International Journal of Computational Science and Mathematics.
ISSN 0974-3189 Volume 3, Number 1 (2011), pp. 73-92

http://www.irphouse.com

On Some Graph Theoretical Techniques for Biological Networks

V. Yegnanarayanan

*Senior Professor, Department of Mathematics
Velammal Engineering College,
Ambattur-Red Hills Road, Chennai-600066, India*

Abstract

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. while the focus of this paper 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 analyzing 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 focused 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 coupled with modern travel patterns, underline the vital nature of this issue.

On a more theoretical level, several recent studies have indicated that networks from a broad range of application areas share common structural properties. Furthermore, a number of the properties observed in such real world networks are incompatible with those of the random graphs which had been traditionally employed as modeling tools for complex networks The latter observation naturally poses the challenge of devising more accurate models for the topologies observed in biological and technological networks, while the former further motivates the development of analysis tools for complex networks. The common structural properties shared by diverse

networks offers the hope that such tools may prove useful for applications in a wide variety of disciplines. Within the fields of Biology and Medicine, applications include the identification of drug targets, determining the role of proteins or genes of unknown function, the design of effective containment strategies for infectious diseases, and the early diagnosis of neurological disorders through detecting abnormal patterns of neural synchronization in specific brain regions. Recent advances in the development of high-throughput techniques in molecular biology have led to an unprecedented amount of data becoming available on key cellular networks in a variety of simple organisms. Broadly, three classes of bio-molecular networks have attracted most attention to date: metabolic networks of biochemical reactions between metabolic substrates; 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.

The large amount of data now available on these networks provides the network research community with both opportunities and challenges. On the one hand, 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. A major motivation for the study of biological networks is the 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 analyze 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. Also, recent work suggesting connections between abnormal neural synchronization and neurological disorders such as Parkinson's disease and Schizophrenia 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, and on which the techniques discussed throughout this paper are based, is graph theory Alongside the potential benefits of applying graph theoretical methods in molecular biology, it should be emphasized that 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 and the need to incorporate dynamics and network topology in the analysis of biological systems are just a sample of the obstacles to be overcome if we are to successfully understand the fundamental networks involved in the operation of living cells. Another important issue, which we shall discuss at various points is that the structure of biological and social networks is often inferred from sampled sub networks. The precise impact of sampling on the results and

techniques published in the recent past needs to be understood if these are to be reliably applied to real biological data. Motivated by the considerations outlined above, 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 paper 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. In this paper, we give an overview of the use of graph theoretical techniques in Biological networks. In particular, we discuss recent work on identifying and modeling the structure of bio-molecular networks, as well as the application of centrality measures to interaction networks Work on the link between structural network properties with emphasis on disease propagation.

Keywords: graphs, biological networks.

Introduction

Biological networks are abstract representations of biological systems, which capture many of their essential characteristics. In the network, molecules are represented by nodes, and their interactions are represented by edges (or arrows). The cell can be viewed as an overlay of at least three types of networks, which describes protein-protein, protein-DNA, and protein-metabolite interactions. Inherent in this description is suppression of detail: many different mechanisms of transcription regulation, for example, may be described by a single type of arrow. Furthermore, the interactions can be of different strengths, so there should be numbers or weights on each arrow. Whenever two or more arrows converge on a node, an input function needs to be specified (for example, AND or OR gates). At present, many of the connections, numbers and input functions are not known. However, something can still be learned even from the very incomplete networks currently available.

First, the network description allows application of tools and concepts developed in fields such as graph theory, physics, and sociology that have dealt with network problems before Second, biological systems viewed as networks can readily be compared with engineering systems, which are traditionally described by networks such as flow charts and blueprints. Remarkably, when such a comparison is made, biological networks are seen to share structural principles with engineered networks.

Here are three of the most important shared principles, modularity, robustness to component tolerances, and use of recurring circuit elements. The first principle, modularity, is an oft-mentioned property of biological networks. For example, proteins are known to work in slightly overlapping, coregulated groups such as pathways and complexes. Engineered systems also use modules, such as subroutines in software and replaceable parts in machines. The following working definition of a module is proposed based on comparison with engineering: A module in a network is a set of nodes that have strong interactions and a common function. A module has defined input nodes and output nodes that control the interactions with the rest of the network. A module also has internal nodes that do not significantly interact with nodes outside the module.

Modules in engineering, and presumably also in biology, have special features that make them easily embedded in almost any system. For example, output nodes should have "low impedance," so that adding on additional downstream clients should not drain the output to existing clients (up to some limit). Why does modularity exist in biological networks? It is important to realize that not all networks that evolve by tinkering are modular. A well-studied example is computer science neural networks (NNs). NNs are a set of interconnected nodes, each of which has a state that depends on the integrated inputs from other nodes. As do protein signaling networks, NNs function to process information between input and output nodes. In a way analogous to biological networks, NNs are optimized by an "evolutionary" tinkering process of adding and removing arrows and changing their weights until the NN performs a given computational goal (gives the "correct" output responses to input signals). Unlike biological networks, however, NNs are nonmodular. They typically have a highly interconnected architecture in which each node participates in many tasks. Viewed in this perspective, the modularity of biological networks is puzzling because modular structures can be argued to be less optimal than NN-style, non modular structures. After all, modules greatly limit the number of possible connections in the network, and usually a connection can be added that reduces modularity and increases the fitness of the network. This is the reason that NNs almost always display a nonmodular design.

A clue to the reason that modules evolve in biology can be found in engineering. Modules in engineering convey an advantage in situations where the design specifications change from time to time. New devices or software can be easily constructed from existing, well-tested modules. A nonmodular device, in which every component is optimally linked to every other component, is effectively frozen and cannot evolve to meet new optimization conditions. Similarly, modular biological networks may have an advantage over nonmodular networks in real-life ecologies, which change over time: Modular networks can be readily reconfigured to adapt to new conditions. The second common feature of engineering and biological networks is robustness to component tolerances. In both engineering and biology, the design must work under all plausible insults and interferences that come with the inherent properties of the components and the environment. Thus, *Escherichia coli* needs to be robust with respect to temperature changes over a few tens of degrees, and no circuit in the cell should depend on having precisely 100 copies of protein X and not 103. This point has been made decades ago for developmental systems and metabolism. The fact that a gene circuit must be robust to such perturbations imposes severe constraints on its design: Only a small percentage of the possible circuits that perform a given function can perform it robustly.

Recently, there have been detailed experimental-theoretical studies that demonstrate how particular gene circuits can be robust, for example, in bacterial chemotaxis and in fruit-fly development. The third feature common to engineering and biological networks is the use of recurring circuit elements. An electronic device, for example, can include thousands of occurrences of circuit elements such as operational amplifiers and memory registers. Biology displays the same principle, using key wiring patterns again and again throughout a network. Metabolic networks use regulatory circuits such as feedback inhibition in many different pathways.

The transcriptional network of *E. coli* has been shown to display a small set of recurring circuit elements termed “network motifs” Each network motif can perform a specific information processing task such as filtering out spurious input fluctuation , generating temporal programs of expression or accelerating the throughput of the network. Recently, the same network motifs were also found in the transcription network of yeast. It is important to stress that the similarity in circuit structure does not necessarily stem from circuit duplication. Evolution, by constant tinkering, appears to converge again and again on these circuit patterns in different nonhomologous systems, presumably because they carry out key functions. Network motifs can be detected by algorithms that compare the patterns found in the biological network to those found in suitably randomized networks. This is analogous to detection of sequence motifs as recurring sequences that are very rare in random sequences. Network motifs are likely to be also found on the level of protein signaling networks . Once a dictionary of network motifs and their functions is established, one could envision researchers detecting network motifs in new networks just as protein domains are currently detected in the sequences of new genes. Finding a sequence motif (e.g., a kinase domain) in a new protein sheds light on its biochemical function; similarly, finding a network motif in a new network may help explain what systems-level function the network performs, and how it performs it.

Will a complete description of the biological networks of an entire cell ever be available? The task of mapping an unknown network is known as reverse-engineering. Much of engineering is actually reverse engineering, because prototypes often do not work and need to be understood in order to correct their design. The program of molecular biology is reverse-engineering on a grand scale. Reverse engineering a nonmodular network of a few thousand components and their nonlinear interactions is impossible (exponentially hard with the number of nodes). However, the special features of biological networks discussed here give hope that biological networks are structures that human beings can understand. Modularity, for example, is at the root of the success of gene functional assignment by expression correlations. Robustness to component tolerances limits the range of possible circuits that function on paper to only a few designs that can work in the cell. This can help theorists to home in on the correct design with limited data. Network motifs define the few basic patterns that recur in a network and, in principle, can provide specific experimental guidelines to determine whether they exist in a given system. These concepts, together with the current technological revolution in biology, may eventually allow characterization and understanding of cell-wide networks, with great benefit to medicine. The similarity between the creations of tinkerer and engineer also raises a fundamental scientific challenge: understanding the laws of nature that unite evolved and designed systems.

Through examples of large complex graphs in realistic networks, research in graph theory has been forging ahead into exciting new directions. Graph theory has emerged as a primary tool for detecting numerous hidden structures in various information networks, including Internet graphs, social networks, biological networks, or, more generally, any graph representing relations in massive data sets. How will we explain from first principles the universal and ubiquitous coherence in the structure of these realistic but complex networks? In order to analyze these large sparse graphs, we use combinatorial, probabilistic, and spectral methods, as well as new and improved tools to analyze these networks. The examples of these networks have led us to focus on new, general, and powerful ways to look at graph theory.

Understanding complex systems often requires bottom up approach, breaking the system into small and elementary constituents and mapping out the interactions between these components. In many cases, the myriads of components and interactions are best characterized as networks. For example society is a network of people connected by various links, including friendships, collaborationships, sexual contacts or scientific co-authorships. Electronic communication relies on two very different networks: the physical network wiring the routers together (internet) and the web of homepages links by URLs. Airline, cell-phone, power-grid or business networks represent further examples of complex networks of technological, scientific or economic interest.

In biological systems networks emerge in many disguises, from food webs in ecology to various biochemical nets in molecular biology. In particular, wide range if interactions between genes proteins and metabolites in a cell are best represented by various complex networks. During the last decade, genomics has produces an incredible quantity of molecular interaction data, contributing to maps of specific cellular networks. The emerging fields of transcriptomics and proteomics have the potential to join the already extensive data sources provided by the genome wide analysis of gene expression at the mRNA and protein level.

Networks offer us a new way to categorize systems of very different origin under a single framework. This approach has uncovered unexpected similarities between the organization of various complex systems, indicating that the networks describing them are governed by generic organization principles and mechanisms. Understanding the driving forces which invest different networks with similar topological features enables system biology to combine the numerous details about molecular interactions into a single framework, offering means to address the structure of the cell as a whole.

Mathematical Preliminaries

The basic mathematical concept used to model networks is a graph. In this section, we shall introduce certain principal notations and recall some basic definitions and facts from graph theory. Furthermore, the notation and nomenclature introduced here will enable us to discuss the various biological networks encountered throughout the paper in a uniform and consistent manner. Throughout, R, R^n and $R^{m \times n}$ denote the field of real numbers, the vector space of n-tuples of real numbers and the space of m x n matrices with entries in R respectively. A^T denotes the transpose of a matrix A in $R^{m \times n}$ and A belonging to $R^{n \times n}$ is said to be symmetric if $A = A^T$. For finite sets S, T, S x T denotes the usual Cartesian product of S and T, while |S| denotes the cardinality of S.

Directed and Undirected Graphs

The graphs or networks which we shall encounter can be divided into two broad classes: directed graphs and undirected graphs. Formally, a finite directed graph, G, consists of a set of vertices or nodes, V(G), V(G) = {v_1; : : : ; v_n}; together with an edge set, E(G) which is a subset of V(G) x V(G). Intuitively, each edge (u; v) in 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 would represent genes with edges denoting the interactions between them. This would be a directed graph because, if gene A regulates gene B, then there is a natural direction associated with the edge between the corresponding nodes, starting at A and finishing at B. Directed graphs also arise in the study of neuronal networks, in which the nodes represent individual neurons and the edges represent synaptic connections between neurons.

An undirected graph, G, also consists of a vertex set, V(G), and an edge set E(G). However, there is no direction associated with the edges in this case. Hence, the elements of E(G) are simply two element subsets of 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 [47], 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 focused on the protein-protein interaction networks of various simple organisms [92, 151]. These networks describe the direct physical interactions between the proteins in an organism's proteome and there is no direction associated with the interactions in such networks. Hence, PPI networks are typically modelled as undirected graphs, in which nodes represent proteins and edges represent interactions.

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 inV(G), the set of all neighbours of u is denoted N(u) and given by N(u) = {v in V(G) : uv belongs to E(G)}.

Node-degree and the Adjacency Matrix

For an undirected graph G, we shall write deg(u) for the degree of a node u in 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) = |N (u)|. In a directed graph G, the in-degree, $deg_{in}(u)$ (out-degree, $deg_{out}(u)$) 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 $v_1,\dots,v_n$. Then the adjacency matrix, A, of G is given by $a_{ij} = 1$ if v_iv_j belongs to E(G) and 0 if v_iv_j does not belongs to E(G) --(1). Thus, the adjacency matrix of an undirected graph is symmetric while this need not be the case for a directed graph.

Paths, Path Length and Diameter

Let u, v be two vertices in a graph G. Then a sequence of vertices $u = v_1; v_2; : : : ; v_k = v$; such that for i = 1,…, k - 1: (i) v_iv_{i+1} belongs to E(G); (ii) $v_i \neq v_j$ for $i \neq j$ is said to be a path of length k -1 from u to v. The geodesic distance, or simply distance, d(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 $d(u; v) = \infty$. If for every pair of vertices, u, v in 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 d(u,v) taken over all pairs of distinct nodes. u,v in V(G) which are connected by at least one path.

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 $C_u = 2e/k(k -1)$ --: (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.

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 <f> denotes the average value of f over all nodes in the graph.

Identification and Modelling of Bio-molecular Networks

Due to recent advances in high-throughput technologies for biological measurement, there is now more data available on bio-molecular networks than ever before. This has made it possible to study such networks on a scale which would have been impossible two decades ago. In fact, large-scale maps of protein interaction networks [197, 125, 186, 67, 151, 117], metabolic networks [97, 140] and transcriptional regulatory networks [114, 177] have been constructed for a number of simple organisms. Motivated by these developments, there has been a significant amount of work done on identifying and interpreting the key structural properties of these networks in recent years. We shall give here an overview of the main aspects. In particular, we shall describe the principal graph theoretical properties of bio-molecular networks which have been observed in experimental data. We shall also discuss several mathematical models that have been proposed to account for the observed topological properties of these networks.

Structural Properties of Biological Networks

In this subsection, we shall concentrate on the following three aspects of network structure, which have received most attention in the last few years:(i) Degree distributions; (ii) Characteristic path lengths; (iii) Modular structure and local clustering properties. For each of these, we shall describe recently reported findings for protein interaction, metabolic and transcriptional regulatory networks in a variety of organisms.

Degree Distributions

Much of the recent research on the structure of bio-molecular and other real networks has focused on determining the form of their degree distributions, P(k); k = 0,1…, which measures the proportion of nodes in the network having degree k. Formally, $P(k) = n_k /n$ 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, 12] that the degree distributions of the Internet and the WWW are described by a broad-tailed power law of the form $P(k) \approx k^{-\upsilon}$ $\upsilon > 1$ --- (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, <k>, of the network, and that P(k) would decay exponentially as | k- <k>| 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 <k>, 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 scalefree. For instance, in [97], the degree distributions of the central metabolic networks of 43 different organisms were investigated using data from the WIT database [140]. 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 $2 < \upsilon < 3$. Similar studies on the degree distributions of protein interaction networks in various organisms have also been carried out. In [200], the protein interaction network of S. cerevisiae was analysed using data from four different sources. As is 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 [70]. 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 $P_{in}(k) \approx e^{-\beta k}$ [13, 74, 60].

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 [2], and is often referred to as the small world property [192]. This phenomenon has now been observed in metabolic, genetic and protein interaction networks. For instance, in [189, 97], 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 [177], 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 [187, 201, 200]. 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.

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. In [152], 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.

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 [191]. More recently, driven by the phenomenal growth of the World Wide Web, 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.

There is now a large body of data available on bio-molecular networks, and there has been considerable interest in studying the structure of these networks and relating it to biological properties in the recent past. In particular, 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 [184, 96].

In this section, we shall describe several measures of network importance or centrality that have been applied to protein interaction and transcriptional regulatory networks in the recent past. We shall place particular emphasis on the efforts to assess the biological significance of the most central genes or proteins within these networks.

Classical Centrality Measures

In this subsection, we shall discuss four classical concepts of centrality which have recently been applied to biological interaction networks:(i) Degree centrality;(ii) Closeness centrality; (iii) betweenness centrality;(iv) Eigenvector centrality.

Degree Centrality

Degree centrality is the most basic of the centrality measures.. The idea behind using degree centrality as a measure of importance in network is the following:”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 V(G) is given by $C_d(u) = deg(u)$ --(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 normalized to lie in the interval [0; 1]. 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 [4]. This naturally leads to the hypothesis that hub nodes in protein interaction networks and genetic regulatory networks may represent essential genes and proteins. In [95], 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 [178, 197]. 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 [201]. 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 nonessential 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 [201] 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 [201, 95]. However, the precise nature of this relationship is far from straightforward. For instance, using a network constructed from data published in [92, 178], the author of [194] 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 [75] 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 [75] 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.

Closeness Centrality Measures

We shall now discuss closeness centrality measures which are defined in terms of the geodesic distance, d(u, v) between nodes in a graph or network. The basic idea behind this category of measures is the following: An important node is typically “close" to, and can communicate quickly with, the other nodes in the network. In the recent paper [196], three closeness measures, which arise in the context of resource allocation problems, were applied to metabolic and protein interaction networks. The specific measures considered were eccentricity, status, and centroid value.The eccentricity, $C_e(u)$, of a node u in a graph G is given by $C_e(u)$ = max d(u, v); v in V(G) ---(6) and the centre of G is then the set C(G) = {v in V(G) : $C_e(v)$ = min $C_e(w)$ w in V(G)}---(7) Thus, the nodes in C(G) are those that minimize the maximum distance to any other node of G. The status, $C_s(u)$, of a node v is given by $C_s(u)$ = ∑ d(u; v) v in V(G) ---(8) and the median of G is then the set M(G) = {v in V(G) : Cs(v) = min $C_s(w)$ w in V(G)}--- (9). The nodes in M(G) minimize the average distance to other nodes in the network. The final measure considered in [196] 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 [196] 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, C(G), and the median, 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, 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 eccentricity 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.

Betweenness Centrality Measures

In [64], 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 V(G), let σ_{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 V(G), let V(u) denote the set of all ordered pairs, (u, v) in V(G) x V(G) such that u; v;w are all distinct. Then, the betweenness centrality of w, $C_b(w)$, is given by $C_b(w) = \sum \sigma_{uv}(w)/ \sigma_{uv}$ (u,v)in V(w)--- (10) Recently, in [99] 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 nonessential proteins.

The authors pointed out that this was not consistent with the scale-free BA model or with the more biologically motivated DD models proposed in [170, 181]. Furthermore, there was considerable variation in the value of $C_b(u)$ for proteins u with the same degree. This naturally raises the following question: if two proteins, u, v have the same degree k but $C_b(u) > C_b(v)$, is u more likely to be essential than v? However, no clear evidence to support this hypothesis was found in the data considered in [99]. In the present context, it is worth noting the work in [136] 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 new concept of betweenness centrality has yet to be applied to bio-molecular networks in a systematic way.

Eigenvector Centrality Measures

As with many of the measures considered in this section, 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 [26, 27, 191, 28]. 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 A is the adjacency matrix of a network G with $V(G) = \{v_1,\ldots,v_n\}$, and $\rho(A) = \max \mid \lambda \mid \lambda$ in $\sigma(A)$ is the spectral radius of A, then the eigenvector centrality score, $C_{ev}(v_i)$ of the node v_i is given by the ith co-ordinate, x_i, of a suitably normalized eigenvector x satisfying Ax = ρ(A)x: In the recent paper [57], 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 Before concluding our discussion of the classical centrality measures and their possible application to the identification of essential genes or proteins, it is worth noting the following points about eigenvector centrality. (i) In order for the definition above to uniquely specify a ranking of the nodes in a network it is necessary that the eigenvalue ρ(A) has geometric multiplicity one. For general networks, this need not be the case. However, if the network is connected then it follows from the Perron- Frobenius Theorem for irreducible non-negative matrices [17, 86] that this will be the case. (ii) 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 [32, 111]. The HITS algorithm for the ranking of web pages, proposed by Kleinberg [105], also relies on similar reasoning.

Graph Theoretical Approaches to Identifying Functional Modules

A graph clustering algorithm for identifying families of related nodes in networks was described in [55], where 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. The TRIBE-MCL algorithm for detecting communities of related nodes within this graph was then described. This technique is based on Markov chain clustering , and identifies communities through iterating two different mathematical operations of inflation and expansion .The core concept behind this method is that families of related nodes are densely interconnected and hence there should be more “long" paths between pairs of nodes belonging to the same family than between pairs of nodes belonging to distinct families. Subsequently, in [146] 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. Further approaches to the determination of functional modules within biological networks have been described in [149, 166]. The technique in [149] 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 [166, 165] 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 [11, 90]. The techniques described in these papers are not based on a graph theoretical analysis of network topology however; in fact, they rely on analyzing gene expression data across different experimental conditions and determining sets of genes which are regulated by common transcription factors.

Network Structure and Disease Propagation

We shall consider here the impact of network structure on disease propagation models. Given that several of the novel network properties considered in the recent past have been observed in social networks and in networks of human sexual contacts [118], it is natural to ask what effect these properties have on the spread of disease through such networks. Given the emergence of new virulent diseases such as the SARS virus and the Asian bird flu, the importance of understanding the interaction between network structure and the dynamics of disease propagation cannot be over-emphasized.. First, we shall discuss recent numerical and theoretical work on the effect of different degree distributions on the behaviour of classical epidemic models, with particular emphasis on the effect of power-law distributions on the so-called epidemic threshold. We shall then discuss extensions of this basic line of research which have attempted to take into account finite-size effects correlations between the degrees of connected nodes. Finally, we shall discuss a number of other issues pertaining to disease spread on networks, including the containment of epidemics on different network topologies and the evolution of different disease strains.

Scale-free Networks and Epidemic Thresholds

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. A detailed discussion of the properties of all of these models is well beyond the scope of the current document, and the interested reader should consult [8, 78]. Here, we shall confine our discussion to results concerned with two basic models of disease spread: the Susceptible-Infected-Susceptible or SIS model and the Susceptible-Infected-Removed or SIR model. Much of the recent work on disease propagation through networks has focussed on these two core models. 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 λ 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 υ 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 [130, 30] $dS/dt = -\lambda SI$ ---(18) ; $dI/dt = \lambda SI - \upsilon I$; $dR/dt = \upsilon I$: 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 or connection between individuals, 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, <k>, of the network. This assumption is only reasonable for networks whose degree distributions are narrow, meaning that the coefficient of variation, $C_V = (<k^2> / (<k^2> - 1))^{1/2}$ is very small.

Conclusions and Directions for Future Research

The need for a more systematic approach to the analysis of living organisms, alongside the availability of unprecedented amounts of data, has led to a considerable growth of activity in the theory and analysis of complex biological networks in recent years. Networks are ubiquitous in Biology, occurring at all levels from biochemical reactions within the cell up to the complex webs of social and sexual interactions that govern the dynamics of disease spread through human populations. Over the last few years, several core themes and questions in biological network analysis have arisen from pressing problems in Biology and Medicine. For instance, while the research on bio-molecular and neurological networks is still at a relatively early stage, a comprehensive understanding of these networks is needed to develop more sophisticated and effective treatment strategies for diseases such as Cancer and Schizophrenia. Other aspects of this line of research have been motivated by the need to determine the biological role of un annotated genes or proteins. On the other hand, at the level of social networks, future approaches to epidemic containment will need to take into account the interplay between network topology and dynamics. Our aim in this article has been to provide as comprehensive an overview as possible of the uses of Graph Theory and Network Analysis within Biology, and to point out problems in Graph Theory that arise from the study of biological networks. Specifically, we concentrated on the following five broad topics.

Structural identification and modelling of bio-molecular networks

Recent advances in high-throughput techniques have led to the construction of maps of protein- protein interaction, transcriptional regulatory and metabolic networks for a variety of organisms. Numerical investigations of the properties of these network maps, indicate that they tend to have short characteristic path lengths, high clustering coefficients and scale-free degree distributions. Motivated by these observations, mathematical models such as the Barabasi-Albert scale-free network and Duplication-Divergence models have been proposed for protein interaction and genetic networks. However, the experimental techniques on which these network maps are based are prone to high rates of false positive errors, and typically only cover a fraction of the network's nodes. The development of more accurate and reliable experimental methodologies is of course of vital importance for future research on the structure of bio-molecular networks. On a more theoretical level, two of the most significant issues that need to be addressed in this area are the sampling properties of complex networks and the impact of data inaccuracies on the identification of network statistics such as the degree distribution.

Centrality measures and essentiality in gene and protein networks

Much of the research on applying centrality measures to bio-molecular networks has focused on the prediction of gene or protein essentiality. In most of the studies discussed the centrality score of a node was found to be indicative of its likelihood to be essential. In particular, this appears to be true for degree centrality, betweenness centrality and eigenvector centrality measures. However, there is no clear evidence that the more complex centrality measures perform any better than degree centrality. A major source of open problems in this area is the robustness of centrality measures to data inaccuracies. Once again, this issue is very important for the reliable application of these techniques to biological data.

Network structure and epidemic dynamics

The interplay between epidemic dynamics and network structure is vital for understanding and containing the spread of infectious diseases. The numerical studies and mean field analyses have shown that a scale-free topology can significantly reduce the epidemic threshold, making the outbreak of epidemics more likely in networks with such a structure. Network topology also has an impact on the effectiveness of immunization schemes for containing epidemic outbreaks. In particular, for networks with a scale-free topology, the targeted immunization of nodes of high degree offers substantial improvements over uniform random immunization. Of course, the reliable identification of social network structure is vital for the practical implementation and interpretation of such results. One important direction for future research in this area is the extension of recent results to incorporate the effects of sampling and data noise on epidemic dynamics on networks and containment strategies. To finish, it is our hope that this article will be of assistance to the broad community of researchers working on the study of biological networks, by highlighting recent advances in the field, as well as significant issues and problems that still need to be addressed.

References

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

[2] R. Albert, H. Jeong, and L. Barabasi. Error and attack tolerance of complex networks. Nature, 406:378-382, 2000.

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

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

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

[6] A. Berman and R.J. Plemmon. Non-negative matrices in the mathematical sciences. SIAM classics in applied mathematics, 1994.

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

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

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

[10] F. Brauer and C. Castillo-Chavez. Mathematical Models in Population Biology and Epidemiol-ogy. Springer-Verlag, 2000.

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

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

[13] E. Estrada. Virtual identi_cation of essential proteins within the protein interaction network of yeast. http://arxiv.org/abs/q-bio.MN/0505007, 2005.

[14] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relatinships of the Internet topology. In SIGCOMM, 1999.

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

[16] L. Giot et al. A protein interaction map of drosophila melanogaster. Science, 302:1727-1736, 2003.

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

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

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

[20] H. Hethcote. The mathematics of infectious diseases. SIAM Review, 42(4):599-653, 2000.[21] R. Horn and C. Johnson. Matrix Analysis. Cambridge University Press, 1985.

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

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

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

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

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

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

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

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

[29] T. Lee et al. Transcriptional regulatory networks in saccharomyces cerevisiae. Science, 298:799-804, 2002.

[30] S. Li et al. A map of the interactome network of the metazoan C. elegans. Science, 303:540-543,2004.

[31] F. Liljeros et al. The web of human sexual contacts. Nature, 411:907, 2001.

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

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

[34] J. D. Murray. Mathematical Biology, Volume 1. Springer-Verlag, 2002.

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

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

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

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

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

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

[41] E. Segal et al. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics, 19(Supp 1):i273-282, 2003.

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

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

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

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

[46] A. Vazquez et al. Modeling of protein interaction networks. ComPlexUs, 1:38{46, 2003.[48] B. Vogelstein, D. Lane, and A. Levine. Sur_ng the p53 network. Nature, 408:307-310, 2000.

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

[48] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Molecular Biology and Evolution, 18(7):1283-1292, 2001.

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

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

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

[52] S. Wuchty. Interaction and domain networks of yeast. Proteomics, 2:1715-1723, 2002.

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

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

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

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