Chapter 1
A First Encounter

Abstract The first chapter of the book gives a short overview of what network analysis does and why it is considered to be a vital part of complex system science: the network analytic framework allows to represent the interaction structure of a complex system as a complex network. The network's structure can then be analyzed by the application of several structural measures. However, there are two different branches in network analysis that either use the resulting values to find so-called *universal features* of complex systems or to allow a *contextual, semantic analysis*. The latter focuses on the connection between structure and function of a network with respect to the complex system of interest and some specific research question. There is a caveat, though: while, in principle, structural measures can be applied to all kinds of networks, if one is only searching for universal features, their results are not always interpretable with respect to a predefined research question. The term "network analysis literacy" is introduced to describe the knowledge of when to apply which measure to yield an interpretable result with respect to the complex system of interest.

1.1 Introduction to Network Analysis

Networks impress by their visual and intuitive quality: everyone of us is entangled in various friendship networks and business relationships, and the prospect of understanding the seemingly complex and erratic net of our personal relationships is an exciting one. Similarly, looking at scientific data in a new way, finding simple patterns that chip away individual noise to extract the main functional groups of entities in the complex system at hand, is surely one of the most gratifying moments in every scientist's life. Network analysis seems to be one of the most promising frameworks within which these two aspects, our personal life and our academic interest, can be combined, analyzed, and maybe even be understood. This prospect and the many exciting articles in journals such as Science, Nature, and PNAS, together with the interdisciplinary applicability of network analysis to various data sets and questions, has led to a tremendous interest in the methods provided by network analysis: Fig. 1.1 shows the dramatic increase of the number of articles with the keywords "network

K.A. Zweig, *Network Analysis Literacy*, Lecture Notes in Social Networks,
DOI 10.1007/978-3-7091-0741-6_1

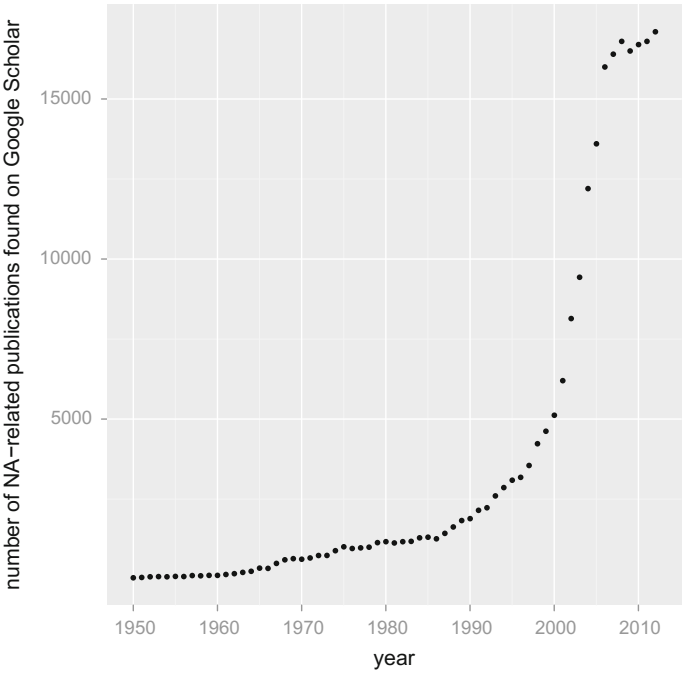

Fig. 1.1 Number of articles published in the given year containing the exact phrases "network analysis" and "complex network" as given by Google Scholar on the 12th of October, 2013

analysis" or "complex networks" as found by Google Scholar.[1] Starting from about 100 articles (as found on Google scholar) in the 1950s, the last years saw more than 15,000 articles with these terms.[2] This chapter gives a broad first encounter with network data by showing the first steps in analyzing a new set of network data.

The following chapters present a classic and widely used part of the toolkit in network analysis; but more importantly, they elaborate on the questions that are necessary to be answered in order to decide whether a given method is meaningful to the research question. One main caveat in network analysis is that once data is transformed into a graph representation, one can in principle apply any of the hundreds of network analytic methods to it—but not every method will compute meaningful and interpretable results with respect to the given data and the question

[1] For each year from 1950 to 2012, a Google Scholar search with both terms, connected by an "OR" was conducted. The number of results displayed was taken as the data point for the given year. The number of results is unlikely to hit the number of published articles in any way but gives at least an indication of the strongly increased interest in the topic.

[2] Note that double counting is as likely as an underestimation of the number of articles: articles with this topic may, for example, have been overlooked because they were published in a non-public journal which Google Scholar might not have access to. Again, the number given by Google scholar is only an indication of how many articles really have been published.

to answer. This book is thus not so much about introducing measures, many more can be found in the books by Wasserman and Faust [24], Newman [18], Borgatti et al. [6], or the book edited by Brandes and Erlenbach [7]. This book rather focuses on this last part, which I call *network analysis literacy*: it aims to empower its readers to know when to use which method so that they can quickly delve into the exciting analysis of networks themselves. Be warned that the technique by which this is accomplished follows the Socratic method which, in general, poses more questions than gives definite answers.

So, what is the first step in network analysis? One basic phase is the transformation of relational data into a complex network representation as described in the next section.

1.2 Data

The first question you might have is: what kind of data can actually be meaningfully represented as networks? A first answer is: almost any kind of data. The basic requirement is that there is a **distinct set of entities**, e.g., humans, organizations, proteins, computers, or books. The second requirement is that there is a known **relationship** between these entities. The information of whether any two entities are in the given relationship or not needs to be known for a large part of the entities since otherwise any kind of analysis will be quite shaky. Some obvious relationships between persons are: friendship, kinship, or employee-employer-relationships. Another interesting type of relationship is membership: it is a relationship between two different kinds of entities, namely persons and institutions, but that can also be represented by a network.

Relationships between non-human entities are equally abundant, and in many cases the resulting structures are also termed *networks* in our day-to-day language: examples are metabolic networks, protein-protein-interaction networks, neural networks, street networks, or computer networks. All of the above examples might be considered 'natural networks' but are there more abstract relationships that can also be represented as complex networks?

Interestingly, mathematicians have a very general understanding about what is a relation and what is not: in a mathematical sense two books can be defined to be "related" because their cover was created by the same designer. This "relatedness" does not mean that they are necessarily related in any colloquial sense: their content can of course be very different! Nonetheless, in a mathematical sense, the relation is meaningfully defined and can be easily checked by an external observer. In mathematics, a *relation* is simply defined as a subset of pairs of entities: $R \subseteq O \times O$, where $O \times O$ denotes the set of all possible pairs from some set of entities or objects O.

> **Note 2.** Mathematically, a *relation* R on a given set of entities or objects is just an arbitrary choice of pairs of these entities (objects), denoted by $R \subseteq O \times O$. In principle, any relation can be represented as a graph.

This is on the one hand much more general than the day-to-day notion of a relationship but on the other hand much less intuitive: a *relation* does not need to stand in any correlation to a real-world *relationship*; it can even represent a relationship that would not be seen as meaningful in the real world: for example, all humans with the same first name can be represented by a relation or all humans which share the same last digit of their ID-card number. Mathematical relations can also (meaningfully) be derived from other relations: One can build a second network based on the connection structure of another network by, for example, connecting two persons with each other if they share at least 8 friends in a friendship network. In this second network, there might be two persons that are connected because they share enough friends but which are not befriended themselves, and vice versa.

So, in this book, a *relationship* is something that can be observed in the real world, a *relation* is the mathematical structure which possibly represents a relationship. However, not all relations are associated with any relationship and the same relation, i.e., the same subset of pairs of a given set of elements, can represent different relationships. How is now a relationship turned into a complex network? This is discussed in the following.

1.2.1 From Relationship to Graph

In the moment a set of entities and a relationship of interest has been defined, there is a range of decisions to be made, to turn the concept of that relationship into a procedure that decides, for each pair of entities, whether they are in the associated mathematical relation or not. In most cases, when data is turned into a network representation, several decisions have to be made: if the relationship of interest contains a direction, is it necessary to include this information in the mathematical relation? Are there different levels of intensity of a given relationship and is it necessary to differentiate between them, by assigning weights to the pairs in the relation? Each of these decisions changes the set of available structural measures and the interpretation of the measure if it is applied to the network. Chapter 5 will explain in detail how data can be turned into networks. In any case, mathematically, a graph is the combination of a set of elements and a relation defined on these elements.

> **Note 3.** What is the difference between a (complex) *network* and a *graph*? The quick answer is that a graph is the *abstract representation* of a relation between entities while a network combines the graph with additional information about the entities and their **relationship** represented by the graph.

In most cases, a *complex network* represents only one set of entities (sometimes two) and **one relationship** between the entities, with some limited options on the attributes assigned to the (mathematical) relation and usually no attributes assigned to the elements. On the graph level, the elements are called *nodes* or *vertices*, and the pairs of nodes contained in the relation are called *edges*. The graph can indicate whether the relationship is directed by either containing a symmetric or asymmetric relation. In the first case, whenever (a, b) is contained in the relation, so is (b, a): for example, the graph can store information of whether Tim is father of Tom or vice versa (or none of that). It can also represent weights of the relationship, by assigning a weight to each element in the relation. Again, the graph itself does not store information about which entity is represented by which node, it is oblivious of any identity of the node. Thus, the *graph* is the more abstract representation which mainly concentrates on the connection structure.

The *network* makes the connection between the graph and the complex system whose interactions it represents. Especially, the network assigns entities to nodes. Furthermore, it is the set of all descriptions and observations of the system in which the entities and their relationship is valid. It can contain observations on the entities like the age and gender of a human actor or the year of publication of a film. It can also contain more than one type of relationship between the actors, or additional observations about the relationships between entities, like the duration of all calls between mobile phone users.

In summary, a *complex network* is a graph, in which the set of elements is associated with a set of entities, and in which the relation between these elements represents a relationship between the corresponding entities.

The distinction between a network and its graph is often not very important, and thus *network* and *graph* are used quite interchangeably in most articles and also in this book.

> **Note 4.** The promise of network analysis is that the abstraction of a complex system as represented by a complex network and its underlying graph still allows to infer something about the complex system of interest. That is actually a strong assumption and later chapters (e.g., Chaps. 10 and 14) elaborate preconditions to enable this transfer.

So, what are the first steps after the data is represented as a network? The following section shows some typical approaches to get a first impression of a new data set on the example of a movie-co-rating network.

Table 1.1 The movie-movie-similarity network contains 494 films and represents 9796 relationships between them

Statistics	Value
n	494
m	9796
$\rho(G)$	0.08

1.2.2 First Probes into the Data

The movie-co-rating network is a network deduced from a so-called *bipartite graph*: As indicated above, some data sets describe a relationship between two different entities, for example, how customer of a video rental store rate the films they rented. Such a data set documents a relationship between customers and the films they rated, but there is no direct relationship between any two customers or between any two films. This kind of data is represented by a bipartite graph, that is, one that can be split in two parts such that all known relationships are only between entities from different parts. Based on such data, one can compute a similarity measure between the films that quantifies whether the films have been more often liked by the same persons than expected or not; this technique is called a *one-mode projection* of a bipartite graph and described in Sect. 13.5.

Such a one-mode projection is the basis for the following demonstrations. It has been created such that the relation is undirected and it is assumed that a pair of films connected by an edge are similar in content. The data comes in a format[3] that is readable by various software applications, e.g., Gephi[4] which is well suited for visually exploring a medium-sized graph [22]. Similarly well suited are yEd [27], visone [13], or Cytoscape [26].

The very first useful information about the data is how many films it contains and how many relationships between them exists. In most visualizations of graphs, this information is given immediately when the graph is opened and displayed. In general, the number of entities is denoted by n and the number of relationships is denoted by m. With around 500 nodes and 10,000 edges, the network is of medium size. From these two basic statistics the so-called *density* of relationships can be computed as another, first inspection into the graph. It is defined as the number of existing relationships divided by the number of possible relationships: in principle, every pair of entities could be related to each other, thus, the number of possible relationships is given by $n(n-1)/2$. The density in the given data with $n = 494$ and $m = 9796$ can thus be computed to be 0.08. This density can also be interpreted as the *probability* that a randomly chosen pair of movies is related; this probability is obviously very small. Table 1.1 summarizes the basic statistics.

[3] Sections 3.6 and 3.7 discusses various graph data formats and how they can be transformed into each other.

[4] Freely downloadable from http://gephi.org/.

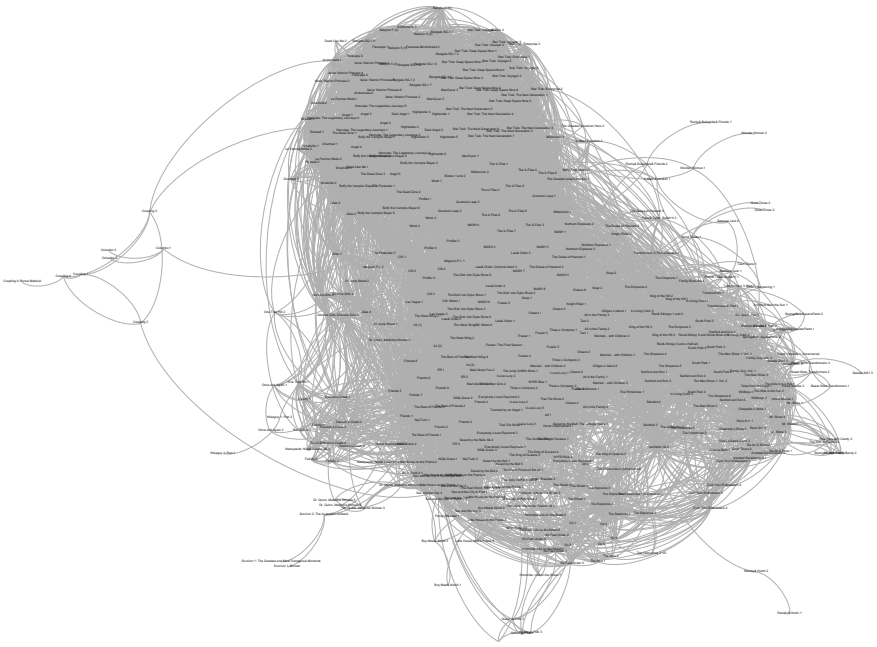

Fig. 1.2 A visualization of the series-series-similarity network which essentially looks like an insect, neatly packed by a spider for an early dinner

The next step is to visualize the data. How to best visualize relational data is a difficult question and there is a large community of scientists working on that topic.[5] For a review on many of the methods and software implementing them, see, e.g. [5, 14, 16, 19]. Most of the layout algorithms work best for small graphs with up to 100–200 nodes and a small density of around 0.01. However, even with up to 10,000 nodes a visualization can be helpful. In the given data, its density is the main problem, so the visualization of the full network looks a bit like a hairball (see Fig. 1.2).[6]

In general, layout algorithms try to find a position for each node such that most nodes are placed near their neighbors, that is, those nodes with which they are connected by an edge. Many of the algorithms are based on a force-directed layout approach in which connected nodes attract each other and unconnected nodes do not affect each other or repulse each other. Classic approaches of this kind are the popular Fruchterman-Reingold algorithm [10] and the Kamada-Kawai algorithm [15].

[5]The main conference for graph drawing related articles is the *International Symposium of Graph Drawing* and the main journal is the *Journal of Graph Algorithms and Applications*. An impressive free online archive of graph drawing related papers, the *Graph Drawing E-Print Archive* (**GDEA**), can be found at: http://gdea.informatik.uni-koeln.de.

[6]The figure was produced with the Force Atlas layout algorithm implemented in Gephi [22]. Subsequent processing of the figure was done in Inkscape [23].

In such a case it can make sense to reduce the data set to a clearly defined subgraph by choosing a suitable subset of nodes. Figure 1.2 presents the network where only movies which are part of series were selected, together with their similarity relationships. This layout still produces a quite dense and complicated representation. However, zooming into the figure reveals that the layout algorithm managed to place parts of the same series close to each other, and that similar series are located in the same area as well (see Fig. 1.3): series like *X-files* and *Buffy*, which are about supernatural forces, are next to each other and side-to-side with science-fiction series like the *StarTrek* and *Star Gate* series. Since *X-files* also shows crime-related aspects, it seems to be meaningful that its seasons are adjacent to other crime series like *CSI, Monk*, or *Profiler*. It is important to note that layout algorithms do not take into

Fig. 1.3 A clip of the network shown in Fig. 1.2. Meaningful patterns seem to emerge, as different seasons of the same series are positioned closed to each other

account any external information. In this case, the algorithm was not aware of any series titles, genre information, or any other information on the corresponding films. It computed the final positions just based on the connections of each of the nodes and tried to place nodes such that they are close to their neighbors and distant to non-neighbors. If such a content-oblivious algorithm is able to find a layout such that humans find intuitive patterns in the network, this is a good sign for the algorithm's abilities. But furthermore it is a sign of a special structure in the network data: we find that many complex networks are clustered, i.e., that it is possible to find groups of entities that are much more related to each other than to entities in other groups. By removing the edges in the visualization and highlighting the labels of all parts of the same series in the same color, the effect becomes even more drastic (Fig. 1.4).

With a large network, instead of choosing a meaningful subgraph like the series graph within the movie graph, it can also be insightful to visualize a random subgraph. Figure 1.5 shows such a visualization of an email contact network that is described below.

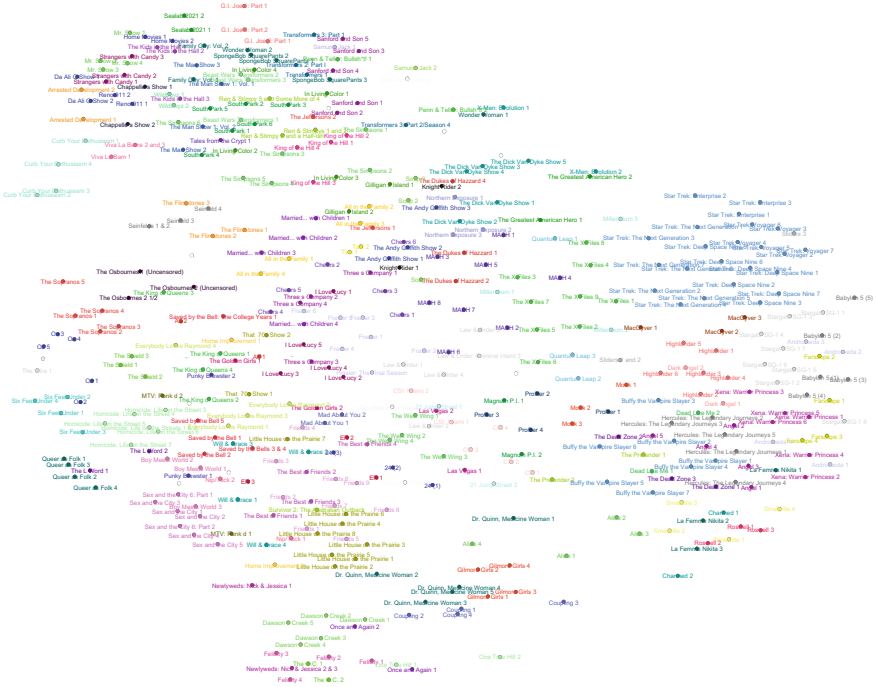

Fig. 1.4 All film titles that are part of the same series, have the *same color*. The edges are hidden

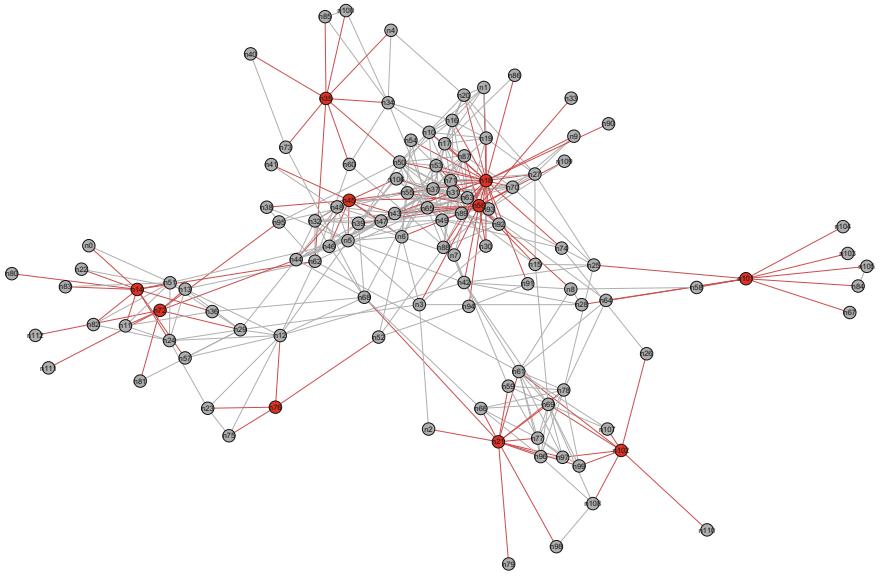

Fig. 1.5 A visualization of a random subgraph of an email contact network [11]

1.2.3 Measuring Indirect Effects

The most important reason for using network analytic methods is to model and analyze *indirect effects*. Basically, all direct effects in relations can be modeled and analyzed by methods from classical statistics—the advantage of using graphs as an abstraction is that in graphs one can compute how far apart two nodes are and thus how likely an **indirect effect** is between them.

The notion of *distance* in a network is a very intuitive one: the connection between two nodes by an edge is seen as a 'street' from the first node to the second. The distance between two nodes is then defined as the minimal number of 'streets' (edges) that need to be traversed to get from the first to the second node (s. Chap. 3 for a formal definition). This information can be of interest for various kinds of networks, e.g., real street networks, but also more abstract networks such as email contact networks. In the latter case, the distance between two nodes gives a notion of how many emails must be sent between direct acquaintances to get a rumor from the first sender to the last recipient. Sometimes, the average distance is not as interesting as the *maximal distance* between any two nodes, the so-called *diameter*. Exemplary, both measures are computed on two sample graphs: the first network represents the physical connections between so-called *autonomous systems*[7] in the Internet on the

[7] An autonomous system comprises a set of computers that are organized by a distinct entity, an Internet service provide, a company, or a university.

2nd January, 2000,[8] and the second network represents the email contact network of members of the university Rovira y Virgili in Tarragona, as compiled by Guimerá et al.[9] In both of these networks, the maximal distance describes how many edges a message needs to traverse to connect the two most far apart nodes with each other. Similarly, the *average distance* describes the expected distance between any two nodes chosen at random. For the email contact network, the maximal distance of any two members in the network is 8, and the average distance is 3.6. For the autonomous system network, the maximal diameter is 9, and the average distance is 3.7. The very similar average distance is surprising, as the order of the two networks is quite different: the email contact network represents the email contacts of 1,133 persons while the network between autonomous systems comprises 6,476 of these entities. It is, however, a common finding that many real-world networks have a comparably small distance, as discussed in Chap. 6.

To understand whether an observed average distance is unusual or unexpected, it is necessary to compare the observed result with the one in an appropriately chosen random graph. This question is discussed in detail all over the book, starting in Chap. 7.

1.2.4 Distributions

The measures introduced so far, like the density or the diameter, result in a single number. The distance is actually a measure between pairs of nodes and if it is computed for all pairs of nodes, it can be represented as a *distribution of values*. Figure 1.6 shows the distance distributions of the two networks used above, the autonomous system network and the email contact network. Although the order of the two networks differs by a factor of 30, the shape of the distributions is remarkably similar. If such a similar behavior of networks with very different origin can be quantified and proven, physicists speak of a *universal behavior*. Such a universal behavior can be supposed to be of importance for a network and the complex system the network is embedded in. The logic behind this assumption is that—without a benefit—a network's structure may be more or less random and only show structures expected in the case of randomness. Only if the emergence of a non-expected structure is beneficial, the system has some incentive to retain this non-expected structure. If it is furthermore possible to find a general mechanism which is simple and which causes this behavior, it is the first step

[8]The data was retrieved from http://snap.stanford.edu/data/as.html, and compiled by Leskovec et al. [17].

[9]Retrieved from http://deim.urv.cat/~aarenas/data/xarxes/email.zip [11]. The data only contains the biggest connected component.

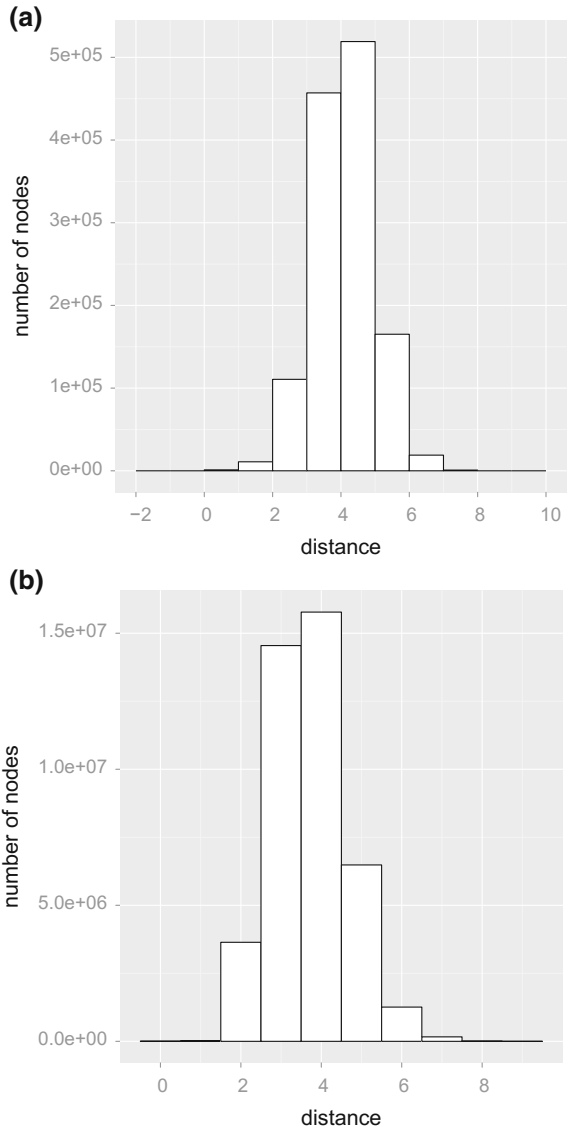

Fig. 1.6 Distance distribution of an email contact network between members of the university Rovira y Virgili [11]. Distance distribution of the connection network between autonomous systems in January, 2000 [17]

in understanding the self-organization of large and complex networks.[10] The *degree distribution* of networks was one of the first distributions of large networks that was analyzed in detail, and some of its possible shapes are discussed in Sect. 6.4.

[10]Of course, just because one kind of mechanism produces the behavior, it does not imply that all systems that show the behavior need to be built by this mechanism. See Chap. 12 for examples of this observation.

1.3 Network Analysis Literacy: A Primer

The first encounter with complex network analysis often leads to a strong excitement over the new methods and their ability to explain complex features of the system of interest. Moreover, due to the high availability of ready-to-use software packages and software applications, knowing the name of a method is already enough to apply it to the data of interest. It is in most cases not necessary anymore to implement a method yourself before applying it. However, this has led to various applications of measures in situations where the result is not easily interpretable anymore, as discussed in Chaps. 2 and 14.

This book's intent is to enable the reader to better understand when to apply which network analytic method to the observed data. The main reason for why this book is necessary is that too many methods are out there, searching for an application in vain, and too few of them are too popular and thus applied all the time. The second reason is that these popular measures are not known anymore by their structural formula, but mostly by a textual description like the following: "The betweenness centrality quantifies the centrality of nodes by measuring how often each node is on a shortest path between any other pair of nodes". First: such a textual description is very often not accurate enough. In the example, the textual description is not entirely correct as the classic betweenness centrality normalizes these values per pair of nodes. Thus, the implication of how the measure needs to be implemented is wrong (Chap. 9). Second, the textual description is so close to day-to-day-terms with a very specific meaning ("being central to a system") that most humans cannot resist to interpret a given result in that vein: "the node with the highest betweenness centrality is the most central node for the complex system of interest". The interpretation of centrality indices is a main focus of this book and for example discussed in Chap. 14. To give a feeling for how our mind induces meaning where no meaning is, the next section presents an analogy regarding the interpretation of a 2D-layout of a given graph, its *visualization*.

1.3.1 Visualizations

An important effect of many force-directed layout algorithms is that nodes that are considered to be *peripheral* in a network, are also often placed at the border of a layout. Along the same lines, the algorithms often put nodes that are close to most nodes in the center of a drawing. In other words, a visualization of a given graph might trigger the following chain of logical implications:

$$
\begin{aligned}
&\text{centrality in the drawing} \\
=\ &\text{centrality in the network} \\
=\ &\text{centrality of the represented entity} \\
&\text{in the complex system of interest?}
\end{aligned}
$$

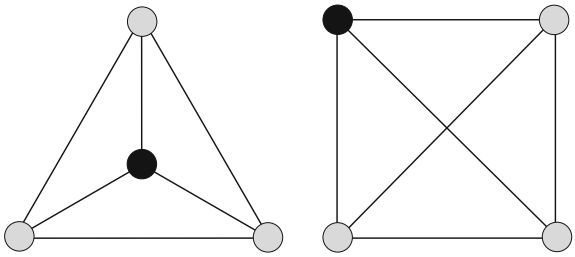

Fig. 1.7 In the *left* visualization, the *black* vertex seems to be more important for the network than in the *right* visualization. A closer look reveals that both visualizations show the same network

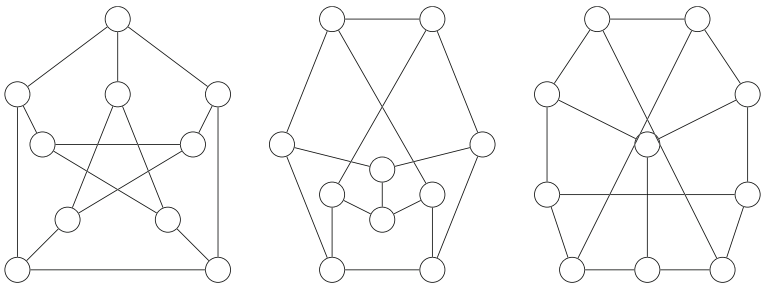

Fig. 1.8 All visualizations show exactly the same graph, namely the Petersen graph. While in Fig. 1.7 a closer look revealed that the displayed graphs are the same, this figure shows that this is not anymore the case for moderately larger networks

While for most algorithms there is a loose correlation between position in 2D and the centrality of a node, this intuitive behavior is not reliable in all individual cases. Thus, the position of a node in a 2D layout cannot always be correlated with the perceived "centrality" of the node in the network. Figure 1.7 shows the visualization of two networks: the left has a clear center in which the black node is placed. This seems to imply that the black node is especially central for the network. In the right layout, the drawing is symmetric and it seems that all nodes contribute equally to the network's structure. Of course, a closer look reveals that both layouts show exactly the same graph, i.e., there is a one-to-one correspondence between the nodes on both sides such that all corresponding nodes have exactly the same relationships to each other.[11]

An even stronger point is made in Fig. 1.8 in which three very different visualizations of **the same** graph are shown. It is difficult **not** to interpret the drawings: the first one seems to suggest that there are two distinct groups of five nodes that behave equivalently in their connection pattern within their respective group and between the groups. The right one again seems to indicate that there is one node which is more central than the others. The middle figure is more complex and does not lend itself to a quick interpretation.

[11] Such graphs are said to be *isomorphic*. See p. 178 for a formal definition.

These three example show again that the human eye is quickly deceived and that our brain is wired to interpret a node's positions in a 2D layout as its functional role in the network. Typical impressions based on a visualization are: "that node is an outsider, that one is central to the network". In the example shown in Fig. 1.8, either none or all of these interpretations are true as the same graph is shown in all three figures. However, the underlying graph is already so complex that it is not easy to verify that all layouts show the same graph—even after a very close look. And this network just contains 10 nodes!

> **Note 5.** All of these examples demonstrate that a visualization of a network can be both revealing and deceiving. This is why Gephi [22] with its beautiful visualizations and other visualization tools is perfect for exploration and hypothesis building; it is also the reason why statistical software packages or self-tailored applications are needed to collect **quantifiable evidence** that a given hypothesis is true.

In that vein: is there any node that is more central than the other nodes in the graph shown in Fig. 1.8? The graph is very famous in graph theory, it is the so-called *Petersen graph*. One way to construct it is to take all possible pairs of the numbers from 1 to 5 and to connect two of these pairs if and only if they do not share a number. That is, $(1, 3)$ is connected to $(2, 4)$ but not to $(3, 5)$. It is clear, that the "name" or *label* of the node determines its connection pattern. Thus, if in all labels, say, 2 and 3 are swapped, the relationships between the old nodes and the newly labeled nodes would still be the same. For example, under the old labeling, let $((1, 3), (2, 4))$ be connected. Then, under the new labeling, this edge between the old nodes $(1, 3)$ and $(2, 4)$ would now be between labels called $(1, 2)$ and $(3, 4)$. It would still be a valid connection in the sense of the Petersen graph. In general, if none of the two labels connected by an edge contain a 2 or a 3, nothing would change under the relabeling. If one label contains at least a 2 or 3, but the other does not, the connection is still valid after relabeling the node. There is no edge between nodes whose label share a number, thus, after relabeling there would also be no edge. If there is an edge and one label contains a 2 and one a 3, after relabeling, they would still not share a number and thus, the connection is still valid. This relabeling shows that all nodes have exactly the same kind of connection pattern and thus, there is no vertex being more central than the others. Only such a more insightful analysis is able to test the question of whether one of the nodes is more central than the others—the visualizations were not able to tell us.

Having stated that quantification is necessary to test hypotheses on a graph's structure, it is necessary to understand that there are two ways to deal with the resulting values: in the first approach, the question is only whether the numbers match if computed on networks from very different systems. In the second approach,

the numbers are interpreted in the context from which the network at hand was created. These two approaches are quickly summarized here and discussed in detail in Chap. 2.

1.4 Approaches to Network Analysis

In today's big field of "network science" and the older field of "social network analysis" there are two very different perspectives pursued by scientists: the first is a purely structural one in which models with as few assumptions as possible are built to explain as many of those structural features which are commonly found in complex networks (Fig. 1.9a). An example that most readers will have heard of are the so-called *small-worlds* found in nearly all complex systems.[12] Networks called a *small-world* share a small average distance and the feature that there is a pronounced local structure of the network in which most neighbors of most nodes are also connected to each other. Small-worlds and other *universal structures* are in depth discussed in Chap. 6. Structural features like these that are found in many complex networks are thought to be "universal" and to require "universal laws"; in that perspective, building a simple model of how complex networks evolve that is able to generate these universal features is a worthwhile endeavor. This perspective of universal structures came into focus after two seminal papers from statistical physicists were published [3, 25]. However, the second perspective is more common: here, social and other complex systems are represented as a complex network, structurally analyzed, and, finally, the found structure is interpreted with respect to its functionality in the complex system of interest (Fig. 1.9b).

In most cases, the first approach is driven by the availability of data (**data-driven approach**): one of the reasons why the field of complex network analysis emerged was that more and more data concerning complex relationships were freely available online [2, Sect. 1.3].[13] The second approach is in most cases **hypothesis-driven**, i.e., based on a hypothesis of how structure of a complex network and its function are related, together with a framework of how a certain kind of data is obtained to be represented by a network and structurally analyzed. As I will discuss throughout the book, starting in Chap. 2, only the latter can result in meaningful, semantic analyses that give an insight in the complex system of interest.

For a newcomer to the field, it is often not easy to disentangle these two perspectives, their approaches, and their respective methods. However, mixing them, i.e., using a purely structural model to explain the function of nodes or subgraphs in a given network, often results in misinterpretations. Similarly, using a structural measure in a context where it cannot be meaningfully interpreted with respect to the complex network of interest, may lead to unintended consequences.

---

[12] If you, dear reader, have not yet heard of it, go and read the famous paper by Watts and Strogatz [25]. See you later!

[13] The book is freely available at http://barabasi.com/networksciencebook/.

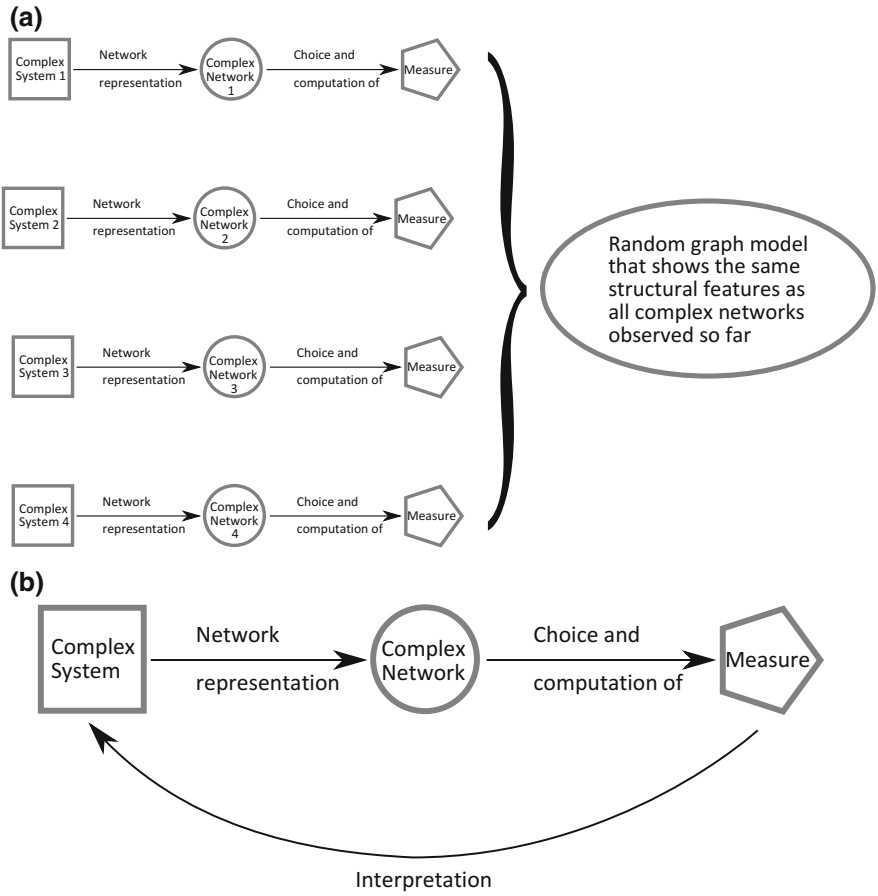

Fig. 1.9 There are two basically very different approaches to network analysis: **a** in the first approach, many different networks from very different contexts are structurally analyzed. If they share an important, non-trivial structural feature, a simple model is searched for to explain the evolution of this "universal" structural feature. **b** The second approach focuses on a given complex system, one network representation of it, analyzes it and tries to interpret the findings with respect to the complex system of interest

1.5 Outlook

Barabási has stated the potential of complex network analysis often, such as in a review from 2012 with the title "The network takeover":

> (…) a new network-based paradigm is emerging that is taking science by storm. It relies on data sets that are inherently incomplete and noisy. It builds on a set of sharp tools, developed during the past decade, that seem to be just as useful in search engines as in cell biology. It is making a real impact from science to industry. Along the way it points to a new way to handle a century-old problem: complexity [1].

This high potential to understand complex systems and finally explain and hopefully solve humanity's large problems, leads to a high responsibility. Wrong network analysis has lead to the discrediting of a famous climate scientist (read more in Sect. 15.5) and has suggested ways to support anti-HIV-campaigns that did not prove useful because the basic research was not well founded (read more in Sect. 14.7 and in Carter Butts' article on "Revisiting the foundations of network analysis" [8]). This book is intended to discuss the connection between a research hypothesis, a measure, and the interpretation of a measure's results for the complex system of interest, an aspect which I call *network analysis literacy*. If you are a data expert, this endeavor requires some motivation of you, my dear reader, to get used to mathematical equations and to really understand what they are measuring. If you are a method expert, this book encourages you to better understand the research question of your data expert or to see the limits of a given data set provided by someone else. This is necessary to choose the right method, in order to get interpretable results.

> **Note 6.** While it is absolutely true that the result of a formula is never wrong in the sense of "different than what it is supposed to be", the **application of the formula** might be a mismatch with the **intention** of what is to be measured.

1.6 Recommended Reading

There are many introductory articles and chapters to complex network analysis and all of them are worthwhile to get an overview. Consider to read the first chapter of the following textbooks:

1. Stanley Wasserman and Katherine Faust: "Social Network Analysis—Methods and Applications" [24].
2. Albert-László Barabási has worked on an e-book called "Network Science", which is now finally also published as a textbook [2]. It is still available for free at http://barabasilab.neu.edu/networksciencebook. Next to an introduction to the field, he has also added a personal introduction that gives an insight into his own interest in the field.
3. David Easley and Jon Kleinberg co-authored a book called "Networks, Crowds, and Markets". Here, the focus is more on processes taking place on networks, a very interesting book as it was written by a computer scientist and an economist. Available online as well [9].
4. Mark E.J. Newman published a book called "Networks—an Introduction". A big book where all the important equations introduced in the last years are explained in an accessible way, even for non-physicists. Some mathematical knowledge is helpful, though [18].

5. Marina Hennig et al.: "Studying social networks—a guide to empirical research" and Christina Prell: "Social network analysis" both have a more hands-on view on how to actually do **social network analysis**, including the construction of questionnaires to observe social relationships [12, 20].
6. Ulrik Brandes and Thomas Erlebach also provide an interesting view on complex network science in the introduction to the book they edited on methods in network analysis, contrasting the views of computer scientists and physicists [7].

There are some classic books to learn more about graph visualization: the one by Kaufmann and Wagner [16] and the one by di Battista et al. [5] focus on algorithms for the visualization of graphs, while the one by Jünger and Mutzel focuses on software for graph visualization [14]. In this book I mainly use `yEd` by yWorks [27] and `Gephi` [4, 22]. Another very worthwhile software is `visone` [13] which allows for a direct connection to the statistical analysis software `R` [21].

References

1. Barabási A-L (2012) The network takeover. Nat Phys 8:14–16
2. Barabási A-L (to be published) Network science. Cambridge University Press, Cambridge
3. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
4. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the third international AAAI conference on Weblogs and Social Media
5. Di Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph drawing: algorithms for the visualization of graphs. Prentice Hall
6. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323:892–895
7. Brandes U, Erlebach T (eds) (2005) Network analysis—methodological foundations. LNCS, vol 3418. Springer
8. Butts CT (2009) Revisiting the foundations of network analysis. Science 325(5939):414–416
9. Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge
10. Fruchtermann TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21(11):1129–1164
11. Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68:065103
12. Hennig M, Brandes U, Pfeffer J, Mergel I (2012) Studying social networks—a guide to empirical research. Campus
13. http://visone.info/index.html
14. Jünger M, Mutzel P (eds) (2004) Graph drawing software. Springer, Berlin
15. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inf Process Lett 31(1):7–15
16. Kaufmann M, Wagner D (eds) (2001) Drawing graphs: methods and models. Springer, Heidelberg
17. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters, and possible explanations. In: Proceedings of the 11th ACM SIGKDD
18. Newman ME (2010) Networks: an introduction. Oxford University Press, New York

19. Nishizeki T, Saidur Rahman M (2004) Planar graph drawing. World Scientific Publishing Co. Pte. Ltd., Singapore
20. Prell C (2011) Social network analysis. SAGE Publications Ltd., London
21. http://cran.r-project.org/
22. https://gephi.org/
23. http://www.inkscape.org/de/
24. Wasserman S, Faust K (1999) Social network analysis—methods and applications, revised, reprinted edn. Cambridge University Press, Cambridge
25. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
26. www.cytoscape.org
27. yWorks. yEd—Java^{TM} graph editor. http://www.yworks.com/en/products_yed_about.htm

http://www.springer.com/978-3-7091-0740-9

Network Analysis Literacy
A Practical Approach to the Analysis of Networks
Zweig, K.A.
2016, XXIII, 535 p. 126 illus., 14 illus. in color.,
Hardcover
ISBN: 978-3-7091-0740-9