This is a preprint of Katy Börner, Soma Sanyal and Alessandro Vespignani (2007) Network Science. In Blaise Cronin (Ed) Annual Review of Information Science & Technology, Volume 41. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 12, pp. 537-607.

Network Science

Katy Börner
School of Library and Information Science, Indiana University, Bloomington, IN 47405, USA
katy@indiana.edu

Soma Sanyal
School of Library and Information Science, Indiana University, Bloomington, IN 47405, USA
ssanyal@indiana.edu

Alessandro Vespignani
School of Informatics, Indiana University, Bloomington, IN 47406, USA
alexv@indiana.edu

This is a preprint of Katy Börner, Soma Sanyal and Alessandro Vespignani (2007) Network Science. In Blaise Cronin (Ed) Annual Review of Information Science & Technology, Volume 41. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 12, pp. 537-607.

1. Introduction

This chapter reviews the highly interdisciplinary field of network science, a science concerned with the study of networks, be they biological, technological, or scholarly networks. It contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science, and computer science. A coherent theoretical framework including static and dynamical modeling approaches is provided along with discussion of non-equilibrium techniques recently introduced for the modeling of growing networks. The chapter also provides a practical framework by reviewing major processes involved in the study of networks such as network sampling, measurement, modeling, validation and visualization. For each of these processes, we explain and exemplify commonly used approaches. Aiming at a gentle yet formally correct introduction of network science theory, we explain terminology and formalisms in great detail. Although the theories come from a mathematical, formulae laden world, they are highly relevant for the effective design of technological networks, scholarly networks, communication networks, and so on. We conclude with a discussion of promising avenues for future research.

At any moment in time, we are driven by and are an integral part of many interconnected, dynamically changing networks[1]. Our neurons fire, cells are signaling to each other, our organs work in concert. The attack of a cancer cell might have an impact on all of these networks and it also impacts our social and behavioral networks if we become conscious of the attack. Our species has evolved as part of diverse ecological, biological, social, and other networks over thousands of years. As part of a complex food web, we learned how to find prey and to avoid predators. We have created advanced socio-technical environments in the shape of cities, water and power systems, street and airline systems. In 1969, researchers started to interlink computers leading to the largest and most widely used networked infrastructure in existence: the Internet. The Internet facilitated the emergence of the World-Wide Web, a virtual network that interconnects billions of Web pages, datasets, services and human users. Thanks to the digitization of books, papers, patents, grants, court cases, news reports and other material, along with the explosion of Wikipedia entries, e-mails, blogs, and such, we now have a digital copy of a major part of humanity's knowledge and evolution. Yet, although the amount of knowledge produced per day is growing at an accelerating rate, our main means of accessing mankind's knowledge is search engines that retrieve matching entities and facilitate local search based on connections, for example, references or Web links. But, it is not only factual knowledge that matters. The more global the problems we need to master as a species, the more we need to identify and understand major connections, trends, and patterns in data, information and knowledge. We need to be able to measure, model, manage, and understand the structure and function of large, networked physical and information systems.

Network science is an emerging, highly interdisciplinary research area that aims to develop theoretical and practical approaches and techniques to increase our understanding of natural and man made networks. The study of networks has a long tradition in graph theory and discrete mathematics (Bollobas, 1998; Brandes & Erlebach, 2005), sociology (Carrington, Scott, & Wasserman, 2004; Wasserman & Faust, 1994), communication research (Monge & Contractor, 2003), bibliometrics/scientometrics (Börner, Chen, & Boyack, 2003; Cronin & Atkins 2000), Webometrics/cybermetrics (Thelwall, 2004), biology (Barabási & Oltvai, 2004; Hodgman 2000), and more recently physics (Barabási, 2002; Buchanan, 2002; Dorogovstev & Mendes, 2003; Pastor-Satorras & Vespignani, 2004; Watts, 1999). Consequently, there is impressive variety in the work styles, approaches and research interests among network scientists. Some specialize in the detailed analysis of a certain type of network, for example, friendship networks. Others focus on the search for common laws that might influence the structure and dynamics of networks across application domains. Some scientists apply existing network measurement, modeling and visualization algorithms to new datasets. Others actively develop new measurements and modeling algorithms. Depending on their original field of research, scientists will emphasize theory development or the practical effects of their results and present their work accordingly. Data availability and quality differ widely from large but incomplete and uncertain datasets to high quality datasets that are too small to support meaningful statistics. Some research questions require descriptive models to capture the major features of a (typically static) dataset, others demand process models that simulate, statistically describe, or formally reproduce the statistical and dynamic characteristics of interest. This variety, coupled with a lack of communication among scientists in different domains has lead to many parallel, unconnected strands of network science research and a diversity of nomenclature and approaches.

Today, the computational ability to sample and the scientific need to understand large-scale networks call for a truly interdisciplinary approach to network science. Measurement, modeling, or visualization algorithms developed in one area of research, say physics, might well increase our understanding of biological or social networks. Datasets collected in biology, social science, information science and other fields are used by physicists to identify universal laws. For example, unexpected similarities between systems as disparate as social networks and the Internet have been discovered (Albert & Barabási, 2002; Dorogovstev & Mendes, 2002; Newman, 2003). These findings suggest that generic organizing principles and growth mechanisms may give rise to the structure of many existing networks.

Network science is a very young field of research. Many questions have still to be answered. Often, the complex structure of networks is influenced by system-dependent local constraints on node interconnectivity. Node characteristics may vary over time and there may be many different types of nodes. The links between nodes may be directed or undirected, and may have weights and/or additional properties that may change over time. Many natural systems never reach a steady state and non-equilibrium models need to be applied to characterize their behavior. Furthermore, networks rarely exist in isolation but are embedded in "natural" environments (Strogatz, 2001).

This chapter reviews network science by introducing a theoretical and practical framework for the scientific study of networks. Although different conceptualizations of the general network science research process are possible, we adopt the process depicted in Figure 1. A network science study typically starts with an hypothesis or research question, for example, does the existence of the Internet have an impact on social networks or citation patterns? Next, an appropriate dataset is collected or sampled and represented and stored in a format amenable to efficient processing. Subsequently, network measurements are applied to identify features of interest. At this point the research process may proceed on parallel tracks concerning the analysis and/or modeling of the system at hand. Given the complexity of networks and the obtained results, the application of visualization techniques for the communication and interpretation of results is important. Interpretation frequently results in the further refinement (for example, selection of different parameter values or algorithms) and re-run of sampling, modeling, measurement and visualization stages. As indicated in Figure 1, there is a major difference between *network analysis* that aims at the generation of descriptive models which explain and describe a certain system and *network modeling* that attempts to design process models that not only reproduce the empirical data but can also be used to make predictions. The latter models provide insights into why a certain network structure and/or dynamics exist. They can also be run with different initializations or model parameters to make predictions for "what if" scenarios, for example: If the National Science Foundation (NSF) decided to double its budget over the next five years, what would be the impact in terms of numbers of publications, patents and citations?

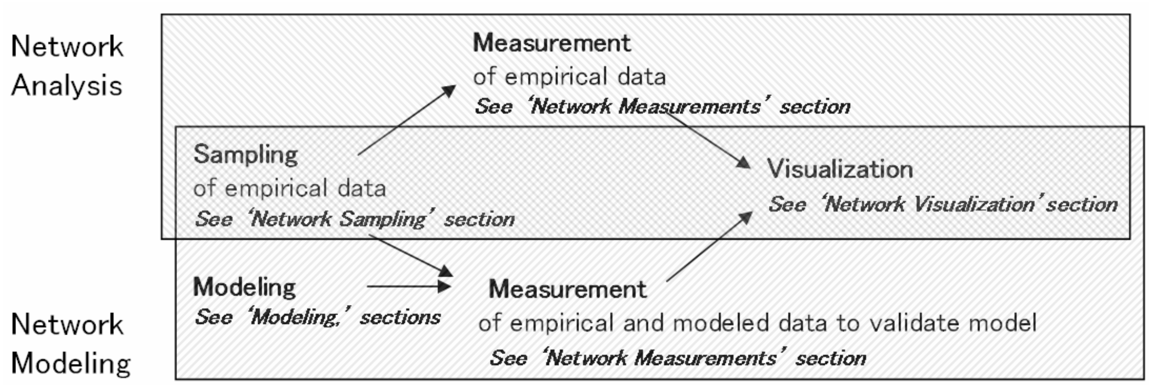


Figure 1. General network science research process.

Figure 1 also indicates which sections of this chapter explain the different stages of the research process. This chapter aims to provide a gentle introduction to the affordances and needs that the different network science disciplines pose. The background knowledge, pre-conceptualizations and the way of conducting science that the different disciplines employ vary widely. Yet, being able to translate among the different conceptualizations and terminologies and to identify similarities and differences among network science algorithms, concepts and approaches is the basis for effective collaboration and the exchange of techniques and practices across disciplinary boundaries. Whenever possible, we will point out commonalities and differences, alternative terminology and the relevance of alien looking concepts to core information science questions such as: How does one ensure that technological infrastructures (Internet, WWW) are stable and secure? What network properties support/hinder efficient information access and diffusion? What is the structure of scholarly networks, how does it evolve and how can it be used for the efficient communication of scholarly knowledge?

The remainder of this review is organized as follows: Section 2 introduces notions and notations used throughout this chapter. Section 3 discusses the basics of network sampling as the foundation of network analysis or modeling. Section 4 presents basic measurements and some examples. Section 5 discusses the major elements of a unifying theoretical framework for network science that aims to contrast, compare and integrate major techniques and algorithms developed in diverse fields of science. Section 6 reviews dynamic network models. Section 7 provides an overview of network visualization techniques as a means of interpreting and effectively communicating the results of network sampling, measurement and/or modeling. Section 8 discusses challenges and promising avenues for future research.

2. Notions and Notations

In this section we provide the basic notions and notations needed to describe networks. Not surprisingly, each field concerned with network science has its own nomenclature. The natural framework for a rigorous mathematical description of networks, however, is found in graph theory and we adopt it here. Indeed, graph theory can be traced back to the pioneering work of Euler to solve the Königsberg bridges problem (Euler, 1736). Building on the introduction of the *random graph model* by Erdös and Rényi (1959) (see also the section on modeling static networks) it has reached a maturity in which a wealth of rigorous mathematical yet practically relevant results is available for the study of networks. The main sources for the subsequent formalizations are the books by Chartrand & Lesniak (1986) and Bollobas (1998). It is our intention to select those notions and notations that are easy to understand for the general *ARIST* audience and sufficient to introduce the basic measurements, models and visualization techniques introduced in the subsequent sections.

2.1 Graphs and Subgraphs

Networks—subsequently also called graphs—have a certain structure (or topology) and can have additional quantitative information. The structure might be rooted or not and directed or undirected. Quantitative information about types, weights or other attributes for nodes and edges might exist. This section introduces different types of networks, their definition and representation. We start with a description of graph structure.

2.1.1 Undirected graphs

An undirected graph *G* is defined by a pair of sets *G = (V,E)*, where *V* is a non-empty countable set of elements, called *nodes* or *vertices* and *E* is a set of *unordered* pairs of different nodes, called *edges* or *links*. We will refer to a node by its order *i* in the set *V*. The edge (*i, j*) joins the nodes *i* and *j*, which are said to be *adjacent*, *connected*, or *neighbors*. The total number of nodes in the graph equals the cardinality of the set *V* and is denoted as *N*. It is also called the *size* of the graph. The total number of edges equals the cardinality of the set *E* and is denoted by *M*. For a graph of size *N*, the maximum number of edges is *N(N-1)/2*. A graph in which all possible pairs of nodes are joined by edges, that is, *M = N(N-1)/2*, is called a *complete N-graph*. Undirected graphs are depicted graphically as a set of dots, representing the nodes, joined by lines between pairs of nodes that represent the corresponding edges, see Figure 2a-d

2.1.2 Directed graphs

A directed graph *D*, or digraph, is defined by a non-empty countable set of nodes *V* and a set of *ordered* pairs of different nodes E_D that are called directed edges. In a graphical representation, the ordered nature of the edges is usually depicted by means of an arrow, indicating the direction of an edge, see also Figure 2e and 2f. Note that the presence of an edge from *i* to *j*, also referred to as *i < j*, in a directed graph does not necessarily imply the presence of the reverse edge *i > j*. This fact has important consequences for the connectedness of a directed graph, as we will discuss later in this section.

2.1.3 Trees

A tree graph is a hierarchical graph where each edge (known as a child) has exactly one parent (node from which it originates). If there is a parent node from which the whole structure arises then it is known as the *rooted tree.* It is easy to prove that the number of nodes in a tree equals the number of edges plus one, that is, *N = E+1*. The deletion of any edge will break a tree into disconnected components.

2.1.4 Multigraphs

The definition of both graphs and digraphs do not allow the existence of *loops* (edges connecting a node to itself) nor *multiple edges* (two nodes connected by more than one edge). Graphs with either of these two elements are called *multigraphs* (Bollobas, 1998). Most networks of interest to the *ARIST* readership are not multigraphs. Hence subsequently we discuss definitions and measures which are applicable to undirected graphs and directed graphs but not necessarily to multigraphs.

2.1.5 Graph Representation

From a mathematical point of view, it is convenient to define a graph by means of an *adjacency matrix* $x = \{x_{ij}\}$. This is an $N \times N$ matrix defined such that $x_{ij} = 1$ if $(i, j) \in E$ and $x_{ij} = 0$ if $(i, j) \notin E$. For undirected graphs the adjacency matrix is symmetric, $x_{ij} = x_{ji}$, and therefore it conveys redundant information. For directed graphs, the adjacency matrix is not necessarily symmetric. Figure 2 shows the adjacency matrices and graphical depictions for four undirected (a-d) and two directed graphs (e and f). Note that the adjacency matrix is also called a *sociomatrix* in the social network literature.

2.1.6 Subgraphs

A graph $G' = (V', E')$ is said to be a *subgraph* of the graph *G = (V, E)* if all the nodes in V' belong to *V* and all the edges in E' belong to *E*, that is, $E' \subseteq E$ and $V' \subseteq V$. The graphs in Figure 2b, d, and f are subgraphs of the graphs shown in Figure 2a, c and e, respectively. A *clique* is a complete *n*-subgraph of size *n* < N. For example, the graph in Figure 2b is a *3*-subgraph of the complete N-graph shown in Figure 2a.

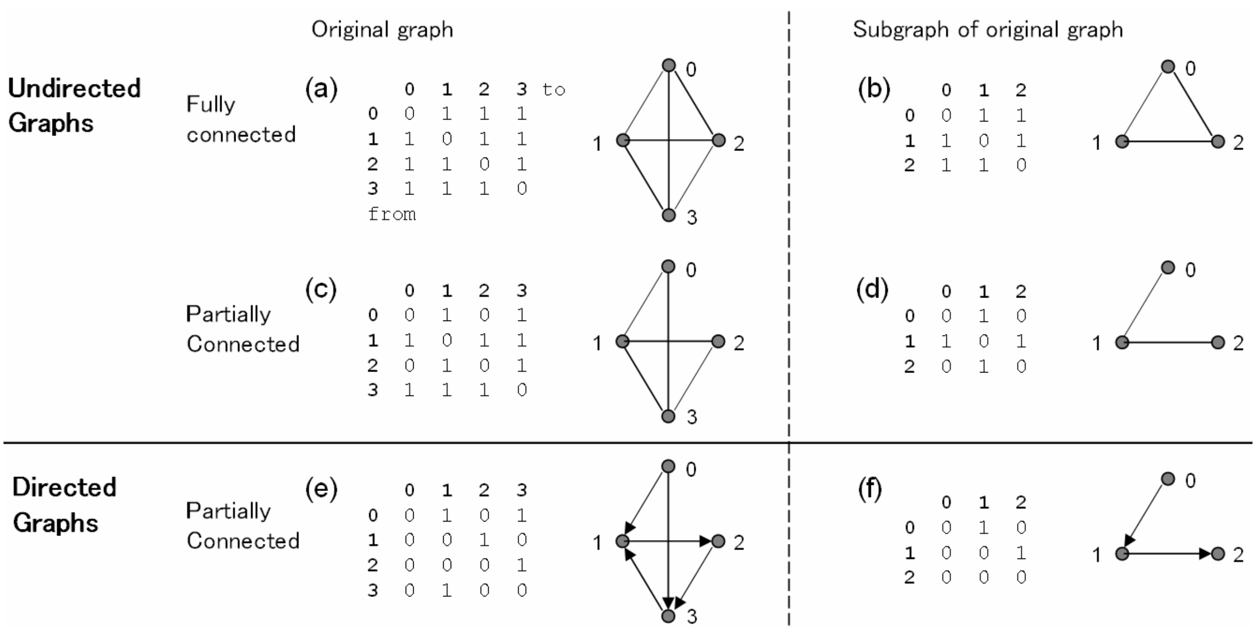


Figure 2: Adjacency matrix and graph presentations of different undirected and directed graphs.

The definitions so far have been qualitative describing the structure of a graph. However, we can also have quantitative information about a graph such as weights for edges.

2.1.7 Weighted Graphs

Many real networks display a large heterogeneity in the capacity and intensity values of edges. For example, in social systems, the strength and frequency of interactions is very important in the characterization of the corresponding networks (Granovetter, 1973). Similarly, the amount of traffic among Internet routers (Pastor-Satorras, & Vespignani, 2004) or the number of passengers using different airlines (Barrat, Barthelemy, Pastor-Satorras, & Vespignani, 2004; Guimera, Mossa, Turtschi, & Amaral, 2005) are crucial quantities in the study of these systems.

Where data are available, it is therefore desirable to go beyond the mere topological representation and to construct a weighted graph where each edge *(i,j)* is associated with a weight w_{ij} representing the intensity or value of the connection. As with the adjacency matrix $x = \{x_{ij}\}$, it is possible to define a weighted adjacency matrix $W = \{w_{ij}\}$. Like the adjacency matrix, the weighted adjacency matrix can be used to represent undirected weighted graphs where $w_{ij}=w_{ji}$ and directed weighted graphs with $w_{ij}\neq w_{ji}$ (however this may not be true always). Altogether, the weighted graph representation provides a richer description because it considers the topology along with quantitative information.

2.1.8 Bipartite Graphs

A simple undirected graph is called bipartite if it has two distinctly different sets of nodes which can be decomposed into two independent sets. It is often represented as $G=(V_1+V_2,E)$, where V_1 and V_2 are the two independent sets.

2.2 Graph Connectivity

There is a standard set of nodes, edge and graph measurements that is commonly used in graph theory and introduced in this subsection. The section on network measurements reviews additional measurements commonly used by network scientists. Table 2 in the section on discussion and exemplification of network measurements depicts common measures.

2.2.1 Node Degree

In undirected graphs, the degree k of a node is termed the number of edges connected to it. In directed graphs, the degree of a node is defined by the sum of its in-degree and its out-degree, $k_i = k_{in,i} + k_{out,i}$, where the *in-degree* $k_{in,i}$ of the node i is defined as the number of edges pointing to i; its *out-degree* $k_{out,i}$ is defined as the number of edges departing from i. In terms of the adjacency matrix, we can write

$$k_{in,i}=\sum_j A_{ji}\ ,\ \ k_{out,i}=\sum_j A_{ij}\ . \tag{1}$$

For an undirected graph, with a symmetric adjacency matrix, $k_{in,i}=k_{out,i}\equiv k_i$ holds. For example, node 1 in Figure 2a has a degree of three. Node 1 in Figure 2e has an in-degree of two and an out-degree of one.

2.2.2 Nearest Neighbors

The nearest neighbors of a node i are the nodes to which it is connected directly by an edge, so the number of nearest neighbors of the node is equal to the node degree. For example, node 1 in Figure 2a has nodes 0, 2, and 3 as nearest neighbors.

2.2.3 Path

A *path* P_{i_0,i_n} that connects the nodes i_0 and i_n in a graph $G = (V, E)$ is defined as an ordered collection of n+1 nodes $V_P=\{i_0,i_1,...,i_n\}$ and n edges $E_P=\{(i_0,i_1),(i_1,i_2),...,(i_{n-1},i_n)\}$, such that $i_\alpha\in V$ and $(i_{\alpha-1},i_\alpha)\in E$, for all α. The *length* of the path P_{i_0,i_n} is n. For example, the path in Figure 2f that interconnects nodes 0, 1, and 2 has a length of two.

2.2.4 Cycle

A *cycle* is a closed path $(i_0=i_n)$ in which all nodes and all edges are distinct. For example, there is a path of length three from node 1 to node 2 to node 3 and back to node 1 in Figure 2e. A graph is called *connected* if there exists a path connecting any two nodes in the graph, see, for example, Figure 2a and 2b.

2.2.5 Reachability

A very important issue is the *reachability* of different nodes, that is, the possibility of going from one node to another following the connections given by the edges in a network. A node is said to be reachable from another node if there exists a *path* connecting the two nodes, even if it goes through multiple nodes in between.

2.2.6 Shortest Path Length

The shortest path length ℓ_{ij} is defined as the length of the shortest path going from nodes *i* to *j*. In the following, we will use l_S to refer to a continuous variable which may represent any value of length.

2.2.7 Diameter

The diameter d_G is defined as the *maximum shortest path length* l_S in the network. That is, the diameter is the longest of all shortest paths among all possible node pairs in a graph. It states how many edges need to be traversed to interconnect the most distant node pairs.

2.2.8 Size

The *size* of a network is the *average shortest path length* $\langle l_s \rangle$, defined as the average value of ℓ_{ij} over all the possible pairs of nodes in the network. Because some pairs of nodes can have the same value for the shortest path length, we can define $P_l(l_s)$ as the probability of finding two nodes being separated by the same shortest length l_s. The size of the network can then be obtained by using this probability distribution as well as the individual path lengths between different nodes.

$$\langle \ell_s \rangle = \sum_{\ell} \ell_s P_\ell(\ell_s) \equiv \frac{2}{N(N-1)} \sum_{i<j} \ell_{ij} \ . \qquad (2)$$

The average shortest path length is also called *characteristic path length*. In the physics literature, the average shortest path length has been also referred to as the *diameter* of a graph. By definition, $\ell_{ij} \leq l_S$ holds. If the shortest path length distribution is a well behaved and bounded function, that is, a continuous function that has a defined starting and end point, then it is possible to show heuristically that in many cases the characteristic path length and the shortest path length have the same increasing behavior with increasing graph size.

2.2.9 Density

The *density* of a graph is defined as the ratio of the number of edges in the graph to the square of the total number of nodes. If the number of edges in a graph is close to the maximum number of edges possible between all the nodes in the graph, it is said to be a dense graph. If the graph has only a few edges, it is said to be a sparse graph.

2.2.10 Graph Components

A *component C* of a graph is defined as a connected subgraph. Two components $C_1 = (V_1, E_1)$ and $C_2 = (V_2, E_2)$ are disconnected if it is not always possible to construct a path $P_{i,j}$ with $i \in V_1$ and $j \in V_2$. A major issue in the study of graphs is the distribution of components, and in particular the existence of a *giant component G*, defined as a component whose size scales with the number of nodes of the graph, and therefore diverges in the limit $N \to \infty$. The presence of a giant component implies that a large fraction of the graph is connected, in the sense that it is possible to find a way across a certain number of edges, joining any two nodes.

The structure of the components of directed graphs is somewhat more complex as the presence of a path from the node *i* to the node *j* does not necessarily guarantee the presence of a corresponding path from *j* to *i*. Therefore, the definition of a giant component becomes fuzzy. In general, the component structure of a directed graph can be decomposed into a *giant weakly connected component* (GWCC), corresponding to the giant component of the same graph in which the edges are considered as undirected, plus a set of smaller *disconnected components* (DC), see Figure 3. The GWCC is itself composed of several parts due to the directed nature of its edges: The *giant strongly connected component* (GSCC), in which there is a directed path joining any pair of nodes. The *giant IN-component* (GIN), formed by the nodes from which it is possible to reach the GSCC by means of a directed path. The *giant OUT-component* (GOUT), formed by the nodes that can be reached from the GSCC by means of a directed path. Last but not least there are the *tendrils* that contain nodes that cannot reach or be reached by the GSCC (among them, the *tubes* that connect the GIN and GOUT) that form the rest of the GWCC.

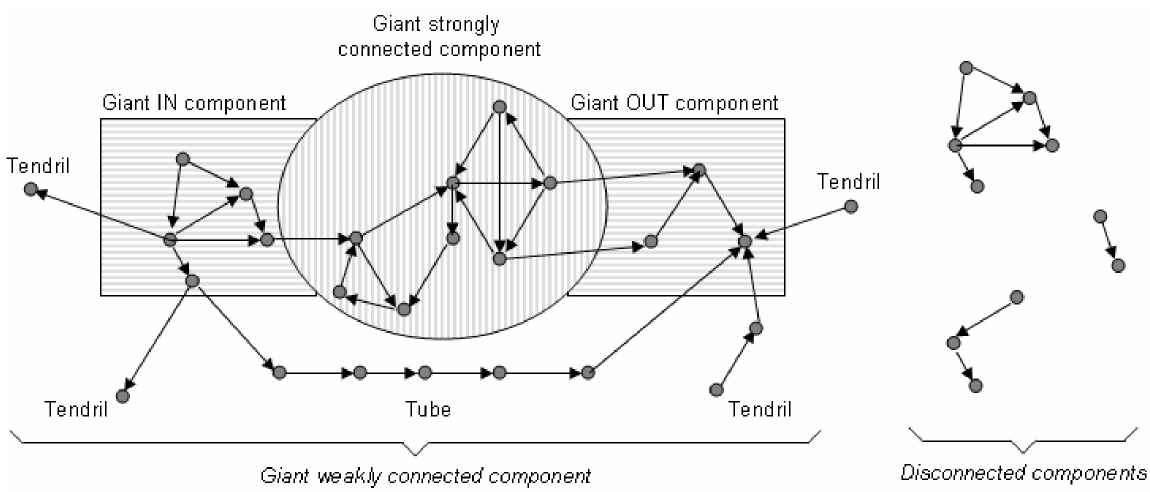


Figure 3. Component structure of directed networks such as the WWW. Adopted from Broder et al. (2000). The component structure of directed graphs has important consequences for the accessibility of information in networks such as the World-Wide Web (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, et al., 2000; Chakrabarti, Dom, Gibson, Kleinberg, Kumar, Raghavan, et al., 1999).

3. Network Sampling

Using the foregoing notions and notations, this section provides a short discussion of the issues related to the gathering of network data. Different application domains have very different affordances ranging from the size, type and richness of network data to the scientific questions that are asked. In some application domains it is relatively easy to gain access and work with a complete network dataset such as social network studies of smaller social groups, for example, all school children in a certain grade at a certain school. However, for many applications the acquisition of a complete network dataset is impossible due to time, resource or technical constraints. In this case, network sampling techniques are applied to acquire the most reliable dataset that exhibits major properties of the entire network. Network sampling thus refers to the process of acquiring network datasets and the discussion of statistical and technical reliability. Sampling may be based on the *features of nodes and or links* or based on the *structure of the network.* For example, a dataset could be compiled by selecting "all papers that cite a set of core papers" or "all Web pages that link to the home page of a certain research group." Sampling based on node and edge features refers to the selection of a subset of nodes and/or edges that match or exceed a certain attribute value. For example, in some application domains it is reasonable to select a set of nodes with certain attributes, for example, "all Web pages of universities in California," "all papers that have been cited at least once," or "all computers that had a computer virus in the last year." Sampling based on the structure of a network is very common in Internet studies, large-scale social network analysis, semantic network studies, Webometrics and scientometrics. Here the structure of the network (and not the attribute values of single nodes or edges) is exploited to acquire a meaningful subset of a network. Link-tracing designs such as *snowball sampling* are applied to acquire information about all nodes connected to one given node. Link tracing can be performed in a recursive manner resulting quickly in rather large datasets. This sampling strategy is considered the most practical way of acquiring social network data of hidden and hard-to-access human populations or of datasets of unknown size. Crawling strategies to gather WWW data rely on exhaustive searches by following hyperlinks. Internet exploration consists in sending probes along the computer connections and storing the physical paths of these probes. These techniques can be applied recursively or repeated from different vantage points in order to maximize the discovered portion of the network. That is, an initial dataset is acquired in a "first wave." And a subset of the features or nodes of this first wave dataset is used as a query/starting point for the "second wave" sampling.

It is clear that sampling techniques may introduce statistical biases. Therefore, a large number of *model based techniques*, such as *probabilistic sampling design* (Frank, 2004) developed in statistics, provide guidance in the selection of the initial datasets. These techniques try to quantify the statistical biases that may be introduced during the sampling stage. Better knowledge of these biases helps us to draw inferences that have less uncertainty, which in turn increases the confidence in the tests.

In many cases, however, the discovery process is constrained by the available techniques. For example, crawling strategies on the Internet usually have intrinsic biases due to the directed nature of the exploration that cannot be avoided. These biases may lead to wrong conclusions. For example, even though it is widely known that the Internet has a power-law degree distribution, it is possible to show that sampling biases can cause a Poissonian degree distribution to appear as a power law distribution (Clauset & Moore, 2005). So, it is difficult to describe whether the Internet is truly a power law distribution or not. For this reason, each particular sampling process requires a careful study of the introduced biases and the reliability of the obtained results. The recent explosion in large-scale data gathering has spurred several studies devoted to the bias contained in the sampling of specific information networks (Clauset & Moore, 2005; Dall'Asta, Alvarez-Hamelin, Barrat, Vazquez, & Vespignani, 2005; Lakhina, Byers, Corvella, & Xie, 2002; Petermann & De Los Rios, 2004).

Finally, there are other sources of biases relating to the intrinsic experimental error of specific sampling methods. In some cases, this causes a false positive or negative on the presence of a node or edge. High throughput techniques in biological network measurements, such as in experiments for detecting protein interactions (Bader & Hogue, 2002; Deane, Salwinski, Xenarios, & Esenberg, 2002), are a case in point. For these reasons, it is important to test the results obtained against null models, which are pattern generating models that replace the mechanisms thought to be responsible for a particular pattern with a randomization. The randomization produces a null statistical distribution for the aspect of the pattern controlled by the replaced mechanism. The observed empirical values are compared with the null distribution which is then used to assess the importance of the replaced mechanism. So, in all these sampling cases a careful scrutiny and examination of the data quality and the test of the results obtained against null models are important elements for the validation of the corresponding network analyses.

4. Network Measurements

Basic measurements for the characterization of networks can be divided into measures for properties of nodes and edges, local measures that describe the neighborhood of a node or the occurrence of subgraphs and motifs and global measures analyzing the interconnectivity structure and statistical properties of the entire network. Note that some node/edge measures as well as some local measures require the examination of the complete network under examination. We next review the standard set of measures and statistical observables commonly used in network science. The section concludes with a discussion of network types and an exemplification of the different measures.

4.1 Node and Edge Properties

4.1.1 Nodes

There exists a multitude of measures that characterize node properties (Hanneman & Riddle, 2005). The *degree* of a node (see definition in the section on graph connectivity) is a very basic indicator of the centrality of a node. Obviously, it is a local measure that does not take into account the global properties of the network. The *Bonacich power index* not only takes into account the degree of a node but also the degree of the nodes connected to a node. For example, the more connections a social actor has in its neighborhood, the more central/powerful it is. *Closeness centrality* approaches compute the distance of a node to all others. *Reach centrality* computes what portion of all other nodes can be reached from a node in one step, two steps, three steps, and so on. The *eigenvector approach* is an attempt to find the most central node in terms of the "global" or "overall" structure of the network. It uses factor analysis (Kim & Mueller, 1978) to identify "dimensions" of the distances among nodes. The dimensions are associated with an unit "eigenvector." The location of each node with respect to each dimension is called an "eigenvalue." Each unit eigenvector is associated with an eigenvalue. Once the unit eigenvectors and their corresponding eigenvalues are known, one can construct a "general eigenvector" as a matrix whose columns are the unit eigenvectors. The collection of eigenvalues is then expressed as a diagonal matrix associated with the general eigenvector. It is assumed that the first dimension captures the global aspects of distances among nodes and the higher dimensions capture more specific, local sub-structures. *Betweenness centrality* is a measure that aims to describe a node's position in a network in terms of the flow it is able to control. As an example, consider two highly connected subgraphs that share one node but no other nodes or edges. Here, the shared node controls the flow of information, for example, rumors in a social network. Any path from any node in one subgraph to any node in the other subgraph leads through the shared node. The shared node has a rather high betweenness centrality. Mathematically, the betweenness centrality is defined as the number of shortest paths between pairs of nodes that pass through a given node (Freeman, 1977). More precisely, let $L_{h,j}$ be the total number of shortest paths from h to j and $L_{h,i,j}$ be the number of those shortest paths that pass through the node i. The betweenness b of node i is then defined as $b_i = \sum L_{h,i,j} / L_{h,j}$, where the sum runs over all h,j pairs with $j \neq h$. An efficient algorithm to compute betweenness centrality was reported by Brandes (2001). The betweenness centrality is often used in transportation networks to provide an estimate of the traffic handled by different nodes, assuming that the frequency of use can be approximated by the number of shortest paths passing through a given node. It is important to stress that while the betweenness centrality is a local attribute of any given node, it is calculated by looking at all paths among all nodes in the network and therefore it is a measure of the node centrality with respect to the global topology of the network.

The above definitions of centrality rely solely on topological elements. When data on the edge weights w is available, then the centrality of a node can be computed based on the intensity or flows associated with the node. This type of centrality is commonly called the *strength* s of a node i and is formally defined as $s_i = \sum_j w_{ij}$.

4.1.2 Edges

The *betweenness centrality of edges* can be calculated analogously to the node betweenness as the number of shortest paths among all possible node pairs that pass through a given edge. Edges with the maximum score are assumed to be important for the graph to stay interconnected. These high scoring edges are the "weak ties" that interconnect clusters of nodes. Removing them frequently leads to unconnected clusters of nodes. The importance of weak ties was first examined by Granovetter (1973). Weak ties are particularly important for decreasing the average path length among nodes in a network, for speeding up the diffusion of information or for increasing the size of one's network for a given path length. However, networks with many weak ties are more fragile and less clustered.

4.2 Local Structure

This subsection discusses commonly used local network measures that describe the level of cohesiveness of the neighborhood of a node/edge and the occurrence of specific patterns or structures such as cliques and components.

4.2.1 Clustering Coefficient

The *clustering coefficient C* indicates the degree to which k neighbors of a particular node are connected to each other. It can be used to answer the question "are my friends also friends of each other?" The clustering coefficient should not be confused with measures used to identify how good a particular clustering of a dataset is, for example, in how far the similarity between clusters is minimal while similarity within a cluster is maximal. The clustering coefficient is commonly used to identify whether a network is a lattice, small world, random network or a scale-free network.

Two definitions of C are commonly used. Both use the notion of a *triangle D* that denotes a clique of size three, that is, a subgraph of three nodes that is fully connected. Basically, this means looking at cases where the node i has a link to node j and j has a link to m, then ask whether i is linked to m or not. If i is linked to m then we have a *triangle D*. Three nodes may also be connected without forming a triangle, there can be a single node connected to an unordered pair of other nodes. These are known as "connected triples." The clustering coefficient is then defined as a ratio of the number of triangles to the number of connected triples in the network:

$$C = \frac{3 \times (\mathit{number\ of\ triangles})}{(\mathit{number\ of\ connected\ triples\ of\ nodes})} . \quad (3)$$

The factor three is due to the fact that each triangle is associated with three nodes. This can be expressed in a more quantitative way for a node i which has a degree k_i. The total number of connected triples in the graph can be obtained by summing over all possible combinations that the neighbors can have which is given by $k_i\,(k_i -1)/2$. The clustering coefficient for *undirected graph*s is then defined by

$$C = \frac{3 \times \Delta}{\sum_i k_i (k_i - 1)/2} . \quad (4)$$

This definition corresponds to the concept of *fraction of transitive triples* used in sociology. To obtain a statistical measure for any quantity we have to deal with a large collection of graphs (which are basically similar); these are called *ensembles* of graphs. Equation 3 then needs to be modified to consider the averages of the two quantities yielding the clustering coefficient as:

$$\langle C \rangle = \frac{6 \times \langle \Delta \rangle}{\left\langle \sum_i k_i (k_i - 1) \right\rangle}. \tag{5}$$

An alternative definition of the clustering coefficient has been introduced by Watts and Strogatz (1998) for the analysis of small-world networks (see also discussion in the section on network types). Assume there is a node i with degree k_i and let e_i denote the number of edges existing between the k_i neighbors of i. The clustering coefficient, C_i, of i, is then defined as the ratio between the actual number of edges among its neighbors e_i and the maximum possible value of edges possible between its neighbors which is $k_i (k_i -1)/2$, thereby giving us

$$C_i = \frac{2e_i}{k_i (k_i - 1)}. \tag{6}$$

Thus, this clustering coefficient C_i measures the average probability that two neighbors of the node i are also connected. Note that this local measure of clustering has meaning only for $k_i > 1$.

For $k_i \leq 1$ we define $C_i \equiv 0$ and following work by Watts and Strogatz (1998) the clustering coefficient of a graph $\langle C_{WS} \rangle$ is defined as the average value of C_i over all the nodes in the graph

$$\langle C_{WS} \rangle = \frac{\sum_i C_i}{N}, \tag{7}$$

where N is the total number of nodes.

The two definitions give rise to different values of a clustering coefficient for a graph (see Table 2 in the discussion and exemplification section). Hence, the comparison of clustering coefficients among different graphs must use the very same measure. However, both measures are normalized and bounded to be between 0 and 1. The closer C is to one the larger is the interconnectedeness of the graph under consideration (see also discussion in the section on network types and Figure 7). Because the clustering coefficient considers the neighbors of a node and not its degree alone, it gives us more information about the node. This can be illustrated by a simple example. A scientist (say i) collaborating with a large number of other scientists in only one discipline will have many collaborators who are also collaborating among themselves. However, a scientist (say j) who collaborates with other scientists in many different disciplines will have fewer collaborators collaborating among themselves. Although the important nodes (scientist i and scientist j) in both these networks may have the same degree (number of collaborators), the network of the collaborators of scientist i will have a larger clustering coefficient than the network of collaborators of scientist j.

Similar to the clustering coefficient which analyzes the density of triangles, the study of the density of *cycles* of n connected nodes (for example, rectangles) is another relevant approach to understanding the local and global cohesiveness of a network (Bianconi & Capocci, 2003; Zhou & Mondragon, 2004).

4.2.2 Motifs

Most networks are built up of small patterns, called motifs. Motifs are local patterns of interconnections that occur throughout a network with higher probability than in a completely random network. They are represented as subgraphs and contribute to the hierarchical set-up of networks (Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii, & Alon, 2002; Shen-Orr, Milo, Mangan, & Alon, 2002;

Vazquez, de Menezes, Oltvai, & Barabási, 2004). They have also been identified as relevant building blocks of network architecture and evolution (Wuchty, Oltvai, & Barabási, 2003). There exist diverse approaches to identify cliques and subgraphs in a graph. Bottom-up approaches include cliques, *n*-cliques, *n*-clans, *k*-plexes, or *k*-cores. The bottom-up approach tries to explore how large networks can be built up out of small and tight components. In the simplest case, the complete network is build out of cliques or fully connected subgraphs. However, not all networks can be built using this limited set of building blocks. In the *n*-clique approach, the definition is relaxed to allow nodes to be connected over a longer path length. Here, the *n* stands for the length of the path that interconnects nodes. In some cases, this approach tends to find long and stringy subgraphs rather than tight and discrete ones. The *n*-clans approach tries to overcome this problem, by requiring that connections to new nodes of a subgraph can only made via existing nodes. The *k*-plexes approach was introduced to relax the strong clique definition by stipulating that a node could become a member of a particular subgraph if it had connections to all but *k* members of the subgraph. It is similar to the he *k*-core approach that requires that all members have to be connected to *k* other members of the subgraph.

Apart from bottom-up approaches there exist diverse top-down approaches that help determine components, cut points, blocks, lambda sets and bridges or factions. *Components* were defined in the section on graph connectivity. *Cut points* are nodes that upon their removal lead to a disintegration of a network into unconnected subgraphs. The resulting divisions into which cut points divide a graph are called *blocks*. Instead of the weak points one can also look for certain connections that link two different parts, these are the *lambda sets* and *bridges*. A node that is well connected to nodes in many other groups is called a *hub*.

4.2.3 Modules and Community Detection

In directed networks, the edge directionality introduces the possibility of different types of local structures (see component structure of directed networks in Figure 3). The characterization of local structures and communities is particularly relevant in the study of the World-Wide Web where a large number of studies deal with the definition and measurement of directed subgraphs (Adamic & Adar, 2003; Flake, Lawrence, & Giles, 2000; Gibson, Kleinberg, & Raghavan, 1998; Kleinberg & Lawrence, 2001; Kumar, Raghavan, Rajagopalan, & Tomkins, 1999). One mathematical way to account for these local cohesive groups is to look at the number of *bipartite cliques* present in the graph (Dill, Kumar, McCurley, Rajagopalan, Sivakumar, & Tomkins, 2002; Kumar et al., 1999). A bipartite clique $K_{n,m}$ identifies a group of *n* nodes, all of which have a direct edge to the same *m* nodes. Naively, we can think of the set as a group of "fans" with the same interests and thus their Web pages point to the same set of relevant Web pages of their "idols," see Figure 4a.

Another way to detect communities is to look for subgraphs where nodes are highly interconnected among themselves and poorly connected with nodes outside the subgraph. Figure 4b depicts within community links as full lines and between community links by a dashed line. In this way, different communities can be determined with respect to varying levels of cohesiveness, for example, based on the diameter of the subgraphs representing the communities. In general, the Web graph presents a high number of bipartite cliques and interconnected subgraphs, all identified by an unusually high density of edges.

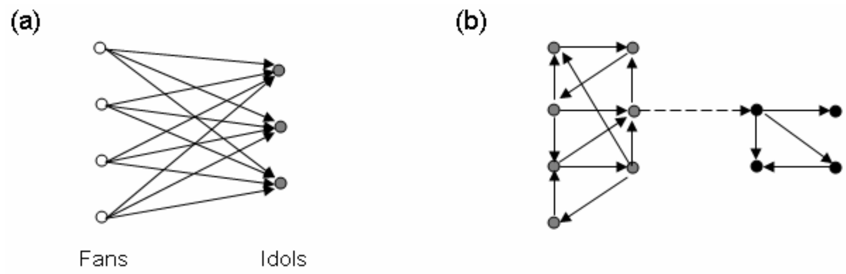


Figure 4. (a) A clique $K_{4,3}$ in which four pages of fans (white nodes) point to the same set of three pages, the idols (in gray). (b) A community of nodes (in gray) weakly connected to other nodes (in black) of the network. The dashed edge denotes the "weak link" with the highest betweenness centrality value. In a community, each node has a higher density of edges within the set than with the rest of the network. Adopted from Kleinberg and Lawrence (2001).

Many networks exhibit a considerable degree of modularity (Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002). That is, the complete network can be partitioned into a collection of modules, each being a discrete entity of several nodes which performs an identifiable task, separable from the tasks of the other modules. Clustering techniques can be employed to determine major clusters. They comprise both nonhierarchical methods (for example, single pass methods or reallocation methods), as well as hierarchical methods (for example, single-link, complete-link, average-link, centroid-link, Ward), and linkage based methods (de Jong, Thierens, & Watson, 2004). Nonhierarchical and hierarchical clustering methods typically work on attribute value information. For example, the similarity of social actors might be judged based on their hobbies and ages. Nonhierarchical clustering typically starts with information on the number of clusters that a dataset is expected to have and sorts the data items into clusters such that an optimality criterion is satisfied.

Hierarchical clustering algorithms create a hierarchy of clusters grouping similar data items. Clustering starts with a set of singleton clusters, each containing a single data item. The number of singleton clusters equals the number of data items N. The two most similar clusters over the entire set are merged to form a new cluster that covers both. Merging of clusters continues until a single, all-inclusive cluster remains. At termination, a uniform, binary hierarchy of $N-1$ partitions results. Frequently, only a subset of all partitions is selected for further processing.

Linkage based approaches exploit the topological information of a network to identify dense subgraphs. They include measures such as betweenness centrality of nodes and edges (Girvan & Newman, 2002; Newman & Girvan, 2004); (see the section on node and edge properties), superparamagnetic clustering (Blatt, Wiseman, & Domany, 1996, 1997; Domany, 1999), hubs and bridging edges (Jungnickel, 1994) (similar to the bridges described previously in motifs), and others. Recently, a series of sophisticated overlapping and nonoverlapping clustering methods has been developed, aiming to uncover the modular structure of real networks (Palla, Derenyi, Farkas, & Vicsek, 2005; Reichardt & Bornholdt, 2004).

4.2.4 Structural Equivalence

The local network structure of a node determines not only the degree of this node but also also, for example, whether my neighbors are also connected, what nodes are reachable, in how many steps. Being part of a large clique is different from being a node on a grid lattice. A short path length to hub nodes is beneficial for spreading information. In many cases, sub-networks of similar structure can be assumed to exhibit similar properties and to support similar functionality. Two nodes are said to be *structurally equivalent* if they have the same relationships to all other nodes in the network. Two nodes are said to be *automorphically equivalent* if they are embedded in local sub-networks that have the same patterns of ties, that is, "parallel" structures. Two nodes are said to be *regularly equivalent* if they have the same kind of ties with members of other sets of nodes that are also regularly equivalent. There exist diverse approaches to determine the *structural equivalence,* the *automorphic equivalence*, or the *regular equivalence* of sub-networks and they use popular measures such as the Pearson correlation coefficient, Euclidean distances, rates of exact matches, or Jaccard coefficient to determine the correlation between nodes (Chung & Lee, 2001).

4.3 Statistical Properties

A statistical analysis is beneficial when one is interested in the characteristics of the entire network rather than the characteristics of single nodes or sub-networks. This is especially relevant in the case of very large networks where local descriptions often do not suffice to answer scientific or practical questions. For example, to study the spreading of computer viruses in the Internet, the complete network has to be analyzed (see section 6 for details on virus spreading models).

Next, we introduce the statistical distributions of the various quantities defined in the previous sections to describe the aggregate properties of the many elements that compose a network.

4.3.1 Node Degree Distribution

The degree distribution *P(k)* of an undirected graph is defined as the probability that any randomly chosen node has degree *k*. Because each edge end contributes to the degree of a node, the average degree $\langle k \rangle$ of an undirected graph is defined as the number of all edges divided by the number of all nodes times two:

$$\langle k \rangle = \sum_k kP(k) \equiv \frac{2E}{N} . \tag{8}$$

A *sparse* graph (defined earlier) has an average degree $\langle k \rangle$ that is much smaller than the size of the graph, that is, $\langle k \rangle << N$.

In the case of directed graphs, one has to consider the in-degree $P(k_{in})$ and out-degree $P(k_{out})$ distributions, defined as the probability that a randomly chosen node has in-degree k_{in} and out-degree k_{out}, respectively. Given that an edge departing from any node must arrive at another node, the average in-degree and out-degrees are equal:

$$\langle k_{in} \rangle = \sum_{k_{in}} k_{in} P(k_{in}) = \langle k_{out} \rangle = \sum_{k_{out}} k_{out} P(k_{out}) \equiv \frac{<k>}{2} . \tag{9}$$

Because we are dealing with statistical probabilities here, there will always be some fluctuations in the degree distribution. Highly heterogeneous networks will have large fluctuations from the average value. Homogeneous networks, for example, a ring lattice, will have low fluctuations. The standard method for measuring the *heterogeneity* of a network is to study the *moments* of the degree distribution. The moment is nothing else than a property of a probability distribution. The *n*-th moment of the degree distribution is formally defined as

$$\langle k^n \rangle = \sum_k k^n P(k) . \tag{10}$$

Note that the second moment of the degree distribution $\langle k^2 \rangle$ governs the *variance* of the distribution. It indicates how close we are to the average value of the distribution. As will be shown in the section on network types, the level of heterogeneity of the degree distribution defines different network types.

4.3.2 Degree Correlation Function

Some networks show degree correlations among neighboring nodes. For example, experts seem to prefer collaborations with other experts. That is, highly connected nodes are interconnected, this is also called *assortative mixing*. In biological and technological networks we find a hierarchical arrangement in which high degree nodes provide connectivity to low degree nodes, also called *disassortative mixing* (Newman, 2002). In mathematical terms, the degree correlation can be measured via the *average nearest neighbor's degree* $k_{nn,i}$ of a node *i*:

$$\bar{k}_{nn}(k) = \frac{1}{N_k} \sum_i k_{nn,i} \delta_{k_i,k} \tag{11}$$

where the sum runs over all nearest neighbor nodes of node *i*. The average degree of the nearest neighbors $k_{nn}(k)$ for all nodes of degree *k* can then be defined as

$$\bar{k}_{nn}(k) = \frac{1}{N_k} \sum_i k_{nn,i} \delta_{k_i,k} \tag{12}$$

where the sum runs over all possible nodes, N_k is the total number of nodes with degree k_i, and $\delta_{k_i,k}$ is the Kronecker delta, which has values $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$. The correlations among the degree of connected nodes can then be expressed as

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k), \tag{13}$$

where $P(k'|k)$ is the conditional probability that an edge of a node with degree *k* is pointing to a node with degree k'. If no degree correlations among neighbor nodes exists then $\bar{k}_{nn}(k)$ is independent of *k*. Uncorrelated random networks provide an example. In the presence of correlations, the behavior of $\bar{k}_{nn}(k)$ identifies two general classes of networks (Newman, 2002): If $\bar{k}_{nn}(k)$ is a function which increases with increasing *k*, nodes with high degree have a larger probability to be connected with large degree nodes indicative for *assortative mixing*. On the contrary, a decreasing behavior of $\bar{k}_{nn}(k)$ defines a *disassortative mixing*, in the sense that high degree nodes have a majority of neighbors with low degree; the opposite holds for low degree nodes.

4.3.3 Node Betweenness Distribution

Similarly, it is possible to characterize betweenness statistically by considering the probability distribution $P_b(b)$ that a node has betweenness *b*. As with Equation 9, it is now possible to obtain different properties of the distribution by defining the moments of the distribution. The *n-th* moment of distribution $\langle b^n \rangle$ is then defined as

$$\langle b^n \rangle = \sum_b b^n P_b(b) \equiv \frac{1}{N} \sum_i b^n{}_i \ . \tag{14}$$

As explained before for Equation 9, the distribution moments quantify the level of heterogeneity of the networks for the betweenness property of the nodes. As before, the significance of the observed average behavior can be quantified by using the second moment of the distribution (see also the section on network types).

4.3.4 Average Clustering Coefficient

The average clustering coefficient can be used to determine whether a type of network is modular or hierarchical on a global statistical level (Pastor-Satorras & Vespignani, 2004; Ravasz et al., 2002). It is determined by computing the clustering coefficient of smaller subgraphs. The subgraphs are selected in a way that ensures they have the same average degree distribution. In mathematical terms, the average clustering coefficient $\langle C(k) \rangle$ of nodes with degree *k* is defined as:

$$\langle C(k) \rangle = \frac{1}{N_k} \sum_i C_i \delta_{k_i,k} \tag{15}$$

where N_k is the total number of nodes with degree *k*, the sum runs over all possible nodes and $\delta_{k_i,k}$ is the Kronecker delta as defined for Equation 11. In many real networks, $\langle C(k) \rangle$ exhibits a highly non-trivial behavior with a power-law decay as a function of *k*, signaling a hierarchy in which most low degree nodes belong to well interconnected communities (high clustering coefficient); hubs connect many nodes that are not directly connected (small clustering coefficient) (Pastor-Satorras & Vespignani, 2004; Ravasz et al., 2002) (see also the section on network types).

4.3.5 Distribution of Node Distances

There are two main statistical characterizations of the distribution of node distances. The first one simply considers the probability distribution $P_\ell(\ell)$ of finding two nodes separated by a distance ℓ. A second indicator, the so called *hop plot* $M(\ell)$, is expressed as the average number of nodes within a distance less than or equal to ℓ from any given node:

$$M(\ell) = N \sum_{\ell'=0}^{\ell} P_\ell(\ell'). \tag{16}$$

At $\ell = 0$ we find the starting node, and thus $M(0) = 1$. At $\ell = 1$ we have the starting node plus its nearest neighbors and therefore $M(1) = k + 1$. If the graph is connected and l_S is the maximum shortest path length, then $M(l_S) = N$ holds. Because the average number of nodes within a certain distance is very different for a regular network, a random network and a small world network, the hop plot is very useful in studying the structure of the network. Note that an increase in the number of nodes within a certain distance increases the hop plot value. Therefore, the hop plot is often referred to as a measure of the *average mass* of a graph. The hop plot can also be related to the spanning tree construction used in mathematics and statistics.

4.4 Network Types

The statistical properties identified in the previous section make possible the detailed analysis of the structural and functional properties of large networks. They form the basis to search for large-scale regularities and asymptotic patterns that can be considered manifestations of the underlying laws that govern the dynamics and evolution of complex networked systems (Albert & Barabási, 2002; Dorogovstev & Mendes, 2002; Newman, 2003; Wolf, Karev, & Koonin, 2002). For instance, many real world networks show the *small-world* property, which implies that the network has an average characteristic path length that increases very slowly with the number of nodes (logarithmically or even slower), despite showing a large degree of local interconnectedness that is typical of more ordered lattices. These two opposite tendencies – the intrinsic randomness with a memory of local ordering – were first reconciled by Watts and Strogatz's (1998) *small-world model*. This model starts with a regular ring lattice of N nodes in which each node is symmetrically connected to its $2m$ nearest neighbors in clockwise and counterclockwise sense. Then, for every node, each edge is rewired with probability p, and preserved with probability $1-p$. The rewiring connects the edge endpoint to a randomly chosen node, avoiding self connections. The parameter p therefore tunes the level of randomness present in the graph, keeping the number of edges constant. Using this model, one can analyze the characteristic path length $\langle l \rangle$ and clustering coefficient $\langle C \rangle$ of a network as a function of the rewiring probability p (see Figure 5). A *regular lattice* with p close to 0 has a high characteristic path length and a high clustering coefficient. A *small world network* with intermediate p has a low characteristic path length and a high clustering coefficient. A *random network* with p close to 1 has a low characteristic path length and a low clustering coefficient. Therefore, there is a broad region of p in which both properties – low characteristic path and high clustering – are found at the same time. In particular, it has been shown (Barrat & Weigt, 2000; Barthélémyy & Amaral, 1999a, 1999b) that this region depends on the size of the network and that in the case of infinite networks any infinitesimal presence of randomness is enough to yield a small-world network.

Note that a small world network can also be generated by adding edges instead of rewiring existing edges (Barabási, Albert, & Jeong, 1999).

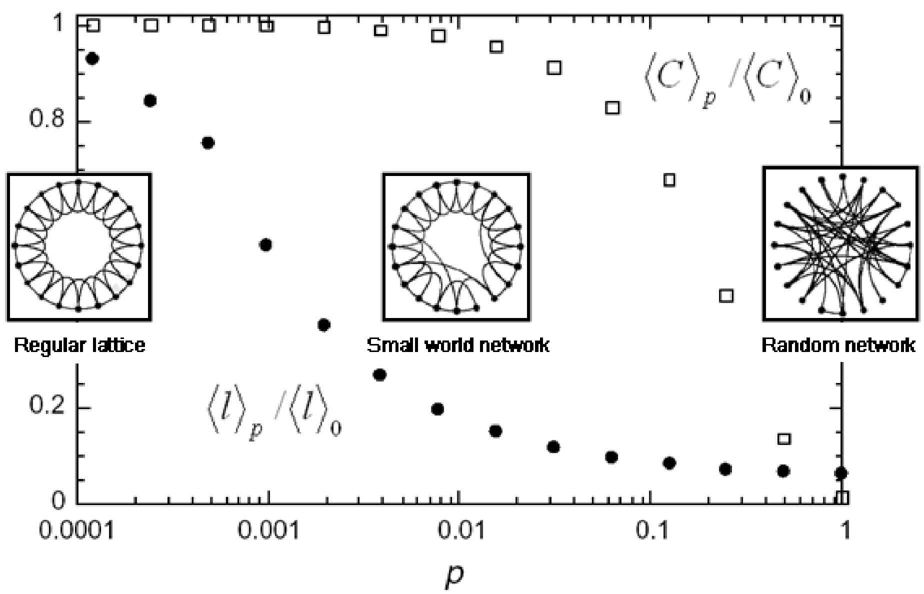


Figure 5. Characteristic path length and clustering coefficient as a function of the rewiring probability p for the Watts-Strogatz model. The characteristic path length and clustering coefficient are normalized by the initial shortest path length $\langle l \rangle_o$ (filled circles) and clustering coefficient $\langle C \rangle_o$ (open squares) for the original regular lattice with p=0. Adopted from Watts and Strogatz (1998).

Another important recent finding is that many networks are characterized by the presence of "hubs", that is, nodes with a large number of links to other nodes. This implies that many networks are extremely heterogeneous, their topology being dominated by a few hubs which link to a majority of the less connected nodes (Albert & Barabási, 2002). This has lead to the definition of two broad classes of networks, depending on the statistical properties of the degree distribution. The first are the so called *homogeneous networks*. These networks exhibit a degree distribution with an exponentially fast decaying tail such as *Poissonian distributions* (see Table 2). The second are *scale-free networks* with heterogeneous connectivity pattern. These networks have a heavy-tailed degree distribution, that is, the probability that a given node has degree *k* is in many cases well approximated by a *power-law distribution* $P(k) \sim k^{-g}$ (see general degree distribution of a scale-free network in Table 2).

An interesting fact about a heavy-tail distribution is that there is a finite probability of finding nodes with a degree value much larger than the average $\langle k \rangle$. This leads to errors if we assume that every node in the network has degree $\langle k \rangle$.

As mentioned before, the heterogeneity of a network is measured by the moments of the degree distribution (Equation 10). In particular, the second moment of the degree distribution obtained by putting $n = 2$ in Equation 10 controls the standard deviation defined by $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$. The standard deviation tells us how diverse the different values of the degree distribution actually are and so is a very important quantity for networks with a heavy-tailed distribution. Fluctuations in systems with a power-law exponent $2 \leq \gamma \leq 3$ are unbounded (they can be infinitely large) and depend only on the system size. The absence of any intrinsic scale for the fluctuations implies that the average value is not a characteristic scale for the system. This also holds for $\gamma \leq 2$. In other words, the average behavior of a *scale-free* system is not typical. That is, when picking up a node at random it will have a low degree most of the time. However, there is an appreciable probability of finding nodes with very large degree values. All intermediate values are possible and knowing the average node degree does not help to describe the degree distribution. This is clearly different from *Poissonian* distributions with fast decaying tails, in which the average *k* value is very close to the maximum of the distribution and represents the most probable value for the degree of a node.

Scale-free networks can be constructed by the use of generalized random graph, *p**-models and many other techniques (Holland & Leinhardt, 1981; Molloy & Reed, 1995; Park & Newman, 2004) (see also the section on modeling static networks). A dynamical approach to modeling scale-free networks was introduced by Barabási and Albert (1999). This novel type of network model can simulate large networks that rapidly evolve by the continuous addition of new nodes. The *Barabási -Albert model* is based on the *preferential attachment* mechanism observed in many real world networks and also known as the *rich get richer* phenomenon, the *Mathew effect* (Merton, 1968), the *Gibrat principle* (Simon, 1955), or *cumulative advantage* (Price, 1976). The model defines a simple class of growing models based on the following two rules:

- Growth: The network starts with a small core graph of m_0 connected nodes. The nodes could be fully connected or any other core graph density except zero could be used. The initial number of edges does not influence the properties of the network in the limit. Every time step we add a new node, and connect it with *m* edges ($m < m_0$) to already existing nodes.
- *Preferential attachment*: The new edges are connected to an existing *s-th* node with a probability proportional to the degree of the node k_s.

These rules define a dynamical algorithm model that can be easily implemented in computer simulations and, starting from a connected initial core graph, generates connected graphs with fixed average degree $\langle k \rangle = 2m$ and a power-law degree distribution. The interest raised by the Barabási-Albert construction resides in its capacity to generate graphs with a power-law degree distribution and small-world properties from very simple dynamical rules. Other features, however, such as the clustering coefficient or the fact that older nodes are always the most connected ones, do not match what we observe in real world networks.

Many extensions of the Barabási-Albert model have been proposed. They extend the original model to account for local geographical factors, rewiring among existing nodes, or age effects (Albert & Barabási, 2002). The importance of the Barabási-Albert model is at the conceptual level. It introduces a simple paradigm that suffices to exemplify the network self-organization which spontaneously generates highly non-trivial topological properties. In addition, the model shifts the focus to microscopic dynamical rules as the starting point of the modeling strategy (see discussion in the sections on network modeling and modeling dynamics of networks).

4.5 Discussion and Exemplification

This section links the terminology and approaches introduced in the four previous sections in a more systematic manner. It also presents an illustrative example to point out the commonalities and differences among regular, random, small world and scale-free networks. Table 1 gives an overview of the different terminology used in different scientific disciplines and their interrelations. Obviously, the different disciplines have developed similar techniques in parallel. It is just now that those commonalities are being discovered. Confusion of terminology is easy and commonplace.

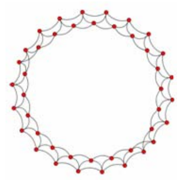
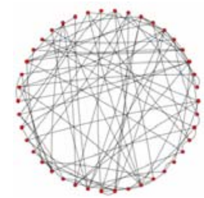
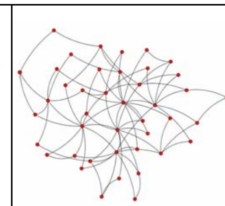
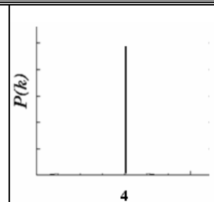
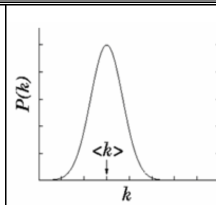
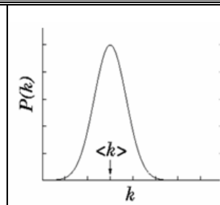
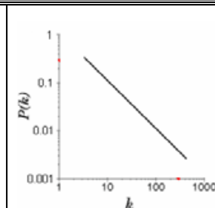
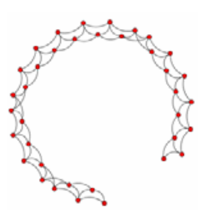
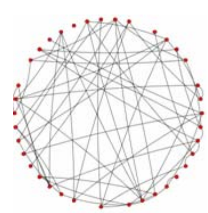
Table 1. Network measurements terminology in mathematics/physics and statistics/social network analysis

Discipline	Mathematics / Physics	Statistics/Social Network Analysis
Terminology Used	Adjacency matrix	Sociomatrix
	Average shortest path length or Diameter	Characteristic path length
	Clustering coefficient	Fraction of transitive triples

Table 2 lists properties of a random, a small world and a scale-free network for means of comparison. The scale free network was generated using the Barabási-Albert model introduced in the section on network types. The model was initialized with a core of two nodes and one edge and 40 time steps. In each time step, one node is added and connected via two edges to nodes already present in the network. There is a preference to connect to highly connected nodes. The lattice, small world and random networks were generated using the Watts-Strogatz, also introduced in the section on network types. The model was initialized with 42 nodes, a node degree of 4, and a rewiring probability of 0.0, 0.5 and 1.0 respectively. Note that all networks have approximately the same number of nodes and edges. All three networks are undirected, fully connected and there are neither parallel edges nor loops.

Table 2. Network measures exemplified for a regular lattice, a small world network, a random network, and a scale-free network.

Network type	Regular lattice	Small world	Random	Scale-free / Heavy-tail

Layout				
# nodes	42	42	42	42
# edges	84	84	84	81
Diameter	11	5	5	5
Characteristic path length	5.6	2.9	2.76	2.6
Clustering Coeff. (Eq. 4.)	1	0.31	0.21	0.26
Clustering Coeff. (Eq. 6.)	0.5	0.16	0.13	0.22
Average degree	4	4	4	3.86
General degree distribution				
After removal of the five most highly connected nodes				

Also presented are the general degree distributions for all four network types. In the regular lattice, all nodes have a degree of four and hence the degree distribution has a *Dirac* delta function at four. For random graphs *P(k)* is Poissonian – it is strongly peaked at $k = <k>$ and decays exponentially for large *k*. The network is rather homogeneous, that is, most nodes have approximately the same number of links. An example of a network with a Poisson distribution is a highway network in which nodes represent cities and links represent streets that interconnect cities. The degree distribution of a scale-free network decays as a power law $P(k) \sim k^{-g}$ at large *k*. The majority of nodes have one or two edges, but a few nodes have a large number of edges and act as hubs. An example of a scale-free network is an airline network where nodes represent cities and links represent airline flights interconnecting them. The last column of Table 2 shows the network after removal of the five most highly interconnected nodes. This is relevant for the discussion of network attacks presented in the section on modeling dynamics of networks.

5. Network Modeling

The section on network types introduced different network types but also two simple network models that generate small world and scale-free networks. Here, we provide a general review of diverse network modeling approaches. A detailed exposition of all modeling approaches is far beyond the scope of the present review. The interested reader is recommended to consult Wasserman and Faust (1994) and Carrington et al. (2004) for a review of social network analysis and Kumar, Raghavan, Rajagopalan, Sivakumar, Tomkins, and Upfall (2000) and Pastor-Satorras and Vespignani (2004) for a computer science or physics driven approach on modeling the Internet. The attempt here is to show how the modeling paradigms developed in different disciplines can be unified in a common conceptual framework. It is our hope that this will encourage researchers to study approaches developed in other disciplines, help interrelate approaches and generally promote the application and comparison of approaches across disciplinary boundaries. Selected concepts will be presented in considerable mathematical detail to give the interested reader a gentle introduction to the strong quantitative foundations that most modeling techniques have. The first sub-section starts with an introduction of models which assume that the network to be simulated is static or in "equilibrium." Examples are networks of co-occurring words in a text or scholarly networks at a given point in time. The next sub-section introduces models that aim to model the dynamic evolution of networks. These models are called dynamical models or non-equilibrium models. The third sub-section discusses modeling frameworks and model validation. Note that models of the dynamics on networks, for example, the spreading of computer viruses on the Internet or the diffusion of knowledge in paper-citation networks, are discussed in the section on modeling dynamics of networks. An overview of the diverse modeling approaches and their applicability is provided in section discussion and model validation (see also Table 4).

5.1 Modeling Static Networks

Mathematicians, statisticians and physicists have made major contributions to models that capture the structure of networks. Interestingly, it is only now, that major commonalities among the rather abstract mathematical theories developed by mathematicians and statisticians and theories describing physical systems developed by physicists are being uncovered. This section reviews known commonalities and uncovers previously unknown commonalities of graph theoretic approaches to network modeling.

Statistical graph models have been used for almost sixty years to obtain a quantitative examination of the stochastic properties of networks: "stochastic" here refers to the fact that a probabilistic or random process is involved. We will also refer to various ensembles, such as the statistical ensemble of graphs, which means a group of all possible graphs having the same number of nodes, edges and probability of connection. Statistical ensembles are a collection of similar objects that differ from one other only because of some probabilistic process that defines the collection.

Next, we review the pioneering work on random graph models, then introduce the class of exponential random graphs and finally show that this class has interesting similarities with the statistical mechanics approach developed by physicists.

5.1.1 Static Random Graph Models and Topology Generators

Approaches such as the paradigmatic *Erdos-Rényi model* (1959) and the *Molloy-Reed* *construction* (1995) are the simplest conceivable ones and have been used as the basic modeling paradigm in several disciplines. They are characterized by an absolute lack of knowledge of the principles that guide the creation of edges between nodes. Lacking any information, the simplest assumption one can make is that is makes sense to connect pairs of nodes at random with a given connection probability p. The first theoretical model of random networks was proposed by Erdos and Rényi in the early 1960s. In its original formulation, the undirected graph $G_{N,E}$ is constructed starting from a set of N different nodes which are joined by E edges whose ends are selected at random among the N nodes. A variation of this model was proposed by Gilbert (1959). Here, the graph $G_{N,p}$ is constructed from a set of N different nodes in which each of the $N(N-1)/2$ possible edges is present with probability p (the connection probability) and absent with probability $1-p$. Both these models generate random graphs whose important properties can then be calculated. For instance, to compute the average degree, the average number of edges $\langle E \rangle$ generated in the construction of the graph is calculated and is found to be $\langle E \rangle = \frac{1}{2} N(N-1)p$.

5.1.2 Exponential Random Graph Family

A more general and well founded group of models, in some respects, the mathematically and conceptually more sophisticated one, is represented by the *exponential random graph family* (Frank & Strauss, 1986; Holland & Leinhardt, 1981; Strauss, 1986; Strauss & Ikeda, 1990; Wasserman & Pattison, 1996). In physicists' terminology with *exponential random graph* one often refers to graphs with a *Poissonian* degree distribution; this is not the case of the exponential random graph family that may account for different degree distributions. In the statistical and social sciences literature these models are also referred to as *Logit models, p*-models,* and *Markov random graphs* depending on the specific assumptions and methods used in the model construction.

This group of models treats the adjacency matrix $x = \{x_{ij}\}$ characterizing a graph of cardinality N as a random matrix whose realization occurs with probability

$$P(x) = \frac{\exp(\sum_i \theta_i z_i(x))}{\kappa(\theta_i)}, \tag{17}$$

where θ_i is a set of model parameters and $z_i(x)$ is a set of network statistical observables, for example, the average degree of the graph <k> or probability distribution of attributes.

In a co-authorship network, the set of model parameters θ_I might represent the likelihood of two authors collaborating due to their geographical proximity. The term $z_i(x)$ might represent the average number of collaborators in a field. Once the relevant statistics and assumptions are included in the model, the parameter θ_i has to be estimated by comparison with the real data. This has spurred the development of a wide array of parameter estimation techniques such as pseudo-likelihood estimation and *Monte Carlo* maximum likelihood estimation

(Frank & Strauss, 1986; Strauss, 1986; Strauss & Ikeda, 1990; Wasserman & Pattison, 1996). Monte Carlo methods are used to obtain an approximate solution to a variety of mathematical problems which deal with random processes. The function $\kappa(\theta_i)$ ensures the correct normalization once we sum the probability distribution over all possible graphs x allowed in the *sample space*, also called *phase space* in physics and engineering. This space can be conceptually schematized as hyper-dimensional space in which each coordinate represents the possible values that each degree of freedom (in the case of a graph the variables x_{ij}) may have Figure 6).

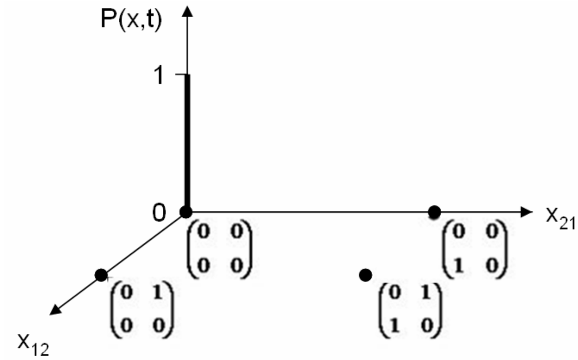


Figure 6. Sample space or phase space of the possible realization of a directed network with two nodes. Each x_{ij} axis represents the existence or not of the corresponding edge. Each point in the phase space thus represents a possible microscopic realization of the network, corresponding to the relative adjacency matrix. Using a third axis, it is possible to report the corresponding occurrence probability P(x,t) associated with each configuration. The dimensionality of the sample space increases with the number of nodes. In principle it is possible to associate each network of size N to an N^2 dimensional hyper-space. Exemplarily depicted is the phase space realization for an unconnected network.

Note that Equation 18 defines the probability for the occurrence of an edge for any pair of nodes described by the adjacency matrix *x*. Therefore, the individual element x_{ij} do not explicitly occur in the equation. Instead, the matrix *x* as a whole is considered.

For co-authorship networks, adding a new author node increases the degrees of freedom for the existing authors as they have the possibility of collaborating with one more person. So the degrees of freedom are related to the particular network that is being studied. For the co-authorship network, this means that if an author *A* collaborates with both *B* and *C*, the point representing this state would be different from the point in the phase space where *A* collaborates with *B* and not with *C*. The dimension of the phase space is the sum of the number of authors and *P(x)*. The complete phase space consists of points representing all possible co-authorships of all authors.

5.1.3 Statistical Mechanics of Networks

These models assume that the probability for a system, for example, a network, to be in a specific microscopic configuration *x* is given by the distribution *P*(*x*) that maximizes the *Gibbs* entropy (Parisi, 1988)

$$S[P] = -\sum_{x} P(x) \ln P(x) \, , \tag{18}$$

where the sum is over all possible stochastic realizations allowed. The Gibbs entropy *S[P]* is a measure of the disorder encoded in the probability distribution. This is similar to the *information entropy* introduced by Shannon (1948). Shannon's measure describes how much information can be processed when it is transmitted through a discrete *Markovian* process. The Shannon-Weaver model of communication (Shannon & Weaver, 1949) has led to several theoretical models in information science (Lynch, 1977). Information scientists have used these concepts in developing models for analyzing the hyperbolic distribution of letters in English text (Lynch, 1977) or the scholarly referencing process (Rettig, 1978). The general similarities and differences between the information theoretic approach and the statistical physics approach have been discussed in detail elsewhere (Jaynes, 1957). Just as the best choice for the probability distribution in information theory is the one that maximizes the information entropy (Shannon's maximum entropy assumption), in physics one expects that in the equilibrium state[2] the statistical disorder reaches its maximum. In the context of physical systems, this assumption can be more formally stated and related to the microscopic dynamics of the physical system. Again, similar to information theory, the maximization of the entropy is constrained by the statistical observables $z_i(A)$, for which one assumes there are statistical estimates:

$$\langle z_i \rangle = \sum_x P(x) z_i(x) \tag{19}$$

and the normalization condition $\sum_x P(x) = 1$. Whenever one has to obtain the maximum of a function based on several constraints, the standard mathematical technique used is that of the *Lagrange multipliers*. Each constraint $\langle z_i \rangle$, is associated with an as yet undetermined constant α which is known as the *Lagrange* multiplier, with α_0 being the multiplier relative to the normalization condition. Note that this is similar to the model parameters in the *exponential family of random graphs* approach which had to be determined from the actual data. The derivative of the distribution must be zero at the maximum, subject to the constraint conditions. The distribution must therefore satisfy the equation

$$\frac{\partial}{\partial P(x)}\left[S[P] + \alpha_0\left(1 - \sum_x P(x)\right) + \sum_x \alpha_i \left(\langle z_i \rangle - \sum_x P(x) z_i(x) \right) \right] = 0 \tag{20}$$

for all possible realizations x. This simple derivative yields the equation:

$$\ln P(x) + 1 + \alpha_0 + \sum_i \alpha_i z_i(x) = 0 \tag{21}$$

that gives the solution

$$P(x) = \frac{\exp(-\sum_i \alpha_i z_i(x))}{Z(\alpha_i)}, \tag{22}$$

where the normalization condition imposes

$$Z(\alpha_i) = e^{\alpha_0 + 1} = \sum_x e^{(-\sum_i \alpha_i z_i(x))}. \tag{23}$$

Finally, the explicit values of the parameters α_i are found by imposing the self-consistent condition on the statistical observables

$$\langle z_i \rangle = \sum_x z_i(x) \frac{\exp(-\sum_i \alpha_i z_i(x))}{Z(\alpha_i)} \tag{24}$$

for all the observables z_i used in the model construction.

5.1.4 Comparison

From the previous discussion it can be easily seen that the exponential family of distributions is equivalent to the statistical mechanics of Gibbs for networks used by physicists (Burda, Jurkiewicz, & Krzywicki, 2003; Burda & Krzywicki, 2003; Dorogovtsev, Mendes, & Samukhin, 2003; Farkas, Derenyi, Palla & Vicsek, 2004; Fronczak, Fronczak, & Holyst, 2005; Krzywicki, 2001; Park & Newman, 2004). Indeed, Equation 17 can be obtained from Equation 22 by a simple substitution of $\theta_i = -\alpha_i$ and $\kappa(\theta_i) = Z(\theta_i)$ yielding an identical probability distribution $P(x)$. In physics, the statistical weight of each system configuration $H(x) = \sum_i \alpha_i z_i(x)$ is named *Hamiltonian* and the function Z is called the *partition* *function*. The Hamiltonian and the partition function are used to completely describe all properties of a system under study.

For example, in a co-authorship network with N author nodes, the partition function gives the sum over all the possible graphs in the network. The Hamiltonian describes the constraints on the co-author relations in the network. Together the partition function and the Hamiltonian of a network tell us the structure of the network (Berg & Lassig, 2002). Analogies can be pushed further and it is possible to show that different statistical constraints correspond to different statistical ensembles in the statistical mechanics definition. Moreover, it is possible to show that this formalism also describes random graphs such as the Erdös and Rényi one. For instance, the random graph family $G_{N,p}$ can be recovered by imposing as a constraint the corresponding value $< E >$ (see Park and Newman [2004) for details]. The *exponential random graph* family used in statistics thus corresponds to the equilibrium ensembles of statistical mechanics developed in physics. Table 3 presents a comparison of the terminology used in the different disciplines.

Table 3: Modeling terminology in mathematics / statistics and physics

Discipline	Mathematics / Statistics	Physics (Statistical Mechanics)
Terminology Used	All graphs in the same *exponential random graph* family	Equilibrium ensembles
	Sample space	Phase space
	Probability of occurrence of a graph in the exponential random graph family $P(x) = \frac{\exp(\sum_i \theta_i z_i(x))}{\kappa(\theta_i)}$	Probability of a system being in an equilibrium state based on the maximum entropy principle $P(x) = \frac{\exp(-\sum_i \alpha_i z_i(x))}{Z(\alpha_i)}$
	Set of statistical observables z_i which define the graph structure.	Set of statistical observables z_i which constrain the physical system.
	Set of model parameters θ_i that are necessary to generate the graph (obtained from data).	Set of parameters α_i corresponding to the constraints of the system known as *Lagrange* multipliers.
	Normalization factor $\kappa(\theta_i)$ that equals the sum of all possible graphs in an ensemble or group.	Partition function $Z(\alpha_i)$ of the system that normalizes the probability distribution in the phase space.

5.2 Modeling Evolving Networks

The modeling approaches we have just introduced in the previous section are focused on the stationary properties of the network for which they derive the probability distribution in the phase space. However, many networks are not static, but evolve over time. The creation of a social relation, the introduction of a hyperlink to a Web page, or the peering of two Internet service providers are dynamical events that shape the evolution of a network based on local interactions among individual nodes.

The *exponential random graph family* framework has been adapted to introduce network evolution by the addition (or deletion) of edges within a fixed number of nodes (Banks & Carley, 1996; Sanil, Banks, & Carley, 1995). Based on this, numerical techniques have also been developed for estimating the model distribution parameters (Snijders, 2002). The latter have been implemented in Snijders's *Simulation Investigation for Empirical Network Analysis* (SIENA) package (http://stat.gamma.rug.nl/snijders/siena.html).

The dynamical evolution of networks can be generally modeled by formally introducing a time variable *t* that indicates the changes of the network quantities in time. In this general approach the number of nodes can vary with time and we do not need to have the constraint of a fixed number of nodes. However, a note of caution is in order here. Because the dynamical approach is extremely dependant on the specific network dynamics, it is potentially risky and, unless we have precise experimental information on the system dynamics, it does not result in quantitatively accurate predictions. Moreover, it does not provide a systematic theoretical framework, each model focusing only on some specific features of the system of interest. On the other hand, the study of the dynamics is well suited to identifying general growth mechanisms out of seemingly very different dynamical rules.

5.2.1 Master Equation Approach

We now proceed to explain the dynamical modeling perspective, in which the probability of a particular network realization *x* at time *t* is given by the distribution $P(x,t)$. The *master equation* approach, developed in physics, can be employed to express the temporal evolution of the probability distribution. This approach assumes that a network has a particular realization in each time step and that the change in the realization over time is controlled by the microscopic dynamics of the model. The microscopic dynamics are expressed as a *rate* $r_{(x\to y)}$ that gives the transition from a particular realization at a certain time *t* to a realization *y* at a later time $t+\partial t$. The master equation then expresses the temporal change in the distribution as a linear differential equation of the form

$$\partial_t P(x,t) = \sum_{y\neq x}\left[P(y,t)r_{(y\to x)} - P(x,t)r_{(x\to y)}\right]. \tag{25}$$

The transition rates have to satisfy the relation $\sum_Y r_{(x\to y)} = 1$ which means that the sum of the rate of all possible configurations *Y* must be unitary. This condition is necessary as the transition between any two realizations is probabilistic and the total probability of any processes must be unitary. The probability distribution $P(x,t)$ is normalized and because the sum of the rates is unitary normalization is preserved on both sides of the equation.

As an example, we model the degree distribution of a paper node in an evolving paper citation network. Over time the paper receives more citations and its degree distribution changes. Here the term $P(x,t)$ in Equation 25 corresponds to $P(k,t)$ where *k* is the degree of the paper node at time *t*. If a new paper cites this node, then the distribution $P(k,t)$ changes to $P(k+1,t)$. These are the *x* and *y* realizations described in Equation 25. Whether a paper is going to be cited or not might depend on many factors such as its age or the number of citations it has already accumulated. All these factors make up the microscopic dynamics that is used to calculate the rate $r_{(x\to y)}$. For simplicity, we assume that there is only one factor that determines the change and that is the degree distribution *k*. So, $r_{(x \to y)}$ is a function of *k* only and we denote this by *p(k)*. *p(k)* can simply be *k,* but in real networks there are often other factors involved and hence it is usually a far more complex function of *k*. The other rate $r_{(x \to y)}$ will then be given by *1-p(k)* to maintain the unitary condition. Thus all the terms in Equation 25 are defined and we can solve the equation to obtain the probability distribution $P(x,t)$ for any state of the system.

5.2.2 Master Equation Approach for Equilibrium Networks

Some systems reach a stationary state over time in which the probability distribution of finding the system in any given configuration does not depend on time anymore, i.e., $P_s(x) = \lim_{t \to \infty} P(x,t)$. Equilibrium systems are those systems in which the stationary probability $P_s(x)$ obey the maximum entropy principle, yielding $P_s(x) = P(x)$ where $P(x)$ is given by Equation 17 or equivalently Equation 22. In this case the solution to the master equation is simply obtained by imposing the *detailed balance condition* ensuring that the probability of leaving a state and arriving in it from another state is the same individually for every pair of states that the system can have. This simply reads as

$$P(y) r_{(y \to x)} = P(x) r_{(x \to y)} \forall y, x, \qquad (26)$$

where the right most symbol means that the relationship holds for all values of *y* and *x* (for a description of the equation elements see Equation 25). The *detailed balance condition* allows one the assignment of rates according to the relation

$$\frac{r_{(y \to x)}}{r_{(x \to y)}} = \frac{P(x)}{P(y)} = \exp\left(\sum_i \theta_i \left[z_i(x) - z_i(y) \right] \right), \qquad (27)$$

which defines the dynamic and ensures the convergence to the correct equilibrium probability distribution in the stationary state.

From a computational point of view these kinds of techniques are at the core of *Monte Carlo* simulations both in the statistics and the physics literature. They circumvent the explicit calculation of the partition function *k* (see Equation 23) that is often hard or impossible. Because Equation 27 describes ratios between transition rates, it suffices to have one of those rates as the reference time scale to obtain all the others in terms of the equilibrium distribution $P(x)$. The rates may be used to produce simulations of the evolution of the network and find the values of the parameters θ_i that better fit the real data when the analytical calculations are too complicate.

5.2.3 Master Equation Approach for Non-Equilibrium Networks

Unfortunately, not all systems have a stationary solution that is given by equilibrium $P(x)$. For instance systems may achieve a stationary state without satisfying the *detailed balance condition*. Remember that the stationary distribution means that there is no change in the distribution with time whereas the detailed balance condition means that there is no change in the distribution for each pair of states.

These cases define *non-equilibrium systems* that still have a stationary state but whose dynamics does not allow a detailed balance and a maximum entropy calculation. In addition, there exist networks with a continuously increasing number of nodes and edges. For those, the dimensionality of the phase space increases continuously, rendering equilibrium calculation unfeasible.

For *non-equilibrium systems*, it is more convenient to rely on approaches dealing directly with the master equation that does not need an exact solution and for which many approximate and numerical techniques exist.

In the master equation approach, it is crucial to consider transition rates or probabilities reflecting the actual dynamics of a system under consideration. In order to model a system, the dynamical laws governing the evolution of the system have to be known. If the dynamical laws are unknown then the dynamical approach is often a difficult exercise in which rough assumptions and uncontrolled approximations have to be made. Yet, it has the advantage of being more intuitive and suitable to large-scale computer simulations. An example is Krapivsky and Redner's (2005) application of the master equation approach to model paper citation networks. In general, the master equation cannot be exactly solved and it is more practical to work with a specific projection of the probability distribution, such as the degree distribution or any other statistical observables in the network.

A continuously growing citation network in which new nodes appear and wiring processes take place (Krapivsky & Redner, 2003) can also be modeled in a way similar to that described before (near Equation 24). For the sake of simplicity we also assume that once an edge is established it will not rewire (this does hold true for a citation network as citation links once established do not change). Further assume, that we are interested in the node degree distribution (which changes over time due to increase in citations) specified by the number N_k of nodes with degree *k*. The master equation for such a growing scheme is given by:

$$\partial_t N_k = r_{(k-1\to k)} N_{k-1} - r_{(k\to k+1)} N_k + \delta_{k,m}. \qquad (28)$$

Here, the first term on the right corresponds to processes in which a node with *k*-1 links is connected to a new node, thus yielding a gain to the number N_k. The second term corresponds to nodes with degree *k* that acquire a new edge, thus representing a loss for N_k. The last term, the Kronecker delta (defined in Equation 11), corresponds to the entering of the new node with degree *m*. The eventual solution of this equation depends on the rates $r_{(k-1\to k)}$ and $r_{(k\to k+1)}$ that specify the network dynamics.

This approach was used by Redner (2005) to model a *Physical Review* citation network. He was able to show that the growth in the average number of references per paper (the out-degree distribution) obtained from the model was consistent with the actual data. However, Krapivsky and Redner (2005) showed the growth predicted by their model was not robust for different model parameters. It will be worthwhile to test their model against other datasets.

Another possibility is to study the average degree values $k_s(t)$ of the $s-th$ node at time *t* as proposed by Albert and Barabási (2002), Dorogovstev and Mendes (2003), Newman (2003), and Pastor-Satorras and Vespignani (2004). For the sake of analytical simplicity the degree *k* and the time *t* are assumed to be continuous variables. Here, the properties of the system can be obtained by studying the differential equation governing the evolution of $k_s(t)$ over time. This equation can be formally obtained by assuming that the degree growth rate of the $s-th$ node increases proportionally to the attachment probability $\Pi[k_s(t)]$ that an edge is attached to it. In the simplest case, edges are only coming from new nodes. Here the rate equation reads:

$$\frac{\partial k_s(t)}{\partial t} = m\Pi[k_s(t)], \qquad (29)$$

where the proportionality factor *m* indicates the number of edges emanating from every new node. This equation is constrained by the boundary condition $k_s(s) = m$, meaning that at the time of their introduction, all nodes have degree *m*. In this formulation all the dynamic information is contained in the probability $\Pi[k_s(t)]$. The properties of each model are defined by the explicit form of the probability $\Pi[k_s(t)]$. Both formulations allow the calculation of the degree distribution, degree correlation and clustering functions. The projection could be considering other quantities such as the number of nodes $N(k|\ell)$ with degree *k* which share an edge with a node of degree ℓ and so forth, but this implies an increasing complication for the dynamical equations. Dynamics might be complicated by other dynamical processes, too, such as edge removal, rewiring and inheritance, as well as node disappearance. For a review of dynamical models see Watts (1999), Barabási (2002), Buchanan (2002), Dorogovtsev and Mendes (2003), Pastor-Satorras and Vespignani (2004) and references therein.

As examples of the above approach let us first consider the case of two models that contains the *preferential attachment* mechanism introduced in the section on network types: the *Barabási-Alber*t model and the *copy model*. The *Barabási-Alber*t model assumes that a new node is linked to an already existing node *s* with a probability proportional to its degree k_s. This immediately produces a probability of attraction

$$\Pi[k_s(t)] = \left[\frac{k_{in,s}(t)}{\sum_j k_j(t)}\right] \tag{30}$$

where the sum of all the degrees of all nodes in the network is the required normalization factor in the denominator. This class of models has been studied in detail as a candidate for a general mechanism to generate *power-law* degree distributions in growing networks. A second, completely different mechanism—the *copy model*—has been proposed in the context of WWW simulations as a mechanism for the generation of skewed degree distributions (Kumar et al., 2000). The copy model was inspired by the fact that creators of new Web pages tend to copy hyperlinks from already existing Web pages with similar content. This mechanism was translated into a growing model in which at each time step a new node (Web page) is added to the network and a *prototype node* is selected at random among the already existing nodes. The outgoing edge of the new node is then distributed based on a *copy factor* a which is constant for all new nodes. A new edge is rewired with probability α to a randomly chosen node of the network. With probability $1-\alpha$ it is attached to a node already having a common edge with the prototype node. At first sight the model seems completely unrelated to the preferential attachment mechanism, but a closer look reveals interesting similarities. This second process of attaching actually increases the probability of high degree nodes to receive new incoming edges.

As an example, let us focus on a generic network node and calculate its probability of receiving an edge during the addition of a new node. Given that a new node will add *m* new edges, a random node in the network is chosen with probability α. Thus any node has a probability α / N to receive an edge, where *N* is the size of the network. With probability $1-\alpha$, the node which is pointed to by one of the edges of the prototype node is selected. The probability that an existing node is linked to the new node equals the number of incoming edges of that node divided by the sum of all node degrees in the network, that is, $k_s(t)/\sum_j k_j(t)$. By combining the two terms we determine that the probability of receiving an edge is

$$\Pi[k_s(t)] = m\left[\frac{\alpha}{N(t)} + (1-\alpha)\frac{k_{in,s}(t)}{\sum_j k_j(t)}\right]. \tag{31}$$

When comparing Equations 30 and 31, we see that the second term on the right hand side of Equation 31 is very similar to the right hand side of Equation 30, which indicates that both have a preferential attachment component. Hence, both generate networks that exhibit *power-law* degree distributions using very different mechanisms. It is one of the strengths of the dynamical approach that systems with shared properties at the macroscopic level often also exhibit shared elements in their description at the microscopic level.

5.2.4 Agent Based Modeling

If a stationary solution satisfying all the constraints of the system cannot be found analytically, numerical simulations and agent based modeling (ABM) approaches are the only viable alternative. Numerical simulations are widely used in physics and biology as a probe to study the behavior of very complicate models not amenable of analytical solutions. Stochastic simulations have indeed widely used in the recent activity on complex networks (Albert and Barabási, 2002; Dorogovstev and Mendes 2003; Newman 2003; Pastor-Satorras and Vespignani 2004). The availability of large-scale, dynamic network datasets (the Internet, the WWW, etc.) and the need to understand, manage, secure these networks has fueled a major increase of realistic ABM research and lead to a change of the modeling perspective. The focus is on single individuals or elements of the system, including the most possibly complete description of the reality. Here ABM can be applied to model local interactions in large-scale computer simulations which ideally generate a network that shows global properties observable in real world systems. Such models have been quite successful in simulating the co-evolution of paper-citation and co-author networks (Börner, Maru, & Goldstone, 2003), analyzing trade and commerce networks spanning different locations (McFadzean & Tesfatsion, 1997) and other social and ecological processes (Gimblett, 2002).

5.3 Discussion

Table 4 provides an overview of the models that are discussed in this section and the subsequent section on modeling dynamics on networks.

Table 4: Network properties modeled and applicable models developed in mathematics / statistics / social network analysis and physics

Section	Network Properties Modeled	Mathematics / Statistics / Social Network Analysis	Physics (Statistical Mechanics)
Modeling static networks	Structural properties	**Static random graph models and topology generators** **Exponential random graph family** (e.g., Logit models, p^*-models, and Markov random graphs) all these graphs have a Poissonian degree distribution.	**Statistical mechanics models using Gibbs entropy** maximization via Lagrange multipliers

Modeling evolving networks	Evolution & structure (equilibrium)	**Exponential random graph family** for fixed number of nodes, edges are changed over time.	**Master Equation**
	Evolution & structure (non-equilibrium)		**Master Equation** **ABM**
Modeling dynamics on networks	Dynamical processes over network		**Master Equation** **ABM**

The *static random graph models and topology generators* as well as the *exponential random models* (introduced in the section on modeling static networks) and the very similar *statistical mechanics models* build on solid statistical foundations and have been mathematically and conceptually developed for many years. However, they are less intuitive and in many practical instances they present us with intractable technical problems. For example, they cannot be applied to model networks whose size is rapidly changing or to non-equilibrium networks. In these cases, the dynamical approach, even if based on a large number of assumptions, is the only viable option. This is especially true if we are interested in studying very large-scale networks for which global equations cannot be specified but the local interactions – the microscopic dynamics – of nodes is known.

In many ways, the recent explosion in dynamical modeling approaches is a consequence of the informatics revolution. The advent of high throughput biological experiments, the possibility of gathering and handling massive datasets on large info-structures and tracking the relations and behavior of million of individuals has challenged the community to characterize and model networks of unprecedented sizes. Today, the Internet comprises more than 10^4 service providers with 10^5 routers keeping track of the behavior of 10^6 users, and its size is continuously increasing. WWW crawls offer maps of the Web with more than 10^7 nodes. In addition, networks of similar size and dynamical characteristics are gathered every day for communication infrastructures such as mobile telephone and ad hoc networks, transportation networks and digital documents. In biomedical research, we are witnessing a paradigm shift with an increasing focus on the so-called system's biology and the many large interaction networks that may be measured by taking advantage of high throughput experiments. In almost all cases, the dynamical features of these systems cannot be neglected because we are typically dealing with networks growing exponentially in time because of their intrinsic dynamics. In this context, dynamical modeling offers an intuitive way to understand the evolution and non-equilibrium properties of these networks and enables the construction of basic generators that capture the rapidly evolving dynamics of such systems.

The availability of large-scale datasets and the ability to run large-scale simulations pose new conceptual questions. One posed by physicists addresses the "universality" of network properties, for example, *small world* or *scale-free* degree distributions (see also discussion in the section on network types). In recent years, networks from diverse domains and serving very different functions have been analyzed. Many of these appear to share similar properties when the total number of nodes in the network is very large. This raised the issue of the supposed emergence of general and common self-organizing principles that go beyond the particulars of individual systems.

5.4 Model Validation

All models make assumptions that reflect our understanding of the world and our theoretical and/or practical interests. *Statistical modeling* based on maximum entropy considerations is most suitable for taking account of the statistical observables at hand; *dynamic modeling*, however, is well suited for large-scale and evolutionary properties. Both types of models need to be validated against empirical data: they do this in very different ways.

Statistical models such as the exponential random graph modeling or equilibrium statistical physics use empirical data to obtain the parameters (q_i in Equation 17 and a_i in Equation 19) necessary for generating the probability distribution of a network. The obtained distributions then provide statistical predictions that can be validated through new measurements on the dataset. Dynamical models define the local dynamics of a network, for example, what papers a new paper should cite, based on information that is in general not related to the statistical observables of the complete network. Here the distribution of the network is assumed to emerge based on the local dynamics and the properties of the generated network are compared with the properties of the empirical network. Hence, the network properties are not an input to the model but are instead used in the model validation process.

Generally, it is impossible to model all properties of a network in realistic detail. Suitable approximations have to be made. Depending on which questions the model wants to tackle, different properties will be considered prominent or neglected. In this sense, all models are incomplete and most address only a limited set of questions. As we will show in the next section, these considerations also apply to models that aim to reproduce the dynamics occurring on networks.

6. Modeling Dynamics on Networks

Networks provide the substrate on top of which the dynamical behavior of a system unfolds. At the same time, the various dynamical processes affect the evolution of a network's structure. Network structure, its evolution over time and the network usage are mutually correlated and need to be studied in concert.

To give an example, epidemiologists, computer scientists and social scientists use very similar models to study spreading phenomena such as the diffusion of viruses, knowledge or innovations. Detailed knowledge of the contact networks defining the interactions of the nodes at various scales, that is, path lengths over which neighboring nodes interact, is required to model these systems. Similarly, in technological networks, for example, the power grid, the Internet, or transportation systems, it is crucial to understand the dynamics of information or traffic flow taking place along the nodes. The resilience of a network depends on basic dynamical processes, as the failure of one network component increases the burden on other elements, potentially overloading them, and disrupting their functions as well. To model dynamics on networks, the theoretical framework presented in the section on network modeling needs to be extended so that the impact of the various network characteristics on the basic features of the dynamical processes can be investigated in a systematic way.

As has been mentioned, diffusion modeling is very similar across different applications, such as the spreading of viruses, diseases, rumors, knowledge, or fashion. For instance, epidemiological models

(Pastor-Satorras & Vespignani, 2001) are based on categorizing people as "susceptible," "infected" and "removed." For rumor spreading models (Daley, Gani & Cannings, 1999), people may be categorized as "ignorant," "spreaders," and "stiflers." For knowledge diffusion purposes they are known as "innovators," "incubators" and "adopters." If we are interested in the spreading of computer viruses, then epidemiological models can be readily applied even through the virus host is now a computer instead of a living being (Bettencourt, Cintron-Arias, Kaiser, & Castillo-Chavez, 2005; Tabah, 1999).

In each of these models, transitions from one state to another are probabilistically obtained from contacts among individuals in the different categories. Subsequently, we explain the diffusion process using epidemiological models as these are well studied and easy to understand. It is our hope that the interested reader will keep in mind the relation of "susceptible" and "infected" to the rumor spreading and knowledge diffusion analogies previously discussed.

6.1.1 Master Equation Approach to Dynamical Processes on Networks

The master equation approach introduced in the section on modeling evolving networks can also be applied to study epidemiological models and other dynamical processes. Let us consider the susceptible-infected-susceptible (SIS) model. The SIS model is a simple model used to the study of infectious diseases leading to an endemic state with a stationary and constant value for the prevalence of infected individuals, that is, the degree to which the infection is widespread in the population. In the SIS model, each node in a network is characterized by a specific state or a vector of states s_i.

In the simplest case, each node can only exist in two discrete states, namely, susceptible ($s_i = S$) and infected ($s_i = I$). The respective *phase space* (see also Figure 6) for a network with two nodes is shown in Figure 7.

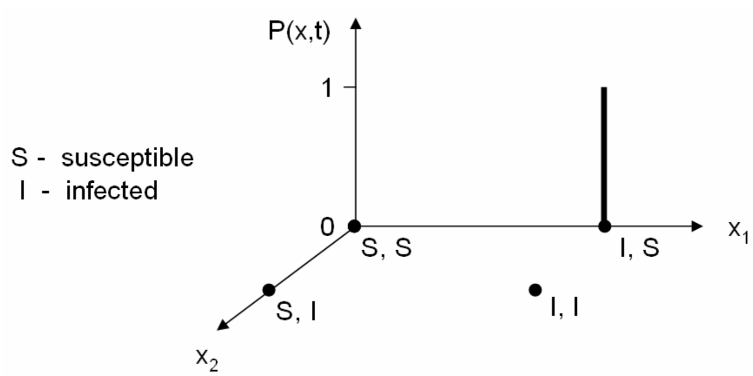


Figure 7. Phase space of realizations for two nodes that can be in susceptible or infected state. Each axis represent the existence or not of a node property such as S and I. Exemplarily shown is a configuration in which node x_1 is infected and x_2 is susceptible, that is, the probability of I, S is one and all other probabilities are 0.

In the following, we exemplify how the SIS model can be implemented for a network with N nodes using the master equation approach. For the sake of simplicity let associate to the infected state the variable 1 and to the susceptible state the variable 0. Let $\{\sigma^a\} = (\sigma_1 = 1, \sigma_2 = 0, ..., \sigma_N = 1)$ denote a particular configuration a specifying the state of each node i and let $P(\{\sigma^a\}, t)$ represents the probability that the system is in state a. In each time step, the state of any node can change. Hence in each time step, the system might be in a different configuration, for example, there might be a time step in which a majority of nodes is susceptible and other time step in which the majority of nodes is infected. Let $w_{(\{\sigma^a\} \to \{\sigma^b\})}$

denote the transition probability to go from state *a* to state *b*. The overall dynamics of the network can then be written down in the form of a *master equation*:

$$\partial_t P(\{\sigma^a\},t) = \sum_{\sigma^b \neq \sigma^a} \left\lfloor P(\{\sigma^b\},t) w_{(\{\sigma^b\}\to\{\sigma^a\})} - P(\{\sigma^a\},t) w_{(\{\sigma^{Sa}\}\to\{\sigma^b\})} \right\rfloor. \quad (32)$$

The transition probabilities $w_{(\{\sigma^a\}\to\{\sigma^b\})}$ are a function of the probability that susceptible nodes get infected by their neighbors (0→1) and that an infected individual is cured (1→0).

The probability that a susceptible node acquires the infection from any given neighbor in an infinitesimal time interval dt is λdt, where λ defines the virus *spreading rate* (see Figure 9). At the same time, infected nodes are cured and become susceptible again with probability μdt. Individuals thus run stochastically through the cycle susceptible $\to$ infected $\to$ susceptible, hence the name of the model. The SIS model does not take into account the possibility of individuals removal due to death or acquired immunization, which would lead to the so called susceptible-infected-removed (SIR) model. Recent epidemic modeling (Pastor-Satorras & Vespignani, 2004) simulating the spreading of computer viruses shows that the SIR model is particularly suited for the initial stages of a computer virus attack. This is because infected computers are switched off as soon as the virus is detected and return to the network only when they have been screened by an antivirus. However, researchers have found that in the long run the clean-up stage reaches a stationary steady state in which the SIS model better represents the overall endemic state. The SIR model has also been used to model knowledge diffusion through blogspaces (Gruhl, Guha, Tomkins, & Liben-Nowell, 2004). In the blogspace, a “virus” resembles a “topic” that might be an URL, phrase, name, or any other representation of a meme that can be tracked from page to page. A blogspace is assumed to be “susceptible” to a topic and may get “infected” by it. Subsequently, it may become immune or be “removed” from the topic. The authors make the assumption that all occurrences of a topic except the first are the result of communication via edges in the network, that is, the topic is not discussed offline, spread via news, or by other means. Because blogspaces can be overwritten by the authors at a later time, the SIR model was extended to the Susceptible – Infected – Removed (but temporarily)—Susceptible again (SIRS) model to include this property. All these models, the SIR and SIRS are extensions of the SIS model.

In the SIS model, the virus spreads by infecting its neighboring nodes. Hence, the connectivity of the nodes in the network influences the transition probability *w* of each node. The transition probabilities for a random network differ from those for a small world network. Referring back to Equation 32, the global evolution of $P(\{\sigma^a\},t)$ (given by the left hand side of the equation) depends on the transition probability *w* of each node (present in the terms on the right hand side of the equation). The master equation considers the network connectivity pattern by means of the transition probabilities.

As discussed in the section on modeling evolving networks, the complete solution of the master equation is rarely achievable even for very simple dynamical processes. Again, we have the same two options to model such systems: the *continuum approach* and the *agent-based modeling* approach.

The continuum approach averages over all nodes in each category defined by the possible states of the nodes. Densities of the different node states instead of total numbers are used so that equations become independent of the total number of nodes in the system. At time *t*, the density of infected nodes is represented by $I(t)$. Assuming that the total density is 1, the density of susceptible nodes is $1-I(t)$. It is also assumed that all nodes have the same degree *k*, that is, the same number of neighbors. We thus have a regular network. Based on these assumptions, we can write down the change in the average density of infected nodes over time for the SIS model as

$$\partial_t I = -\mu I(t) + \lambda k I(t)(1 - I(t)) \,. \tag{33}$$

In Equation 32, the first term on the right-hand-side considers infected nodes spontaneously recovering with unit rate μ. The second term represents the rate at which new infected nodes are generated in the network, that is, the density of healthy nodes acquiring the infection. It is proportional to the infection spreading rate λ, the density of susceptible nodes that might become infected $(1-I(t))$ and the number of (potentially infected) nodes in contact with a node given by the number of edges *k*. This last factor assumes the *homogeneous mixing hypothesis* which asserts that the force of the infection—the per capita rate of acquisition of the disease for the susceptible nodes – is proportional to the average number of contacts with infected individuals that is approximated as $kI(t)$.

Interestingly, even such a simple approximation of the SIS model can be used to central concepts in the modeling of spreading phenomena. Imposing the stationary condition $dI(t)/dt = 0$, we can obtain a nonzero stationary solution only if $\lambda > \mu/k$. This inequality defines the epidemic threshold λ_C below which the epidemic will die in a finite time; that is, $I = 0$ (see also Figure 8). Mathematically, it states that epidemics may propagate throughout the network only if the rate of contagion is sufficiently high to sustain or increase the number of infected nodes. This result may be recovered in many other models with complicated interactions. Finally, it is worth remarking that the continuum approach may be extended to consider stochastic formulations that allow the analytic inspection of the effect of noise in the dynamical process.

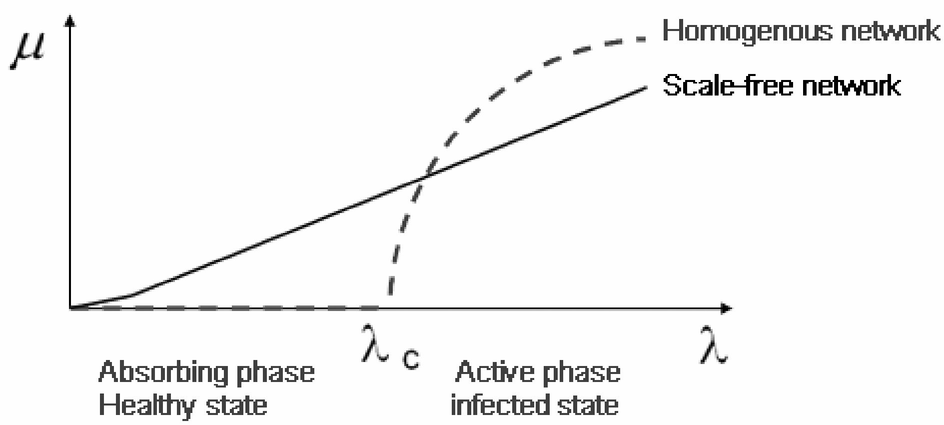


Figure 8. Schematic diagram of the SIS model for a homogenous networks and a scale free network. As can be seen, there exists no absorbing phase or healthy state for scale-free networks.

We must also keep in mind that the continuum approach is intrinsically considering a coarse grained perspective that does not take into account individual heterogeneity or other possible fluctuations. These include certain nodes in the system being more susceptible than others to attacks, for example, a computer virus may be targeted to attack specific operating systems or specific software. Possible fluctuations may also arise due to the non-uniformity in the degree distribution of the nodes or the change in behavior after infection has occurred, for example, staying in bed when infected with a cold instead of spreading it at school.

6.1.2 Agent-Based Modeling Approach

In most real world networks node attributes differ from node to node. There might also be sub-networks within a larger network that exhibit very different structure and behavior. For example, a computer network may be made of computers having different kinds of operating systems, which respond to a virus attack in different ways. In these cases, the continuum approach might not lead to solvable equations because of the large amount of parameters to be included in the analytical description. ABMs that can simulate non-heterogeneous nodes and local sub-networks need to be applied. Because ABMs can actually specify the interaction between the nodes to some extent, they can simulate complex and varied interactions between the individual nodes.

When modeling SIS using ABM, each individual node is again assumed to be either susceptible or infected. At each discrete time step, a model-specific update procedure is applied to each node. A node changes its state depending on the state of its neighboring nodes. Whether a node gets infected or not is a probabilistic process that can be simulated using *Monte Carlo* methods in which random number generators are used to simulate the random events of the dynamic process. The probabilities for changing from one state to another are the same. Because the simulations are based on random number generators we do not know exactly which nodes will become infected or healed. However, when applied over multiple time steps, this approach ideally leads to a system behavior that resembles the dynamics of a simulated real world system. Also, there is the added advantage that all attributes of each node can be determined in each time step and saved for further analysis.

Although extremely powerful, ABMs are often very intricate and the effect of any modeling assumption or parameter is not easy to study. ABMs in general have very few analytical tools by which they can be studied, and often no backward sensitivity analysis can be performed because of the large number of parameters and dynamical rules incorporated. This calls for a careful balance of the details included in the model and the interpretation of results obtained from computer simulations.

Ideally, the modeling approach to dynamical processes includes both methodologies at once. The microscopic model is formulated and the continuum equations are accordingly derived. The analytical predictions are then tested with ABM techniques in order to asses the role of noise and recover the obtained predictions.

6.1.3 Importance of Network Topology for Diffusion Processes

It is important to stress that the network topology, that is, the contact pattern among nodes, heavily influences the properties of dynamical processes. In the case of epidemic modeling, it is understood that there is no one-fits-all social network that might, even approximately, function as the prototypical substrate for epidemic modeling. Recently, major progress has been made in the understanding of disease spread through a wide array of networks with complex topological properties, for example, small world and scale-free distributions. It has been shown that scale-free networks do not have any epidemic threshold below which the infection cannot initiate a major outbreak (May & Lloyd, 2001; Pastor-Satorras & Vespignani 2001, 2002). In other words, the epidemic threshold above which the epidemics can spread is zero in the limit of an infinitely large network (see Figure 8). This new scenario is of practical interest in computer virus diffusion and the spreading of diseases in heterogeneous populations (Liljeros, Edling, Amaral, Stanley, & Aberg, 2001; Schneeberger, Mercer, Gregson, Ferguson, Nymukapa, Anderson, et al., 2004). It also raises questions on how to protect a network and how to find optimal strategies for the deployment of immunization resources. Based on the notion of a critical threshold in a homogeneous network, the usual strategy is to randomly immunize a certain percentage of the population to decrease the epidemic transmission rate. However, this will not work for a scale-free network, as is evident from Figure 8. A better strategy is to give immunization to highly connected individuals. Indeed, it is possible to show in mathematical terms that the immunization of a tiny fraction of the most connected individuals decreases the spreading of epidemics dramatically (see also the response of scale-free networks to the attack of highly connected nodes in Figure 9).

We have focused the previous discussion on epidemic modeling but it is clear that many of the approaches and insights can be readily transferred to many other processes such as the spread of ideas, scholarly knowledge and information. Heavy-tailed distributions have been observed in co-authorship networks (Newman, 2001) and paper-citation networks (Redner 1998; Börner et al. 2003). If we want to understand the diffusion of knowledge through co-author collaborations and paper citation linkages then an understanding of how epidemics spread may guide us to improved models of knowledge diffusion. Obviously, the goals are very different: we are typically interested in minimizing the spread of computer viruses and maximizing the spread of good ideas. The latter might be achieved by infecting the most connected individuals with the "idea."

Finally, we would like to point out the analogy between diffusion processes and search processes. Page limits do not allow us a detailed discussion of a subject that has received considerable attention in recent years; it is clear, however, that the topological properties of networks affect search and retrieval results. Search strategies that take into account the structure of a network have been demonstrated to be superior over those that do not. Two important quantitative measures for information retrieval are *recall* and *precision*. Recall is defined as the number of relevant documents retrieved as a fraction of all relevant documents and precision is defined as the total number of relevant documents as a fraction of all the documents retrieved. The best results in terms of recall and precision are achieved if global knowledge of all nodes and edges is available. In this case, the shortest path from the starting node to the target node can be computed and used to retrieve the desired information. In real world search scenarios, such global knowledge is rarely available. Here, knowledge about the general network properties can be exploited to improve search results. Examples are Kleinberg's (2000) work on search in *small world* networks the work on search in *scale-free* networks by Adamic, Lukose, Puniyani, and Huberman (2001). A comparison of topological properties and search performance in structured and unstructured peer-to-peer networks was presented by Fletcher, Sheth, and Börner (2004).

6.1.4 Network Stability, Optimization, and Protection

Deep understanding of dynamical processes on networks has lead to increased attention being given to the stability, optimization and protection of networks. These problems are emerging as fundamental issues in homeland security and reliability assessment of critical infrastructure networks, among others.

A first empirical analysis of the robustness of large-scale networks in the event of failures can be obtained by studying the topological response to the removal of nodes or edges. In this case, nodes are divided in two simple classes, one referring to functional nodes and the other to malfunctioning and thus removed nodes. Focusing on the effect of node removal, and assuming that all nodes are equally susceptible to failure, an instructive experiment can be performed on a connected graph of size N by looking at the effect achieved by removing a fraction g of randomly selected nodes. In order to monitor the response of the network to the damage, one can control several topological quantities related to network connectivity. A first and natural quantity to study is the size S_g of the largest component of connected nodes in the network after damage with respect to the size of the undamaged network. In particular, a ratio $S_g / N > 0$ indicates that a macroscopic fraction of nodes is still able to communicate. On the other hand, $S_g / N \cong N^{-1}$ signals that the whole network has been *compromised* by a fragmentation into very small disconnected components. A second quantity is the diameter of the network as a function of the fraction of removed nodes (Albert, Jeong, & Barabási, 2000). A natural question to ask in this context concerns the maximum amount of damage that the network can take, that is, the *threshold* value of the fraction of nodes that can be removed before the network functionality abruptly drops to zero. In regular meshes and random graphs with an exponentially fast decaying degree distribution there exists a threshold value g_c denoting the number of nodes that need to be removed before a network can be considered compromised.

Scale-free and heavy-tailed networks instead present two faces in front of component failures: they are extremely robust when faced with the loss of a large number of randomly selected nodes, but extremely fragile in response to a targeted attack (see Figure 11). When removing nodes at *random*, chances are that the largest fraction of deleted elements will have a very small degree. Hence, their deletion disconnects only a small number of adjacent edges and the overall damage of the network is limited. In order to do major damage, almost all nodes have to be randomly removed (Albert et al., 2000; Cohen, Erez, Ben-Avraham & Havlin, 2000). By way of contrast, a *targeted attack* of high degree nodes has a very disruptive effect. Scale-free networks are more vulnerable to a targeted attack than random graph models or regular meshes.

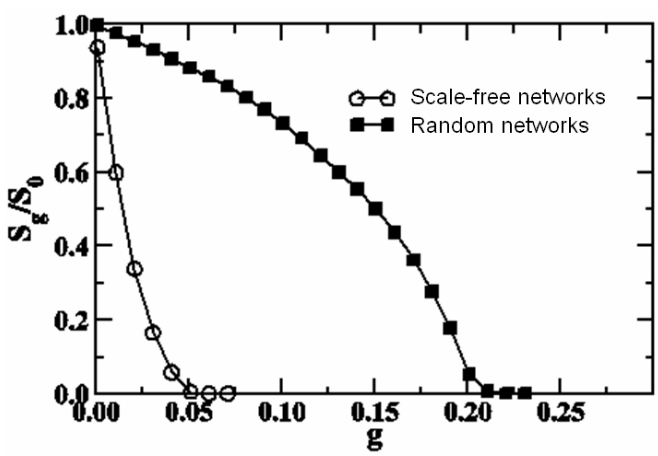


Figure 9. Topological resilience to targeted attacks of the scale-free Internet Router level network and an Erdős-Rényi random graph with the same average degree. As can be seen, the scale-free network is the more fragile. Even a removal density as low as g=0.05 suffices to fragment the whole network.

6.1.5 Propagation and Adaptation

Not all dynamical processes, however, concern the change of state of the nodes in a network. In many cases, we have dynamical entities such as people, information packets, energy or matter flowing through a network. Here, the dynamic description focuses on the dynamical entities and the state of each node depends on the entities present at that node. In all cases, a straightforward generalization of continuum and ABM approaches is possible. Both models allow the study of the dynamics of information or traffic flow taking place over a network. In particular, it is possible to study the robustness of networks as a dynamical process, which takes into account the time response of elements to different damage configurations. For instance, after any router or connection fails, the Internet responds very quickly by updating the routing tables of the routers in the neighborhood of the failure point. In general this adaptive response is able to circumscribe the damage but in some cases failures may cascade through the network, causing far more disruption than one would expect from the initial cause (Lee, 2005; Moreno, Pastor-Satorras, Vazquez, & Vespignani, 2003; Motter & Lai, 2002). This is typical of complex systems where emergent properties imply that events and information spread over a wide range of length and time scales.

This also means that small variations generally have a finite probability of triggering a system-wide response, the so called *critical behavior*. This happens through chains of events that eventually may involve a macroscopic part of the system and, in some cases, lead to a global failure. It is important to realize that in a large networked system this property is inherent to the system's complexity and cannot be changed by using local reinforcements or technological updates. We can vary the proportion of small or large events, but we have to live with appreciable probabilities for very large events: we must deal with the inherent complexity of the real world.

6.1.6 Coupling of Dynamics and Network Topology

As we have seen, the dynamical processes and the underlying topology are mutually correlated and it is very important to define appropriate quantities and measures capable of capturing their impact on the formation of complex networks. To carry out this task, we need to develop large empirical datasets that simultaneously capture the topology of the network and the time-resolved dynamics taking place on it. At the same time, a modeling paradigm that considers the dynamical processes on top of the evolving network is needed.

The recognition that many real world networks have nodes and edges of different strength and weights and are better described as weighted graphs is fostering the development of models that couple strengths features with the dynamical evolution of a network (Barrat et al., 2004). Moreover, modeling techniques based on the topology of the network incorporating only the net effect of regulatory interactions between components can provide a starting point for understanding the downstream impact of mutations or new drugs in biological networks. This is the case of the Boolean descriptions of networks in which the state of each component is either 1 (ON) or 0 (OFF). In this sort of model, time is divided into discrete steps and the next state for each node in the control network is determined by a Boolean function of its state and the state of the nodes that influence it. This mapping defines a discrete dynamical system that is much easier to analyze than the differential equations described in the master equation approach. The Boolean function for each node is determined from its state and the known activating and inhibiting interactions between nodes. When both activators and inhibitors act on a node, we assume that the inhibition is dominant; the node will turn off. The first step in validating a model like this is to determine whether it reproduces the normal behavior of the system. Recent evidence (Kauffman, Peterson, Samuelsson, & Troein, 2003) suggests that a Boolean model correctly integrates the topology and the nature of interactions in a gene control network and can also produce important insights into the dynamics of these networks. Although Boolean networks were initially proposed for genetic networks they have since been used to study other network types as well. They are easy to handle computationally and even dynamics can be modeled on these networks by allowing the Boolean variables to evolve in a continuous time.

7. Network Visualization

Visualization techniques can be applied to communicate the results of network measurement, modeling and validation or to visually compare the structure and dynamics of empirical and simulated networks. Techniques range from well designed tables that support easy comparison, via standard or customized graphs, to the layout of networks and the visualization of network dynamics. This section focuses on the visualization of network structure and dynamics and the research challenges that arise due to the size and complexity of real networks and the diversity of the network science applications.

We now introduce the basics of visualization design, give an overview of major matrix, tree, and graph layout algorithms, and discuss the visualization of network dynamics as well as interactivity design.

7.1 Visualization Design Basics

The design of effective visualizations that support visual exploration and decision making requires detailed knowledge about the intended user group and their information needs. This knowledge together with knowledge about human visual perception and cognitive processing constrain the "solution space" and guide the design of effective visualizations. Combined with existing knowledge on network sampling, measuring and modeling it provides a solid basis for the design of effective visualizations. Here, we provide information on how to acquire knowledge on users and their tasks, give pointers to research results on human visual perception and cognitive processing, and postulate basic network visualization design guidelines.

7.1.1 User and Task Analysis

Detailed knowledge of users and their tasks is needed to design visualizations that are legible and informative. If one does not understand how a user conceptualizes his/her world, what s/he needs to see in what context and when, then there will be little hope that the visualization is more than "eye candy." Even if you are the primary user of the visualization, for example, you want to confirm your scientific hypothesis or answer practical questions, it is advisable that you go through the process of externalizing your tasks and information needs.

Information on users and their tasks can be acquired via interviews, questionnaires, observation, or analysis of existing documents and manuals. Excellent introductions on how to conduct a user/tasks analysis can be found in interface design textbooks, for example, Hackos and Redish (1998). Issues to bear in mind include the following:

- Who are your users (profession, location, gender, age, or lifestyle preferences)?
- What is their level of technical and subject expertise? The visual language used will have to match the user's understanding.
- What is the visualization context? Describe your users' physical and social environments. Note any environmental challenges such as poor lighting or noise, and any technical challenges such as screen size, resolution, color quality and number of displays. Determine what hardware and browser software, monitors and screen resolutions your audience uses.
- Describe scenarios of use or those situations or circumstances in which the visualizations may be used.
- Exactly what do your users need to understand, discover or communicate; and in what sequence?

Make sure you know what task the visualization is intended to support. Examples are the identification of trends in the data, outliers, maxima and minima, boundaries, clusters and structure, dynamics, and related information. Note that each of these tasks potentially demands a very different visualization design. Also keep in mind that gaining insight typically employs different stages of inquiry such as observation, exploration, model construction, simulation, verification, interpretation and communication of results (Hanson, 1958; Popper, 1959).

7.1.2 Human Visual Perception and Cognitive Processing

Human visual perception and processing capabilities are nearly constant. What you learn about them today will still be valid in 50 years unless scientific evidence falsifies current knowledge. Hence, it makes sense to acquire detailed knowledge on human perception and processing and to use it to constrain the quite large design solution space. Books by MacEachren, (1995), Palmer (1999) and Ware (2000) are excellent resources for a detailed examination of human perception and cognition. It is beyond the scope of this chapter, however, to provide a comprehensive review of research in this area.

7.1.3 Basics of Network Visualization Design

Knowledge on users and their tasks, as well as on human visual perception and cognitive processing, form the basis for the design of effective visualizations. In general, visualization design comprises decisions about (a) employed metaphors and reference systems, (b) the number and type of data layers, and (c) visual mappings. There is a strong interplay among the three elements, for example, the selection of a different metaphor might very well influence the number of data layers and the visual mappings employed. Hence all three elements should be dealt with in concert. Extensive knowledge of existing algorithms and visualizations (e.g., (Di Battista, Eades, Tamassia, & Tollis, 1999; Freeman, 2000; Herman, Melancon, & Marshall, 2000), close collaboration with users, thorough testing and patient (re)design of visualizations will provide the best results.

7.1.4 Metaphors and Reference Systems

Metaphors should be selected so that they best match the conceptualization and information needs of the intended user group(s). As for the visualization of networks, diverse metaphors have been suggested such as time lines, subway maps, galaxy visualizations of networks, or the overlay of nodes and edges on reference systems such as geospatial maps.

Reference systems refer to temporal, geospatial, semantic and other substrates that can be used to contextualize and ease the understanding of network layouts. If time is important and a 2D layout is desirable, then using one axis to order nodes, for example, by time, is appropriate. An example of time-ordered network layouts are *Historiograph* visualizations of paper-citation networks generated by Garfield's HistCite^{TM} tool (Garfield, Sher, & Torpie, 1964; Pudovkin & Garfield, 2002); (see Figure 10, left). If a highway, airline or Internet traffic networks need to be visualized, then a geospatial substrate map might be best (see Figure 10, middle). In some cases, for example, when visualizing a co-author network, a free layout of nodes that reveals the topology of the network might work best (see Figure 10, right).

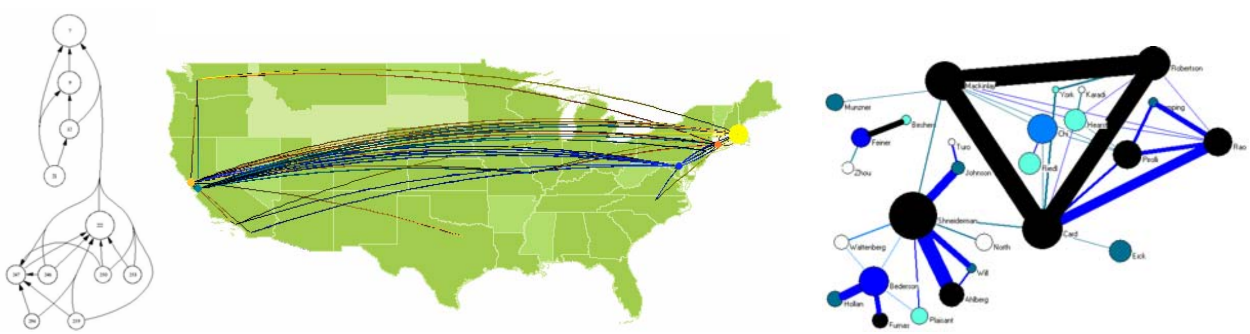


Figure 10. Temporal reference system used to display a citation network as so called Historiograph (left), layout of paper citations on a geospatial map of U.S. (middle), and semantic space of co-author relations (right).

7.1.5 Visual Layers

V*isual layers* ease the readability of network visualizations. In most cases, a network visualization will comprise a base map (e.g., a map of the U.S.), an information overlay (e.g., router nodes and edges representing Internet traffic), labels (e.g., names of major cities), and a legend (e.g., a title, short explanation of unique features, a description of all visual encodings). Note that the credibility of any visualization can be considerably improved if the name of the map maker and date of creation is given and information on the displayed dataset and its manipulation is provided.

Last but not least, many visualizations benefit from being interactive (see the sub-section on interaction and distortion techniques). Interactivity design can be conceptualized as an additional layer that is receptive to and reflects user actions, for example, by indicating the zoom level.

7.1.6 Visual Mappings

Given appropriate metaphors, reference systems and visual layers, one needs to define the following: What data entities should be represented as nodes? What relationships are important and should be represented as edges? What node/edge attributes are important and need to be visually encoded? Are there any subnetworks or "backbone" structures that need to be made visible? What subset of nodes, edges, subgraphs, and backbones needs to be labeled and how? If the network is large, then one also has to decide what data can be omitted to provide users with a meaningful overview of the dataset and to enable the user to gain access to the omitted data.

In some cases the answers to these questions are straightforward. In others, considerable thought is required to come up with the right conceptualization and representation. Examples that inspired subsequent breakthroughs are Euler's rendering of the Königsberg bridges problem as a graph in which nodes represent land masses and edges represent bridges (Euler 1736) or Moreno's (1934) first visualizations of social networks by graph structures. Moreno's so called *Sociograms* used directed edges, color, different node shapes, and the location of nodes to show the status of a person, the relationships among a group of people or to stress structural features of a network (Moreno, 1943).

Today, there exists a multitude of software libraries and tools that make it easy to analyze a network and to generate a network visualization. Some of the tools support dynamic changes of the mapping between data and their visual representation. A comprehensive review and comparison of software for social network analysis can be found in Huisman and Van Duijn (2005).

7.2 Matrix Visualization

As discussed in the sub-section on graphs and subgraphs, graphs are commonly represented by adjacency matrices (see Figure 2 for examples). The adjacency matrices can be visualized by dense pixel displays, also called structure plots, which use the space created by an ordered list of all nodes as a typically two-dimensional reference system. The existence of an edge between two nodes *(a,b)* is indicated by the shading of the area *(i,j)* where *i* is the row for *a* and *j* is the column for *b*.

Figure 11 a-f shows dense pixel displays for the six graphs given in Figure 2. Our visual system quickly identifies the symmetrical nature of the interlinkage patterns that is indicative of undirected graphs (see Figure 11a-d). Only the lower or upper half of the matrix needs to be displayed. Directed graphs can quickly be identified by their non-symmetric nature (see Figure 11e and f).

Dense pixel displays can be used to display the structure of very large graphs. Figure 11g shows a medium sized graph in which the existence of an edge between two nodes is indicated by the shading of exactly one pixel. Networks that have more nodes than there are pixels on a monitor can be represented by averaging over a certain number of nodes and edges; for example, when displaying the interlinkage pattern of 10,000 nodes using a space of 1,000 x 1,000 pixels, each pixel represents the linkage density of 10 x 10 nodes. Edge weights can be represented by using not only black and white pixel values but grey tones or color.

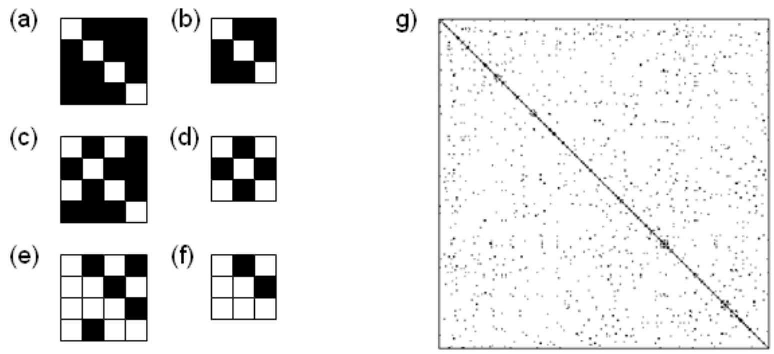


Figure 11. Dense pixel displays structure plots of small and larger graphs. Dense areas in the matrix reveal graph structure. For example, the high density of node linkages in the diagonal of the plot in Figure 11g indicates that the majority of nodes has links to themselves. This network property is very common, for example, in citation networks where it indicates a high level of self citation. Vertical or horizontal lines can be easily spotted, representing nodes with high in, out, or total degree. Other dense pixel areas might indicate clusters.

Obviously, the ordering of the nodes has a strong effect on the patterns that are visible. In 1972, Hartigan introduced block clustering (Hartigan, 1972). The concept of reordering was first introduced by Bertin (1981).Generalized association plots were applied and generalized by Chen (1999). Blockmodeling (Doreian, Batagelj, & Ferligoj, 2005) is an empirical technique that can be used to reduce a large, potentially incoherent network into a smaller comprehensible structure that can then be visualized and interpreted more readily. Current research seeks to develop reordering algorithms that help reduce noise and emphasize structural features (Mueller, 2004). Common choices are ordering by degree, by connected components, by core number or core levels, and according to other node properties and otherwise identified clusters.

7.3 Tree Layout

Many networks are trees. Examples of trees are family trees, phylogenetic trees, organizational charts, classification hierarchies or directory structures. Diverse algorithms exist to layout trees (see Figure 12). Depending on the dataset, users and their tasks different algorithms are appropriate.

Dendrograms are a simple yet effective method for the representation of tree data. Most commonly, dendrograms are drawn in a *Cartesian* layout, as an upright or left to right tree. The branching tree-like diagram effectively represents the hierarchical relationships among nodes. The length of edges might vary to represent edge attribute values. An illustrative use of dendrograms is the display of phylogenetic trees.

Radial tree layout (Di Battista et al., 1999; Herman et al., 2000) supports visualization and manipulation of large hierarchies. In a radial tree, the focused node is placed in the center of the display and all other nodes are rendered on appropriate circular levels around the selected node. The further away a node is from the center, the smaller it is rendered. This way, focus and context of very large tree structures can be displayed on a screen of limited size.

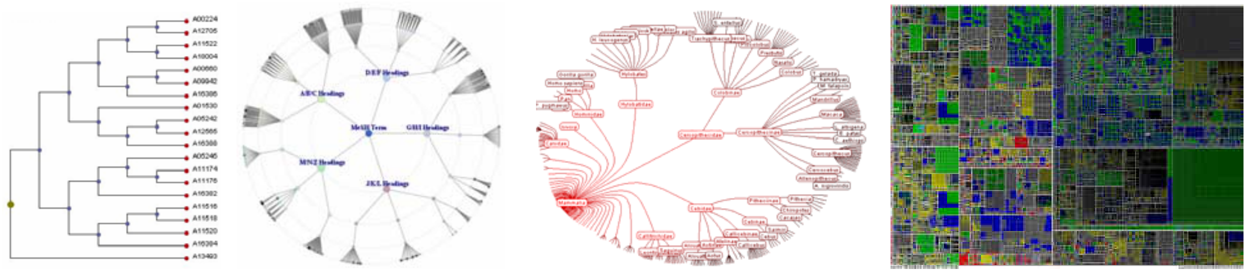


Figure 12. Dendrogram (left), radial layout (middle left), hyperbolic tree layout (middle right), and treemap layout (right).

Hyperbolic tree layout is based on *Poincaré's* model of the (hyperbolic) non-*Euclidean* plane. Lamping, Rao, and Pirolli (1995) rediscovered hyperbolic spaces in 1995 for information visualization and Munzner (1998) developed the first 3D hyperbolic viewer. In a hyperbolic tree, the root is placed at the center; the children are placed at an outer ring in equal distance to their parents. The circumference jointly increases with the radius and more space becomes available for the growing number of intermediate and leaf nodes. Whereas the radial tree uses a linear layout, the hyperbolic layout uses a nonlinear (distortion) technique to accommodate focus and context for a large number of nodes.

In radial tree and hyperbolic tree layouts, node overlapping is prevented by assigning an open angle for each node. All children of a node are then laid out in this open angle. Frequently, the tree visualization is interactive – users can click on a node to initiate its fluent movement into the center or can grab and reposition a single node.

Treemaps, developed in the Human-Computer Interaction Lab at the University of Maryland (Shneiderman 1992, 2005), trace their ancestry back to *Venn* diagrams (Venn, 1894/1971). They use a space filling technique to map a tree structure (for example, a file directory) into nested rectangles with each rectangle representing a node. A rectangular area is first allocated to hold the representation of the tree, and this area is then subdivided into a set of rectangles that represent the top level of the tree. This process continues recursively on the resulting rectangles to represent each lower level of the tree, each level alternating between vertical and horizontal subdivision. Upon completion, all child-rectangles are enclosed by their parent-rectangle. Area size and color can be used to encode two node attribute values, for example, file size and age, respectively. Node children can be ordered by area size leading to a better understanding of their size differences. Tree maps have been successfully used to identify large files in nested directory structures or to make sense of stock option trends.

7.4 Graph Layout

Graph layout algorithms can be applied to arbitrary graphs (see Figure 14). They aim to sort a set of randomly placed nodes into a layout that satisfies aesthetic criteria for visual presentation such as non-overlapping, evenly distributed nodes, symmetry, uniform edge lengths, minimized edge crossings, and orthogonal drawings that minimize area, bends, slopes, and angles. The criteria may be relaxed to speed up the layout process (Eades, 1984).

In some cases, it is desirable to order nodes by their attributes, for example, time or size, or by their structural features, for example, their degree. An example is *Historiographs*, introduced in the section on visualization design basics and shown in Figure 10, left. They vertically organize nodes (representing papers) according to their publication date. Nodes are then placed horizontally in a way such that the resulting layout of nodes and edges (representing citation links) fulfill the aesthetic criteria previously discussed.

Force-directed layout (FDL) algorithms were originally introduced by Eades (1984). They are commonly used to display general graphs, both directed and undirected, cyclic and acyclic. Here repulsive forces F_r are applied in inverse proportion to the distance d between any two nodes i and j, and attractive forces F_a in logarithmic proportion to the distance between two nodes linked by an edge:

$$F_r(i,j) = \frac{C_3}{d} \text{ and } F_a(i,j) = C_1 * \log(\frac{d}{C_2}), \qquad (34)$$

were C_1,C_2 , and C_3 are constant values. For all nodes, the algorithm calculates each force against all other nodes, sums them as the force of all connected nodes and moves the node appropriately. This way, a set of randomly placed nodes is sorted into a desirable layout. However, the complexity of the algorithm increases quadratic with the number of nodes, that is, $O(N^2)$, making it unsuitable for large data sets.

Extensions of Eades's algorithm provide methods for the intelligent initial placement of nodes, cluster the data to perform an initial coarse layout followed by successively more detailed placement, and use grid-based systems for dividing up the dataset. For example, Graph EMbedder (GEM) attempts to recognize and forestall non-productive rotation and oscillation in the motion of nodes in the graph as it cools (Frick, Ludwig, Meldhau, 1994). Walshaw's (2000) multilevel algorithm provides a "divide and conquer" method for laying out very large graphs by using clustering. VxOrd (Davidson, Wylie, & Boyack, 2001) uses a density grid in place of pair-wise repulsive forces to speed up execution and achieves computation times in the order of O(N). It also employs barrier jumping to avoid trapping of clusters in local minima. An extremely fast layout algorithm for visualizing large-scale networks in three-dimensional space was proposed by Han and Ju (2003). Today, the algorithm developed by Kamada and Kawai (1989) and Fruchterman and Reingold (1991) are most commonly employed, partially because they are available in the widely used *Pajek* visualization toolkit (de Nooy, Mrvar, & Batagelj, 2005).

Since there are major differences in the visualization of small, medium size and large networks, they are discussed separately.

7.4.1 Small Networks

By small networks, we mean networks that have up to 100 nodes. Examples are social networks, food webs, or import and export among countries. Here, all nodes and edges and many of their attributes can be shown. General mappings for nodes and edges are as follows. The node (area) size is commonly used to encode a primary value such as size, importance, power, or activity level. Node color is often employed to encode secondary values such as intensity or age. Node types are often encoded by node shapes – especially if only a few node types can be defined or node types can be aggregated into a smaller number of main types for which an easily distinguishable shape encoding exists. Sample layouts of one and the same small world network using different layout algorithms are shown in Figure 13.

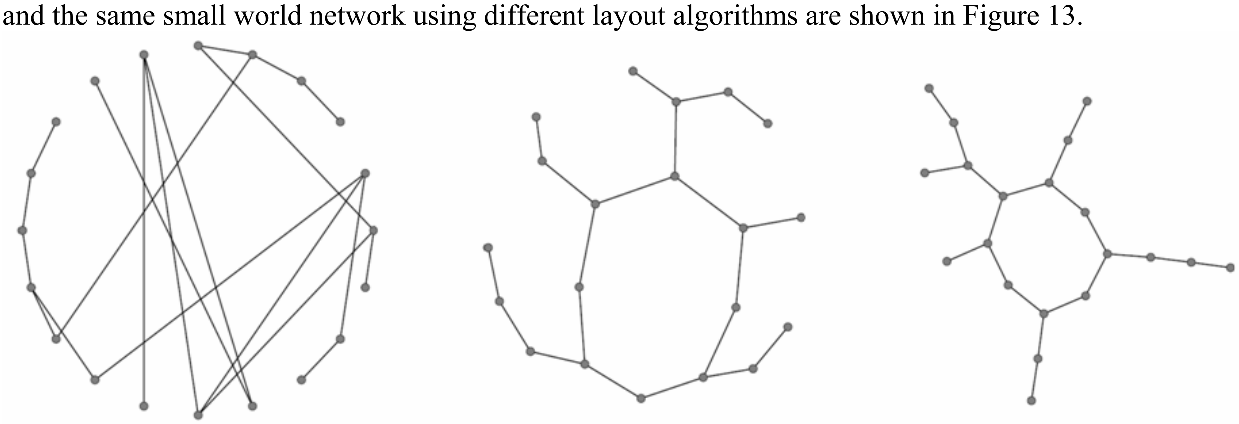


Figure 13. Circle layout (left), Fruchterman-Rheingold layout (middle), and Kamada-Kawai Layout (right) of a small world network.

7.4.2 Medium Sized Networks

By medium sized networks we mean networks that have more than 100 and up to 1,000 nodes. Examples are gene association networks, metabolic networks or economic networks. Most nodes can be shown, but not all their attributes or labels. Typically, it is not possible to show all edges. Frequently, the number of nodes, edges and their displayed attributes need to be reduced intelligently. For example, it might be beneficial to identify major components in a network and to represent all network components of size one and two simply by displaying one and two nodes respectively and using a number next to them to indicate how many components of this size exist. In other cases, it might suffice to determine and depict the giant connected component of the network exclusively and to provide information on the size and number of other components in a tabular format.

Major design strategies include showing only important nodes, edges, labels, and attributes; using most appropriate metaphors and reference systems to lay out nodes spatially, to supply landmarks that guide orientation and navigation, and to provide focus and context.

7.4.3 Large Networks

By large networks we refer to networks that have more than 1,000 nodes. Neither all nodes nor all edges can be shown at once. Sometimes, there are more nodes than pixels. Examples are communication networks such as the Internet, telephone network, wireless network; network applications such as the World-Wide Web, E-mail interaction networks; transportation networks/road maps; relationships between objects in a database such as function/module dependency graphs, knowledge bases; or scholarly networks.

Major challenges comprise the selection of important nodes, edges, subgraphs and backbones and their positioning; the de-cluttering of links; labeling; as well as navigation and interaction design. A major design strategy is the tight coupling of data analysis and visualization.

For example, important nodes, edges or subgraphs can be identified using the measurements introduced in the section on node and edge properties. It is important to show strong and weak links. *Pathfinder network scaling* (Schvaneveldt, 1990) is frequently used to identify the “backbone” of a network. Major network components can be identified using the algorithms introduced in the section on local structure. These components can each be presented as a “super node”, the (area) size of which might represent the number of its nodes.

Hierarchy visualizations of the nested structure of a network or a visualization of major clusters and their interconnections help us to understand the global structure of a network. The cut-out of subnetworks or focus and context visualizations support the examination of local network properties. The focus and context approach shows only one cluster in detail; other clusters are indicated by single nodes to provide context. Careful interactivity design aims to support overview, zoom and filter, as well as the retrieval of details on demand (see the section on interaction and distortion techniques).

7.5 Visualization of Dynamics

Almost all networks are optimized to support diverse dynamic processes. Electricity and transportation systems are optimized to effectively distribute tangible and intangible objects. Friendship networks are often support networks; our brain cells grow in response to the input they receive. As discussed in the section on modeling dynamics on networks, there exist major difference in studying the evolution of networks and studying dynamic processes on networks. Ideally, both could be studied, understood and communicated in concert.

7.5.1 Visualizing Network Evolution

Visualizations that show the evolution of networks in terms of attribute changes or structural changes (decreases or increases in the number of nodes and edges) can be divided into two general types: algorithms that process data on network changes incrementally and algorithms that identify and aim to visualize network changes based on the complete dataset. Examples of incremental visualizations, also called organic visualizations (Fry, 2000), are Fry's Anemone (http://acg.media.mit.edu/people/fry/anemone) and Gnutellavision (Dhamija, Fisher, & Yee, 2000; Yee, Fisher, Dhamija, & Hearst, 2001); (see Figure 14, left). Note that Gnutellavision provides interactive exploration of subregions of a graph from different perspectives.

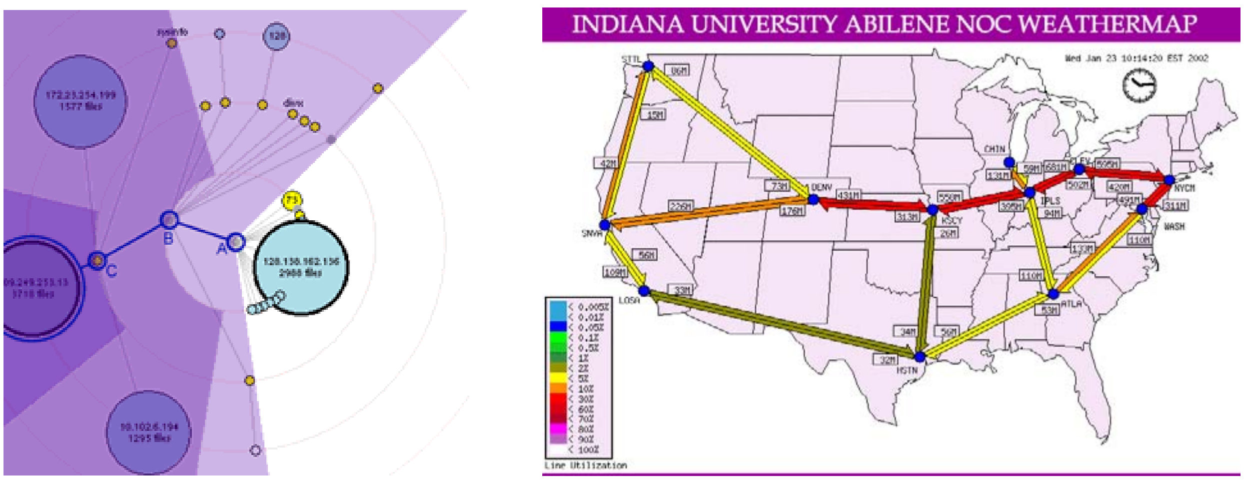


Figure 14. Dynamic network visualizations with fixed substrate map: Abilene network (left) and variable radial tree layout: Gnutellavision (right).

Examples of visualizations that aim to depict the change of a network over time based on a complete dataset are *Netmap* (Standish & Galloway, 2001) which visualizes Tierra's tree of life or routers and their interconnections around a certain host and Chen's *CiteSpace* system (Chen, 2004) which visualizes intellectual turning points in scholarly networks.

7.5.2 Visualizing Dynamics on Networks

In some application domains, the structure of a network is fixed and the flow dynamics over this fixed substrate is of interest. An example is the Abilene Weather forecast map at http://loadrunner.uits.iu.edu/weathermaps/abilene/, given in Figure 14, right. Other examples are weather forecast maps or migration maps that often use a geospatial substrate map. Here the reference system, the network, and the activity need to be depicted. Activity is often indicated by line overlays. Arrow heads can be used to indicate directedness, yet if many arrows go into one node this simply leads to an increase in the size of the node. Hence line thickness, shading, color and other forms of coding are frequently employed to indicate edge directions. High amounts of traffic quickly lead to cluttered displays. Intelligent aggregation methods that identify and visually encode major traffic flows need to be employed. Note that visualizations of network dynamics can be static or animated over time.

7.6 Interaction and Distortion Techniques

Many network datasets are too large to be displayed at once. Often, only the "backbone" or major subgraphs of a network and important "landmark" nodes can be displayed. Additional information associated with single nodes, links or subgraphs might need to be retrieved on demand. Even if large, tiled display walls are available or high density printouts can be used, in many cases a user does not want to see a million nodes at once. Instead, and in accordance with Shneiderman's (1996) information seeking mantra, users prefer to have an overview of the most general structure (major clusters, backbone of the network) first. Then, they may well pan, zoom and filter out subnetworks of interest. Finally, they might request details of a certain subnetwork on demand. Consequently, network visualizations should be designed with interactivity in mind. Most commonly, network visualizations support:

- Conditioning: filter, set background variables and display foreground parameters
- Identification: highlight, color, shape code
- Parameter control: line thickness, length, color legend, time slider, and animation control
- Navigation: Bird's Eye view, zoom, and pan
- Information requests: Mouse over or click on a node to retrieve more details or collapse/expand a subnetwork

When designing interactive visualizations one needs to keep in mind that the bandwidth from computer to human is much higher than the other way round. Ideally, occasional user steering leads to the computer-generated display of visualization sequences that can be readily perceived via the high bandwidth channel of our visual system and cognitively understood. If possible, the user should obtain the illusion of direct control. Hence, visual feedback should be provided within 1/10 second (Shneiderman, 1987). When transitioning from one visualization to the next, it is advantageous to use animation instead of jumps to support object constancy and navigation. Keeping information density almost constant at all zoom levels is important.

Curiosity is an important ingredient of scientific discovery. It can be supported by implementing a universal undo, making it impossible for a user to irrevocably get lost or in flounder. In general, the user needs to be kept "in flow" (Csikszentmihalyi, 1991). Boredom (too little information, too slowly) and anxiety (too much information, too fast) need to be avoided (Bederson, 2004).

8. Discussion and Outlook

As we have seen, networks can be found in biological, social, economic, informational and many other systems and processes. Although the advances that we have witnessed in the past few years have been spectacular, in terms of both their impact on basic science and practical implications, they have highlighted the incompleteness of our knowledge as well. Network science is going to face a number of important challenges and questions in the next few years.

To give a concrete example, let us consider the area of scholarly information access and management that represents an important focus for the readership of this volume. Today, our main means of accessing our collective knowledgebase are search engines. Companies such as Google and Microsoft claim that a few good keywords and local link traversal suffice to make use of mankind's collective knowledge. Search does work well for fact retrieval. Yet, it is instructive to see what coverage a dataset has, what major clusters exist, from which clusters search results were drawn, or how retrieved documents interrelate. Private and commercial entities have expanded great effort to develop directory structures, classification hierarchies, and other means to organize knowledge. However, it appears to be difficult – if not humanly impossible – to design and update an organizational schema comprising hundreds of thousands of classes so that it captures the evolving structure of a rapidly increasing scholarly document data set of potentially millions of entries. Without an "up" bottom when conducting a search and without organizational schemas that expeditiously and comprehensively organize scholarly data we are bound to the ground. Today, our bird's-eye views are at best one meter above the landscape of science whereas a global view would be possible only from a 1,000-meter height. Given nearly constant human abilities, our distance above ground is decreasing as the amount of information is growing. Scientists are forced into narrow specialties, scrutinizing a tiny portion of science's shoreline. They are largely ignorant of a vast hinterland containing mountain ranges of data and vast floodplains of experience. Yet, a more global view of science is required to identify major experts, to learn how knowledge evolves and interrelates, to understand what duplications or complementary approaches exist, what research areas are emerging. Such information is vital for funding agencies, companies and researchers (for example, for setting research priorities) but is also beneficial to science education and appreciation. The study of science by scientific means requires the analysis of terabytes of scholarly data. It requires the measurement and modeling of large-scale coupled networks of authors/inventors/awardees/investors, papers/grants/patents, and so on, and their many interrelations (Börner et al., 2003). These networks grow continuously and they are used to diffuse knowledge, money and reputation. The study of feedback cycles – for example, the fact that authors who publish highly cited papers have a greater chance of having their proposals funded and hence securing more resources to increase their chances of publishing yet more highly cited papers – seems to be particularly important for understanding the structure and dynamics of science.

These considerations translate into a set of theoretical and practical challenges that ranges from the study of multiple overlapping and interacting networks to the design of effective visualizations that show the structure, evolution and dynamics of very large-scale (more than a million nodes) networks. We need to understand the interfacing and interaction of *networks of networks* and to start large-scale measurement projects for gathering empirical data that comprise not only the physical properties of multi-scale networks but also their usage. The theory and tool development required to address the challenges listed above will benefit enormously from a cyberinfrastructure for network science research. This infrastructure will need to provide access to data, services and computing resources, as well as expertise (Börner, 2006).

It is clear that different application domains will pose different challenges due to the availability of data as well as scientific and practical demands. In addition, the different sciences will make very different contributions: mathematicians and statisticians will advance network science theory; physicists will continue their search for universal laws; biologists will aim to uncover the secret of life; social scientists will continue to study the social fabric in which we are embedded; computer scientists and information scientists will develop effective and scalable algorithms and infrastructures; graph drawing experts and designers will aim to improve our ability to visually communicate network structure, evolution, and usage. Network science has true potential to integrate the knowledge acquired in diverse fields of science. Given the ubiquity of networks in our world, the results of the theoretical and practical study of networks might help solve some of the major challenges confronting society. It is our hope that this chapter succeeds in paving the way for an adoption of approaches and theories developed outside information science and computer science yet immediately applicable to information science and computer science problems. We also hope that the chapter inspires new collaborations across scientific disciplines and the development of theoretical approaches with the true potential for practical application.

Acknowledgments

We would like to thank Rebecca Rutherford for her assistance in the conversion of formulas and citation entries in the preparation of the manuscript. We benefitted from discussions with Stanley Wasserman, Ariel Balter, Kevin W. Boyack, Joseph Cottam, Ketan K. Mane, Shashikant Penumarthy, and Elijah Wright. We thank the anonymous reviewers that provided detailed comments on an earlier draft of the chapter. This work is supported by a National Science Foundation grant under IIS-0513650 to the first and third author and an NSF CHE-0524661 and CAREER IIS-0238261 award to the first author. The second author is supported by a James S. McDonnell Foundation grant.

Endnotes

1. By "networks" we refer to any system that allows its abstract/mathematical representation as a graph, that is, a set of nodes and edges.

2. The word *equilibrium* refers to a situation in which the probability distribution describing the possible states is not biased or constrained. This happens when external forces constrain the system to be on a specific subset of the allowed states. This will be properly defined in the context of dynamical modeling, see sub-section on modeling evolving networks.

References

Adamic, L., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, *25*, 211–230.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., & Huberman, B. A. (2001). Search in power-law networks. *Physical Review E*, *64*, 46135.

Albert, R. & Barabási, A. R. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance in complex networks. *Nature*, *406*, 378.

Bader, G. D., & Hogue,C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, *20*, 991–997.

Banks, B. D., & Carley, K. (1996). Models for network evolution. *Journal of Mathematical Sociology*, *21*, 173–196.

Barabási, A.-L. (2002). Linked: How everything is connected to everything else and what it means: Perseus Books Group.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Barabási, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, *272*, 173.

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, I, 101–113.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 3747–3752.

Barrat, A., & Weigt, M. (2000). On the properties of small-world network models. *European Physical Journal B*, *13*, 547–560.

Barthélémy, M., & Amaral, L. A. N. (1999a). Erratum: Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, *82*, 5180.

Barthélémy, M., & Amaral, L. A. N. (1999b). Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, *82*, 3180–3183.

Bederson, B. B. (2004). Interfaces for staying in the flow. *Ubiquity*, *5*(27) http://www.acm.org/ubiquity/views/v5i27_bederson.html.

Berg, J., & Lassig, M. (2002). Correlated random networks. *Physical Review Letters*, *89*, 228701–228705.

Bertin, J. (1981). *Graphics and graphic information-processing*. Berlin: Walter de Gruyter.

Bettencourt, L.M.A., Cintron-Arias, A., Kaiser, D.I., & Castillo-Chavez, C. (2005) The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models. arXiv:physics. 0502067.

Bianconi, G., & Capocci, A. (2003). Number of loops of size h in scale-free networks. *Physical Review Letters*. *90*, 078701-1 –078701-4.

Blatt, M., Wiseman, S., & Domany, E. (1996). Superparamagnetic clustering of data. *Physical Review Letters*, *76*, 3251–3254.

Blatt, M., Wiseman, S., & Domany, E. (1997). Data clustering using a model granular magnet. *Neural Computing*, *9*, 1805–1842.

Bollobas, B. (1998). *Modern graph theory*. New York: Springer.

Börner, K. (2006). Semantic association networks: Using Semantic Web technology to improve scholarly knowledge and expertise management. In V.G. C. Chen (Ed.), *Visualizing the Semantic Web* (pp. 183–198). Berlin: Springer.

Börner, K., Chen, C., & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science & Technology*, *37*, 179–255.

Börner, K., Maru, J. T., & Goldstone, R. L. (2003). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl_1), 5266–5273.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*, 163–177.

Brandes, U., & Erlebach, T. (2005). *Network analysis: Methodological foundations*. Berlin: Springer.

Broder, A. Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Computer Networks*, *33*, 309–320.

Buchanan, M. (2002). *Nexus: Small worlds and the groundbreaking science of networks*. New York: Norton.

Burda, Z. and Krzywicki, A. (2003) Uncorrelated random networks. Physical Review E, 67. 046118.

Burda, Z., Jurkiewicz, J. and Krzywicki, A. (2003) Statistical mechanics of random graphs. cond-mat/0312494.

Carrington, P., Scott, J., & Wasserman, S. (2004). *Models and methods in social network analysis*. New York: Cambridge University Press.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S. R., Raghavan, P., et al. (1999). Hypersearching the Web. *Scientific American*, *280*(6), 44–52.

Chartrand, G., & Lesniak, L. (1986). *Graphs and digraphs*. Menlo Park, CA: Wadsworth & Brooks/Cole.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl. 1), 5303–5310.

Chen, C. H. (1999). Extensions of generalized association plots (GAP). Paper presented at the Annual Meeting of the American Statistical Association, Baltimore, MD.

Chung, Y. M., & Lee, Y. J. (2001). A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, *52*, 283–296.

Clauset, A., & Moore, C. (2005). Accuracy and scaling phenomena in Internet mapping. *Physical Review Letters*, *94*, 018701.

Cohen, R., Erez, K., Ben-Avraham, D., & Havlin, S. (2000). Resilience of the Internet to random breakdowns. *Physical Review Letters*, *85*, 4626–4629.

Cronin, B., & Atkins, H. B. (Eds.). (2000). *The web of knowledge: A Festschrift in honor of Eugene Garfield*. Medford, NJ: Information Today.

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience*. New York: HarperCollins.

Daley, D. J., Gani, J., & Cannings, C. (1999) Epidemic modeling: An introduction. Cambridge University Press.

Dall'Asta, L., Alvarez-Hamelin, I., Barrat, A., Vazquez, A., & Vespignani, A. (2005). Traceroute-like exploration of unknown networks: A statistical analysis. In A. López-Ortiz & A. Hamel (Eds.). *Combinatorial and algorithmic aspects of networking* (pp. 140–153). Berlin: Springer.

Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proceedings of the IEEE Symposium on Information Visualization*, 23–30.

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge, UK: Cambridge University Press.

de Jong, E. D., Thierens, D., & Watson, R. A. (2004). Hierarchical genetic algorithms. *Parallel Problem Solving from Nature: Proceedings of the 8th International Conference*, 232–241.

Deane, C. M., Salwinski, L., Xenarios, I., & Esenberg, D. (2002). Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Molecular Cell Proteomics*, *1*, 349–356.

Dhamija, R., Fisher, D., & Yee, K.-P. (2000). *Gnutellavision: Real-time visualization of a peer-to-peer network*. Retrieved March 10, 2006, from http://bailando.sims.berkeley.edu/infovis/gtv

Di Battista, G., Eades, P., Tamassia, R., & Tollis, I. G. (1999). *Graph drawing: Algorithms for the visualization of graphs*. Upper Saddle River, NJ: Prentice Hall.

Dill, S., Kumar, S. R., McCurley, K. S., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2002). Self-similarity in the Web. *ACM Transactions on Internet Technology*, *2*, 205–223.

Domany, E. (1999). Superparamagnetic clustering of data: The definitive solution of an ill-posed problem. *Physics A*, *263*, 158–169.

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized blockmodeling*. Cambridge, UK: Cambridge University Press.

Dorogovstev, S. N., & Mendes, J. F. F. (2002). Evolution of random networks. *Advances in Physics*, *51*, 1079–1187.

Dorogovstev, S. N., & Mendes, J. F. F. (2003). *Evolution of networks*. Oxford, UK: Oxford University Press.

Dorogovstev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2003). Principles of statistical mechanics of random networks. Nuclear Physics B, 666: 396.

Eades, P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, *42*, 149–160.

Erdös, P., & Rényi, P. (1959). On random graphs. *Publicationes Mathematicae*, *6*, 290–297.

Euler, L. (1736). Solutio problematis ad geometriam situs pertinentis. *Commetarii Academiae Scientiarum Imperialis Petropolitanae*, *8*, 128–140.

Farkas, I., Derenyi, I., Palla, G. & Vicsek, T. (2004) Equilibrium statistical mechanics of network structures. Lecture Notes in Physics, 650. 163-187.

Flake, G., Lawrence, S., & Giles, C. L. (2000). Efficient identification of Web communities. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150–160.

Fletcher, G., Sheth, H., & Börner, K. (2004). Unstructured peer-to-peer networks. In G. Moro, S. Bergamaschi, & K. Aberer (Eds.), *Topological properties and search performance* (Lecture Notes in Computer Science, 3601) (pp. 14–27). Berlin: Springer.

Frank, O. (2004). Network sampling and model fitting. In P. Carrington, J. Scott & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 31–56). New York: Cambridge University Press.

Frank, O., & D. Strauss (1986). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Freeman, L. C. (1977). A set of measuring centrality based on betweenness. *Sociometry*, *40*, 35–41.

Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure*, 1(1) http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

Frick, A., Ludwig, A., & Mehldau, H. (1994). A fast adaptive layout algorithm for undirected graphs. *Proceedings of Graph Drawing '94*, 388–403.

Fronczak, P., Fronczak, A. and Holyst, J.A. (2005) Interplay between network structure and self-organized criticality. cond-mat/0509043.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software-Practice & Experience*, *21*, 1129–1164.

Fry, B. (2000). *Organic information design*. Unpublished master's thesis, Massachusetts Institute of Technology.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.

Gibson, D., Kleinberg, J. M., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, 225–234.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, *30*, 1141–1144.

Gimblett, H. R. (Ed.). (2002). *Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes*. Oxford, UK: Oxford University Press.

Girvan, M., & Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 7821–7826.

Granovetter, M. (1973). Strength of weak ties. *American Journal of Sociology*, *78*, 1360–1380.

Gruhl, D., Guha, R., Tomkins, A., & Liben-Nowell, D. (2004). Information diffusion through blogspace. Paper presented at the *Thirteenth International World Wide Web Conference*. Retrieved March 10, 2006, from http://www2004.org/proceedings/docs/1p491.pdf

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2005). Structure and efficiency of the world-wide airport network. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 7794–7799.

Hackos, J. T., & Redish, J. C. (1998). *User and task analysis for interface design*. New York: Wiley.

Han, K., & Ju, B.-H. (2003). A fast layout algorithm for protein interaction networks. *Bioinformatics*, *19*, 1882–1888.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods: Centrality and power*. Retrieved September 19, 2005, from http://faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html

Hanson, N. (1958). *Patterns of discovery*. Cambridge, UK: Cambridge University Press.

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, *67*(337), 123–129.

Herman, I., Melancon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, *6*, 24–43.

Hodgman, T. C. (2000). A historical perspective on gene/protein functional assignment. *Bioinformatics*, *16*, 10–15.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, *76*(373), 33–50.

Huisman, M., & Van Duijn, M. A. J. (2005). Software for social network analysis. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), Models and methods in social network analysis (pp. 270–316). New York: Cambridge University Press.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Jungnickel, D. (1994). *Graphs, networks and algorithms*, Heidelberg, Germany: Springer.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*, 7–15.

Kauffman, S., Peterson, C., Samuelsson, B., & Troein, C. (2003). Random Boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 14796–14799.

Kim, J.-O., & Mueller, C. (1978). *Factor analysis: Statistical methods and practical issues*. Thousand Oaks, CA: Sage.

Kleinberg, J. (2000). The small-world phenomenon: An algorithmic perspective. *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 163-170.

Kleinberg, J., & Lawrence, S. (2001). The structure of the Web. *Science*, *294*, 1849–1850.

Krapivsky, P. L., & Redner, S. (2003). Rate equation approach for growing networks. In R. Pastor-Satorras, M. Rubi, & A. Diaz-Guilera (Eds.), *Statistical mechanics of complex networks* (Lecture Notes in Physics 625) (pp.3– 22). Berlin: Springer.

Krapivsky, P. L., & Redner, S. (2005). Network growth by copying. *Physical Review E*, *71*, 036118-1 – 036118- 8.

Krzywicki, A. (2001) Defining statistical ensembles of random graphs. cond-mat/0110574.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfall, E. (2000). Stochastic models for the Web graph. *Proceedings of the 41st IEEE Symposium on Foundations of Computer Sciences*, 57–65.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks*, *31*, 1481–1493.

Lakhina, A., Byers, J., Crovella, M., & Xie, P. (2002). *Sampling biases in IP topology measurements* (Technical Report BUCS-TR-2002-021). Boston: Boston University Computer Science Department.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 401–408.

Lee, E. J. (2005). Robustness of the avalanche dynamics in data packet transport on scale-free networks. Physical Review E, 71, 056108.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., & Aberg, Y. (2001). The web of sexual contacts. *Nature*, 411, 907–908.

Lynch, M. F. (1977). Variety generation: A Reinterpretation of Shannon's mathematical theory of communications, and its implication for information science. *Journal of the American Society for Information Science*, *28*, 19–25.

MacEachren, A. M. (1995). *How maps work: Representation, visualization and design*. New York: Guildford Press.

May, R. M., & Lloyd, A. L. (2001). Infection dynamics on scale-free networks. *Physical Review E*, *64*, 066112.

McFadzean, D., & Tesfatsion, L. (1997). An agent-based computational model for the evolution of trade networks. *Computing in Economics and Finance* 110. Retrieved March 16, 2006, from http://bucky.stanford.edu/cef97/abstracts/mcfadzean.html

Merton, R. K. (1968). The Matthew effect in science. The reward and communication systems of science are considered. *Science*, *159*(810), 56–63.

Milo, R., Shen-Orr, S. Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, *298*(5594), 824–827.

Molloy, M., & Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures Algorithms*, *6*, 161–174.
Monge, P. R., & Contractor N. (2003). *Theories of communication networks*. New York: Oxford University Press.
Moreno, J. L. (1934). *Who shall survive? A new approach to the problems of human interrelations*. Washington, DC: Nervous and Mental Disease Publishing Company.
Moreno, J. L. (1943). Sociometry and the social order. *Sociometry*, *6*, 299–344.
Moreno, Y., Pastor-Satorras, R.,Vazquez, A., & Vespignani, A. (2003). Critical load and congestion instabilities in scale-free networks. *Europhysics Letters*, *62*, 292.
Motter, A. E., & Lai, Y. C. (2002). Cascade based attacks on complex networks. *Physical Review E*, *66*, 065102.
Mueller, C. (2004). Sparse matrix reordering algorithms for cluster identification. Retrieved March 20, 2006, from http://www.osl.iu.edu/~chemuell/projects/bioinf/sparse-matrix-clustering-chris-mueller.pdf
Munzner, T. (1998). Exploring large graphs in 3D hyperbolic space. *IEEE Computer Graphics and Applications*, *18*(4), 18–23.
Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.
Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.
Newman, M. E. J. (2001). The structure of scientific network collaborations. *Proceedings of the National Academy of Sciences of the USA*, 98(2), 404-409.
Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*, 208701.
Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818.
Palmer, S. E. (1999). *Vision science: From photons to phenomenology*. Cambridge, MA: MIT Press.
Parisi, G. (1988). *Statistical field theory*. Redwood City, CA: Addison-Wesley.
Park, J., & Newman, M. E. J. (2004). The statistical mechanics of networks. *Physical Review E*, *70*, 066117.
Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, *86*, 3200–3203.
Pastor-Satorras, R., & Vespignani, A. (2002). Epidemic dynamics and endemic states in complex networks. *Physical Review E*, *63*, 036104.
Pastor-Satorras, R., & Vespignani, A. (2004). *Evolution and structure of the Internet: A statistical physics approach*. Cambridge, UK: Cambridge University Press.
Petermann, T., & De Los Rios, P. (2004). Exploration of scale-free networks. *European Physical Journal B*, *38*, 201–204.
Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.
Price, D. J. D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*, 292–306.
Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Association of Information Science and Technology*, *53*, 1113– 119.
Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*, 1551–1555.
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, *4*, 131–134.
Redner, S. (2005). Citation statistics from 110 years of *Physical Review*. *Physics Today*, *58*, 49–54.
Reichardt, J., & Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a

Potts model. *Physical Review Letters*, *93*(21), 218701.

Rettig, J. L. (1978). A theoretical model and definition of the reference process. *RQ*, *18,* 19–29.

Sanil, A., Banks, D., & Carley, K. (1995). Models for evolving fixed node networks: Model fitting and model testing. *Social Networks*, *17*, 65–81.

Schneeberger, A., Mercer, C. H., Gregson, S. A., Ferguson, N. M., Nyamukapa, A., Anderson, R. M., et al. (2004). Scale-free networks and sexually transmitted diseases: A description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sexually Transmitted Diseases*, *31*, 380–387.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood. NJ: Ablex.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423,623–656.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shen-Orr, S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, *31*, 64–68.

Shneiderman, B. (1987). Designing the User Interface: Strategies for Effective Human-Computer Interaction: Addison-Wesley Pub Co.

Shneiderman, B. (1992). Tree visualization with tree-maps: A 2-d space filling approach. *ACM Transactions on Graphics*, *11*, 92–99.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, 336–343.

Shneiderman, B. (2005). Treemaps for space-constrained visualization of hierarchies. Retrieved August, 2005, from http://www.cs.umd.edu/hcil/treemap-history/index.shtml

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425–450.

Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*, http://www.cmu.edu/joss/content/articles/volume3/Snijders.pdf.

Standish, R. K., & Galloway, J. (2002). *Visualising Tierra's tree of life using Netmap*. Retrieved March 10, 2006, from http://parallel.hpc.unsw.edu.au/rks/docs/netmap

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review*, *28*, 513–527.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, *85*, 204–212.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*, 268–276.

Tabah, A.N. (1999) Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science & Technology*, 34, 249 -286.

Thelwall, M. (2004). *Link analysis: An information science approach*. Amsterdam: Academic Press.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Vazquez, A., de Menezes, M. A., Oltvai, Z. N., & Barabási, A.-L. (2004). *Predicting the location and behavior of metabolic switches based on an optimal growth efficiency principle*. Technical Report, University of Notre Dame.

Venn, J. (1971). *Symbolic logic*. Bronx, NY: Chelsea Publishing Company.(Original work published 1894)

Walshaw, C. (2000). A multilevel algorithm for force-directed graph drawing. *Proceedings of the 8th International Symposium Graph Drawing*, 171–182.

Ware, C. (2000). *Information visualization: Perception for design*. San Francisco: Morgan Kaufmann.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks. *Psychosometrika*, *61*, 401–426.

Watts, D. J. (1999). *Small world*. Princeton, NJ: Princeton University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.

Wolf, Y. I., Karev, G., & Koonin, E. V. (2002). Scale-free networks in biology: New insights into the fundamentals of evolution? *Bioessays*, *24*, 105–109.

Wuchty, S., Oltvai, Z. N., & Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, *35*, 176–177.

Yee, K.-P., Fisher, D., Dhamija, R., & Hearst, M. (2001). Animated exploration of graphs with radial layout. *Proceedings of the IEEE Symposium on Information Visualization*, 43-50.

Zhou, S., & Mondragon, R. J. (2004). Accurately modeling the Internet topology. *Physical Review E*, *70*, 066108-1–066108-8.