is twice the number of edges, $\sum_{i \in V} \deg(i) = 2m$. Argue that this holds for all undirected graphs. Argue also, that the number of guests at a party who shake hands with an odd number of other guests is even, and that there are at least two guests who shake hands with the exact same number of other guests (although not necessarily with the same other guests).
3. Given a population sample of m men and f women, let the relation "is or has been married to" be represented in a rectangular $m \times f$-matrix with entries in the range $\{0, 1\}$.
   - What are the disadvantages of this two-mode network representation?
   - What is the maximum number of 3-cliques in such a network?
   - If we use a one-mode network representation instead, what are the dimensions of the corresponding adjacency matrix?
4. List all possible triad types for undirected networks.

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

5. A dyad consists of two actors and allows for two different orderings. Similarly, a triad consists of three actors and allows for six different orderings. Discuss why triadic configurations in dyadic data form a special case of *triadic data*, and give an example of triadic data that cannot be represented as triads of a network. Come up with two definitions of your own for what might be called *three-mode data* and discuss the differences.
6. Why is closeness centrality ill-defined on (directed) graphs that are not (strongly) connected? Can you think of alternative centrality measures that are based on the same idea but applicable to any (directed) graph?
7. It is often observed that random graphs and graphs with a core-periphery structure yield highly correlated values for degree, closeness, and betweenness centrality. Can you explain this finding? Construct a small undirected graph in which the most central vertices according to degree, closeness and betweenness are all different.
8. Why do structurally equivalent vertices have the same eigenvector centrality? Does this also hold for vertices that are regular equivalent?
9. Explain why graphs with the same triad census can have different clustering coefficient distributions. Give an example.
10. Give lower and upper bounds on the number of edges in the k-core of an undirected graph. Can you do the same for a clique of k vertices?
11. What is the difference between the triad census and the coefficients of triadic statistics in ERGMs?
12. Consider an ERGM for graphs with $n = 42$ vertices. Let the edge count $m(G)$ be its only statistic $s_1(G)$ with corresponding parameter $\theta_1 = -0.7$. Which graph has maximum probability, and what is (roughly) the expected number of edges of a graph sampled from this distribution?

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH

Copyright @ 2012. Campus Verlag.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

© Campus Verlag GmbH