# Graph Theory and Networks in Biology

Oliver Mason and Mark Verwoerd
Hamilton Institute, National University of Ireland
Maynooth, Co. Kildare, Ireland
{oliver.mason, mark.verwoerd}@nuim.ie

January 17, 2007

**Abstract**

In this paper, we present a survey of the use of graph theoretical techniques in Biology. In particular, we discuss recent work on identifying and modelling the structure of bio-molecular networks, as well as the application of centrality measures to interaction networks and research on the hierarchical structure of such networks and network motifs. Work on the link between structural network properties and dynamics is also described, with emphasis on synchronization and disease propagation.

## 1 Introduction and Motivation

The theory of complex networks plays an important role in a wide variety of disciplines, ranging from communications and power systems engineering to molecular and population biology [1–8]. While the focus of this article is on biological applications of the theory of graphs and networks, there are also several other domains in which networks play a crucial role. For instance, the Internet and the World Wide Web (WWW) have grown at a remarkable rate, both in size and importance, in recent years, leading to a pressing need both for systematic methods of analysing such networks as well as a thorough understanding of their properties. Moreover, in sociology and ecology, increasing amounts of data on food-webs and the structure of human social networks are becoming available. Given the critical role that these networks play in many key questions relating to the environment and public health, it is hardly surprising that researchers in ecology and epidemiology have focussed attention on network analysis in recent years. In particular, the complex interplay between the structure of social networks and the spread of disease is a topic of critical importance. The threats to human health posed by new infectious diseases such as the SARS virus and the Asian bird flu [9, 10], coupled with modern travel patterns, underline the vital nature of this issue.

Within the fields of Biology and Medicine, potential applications of network analysis include identifying drug targets, determining the role of proteins or genes of unknown function [11, 12], designing effective containment strategies for infectious diseases [13], and providing early diagnosis of neurological disorders through detecting abnormal patterns of neural synchronization in specific brain regions [14]. The development of high-throughput techniques in molecular biology have led to an unprecedented amount of data becoming available on cellular networks in a variety of simple organisms [15, 16]. Broadly speaking, three classes of such bio-molecular networks have attracted most attention to date: metabolic networks of biochemical reactions; protein interaction networks consisting of the physical interactions between an organism's proteins; and the transcriptional regulatory networks which

describe the regulatory interactions between different genes. At the time of writing, the central metabolic networks of numerous bacterial organisms have been mapped [17–19]. Also, large scale data sets are available on the structure of the protein interaction networks of *S. cerevisiae* [15, 20], *H. pylori* [21], *D. melanogaster* [22] and *C. elegans* [16, 23], and the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* have been extensively studied [24–26].

Thus, it is now possible to investigate the structural properties of networks in living cells, to identify their key properties and to hopefully shed light on how such properties may have evolved biologically. Given the special nature of biological systems, there is a pressing need for tailored analysis methods which can extract meaningful biological information from the data becoming available through the efforts of experimentalists. This is all the more pertinent given that the network structures emerging from the results of high-throughput techniques are too complex to analyse in a non-systematic fashion. A knowledge of the topologies of biological networks, and of their impact on biological processes, is needed if we are to fully understand, and develop more sophisticated treatment strategies for, complex diseases such as cancer [27]. Also, recent work suggesting connections between abnormal neural synchronization and neurological disorders such as *Parkinson's disease* and *Schizophrenia* [14] provides strong motivation for studying how network structure influences the emergence of synchronization between interconnected dynamical systems.

The mathematical discipline which underpins the study of complex networks in Biology and elsewhere is *graph theory* [28]. The complexity of the networks encountered in cellular biology and the mechanisms behind their emergence presents the network researcher with numerous challenges and difficulties. The inherent variability in biological data, the high likelihood of data inaccuracy [29] and the need to incorporate dynamics and network topology in the analysis of biological systems are just some of the obstacles to be overcome if we are to successfully understand the fundamental networks involved in the operation of living cells.

A substantial literature dedicated to the analysis of biological networks has emerged in the last few years, and some significant progress has been made on identifying and interpreting the structure of such networks. Our primary goal in the present article is to provide as broad a survey as possible of the major advances made in this field in the recent past, highlighting what has been achieved as well as some of the most significant open issues that need to be addressed. It is particularly hoped that the article will serve as a useful introduction to the field for those unfamiliar with the literature.

In the interests of clarity, we shall now give a brief outline of the main topics covered throughout the rest of the paper. In Section 2, we shall fix the principal notations used throughout the paper, and briefly review the main mathematical and graph theoretical concepts that are required in the remainder of the article. In Section 3, we shall discuss recent findings on the structure of bio-molecular networks and discuss several models, including *Scale-Free* graphs and *Duplication-Divergence* models, that have been proposed to account for the properties observed in real biological networks. Section 4 is concerned with the application of *measures of centrality or importance* to biological networks, and on the connection between the centrality of a gene or protein and its likelihood to be *essential* for an organism's survival. In Section 5, we shall discuss *motifs* and *functional modules* in bio-molecular networks.

In Sections 6 and 7, we shall discuss two major topics at the interface between dynamics and network theory. Section 6 is concerned with the phenomenon of synchronization in networks of inter-connected dynamical systems and its relevance in biological contexts. Particular attention will be given to suggested links between *patterns of synchrony* and *neurological disorders*. In Section 7, we shall discuss some

recent work on how the structure of a social network affects the behaviour of disease propagation models, and discuss the epidemiological significance of these findings. Finally, in Section 8 we shall present our concluding remarks and highlight some possible directions for future research.

# 2 Definitions and Mathematical Preliminaries

Throughout, $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ denote the field of real numbers, the vector space of $n$-tuples of real numbers and the space of $m \times n$ matrices with entries in $\mathbb{R}$ respectively. $\mathbf{A}^T$ denotes the transpose of a matrix $\mathbf{A}$ in $\mathbb{R}^{m\times n}$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be symmetric if $\mathbf{A} = \mathbf{A}^T$.

For finite sets $S, T$, $S \times T$ denotes the usual Cartesian product of $S$ and $T$, while $|S|$ denotes the cardinality of $S$.

## 2.1 Directed and Undirected Graphs

The concept of a *graph* is fundamental to the material to be discussed in this paper. The graphs or networks which we shall encounter can be divided into two broad classes: *directed graphs* and *undirected graphs*, as illustrated in Figure 1.


Figure 1: An example of a directed graph (left) and an undirected graph (right), comprising two nodes and one edge.

Formally, a finite *directed graph*, $G$, consists of a set of *vertices* or *nodes,* $\mathcal{V}(G)$,

$$\mathcal{V}(G) = \{v_1, \ldots, v_n\},$$

together with an *edge* set, $\mathcal{E}(G) \subseteq \mathcal{V}(G) \times \mathcal{V}(G)$. Intuitively, each edge $(u, v) \in \mathcal{E}(G)$ can be thought of as connecting the starting node $u$ to the terminal node $v$. For notational convenience, we shall often write $uv$ for the edge $(u, v)$. We shall say that the edge $uv$ *starts* at $u$ and *terminates* at $v$. For the most part, we shall be dealing with graphs with finitely many vertices and for this reason, we shall often omit the adjective finite where this is clear from context.

In Biology, transcriptional regulatory networks and metabolic networks would usually be modelled as directed graphs. For instance, in a transcriptional regulatory network, nodes represent genes with edges denoting the interactions between them. As each such interaction has a natural associated direction, such networks are modelled as directed graphs.

An *undirected graph*, $G$, also consists of a vertex set, $\mathcal{V}(G)$, and an edge set $\mathcal{E}(G)$. However, there is no direction associated with the edges in this case. Hence, the elements of $\mathcal{E}(G)$ are simply two-element subsets of $\mathcal{V}(G)$, rather than ordered pairs as above. As with directed graphs, we shall use the notation $uv$ (or $vu$ as direction is unimportant) to denote the edge $\{u, v\}$ in an undirected graph. For two vertices, $u, v$ of an undirected graph, $uv$ is an edge if and only if $vu$ is also an edge. We are not dealing with multi-graphs [28], so there can be at most one edge between any pair of vertices in an undirected graph. The number of vertices $n$ in a directed or undirected graph is the *size* or *order* of the graph.

In recent years, much attention has been focussed on the protein-protein interaction (PPI) networks of various simple organisms [15, 21]. These networks describe

the physical interactions between an organism's proteins and are typically modelled as undirected graphs, in which nodes represent proteins and edges represent interactions. We shall say more on PPI and other bio-molecular networks in the next section.

An edge, $uv$ in a directed or undirected graph $G$ is said to be *an edge at* the vertices $u$ and $v$, and the two vertices are said to be *adjacent* to each other. In this case, we also say that $u$ and $v$ are *neighbours*. For an undirected graph, $G$ and a vertex, $u \in \mathcal{V}(G)$, the set of all neighbours of $u$ is denoted $\mathcal{N}(u)$ and given by

$$\mathcal{N}(u) = \{v \in \mathcal{V}(G) : uv \in \mathcal{E}(G)\}.$$

## 2.2 Node-degree and the Adjacency Matrix

For an undirected graph $G$, we shall write $\deg(u)$ for the degree of a node $u$ in $\mathcal{V}(G)$. This is simply the total number of edges at $u$. For the graphs we shall consider, this is equal to the number of neighbours of $u$,

$$\deg(u) = |\mathcal{N}(u)|.$$

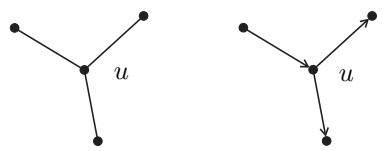
In a directed graph $G$, the *in-degree*, $\deg_{\text{in}}(u)$ (*out-degree* , $\deg_{\text{out}}(u)$) of a vertex $u$ is given by the number of edges that terminate (start) at $u$.

Suppose that the vertices of a graph (directed or undirected) $G$ are ordered as $v_1, \ldots, v_n$. Then the adjacency matrix, $\mathbf{A}$, of $G$ is given by

$$a_{ij} = \begin{cases} 1 & \text{if } v_i v_j \in \mathcal{E}(G) \\ 0 & \text{if } v_i v_j \notin \mathcal{E}(G) \end{cases} \tag{1}$$

Thus, the adjacency matrix of an undirected graph is symmetric while this need not be the case for a directed graph. Figure 2 illustrates this.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$




$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2: The adjacency matrix of an undirected graph is symmetric; that of a directed graph generally is not. In this example, we have that $\deg(u) = 3$ for the undirected graph and $\deg_{\text{in}}(u) = 1$, $\deg_{\text{out}}(u) = 2$ for the directed graph.

## 2.3 Paths, Path Length and Diameter

Let $u, v$ be two vertices in a graph $G$. Then a sequence of vertices

$$u = v_1, v_2, \ldots, v_k = v,$$

such that for $i = 1, \ldots, k-1$:

(i) $v_i v_{i+1} \in \mathcal{E}(G)$;

(ii) $v_i \neq v_j$ for $i \neq j$

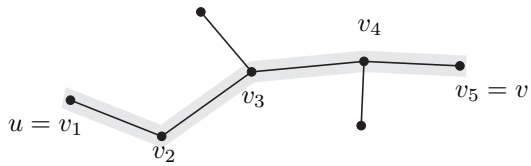



Figure 3: A path of length 4.

is said to be a path of length $k-1$ from $u$ to $v$. Figure 3 contains an example of a path of length 4.

If the requirement (ii) that the vertices are distinct is removed, then we obtain the definition of a *walk*. A path or walk with in which the initial vertex $v_1$ and final vertex $v_k$ are identical is said to be *closed*.

The *geodesic distance*, or simply distance, $\delta(u, v)$, from $u$ to $v$ is the length of the shortest path from $u$ to $v$ in $G$. If no such path exists, then we set $\delta(u, v) = \infty$. If for every pair of vertices, $u, v \in \mathcal{V}(G)$, there is some path from $u$ to $v$, then we say that $G$ is *connected*. The average path length and diameter of a graph $G$ are defined to be the average and maximum value of $\delta(u, v)$ taken over all pairs of distinct nodes, $u, v$ in $\mathcal{V}(G)$ which are connected by at least one path.

## 2.4 Clustering Coefficient

Suppose $u$ is a node of degree $k$ in an undirected graph $G$ and that there are $e$ edges between the $k$ neighbours of $u$ in $G$. Then the clustering coefficient of $u$ in $G$ is given by [30]

$$C_u = \frac{2e}{k(k-1)}. \tag{2}$$

Thus, $C_u$ measures the ratio of the number of edges between the neighbours of $u$ to the total possible number of such edges, which is $k(k-1)/2$. The average clustering coefficient of a graph $G$ is defined in the obvious manner.

## 2.5 Weighted Networks and Graphs

While most of the work described in this paper is concerned with graphs and networks for which an edge is either present or not, in many practical applications it is also of interest to quantify the "strength" of an interaction when it is present. The natural way to model such interactions is with a *weighted graph*. Here, a weight $w_{ij}$ is assigned to each pair of vertices $v_i, v_j$ in the network. The unweighted graphs that we have defined are a special case of this definition in which $w_{ij}$ takes only the values 0 and 1.

Weighted graphs can arise in a number of situations in Biology. For instance, if the strengths of the interactions in a PPI or genetic regulator network are taken into account or if we incorporate the speed of the reactions in a metabolic network, then weighted networks are the most suitable modelling paradigm.

Recently, several authors have begun to study weighted complex networks and various generalisations of the concepts discussed above have been introduced. For instance, weighted versions of the clustering coefficient have been defined and studied in [31–33]. The concept of *strength* has also been defined [31] to generalise notions of centrality (see Section 4) and the degree distribution (see Section 3) to

weighted networks. Furthermore, models for the evolution of weighted networks have been proposed in the spirit of those described for the unweighted case in Section 3 [34] and a more general definition of *network motifs* (Section 5) has been proposed in [32].

### 2.6 Statistical Notations

Throughout the paper, we shall often be interested in average values of various quantities where the average is taken over all of the nodes in a given network of graph. For some quantity, $f$, associated with a vertex, $v$, the notation $\langle f \rangle$ denotes the average value of $f$ over all nodes in the graph.

## 3 Identification and Modelling of Bio-molecular Networks

Broadly speaking, three classes of bio-molecular networks have been studied in depth: protein interaction networks, metabolic networks and transcriptional regulatory networks.

*Protein interaction networks* are typically modelled as undirected graphs, in which nodes correspond to proteins and an edge represents a physical interaction between two proteins. Large-scale maps of protein interaction networks [21–23, 29, 35] have been constructed recently using high-throughput approaches such as yeast-2-hybrid screens [15] or mass spectrometry techniques [36] to identify protein interactions. Data on protein interactions are also stored in databases such as the database of interacting proteins (DIP) [37]. While the accuracy of these techniques is known to be questionable (and both techniques are prone to detect interactions which may never take place in the cell), more reliable networks can be constructed by combining data from different sources and applying multiple criteria for the identification of interactions [38]. For example, data from the results of different high-throughput protein interaction experiments can be combined to obtain more reliable results [29, 39]. Other types of data such as co-expression in microarray experiments can be combined with protein interaction data-sets to add confidence to interactions predicted by high-throughput methods [40, 41]. In the recent past, there has also been considerable interest in the development of computational approaches for the prediction of protein interactions [42].

*Metabolic networks* describe the intricate web of bio-chemical reactions within a cell through which substrates are transformed into products through reactions catalysed by enzymes. As with protein interaction networks, genome-scale metabolic networks have been constructed for a variety of simple organisms including *S. cerevisiae* and *E. coli* [17, 43–47], and are stored in databases such as the KEGG [17] or BioCyc [18] databases. A common approach to the construction of such networks is to first use the annotated genome of an organism to identify the enzymes in the network and then to combine bio-chemical and genetic information to obtain their associated reactions [46, 48, 49]. While efforts have been made to automate certain aspects of this process, there is still a need to validate the networks generated automatically manually against experimental biochemical results [50, 51]. For metabolic networks, significant advances have also been made in modelling the reactions that take place on such networks. One of the most widely used techniques in this area is *Flux Balance Analysis* [48, 52] which combines mass balance and other regulation constraints with optimality conditions to predict steady-state reaction fluxes in metabolic networks. Such predictions can then be experimentally verified and used to refine and improve the underlying network model. We shall have more to say on FBA and metabolic networks later in the article.

*Transcriptional regulatory networks* describe the regulatory interactions between genes. Here, nodes correspond to individual genes and a directed edge is drawn from gene A to gene B if A (or its product) regulates the activity of gene B. Network maps have been constructed for the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* [24, 53–55] and are maintained in databases such as RegulonDB [24] and EcoCyc [55]. Such maps are usually constructed through a combination of high-throughput genome location experiments and literature searches [53]. While much of our discussion shall focus on these three classes of networks, it should be noted that many other variations are possible and have been studied, including epistatic networks [56] describing the interactions through which different genetic mutations either aggravate or buffer each other's effects on an organism, protein domain networks [57] and various integrated networks of genetic, protein and metabolic interactions [58].

## 3.1 Structural Properties of Biological Networks

The most widely studied topological features of bio-molecular networks are their *degree distributions, characteristic path lengths and clustering properties.*

### 3.1.1 Degree Distributions

The degree distribution of a network, $P(k), k = 0, 1, \ldots$, measures the proportion of nodes in the network having degree $k$. Formally,

$$P(k) = \frac{n_k}{n},$$

where $n_k$ is the number of nodes in the network of degree $k$ and $n$ is the size of the network. It was reported in [59, 60] that the degree distributions of the Internet and the WWW are described by a broad-tailed power law of the form[1],

$$P(k) \sim k^{-\gamma}, \qquad \gamma > 1 \tag{3}$$

Networks with degree distributions of this form are now commonly referred to as *scale-free networks.* This finding initially surprised the authors of these papers as they had expected to find that the degree distributions were Poisson or Gaussian. In particular, they has expected that the degrees of most nodes would be close to the mean degree, $\langle k \rangle$, of the network, and that $P(k)$ would decay exponentially as $|k - \langle k \rangle|$ increased. For such networks, the mean degree can be thought of as typical for the overall network. On the other hand, the node-degrees in networks with broad-tailed distributions vary substantially from their mean value, and $\langle k \rangle$ cannot be thought of as a typical value for the network in this case.

Following on from the above findings on the WWW and the Internet, several authors have investigated the form of the degree distributions, $P(k)$, for various biological networks. Recently, several papers have been published that claim that interaction networks in a variety of organisms are also scale-free. For instance, in [43], the degree distributions of the central metabolic networks of 43 different organisms were investigated using data from the WIT database [44]. The results of this paper indicate that, for all 43 networks studied, the distributions of in-degree, $P_{in}(k)$, and out-degree, $P_{out}(k)$, have tails of the form (3), with $2 < \gamma < 3$.

Similar studies on the degree distributions of protein interaction networks in various organisms have also been carried out. In [61], the protein interaction network of *S. cerevisiae* was analysed using data from four different sources. As is

[1] In fact, the form $P(k) \sim (k + k_0)^{-\gamma} e^{-k/k_c}$ with offset $k_0$ and an exponential cutoff $k_c$ is more usually fitted to real network data.

often the case with data of this nature, there was little overlap between the interactions identified in the different sets of data. However, in all four cases, the degree distribution appeared to be broad-tailed and to be best described by some form of modified power law. Similar findings have also been reported for the protein interaction networks of *E. coli, D. melanogaster, C. elegans* and *H. pylori* in the recent paper [62]. Note however that for transcriptional regulatory networks, while the outgoing degree distribution again appears to follow a power law, the incoming degree distribution is better approximated by an exponential rule of the form $P_{in}(k) \sim e^{-\beta k}$ [2, 63, 64].

At this point, it is important to record some remarks on the observations of scale-free topologies in biological interaction networks. First of all, the broad-tailed degree distributions observed in these networks is not consistent with the traditional random graph models which have been used to describe complex networks [1, 65]. In these models, node-degrees are closely clustered around the mean degree, $\langle k \rangle$, and the probability of a node having degree $k$ decreases exponentially with $|k - \langle k \rangle|$. However, in scale-free networks, while most nodes have relatively low degree, there are significant numbers of nodes with unusually high degree - far higher than the mean degree of the network. Such nodes are now usually referred to as *hubs.* It has been noted [66] that the scale-free structure has implications for the robustness and vulnerability of networks to failure and attack. Specifically, while removing most of the nodes in a scale-free network will have little effect on the network's connectivity, the targeted removal of hub nodes can disconnect the network relatively easily. This has led to the suggestion that genes or proteins which are involved in a large number of interactions, corresponding to hub nodes, may be more important for an organism's survival than those of low degree. The connection between network topology and the biological importance of genes and proteins has been extensively studied recently and we shall describe this strand of research in detail in Section 4.

A second important point is that all of the analysis described above has been carried out on *sampled subnetworks* rather than on a complete network. Thus, the protein interaction networks which have been studied usually do not contain the complete set of proteins of an organism[2] and the interactions included in these networks are far from complete. Thus, the conclusions being drawn are based on a subnetwork containing only a sample of the nodes and edges of the complete network. While some studies have indicated that the statistical properties of interaction networks may be robust with respect to variations from one data set to another, the impact of sampling and inaccurate/incomplete information on the identified degree distributions is an important issue which is not yet fully understood. For instance, in [67] it was shown using a model of protein interaction networks that an approximate power law distribution can be observed in a sampled sub-network while the degree distribution of the overall network is quite different. Further evidence of the need for caution in drawing conclusions about the overall structure of biological networks based on samples has been provided in [68, 69], where results on the sampling properties of various types of network models were presented. For instance, in [68], a sampling regime based on the construction of *spanning trees* [28] was studied. Here, starting from a source vertex $v_0$, a tree $T$ is constructed by first adding the neighbours of $v_0$ to $T$ and then selecting one of these, and repeating the process. In this paper, approximate arguments were presented to show that such a sampling regime can lead to a subnetwork with degree distribution of the form $P(k) \sim 1/k$ even when the complete network has a Poisson degree distribution.

[2]although the network for *S. cerevisiae* by now covers the organism's proteome fairly comprehensievely

### 3.1.2 Diameter and Characteristic Path Length

Several recent studies have revealed that the average path lengths and diameters of bio-molecular networks are "small" in comparison to network size. Specifically, if the size of a network is $n$, the average path length and diameter are of the same order of magnitude as $\log(n)$ or even smaller. This property has been previously noted for a variety of other technological and social networks [1], and is often referred to as the *small world property* [30]. This phenomenon has now been observed in metabolic, genetic and protein interaction networks. For instance, in [43, 70], the average path lengths of metabolic networks were studied. The networks analysed in these papers had average path lengths between 3 and 5 while the network sizes varied from 200-500. Similar findings have been reported for genetic networks in [71], where a network of approximately 1000 genes and 4000 interactions was found to have a characteristic path length of 3.3, and for protein interaction networks in [61, 72, 73].

In a sense, the average path length in a network is an indicator of how readily "information" can be transmitted through it. Thus, the small world property observed in biological networks suggests that such networks are efficient in the transfer of biological information: only a small number of intermediate reactions are necessary for any one protein/gene/metabolite to influence the characteristics or behaviour of another.

### 3.1.3 Clustering and Modularity

The final aspect of network structure which we shall discuss here is concerned with how densely clustered the edges in a network are. In a highly clustered network, the neighbours of a given node are very likely to be themselves linked by an edge. Typically, the first step in studying the clustering and modular properties of a network is to calculate its average clustering coefficient, $C$, and the related function, $C(k)$, which gives the average clustering coefficient of nodes of degree $k$ in the network. As we shall see below, the form of this function can give insights into the global network structure.

In [19], the average clustering coefficient was calculated for the metabolic networks of 43 organisms and, in each case, compared to the clustering coefficient of a random network with the same underlying degree distribution. In fact, the comparison was with the Barabasi-Albert (BA) model of scale-free networks which we shall discuss in the next subsection. In each case, the clustering coefficient of the metabolic network was at least an order of magnitude higher than that of the corresponding BA network. Moreover, the function $C(k)$ appeared to take the form $C(k) \sim k^{-1}$. Thus, as the degree of a node increases, its clustering coefficient decreases. This suggests that the neighborhoods of low-degree nodes are densely clustered while those of hub nodes are quite sparsely connected. In order to account for this, the authors of [19] suggested a hierarchical modular structure for metabolic networks in which individual modules are comprised of densely clustered nodes of relatively low degree while different modules are linked by hub nodes of high degree.

Similar results for the clustering coefficient and the form of the function $C(k)$ have been reported in [62] for the protein interaction networks of *S. cerevisiae, H. pylori, E. coli* and *C. elegans*, indicating that these undirected networks may also have a modular structure, in which hub nodes act as links or bridges between different modules within the networks. Further evidence for the intermediary role of hub nodes was provided in [74] wherein correlations between the degrees of neighbouring nodes in the protein interaction network and the transcriptional regulatory network of *S. cerevisiae* were investigated. The authors of this paper found clear evidence of such correlation; in fact, for both networks, nodes of high degree are

significantly more likely to connect to nodes of low degree than to other "hubs". This property of a network is referred to as *disassortativity* (for more discussion on degree correlations in biological networks, see [75]). In [76], the relation $C(k) \sim k^{-1}$ was explained as the effect of degree correlations on the calculation of the standard clustering coefficient. Moreover, the usual definition was shown to under-estimate the degree of clustering in networks with such degree correlations, particularly for disassortative networks. A novel measure of clustering was proposed which was not effected by such correlations and did not display the same decay with increasing degree in protein interaction networks.

Finally, note that a high level of local clustering was also observed in the network of synthetic lethal genetic interactions in *S. cerevisiae* studied in [71]. In such a network, an edge is drawn between two genes when a double knock-out experiment, removing both genes, has fatal consequences for the organism. The same study suggested a promising technique for using such a network to predict protein-protein interactions between gene products.

## 3.2 Mathematical Models for Interaction Networks

Given the empirically observed properties of interaction networks discussed above, it is natural to ask whether these can be explained by means of mathematical models based on plausible biological assumptions. Reliable models for the evolution of interaction networks may deepen our understanding of the biological processes behind their evolution. Moreover, such models could be used to assess the reliability of experimental results on network structure and to assist in experimental design. For instance, the strategy for optimally identifying protein-protein interaction (PPI) network structure described in [77] relies on the statistical abundance of nodes of high degree in scale-free networks which we shall discuss in more detail below. Furthermore, this strategy was suggested as a means of determining the PPI network in humans. Note also the work described in [78] on assessing the reliability of network data and predicting the existence of links in a PPI network which have not yet been determined. The methods in this paper were based on properties of the *small-world* network model of Watts and Strogatz introduced in [30] to described social and neurological networks.

To date, several different mathematical models of complex networks have been proposed in the literature. A number of these were not developed with specifically biological networks in mind, but rather to account for some of the topological features observed in real networks in Biology and elsewhere. On the other hand, in the recent past several models for protein interaction and genetic networks have been proposed based on biological assumptions.

### 3.2.1 Classical Models and Scale-free Graphs

In the 1950's, Paul Erdös and Alfréd Rényi introduced their now classical notion of a *random graph* to model non-regular complex networks. The basic idea of the Erdös-Rényi (ER) random graph model is the following. Let a set of $n$ nodes, $\{v_1, \ldots, v_n\}$, and a real number $p$ with $0 \leq p \leq 1$ be given. Then for each pair of nodes, $v_i, v_j$, an edge is placed between $v_i$ and $v_j$ with probability $p$. Effectively, this defines a probability space where the individual elements are particular graphs on $\{v_1, \ldots, v_n\}$ and the probability of a given graph with $m$ edges occurring is $p^m(1-p)^{n-m}$. For background on the mathematical theory of ER graphs, consult [28, 65].

The theory of random graphs has been a highly active field of mathematics for fifty years and many deep theorems about the properties of ER graphs have been established [65]. In particular, the connectivity, degree distribution and average path

length have been extensively studied for this class of networks. For mathematically rigorous results on these topics, consult [65]. One of the most fundamental facts concerning ER graphs is that the degree distribution of a large ER network can be approximated by a Poisson distribution. This means that for ER graphs, the node degrees tend to be tightly clustered around the mean degree $\langle k \rangle$ of the network.

This contrasts with the findings reported in the previous subsection that the degree distributions of many biological networks appear to follow a broad-tailed power law. The same observation has also been made for several man-made networks including the WWW and the Internet. This led Barabasi, Albert and co-workers to devise a new model for the dynamics of network evolution. This model is based on the two fundamental mechanisms of *growth* and *preferential attachment*, and has been the subject of intensive research in the last few years. It is usually referred to as the Barabasi-Albert (BA) model.

The core idea of Barabasi and Albert was to consider a network as an evolving entity and to model the dynamics of network growth. The simple BA model is now well known and is usually described in the following manner [1]. Given a positive integer, $m$, and an initial network, $G_0$, the network evolves according to the following rules (note that this is a discrete-time process):

(i) *Growth:* At each time $j$, a new node of degree $m$ is added to the network;

(ii) *Preferential Attachment:* For each node $u$ in the existing network, the probability that the new node connects to it is proportional to the degree of $u$. Formally, writing $G_j$ for the network at time $j$ and $P(u, j)$ for the probability that the new node added at time $k$ is linked to $u$ in $G_{j-1}$:

$$P(u, j) = \frac{\deg(u)}{\sum_{v \in \mathcal{V}(G_{j-1})} \deg(v)}. \tag{4}$$

The above scheme generates a network whose degree distribution asymptotically approaches the power law $P(k) \sim k^{-\gamma}$ with $\gamma = 3$ [1]. A number of variations on the basic BA model have also been proposed that have power law degree distributions with values of the degree exponent other than three. See for instance, the models for evolving networks described in [79, 80] which give rise to power law degree distributions with exponents in the range $2 < \gamma < +\infty$.

### 3.2.2 Some Issues in the Use of Scale-free Models

While the degree distributions of BA and related scale-free models appear to fit the experimental data on bio-molecular networks more accurately than classical ER networks, there are several issues related to their use that should be noted. In [81], it was pointed out that the commonly used definition of BA graphs is ambiguous. For instance, the question of how to initiate the process of network evolution is not explicitly dealt with in the original papers; how do we connect the new node to the existing nodes with probability proportional to their degrees if all such nodes have degree zero to begin with? This issue can be circumvented by beginning with a network which has no isolated nodes. However, this immediately raises the difficult question of how the choice of initial network influences the properties of the growing network. These issues have been discussed in detail in [81, 82] where more mathematically rigorous formulations of the preferential attachment mechanism for network growth have been presented. A number of formal results concerning degree distributions, network diameter, robustness to node removal and other network properties have also been presented in [83, 84].

It is important to note that a number of other, more practical reservations about the use of the BA and related models in Biology have been raised recently

(See [85] for a general discussion concerning this issue). Firstly, the BA model is not based on specific biological considerations. Rather, it is a mathematical model for the dynamical growth of networks that replicates the degree distributions, and some other properties, observed in studies of the WWW and other networks. In particular, it should be kept in mind that the degree distribution is just one property of a network and that networks with the same degree distribution can differ substantially in other important structural aspects [86].

Many of the results on BA and related networks have only been empirically established through simulation, and a fully rigorous understanding of their properties is still lacking. A number of authors have started to address this issue in the recent past but this work is still in an early stage. Also, as noted above, the definition of BA graphs frequently given in the literature is ambiguous [81].

Another very important point is that the observations of scale-free and power law behaviour in biological networks are based on partial and inaccurate data. The techniques used to identify protein interactions and transcriptional regulation are prone to high levels of false positive and false negative errors [29]. Moreover, the networks being studied typically only contain a sample of the nodes and edges in the complete network for an organism. Thus, we are in effect drawing conclusions about the topology of an entire network based on a *sample* of its nodes, and a noisy sample at that. In order to do this reliably, a thorough understanding of the effect of sampling on network statistics, such as distributions of node degrees and clustering coefficients, is required. Some authors have recently started to address this issue and two highly relevant theoretical results have recently been presented in [68, 69]. It was shown in [69] that subnetworks sampled from a scale free network are not in general scale free, while it is also possible for a sampled subnetwork of a network with Poisson degree distribution (which is certainly not scale-free) to appear to be scale-free [68]. Note that the sampling regimes considered in these two papers were not identical.

Further results of this kind appeared in [87]. In this paper, the large-scale yeast-2-hybrid (Y2H) experiments which have generated many of the existing protein-interaction maps for yeast (and other organisms) were simulated on four network models with degree distributions given by normal, exponential, scale-free and truncated normal distributions respectively. In each case, a percentage of nodes in the network was first selected as *baits* and then a percentage of the interacting partners of these baits was randomly selected. This was done for percentages of bait and edge coverage ranging from 1% to 100%. The degree distributions of the resulting sampled networks were then tested for how closely they resembled a power law. It was found that for low bait and edge coverage, the subnetworks obtained in this way closely resemble scale-free networks irrespective of the original network topology (from the point of view of degree distributions). Hence, the authors argue that none of the four models considered could be definitively ruled out as a model of the complete interaction network. However, it should be noted that, for ER networks the resemblance to scale-free networks only held for levels of bait-edge coverage which are certainly lower than the coverage in current data-sets on the yeast interaction network. More importantly, the analysis in this paper was based on a very small number of sampled networks and more extensive studies are required to make definitive statements. Also, this and other studies of sampled networks address the question of the probability of a sampled network being scale-free given the topology of the entire network. Another important question that needs to be addressed is given the sampled network's topology what is the probability that the entire network is scale-free or otherwise. Nonetheless, this and the other studies illustrate the need for caution about the effects of sampling and data noise when we attempt to draw conclusions about the structure of biological and other real world networks.

The numerical study presented in [87] suggested that subnetworks of scale-free networks appear to be scale-free for a large range of bait-edge coverage levels. This might appear to contradict the theoretical findings in [69] that subnetworks sampled from scale-free networks are not in general scale-free. However, the result in [69] was concerned with the precise functional form of the expected degree distribution for the limiting case of an infinite network, whereas [87] considered empirical measures of goodness-of-fit for specific finite-size networks. Also, the sampling schemes in the two papers are fundamentally different. Similar remarks apply to the results of [68]. While the theoretical results described in [69] are significant and interesting, for practical applications, an understanding of the sampling properties of finite-size networks is required. Moreover, even in the limiting case, a rigorous mathematical theory of the sampling properties of scale-free and other network models, in the spirit of [82] has yet to be developed.

Before moving on to discuss a number of more biologically motivated models for interaction networks, we note the recent paper [88] in which *geometric random graphs* [89] were suggested as an alternative model for protein interaction networks. This suggestion was based on comparing the frequency of small subgraphs in real networks to their frequency in various network models, including geometric graphs. However, as with BA models, there is no clear biological motivation for choosing geometric graphs to model protein interaction networks and, furthermore, the comparisons presented in [88] are based on a very small number of sample random networks. On the other hand, the authors of this paper make the important point that the accuracy of network models is crucial if we are to use these to assess the reliability of experimental data or in the design of experiments for determining network structure.

### 3.2.3 Duplication and Divergence Models

Many of the recent models for network evolution are founded on some variation of the basic mechanisms of growth and preferential attachment. However, there are other, more biologically motivated models which have been developed specifically for protein interaction and genetic regulatory networks. As with the models discussed above, these are usually based on two fundamental processes: *duplication* and *divergence*. The hypothesis underpinning these so-called Duplication-Divergence (DD) models is that gene and protein networks evolve through the occasional copying of individual genes/proteins, followed by subsequent mutations. Over a long period of time, these processes combine to produce networks consisting of genes and proteins, some of which, while distinct, will have closely related properties due to common ancestry.

To illustrate the main idea behind DD models, we shall give a brief description of the model for protein interaction networks suggested in [90]. Given some initial network $G_0$, the network is updated at each time $t$ according to the rules:

(i) *Duplication:* a node $v$ is chosen from the network $G_{t-1}$ at random and a new node $v'$, a duplicate of $v$, is added to the network and connected to all of the neighbours of $v$;

(ii) *Divergence:* for each neighbour, $w$, of $v'$, the edge $v'w$ is removed with probability $q$.

As pointed out in [90], the above scheme effectively introduces a preferential attachment mechanism into the network and generates a power law degree distribution. A number of basic properties of the model and its suitability to model the PPI network of *S. cerevisae* are discussed in this paper also. The same basic model has also been studied more analytically in [91]. In this paper, it was shown that

if $q < 1/2$, then the degree distribution of the DD network is given by a power law whose exponent $\gamma$ satisfies $\gamma < 2$. The authors of this paper also considered some closely related models for the growth of gene networks in the earlier paper [92]. Here it was pointed out that duplication alone will not give rise to a power law degree distribution.

The model described in [90] allowed for self-interacting proteins, where the copy $v'$ can also form a link to the original $v$ with some non-zero probability. However, there are several assumptions associated with the basic scheme described above whose biological validity is questionable.

(i) The new node, $v'$, can only form links to neighbours of the original node $v$ — this restricts the types of mutations allowed for duplicate genes;

(ii) A node can only undergo mutation or divergence at the instant when it is added to the network — this ignores the possibility of genes continuing to mutate long after the duplication event;

(iii) Nodes and edges can only be added to the network and not removed — this clearly places a significant restriction on the types of mutation and evolution possible.

Several extensions of the basic DD model have been proposed to relax some of the assumptions outlined above. For instance, point (i) above has been addressed in [93], while a model that allows for edge additions and removals at a much faster rate than gene duplications has been described and analysed in [94]. Yet another growth model (based on a preferential attachment mechanism) which allows edges to be added and deleted between nodes in the existing network, and for new nodes to be added to the network has been presented in [95]. Finally, the issue in point (iii) has been addressed in the recent paper [96] by a growth-deletion model that allows for the addition and removal of both edges and nodes. As with BA networks, the theory of DD networks is still in a very early stage of development and relatively few mathematically rigorous results have been derived.

## 3.3 Summarizing Comments

Maps of bio-molecular networks are now available for simple organisms, and preliminary results on the structural properties of their protein-protein interaction, transcriptional regulatory and metabolic networks have been reported. It has been widely claimed that networks have scale-free degree distributions, short characteristic path lengths and high clustering coefficients. Traditional random graph models for complex networks are certainly inadequate for describing such networks. Several new mathematical models for the growth of random networks have been proposed in the recent past. These include scale-free models, and the more biologically inspired *Duplication-Divergence* models for gene and protein networks. The mathematical theory of these models is only beginning to be developed and offers many exciting and challenging opportunities for future biologically motivated mathematical research. However, the available data on bio-molecular networks is still far from reliable and does not cover the entire collection of interactions or nodes in such networks. Recent results on the sampling properties of networks with power law and Poisson degree distributions highlight the need for caution when drawing conclusions on global network properties from an analysis of a sampled subnetwork.

# 4 Measures of Centrality and Importance in Biological Networks

The problem of identifying the most important nodes in a large complex network is of fundamental importance in a number of application areas, including Communications, Sociology and Management. To date, several measures have been devised for ranking the nodes in a complex network and quantifying their relative importance. Many of these originated in the Sociology and Operations Research literature, where they are commonly known as *centrality measures* [97]. More recently, schemes such as the PageRank algorithm on which GOOGLE is based, have been developed for identifying the most relevant web-pages to a specific user query.

Recently, several researchers have applied centrality measures to identify structurally important genes or proteins in interaction networks and investigated the biological significance of the genes or proteins identified in this way. Particular attention has been given to the relationship between centrality and essentiality, where a gene or protein is said to be essential for an organism if the organism cannot survive without it. The use of centrality measures to predict essentiality based on network topology has potentially significant applications to drug target identification [11, 27].

## 4.1 Classical Centrality Measures

First of all, we shall discuss four classical notions of network centrality and their use in biological networks: *Degree centrality*; *Closeness centrality*; *Betweenness centrality*; *Eigenvector centrality*.

### 4.1.1 Degree Centrality

The idea behind using degree centrality as a measure of importance in a network is that:

*An important node is involved in a large number of interactions.*

Formally, for an undirected graph $G$, the degree centrality of a node $u \in \mathcal{V}(G)$ is given by

$$C_d(u) = \deg(u). \tag{5}$$

For directed networks, there are two notions of degree centrality: one based on in-degree and the other on out-degree. These are defined in the obvious manner. Degree centrality and the other measures discussed here are often normalised to lie in the interval $[0, 1]$.

As discussed in the previous section, a number of recent studies have indicated that bio-molecular networks have broad-tailed degree distributions, meaning that while most nodes in such networks have a relatively low degree, there are significant numbers of so-called hub nodes. The removal of these hub nodes has a far greater impact on the topology and connectedness of the network than the removal of nodes of low degree [66]. This naturally leads to the hypothesis that hub nodes in protein interaction networks and genetic regulatory networks may represent essential genes and proteins. In [98], the connection between degree centrality and essentiality was investigated for the protein-protein interaction network in *S. cerevisiae*. The analysis was carried out on a network consisting of 1870 nodes connected by 2240 edges, which was constructed by combining the results of earlier research presented in [20, 37]. In this network, 21% of those proteins that are involved in fewer than 5

interactions, $C_d(u) \leq 5$, were essential while, in contrast, 62% of proteins involved in more than 15 interactions, $C_d(v) \geq 15$, were essential.

More recently, similar findings were reported in [73]. Again, the authors considered a network of protein interactions in yeast, this time consisting of 23294 interactions between 4743 proteins. The average degree of an essential protein in this network was 18.7, while the average degree of a non-essential protein was only 7.4. Moreover, defining a hub to be a node in the first quartile of nodes ranked according to degree, the authors of [73] found that over 40% of hubs were essential while only 20% of all nodes in the network are essential.

The above observations have led some authors to propose that, in protein interaction networks, node degree and essentiality may be related [73, 98]. However, the precise nature of this relationship is far from straightforward. For instance, using a network constructed from data published in [15, 20], the author of [99] has claimed that there is little difference between the distributions of node degrees for essential and non-essential proteins in the interaction network of yeast. However, in this network, the degrees of essential proteins are still typically higher than those of non-essential proteins.

In [100] the connection between the degree of a protein and the rate at which it evolves was investigated. The authors reasoned that if highly connected proteins are more important to an organism's survival, they should be subject to more stringent evolutionary constraints and should evolve at a slower rate than non-essential proteins. However, the authors of [100] found no evidence of a significant correlation between the number of interactions in which a protein is involved and its evolutionary rate. Once again, this indicates that while node degree gives some indication of a protein's likelihood to be essential, the precise relationship between essentiality and node degree is not a simple one.

#### 4.1.2 Closeness Centrality Measures

The basic idea behind closeness centrality measures is that:

> *An important node is typically "close" to, and can communicate quickly with, the other nodes in the network.*

In [101], three closeness measures, which arise in the context of resource allocation problems, were applied to metabolic and protein interaction networks. The specific measures considered in this paper were *excentricity, status, and centroid value.*

The excentricity, $C_e(u)$, of a node $u$ in a graph $G$ is given by

$$C_e(u) = \max_{v \in \mathcal{V}(G)} \delta(u, v), \tag{6}$$

and the *centre* of $G$ is then the set

$$\mathcal{C}(G) = \{v \in \mathcal{V}(G) : C_e(v) = \min_{w \in \mathcal{V}(G)} C_e(w)\}. \tag{7}$$

Thus, the nodes in $\mathcal{C}(G)$ are those that minimise the maximum distance to any other node of $G$.

The *status*, $C_s(u)$, of a node $v$ is given by

$$C_s(u) = \sum_{v \in \mathcal{V}(G)} \delta(u, v), \tag{8}$$

and the *median* of $G$ is then the set

$$\mathcal{M}(G) = \{v \in \mathcal{V}(G) : C_s(v) = \min_{w \in \mathcal{V}(G)} C_s(w)\}. \tag{9}$$

The nodes in $\mathcal{M}(G)$ minimise the *average* distance to other nodes in the network.

The final measure considered in [101] is the *centroid* value which is closely related to the status defined above. In fact, these two measures give rise to identical rankings of the nodes in a graph and, for this reason, we shall not formally define centroid value here.

A number of points about the results presented in [101] are worth noting. First of all, on both ER graphs and the BA model of scale-free graphs, all three measures were found to be strongly correlated with node-degree. The measures were then applied to the central metabolic network of *E. coli* and the centre, $\mathcal{C}(G)$, and the median, $\mathcal{M}(G)$, of this network were calculated. The authors reasoned that central nodes represent "cross-roads" or "bottlenecks" in a network and should correspond to key elements of the organism's metabolism. In support of this assertion, the centre, $\mathcal{C}(G)$, contained several of the most important known substrates, including ATP, ADP, AMP and NADP. On the other hand, in the protein interaction network of *S. cerevisiae*, no discernible difference between the excentricity distribution of essential and non-essential proteins was observed. In the same paper, centrality measures were also applied to networks of protein domains where two domains are connected by an edge if they co-occur in the same protein. The nodes with the highest centrality scores in these networks corresponded to domains involved in signal transduction and cell-cell contacts.

### 4.1.3 Betweenness Centrality Measures

In [102], the concept of *betweenness centrality* was introduced as a means of quantifying an individual's influence within a social network. The idea behind this centrality measure is the following:

> *An important node will lie on a high proportion of paths between other nodes in the network.*

Formally, for distinct nodes, $u, v, w \in \mathcal{V}(G)$, let $\sigma_{uv}$ be the total number of geodesic paths between $u$ and $v$ and $\sigma_{uv}(w)$ be the number of geodesic paths from $u$ to $v$ that pass through $w$. Also, for $w \in \mathcal{V}(G)$, let $V(u)$ denote the set of all ordered pairs, $(u, v)$ in $\mathcal{V}(G) \times \mathcal{V}(G)$ such that $u, v, w$ are all distinct. Then, the betweenness centrality (BC) of $w$, $C_b(w)$, is given by

$$C_b(w) = \sum_{(u,v) \in V(w)} \frac{\sigma_{uv}(w)}{\sigma_{uv}}. \tag{10}$$

The authors of [103] found that for scale-free networks with exponent $\gamma \in (2, 3]$, betweenness centrality appears to follow a power law $P(C_b = k) = k^{-\eta}$. Furthermore, their numerical simulations suggested that for all such networks, the exponent $\eta$ either takes a value close to 2 or to 2.2. Based on this observation they classified synthetic and real networks including metabolic and protein interaction networks into two classes. While their finding that protein interaction networks and the metabolic networks of eukaryotes and bacteria fall into one class with the metabolic networks of archaea belonging to the other is interesting, the fact that they did not analyse the robustness of this classification scheme with respect to data uncertainty casts doubt over their conclusions. The relationship between betweenness centrality and node degree in scale-free networks was investigated in [104, 105] in which it was shown that the relationship between betweenness centrality (BC) and node-degree approximately follows the rule $BC \sim k^{(\gamma-1)/(\eta-1)}$. Here $\gamma$ ($\eta$) are the exponents in the power laws for node-degree (betweenness centrality). Thus the betweenness centrality of a node increases with incresing degree.

Recently, in [106] the measure $C_b$ was applied to the yeast protein interaction network and the mean value of $C_b$ for the essential proteins in the network was approximately 80% higher than for non-essential proteins. However, given that considerable variability in betweenness centrality has been observed in the network, this does not necessarily imply that BC is a good indicator of essentiality. It was also noted that the network contained significant numbers of proteins with high betweenness centrality scores but low node degree and that the variability in BC scores was considerably higher for low-degree nodes than for higher-degree nodes. Nonetheless, there was still a positive correlation between node-degree and BC scores. Also, while the graphs plotting the distribution of BC scores against node-degree differed from those for BA and DD network models, it is not clear whether this effect could be explained by high levels of false negatives for low-degree nodes in the interaction network. In view of this, there appears to be insufficient evidence to fully justify the authors' claim that their findings are inconsistent with these network models. The positive correlation of BC scores with node degree has also been observed for transcriptional regulatory networks in the recent paper [107].

In the present context, it is worth noting the work in [108] where a definition of betweennness centrality based on random paths between nodes, rather than on geodesic paths was considered. This centrality measure was motivated by the fact that, in real networks, information does not always flow along the shortest available path between two points. This novel centrality measure and its correlation with other centrality measures on transcriptional regulatory and protein interaction networks was investigated in [109]. No results on its biological significance are contained in this paper however.

#### 4.1.4 Eigenvector Centrality Measures

Eigenvector centrality measures appear to have first arisen in the analysis of social networks, and several variations on the basic concept described here have been proposed [97, 110–112]. This family of measures is a little more complicated than those considered previously and eigenvector centrality measures are usually defined as the limits of some iterative process. The core idea behind these measures is the following.

*An important node is connected to important neighbours.*

In much of the original work presented in the sociology literature, the eigenvector centrality scores of a network's nodes were determined from the entries of the principal eigenvector of the network's adjacency matrix. Formally, if $\mathbf{A}$ is the adjacency matrix of a network $G$ with $\mathcal{V}(G) = \{v_1, \ldots, v_n\}$, and

$$\rho(\mathbf{A}) = \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|$$

is the spectral radius of $\mathbf{A}$, then the eigenvector centrality score, $C_{ev}(v_i)$ of the node $v_i$ is given by the $i^{th}$ co-ordinate, $x_i$, of a suitably normalised eigenvector $\mathbf{x}$ satisfying

$$\mathbf{A}\mathbf{x} = \rho(\mathbf{A})\mathbf{x}.$$

In the recent paper [113], the connection between various centrality measures, including eigenvector centrality, and essentiality within the protein interaction network of yeast was investigated. In this paper, the performance of eigenvector centrality was comparable to that of degree centrality and it appeared to perform better than either betweenness centrality or closeness centrality measures. A number of other centrality measures which we shall mention later in this section were also studied.

In order for the definition of eigenvector centrality given above to uniquely specify a ranking of the nodes in a network it is necessary that the eigenvalue $\rho(\mathbf{A})$ has geometric multiplicity one. For general networks, this need not be the case. However, if the network is strongly connected then it follows from the Perron-Frobenius Theorem for irreducible non-negative matrices [114, 115] that this will be the case. Similar ideas to those used in the definition of eigenvector centrality have recently been applied to develop the Page-Rank algorithm on which the GOOGLE search engine relies [116, 117]. The HITS algorithm for the ranking of web pages, proposed by Kleinberg [118], also relies on similar reasoning.

#### 4.1.5 Other Centrality Measures

Finally for this subsection, we briefly note several less standard centrality measures which have been developed in the last decade or so, with potential applications in the analysis of biological networks. For instance, in [113, 119] the notion of *subgraph centrality* was introduced and the relationship between the subgraph centrality of a protein in the yeast interaction network and its likelihood to be essential was investigated. Loosely speaking, the subgraph centrality of a node measures the number of subgraphs of the overall network in which the node participates, with more weight being given to small subgraphs. Formally, if $\mathbf{A}$ is the adjacency matrix of a network with vertex set, $\mathcal{V}(G) = \{v_1, \ldots, v_n\}$, and we write $\mu_k(i)$ for the $(i, i)$ entry of $\mathbf{A}^k$, then the subgraph centrality of node $v_i$, $C_{sg}(v_i)$ is defined by

$$C_{sg}(v_i) = \sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!}. \tag{11}$$

In [119, 120], it was found that $C_{sg}$ performed better than most other centrality measures, including degree centrality, in predicting essentiality in the yeast protein interaction network.

Other concepts of centrality that have been proposed include *flow betweenness centrality* [121], *information centrality* [122]. For completeness, we also note the recent measure introduced in [123] which ranks nodes according to the effect their removal has on the efficiency of a network in propagating information and the centrality measure based on game theoretic concepts defined in [124]. We shall not discuss these in detail here however as little work on their biological relevance has been done to date.

#### 4.1.6 Comparison of different centrality measures

Given the number of available measures of centrality, the question of their relative efficiency in predicting essentiality arises naturally. Recently, in [120] the performance of the main centrality measures discussed above in predicting essentiality in the PPI network of *S. cerevisiae* was studied. Specifically, eigenvector centrality, degree centrality, betweenness centrality, information centrality, closeness centrality and subgraph centrality were considered. For each measure, the fraction of essential proteins in the top 1%, 5%, 10%, 15%, 20% and 25% according to the centrality measures was calculated. In each case, eigenvector centrality and subgraph centrality performed best, and offered considerable improvements over the other measures when only the top 1% and 5% of proteins are considered. Closeness centrality and betweenness centrality were the weakest indicators of essentiality with degree centrality and information centrality performing comparably in all cases. In all cases however, all of the centrality measures performed significantly better than random selection as indicators of essentiality. Note that further improvement on the performance of subgraph centrality has been achieved by the notion of *bi-partivity* [125]

which quantifies the extent to which a protein is involved in closed walks of even length in the graph.

## 4.2 Alternative Approaches to Predicting Essentiality

In addition to centrality measures, a number of other methodologies for predicting gene or protein essentiality have been proposed in the last few years.

### 4.2.1 Functional Classes and Essentiality

In the Yeast Protein Database (YPD) [16] various functional classes are defined to which the proteins in yeast can be assigned. Using the functional classification of proteins in the Yeast Protein Database (YPD) [16], the authors of [11] studied the relationship between the functions of a protein in the interaction network of yeast and its likelihood to be essential. They found that the probability of essentiality varied significantly between the 43 different functional classes considered. For instance, in one class containing proteins that are required for DNA splicing, the percentage of essential proteins was as high as 60% while only 4.9% of the proteins in the class responsible for small molecule transport were essential. This suggests that to predict essentiality, the functional classification of proteins should be taken into account. However, the fact that many proteins are as yet unclassified is a significant impediment to such an approach.

In the same paper, the nodes within each of the 43 functional classes were ranked according to their degree and, within each class, the degree of a protein was found to be a good indicator of its likelihood to be essential. Genes were also ranked using the standard deviation of their expression levels across a large number of different yeast derivatives: each derivative corresponding to one gene deletion. Some connection between the variability in the mRNA expression of a gene and its likelihood to be essential was observed. Specifically, genes whose expression levels varied little were more likely to be essential. It is hypothesised in [11] that this may be due to robustness mechanisms that maintain the expression levels of essential genes close to a constant level, while those of less important genes are subject to less stringent constraints, and hence can be more variable.

### 4.2.2 Damage in Metabolic and Protein Networks

The concept of *damage* was recently defined for metabolic networks in [126] and then later for protein interaction networks in [127]. In the first of these papers, metabolic networks were modelled as directed bi-partite graphs [28]. Such a graph has two distinct sets of nodes: one contains the metabolites while the nodes of the other set represent the reactions catalysed by the enzymes of the metabolism. Each such enzyme, $v$, is assigned a score $\mathrm{dg}(v)$, its *damage* , which characterises the topological effect of deleting $v$ from the network. Essentially, $\mathrm{dg}(v)$ is the number of metabolites that would no longer be produced if the enzyme $v$ and all the reactions catalysed by it were removed from the network. The following findings about the relationship of this concept to essentiality were reported in [126].

(i) For each value of the damage, $D > 0$, let $f_D$ be the fraction of enzymes, $v$ with $\mathrm{dg}(v) = D$ which are essential. An F-test indicated that there was a statistically significant correlation between $D$ and $f_D$.

(ii) The set of enzymes $v$ for which $\mathrm{dg}(v) \geq 5$ contains 9% of all enzymes and 50% of the essential enzymes.

Based on their findings, the authors of [126] suggested that enzymes with high damage are potential drug targets. However, it should be noted that there exist

several essential enzymes, $v$, for which $\mathrm{dg}(v)$ is quite low and that, conversely, there are also non-essential enzymes with high damage scores.

More recently, in [127] an analogous concept for protein interaction networks was defined and applied to the yeast protein interaction network. The results of this paper indicate that any correlation between damage and essentiality is very weak. On the other hand, the authors of this paper found that if the set of nodes disconnected from the network by the removal of a protein $v$ contains an essential protein, then there is a high probability of $v$ itself being essential.

Finally, we note another measure of importance in biological networks which was recently described in [128]. This measure was based on the notion of *bottle-necks* within networks and its relationship to essentiality was investigated in this paper.

### 4.2.3 Metabolic Networks, Flux Balance Analysis and Essentiality

A serious drawback of centrality measures and some of the approaches described so far is that they rely on purely topological methods and fail to take any of the underlying Biology into account. More detailed models of metabolic networks incorporating the reactions in the network provide a more biologically motivated approach to predicting essentiality and assessing the impact of gene deletions. A popular approach to the modelling of metabolic networks is Flux Balance Analysis (FBA) [48, 52]. Given a vector $\mathbf{x} \in \mathbb{R}^m$ of metabolite concentrations, and a vector $\mathbf{v} \in \mathbb{R}^n$ of reaction fluxes, the network dynamics are described by an equation of the form

$$\dot{\mathbf{x}} = \mathbf{N}\mathbf{v} \tag{12}$$

where $\mathbf{N} \in \mathbb{R}^{m \times n}$ is a rectangular matrix describing the participation of the $m$ metabolites in the $n$ reactions. It follows from (12) that the steady-state fluxes must satisfy $\mathbf{N}\mathbf{v} = \mathbf{0}$. This and other constraints such as those imposed by maximal possible fluxes and the irreversibility of some reactions defines a cone of feasible steady-state fluxes. A key assumption of FBA is that organisms like *E. coli* have evolved so as to optimize some objective function such as growth rate in certain media. This leads to a linear programme which can be solved to find the optimal steady-state fluxes and associated growth rate. The validity of FBA as a modelling paradigm for real systems has been tested against experimental data in [49]. Specifically, the experimental data was consistent with the metabolism of *E. coli* being optimized for growth rate. Further results of this nature have been reported in [129].

FBA has been applied to the metabolism of *E. coli* in [46, 130] and to that of *S. cerevisiae* in [131] to computationally predict the impact of gene deletions on the organism. The deletion of a gene coding for an enzyme is simulated by setting the corresponding reaction flux to zero. The linear programme is then solved again under the new set of constraints (including this additional one) and the impact of the deletion on the optimal growth rate is calculated. The results of the computational study were consistent with experimental findings in 86% of cases for *E. coli* [46] and in 83% of cases for *S. cerevisiae* [131]. Note that the effect of gene deletion can depend on the environmental conditions, and an extended analysis was carried out in [130] which used FBA to simulate gene deletion under varying conditions.

Techniques such as FBA have the advantage of being based on real biological considerations and are more flexible than the purely graph-theoretical approaches based on centrality measures. The ability to consider different environmental conditions and assess the impact of deletions on other objective functions than growth rate are also considerable advantages. A drawback of FBA in the analysis of gene deletions is the assumption of optimality in the mutant. As pointed out in [51, 132], while the wild-type *E. coli* may have evolved to optimize growth rate, it is not as

likely that a mutant created in the laboratory will have managed to do likewise. In these papers, an alternative approach to calculating the growth rate in such mutants is presented which minimizes the adjustment in fluxes after the perturbation rather than calculating a new optimal point. This technique was found to correctly predict the essentiality of certain genes which the straightforward FBA analysis found to be non-essential.

An alternative approach to using detailed metabolic models for the prediction of gene deletion phenotypes relies on the notion of *elementary flux modes* [133]. These are minimal sub-networks of the overall network that can support steady-state operation. In this paper, the effect of deleting gene $i$ was measured by calculating the number of flux modes with positive growth rate that do not involve gene $i$ (denoted $N(\mu, i)$) and calculating the ratio of this number to the total number of flux modes with positive growth rate. Genes for which $N(\mu, i)$ was zero are claimed to be essential. The results of the computational analysis were consistent with experiment in 90% of cases. It was also shown in this paper that some gene deletions can have a serious impact of the number of growth supporting elementary flux modes while having relatively little impact on topological features of the network such as diameter. This suggests that using such topological parameters as a measure of network robustness [66] to random deletions or mutations is not appropriate for biological systems.

The relationship between node-degree in metabolic networks and essentiality was recently investigated in [134]. Note that, in contrast to protein interaction networks, for metabolic networks, gene deletion corresponds to the removal of an edge rather than a node. While the degree of a metabolite does not appear to correlate with the fraction of essential reactions in which it is involved, there is a clear correlation between the degree and the likelihood of lethality when *all* reactions in which the metabolite takes part are removed from the system. Finally, we should note that FBA has also been used in [135] to study the distribution of fluxes across the reactions in the metabolic network of *E. coli* and in [56] to construct an network of epistatic interactions for *S. cerevisiae* which was then used to classify genes into functional modules. We shall refer to this work again in the next section.

## 4.3 Final Thoughts on Essentiality

Finally, we shall discuss a number of issues with the various approaches to predicting essentiality that have been described throughout this section.

### 4.3.1 Marginal Essentiality

While our discussion has focussed on essentiality, a gene or protein may be important to an organism without necessarily being essential. For instance, some sets of non-essential genes are *synthetically lethal,* meaning that the simultaneous removal of the genes in the set kills the organism while individual deletions are non-fatal. In the paper [73], the less restrictive concept of *marginal essentiality* and its relationship to various topological measures was studied in the protein interaction network of *S. cerevisiae*. Here, proteins were classified into five groups based on their marginal essentiality: those with the lowest marginal essentiality scores being assigned to group 1, and those with the highest assigned to group 5. The authors of [73] found that the average degree and clustering coefficient of the nodes in a group increases monotonically with the group number. For instance, the average degree of those proteins assigned to Group 1 is about half of that of the proteins in Group 4. Moreover, defining a hub node to be one in the first quartile of nodes ranked according to degree, they found that less than 10% of the proteins in Group 1 are

hubs while more than 35% of those in Group 5 are hubs. The percentage again increased monotonically with the group number.

### 4.3.2 Fitness Effect and Evolutionary Rate

In [136] it was reported that the degree of a protein in the interaction network of yeast was positively correlated with the *fitness effect* of deleting the gene that encodes the protein. Here, fitness effect measures the reduction in the growth rate of the organism when the gene is deleted. This investigation was motivated by the question of whether the importance of a gene or protein for an organism correlates with the rate at which it evolves. For more information, and varying opinions on this topic, consult [100, 137–141].

### 4.3.3 Sensitivity to Data Errors

The issue of sensitivity to data inaccuracy is of critical importance for all of the techniques described here. It was noted in [127] that the measure damage discussed above is quite sensitive to false negative errors, in which a real interaction between two nodes in a network has not been identified due to experimental error. Clearly, such sensitivity to data noise has serious implications for the practical use of any of the methods described here. In particular, it is important to have a thorough understanding of the effect of missing or inaccurate data on the performance of centrality measures or other approaches to predicting essentiality. While there has been some research into this fundamental issue recently [142–145], more intensive quantitative and theoretical studies are needed before we can reliably apply the techniques discussed here to the problem of essentiality prediction. This issue is all the more important given that much of the data available on bio-molecular networks contains large numbers of false positive and false negative results [29, 75].

### 4.3.4 Essentiality and Modules

Finally for this section, we note the work of [146] on determining the essentiality and cellular function of modules within the yeast PPI network. The results of this paper indicate that the essentiality (or non-essentiality) and functionality of an overall complex is largely determined by a core set of proteins within the complex. Moreover, the essentiality of individual proteins appears to depend on the importance of the modules in which they lie. This suggests that it may be more appropriate to address the question of essentiality at the level of modules rather than individual proteins or genes and motivates the problem of extending centrality measures to deal with groups of nodes.

## 4.4 Summarizing Comments

Several measures of network centrality have been applied to bio-molecular networks, including degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. In particular, a number of recent studies on protein-protein interaction and transcriptional regulatory networks have indicated a link between the centrality score of a gene or protein and its likelihood to be essential for survival. Some measures, such as Subgraph Centrality, appear more effective at predicting essentiality. However, given the nature of biological data, the impact of inaccurate and incomplete data on centrality measures needs to be more fully investigated before definitive statements on their relative effectiveness can be made. On the other hand, techniques such as Flux Balance Analysis are preferable to purely topological measures, as the results of such methods have readily interpretable biological significance.

# 5 Motifs and Functional Modules in Biological Networks

Many bio-molecular networks appear to be modular in nature, leading researchers to investigate this aspect of the structure of real interaction networks and the possible biological mechanisms behind its emergence. Also, it has been suggested that certain small subgraphs, known as motifs, which occur with very high frequency in biological networks are the basic building blocks of these networks and can be used to categorize them. A loose hierarchical structure for bio-molecular networks was proposed in [2, 147]. Within this structure, individual nodes, are first grouped into network motifs. These are in turn grouped into larger modules of functionally related nodes before finally, the modules are themselves connected to form the overall network.

## 5.1 Identification of Network Motifs

The concept of a network motif and a basic scheme for motif detection were described in the paper [148]. Specifically, given a directed network $G$, the motifs in $G$ of size $k$ are identified as follows:

(i) For each possible subgraph, $S$ of size $k$, of $G$ count the number of occurrences, $N_S$, of $S$ in $G$.

(ii) Next randomly generate a large number of networks such that in each random network:

  (a) Each node has the same in-degree and out-degree as in the real network $G$;

  (b) Every subgraph of size $k-1$ occurs with the same frequency as in the real network $G$. Two schemes for generating the random networks are described in [148] and its supporting material.

(iii) A subgraph, $S$, is then said to be a motif of $G$ if it satisfies the following three conditions:

  (a) The probability of $S$ occurring in a random network more often than $N_S$ times is less than some prescribed value $P$ (in [148] $P$ is taken to be .01);

  (b) There are at least four distinct occurrences of $S$ in the network $G$;

  (c) The actual number of occurrences of $S$ in $G$ is significantly larger than the average number of occurrences of $S$ in the randomly generated networks, denoted $\langle N_S^{rand} \rangle$; formally, $N_S - \langle N_S^{rand} \rangle > 0.1 \langle N_S^{rand} \rangle$.

This approach to identifying motifs can be easily adapted to detect motifs in undirected networks such as protein interaction networks [149]. However, the identification of motifs within large complex networks is computationally intensive and, to the best of the authors' knowledge, standard methods are only feasible for motifs containing less than 7 or 8 nodes. In [150] a systematic method of defining network measures or "scalars" which are related to subgraphs and can be used to detect motifs was introduced. However, the precise relationship between "scalars" and subgraphs is not straightforward.

Using the scheme described above, small motifs have been identified in a number of real biological networks. In particular, the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* have been found to have one three-node motif and one four-node motif. These are the so-called *feed-forward* motif and *bi-fan* motif, shown in Figure 4. The feed-forward and bi-fan patterns are also motifs of the neuronal

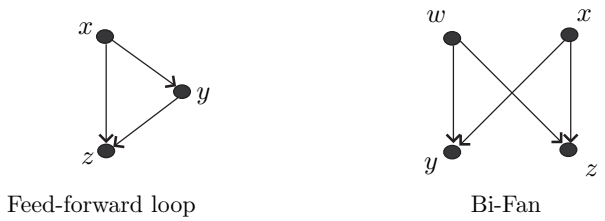



Figure 4: Feed-forward and Bi-Fan motifs of transcriptional networks.

network of the nematode *C. elegans*. This network has an additional four-node motif known as the bi-parallel motif. Other common motifs which have been detected in food webs, electronic circuits and the World-Wide-Web include the *three-chain* , *three and four-node feedback loops* and the *fully-connected triad*, see Figure 5. Note that the network motifs of the transcriptional network of yeast have also been investigated in the paper [53], where the motifs identified have also been related to specific information processing tasks.

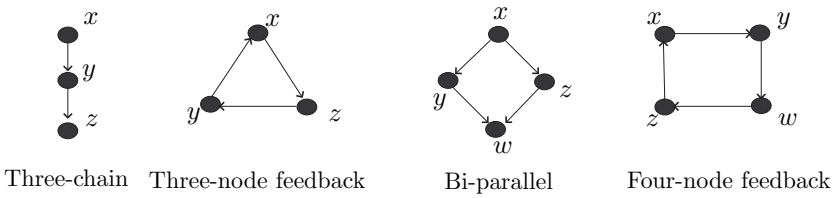



Figure 5: Common motifs in real networks.

Before proceeding, a number of facts about the findings reported in [148] are worth noting. The feed-forward loop and bi-fan motifs have been found in transcriptional regulatory networks and neuronal networks, both of which involve some form of information processing. Also, the motifs found in the food-webs studied are distinct from those found in transcriptional regulatory networks and the WWW, while electronic circuits with distinct functions tend to have different sets of motifs. These observations have led some authors to suggest that there is a connection between a network's motifs and its function, and hence, that complex networks may be classified into distinct functional families based on their typical motifs. For instance, given that information processing is fundamental to both neuronal and transcriptional networks, it is reasonable to suggest that feed-forward loops and bi-fans occur often in such networks because of their suitability for information processing tasks. On the other hand, there is no overlap between the motifs observed in transcriptional networks and those of the functionally unrelated food-web networks.

The approach to motif detection described in this section relies on a *null model* with a fixed degree distribution but otherwise randomly placed edges. Essentially, a motif is a subgraph which is over-represented in a real network when compared with an ensemble of networks with the same degree distribution. The abundance of motifs with respect to this null model, and the observations in the previous paragraph have led to the suggestion that the motifs in PPI and transcriptional networks are biologically significant. Further evidence for such a view was recently

presented in [151]. Here, the motif patterns of *geometric networks,* where links are formed based on the spatial proximity of nodes, were studied analytically and it was shown that simple geometric constraints alone are not sufficient to account for the motifs observed in biological and social networks. However, the validity of the null model described here needs to be considered carefully in the light of the findings presented in [152]. In this study, a null network model was proposed which took both the degree distribution and the hierarchical structure of the network into account (through the form of the function $C(k)$ giving the average clustering coefficient of nodes of degree $k$). Motifs such as those found in real networks were observed to occur naturally in complex networks with degree distributions and hierarchical properties similar to those of biological networks. This highlights the difficulty of drawing biological conclusions based on largely topological considerations. In order to properly assess the biological significance of motifs, a combination of theoretical and experimental work concentrating on the possible biological functions of motifs is required. Some work along these lines will be described in the following subsection.

Note also that the basic approach to motif detection described here is unsuited for weighted networks. In [32], an alternative approach to defining and identifying motifs which can be applied to weighted networks was described. A major advantage of such an approach is that it can take into account the level of confidence in network data through weighting the edges in the network appropriately. Finally for this subsection, we note that the transcriptional network of *E. coli* has been investigated in more detail in [26] and several additional motifs have been identified: *single input modules (SIMs)* ; *dense overlapping regulons* and *negative autoregulatory units.*

## 5.2 Dynamical Properties of Motifs

While the statistical abundance of motifs in bio-molecular networks suggests that they have biological significance, the work described in the last subsection does not explicitly connect motifs with any biological function. A number of recent studies have considered this problem and focussed on dynamical properties of motifs that are of biological relevance. Here, we shall concentrate on the role of the feed-forward loop (Figure 4) in transcriptional regulatory networks, which has attracted most attention in the literature to date.

Feedforward loops (FFLs) in transcriptional networks can broadly be divided into two classes depending on the nature of the individual interactions in the loop. FFLs such as that in Figure 4 consist of three basic interactions, each of which can be negative (repression) or positive (activation). If the overall sign of the indirect path from $x$ to $z$ is the same as the sign of the direct regulation of $z$ by $x$, the loop is *coherent.* Otherwise it is *incoherent.* Note that there are 8 different sign configurations for the FFL motif, 4 coherent and 4 incoherent. In [26], the specific coherent FFL in which all interactions are positive (activation) was studied numerically. Further, in this paper the joint regulation of $z$ by $x$ and $y$ was assumed to follow AND-gate logic, meaning that the concentrations of the transcription factors corresponding to $x$ and $y$ need to both be above threshold levels in order to activate $z$. In this case it was found that this motif can effectively filter out transient or fluctuating input signals. Moreover, it was also shown to respond to persistent activation with a slight delay and to shut down rapidly once the activating signal is removed. Circuits of this type are said to act as *sign-sensitive delays.*

Coherent FFL configurations seem to occur far more frequently than incoherent configurations in real systems such as the transcriptional network of *E. coli* [153], while the loop consisting of three activations is by far the most common in both yeast and E. coli [154]. In [155], a more detailed model of the coherent FFL circuit was described and analysed. Here, the robustness of the model's behaviour with respect to variations in parameter values and external perturbations was investigated. For

instance, the sign-sensitive delay action was found to be quite sensitive to variations in the model's parameters and, while the circuit is quite robust with respect to the size of external perturbations, the duration of the perturbation in comparison to the internal time-scales of the circuit appears to be critical.

Extending the work in [26], a more complete mathematical analysis of the kinetic behaviour of the FFL motif was presented in the paper [154]. Here, all eight configurations (coherent and incoherent) of the signs in the FFL were studied and both AND-gate and OR-gate logic for the joint regulation of $z$ were considered. Using numerical simulations, the steady-state behaviour and response times (the time it takes for $z$ to reach half of its steady state level in response to a step input stimulation of $x$ or $y$) of all of the FFL configurations were analysed. The coherent loops were again found to act as sign-sensitive delay circuits, while the analysis suggested that the incoherent loops could speed up response times in comparison to simple regulatory mechanisms. The work of these papers links motifs to possible biological function and suggests experiments on motifs in real systems such as those described in [153]. Here, the *l-arabinose* utilization circuit in *E. coli* was studied as an experimental example of a coherent FFL system with all interactions positive and AND-gate joint regulation. The dynamics of the system were analysed and the coherent FFL circuit indeed functioned as a sign-sensitive delay element that filters out transient activation signals from a fluctuating environment.

Before finishing our discussion of this topic, we should note a number of other theoretical and experimental investigations of the dynamical properties of network motifs. The negative autoregulatory circuit consisting of a transcription factor that down-regulates its own transcription was studied in [156], where the response times of a simple transcriptional unit (without autoregulation) and a negative autoregulatory circuit were compared.

Here, it was shown theoretically that the response-time (once again the time to reach half of the steady-state output level) of the autoregulatory circuit is shorter than that of the simple transcriptional circuit, with the same steady state. In fact, for very strong auto-repression, the response-time of the auto-regulatory circuit is only one fifth of that of simple transcription. It has also been demonstrated experimentally in the same paper that while a transcriptional circuit without autoregulation has a response-time of approximately one cell-cycle, the response-time for a circuit with negative auto-regulation is about one-fifth of a cell cycle. Finally, we also note the recent work on the kinetics of the single-input module (SIM) motif in *E. coli* [26] and the p53-Mdm2 feedback loop [157].

## 5.3 Evolutionary Conservation, Extensions and Final Thoughts on Motifs

### 5.3.1 Motifs and Evolutionary Conservation

The biological significance of motifs has been considered from a slightly different point of view in [149] where the extent to which motifs in the protein interaction network of yeast are evolutionarily conserved was studied. Specifically, 678 proteins in the yeast PPI network were identified which have orthologs [3] in each of five higher organisms, and for each 2, 3, 4 and 5 node motif, the percentage of motifs which were completely conserved across all of the 5 higher organisms was determined. A sub-graph is completely conserved if all of the proteins in it have orthologs in each of the higher organisms. For the yeast PPI network, motifs which have a higher number of nodes and are more densely interconnected also have a higher rate of conservation. For instance, the completely connected five-node motif has the highest rate of conservation of all motifs with between 2 and 5 nodes.

[3] Orthologs are genes with a common ancestor.

To validate these findings, the same number of orthologs was positioned randomly on the network and the percentages of completely conserved motifs were again calculated. In this case, the rates of conservation were considerably lower, and moreover, the rate of conservation decreased with increasing motif size, in contrast to what was observed for the real orthologs. In particular, for the completely connected five-node motif, the natural rate of conservation was found to be 47.24% while the random conservation rate was as low as .02%. Furthermore, larger, more tightly connected and conserved motifs were found to be more functionally homogeneous. In fact, for a significant number of these, all of the proteins in the complex belonged to at least one common functional class.

Note also that in [158] a correlation between the natural rate of conservation of motifs in the yeast PPI network and the suitability of the motif structure for synchronization of interconnected Kuramoto oscillators was reported. We shall have more to say about the question of synchronization later in the article.

### 5.3.2 Extensions of the Motif Concept

In [159], the significance profile (SP) was proposed as a means of classifying networks. Given a network, $G$, for each possible subgraph, $S$, the number of occurrences of $S$ in a real network $G$ is calculated and compared to the average number of occurrences of $S$ in an ensemble of random networks with the same degree profile as $G$. The $Z$-score for each such subgraph is then calculated as

$$Z_S = \frac{N_S - \langle N_S^{rand} \rangle}{\mathrm{std}(N_S^{rand})} \tag{13}$$

where $N_S$, $\langle N_S^{rand} \rangle$ and $\mathrm{std}(N_S^{rand})$ denote the number of occurrences of $S$ in $G$, and the mean and standard deviation of the number of occurrences of $S$ in the ensemble of random networks respectively. The vector of $Z$-scores for subgraphs of a fixed size is then normalized to give the *significance profile* vector.

$$SP_S = \frac{Z_S}{(\sum_S Z_S^2)^{1/2}}. \tag{14}$$

Significance profiles for subgraphs of sizes three and four are calculated in [159] for a number of real biological networks. While this method has been proposed as a means of identifying different classes of complex networks, it should be noted that some networks with similar SP vectors for three-node subgraphs have distinct four-node SPs. As mentioned in [159], this means that higher order SPs are needed if this technique is to be used effectively to classify networks. Also it is not clear at the moment how to determine the maximal subgraph size required to correctly distinguish network classes using this technique.

Another possible extension of the motif concept was recently suggested in [160]. Here, so-called *topological generalizations* of subgraphs and motifs were introduced based on duplicating certain nodes within the subgraph. Several significant motif generalizations within the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* were identified and possible functions for the observed generalizations were also proposed and investigated on simple mathematical models of transcriptional regulation and neuronal networks. While most of our discussion has focussed on transcriptional networks or protein interaction networks in isolation, this distinction is somewhat artificial, and ultimately the methods described here will need to be extended to more integrated cellular networks. In this context, the work of [58] on identifying motifs within a more complete cellular network, which takes into account both transcriptional interactions and direct protein-protein interactions, and

the study of motifs within an integrated network involving five different interaction types in [161] should be noted.

The reservations about the choice of null-model used to identify motifs discussed in Section 5.1 also apply to the work discussed here.

### 5.3.3 Some Final Thoughts on Motifs

Studying the motifs of a complex biological network can provide useful insights into the both the structure and function of the network. For instance, once we have identified a network's motifs, analysis such as that described above on the dynamical properties of the FFL motif can help us to determine the key functional roles of the network. A knowledge of the motifs of a network is a necessary step in unravelling its hierarchical structure and can be used to help develop more complete models for the evolution of bio-molecular networks than those discussed in Section 3.

Motifs and extensions such as the significance profile could be used to identify distinct categories of complex networks. However, as noted in [159], networks with the same motif profile for three-node subgraphs can have different four-node or higher order motifs and this casts some doubt on how effective these methods are likely to prove as a means of classifying networks. Moreover, the identification of higher order motifs is likely to be very costly from a computational point of view.

It is important to keep in mind that the precise biological significance of the various network motifs which we have discussed is still not fully understood and, while motifs are *statistically* significant subgraphs, there may be other subgraphs within a network, occurring in smaller numbers, that are biologically important. This issue has been debated in [162, 163], and in [164] two biological reasons for the emergence of motifs have been considered: gene duplication and convergent evolution. The findings described in [164] for the transcriptional regulatory networks of *S. cerevisiae* and *E. coli* were not consistent with the hypothesis that motifs have emerged due to widespread duplication of simple structures. This suggests that some mechanism of natural selection may have played a role in choosing the specific motif structures observed in these networks. While this provides some evidence for biological factors playing a role in the emergence of motifs, it is far from conclusive and further analysis of the dynamical properties of motifs such as that discussed in Section 5.2 is needed to more fully understand the role and biological significance of motifs in real networks.

## 5.4 Modular Structure and Function in Biological Networks

Alongside the study of motifs, there has also been significant interest recently in the larger scale organisation of biological networks. In particular, considerable attention has been given to hierarchical and modular aspects of the structure of biological networks [61]. A major motivation for this work is the need to determine the function of the large numbers of proteins or genes, even within simple organisms, whose biological roles are currently unknown.

### 5.4.1 Network Hierarchy and Motif Clusters

The authors of [165] studied how the FFL and bi-fan motifs in the *E. coli* transcriptional regulatory network are integrated into the overall network structure. Their findings suggested that the network is organised hierarchically with motifs being first aggregated into larger *motif clusters,* which are then further combined into so-called *super-clusters* which form the core of the overall network. Each motif cluster primarily consisted of the same motif type. For instance, all but one of the identified feed-forward loops (FFLs) in the network were contained in six FFL clusters,

and similarly, all but one of the bi-fan motifs were contained in two bi-fan clusters. Moreover, these motif clusters combined to form one large super-cluster containing all but one feed-forward loop and one bi-fan motif.

Another approach to investigating the hierarchical and modular structure of the transcriptional network of *E. coli* was described in [166]. Here, five different regulatory levels were identified, such that each node is either self-regulatory or else can only regulate nodes at lower levels. Based on this hierarchical decomposition of the network, a scheme for identifying modules of functionally related genes was described which appears to work quite well in identifying sets of genes with similar functionality. Many of the FFL and bi-fan motifs in this network contained genes responsible for regulating modules with diverse functions. This fact suggests that viewing motifs as the basic building blocks of functional modules cannot be entirely accurate, as, for instance, the same feed-forward loop can be involved in the regulation of numerous different modules.

### 5.4.2 Divisive and Agglomerative Algorithms for Module Identification

The detection of communities and modules in complex networks has been a subject of interest for some time in disciplines such as sociology, communications, and power systems, and a variety of techniques known collectively as *hierarchical clustering* have been developed for this problem [97, 167]. Broadly, these approaches can be divided into two classes: *divisive* and *agglomerative*. In agglomerative techniques, a measure of similarity between pairs of vertices is defined and is then used to hierarchically construct a partition of the network into modules. The core idea of such approaches is that pairs of nodes within the same module should have high similarity scores. These algorithms usually start by taking a graph consisting of the network's nodes with no edges and, at each step, add an edge between the pair of unconnected nodes with the highest similarity score until the desired community structure has been identified. Agglomerative methods have been found to work poorly in some real networks whose community structure is well-known however, and, moreover, they tend not to identify peripheral members of communities.

A recent variation on the theme of agglomerative clustering was presented in [168, 169]. Here, a quantitative measure, $M$, of the "modularity" of a proposed division of a network into modules was defined. Effectively, $M$ measures the difference between the number of edges between the modules in the division, and the expected number if edges were placed at random. For a network containing $n$ nodes, the technique in [168, 169] starts from an initial division into $n$ communities, each containing a single node, and, at each step, joins the two communities that give the greatest increase in the value of $M$. These algorithms worked quite well when applied to real and computer generated networks with a known community structure. However, the networks studied were technological and social networks and a food web and, to the best of our knowledge, there has been no effort to apply them to bio-molecular networks as yet. Also, in contrast to some techniques discussed below, these algorithms are not based on any biological considerations, and the biological significance of any modules which they may identify would need to be investigated.

A different, information-theoretic measure of modularity which applies directly to a network rather than a specific partition of the network has recently been proposed in [170]. Algorithms for splitting a network into modules were also described in the same paper and their effectiveness was tested on real and synthetic network data, with promising results. Note also the approaches based on analysis of the spectrum of the Laplacian matrix of the network described in [171, 172].

In contrast to agglomerative techniques, divisive approaches work by successively removing edges from the original network until a desired partition is obtained. While traditional methods removed edges between pairs of nodes with low

similarity scores, in [173–176] algorithms based on extensions of betweenness and information centrality to edges rather than nodes were introduced. These algorithms proceed by successively removing the edges with the highest betweenness or information centrality scores. The core idea is that edges connecting distinct communities will typically have higher betweenness and information centrality scores than edges within communities. A variant of the concept of edge-betweenness was applied to a bi-partite model of metabolic networks in the paper [177]. While the algorithm did detect some biologically relevant subnetworks, the data used in this paper was somewhat obsolete and there is no explicit justification given for the authors' claim that "the big picture should be fairly insensitive to discrepancies in the database". More recently, algorithms based on edge-betweenness have been applied to datasets on protein interaction networks in yeast and humans in [178] and their robustness to false positives in the datasets was investigated. While the results of this paper are promising, and the authors highlight the important question of when to stop the algorithm (how many edges to remove), the results given are not conclusive, it is far from clear which criterion to use to determine when to stop the algorithm and a complete theoretical analysis of the robustness of the algorithm is lacking.

An advantage of hierarchical algorithms is that they provide a picture of the organisation of bio-molecular, and other, networks at different levels of granularity, and can illustrate the integration of smaller modules into larger more complex modules. Thus, they give a more complete picture of both the modular and hierarchical structure in networks. On the other hand, most of the techniques discussed above are designed for generic networks and have no biological motivation. For this reason, while they may detect densely connected modules within bio-molecular networks, the connection between such modules and biological function is unclear.

Another serious drawback of these techniques is that they divide a network into non-overlapping modules, while in biological networks, a single protein or gene can be involved in multiple functions, and belong to more than one module or protein cluster. This observation also applies in other contexts; for instance, in social networks, there is often significant overlap between different communities or modules. Using the observation that many communities are comprised of cliques (completely connected subgraphs), which are themselves densely interconnected, an approach to module identification that addresses this issue was presented in [179]. The method described in this paper allows for the identification of overlapping modules so that, for instance, a protein may be assigned to more than one functional class.

### 5.4.3 Graph Theoretical Approaches to Identifying Functional Modules

There are also a number of network clustering algorithms which have been specifically designed for biological networks. For example, in [180] the problem of how to cluster proteins in large databases into families based on sequence similarity was considered. The first step in this algorithm was to assign sequence similarity scores to each pair of proteins using an algorithm such as BLAST. A weighted graph was then constructed, whose nodes are proteins and where the weight of an edge between two nodes is the similarity score calculated in the previous step. A weighted adjacency matrix $\mathbf{M}$ for this graph was defined using the edge-weights and normalised to be column-stochastic [115]. The TRIBE-MCL algorithm of this paper is based on *Markov chain clustering* , and identifies communities through iterating two different mathematical operations of *inflation* and *expansion* on the adjacency matrix $\mathbf{M}$. Inflation consists of taking powers of the individual entries in the matrix and re-normalising to remain within the class of column-stochastic matrices. Expansion consists of taking powers of the matrix with respect to the usual matrix

product. The core concept behind this method is that families of related nodes are densely interconnected and hence there should be more "long" paths [4] between pairs of nodes belonging to the same family than between pairs of nodes belonging to distinct families. Subsequently, in [181] this algorithm was used to identify functionally related families in the protein interaction network of *S. cerevisiae*. In fact, the algorithm was applied to the *line-graph* $L(G)$, where the nodes of $L(G)$ are the edges of $G$ and two nodes in $L(G)$ are connected if the corresponding edges in $G$ are incident on a common node in $G$. Three separate schemes of protein function classification were then used to validate the modules identified with this algorithm, and the coherence of functional assignment within these modules was significantly higher than that obtained for random networks obtained by shuffling protein identifiers between modules. This together with further analysis indicated that the identified modules did represent functional families within the network.

In the recent paper [56], the PRISM algorithm for identifying modules of functionally related genes based on analysing *epistatic* networks of interactions was presented. The core idea behind this algorithm is that genes belonging to one functional module should interact with genes in another module in a similar fashion. Using this algorithm, it was possible to group genes with similar functional annotation into the same module even in the absence of a direct interaction between them. Finally, we note that in [182], a technique for identifying *quasi-cliques* in protein interaction networks based on the eigenvectors of the network's adjacency matrix was described and applied to the yeast interaction network. Most of the quasi-cliques identified in this way were found to have homogeneous functional annotation in the MIPS database suggesting that this technique could be useful in assigning function to unannotated proteins.

Further approaches to the determination of functional modules within biological networks have been described in [128, 183]. The technique in [128] relies on searching for *highly connected subgraphs (HCS)* where a HCS of a graph $G$ is a subgraph $S$ for which at least half of the nodes of $S$ must be removed in order to disconnect it. On the other hand, in [183, 184] a procedure is described which identifies modules of related genes in the transcriptional regulatory network of yeast as well as the regulators of each such module. Other approaches to determining functional modules within transcriptional networks have been described in [25, 185]. The techniques described in these papers are not based on a graph theoretical analysis of network topology however; in fact, they rely on analysing gene expression data across different experimental conditions and determining sets of genes which are regulated by common transcription factors.

## 5.5 Predicting Protein or Gene Function from Network Structure

Several direct approaches to assigning functions to unannotated proteins have also been proposed recently. The simplest of these is the so-called *majority rule* which works in the following way [186, 187]. Given a classification scheme with an associated set of functions,

$$\mathcal{F} = \{f_s : 1 \leq s \leq M\},$$

an interaction network, $G$, and an unannotated protein $i$ in $G$, each function, $f_s \in \mathcal{F}$, is assigned a score which is simply the number of times $f_s$ occurs among the annotated neighbours of $i$. The functions with the highest scores are then identified as the most likely functions for the protein $i$. A simple extension of this concept which takes into account nodes other than the immediate neighbours of the unannotated protein was presented in [188]. It should be noted that this approach has the major

[4] Two distinct paths need not consist of disjoint sets of nodes and edges

drawback of relying entirely on the functions of previously annotated proteins, while it can often happen that none of the neighbours of a protein of unknown function have been annotated.

Two more sophisticated approaches to protein function prediction that avoid the above mentioned difficulty were described in [189, 190]. Essentially, these algorithms assign functions to the proteins in an interaction network so as to minimize the number of pairs of interacting proteins with different functional assignments. A key aspect of these approaches is that the optimal global assignment of protein function is not unique. In practice, a number of different optimal solutions are determined, and the frequency with which a given function $f_s$ is assigned to a protein $i$ is interpreted as the probability of the protein having that function.

The work presented in two other recent papers is also worth noting in the present context. Firstly, in [187], the *functional flow* algorithm was described. The core idea of this method is to consider annotated proteins within the network as reservoirs or sources of flow for the functions assigned to them. Each such function then "flows" through the network according to a specified set of rules and the amount of each function at a node when the iterations finish is used to determine the most likely functions for that node. On the other hand, the technique described in [12] is based on the hypothesis that pairs of proteins with a high number of common interaction partners are more likely to share common functions. Formally, for a pair of proteins $i$, $j$, of degrees $n_1$, $n_2$ respectively, with $m$ common interaction partners, the probability $p(i, j, m)$ of them having $m$ common partners if links were distributed randomly is calculated. This method was applied to the protein interaction network of *S. cerevisiae* and, of the 100 pairs of proteins with the lowest value of $p(i, j, m)$, over 95% of them consisted of proteins with similar function. The authors also described how to use these basic ideas to identify modules within an interaction network and validated the method on the yeast interaction data. A related probabilistic approach to using interaction network topology to predict protein function has also been presented in [191].

## 5.6 Summarizing Comments

In many real biological, and technological, networks, certain small subgraphs occur far more frequently than would be expected for randomly wired networks with the same degree distribution. Such subgraphs are known as motifs, and experimental observations have indicated that networks with similar function tend to have similar sets of motifs. This has led researchers to consider a network's motifs as being characteristic of the network in some sense. However, the precise biological significance of motifs is still not completely understood, and it has been suggested that the abundance of certain motifs may be a consequence of the degree distribution and hierarchical structure of real networks. On the other hand, several recent studies on the dynamical properties of simple motifs have provided some insights into their biological significance. In particular, the dynamics of the FFL motif and the auto-regulatory motif in transcriptional networks have been studied and linked to biological function. The modular structure of biological networks has also attracted a considerable deal of attention, and a number of automatic algorithms for the identification of functionally homogeneous modules have been proposed. This line of research and the work on direct methods for predicting gene and protein function are motivated by the need to determine the function of large numbers of unannotated genes and proteins.

# 6 Synchronization

So far, our discussion of complex biological networks has largely focussed on their *structural* properties. We shall next consider the relationship between dynamic behaviour and network topology. Synchronization is a population effect that emerges in complex systems comprising a large number of identical or nearly identical components. In the natural world, synchronization manifests itself across different levels of organization, from groups of organisms (the synchronous flashing of fireflies [192]) down to groups of cells [193, 194]. It has also been implicated in discussions on the binding problem, one of the central problems in the philosophy of mind. Specifically, in this context, synchronization has been put forward as a mechanism that might explain how information, distributed across the brain is integrated to form a coherent perception [195, 196]. Given the variety of applications, the importance of understanding the principles of synchronization is clear. One way to gain such understanding is to try to reproduce this phenomenon *in silico*, using a simple mathematical model of coupled oscillators.

## 6.1 A Model of Synchronization

One of the first detailed mathematical treatments of synchronization was presented by Arthur Winfree. His 1967 paper [197] laid the basis for the work of Kuramoto and others, who helped develop it into a mature mathematical theory with applications in different fields [198–200].

We consider a simplified model of synchronization, introduced by Kuramoto [199, 201, 202]:

$$\dot{\theta}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\theta_i - \theta_j). \tag{15}$$

Here $\theta_i$ and $\omega_i$ respectively denote the phase and intrinsic frequency of oscillator $i$; $K$ is the coupling strength, and $N$ is the number of oscillators. This setting assumes undirected all-to-all coupling. Studies indicate that the emergence of synchronization in the model is robust with respect to variations in the interconnection structure, albeit that the transition dynamics generally depend upon the details of the underlying topology.

A qualitative description of the behaviour of the system (15) is as follows (see [203]). When the interactions are weak, i.e. $K$ is small, the system is in an incoherent state, in which the distribution of the phases $\{\theta_i\}$ is roughly uniform. In this state, each oscillator tends to oscillate at its own intrinsic frequency, $\omega_i$. When the level of interaction is gradually increased, clusters of oscillators emerge, oscillating at a common frequency and (sometimes) phase. When the coupling is still further increased, more and more oscillators join in, leading eventually to a state of full synchronization in which all oscillators are oscillating as one. Note that, strictly speaking, full synchronization is only possible when all the oscillators are identical, i.e. when $\omega_i = \omega_j$ for all $i, j$. The transition from a completely incoherent to a completely coherent state is typically steep, and has an associated critical coupling strength $K_c$, which marks the start of this transition.

The analysis of the Kuramoto model has led to a number of important results. For a comprehensive review, see [199]. The majority of these results only strictly hold in the thermodynamic limit when $N \to \infty$, although some results are available for large but finite populations [204] and in general finite-dimensional models are beginning to receive more attention [196].

## 6.2 Types and Measures of Synchrony

In a system of coupled oscillators such as (15), the emergence of synchronization is easy to detect and quantify. In experiment, this is not quite as easy. The fundamental difficulty is how to extract information about phase and frequency from complicated time series of varying natures. This is a non-trivial problem as the underlying processes are typically non-stationary and, in a strict sense, non-periodic.

Before trying to detect synchronization proper, there are a few other things one can do. For instance, to test for statistical dependence between two time series, one could compute the spectral covariance or *coherence* [205]. In [206, 207] this technique was used to quantify task-specific interactions in the brain. In recent years, it has been suggested that this measure would lack the sensitivity required to detect subtler forms of synchrony, such as phase synchrony, as it would not separate out effects of amplitude and phase.

Other measures of synchrony include phase coherence [196, 208, 209], entropy, and mutual information [210, 211]. These latter measures are particularly popular among experimentalists, who seek to establish, for instance, whether or not a particular phase relationship exists between a given set of experimental variables. The application of these measures is limited by the fact that, in a typical experiment, phase information is not directly accessible, but needs to be extracted from the recorded (noisy) time series using specialized algorithms. This is a nontrivial problem as the time series (e.g. EEG recordings) are generally non-periodic, and hence standard notions of phase do not apply. However, there exist alternative notions of phase that do generalize to non-periodic signals. Based on these notions, computational techniques have been developed that are capable of extracting phase information from arbitrary time series [212]. These techniques have been successfully applied to the analysis of brain data [208–210], revealing interesting patterns of synchrony.

Another factor that might complicate the application of these measures in practice, is the lack of statistics. If prior information about the data was available one could use this to specify what degree of coherence should be considered statistically significant. But in an experimental setting, such information is rarely available. A typical way to overcome this problem is to use schemes which generate ensembles of surrogate data that are in some sense statistically similar to the original time series [210]. An early example of an application of this approach can be found in [209].

For the system of coupled oscillators (15), the standard measure of synchrony is the order parameter [203, 213, 214]:

$$r(t) \quad = \quad \left| \frac{1}{N} \sum_{j=1}^{N} e^{i\theta_j(t)} \right| . \tag{16}$$

Geometrically, the value of the order parameter indicates how well a given set of unit vectors are aligned with respect to one another (with 1 indicating perfect alignment). A slightly more general definition is adopted in [215], incorporating the adjacency matrix to account for the network's local structure. Much the same measure is used again in [216].

## 6.3 Synchronizability and systems of coupled oscillators

The study of systems of coupled oscillators has recently been extended from dealing exclusively with networks with all-to-all coupling to include networks with local connectivity, such as lattices, scale-free and small-world networks.

In [215] the transition behaviour of an appropriately defined order parameter was approximated to good accuracy in large networks of almost arbitrary structure.

Notably, the following expression for the critical coupling strength was derived:

$$K_c = \frac{k_0}{\lambda_N}. \tag{17}$$

Here $K_0$ is a constant, depending on the distribution of the oscillators' intrinsic frequencies, and $\lambda_N$ is the spectral radius of the network's adjacency matrix. Note that this estimate requires (full) knowledge of the adjacency matrix. Subsequently, a less restrictive estimate was obtained after introducing the additional assumption that the eigenvector associated with the spectral radius equals, or is approximately equal to, some scalar multiple of the vector of node degrees. Under these assumptions, the expression for the critical coupling is given as:

$$K_c = k_0 \frac{\langle k \rangle}{\langle k^2 \rangle}, \tag{18}$$

which coincides with the result reported in [216]. For a detailed account of the validity of the various assumptions we refer to the paper [215]. In the above expression (18), $\langle k \rangle$ and $\langle k^2 \rangle$ denote the first and second moments of the node degree distribution, respectively. As pointed out in [216], for scale-free networks with a power law coefficient between 2 and 3, the second moment grows without bound as the number of nodes tends to infinity. This would suggest that, in such networks, there is no critical coupling in the thermodynamic limit; or indeed no threshold for coherent oscillations. This is not true in finite networks [158, 216]. Indeed, in [216] it is reported that there exists a clear dependence between the critical coupling strength and the network size. Related observations have been reported in the literature on disease propagation. Particularly, the absence of an epidemic threshold has been established as a characteristic feature of disease spread models on (infinite) scale-free networks. Finite-size effects have also been discussed in this context [217, 218]. The similarity between the physics of coupled oscillators and models of disease spread has been discussed previously in [216] and elsewhere.

### 6.3.1 Factors that Promote Synchronization

Let us consider what structural properties of a network enhance its synchronizability. We have already seen that in networks with heavy-tailed degree distributions (that is, with large second moments), the critical coupling is generally low. In other words, a network's propensity to synchronize appears to be positively correlated with the heterogeneity of its degree distribution.

Another factor that appears to have significant impact is the *clustering* coefficient, [213]. Indeed, simulation results indicate that networks (Poisson or scale-free) that share the same number of nodes, the same number of edges and the same degree distribution, but have a different average clustering coefficient, can have very different synchronization properties. In particular, it was found that increasing the clustering coefficient of a Poisson network leads to a more gradual transition from incoherence to coherence. For scale-free networks, the effect was more ambiguous in that increased clustering appeared to promote the onset of synchronization at low coupling strengths, suppressing the same at high coupling strengths. For moderate coupling strengths the network would seem to split into several dynamic clusters oscillating at different frequencies. The authors proposed that scale-free networks with high clustering undergo two separate transitions: a first transition to a partially synchronized state, corresponding to the formation of clusters oscillating at distinct frequencies; followed by a second transition to full synchronization when the clusters are tuned to a common frequency.

In [158], it was demonstrated numerically that (finite-size) scale-free networks of Kuramoto oscillators exhibit a phase transition at a coupling strength that is inversely proportional to the average node degree. In the same study, the authors also investigated the 'fitness for synchronization' of particular network motifs, defining fitness as the (normalized) coupling strength at which the probability that a motif synchronizes first exceeds one half. The results suggested that motifs with high interconnectedness are more prone to synchronize. Interestingly, this ability to synchronize was found to be correlated with the motif's natural conservation rate in the yeast protein interaction network (see Section 5.3).

In small-world networks, the onset of synchronization appears to depend strongly on the rewiring probability, especially when this probability is small. In fact, no synchronization whatsoever is observed when this probability tends to zero [219] (in the simulation only relatively sparsely connected networks were considered). Interestingly, the transition behaviour does not appear to change much after the rewiring probability reaches a value of 0.5, suggesting that some form of saturation sets in.

## 6.4 Master Stability Functions

A second important stand of work on synchronization centres around the theory of Master Stability Functions (MSF). The main idea here is as follows. Let $\mathbf{f}, \mathbf{h} : \mathbb{R}^m \mapsto \mathbb{R}^m$ be differentiable, and let $\mathbf{G} \in \mathbb{R}^{N\times N}$ be such that $\sum_j G_{ij} = 0$ for all $i$. Also, let $K > 0$. Consider the system of differential equations

$$\dot{\mathbf{x}}^i = f(\mathbf{x}^i) + K \sum_j G_{ij}\mathbf{h}(\mathbf{x}^i), \qquad i = 1, 2, \ldots N. \tag{19}$$

The theory of Master Stability Functions is concerned with the stability of the *synchronization manifold* $\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^{mN} : \mathbf{x}^i = \mathbf{x}^j \ \forall (i,j)\}$. Observe that $\mathcal{S}$ is an invariant of the system dynamics, that is, if $\mathbf{x}(t_0) \in \mathcal{S}$ for some $t_0 \in \mathbb{R}$ then $\mathbf{x}(t) \in \mathcal{S}$ for all $t \geq t_0$. This is by virtue of the assumption that $\mathbf{G}$ has zero row sums.

In the framework outlined above, the map $\mathbf{f}$ represents the local dynamics, given by $\dot{\mathbf{x}}^i = \mathbf{f}(\mathbf{x}^i)$ (this corresponds to the situation when $K = 0$ in Eqn. (19)). The map $\mathbf{h}$ is an output function that determines which of the local state variables (or what combination thereof) can be accessed from outside (globally). The matrix $\mathbf{G}$ encodes for the network topology and generally coincides with the (normalized) Laplacian [220], or a weighted version thereof. Finally, the parameter $K$ represents the coupling strength.

A typical problem in the MSF framework is to find or modify a coupling scheme $\mathbf{G}$ such that the synchronization manifold is stable for the largest range of coupling strengths. The standard approach is to first linearize the nonlinear systems of ODEs (19) around a point $\mathbf{s}$ on the synchronized manifold. The resulting system of linear ODEs may then be decoupled using a transformation that involves diagonalizing $\mathbf{G}$. Provided that $\mathbf{G}$ is diagonalizable, this results in a system of variational equations

$$\dot{\boldsymbol{\eta}}^i(t) = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{s}) + K\gamma_i \frac{\partial \mathbf{h}}{\partial \mathbf{x}}(\mathbf{s})\right) \boldsymbol{\eta}^i(t) \tag{20}$$

where $\gamma_i$ denotes the $i$-th eigenvalue of $\mathbf{G}$, ordered by magnitude. It follows that the synchronization manifold is stable if the maximum Lyapunov exponent of the generic variational equation

$$\dot{\boldsymbol{\eta}}(t) = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{s}) + z \frac{\partial \mathbf{h}}{\partial \mathbf{x}}(\mathbf{s})\right) \boldsymbol{\eta}^i(t) \tag{21}$$

is negative over the set $\Gamma_{\mathbf{G}} := \{z \in \mathbb{C} : \ z = K\gamma, \ \gamma \in \sigma(\mathbf{G})\}$, where $\sigma(\cdot)$ denotes the spectrum. Equation (21) is called the Master Stability Equation (MSE). The

associated Master Stability Function is the function that maps the complex number $z$ onto its corresponding Maximum Lyapunov Exponent (MLE). If we denote the set of all values of $z$ which render the MLE of (21) negative by $\Lambda$, then a sufficient condition for $\mathcal{S}$ to be stable is that $\Gamma_{\mathbf{G}} \subset \Lambda$.

It has been observed that, for a large class of oscillatory systems, the intersection of the set $\Lambda$ with the reals tends to define an interval [221]. That is, typically, there exist real numbers $\alpha_{\min}$ and $\alpha_{\max}$ such that the Lyapunov exponent associated with the MSE (21) is negative for all real-valued $z$ satisfying $\alpha_{\min} \leq z \leq \alpha_{\max}$. Hence, in case the eigenvalues of $\mathbf{G}$ are real-valued (which they are, for instance, when the underlying graph is undirected), we have that the system (19) is robustly synchronizable (RS), meaning that there exist $K$ such that $\mathcal{S}$ is stable, if

$$\frac{\gamma_N}{\gamma_2} < \frac{\alpha_{\max}}{\alpha_{\min}},$$

where $\gamma_2$ and $\gamma_N$ denote the first nonzero and the largest eigenvalue of $\mathbf{G}$, respectively. Obviously the smaller the ratio between the eigenvalues, the more likely it will satisfy the above inequality. For this reason the said ratio has been proposed as a measure of synchronizability. Note, it does not measure how well the subsystems are synchronized (if at all they synchronize, synchronization is perfect); rather, it measures how robust the synchronized state is against perturbations in the parameters, particularly the coupling strength.

### 6.4.1 Unweighted Networks

Several studies have investigated how synchronizability in regular lattices and Erdös-Renyi networks, compares with synchronizability in small-world or scale-free networks. In [221] it was shown that by randomly adding links to a so called pristine world, a cycle of $N$ nodes wherein each node is coupled to its $2k$ nearest neighbors, robust synchronizability can be significantly improved. This is to be expected in that, as the network tends towards a situation in which it is fully connected, the ratio between the eigenvalues of the associated Laplacian tends to one. More interestingly, therefore, is the question how efficient this procedure is in terms of the cost associated with adding in new links. It turns out that the procedure is very efficient indeed. Small-worlds generated from pristine worlds with low values of $k$ $(1, 2, 4)$ need only relatively few connections added, compared to ER networks and hypercubes for large enough network sizes, to make them robustly synchronizable.

It is natural to ask what properties of small worlds enables them to sustain a robustly synchronized state in the face of perturbations? Intuitively, one could argue that robust synchronizability is correlated with average network distance (this distance being relatively small for small-world networks), and so one would expect that the smaller the average network distance, the easier it is to robustly synchronize the network. In [222], this argument was shown to be false. In particular, it was shown that the ratio $\frac{\gamma_N}{\gamma_2}$ correlates negatively with average distance. Among the networks included in this study were a class of semi-random scale-free networks, a class of aging scale-free networks, and a modified version of the classical Strogatz-Watts Small-World network. Rather than short average network distance, the authors proposed the homogeneity of the degree distribution as an important indicator of the network's propensity to robustly synchronize. Once again, it is important to keep in mind that the eigenvalue ratio as a measure of robust synchronizability has little to do with the *onset* of synchronization, as it is studied, for instance, in the framework of coupled oscillators. This would explain how a relatively heterogeneous degree distribution can promote the onset of synchronization (although a state of full synchronization may be hard to attain in this case) while rendering

the synchronized state more sensitive to parameter variations (and less robustly synchronizable).

### 6.4.2 Weighted networks

The above discussed results primarily applies to settings with unweighted networks. More recently, people have begun looking at synchronization in weighted complex networks. The motivation is that biological, as well as technological networks are typically weighted. A recurring example is that of neuronal networks [223–225].

In [225], the authors propose a coupling scheme in which the weight of an edge incident on a node $i$ is normalised by the degree $d_i$ of that node, so that the sum over all weights associated with the edges incident on that node equals unity. More precisely, they propose a coupling scheme

$$\mathbf{G} := \mathbf{D}_\beta^{-1}\mathbf{L},$$

where $\mathbf{D} := \operatorname{diag}\begin{pmatrix} d_1^\beta & d_2^\beta & \cdots & d_N^\beta \end{pmatrix}$ and $\mathbf{L}$ is the usual Laplacian. They go on to show that, for a variety of different networks, including a class of Scale-Free and Small-World networks, robust synchronizability is optimal when $\beta = 1$. As indicated, this choice of $\beta$ essentially neutralizes the heterogeneity in the degree distribution. As a result, the weighted network has behavioral characteristics similar to those of a random regular network with the same (mean) degree, including good robust synchronizability.

A similar approach was adopted in [223]. However, instead of scaling the weight of an edge by the degree of the node it is incident on, the authors propose to scale an edge by the collective 'load' associated with all the edges connecting to the same node as the given edge. Here, the load of a link is related to the number of shortest paths that make use of this link. Let $l_{ij}$ denote the load on the link from node $i$ to node $j$ (with the convention that $l_{ij} = 0$ if no such link exists). Then the proposed coupling scheme is as follows:

$$\mathbf{G} := \mathbf{D}_\alpha^{-1}\mathbf{L},$$

where in this case, $\mathbf{D}_\alpha := \operatorname{diag}\begin{pmatrix} \sum_j (l_{1j})^\alpha & \sum_j (l_{2j})^\alpha & \cdots & \sum_j (l_{Nj})^\alpha \end{pmatrix}$. The results in the paper indicate that, for a large class of networks, the propensity for robust synchronization is optimal when $\alpha = 1$. Note that the case $\alpha = 0$ corresponds to the situation outlined above when weights are scaled by the degree of the node they are incident on (the case $\beta = 1$).

In both of the above discussed approaches, a key assumption is that $\mathbf{G}$ is diagonalizable. Interestingly, it was shown recently [226] that when weights are assigned to optimize robust synchronizability, the resulting coupling scheme is almost always nondiagonalizable. The authors proposed assigning weights so that the network becomes an oriented spanning tree, subject to constraints such as homogeneity of the node intensities (see below). This endows the network with a hierarchical structure at the top of which is a master oscillator, which entrains a set of slave oscillators, which in turns entrain other oscillators and so on.

The results outlined here demonstrate that, as far as weighted networks go, heterogeneity in the degree distribution need not rule out good robust synchronizability (RS). In fact, what constitutes poor RS is not so much heterogeneity in the degree distribution as heterogeneity in the distribution of node intensities [224], where a node's intensity is defined as the sum of the weights of the links incident on that node. Indeed, it appears that, as long as the distribution of intensities is reasonably homogeneous and the mean degree sufficiently high, good RS is guaranteed.

## 6.5 Synchronization in the brain

Synchronization has been put forward as a putative mechanism that would make possible the integration of distributed neural activity in the brain [227]. Indeed, recent studies suggest that during processing of visual and auditory stimuli, activities of functionally specific brain regions are temporally aligned so as to produce a unified cognitive moment. This would imply that an inability to synchronize, due to abnormalities in the neural circuit for instance, could have severe behavioral implications [208, 228]. An understanding of the mechanics of this phenomenon may thus hold the key to devising new treatments for neurological disorders.

It has been known for quite a while that groups of neurons *within a single sensory modality* such as the visual cortex, selectively synchronize their activities, supposedly to integrate the particular features for which they encode. However, the notion that this same kind of integration would also take place *across different sensory modalities* was discovered only recently. In a study reported by Roelfsema *et al.* [206], five cats were conditioned to press and release a lever in response to particular visual stimuli. Electrodes were implanted at different locations in the motor and visual cortices to monitor the electrical activity during execution of the task. Coupling between these brain areas was investigated using cross-correlation analysis on pairs of LFP (Local Field Potential) traces. Tighter coupling was observed when the animals were engaged in the specific visuomotor task than when engaged in feeding or at rest. Based on these and other findings, Varela *et al.* have suggested that "large-scale synchrony is the underlying basis for active attentive behaviour". [227, 229].

In a more recent study [207], it was investigated how the interactions between selected areas in the hippocampus and amygdala in fear-conditioned mice compare against those in controls. The response of the fear-conditioned group indicated a selective synchronization in the theta frequency range (4-7 Hz) upon presentation of the conditioned stimulus, which was not found in the control group. No significant synchronization was observed in either group during presentation of the unconditioned stimulus. It was argued that these results are indicative of a functional relationship between theta rhythm synchronization and the retrieval and expression of fear.

### 6.5.1 Abnormal Neural Synchrony and Schizophrenia

Assuming synchronization is the mechanism that underlies neural integration, it seems reasonable to suppose that disruptions in neural synchrony would impact one's behaviour. Interestingly, an impaired ability to integrate information has long been identified as one of the symptoms of Schizophrenia. Other symptoms include delusions, hallucinations, and incoherent thoughts, as well as social withdrawal, poor motivation, and apathy [230, 231]. In recent years, it has been proposed that these cognitive and affective impairments may be related to a defect in the mechanism believed to be responsible for the integration of distributed neural activity, that is, to gamma band synchronization [208, 214, 228].

A recent report supports this [208]: when a set of Gestalt images were presented to a group of patients diagnosed with Schizophrenia (SZ) and a group of Normal Control (NC) subjects, a significant difference in neural orchestration between the two was observed. A phase-locking response, persistent among individuals from the NC group, but absent in the SZ group, was hypothesized to reflect a feature-binding mechanism in the visual cortex which would explain the more efficient task performance by healthy individuals.

Further evidence for abnormal neural synchrony in Schizophrenia was reported in [211]. In this study, two groups, patients and controls, were presented with a

set of images depicting six basic human emotions, which they were to recognize. The response of each individual was measured using whole head MEG (Magnetoencephalogram). Local activity was averaged over a region of interest (ROI) and a coherence score was computed as the mutual information (MI) [232] between ROIs. The MI analysis revealed a very organized pattern of linkages for normal subjects, as opposed to the overall disturbed linkages for Schizophrenia patients. At some level, these results agree with the outcome of another study [214], which involved first-degree relatives of patients with Schizophrenia. Gamma-band synchronization was found to be reduced in first-degree relatives with Schizophrenia Spectrum Personality Problems.

### 6.5.2 A Theory of Neural Synchronization?

It has been established that the processing of particular audiovisual stimuli coincides with the temporal synchronization of neural activities in functionally specialized brain regions. In addition there is some evidence that patients with Schizophrenia or related neurological disorders are more likely to display abnormal patterns of synchrony than controls. Meanwhile, the mechanics of this synchronization and its supposed role in the integration of information remain poorly understood. Most experimental studies resort to elementary statistical techniques to conclude with confidence that some form of synchronization takes place. Beyond that, there appears to be a shortage of quantitative models; models that do not just extract information from the data, but indeed attempt to explain the data. With no disrespect for the seminal importance of Kuramoto's work, and that of others' who have contributed to the theory of coupled oscillators, it appears that we are still far removed from effectively applying this theory in the context of the neural synchronization problem. Fortunately, there is reason to believe that this gap is closing fast, considering on the one hand the rate at which measurement techniques are being refined, and, on the other hand, some of the pioneering work that is being done on the theoretical front.

## 6.6 Summarizing Comments

Much of the research on the relation between network structure and synchronizability has focussed on networks of coupled Kuramoto oscillators. The onset of synchronization in complex networks of coupled oscillators appears to be determined by a few key factors, the most important of which is the heterogeneity of the degree distribution. In particular, when the variance of the degree distribution tends to infinity, as is the case, for example, in scale-free networks with a power law exponent between 2 and 3 and network size tending to infinity, the value for the critical coupling tends to zero. For finite-size scale-free networks, the critical coupling is generally nonzero. The theory of Master Stability Functions provides a useful tool in studying the robustness of a network's capacity to synchronize. Results based on this theory indicate that generically, synchronizability can be improved by introducing weights and directionality while maximal synchronizability is attained by balancing node intensities.

One important future direction for research in this area that deserves mention is the study of adaptive weighted networks [233]. Real life networks are hardly ever static. From a modelling perspective it is essential that, when in the years ahead new experimental data become available, which will eventually include detailed dynamic information, one has in place the right models to accommodate the forms of interaction these data may reveal.

# 7 Network Structure and Disease Propagation

The mathematical theory of epidemics has been the subject of intensive research for some time now and several different models for disease spread have been developed [234, 235]. Recently, researchers have begun to investigate how the novel properties observed in social networks and in networks of human sexual contacts [236] effect the behaviour of various models of disease spread.

## 7.1 Scale-free Networks and Epidemic Thresholds

Here, we shall confine our discussion to results concerned with the two basic models of disease spread on which the recent literature on network epidemiology has largely focussed: the *Susceptible-Infected-Susceptible* or *SIS* model and the *Susceptible-Infected-Removed* or *SIR* model. In the SIS model, a population is divided into two groups: the first (S) consists of susceptible individuals, who are not infected but can contract the disease from members of the second group (I) of infected individuals. After a period of time, an infected person recovers and then becomes susceptible again. Hence no immunity is conferred by contracting the disease and the recovered infective can become infected again at a later time. In contrast, in the SIR model, a recovered infective is regarded as being immune to the disease and cannot subsequently become infected again. Hence, the population is divided into three groups in such models: susceptibles (S), infectives (I) and removed or recovered (R).

There are two fundamental parameters associated with any SIS or SIR model: the probability $\lambda$ of an infective passing on the disease to a susceptible with whom they are in contact during the period in which they are infective, and the rate $\nu$ at which an infective recovers. In basic models of population epidemiology, it is assumed that the population is homogeneously mixed. This essentially amounts to assuming that each individual, or node, in the population has the same number of contacts. Under the assumptions of homogeneous mixing and a fixed population size, the standard equations for the SIR model are given by [237, 238]

$$\begin{aligned} \frac{dS}{dt} &= -\lambda SI \\ \frac{dI}{dt} &= \lambda SI - \nu I \\ \frac{dR}{dt} &= \nu I. \end{aligned} \tag{22}$$

Here, the variables $S(t), I(t), R(t)$ represent the total number of individuals in the susceptible, infected and recovered classes respectively at time $t$. From a network point of view, we can consider the population as a graph, $G$, in which each individual is represented by a node and each edge represents a contact through which the disease can spread. In a homogeneously mixed population, each node $v$ in $G$ has the same degree, which would be equal to the mean degree, $\langle k \rangle$, of the network. This assumption is only reasonable for networks whose degree distributions are narrow, meaning that the coefficient of variation, $C_V = \sqrt{\langle k^2 \rangle / \langle k \rangle^2 - 1}$, is very small.

Under the assumption of homogeneous mixing, the quantity $\rho_0 = \langle k \rangle \lambda / \nu$, represents the average number of secondary infections that would result from the introduction of a single infected individual into an entirely susceptible population. In this case, the introduction of an infective into the population will result in an epidemic if the basic reproductive number $R_0 = \rho_0$ is greater than one, while if $R_0 < 1$, the disease will die out. Thus, defining $\lambda_c = \nu / \langle k \rangle$, an epidemic occurs if the spreading rate, $\lambda$ satisfies $\lambda > \lambda_c$ while the disease dies out if $\lambda < \lambda_c$. The constant $\lambda_c$ is usually referred to as the epidemic threshold.

While the assumption of homogeneous mixing might be reasonable for the classical ER random graph models, it is entirely inappropriate for BA and other scale-free networks with broad-tailed degree distributions. The first results concerning epidemic spread on scale-free networks were presented in [239]. Specifically, it was shown that for the SIS model on scale-free networks, surprisingly the epidemic threshold is effectively zero. Similar findings were later presented in [218], where the SIR model on networks with heterogeneous mixing was considered. For such networks, the basic reproductive number $R_0$ is given by the formula

$$R_0 = \rho_0(1 + C_V^2). \tag{23}$$

Now, in the limit as network size tends to infinity, for a scale-free network with degree distribution of the form $P(k) \sim k^{-\gamma}$ with $2 < \gamma < 3$, the coefficient of variation $C_V$ of its node-degrees is infinite (more precisely, the second moment $\langle k^2 \rangle$ diverges as the network size, $n$, tends to infinity, while $\langle k \rangle$ remains finite). Thus, for any non-zero spreading rate $\lambda$, the introduction of an infective into the population can result in an epidemic. This also follows from the following formula for the epidemic threshold for scale-free networks with degree distribution $P(k) \sim k^{-3}$, which was presented in [240] (as well as a number of other sources).

$$\lambda_c = \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{24}$$

Note that this same formula has appeared above in the context of coherent synchronization on random networks (18).

In [239] it was observed that, on a BA scale-free network, the steady state prevalence $P_{\text{ss}}$ [5] depends on $\lambda$ as $P_{\text{ss}} \sim e^{-C/\lambda}$. The same result was subsequently derived using different methods in [218]. Approximate expressions for the fraction of nodes, $I$, in a scale-free network that are ever infected for an SIR model of disease spread (the final epidemic size) have also been presented in [218]. The dependence of $I$ on $\lambda$ for scale-free networks with $2 < \gamma < 3$ followed a power law of the form $C(\lambda)^{1/(3-\gamma)}$. Also, for networks with $\gamma = 3$, the number of infected nodes of low-degree is typically small, while many (essentially all) nodes of high-degree are infected. These findings are in agreement with those described in [241], which indicate that disease spreads in a hierarchical cascade from hub nodes to nodes with intermediate degree to nodes with low degree. These observations clearly have significant implications for the development of containment strategies. Specifically, they suggest that an effective containment strategy would first and foremost target the hubs of a network. Similar recommendations have been made in [242].

Before we proceed, it should be noted that the results discussed in the previous paragraph are based on a number of assumptions. They have been derived for the limiting case of an infinite network or population, and rely on a continuous approximation of the node-degree variable $k$. When finite size effects are taken into account the epidemic threshold does not vanish but in fact takes a positive value [218]. Also, the networks for which the above results were derived do not take any correlation between the degrees of connected nodes into account. Both of these assumptions are clearly invalid for real social networks. Later in this section, we shall describe attempts which have been made to address these limitations.

## 7.2 Impact of Finite Size and Local Structure on Disease Spread

Real networks of social and sexual contacts are finite and, for this reason, a number of authors have studied the dynamics of disease spread on scale-free networks

[5] The steady-state prevalence is the fraction of infected nodes in the steady state.

with finitely many nodes. In [217], the epidemic threshold, $\lambda_c$, and the steady-state prevalence, $P_{\text{ss}}$, for the SIS model on finite scale-free networks were investigated. It was found that $\lambda_c$ is non-vanishing in this case, and formulae approximating the dependence of $\lambda_c$ and $\rho$ on the network size, $n$, were also derived. While the epidemic threshold is non-zero for finite scale-free networks, it is considerably smaller than for a corresponding homogeneous network with the same average degree [217]. In fact for scale-free networks of size larger than 1000, the threshold is at least one order of magnitude smaller than in the homogeneous case. These findings are largely in agreement with the remarks on finite-size effects for SIR models made towards the end of the paper [218]. Note also the findings reported in [243] where the behaviour of the SIS model on two different types of network with scale-free degree distributions was studied numerically. For both network types, the epidemic threshold $\lambda_c$ is non-zero. However, the dependence of $\lambda_c$ on network size and the effect of the spreading rate $\lambda$ on $P_{\text{ss}}$ varied significantly between the two classes of network, even for networks with the same underlying degree distribution. These results demonstrate that it is possible for two networks with the same degree distribution, but different local structures, to exhibit significantly different behaviours with respect to disease propagation.

In order to take more aspects of network structure into account, a number of authors have studied classes of scale-free networks in which the degrees of neighbouring nodes are correlated. Such networks offer a more realistic picture of real social networks in which such correlation is common. In [244] the SIS model was studied on a class of highly-clustered scale-free networks. Numerical simulations indicated that the highly clustered networks behave in a qualitatively different manner than the usual scale-free models, both with respect to the dependence of steady-state prevalence $P_{\text{ss}}$ on $\lambda$ and to survival probability of the disease. Moreover, the authors of this paper argue that for this highly structured class of scale-free networks, there is a non-vanishing epidemic threshold even in the limit as the network size, $n$, tends to infinity. They further conjectured that the value of the threshold depends on the degree correlations within the network rather than on the degree distribution itself.

In [245] the value of the epidemic threshold for a scale-free network was related to the largest eigenvalue of the so-called connectivity matrix $C$, where $C_{kk'} = kP(k'|k)$. Here $P(k'|k)$ represents the probability that a given link emanating from a node of degree $k$ connects to a node of degree $k'$. For networks with no higher order correlations, the epidemic threshold is equal to the reciprocal of the largest eigenvalue of $C$. Based on these results, in [246] conditions for the absence of an epidemic threshold in scale-free networks with arbitrary two-point degree correlation functions $P(k'|k)$ and degree exponents in the range $2 < \gamma \leq 3$ were investigated. The principal result of this paper established that in this case, provided the network possesses no additional, higher order, structure, the epidemic threshold is again zero in the limit of infinite network size. We should also note here the work described in [86, 247] which further investigated the effects of degree correlations and local structure on the dynamics of disease spread in scale-free networks.

## 7.3 Containment Strategies on Heterogeneous Networks

One of the most fundamental issues in epidemiology is how to design effective strategies for containing the outbreak of an infectious disease. One simple strategy is mass vaccination, in which (almost) every individual in the population is vaccinated against a disease, and hence immune to it. While this can be an effective strategy for containing infectious diseases, it is crude and operationally expensive. As a result, there is great interest in alternative strategies which, although perhaps slightly less effective, are much more economical in terms of resources and logistics.

Recently, in [240, 242], the implications of power law degree distributions for the design of immunization programmes was investigated using mean-field approximations and numerical simulations. The first strategy considered was that of uniform random vaccination in which individuals are uniformly selected at random and vaccinated. However, while this strategy can work for homogeneous populations, it is known to be ineffective in the heterogeneous case [234]. The findings in [240, 242] suggest that for scale-free networks, and the SIS model of disease spread, considerable improvements over uniform vaccination can be achieved through targeting hub nodes within a network. In fact, two different approaches of this kind were suggested. In the first of these, nodes are vaccinated with probability proportional to their degree, so that a greater proportion of nodes of high degree are vaccinated than is the case for nodes of low degree. The second strategy aims to specifically target hub nodes by vaccinating all nodes in the network of degree higher than some threshold $k_c$. While this appears to be more cost effective, in terms of how many individuals need to be immunized in order to eventually eradicate the disease, it relies on a fairly complete knowledge of the network's topology, which is not typically available for real social networks.

The selective targeting of hubs requires a fairly good knowledge of a network's degree distribution, and such global information may not always be available. In [248], an alternative strategy was proposed, based on the immunization of random acquaintances. Like uniform immunization, this strategy requires no specific knowledge about the network, and has the added advantage of becoming effective at a much lower penetration rate.

A disease containment strategy for outbreaks of smallpox, based on bi-partite graph [28] models of social networks was described in [13]. The graphs used in this paper have two distinct types of vertices, which correspond to locations and individuals respectively. A containment strategy combining targeted vaccination with early detection appeared to work effectively. Early detection can be accomplished by placing sensors at locations with high degree, that is, locations visited by many people, while efficient vaccination is effected by targeting long-distance travellers. Various factors such as withdrawing infected individuals to their homes, and delays in introducing containment measures can have an impact on the number of deaths caused by a smallpox outbreak. Numerical simulation suggested that the most significant such factor was the early removal of infected individuals to their homes with the next most influential factor being the length of delay in implementing vaccination schemes.

In [249], motivated by the recent emergence of the SARS virus, several intervention strategies for epidemic containment were considered, and the impact of each strategy on the effective reproduction number was determined. In general, the results of the paper suggest that combining different strategies is a good idea, while the strategy of tracing and quarantining the contacts of diagnosed cases was found to be particularly effective. The model studied in this paper incorporated several realistic aspects of social structure. For instance, given that people tend to be more frequently in contact with individuals within their own household than with people from other households, a distinction was drawn between *within-household* transmission and *between-household* transmission. Furthermore, school-children and the rest of the population were considered separately. While the manner of counting secondary infections, and the reproduction number, used in this paper were somewhat non-standard, they have the advantage of being analytically tractable. Parameter values pertaining to the distribution of household sizes were selected in accordance with given census data. Various control strategies were considered, including exposure avoidance, isolating cases at diagnosis, closing schools, quarantining affected household, and contact tracing. If an emerging infection were to enter a juvenile population, closing schools can reduce transmission significantly.

## 7.4 Other Network Models and the General Theory of Disease Spread on Networks

Disease propagation on network topologies other than scale-free topologies has also been considered. For instance, in [250] the impact of dynamically adding long-range links to regular one-dimensional lattices on the spread of disease was studied. Using the SIR model for disease spread, they have shown that the resulting small-world network [30] structure exhibits a shortcut-dependent epidemic threshold. An approximate expression for this threshold in terms of the effective spreading rate and the effective recovery rate was shown to be accurate over a large range of parameter values. The authors also acknowledged the fact, previously stated elsewhere [10, 218], that the basic reproduction number has limited use outside the homogeneous mixing paradigm. They argue that this is particularly true for small-world networks because "the effect of a secondary infection caused by nearest-neighbor transmission is different from the one caused by a long-range jump" [250]. Assuming a spreading probability of one, so that susceptibles in direct contact with infectives will become infected during the next iteration step, it was shown that the epidemic saturation time, i.e. the time it takes for 95% of the susceptible population to become infected, scales with $-log(n_0)$, where $n_0$ is the fraction of nodes initially infected. The scenario of spreading with near certainty would correspond to the onset of an epidemic, and is used by the authors to predict the final epidemic size as well the development of an epidemic from its beginning stages. The dynamics of the SIR model and the related *susceptible-exposed-infected-removed* (SEIR) model on small-world networks were also investigated in the paper [251].

In [252] a computational SIR-type model of global epidemic spread is presented that is based on real air-transportation and census data. The worldwide air-transportation network (WAN) comprises nearly four thousand nodes (airports) and over eighteen thousand connections between them. It has a highly heterogeneous structure, which the authors show to have a major impact on both the epidemic spread pattern and the predictability of the same. Specifically, for the WAN,the epidemic phase, during which nearly all the agents are in the infected state, tends to be relatively short. At the same time, the time it takes for the epidemic to die out is much longer. Also, the predictability of the spreading pattern (the order and degree in which respective cities are affected) appears to be relatively poor, especially during the first few weeks of an outbreak. The authors propose that this is due to the heterogeneity of the connectivity pattern, which would provide a choice of effective spreading channels.

Recently, in [253] analytical techniques were developed which can be used to derive exact solutions for a large class of standard epidemiological models on a variety of networks. These techniques are based on generating functions and allow for great flexibility in terms of assumptions on network structure and degree correlations. Further they can accommodate heterogeneity in transmission rate and infectious period and allow for correlations between parameters such as transmission rate and node degree. The results derived in this paper include formulae for the epidemic threshold and average outbreak size for the network classes considered. More recently, the problem of epidemic spread on random graph models has been studied in a mathematically rigorous fashion within the framework of Markov processes in [254]. Here, the dependence of the final epidemic size and the lifetime of an outbreak on graph parameters such as the spectral radius of the network's adjacency matrix and the isoperimetric number of the network was investigated. Some general theorems as well as results for a variety of graph models including the ER and scale-free models were derived for the SIS and SIR models of disease spread.

The techniques developed in [253] have been applied in [10] in an effort to explain

some puzzling aspects of the recent SARS outbreaks. Specifically, the question of why these outbreaks never led to an epidemic, given the relatively high estimates for the basic reproduction number, was considered. Using purely analytical tools, the authors derive expressions for the likelihood that a small outbreak results in an epidemic in, respectively, an urban network, a power-law network and a Poisson network. They found that "outbreaks are consistently less likely to reach epidemic proportions in the power-law network than in the others". It was also shown that for all three network classes (all with heterogeneous mixing) there is a nonzero probability that an outbreak does not become an epidemic, even when the spreading rate of a disease exceeds the epidemic threshold. By contrast, in the paradigm of homogeneous mixing, an epidemic will occur with certainty whenever the basic reproduction number is greater than unity. It is also worth noting that the likelihood of an outbreak is a monotonically increasing function of the degree of the first infective, and if $\lambda$ is far above the epidemic threshold, the risk of an epidemic is very high even for small initial outbreaks in the case of urban networks.

Finally, we note that the evolution of diseases on local and global networks has been studied in [255]. The basic premise of this work was that different disease strains adapt to compete for resources (susceptible hosts). In the model proposed here, adaptation corresponds to a random mutation of both the transmission rate and the infectious period, which takes place whenever a new infection occurs. As the authors point out, in mean-field models this type of evolution would result in runaway behavior with selection for ever higher transmission rates and ever longer infectious periods. By contrast, both spatial heterogeneity in local networks and the presence of shortcuts in global networks appear to constrain the evolutionary dynamics, to the effect that the rate of adaptation is generally slower (in the case of a global network, the transmission rate even saturates at some finite value) and the variability (in the dynamics) higher than in mean-field models. Simulation results suggest that in networks with many long-distance connections and a low clustering coefficient, disease strains with conservative transmission rates and long infectious periods are most likely to survive. By comparison, for networks with strong local connectivity the fittest strains are those that have high transmission rates and relatively short infectious periods.

### 7.5 Summarizing Comments

The structure of a social network can have a significant impact on the dynamics of disease propagation. In particular, for scale-free networks, in the limiting case of infinitely many nodes, the epidemic threshold is zero. This means that any non-zero spreading rate could lead to an epidemic. This fact has been established for uncorrelated scale-free networks of infinite size. For scale-free networks of finite size, the epidemic threshold is non-vanishing but considerably smaller than for a homogeneously mixed population. Results have recently been derived giving conditions under which the epidemic threshold will be zero for scale-free networks with degree correlations, in the limiting case of networks of infinite size. The dynamical behaviour of epidemics on networks with heterogeneous degree distributions has implications for the design of strategies for containing outbreaks. In particular, the targeting of nodes, or individuals, of high degree can offer significant improvements over random immunization programmes.

## 8 Conclusions and Directions for Future Research

There has been much written in the last few years about the need to move away from a purely reductionist approach to Biology, and to develop an integrative, systems-

oriented analysis paradigm. While ultimately, we may wish to understand the dynamical processes that take place in living organisms, we first need to understand how the components in biological systems interact with each other, and the biological significance of those interactions. Biological network analysis is thus a necessary, and highly important aspect of the general systems-driven approach to Biology.

Recent developments in Biology and Medicine have led to a clear need for biological network analysis. Advances in high-throughput experimental technologies have generated massive amounts of data on bio-molecular networks. Given the size and complexity of these networks, systematic methods are clearly required in order to derive meaningful information from their structure. Without the provision of such methods, the time and money spent on the construction of complete network maps will lead to little more than intricate and unintelligible graphical representations of the interactions within living cells. Moreover, current techniques for the generation of network data are error-prone. Network analysis techniques can be used to assess the accuracy of such data and to help obtain more reliable network maps in the future.

While the subject is still at an early stage of development and there is still much to do, network analysis has already been used, with some promise, to address a variety of biological problems. For instance, the efforts to determine the function or importance of a protein from network structure demonstrate that, notwithstanding the limitations of current data and methodologies, biological information can be derived from the topology of interaction networks. Breaking complex networks into modules and motifs helps to simplify their structure and gives valuable insights into network organization and function. Mathematical models for network growth, such as those described in Section 3, allow us to quantitatively test hypotheses concerning the evolution of PPI and other biological networks and reliable models can be used to test the performance of algorithms in silico. The design of effective containment strategies for disease spread, and of novel therapies for complex neurological disorders are two examples of the potential benefits of the research directions discussed in Sections 6 and 7.

It is clear from our discussion of synchronization and disease spread that network topology can have a major impact on dynamical processes. If we are to develop truly integrated models of biological processes, we need a deeper understanding of issues such as those considered in Sections 6 and 7. The analysis of the dynamics of network motifs, and their connection to biological function, is another example of how valuable insights can be gained from studying the interplay between topology and dynamics. In this case, the theoretical work has suggested novel experiments to deepen our understanding of the organization of living cells.

Despite the progress that has been made in the analysis of biological networks, there are many major issues that still need to be addressed. The unreliable quality and incompleteness of existing data sets is a serious impediment to network research, and the development of improved experimental and statistical techniques to enhance the accuracy of network maps is of vital importance for future research efforts. Network-based methods for experiment-design and the prediction of interactions should play a key role in this work.

Robustness with respect to data inaccuracy is a critical issue for the techniques used to predict essentiality and determine protein function described in Sections 4 and 5. The effect of false positives and false negatives on the performance of these methods needs to be analysed more thoroughly if they are to be used with confidence. The same comment applies to the impact of network sampling on such methods, and there is considerable scope, and need, for more research on these questions. A second major limitation of many existing methods for predicting the importance or function of a gene or protein is that they rely on static, topological

considerations and fail to take into account biological or dynamic information about the nature of a network's interactions. The extension of existing approaches to incorporate such details should form a major part of future research efforts. Flux Balance Analysis and related methods are one example of the type of work that can be done in this direction. Extending existing algorithms to weighted networks would also be a great assistance as it would allow information such as the level of confidence in an interaction or its strength to be included. With regard to network motifs, more detailed analysis of the biological and dynamical properties of motifs is required if their role is to be understood. To date, only the feed-forward loop (FFL) motif has been analysed in any depth, and combined experimental and theoretical work of the type discussed in Section 5.2 should be undertaken for other motifs also.

Practically all of the existing theoretical results on synchronisation and epidemic spread have been derived for the limiting case of infinite-size networks. Obviously, real biological networks are not infinite and both of these phenomena should be studied in more detail on finite networks to help obtain more realistic and applicable results. While interesting theoretical results have been obtained for the synchronisation of coupled oscillators and much has been learnt about the role of synchronisation in neurological disorders, the gap between theory and experiment is still daunting. If we are to make the hoped-for impact in the development of treatments for diseases such as schizophrenia, there is a clear need for more accurate and sophisticated models of neural oscillations.

In finishing, it can be fairly said of biological network analysis that the need for it is clear, the challenges many and the possibilities exciting. It is hoped that this article will be of assistance to a broad community of researchers, by highlighting recent advances in the field, as well as significant issues and problems that still need to be addressed.

**Acknowledgements**

This work was partially supported by Science Foundation Ireland (SFI) grant 03/RP1/I382 and SFI grant 04/IN1/I478. Science Foundation Ireland is not responsible for any use of data appearing in this publication. The authors would like to thank the anonymous reviewers for their thorough and insightful reviews which have helped to improve the paper.

# References

[1] Albert, R., and Barabasi, A. The statistical mechanics of complex networks. Reviews of Modern Physics. 2002;74:47–97.

[2] Barabasi, L., and Oltvai, Z. Network biology: understanding the cell's functional organization. Nature Reviews - Genetics. 2004;5:101–113.

[3] Newman, M. The structure and function of complex networks. SIAM Review. 2003;45(2):167–256.

[4] Dorogovtsev, S., and Mendes, J. Evolution of networks. Advances in Physics. 2002;51:1079–1187.

[5] Alm, E., and Arkin, A. Biological networks. Current Opinion in Structural Biology. 2003;13:193–202.

[6] Albert, R., Jeong, H., and Barabasi, A. Diameter of the World-Wide Web. Nature. 1999;401:130–131.

[7] Bray, D. Molecular networks: the top-down view. Science. 2003;301:1864–1865.

[8] Alon, U. Biological networks: the tinkerer as engineer. Science. 2003;301:1866–1867.

[9] Wang, W., and Ruan, S. Simulating the SARS outbreak in Beijing with limited data. Journal of Theoretical Biology. 2004;227:369–379.

[10] Ancel Meyers, L. et al. Network theory and SARS: predicting outbreak diversity. Journal of Theoretical Biology. 2005;232:71–81.

[11] Jeong, H., Oltvai, Z., and Barabasi, A. Prediction of protein essentiality based on genomic data. ComPlexUs. 2003;1:19–28.

[12] Samanta, M., and Liang, S. Predicting protein functions from redundancies in large-scale protein interaction networks. Proceedings of the National Academy of Sciences. 2003;100(22):12579–12583.

[13] Eubank, S. et al. Modelling disease outbreaks in realistic urban social networks. Nature. 2004;429:180–184.

[14] Schnitzler, A., and Gross, J. Normal and pathological oscillatory communication in the brain. Nature Reviews - Neuroscience. 2005;6:285–295.

[15] Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences. 2001;98(8):4569–4574.

[16] Costanza, M. et al. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Research. 2000;28:73–76.

[17] Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000;28(1):27–30.

[18] Karp, P. et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research. 2005;33(19):6083–6089.

[19] Ravasz, E. et al. Hierarchical organization of modularity in metabolic networks. Science. 2002;297:1551–1555.

[20] Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 2000;403:623–627.

[21] Rain, J. et al. The protein-protein interaction map of Heliobacter Pylori. Nature. 2001;409:211–215.

[22] Giot, L. et al. A protein interaction map of Drosophila Melanogaster. Science. 2003;302:1727–1736.

[23] Li, S. et al. A map of the interactome network of the metazoan *C. elegans*. Science. 2004;303:540–543.

[24] Salgado, H. et al. The comprehensive updated regulatory network of Escherichia coli K-12. BMC Bioinformatics. 2006;7(5).

[25] Ihmels, J. et al. Revealing modular organization in the yeast transcriptional network. Nature Genetics. 2002;31:370–377.

[26] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. Network motifs in the transcriptional regulatory network of Escherichia Coli. Nature Genetics. 2002;31:64–68.

[27] Vogelstein, B., Lane, D., and Levine, A. Surfing the p53 network. Nature. 2000;408:307–310.

[28] Diestel, R. Graph Theory. Springer-Verlag; 2000.

[29] Von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002;417:399–403.

[30] Watts, D., and Strogatz, S. Collective dynamics of small-world networks. Nature. 1998;393:440–442.

[31] Barrat, A. et al. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences. 2004;101(11):3747–3752.

[32] Onnela, J. et al. Intensity and coherence of motifs in weighted complex networks. Physical Review E. 2005;71:065103.

[33] Kalna, G., and Higham, D. Clustering coefficients for weighted networks. In: Adaptation in Artificial and Biological Systems; 2006. .

[34] Barrat, A., Barthelemy, M., and Vespignani, A. Weighted evolving networks: coupling topology and weight dynamics. Physical Review Letters. 2004;92(22):228701.

[35] Mewes, H. et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Research. 2002;30(1):31–34.

[36] Gavin, A. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415:141–147.

[37] Xenarios, I. et al. DIP: the database of interacting proteins. Nucleic Acids Research. 2000;28(1):289–291.

[38] Bork, P. et al. Protein interaction networks from yeast to human. Current Opinion in Structural Biology. 2004;14:292–299.

[39] Bader, G., and Hogue, C. Analysing yeast protein-protein interaction data obtained from different sources. Nature Biotechnology. 2002;20:991–997.

[40] Ge, H. et al. Correlation between transcriptome and interactome data from Saccharomyces cerevisiae. Nature Genetics. 2001;29:482–486.

[41] Jansen, R., Greenbaum, D., and Gerstein, M. Relating whole-genome expression data with protein-protein interactions. Genome Research. 2002;12:37–46.

[42] Valencia, A., and Pazos, F. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology. 2002;12:368–373.

[43] Jeong, H. et al. The large-scale organization of metabolic networks. Nature. 2000;407:651–654.

[44] Overbeek, R. et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Research. 2000;28(1):123–125.

[45] Karp, P. et al. The EcoCyc Database. Nucleic Acids Research. 2002;30(1):56–58.

[46] Edwards, J., and Palsson, B. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics and capabilities. Proceedings of the National Academy of Sciences. 2000;97(10):5528–5533.

[47] Schilling, C. et al. Genome-scale metabolic model of Heliobacter pylori 26695. Journal of Bacteriology. 2002;184:4582–4593.

[48] Kauffman, K., Prakash, P., and Edwards, J. Advances in flux balance analysis. Current Opinion in Biotechnology. 2003;14:491–496.

[49] Edwards, J., Ibarra, R., and Palsson, B. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology. 2001;19:125–130.

[50] Covert, M. et al. Metabolic modeling of microbial strains in silico. Trends in Biochemical Sciences. 2001;26(3):179–186.

[51] Segre, D. et al. From annotated genomes to metabolic flux models and kinetic parameter fitting. Omics. 2003;7(3):301–316.

[52] Reed, J., and Palsson, B. Thirteen years of building constraint-based in silico models of Escherichia coli. Journal of Bacteriology. 2003;185(9):2692–2699.

[53] Lee, T. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002;298:799–804.

[54] Salgado, H. et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Research. 2006;34(1).

[55] Keseler, I. et al. EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Research. 2005;33(1).

[56] Segre, D., De Luna, A., Church, G., and Kishony, R. Modular epistasis in yeast metabolism. Nature Genetics. 2005;37(1):77–83.

[57] Wuchty, S. Scale-free behaviour in protein domain networks. Molecular Biology and Evolution. 2001;18(9):1694–1702.

[58] Yeger-Lotem, E. et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proceedings of the National Academy of Sciences. 2004;101(16):5934–5939.

[59] Faloutsos, M., Faloutsos, P., and Faloutsos, C. On power-law relationships of the Internet topology. In: SIGCOMM; 1999.

[60] Barabasi, L., and Albert, R. Emergence of scaling in random networks. Science. 1999;286:509–512.

[61] Yook, S., Oltvai, Z., and Barabasi, A. Functional and topological characterization of protein interaction networks. Proteomics. 2004;4:928–942.

[62] Goh, K., Kahng, B., and Kim, D. Graph theoretic analysis of protein interaction networks of eukaryotes. Physica A. 2005;357:501–512.

[63] Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. Topological and causal structure of the yeast transriptional regulatory network. Nature Genetics. 2002;31:60–63.

[64] Featherstone, D., and Broadie, K. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays. 2002;24:267–274.

[65] Bollobas, B. Random Graphs. Cambridge University Press; 2001.

[66] Albert, R., Jeong, H., and Barabasi, A. Error and attack tolerance of complex networks. Nature. 2000;406:378–382.

[67] Thomas, A. et al. On the structure of protein-protein interaction networks. Transactions of the Biochemical Society. 2003;31(6):1491–1496.

[68] Clauset, A., and Moore, C. Accuracy and scaling phenomena in Internet mapping. Physical Review Letters. 2005;94:018701.

[69] Stumpf, M., Wiuf, C., and May, R. Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences. 2005;102(12):4221–4224.

[70] Wagner, A., and Fell, D. The small world inside large metabolic networks. Proceedings of the Royal Society - B. 2001;268:1803–1810.

[71] Tong, A. et al. Global mapping of the yeast genetic interaction network. Science. 2004;303:808–813.

[72] Wagner, A. The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. Molecular Biology and Evolution. 2001;18(7):1283–1292.

[73] Yu, H. et al. Genomic analysis of essentiality within protein networks. Trends in Genetics. 2004;20(6):227–231.

[74] Maslov, S., and Sneppen, K. Specificity and stability in topology of protein networks. Science. 2002;296:910–913.

[75] Colizza, V. et al. Characterization and modeling of protein-protein interaction networks. Physica A. 2005;352:1–27.

[76] Soffer, S., and Vazquez, A. Network clustering coefficient without degree-correlation biases. Physical Review E. 2005;71:057101.

[77] Lappe, M., and Holm, L. Unraveling protein interaction networks with near-optimal efficiency. Nature Biotechnology. 2004;22(1):98–103.

[78] Goldberg, D., and Roth, F. Assessing experimentally derived interactions in a small world. Proceedings of the National Academy of Sciences. 2003;100(8):4372–4376.

[79] Dorogovstev, S., Mendes, J., and Samukhin, A. Structure of growing networks with preferential linking. Physical Review Letters. 2000;85(21):4633–4636.

[80] Krapivsky, P., Redner, S., and Leyvraz, F. Connectivity of growing random networks. Physical Review Letters. 2000;85(21):4629–4632.

[81] Bollobas, B. et al. The degree-sequence of a scale-free random graph process. Random Structures and Algorithms. 2001;18:279–290.

[82] Bollobas, B., and Riordan, O. "Mathematical results on scale-free graphs". In: Bornholdt S, Schuster H, editors. Handbook of Graphs and Networks. Wiley; 2002.

[83] Bollobas, B., and Riordan, O. Robustness and vulnerability of scale-free random graphs. Internet Mathematics. 2003;1:1–35.

[84] Bollobas, B., and Riordan, O. The diameter of a scale-free random graph. Combinatorica. 2004;24:5–34.

[85] Fox Keller, E. Revisiting "scale-free" networks. Bioessays. 2005;27:1060–1068.

[86] Volchenkov, D., Volchenkova, L., and Blanchard, P. Epidemic spreading in a variety of scale-free networks. Physical Review E. 2002;66:046137.

[87] Han, J. et al. Effect of sampling on topology predictions of protein-protein interaction networks. Nature Biotechnology. 2005;23(7):839–844.

[88] Przulj, N., Corneil, D., and Jurisica, I. Modeling interactome: scale-free or geometric. Bioinformatics. 2004;20(18):3508–3515.

[89] Penrose, M. Geometric Random Graphs. Oxford University Press; 2003.

[90] Vazquez, A. et al. Modeling of protein interaction networks. ComPlexUs. 2003;1:38–46.

[91] Chung, F. et al. Duplication models for biological networks. Journal of Computational Biology. 2003;10(5):677–687.

[92] Bhan, A., Galas, D., and Dewey, T. A duplication growth model of gene expression networks. Bioinformatics. 2002;18(11):1486–1493.

[93] Sole, R. et al. A model of large scale proteome evolution. Advances in Complex Systems. 2002;5:43–54.

[94] Berg, J., Lassig, M., and Wagner, A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evolutionary Biology. 2004;4:51.

[95] Wagner, A. How the global structure of protein interaction networks evolves. Proceedings of the Royal Society - B. 2002;270:457–466.

[96] Chung, F., and Lu, L. Coupling Online and Offline Analyses for Random Power Law Graphs. Internet Mathematics. 2004;1(4):409–461.

[97] Wasserman, S., and Faust, K. Social Network Analysis: Methods and Applications. Cambridge University Press; 1994.

[98] Jeong, H., Mason, S., Barabasi, A., and Oltvai, Z. Lethality and centrality in protein networks. Nature. 2001;411:41–42.

[99] Wuchty, S. Interaction and domain networks of yeast. Proteomics. 2002;2:1715–1723.

[100] Hahn, M., Conant, G., and Wagner, A. Molecular evolution in large genetic networks: does connectivity equal constraint? Journal of Molecular Evolution. 2004;58:203–211.

[101] Wuchty, S., and Stadler, P. Centers of complex networks. Journal of Theoretical Biology. 2003;223:45–53.

[102] Freeman, L. A set of measures of centrality based on betweenness. Sociometry. 1978;40:35–41.

[103] Goh, K. et al. Classification of scale-free networks. Proceedings of the National Academy of Sciences. 2002;99(20):12583–12588.

[104] Goh, K., Kahng, B., and Kim, D. Universal behaviour of load distribution in scale-free networks. Physical Review Letters. 2001;87(27):278701.

[105] Goh, K., Kahng, B., and Kim, D. Betweenness centrality correlation in social networks. Physical Review E. 2003;67:017101.

[106] Joy, M. et al. High-betweenness proteins in the yeast protein interaction network. Journal of Biomedicine and Biotechnology. 2005;2:96–103.

[107] Potapov, A. et al. Topology of mammalian transcription networks. Genome Informatics. 2005;16(2):270–278.

[108] Newman, M. A measure of betweenness centrality based on random walks. Social Networks. 2005;27:39–54.

[109] Koschutzki, D., and Schreiber, F. Comparison of centralities for biological networks. In: German Conference of Bioinformatics; 2004.

[110] Bonacich, P. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology. 1972;2:113–120.

[111] Bonacich, P. Power and centrality: a family of measures. American Journal of Sociology. 1987;92:1170–1182.

[112] Bonacich, P, and Lloyd, P. Eigenvector-like measures of centrality for asymmetric relations. Social Networks. 2001;23:191–201.

[113] Estrada, E. Virtual identification of essential proteins within the protein interaction network of yeast. http://arxivorg/abs/q-bioMN/0505007. 2005.

[114] Berman, A., and Plemmons, R. Non-negative matrices in the mathematical sciences. SIAM classics in applied mathematics; 1994.

[115] Horn, R., and Johnson, C. Matrix Analysis. Cambridge University Press; 1985.

[116] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems. 1998;30(1-7):107–117.

[117] Langville, A., and Meyer, C. A survey of eigenvector methods for web information retrieval. SIAM Review. 2005;47(1):135–161.

[118] Kleinberg, J. Authoritative sources in a hyperlinked environment. In: 9th ACM-SIAM Symposium on Discrete Algorithms; 1998.

[119] Estrada, E. Subgraph centrality in complex networks. Physical Review E. 2005;71:056103.

[120] Estrada, E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics. 2006;6:35–40.

[121] Freeman, L., Borgatti, S., and White, D. Centrality in valued graphs: a measure of betweenness based on network flow. Social Networks. 1991;13:141–154.

[122] Stephenson, K., and Zelen, M. Rethinking centrality: methods and examples. Social Networks. 1989;11:1–37.

[123] Latora, V., and Marchiori, M. A measure of centrality based on the network efficiency. http://arxivorg/abs/cond-mat/0402050. 2004.

[124] Gomez, D. et al. Centrality and power in social networks: a game theoretic approach. Mathematical Social Sciences. 2003;46:27–54.

[125] Estrada, E. Protein bipartivity and essentiality in the yeast protein-protein interaction network. to appear in Journal of Proteome Research. 2006.

[126] Lemke, N. et al. Essentiality and damage in metabolic networks. Bioinformatics. 2004;20(1):115–119.

[127] Schmith, J. et al. Damage, connectivity and essentiality in protein-protein interaction networks. Physica A. 2005;349:675–684.

[128] Przulj, N., Wigle, D., and Jurisica, I. Functional topology in a network of protein interactions. Bioinformatics. 2004;20(3):340–348.

[129] Ibarra, R., Edwards, J., and Palsson, B. Escherechia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002;420:186–189.

[130] Edwards, J., and Palsson, B. Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics. 2000;1:1.

[131] Duarte, N., Herrgard, M., and Palsson, B. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Research. 2004;14:1298–1309.

[132] Segre, D., Vitkup, D., and Church, G. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences. 2002;99(23):15112–15117.

[133] Stelling, J. et al. Metabolic network structure determines key aspects of functionality and regulation. Nature. 2002;420:190–193.

[134] Mahadevan, R., and Palsson, B. Properties of metabolic networks: structure versus function. Biophysical Letters. 2005;88:L07–L09.

[135] Almaas, E. et al. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature. 2004;427:839–843.

[136] Fraser, H. et al. Evolutionary rate in the protein interaction network. Science. 2002;296:750–752.

[137] Jordan, I. et al. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Research. 2002;12:962–968.

[138] Yang, J., Gu, Z., and Li, W. Rate of protein evolution versus fitness effect of gene deletion. Molecular Biology and Evolution. 2003;20(5):772–774.

[139] Pal, C., Papp, B., and Hurst, D. Communication on "Protein dispensability and rate of evolution". Nature. 2003;421:496–497.

[140] Hirsh, A., and Fraser, H. Reply to communication on "Protein dispensability and rate of evolution". Nature. 2003;421:497–498.

[141] Hirsh, A., and Fraser, H. Protein dispensability and rate of evolution. Nature. 2001;411:1046–1049.

[142] Costander, E., and Valente, T. The stability of centrality measures when networks are sampled. Social Networks. 2003;25:283–307.

[143] Zemljic B., and Hleben, V. Reliability of measures of centrality and prominence. Social Networks. 2005;27:73–88.

[144] Borgatti, S., Carley, K., and Krackhardt, D. On the robustness of centrality measures under conditions of imperfect data. Social Networks. 2006;28:124–136.

[145] Scholz, J., Dejori, M., Stetter M., and Greiner M. Noisy scale-free networks. Physica A. 2005;350:622–642.

[146] Dezso, Z., Oltvai, Z., and Barabasi, A. Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccaromyces Cerevisiae. Genome Research. 2003;13:2450–2454.

[147] Babu, M. et al. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology. 2004;14:283–291.

[148] Milo, R. et al. Network motifs: simple building blocks of complex networks. Science. 2002;298:824–827.

[149] Wuchty, S., Oltvai, Z., and Barabasi, A. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nature Genetics. 2003;35(2):176–179.

[150] Ziv E. et al. Systematic identification of statistically significant network measures. Physical Review E. 2005;71:016110.

[151] Itzkovitz, S., and Alon, U. Subgraphs and network motifs in geometric networks. Physical Review E. 2005;71:026117.

[152] Vazquez, A. et al. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proceedings of the National Academy of Sciences. 2004;101:17940–17945.

[153] Mangan, S., Zaslaver, A., and Alon, U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. Journal of Molecular Biology. 2003;334:197–204.

[154] Mangan, S., and Alon, U. Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences. 2003;100(21):11980–11985.

[155] Hayot, F., and Jayaprakash, C. A feedforward loop in transcriptional regulation: induction and repression. Journal of Theoretical Biology. 2005;234:133–143.

[156] Rosenfeld, N., Elowitz, M., and Alon, U. Negative autoregulation speeds the response times of transcription networks. Journal of Molecular Biology. 2002;323:785–793.

[157] Lahav, G. et al. Dynamics of the p53-Mdm2 feednack loop in individual cells. Nature Genetics. 2004;36(2):147–150.

[158] Vega, Y., Vàzquez-Prada, M., and Pacheco A. Fitness for synchronization of network motifs. Physica A. 2004;343:279–287.

[159] Milo, R. et al. Superfamilies of evolved and designed networks. Science. 2004;303:1538–1542.

[160] Kashtan, N. et al. Topological generalizations of network motifs. Physical Review E. 2004;70:031909.

[161] Zhang, L. et al. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. BMC Journal of Biology. 2005;4:6.

[162] Artzy-Randrup, Y. et al. Comment on "Network motifs, simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science. 2004;305:1107c.

[163] Milo, R. et al. Response to comment on "Network motifs, simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science. 2004;305:1107d.

[164] Conant, G., and Wagner, A. Convergent evolution of gene circuits. Nature Genetics. 2003;34(3):264–266.

[165] Dobrin, R. et al. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. BMC Bioinformatics. 2004;5:10.

[166] Ma., H., Buer, J., and Zeng, A. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics. 2004;5:199.

[167] Newman, M. Detecting community structure in networks. European Physics Journal B. 2004;38:321–330.

[168] Newman, M. Fast algorithm for detecting community structure in networks. Physical Review E. 2004;69:066133.

[169] Clauset, A., Newman, M., and Moore, C. Finding community structure in very large networks. Physical Review E. 2004;70:066111.

[170] Ziv, E., Middendorf, M., and Wiggins C. An information-theoretic approach to network modularity. Physical Review E. 2005;71:046117.

[171] Capocci, A. et al. Detecting communities in large networks. Physica A. 2005;352:669–676.

[172] Donetti L., and Munoz, M. Detecting network communities: a new systematic and efficient algorithm. http://arXiv:cond-mat/0404652. 2004.

[173] Girvan, M., and Newman, M. Community structure in social and biological networks. Proceedings of the National Academy of Sciences. 2002;99(12):7821–7826.

[174] Newman, M., and Girvan, M. Finding and evaluating community structure in networks. Physical Review E. 2004;69:026113.

[175] Radicchi, F. et al. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences. 2004;101:2658–2663.

[176] Fortunato S., Latora V., and Marchiori, M. Method to find community structures based on information centrality. Physical Review E. 2004;70:056104.

[177] Holmes, P., Huss, M., and Jeong, H. Subnetwork hierarchies of biochemical pathways. Bioinformatics. 2003;19(4):532–538.

[178] Dunn R., and Dudbridge, F., and Sanderson, C. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics. 2005;6:39.

[179] Palla G., et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435:814–818.

[180] Enright, A., van Dongen, S., and Ouzounis, C. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002;30(7):1575–1584.

[181] Pereira-Leal, J., Enright, A., and Ouzounis, C. Detection of functional modules from protein interaction networks. PROTEINS: Structure, Function and Bioinformatics. 2004;54:49–57.

[182] Bu, D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research. 2003;31(9):2443–2450.

[183] Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics. 2003;34(2):166–176.

[184] Segal, E. et al. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics. 2003;19(Supp 1):i273–i282.

[185] Bar-Joseph, Z. et al. Computational discovery of gene modules and regulatory networks. Nature Biotechnology. 2003;21(11):1337–1342.

[186] Schwikowski, B., Uetz, P., and Fields, S. A network of protein-protein interactions in yeast. Nature Biotechnology. 2000;18:1257–1261.

[187] Nabieva, E. et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005;21:i302–i310.

[188] Hishigaki, H. et al. Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast. 2001;18:523–531.

[189] Vazquez, A. et al. Global protein function prediction from protein-protein interaction networks. Nature Biotechnology. 2003;21(6):697–700.

[190] Karaoz, U. et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proceedings of the National Academy of Sciences. 2004;101:2888–2893.

[191] Letovsky, S., and Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003;19:i197–i204.

[192] Buck, J. Synchronous rhythmic flashing of fireflies, II. The Quarterly Review of Biology. 1988;63(3):265–289.

[193] Glass, L. Synchronization and rhythmic processes in physiology. Nature. 2001;410:277–284.

[194] Keener, J., and Sneyd, J. Mathematical Physiology. Marsden JE, Sirovich L, editors. Interdisciplinary Applied Mathematics. Springer-Verlag, New York, Inc.; 1998.

[195] Engel, A., and Singer, W. Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences. 2001;5(1):16–25.

[196] Schnitzler, A., and Gross, J. Normal and Pathological Oscillatory Communication in the Brain. Nature Reviews Neuroscience. 2005;6:285–296.

[197] Winfree, A. Biological rhythms and the behavior of populations of coupled oscillators. Journal of Theoretical Biology. 1967;16:15–42.

[198] Strogatz, S. Arthur Winfree (obituary). SIAM news. 2003;36(1).

[199] Acebrón, J. et al. The Kuramoto model: A simple paradigm for synchronization phenomena. Reviews of Modern Physics. 2005;77:137–185.

[200] Strogatz, S. Exploring complex networks. Nature. 2001;410:268–276.

[201] Kuramoto, Y. In: Arakai, H., editor. International Symposium on Mathematical Problems in Theoretical Physics. vol. 39 of Lecture Notes in Physics. Springer, New York; 1975. .

[202] Kuramoto, Y., and Nishikawa, I. Statistical macrodynamics of large dynamical systems: case of a phase transition in oscillator communities. Journal of Statistical Physics. 1987;49:569–605.

[203] Strogatz, S. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. Physica D. 2000;143:1–20.

[204] Jadbabaie, A., Motee, N., and Marahona, M. On the Stability of the Kuramoto Model of Coupled Nonlinear Oscillators. In: Proceedings of the American Control Conference, Boston, Massachusetts; 2004. .

[205] Peebles Jr., P. Probability, Random Variables, and Random Signal Principles. McGraw-Hill; 2000.

[206] Roelfsema, P., Engel, A., König, P., and Singer, W. Visuomotor integration is associated with zero time-lag synchronization among cortical areas. Nature. 1997;385:157–161.

[207] Seidenbecher, T., Laxmi, T., Stork, O., and Pape, H. Amygdalar and Hippocampal Theta Rhythm Synchronization During Memory Retrieval. Science. 2003;301:846–850.

[208] Spencer, K. et al. Abnormal Neural Synchrony in Schizophrenia. Journal of Neuroscience. 2003;23(19):7407–7411.

[209] Lachaux, J. et al. Measuring Phase Synchrony in Brain Signals. Human Brain Mapping. 1999;8:194–208.

[210] Hurtado, J., Rubchinsky, L., and Sigvardt, K. Statistical Method for Detection of Phase-Locking Episodes in Neural Oscillations. Journal of Neurophysiology. 2004;91:1883–1898.

[211] Ioannides, A. et al. Real-time neural activity and connectivity in healthy individuals and schizophrenia patients. NeuroImage. 2004;23:473–482.

[212] Pikovsky, A., Rosenblum, M., and Kurths, J. Synchronization: a universal concept in nonlinear sciences. vol. 12 of Cambridge Nonlinear Sciences. Cambridge University Press; 2001.

[213] McGraw, P., and Metzinger, M. Clustering and the synchronization of oscillator networks. Physical Review E. 2005;72.

[214] Hong, H. et al. Factors that predict better synchronizability on complex networks. Physical Review E. 2004;69:067105.

[215] Restrepo, J., Ott, E., and Hunt, B. Onset of synchronization in large networks of coupled oscillators. Physical Review E. 2005;71:036151.

[216] Ichinomiya, T. Frequency synchronization in a random oscillator network. Physical Review E. 2004;70:026116.

[217] Pastor-Satorras, R., and Vespignani, A. Epidemic dynamics in finite size scale-free networks. Physical Review E. 2002;65:035108.

[218] May, R., and Lloyd, A. Infection dynamics on scale-free networks. Physical Review E. 2001;64:066112.

[219] Hong, H., Choi, M., and Kim, B. Synchronization on small-world networks. Physical Review E. 2002;65:026139.

[220] Chung, F. Spectral Graph Theory. American Mathematical Society; 1994.

[221] Barahona, M., and Pecora, L. Synchronization in Small-World Systems. Physical Review Letters. 2002;89(5).

[222] Nishikawa, T. et al. Heterogeneity in Oscillator Networks: Are Smaller Worlds Easier to Synchronize? Physical Review Letters. 2003;91:014101.

[223] Chavez, M. et al. Synchronization is Enhanced in Weighted Complex Networks. Physical Review Letters. 2005;94:218701.

[224] Zhou, C., Motter, A., and Kurths, J. Universality in the Synchronization of Weighted Random Networks. Physical Review Letters. 2006;96:034101.

[225] Motter, A., Zhou, C., and Kurths, J. Network Synchronization, diffusion and the paradox of heterogeneity. Physical Review E. 2005;71:016116.

[226] Nishikawa, T., and Motter, A. Synchronization is optimal in nondiagonalizable networks. Physical Review E. 2006;73:065106.

[227] Varela, F. et al. The brainweb: phase synchronization and large-scale integration. Nature Reviews Neuroscience. 2001;2:229–239.

[228] O'Donnell, B. et al. EEG synchronization deficits in schizophrenia spectrum disorders. International Congress Series. 2002;1232:697–703.

[229] Rodriguez, E. et al. Perception's shadow: long-distance synchronization of human brain activity. Nature. 1999;397:430–433.

[230] Sawa, A., and Snyder, S. Schizophrenia: Diverse Approaches to a Complex Disease. Science. 2002;296:692–695.

[231] Kandel, E., Schwarz, J., and Jessell, T., editors. Principles of Neural Science. McGraw-Hill; 2000.

[232] Cover, T., and Thomas, J. Elements of Information Theory. John Wiley and Sons, Inc.; 1991.

[233] Zhou, C., and Kurths, J. Dynamical Weights and Enhanced Synchronization in Adaptive Complex Networks. Physical Review Letters. 2006;96:164102.

[234] Anderson, R., and May, R. Infectious diseases of humans: dynamics and control. Oxford University Press; 1991.

[235] Hethcote, H. The mathematics of infectious diseases. SIAM Review. 2000;42(4):599–653.

[236] Liljeros, F. et al. The Web of Human Sexual Contacts. Nature. 2001;411:907–908.

[237] Murray, J. Mathematical Biology, Volume 1. Springer-Verlag; 2002.

[238] Brauer, F., and Castillo-Chavez, C. Mathematical Models in Population Biology and Epidemiology. Springer-Verlag; 2000.

[239] Pastor-Satorras, R., and Vespignani, A. Epidemic spreading in scale-free networks. Physical Review Letters. 2001;86(14):3200–3203.

[240] Pastor-Satorras, R., and Vespignani, A. Immunization of complex networks. Physical Review E. 2002;65:036104.

[241] Barthelemy, M. et al. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of Theoretical Biology. 2005;235:275–288.

[242] Dezso, Z., and Barabasi, A. Halting viruses in scale-free networks. Physical Review E. 2002;65:055103.

[243] Hwang, D. et al. Thresholds for epidemic outbreaks in finite scale-free networks. Mathematical Biosciences and Engineering. 2005;2(2):317–327.

[244] Eguiluz, V., and Klemm, K. Epidemic threshold in structured scale-free networks. Physical Review Letters. 2002;89:108701.

[245] Boguna, M., and Pastor-Satorras, R. Epidemic spreading in correlated complex networks. Physical Review E. 2002;66:047104.

[246] Boguna, M., Pastor-Satorras, R., and Vespignani, A. Absence of epidemic threshold in scale-free networks with degree correlations. Physical Review Letters. 2003;90:028701.

[247] Moreno, Y., Gomez, J., and Pacheco, A. Epidemic incidence in correlated complex networks. Physical Review E. 2003;68:035103.

[248] Cohen, R., Havlin, S., and ben Avraham, D. Efficient Immunization Strategies for Computer Networks and Populations. Physical Review Letters. 2003;91:247901.

[249] Becker, N. et al. Controlling emerging infectious diseases like SARS. Mathematical Biosciences. 2005;193:205–221.

[250] Saramaki, J., and Kaski, K. Modeling development of epidemics with dynamic small-world networks. Journal of Theoretical Biology. 2005;234:413–421.

[251] Verdasca, J. et al. Recurrent epidemics in small-world networks. Journal of Theoretical Biology. 2005;233:553–561.

[252] Colizza, V. et al. The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the USA. 2006;103(7):2015–2020.

[253] Newman, M. Spread of epidemic disease on networks. Physical Review E. 2002;66:016128.

[254] Ganesh, J., Massoulie, L., and Towsley, D. The effect of network topology on the spread of epidemics. In: IEEE Infocom; 2005.

[255] Read, J., and Keeling, M. Disease evolution on networks: the role of contact structure. Proceedings of the Royal Society - B. 2003;270:699–708.