Imperial College London
Department of Computing

Disease Re-classification via Integration of Biological Networks

Kai Sun, Chris Larminie, Nataša Pržulj

June 2011

Abstract

Currently, human diseases are classified as they were in the late 19th century, by considering only symptoms of the affected organ. With a growing body of transcriptomic, proteomic, metabolomic and genomics data sets describing diseases, we ask whether the old classification still holds in the light of modern biological data. These large-scale and complex biological data can be viewed as networks of inter-connected elements.

We propose to redefine human disease classification by considering diseases as systems-level disorders of the entire cellular system. To do this, we will integrate different types of biological data mentioned above. A network-based mathematical model will be designed to represent these integrated data, and computational algorithms and tools will be developed and implemented for its analysis. In this report, a review of the research progress so far will be presented, including 1) a detailed statement of the research problem, 2) a literature survey on relative research topics, 3) reports of on-going work, and 4) future research plans.

Contents

List of Tables

List of Figures

1. Introduction

Most human diseases can be viewed as a consequence of a breakdown of cellular processes. However, the relationships between diseases and the molecular interaction networks underlying them remain poorly understood. As more and more transcriptomic, proteomic, metabolomic and genomics data sets become public available, it is beneficial to improve our understanding of human diseases and diseases relationship based on these new system-level biological data.

We propose to address the question of integrating different types of large scale biological data with the aim to redefine disease classification and relationships among diseases. Network-based mathematical models and computational tools will be designed and implemented for disease classification, and will hopefully lead to improvement of therapeutics in collaboration with GlaxoSmithKline (GSK).

In consideration of reliable diagnosis and treatment, an accurate classification of human disease is essential in the area of biology and medical science. Contemporary classification of human disease dates to the late 19th century, and derives from observational correlation between pathological analysis and clinical syndromes [1]. Published by the World Health Organization (WHO), the *International Statistical Classification of Diseases and Related Health Problems* (also known by the abbreviation ICD) is considered as an international standard diagnostic classification, and is used worldwide for all general epidemiological and many health management purposes [2].

However, with the identification of molecular underpinnings of many disorders, as well as the further influences of definitive laboratory tests in the overall diagnostic paradigm, this classification approach has been considered as both a lack of sensitivity in identifying preclinical disease and a lack of specificity in defining disease unequivocally [1]. A precise disease classification consistent with modern systems-level data, such as proteomic, metabolomic and genomics data, will result in better understanding of basis for disease susceptibility and environmental influence and higher therapeutic efficacy of disease treatment.

The behavior of most complex systems emerges from the orchestrated activity of many components that interact with each other through pairwise interactions [3]. The components can be abstracted as nodes, which are linked by edges indicating the interactions between these components, and these nodes and edges form a network. As most of biological data are large-scale and complex, networks have been applied to model these data, leading to the development of novel quantitative biological network analysis methods.

Hence, we propose to take a network science approach to presenting various slices of systems-level biological information in an integrated way that will allow mining of these complex data. The integrated biological data are proposed to form a hybrid network and new techniques for analyzing the network will be developed and implemented. We will use theoretical insights from graph theory, the mathematics of complicated networks, along with modern probabilistic models and scientific computing approaches for developing these new techniques. Additionally, graphlet-based approaches will be applied to the hybrid model to uncover biological function from the model's topology and structure.

In the rest of the report, we will first give a review of literature and relevant research associated with biological network analysis (Chapter 2). Relative topics such as different types of biological networks, random models, global and local properties of networks, will be fully explained. Graphlet-based methods will be recalled, along with their applications for node comparison, network comparison and network alignment.

Chapter 3 will focus on details of on-going projects. In the session biological network integration, the collection of relevant data will be covered, including detailed descriptions for some public biological databases that will be used in the project. Additionally, the structure of integrated network model will be discussed. In the session network analysis and modeling, the research work done for the response letter to paper "How threshold behavior affects the use of subgraphs for network comparison" will be presented.

Future research plans will be covered in Chapter 4. We will list the main challenges of the project, and point out the new techniques to be developed for solving these research problems. Furthermore, an outline of the dissertation will be included in this chapter.

Chapter 5 will summarize the main points of the report. Other relative work of the project, such as supplemental materials of Chapter 2 and Chapter 3, can be found in the appendix.

2. Literature Review

2.1. Data: Biological Networks

Networks have been used to represent many real-world phenomena including biological systems. These networks are commonly modeled by *graphs*. A graph is defined as a set of objects, called nodes, along with pairwise relationships that link the nodes, called edges. There are many different types of biological networks representing various biological phenomenons. One popular example is protein-protein interaction network, which models the physical interactions among proteins in the cell.

In this section, we will give a introduction of different types of biological networks. Detailed descriptions of molecular interaction networks will be given in section 2.1.1. These molecular interaction networks include protein-protein interaction network, transcriptional regulation network, metabolic network, cell signalling network and genetic interaction network. In section 2.1.2, we will present some other types of biological networks, for example, disease-gene association network and drug-target association network.

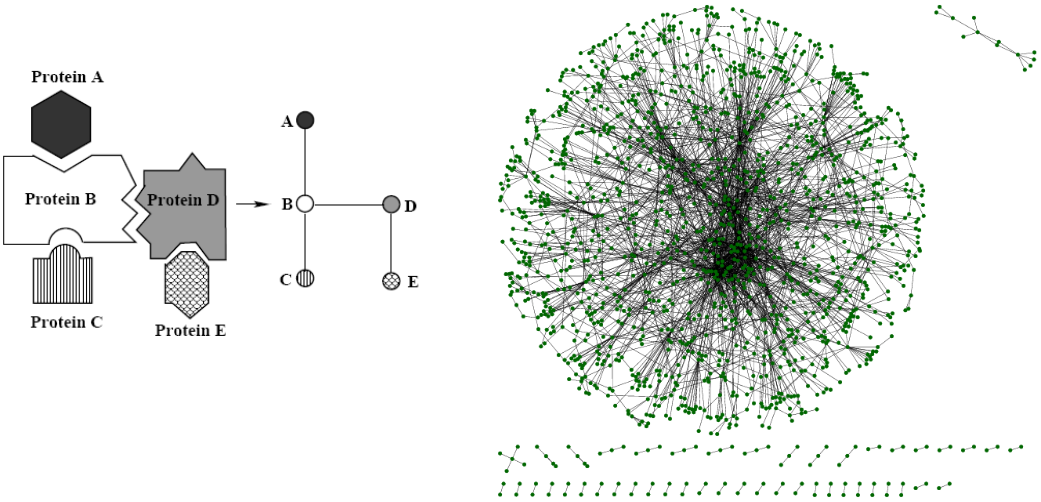
2.1.1. Molecular Interaction Networks

Molecular interaction networks have been used to model interactions between biological molecules. In these networks, nodes represent biological molecules such as genes, proteins, metabolites, etc., and edges represent physical, chemical, or functional interactions between the biological molecules. Analyses of these molecular interaction networks will lead to better understanding of entire cellular system.

Protein-protein Interaction Networks

Proteins are important macromolecules of life and understanding the collective behavior of their interactions is of biological importance [4]. Interaction between two proteins occurs when they physically bind together, often to perform biological function. As the core of the entire interactomics system, protein-protein interactions (PPIs) are of central importance for virtually every process in a living cell. Studies of PPIs can lead to further understanding of diseases, along with the development of therapeutic approaches.

Networks are used to model PPIs. In PPI networks, nodes are proteins and undirected edges exist between pairs of nodes corresponding to proteins that can physically bind to each other. Figure 2.1a demonstrates a schematic representation of a PPI network. Some stable protein interactions form protein complexes, which are groups of proteins that together perform a certain cellular function. There is evidence suggests that protein complexes correspond to dense subgraphs in PPI networks [5, 6, 7].



(a) A schematic representation of a PPI network

(b) A human PPI network

Figure 2.1.: (a) A schematic representation of a PPI network. (b) A human PPI network with 2,667 interactions amongst 1,529 proteins. PPIs are obtained from U. Stelzl's study in 2005 [8], and the PPI network is visualized by using Cytoscape 2.8.1 [9]

Nowadays, large-scale (high-throughput) experimental techniques (HT) have been applied to detect PPIs. The two techniques commonly used are yeast two-hybrid (Y2H) screening [10, 11, 12, 8, 13] and Mass Spectrometry (MS) of purified complexes [14, 15, 16, 17]. Compared to traditional small-scale biochemical techniques (SS), HT methods are more standardized, and offer an unbiased view of the entire proteome [4]. Recently, partial PPI networks of some organisms, for example, *Homo Sapiens* (human),*Saccharomyces Cerevisiae* (yeast), *Caenorhabditis Elegans* (nematode worm) and *Drosophila Melanogaster* (fruitfly), have been produced. Figure 2.1b shows a human PPI network with 2,667 interactions amongst 1,529 proteins. However, due to limitations in experimental techniques, current PPI data sets are noisy and largely incomplete. Additionally, sampling and data collection biases introduced by human make the PPI networks quite sparse with some parts being more dense than others (e.g., parts relevant for human disease) [18].

Transcriptional Regulation Networks

Transcriptional regulation networks are biochemical networks responsible for regulating the expression of genes in cells [19]. In a transcriptional regulation network, nodes represent genes, and directed edges are interactions through which the products of one gene affect those of another. Figure 2.2 illustrates how to model gene regulation as a network. As shown in the figure, if transcription factor X, which is protein product of gene X, binds regulatory DNA regions of gene Y to regulate the production rate of protein Y, then this process can be modeled as a simple network which contains a directed edge from node X to node Y.

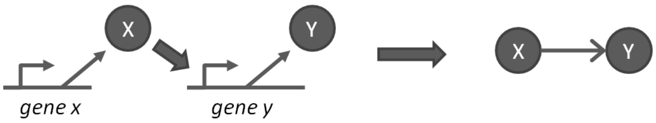


Figure 2.2.: A schematic representation of a transcription regulation network. The figure is reproduced from [20]

Metabolic Network

Metabolism is the set of biochemical reactions that allow living organisms to grow and reproduce, maintain their structures, and respond to their environments. These biochemical reactions are organized into various of metabolic pathways, which are the series of successive biochemical reactions for a specific biological function. In a metabolic pathway, one metabolite (small molecules such as Amino acids) is transformed through a series of steps into another metabolites, catalyzed by a sequence of enzymes.

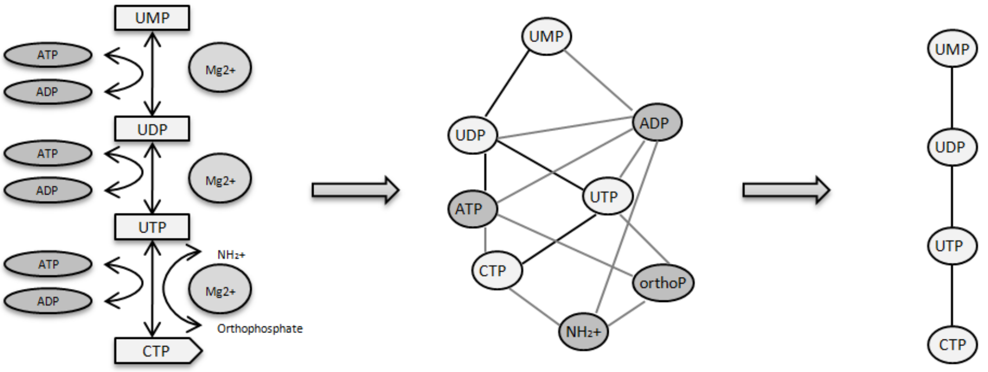


Figure 2.3.: A schematic representation of a metabolic network. The figure shows how to model a simple metabolic pathway (catalyzed by Mg^{2+}-dependant enzymes) as a network. The figure is reproduced from [3]

Metabolism can be modeled as metabolic networks. In a metabolic network, nodes correspond to metabolites and enzymes, and edges are biochemical reactions that convert one metabolite into another. The example shown in Figure 2.3 demonstrates how to model a simple metabolic pathway (catalyzed by Mg^{2+}-dependant enzymes) as a network. The metabolic pathway illustrated by the first network in figure 2.3 can be modeled as undirected graph if all interacting metabolites are considered equally. Furthermore, if co-factors are ignored, the network can be simplified as a four-node path only connecting the main source metabolites to the main products.

Cell Signalling Network

Cell signalling can be considered as a complex communication system that governs basic cellular activities. The function of communicating with the environment is achieved through a number of pathways that receive and process signals, not only from the external environment but also from different regions within the cell [21]. These pathways are ordered sequences of signal transduction reactions in a cell, and can form the cell signalling network. In the cell signalling networks, nodes are genes and the edges shows order of signal transduction reactions in the cell.

Genetic Interaction Network

Proteins or genes can be linked and form a network not only by their physical interaction, but also their functional associations. The functional association of genes refer to the phenomenon whereby the mutation of one gene affects the phenotype associated with the mutation of another gene. For example, Two non-essential genes that cause lethality when mutated at the same time form a synthetic lethal interaction. Genetic interaction networks are used modeled the functional association of gene, in which nodes correspond to genes and edges indicate functional associations of genes.

Large-scale genetic interactions have been detected in model organisms, like *Escherichia Coli* (bacterium), *Baker's yeast*, and *Schizosaccharomyces Pombe* (fission yeast). For example, Tong *et al.* constructed an yeast genetic interaction network containing 1000 genes and 4000 interactions [22]. Though there are not many studies on human genetic interactions, knowledge of the genetic interaction networks of other organisms may be relevant to our understanding of complex human diseases [23].

2.1.2. Disease Networks and Drug Networks

Besides molecular-level data such as protein-protein interactions and genetic interactions, many other biological data can be modeled as networks. For example, a human disease network can be constructed based on the associations between disorders and disease genes. Another example is drug-target association network, which is built to represent the interactions between drugs and their target proteins. These biological networks play an important role in network medicine as well as systems pharmacology.

Disease Network

Disease-gene association network is a network that connects genetic disorders and all known disease genes in the human genome. Figure 2.4 gives an example of a diseasome bipartite network, in which a disorder and a gene are connected by an edge if mutations in that gene lead to the specific disorder [24]. The disease-gene association network has two projections. The first projection is a disease network, in which two genetic disorders are connected if there is a gene that is implicated in both. The second projection is a disease gene network, in which two genes are linked by an edge if they are involved in the same disorder.

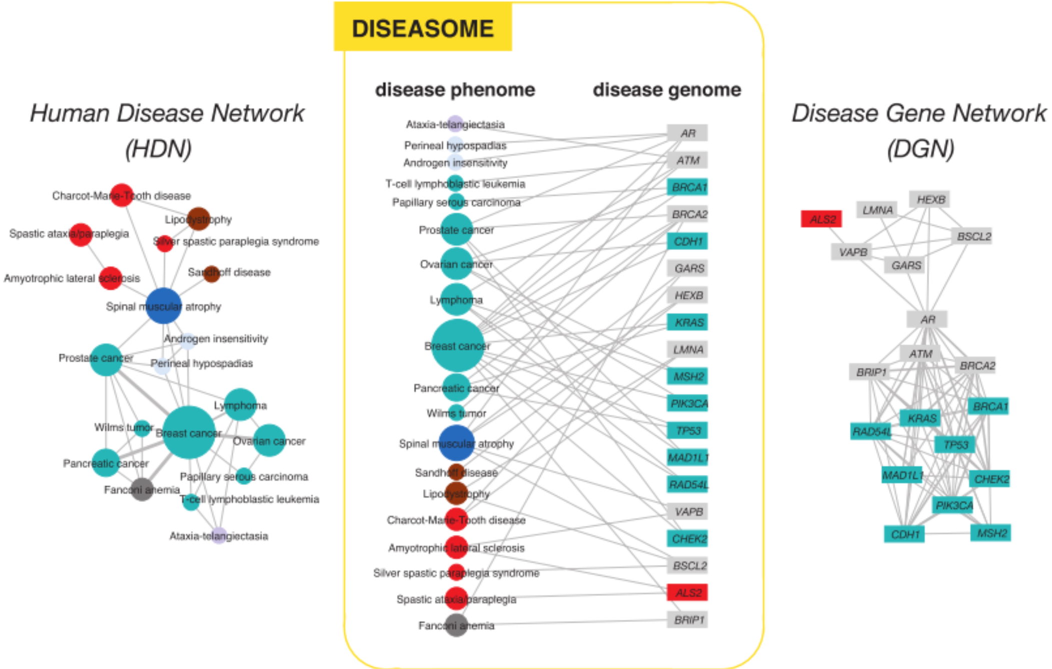


Figure 2.4.: Left: human disease network. Center: disease-gene association network. Right: disease gene network. Circles and rectangles correspond to human diseases and disease genes, respectively. In the disease-gene association network, the size of a circle is proportional to the number of genes participating in the corresponding disorder, and the color corresponds to the disorder class to which the disease belongs. In the human disease network, the width of a link is proportional to the number of genes that are implicated in both diseases. In the disease gene network, the width of a link is proportional to the number of diseases with which the two genes are commonly associated. The figure is taken from [24].

.

In Goh *et al.*'s study, the disease-gene association information was obtained from the Online Mendelian Inheritance in Man (OMIM, see section 3.1.1 for details). The disease-gene association network can also be constructed based on the information obtained by systematic literature mining methods [25]. In Li *et al.*'s study, diseases were associated to biological pathways where disease genes were enriched and linked together based on shared pathways. The human disease network constructed by using this method offers a pathway-based view of the relationship between disorders.

However, it is noticed that in the human disease network shown in figure 2.4, metabolic diseases are the most disconnected class in the network. To gain insights of the relationship between diseases and molecular interaction networks, Lee *et al.* proposed a metabolic disease network in which nodes were diseases and two diseases were linked if mutated enzymes associated with them catalyze adjacent metabolic reactions [26]. The metabolic disease network was constructed based on the metabolic reactions information obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG, see section 3.1.1 for details) and a database of Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions (BiGG), as well as disease-gene association information obtained from OMIM. Further more, Medicare records of 13,039,018 elderly patients in the US were analyzed to exam the co-occurrences of diseases which were linked in the metabolic disease network.

Drug-target Network

Most drugs act by binding to specific proteins, thereby changing their biochemical and/or biophysical activities, with multiple consequences on various functions. These specific proteins are called drug targets, and the identification of interactions between drugs and target proteins is a key area in genomic drug discovery. Network analyses of drug action have been used in field of systems pharmacology, for understanding the mechanisms underlying the multiple actions of drugs as well as drug discovery for complex diseases [27].

Drug-target association network is a network connects drugs and target proteins. Similar to the disease-gene association network, drug-target association network can have two projections. In the first projected network, nodes are drugs and drugs are connected if they share a common protein target. In the second projected network, nodes are protein targets and two protein targets are linked by an edge if they are affected by the same drug. Yıldırım *et al.* built a bipartite graph composed of US Food and Drug Administration (FDA)-approved drugs and proteins linked by drug-target binary associations, shown by figure 2.5 [28]. Lists of drugs and corresponding targets obtained from the DrugBank database (see section 3.1.1 for details) were used to construct the drug-target association network.

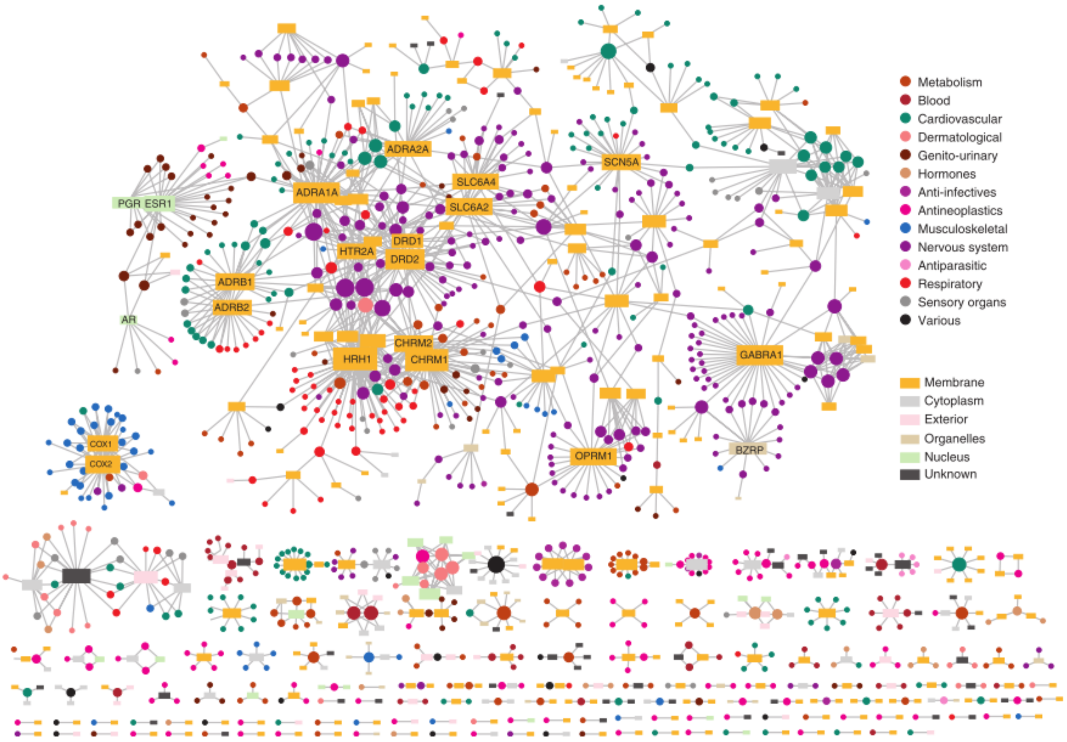
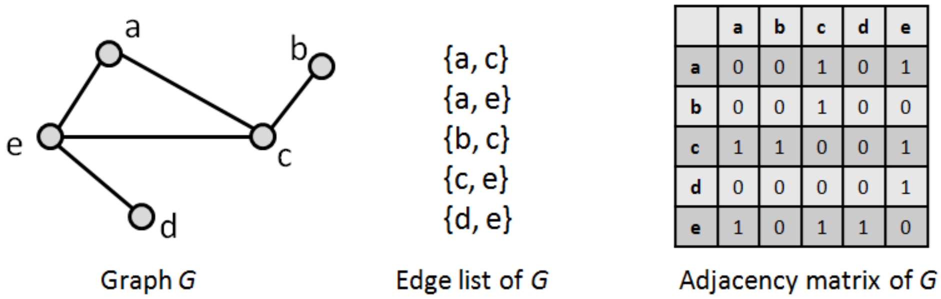


Figure 2.5.: Drug-target network. Circles and rectangles correspond to drugs and target proteins, respectively. The size of the drug node is proportional to the number of targets that the drug has, while the size of the protein node is proportional to the number of drugs targeting the protein. Drugs are colored according to their Anatomical Therapeutic Chemical Classification, and proteins are colored according to their cellular component obtained from the Gene Ontology database. The figure is taken from [28].

2.2. Methods: Graph Theory for Network Analysis

Theoretical insights from graph theory have been successfully applied to various of network analysis tasks, including network comparison, network alignment, as well as network integration. Recall that a graph is defined as a set of objects, called nodes, along with pairwise relationships that link the nodes, called edges. In graph theory, a graph is usually denoted by $G(V, E)$, where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. $|V|$ is used to denote the number of nodes, and $|E|$ is used to denote the number of edges. $V(G)$ is used to represent the set of nodes and $E(G)$ is used to represent the set of edges. There are two main standards for representing network data, namely edge list and adjacency matrix. An edge list is simply a list of edges in the network. An adjacency matrix is an $n \times n$ matrix where the entry a_{ij} is 1 corresponding to the presence of an edge connecting node i to node j and 0 corresponding to the absence of an edge connecting node i to node j. Figure 2.6 demonstrates an example of how to represent a network G by edge list and adjacency matrix.



Graph *G*

{a, c}
{a, e}
{b, c}
{c, e}
{d, e}

Edge list of *G*

	a	b	c	d	e
a	0	0	1	0	1
b	0	0	1	0	0
c	1	1	0	0	1
d	0	0	0	0	1
e	1	0	1	1	0

Adjacency matrix of *G*

Figure 2.6.: Two main standards for representing network data.

In this section, we will give an introduction to network analysis and modeling methods that are commonly applied to biological networks. In section 2.2.1, we will talk about the main computational concepts of network comparison, including global and local network properties. Then in section 2.2.2, we will describe several main network models and illustrate their use to solve real biological problems. In section 2.2.3 and section **??**, we will give an overview of the major approaches for network alignment and network integration, respectively. Finally, some software tools for network analysis will be presented in section 2.2.4.

2.2.1. Network Comparison

Network comparison aims to identify similarities and differences between data sets or between data and models. It is regarded as an essential part of biological network analysis. The task of large-scale network comparison brings on subgraph isomorphism problem, which is to determine whether a graph G contains a subgraph that is isomorphic to graph H. However, subgraph isomorphism problem is *NP-complete*, which means that no efficient algorithm is known for solving it [29].

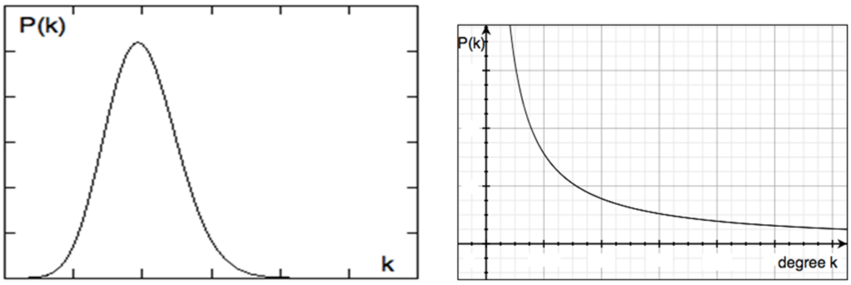
Hence, some computable heuristics that we call *network properties*, are proposed for biological network comparison. Network properties can be roughly and historically divided into two categories, *global network properties* and *local network properties.*

Global network properties

Global properties include *degree distribution*, *clustering coefficient*, *average diameter* and various forms of network *centralities*. These network properties offer an overall view of the network.

- Degree distribution. The degree of a node in a network is defined as the number of edges the node has to other nodes. Let *P(k)* be the percentage of nodes of degree k in the network. The degree distribution is the distribution of *P(k)* over all k.

The simplest network model, for example, Erdös-Rény random graph (see section 2.2.2 for details), has a Poisson degree distribution (figure 2.7a). However, it has been noticed that most networks in the real world have degree distributions that approximately follow a power law (figure 2.7b). These networks are called scale-free networks [30] (see section 2.2.2 for details).



(a) Poisson distribution

(b) Power law distribution

Figure 2.7.: Poisson distribution and power law distribution.

- Clustering coefficient. A network shows clustering if the probability of a pair of nodes being adjacent is higher when the two nodes have a common neighbor [31]. The clustering coefficient C_i of a node i is the proportion of number of edges between the nodes within its neighborhood (denoted by E_i) divided by the number of edges that could possibly exist between them,

$$C_i = 2E_i/k_i(k_i - 1), \tag{2.1}$$

  where k_i is the number of neighbors of i. The average clustering coefficient is defined as the average of the clustering coefficients of all the nodes i in the network [32],

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i. \tag{2.2}$$

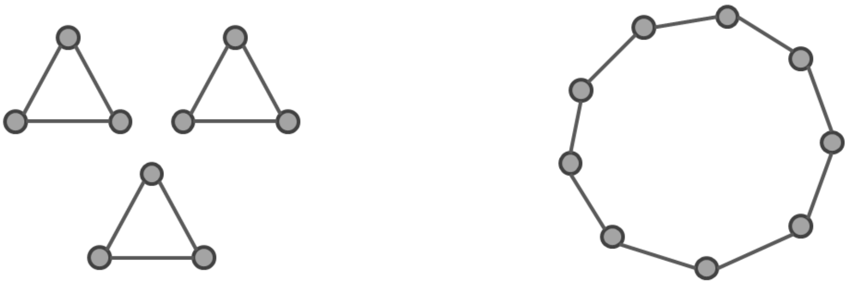
  where n is the number of nodes. The distribution of the average clustering coefficients of all nodes of degree k in the network over all k is called *clustering spectrum*.

- Average diameter. The average diameter of a network is the average of shortest path lengths over all pairs of nodes in a network. Most large-scale real-world networks have small diameters, referred to as the small-world property [32].

- Node centralities. There are various measures of the centrality of a node in a network. For example, *degree centrality* is defined as the number of edges incident upon a node, which means high-degree nodes have higher degree centrality than other nodes. Another example is *closeness centrality*, which defines nodes with short paths to all other nodes in the network have high closeness centrality. These centrality measures are proposed to determine the topological importance of nodes.

For instance, in PPI networks, nodes with high degree centrality are considered to be biologically important.

The global network properties mentioned above have been widely used for biological network comparison. However, these measures are not powerful enough to precisely describe a network's topology, as networks with exactly the same value for one network property can have very different structure. One straightforward example is, considering a network G which contains 3 triangles, and another network H which contains a 9-node circle, it is not difficult to find out the two networks have the same number of nodes, the same number of edges and the same degree distribution. However, these two networks have very different structure, as demonstrated by figure 2.8.



(a) A network contains 3 triangles.

(b) A network contains 9-node circle.

Figure 2.8.: Examples of Networks which have the same size and degree distribution but very different structure. (a) A network contains 3 triangles. (b) A network contains 9-node circle. Figures are reproduced from [18].

Furthermore, due to the incompleteness and biases in current biological data, these global network properties may even mislead the understanding of biological network topology. Though high throughput experimental methods have yielded large amounts of biological network data, these networks are currently largely incomplete. As all these global network properties are calculated based on the entire network, they do not tell us much about the structure of these incomplete networks. Additionally, the current biological data may contain biases introduced by sampling techniques which are used to obtain these biological data. For example, in bait-prey experiments for PPI detection, if the number of baits is much smaller than the number of preys, all of the baits will be detected as hubs, and all of the preys will be of low degree [18]. Therefore, it is essential to perform local statistics on these networks.

Local network properties

As mentioned above, many real-world networks share global network properties such as small-world and scale-free. Despite these global similarities, networks from different fields can have very different local structure [33]. Generally speaking, the local properties of networks include *network motifs* and *graphlets*. Both motifs and graphlets can be considered as small building blocks of complex networks, and they are widely used in biological network analysis with the aim to uncover local structure of networks.

The network motifs are those subgraphs that recur in a network at frequencies much higher than those found in randomized networks [20, 19]. In Milo *et al.*'s study, several networks including transcriptional regulation network, food webs, neuron connectivity network, electronic circuits and World Wide Web, were scanned for all possible 3-node and 4 node subgraphs (all 3-node subgraphs are listed in figure 2.9), and the number of occurrences of each subgraph was recorded.

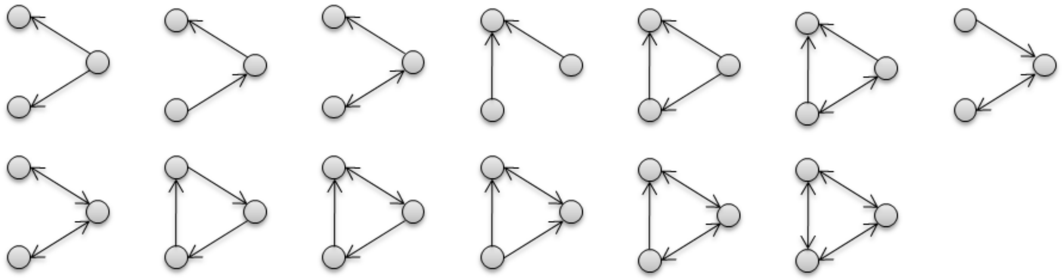


Figure 2.9.: All 13 types of three-node connected subgraphs. The figure is reproduced from [20].

The identified motifs are insensitive to noise, since they do not change after addition, deletion, or rearrangement of 20% edges to the network at random [20]. It is shown that different networks may have different motifs. Moreover, in biological networks, these motifs are suggested to be recurring circuit elements that carry out key information-processing tasks [19, 34, 35].

Based on network motifs, an approach to study similarity in the local structure of networks was proposed in [36]. A real network was compared to a set of randomized networks with the same degree sequence to calculate the *significance profile* (SP). For each 3-node and 4-node subgraph i, the statistical significance was described by the Z *score*,

$$Z_i = (N_{real_i} - < N_{rand_i} >)/std(N_{rand_i}), \tag{2.3}$$

where N_{real_i} is the number of times i appeares in the real network, and $< N_{rand_i} >$ and $std(N_{rand_i})$ are the mean and standard deviation of its appearances in the set of randomized networks. The SP was defined as the vector of normalized Z score,

$$SP_i = Z_i/(\sum Z_i^2)^{1/2}. \tag{2.4}$$

By using this method, several superfamilies of previously unrelated networks were found with very similar SPs [36].

However, it is noticed that motif-based approaches ignore subnetworks which recur at low or average frequencies in a network, and thus are not sufficient for full-scale network comparison [37]. Moreover, the detection of motifs is highly depending on the choice of the appropriate null model. For example, if the Erdös-Rény random graph (see section 2.2.2 for details) is chosen as the null model, every dense subgraph would be identified as a motif since they do not exist in the ER model network.

Graphlets have been introduced to measure of local structure of network, based on the frequencies of occurrences of all small induced subgraphs in a network. A subgraph S of graph G is induced if S contains all edges that appear in G over the same subset of nodes. A *graphlet* is defined as a small, connected and induced subgraph of a larger network [7, 37]. Figure 2.10 lists all 30 graphlets on 2 to 5 nodes. By taking into account the "symmetries" between nodes of a graphlet, there contain 73 topologically unique node types across these graphlets, called *automorphism orbits*. In figure 2.10, orbits are numbered from 0 to 72, and in a particular graphlet, nodes belonging to the same orbit are of the same shade.

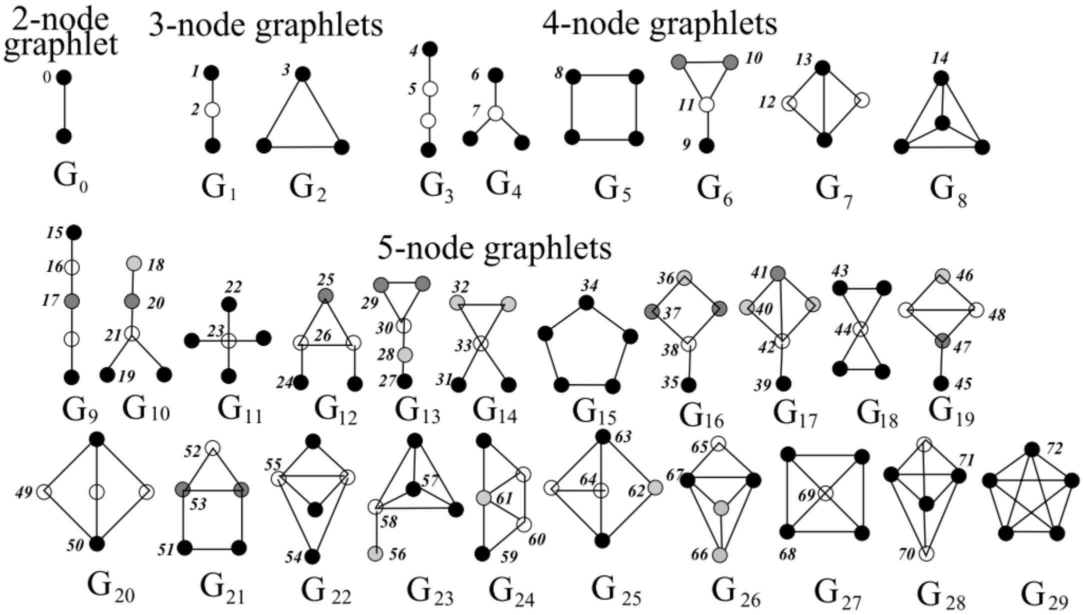


Figure 2.10.: Graphlets with 2-5 nodes $G_0, G_1, \ldots, G_{29}$. The automorphism orbits are numbered from 0 to 72, and the nodes belonging to the same orbit are of the same shade within a graphlet [37, 38].

To uncover the structure of biological network, many graphlets-based methods have been developed for different network analysis tasks. *Graphlet degree vector* (denoted by GDV), which is a generalization of node degree, has been used to measure the similarity between nodes in a network [37]. Recall that the degree of a node is defined as the number of edges incident to that node. The graphlet degree vector of a node is a 73 dimensional vector, and the ith element of GDV u_i counts how many times the node u is touched by the particular automorphism orbit i. An example demonstrating the calculation of the GDV is shown in figure 2.11. Obviously, GDV captures more structural details than node degree.

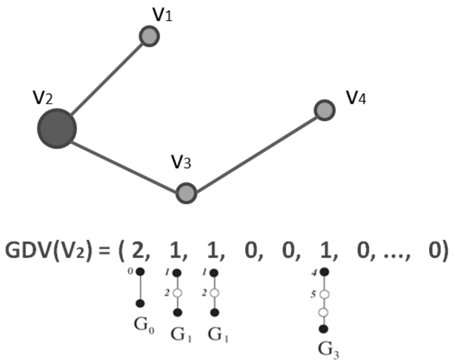


Figure 2.11.: An example demonstrates the calculation of GDV. The GDV of node V_2 is (2,1,1,0,0,1,0...,0), as V_2 is touched twice by orbit 0, once by orbit 1, orbit 2 and orbit 5.

Based on GDVs, the *signature similarity* $S(u,v)$ of two nodes u and v is computed as,

$$S(u,v) = 1 - \frac{1}{\sum_{i=0}^{72} w_i}(\sum_{i=0}^{72} w_i \times (\frac{|log(u_i+1) - log(v_i+1)|}{log(max(u_i,v_i))+2})). \tag{2.5}$$

where w_i is the weight of orbit i that accounts for dependencies between orbits [38]. Signature similarities have been applied to PPI networks to detect the similarities between proteins. It is shown that topologically similar proteins under the measure of GDVs perform the same biological function [38]. Furthermore, homologous proteins in a PPI network have a statistically significantly higher GDV similarity than non-homologous proteins [39].

As described in section 2.2.1, the degree distribution of a network is the distribution of *P(k)* over all *k*, where *P(k)* is the percentage of nodes of degree k in the network. In [37], the notion of the degree distribution was generalized to *graphlet degree distribution* (GDD). For each of the 73 automorphism orbits shown in figure 2.10, the number of nodes touching this orbit k times is counted, for each value of *k.* That means, there is an associated degree distribution for each of the 73 automorphism orbits. The spectrum of these graphlet degree distribution measures the local structure property of a network.

Let $d_G^j(k)$ be the sample distribution of the node counts for a given degree k in a network G and for a particular automorphism orbit j. The sample distribution is scaled by $1/k$ to decrease the contribution of larger degrees in a GDD, and normalized to give a total sum of 1,

$$N_G^j(k) = \frac{d_G^j(k)/k}{\sum_{l=1}^{\infty} d_G^j(l)/l}. \tag{2.6}$$

To compare two network G and H, for a particular orbit j, the distance between the two scaled and normalized distribution is defined as,

$$D^j(G,H) = \frac{1}{\sqrt{2}}(\sum_{k=1}^{\infty}[N_G^j(k) - N_H^j(k)]^2)^{\frac{1}{2}}. \tag{2.7}$$

The distance is scaled by $1\sqrt{2}$ to be between 0 to 1 [40]. The arithmetic agreement between these two network is,

$$GDD_{arith} = \frac{1}{73}\sum_{j=0}^{72}(1 - D^j(G,H)). \tag{2.8}$$

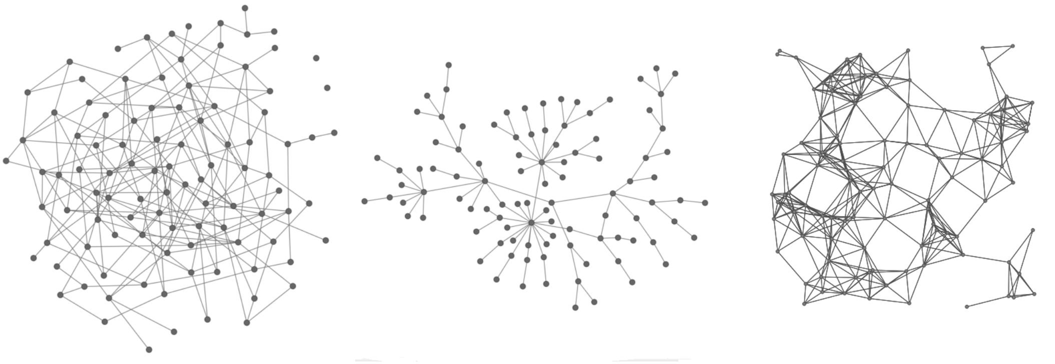
The geometric agreement between these two network is,

$$GDD_{geo} = (\prod_{j=0}^{72}(1 - D^j(G,H)))^{\frac{1}{73}}. \tag{2.9}$$

The topological similarity between two networks can be measured based on their GDD agreement. Furthermore, GDD agreements have been used to search network model that best fit the real-world networks. It is shown that most of the PPI networks are better modeled by GEO models than by ER, ER-DD or SF models [37].

2.2.2. Network Models

Network models are crucial for network motif identification, as well as finding cost-effective strategies for completing interaction maps, which is an active research topic [18]. There are several network models that are commonly used for biological network analysis, namely Erdös-Rény random graph (denoted by ER) [41, 42], Erdös-Rény random graph with the same degree distribution as the data networks (denoted by ER-DD), scale-free network (denoted by SF) [30], geometric random graph (denoted by GEO) [43], geometric gene duplication and mutation model (denoted by GEO-GD) and stickiness index-based network model (denoted by STICKY) [44].



(a) Erdös-Rény random graph

(b) Scale-free network

(c) Geometric random graph

Figure 2.12.: Examples of model networks. (a) An Erdös-Rény random graph. (b) A scale-free network. (c) A geometric random graph. Figures are taken from [18].

- Erdös-Rény random graphs. Proposed in the late 1950s, Erdös-Rény random graph is considered as the earliest random network model. In this model, a graph is constructed by connecting nodes randomly, which means edges are chosen from the $n(n-1)/2$ possible edges with the same probability p, where n is the number of nodes. Figure 2.12a gives an example of ER random network. Even though ER random models are not expected to fit the real networks well, they form a standard model to compare the data against, as many properties of ER can be proven theoretically [18, 45].

- Generalized random graphs. It's noticed that the degree distribution of a real networks always follows a power law distribution, while a ER random model has a binomial degree distribution, which can be approximated with Poisson distribution when the number of nodes is large. As a variation of the ER model, ER-DD model preserves the degree distribution of data. ER-DD models can be generated by using "stubs method" proposed in [46].

- Scale-free networks. A scale-free network is a network in which the probability of number of links per node follows a power law distribution $P(k) = k^{-\gamma}$, where k is the number of links per node and γ is a parameter whose value is typically in the range $2 < \gamma < 3$ [30]. Starting from three connected nodes, a scale-free network can be produced by preferential attachment. When a new node is adding into the network, it prefers to attach to the more connected nodes. As a result of this process, a few highly connected nodes (hubs) form in the SF model, as shown in Figure 2.12b.

- Geometric random graphs. To construct a GEO model, nodes are uniformly randomly distributed in a metric space, and two nodes are connected by an edge if the distance (can be Euclidean distance, Chessboard distance, Manhattan distance, etc.) between them is within a chosen radius r [43]. Thought GEO models follow Poisson degree distribution, which is not consistent with real networks, GEO models are shown to provide the better fit to the currently available PPI networks than ER, ER-DD and SF models, according to local network structure [37, 47, 48].

- Geometric gene duplication and mutation model. For better understanding of biological networks, especially PPI networks, GEO-GD models, which are GEO models that incorporate the principles of gene duplications and mutations, is proposed recently [49]. Starting from a small seed network, nodes are duplicated and placed at the same point in biochemical space as its parent. Controlled by natural selection, these nodes are either eliminated one, or slowly separated in the biochemical space. This process allows the child node to inherit most of the interactions of its parent node, along with some new interactions. GEO-GD models have been shown as well-fitting network models for currently available PPI networks [49].

- Stickiness index-based network model. The STICKY model is a random graph model that inserts a connection according to the degree, or "stickiness", of the two proteins involved [44]. The model is motivated by the assumption that a high degree protein has many binding domains and a pair of proteins is more likely to interact if they both have high stickiness indices, where the stickiness index of a protein can be defined as the normalized degree of a protein. The probability of an edge between two nodes is the product of their stickiness indices. It is shown that given a PPI network's underlying degree information, stickiness model better fits the network than random graphs that match the degree distribution of the network [44].

Table 2.1 indicate a comparison of global network properties (degree distribution, clustering coefficient and average diameter) between the network models described above and the real-world networks. Many real-world networks are small-world (hence they have small average diameter and high clustering coefficient) and scale-free (hence they have power law distribution). GEO-GD, STICKY models are better fit real-world networks according to these global network properties.

	Real	ER	ER-DD	SF	GEO	GEO-GD	STICKY
DD	Power law	Poisson	Power law	Power law	Poisson	Power law	Power law
CC	High	Low	Low	Low	High	High	High
AD	Small	Small	Small	Small	Small	Small	Small

Table 2.1.: Comparison of global network properties between real networks and random models. Here, DD stands for degree distribution, CC stands for clustering coefficient and AD stands for average diameter.

2.2.3. Network Alignment

Network alignment is the problem of finding similarities between the structure or topology of two or more networks. The aim of network alignment is to find the best way to fit network G into network H [50]. Figure 2.13 presents an example of an alignment of two networks H and G.

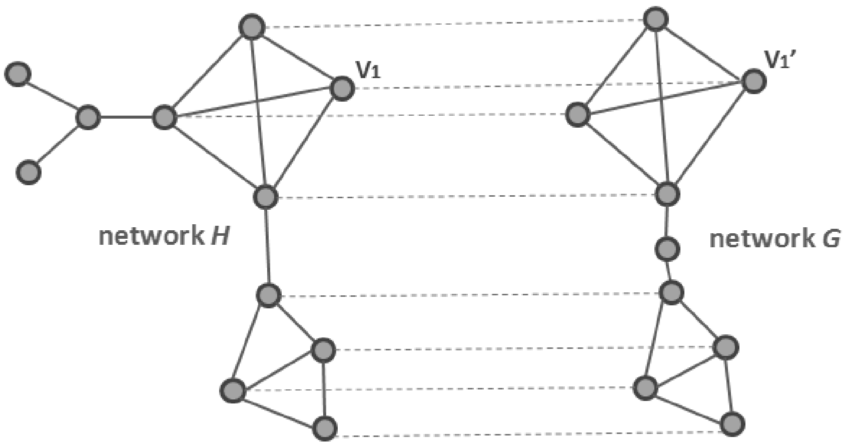


Figure 2.13.: An example of an alignment of two networks.

The alignment of biological networks is the process of comparison of two or more biological networks of the same type to identify subnetworks that are conserved across species and hence likely to present true functional modules [50]. Analogous to genomic sequence alignments, biological network alignments can be useful for knowledge transfer, as we may know a lot about some nodes in one network and almost nothing about the aligned, topologically similar nodes in the other network [18].

The network alignment problem is related to the subgraph isomorphism problem, which is NP-complete (see section 2.2.1 for details). Hence, various computable heuristics have been devised for biological network alignment. These heuristics can be roughly divided into two catalogues, namely *local network alignment* and *global network alignment*. To align two networks, a local alignment maps independently each local region of similarity, while a global network alignment uniquely maps each node in the smaller network to only one node in the larger network.

PathBLAST, which is the earliest network alignment algorithm, was developed to identify protein pathways and complexes conserved by evolution [51]. To align two PPI networks of different species, PathBLAST made used of both network topology and protein sequence similarity of two networks. PathBLAST has been used to identify orthologous pathways between yeast S. Cerevisiae and bacteria H. pylori. In [52], PathBLAST was extended to detect conserved protein clusters.

GRAph ALigner (GRAAL) is a global network alignment algorithm based solely on network topology. A seed-and-extend approach is used in GRAAL algorithm. According to [48], the steps that GRAAL aligns two networks can be summarized as the following.

- The densest parts of the networks are first aligned. GRAAL chooses a pair of nodes which has the smallest cost as the initial seed, and aligns then together. The cost of aligning a node v in network G and a node u in network H is defined as,

$$C(v,u) = 2 - ((1-\alpha) \times \frac{deg(v) + deg(u)}{max\_deg(G) + max\_deg(H)} + \alpha \times S(v,u)). \qquad (2.10)$$

where $deg(v)$ and $deg(u)$ are the degree of node v and u, respectively, $max\_deg(G)$ is the maximum degree of nodes in G, $max\_deg(H)$ is the maximum degree of nodes in H, $S(v,u)$ is the signature similarity of v and u, and α is a parameter in the range of [0,1]. The value of α controls the contribution of node degrees and node signature similarity to the cost function.

- After aligning the seed nodes, GRAAL builds the spheres of all possible radii around u and v, where a sphere $S_G(v,r)$ of radius r around node u is defined as the set of nodes $\{X\}$ that the length of the shortest path between v and $x(x \in X)$ is r.

- For each r, the nodes in $S_G(u,r)$ and $S_H(v,r)$ are greedily aligned by searching for unaligned pair of nodes which has the smallest cost according to equation 2.10.

- When all spheres of the seed nodes have been aligned, if there are unaligned nodes in both networks, GRAAL searches for a new pair of nodes as a new seed (details are described in [48]) and repeats the greedy alignment process, until each node of G is aligned to exactly one node in H.

GRAAL has been applied to align the PPI network of yeast and human, and the alignment result has shown that there are very strong enrichment for the same biological function in both yeast and human PPI networks [48].

2.2.4. Software Tools for Network Analysis

These days, various of software tools have been developed to perform different biological network analysis tasks. Some of them are used for network analysis and modeling. For example, mfinder/mDraw [53], MAVisto [54], and FANMOD [55] were developed for detecting motifs in networks, Pajek [56] was developed for analyzing global network properties, and tYNA [57] was developed for analyzing some global and local network properties. Meanwhile, some of these software tools are used for network alignment and comparison, including NetAlign [58] and PathBLAST [51], which were developed for comparing PPI networks via network alignment. Some of these tools are applied to find and visualize clusters in networks, such as CFinder [59].

One network visualization and analysis software is Cytoscape [9]. It is an open source bioinformatics platform, and these year, it has become one of the most commonly used network analysis tool in the world. The core distribution of Cytoscape provides a basic set of features for data integration and visualization. Additionally, there are many additional features are available as plugins. Though Cytoscape is originally developed for biological research, it can also be used for analyzing other types of networks, such as social networks.

GraphCrunch [60] is an open source software tool for analyzing large biological and other real-world networks and comparing them against random graph models. It can generate random networks with the number of nodes and edges within 1% of those in the real-world networks for user-specified random graph models. The generators of ER, ER-DD, GEO, SF, and STICKY models have been implemented in the software. GraphCrunch can be used to evaluate the fit of a variety of network models to real-world networks, with respect to a series of global network properties and local network properties. The global network properties implemented in GraphCrunch include degree distribution, clustering coefficient, clustering spectrum, average diameter, spectrum of shortest path lengths. The local network properties implemented are RGF-distance [61] and GDD agreement. GraphCrunch is available at `http://www.ics.uci.edu/~bio-nets/graphcrunch/`.

GraphCrunch 2 [62] is an update version of GraphCrunch 2. Besides the model networks implemented in GraphCrunch, GEO-GD model and SF-GD model are also implemented in GraphCrunch. Also, it implements GRAAL algorithm for network alignment, as well as an algorithm for clustering nodes within a network based solely on their topological similarities. GraphCrunch 2 is available at `http://bio-nets.doc.ic.ac.uk/graphcrunch2/`.

3. Methodology and On-going Work

This chapter will focus on details of on-going projects. In section 3.1, the collection of relevant data will be covered, along with detailed descriptions for the biological databases that will be used in the project. Additionally, the structure of the integrated molecular interaction network and the hybrid model will be discussed. In section 3.2, the research work done for the response letter to paper "How threshold behavior affects the use of subgraphs for network comparison" will be presented.

3.1. Biological Network Integration

Network integration is the process of combining several networks, encompassing interactions of different types over the same set of elements, to study their interrelations [63]. With the development of high-throughput experimental technique, a growing body of biological data is identified and various biological databases becomes public available. The information contained in these databases can be used to construct different types of biological networks, as mentioned in section 2.1.1 and section 2.1.2. Because each type of these biological network lends insight into a different slice of biological information, integrating different network types may paint a more comprehensive picture of the overall biological system under study [63].

3.1.1. Data Collection

Aiming to redefine human disease classification and relationships between diseases, we propose to integrate different types of large-scale biological information into a hybrid network model. These biological data include molecular interaction networks, disease-gene association information, drug-target association information and electronic patient medical records. Furthermore, the molecular interaction network includes transcriptional regulation network, metabolic network, cell signaling network, protein-protein interaction network and genetic interaction network.

Different types of biological information have been collected from different databases, as listed in table 3.1. Note that though some of the databases contain information for many model species, currently we only consider the biological information of human.

Database	Biological information contained in the database
BioGRID	Protein-protein interaction network, genetic interaction network
HPRD	Protein-protein interaction network
KEGG	Transcriptional regulation network, metabolic network, cell signaling network
OMIM	Disease-gene association information
Orphanet	Disease-gene association information, drug-target association information
DrugBank	Drug-target association information
ChEMBL	Drug-target association information
THIN	Electronic medical records

Table 3.1.: Databases planned to be integrated into the hybrid network model. BioGRID stands for the Biological General Repository for Interaction Datasets, HPRD stands for Human Protein Reference Database, KEGG stands for Kyoto Encyclopedia of Genes and Genomes, OMIM stands for Online Mendelian Inheritance in Man, and THIN stands for The Health Improvement Network.

Biological General Repository for Interaction Datasets (BioGRID)

The Biological General Repository for Interaction Datasets (BioGRID) [64] is a freely accessible biological database that contains physical interactions (PPIs) and genetic interactions of major model organism species. The physical and genetic interactions in BioGRID are curated from focused studies reported in the primary literature, and are updated monthly.

In BioGRID 3.1.74 (released in March 2011), 35,386 non-redundant protein-protein interactions among 8,920 human proteins are recorded. However, due to the technical difficulty in detecting human genetic interactions, the number of human genetic interactions recorded in BioGRID 3.1.74 is only 493, which is obviously not sufficient. Thus, network alignment methods may be applied to transfer the knowledge of the genetic interaction networks of other model organisms, like yeast, to the human genetic interaction network.

Human Protein Reference Database (HPRD)

Human Protein Reference Database (HPRD) [65] is a public accessible human protein database developed by scientists from the Institute of Bioinformatics in Bangalore, India and the Pandey lab at Johns Hopkins University in Baltimore, USA. HPRD includes protein-protein interactions, post-translational modifications, enzyme-substrate relationships and disease associations, and all these information is manually extracted from the literature by expert biologists [65].

The current release of HPRD is HPRD 9 (released in April 2010). It contains 37,039 non-redundant physical interactions among 9,465 human proteins.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

The biological information of transcriptional regulation networks, metabolic networks and cell signalling networks are collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) [66]. This public available database has been developing by scientists in Kyoto university and the University of Tokyo since 1995, and nowadays, KEGG has become one of the most widely used biological databases in the world.

The molecular interaction networks can be construct from KEGG PATHWAY database. KEGG PATHWAY contains pathway maps, which are manually created from published materials, of metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and drug development. There are 236 human pathway maps recorded in the latest release of KEGG (up to June 2011).

Online Mendelian Inheritance in Man (OMIM)

The disease-gene association network can be constructed from Online Mendelian Inheritance in Man (OMIM) [67]. Mendelian Inheritance in Man was started in the early 1960s, and its online version OMIM was created by a collaboration between the National Library of Medicine and Johns Hopkins University in 1995.

OMIM is a comprehensive knowledgebase of human genes and genetic disorders. It consists of overviews of genes and genetic phenotypes, particularly disorders, and is useful to students, researches, and clinicians [67]. Up to June 2011, OMIM has 13,642 entries describing genes with known sequence and 6727 entries describing phenotypes.

Orphanet

Orphanet is a database dedicated to information on rare diseases and orphan drugs. It was created upon request of the French Ministry of health and the National Institute of Health and Medical Research, with the aim to improve management and treatment of rare diseases. The rare disease information contained in Orphanet is listed in table 3.2.

Data	Availability	Description
List of rare diseases	Free	List including preferred name, synonyms, alpha number.
Identity card of diseases	Free	Table including Orpha number of the disease, MIM number, ICD-10 code, etc.
Classification of rare diseases	Not free	The clinical classification of rare diseases.
Table of causative genes	Not free	Table with Orpha number of the disease and linked causative genes.
Orphan drugs	Free	Table with Orpha number of the diseases for which the substance is indicated.

Table 3.2.: Rare disease information contained in Orphanet. Note that only the data relative to the project is listed in the table.

Currently there are 5,954 rare diseases recorded in Orphanet. 2,365 diseases are linked to 2,364 genes, and more than 200 diseases are linked to 875 substances.

DrugBank

The DrugBank database is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information [68], released by University of Alberta. These days, DrugBank has become widely used by pharmacists, medicinal chemists, pharmaceutical researchers, clinicians, educators and the general public [69].

The latest version DrugBank 3 (released in January 2011) includes 6,825 drug entries (1,431 FDA-approved small molecule drugs, 134 FDA-approved biotech (protein/peptide) drugs, 83 nutraceuticals and 5,210 experimental drugs). 4,434 non-redundant protein (i.e. drug target, enzyme, transporter, carrier) sequences are linked to the drug entries.

ChEMBL

ChEMBL is another public accessible database that contains drug-target association information. Established by the European Bioinformatics Institute (EBI), ChEMBL contains information on protein targets and their associated bioactive small molecules. Current release ChEMBL-09 (released in February 2011) contains 658,075 drug-like compounds and their protein targets.

The Health Improvement Network (THIN)

The Health Improvement Network (THIN) is a research database that contains electronic patient medical records. These records cover more than three million anonymized patients in the UK. The THIN database is developed by In Practice Systems Ltd (INPS) and CSD Medical Research. The content of the THIN database can be organized into the following categories, as shown in table 3.3

Type of records	Desciption
Patient records	Information on patient characteristics and registration details
Medical records	Information on symptoms, diagnoses and interventions
Therapy records	Information on details of prescriptions issued to patients
AHD records	Information on preventative health care immunizations and test results
Consult records	Information on consultation details
Staff records	Information on staff (clinician, nurse, etc.) details
PVI records	Information on postcode-based socioeconomic, ethnicity and environmental indicators

Table 3.3.: Catalogues of THIN database. Here, AHD stands for additional health data, and PVI stands for postcode linked variables

Though the THIN database is not public available, we have collected some sample data from CSD EPIC (see appendix A for details of the sample data). Table 3.4 lists the statistics of sample data, compared to the whole THIN database. The sample data contains approximately 0.02% of the whole THIN database.

Type of records	Number of records in sample data	Number of records in THIN
Patient records	2,000	9.15 million
Medical records	105,798	454 million
Therapy records	203,666	697 million
AHD records	140,320	644 million
Consult records	200,566	763 million
Staff records	3,645	(no statistics)
PVI records	4,016	(no statistics)

Table 3.4.: Statistics of THIN sample data, compared to the whole dataset.

ICD-10

The 10th revision of International Statistical Classification of Diseases and Related Health Problems (ICD-10) is considered as an international standard diagnostic classification, and is used worldwide for all general epidemiological and many health management purposes [2].

Different from the other collected databases, ICD-10 is not used to construct the integrated network, but to evaluate and validate our disease re-classification results. As ICD-10 is based on diagnosis and symptoms, and our new disease classification will be based on system-level biological networks, the similarity and difference between these two disease classification may lead to better understanding of human diseases.

Relationship among databases

The databases mentioned above are not isolated from each other. They can be integrated together based on the biological information they share. For example, OMIM and BioGRID can be linked by human genes, as genes are contained in both databases. Figure 3.1 illustrates the relationship among these databases.

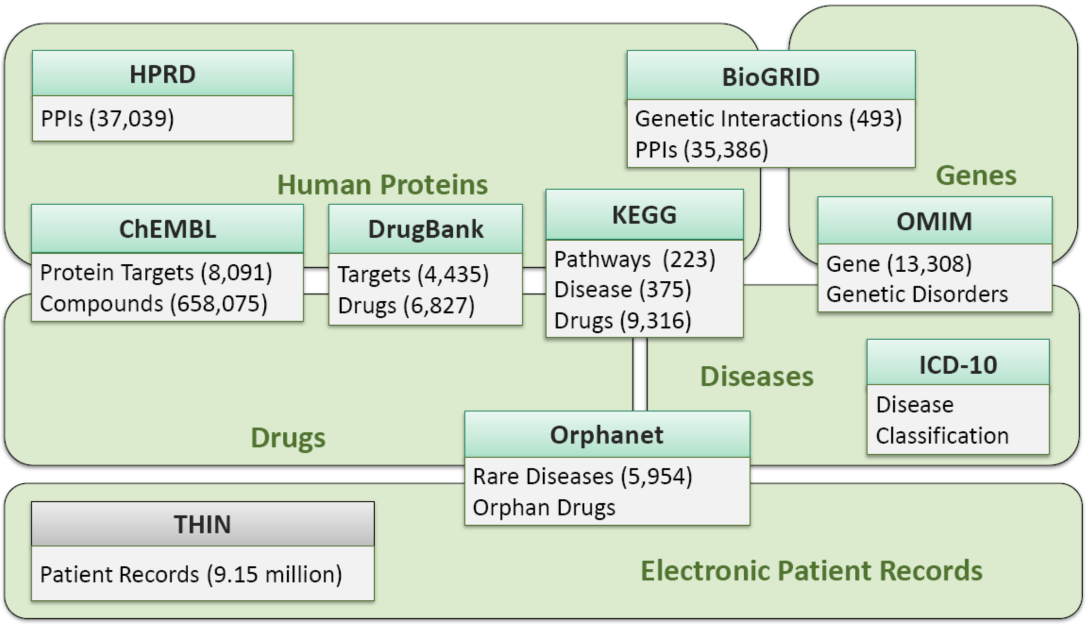


Figure 3.1.: Relationship among different types of databases. The large blocks are entities that will be integrated into the hybrid network, and the small blocks stand for the databases. The block of THIN is shaded as grey to indicate that we have not collected the data. A database is placed across entities if the database contains information on these entities. For example, ChEMBL is put between Human Proteins and Drugs, as it contains information on protein targets (belongs to human proteins) and compounds (belongs to drugs). The numbers in brackets show the statistics of the databases.

Besides the biological information mentioned above, some other data sources are also considered to be integrated into the integrated network. One potential data source is the human disease network obtained by literature mining methods [25]. Scientific literature remains a major source of valuable information, hence tools for mining such data and integrating it with other sources are of vital interest and economic impact [70]. Other databases such as side effect resource (SIDER) and Reactome may be beneficial to the project as well. SIDER [71] is a public data source that connects 888 drugs to 1,450 side effects (phenotypic responses of the human organism to drug treatment), and Reactome [72] is an expert-authored, peer-reviewed knowledgebase of reactions and pathways that contains 5234 human proteins, 3,958 protein complexes, 4,247 reactions and 1,116 pathways in it's current release.

3.1.2. Integration of Molecular Interaction Networks

Disease can be considered as the result of a modular collection of genomic, proteomic, metabolomic, and environmental networks that interact to yield the pathophenotype [1]. An understanding of the functionally relevant genetic, regulatory, metabolic, and protein-protein interactions in a cellular network will play an important role in understanding the pathophysiology of human diseases [73]. Hence, we consider diseases as systems-level disorders of the entire cellular system, and propose to improve our understanding of human diseases and diseases relationship based on system-level molecular data.

Therefore, to construct the hybrid model, firstly we plan to build a molecular interaction network, by integrating transcriptional regulation network, cell signalling network, metabolic network, protein-protein interaction network and genetic interaction network.

A schematic representation of the integrated molecular interaction network is shown by figure 3.2. This integrated network contains two types of nodes, corresponding to genes (or their protein products) and metabolites, and five types of edges, corresponding to gene regulation, cell signalling, metabolism, protein-protein interaction and genetic interaction (genetic interactions are not shown in figure 3.2 as there are no sufficient human genetic interaction data available so far). Furthermore, in this network, some of the edges are undirected, such as protein-protein interaction, while some of the edges are directed, such as transcriptional regulation. The integrated molecular interaction network then forms the bottom layer of the hybrid model (see section 3.1.3 for details).

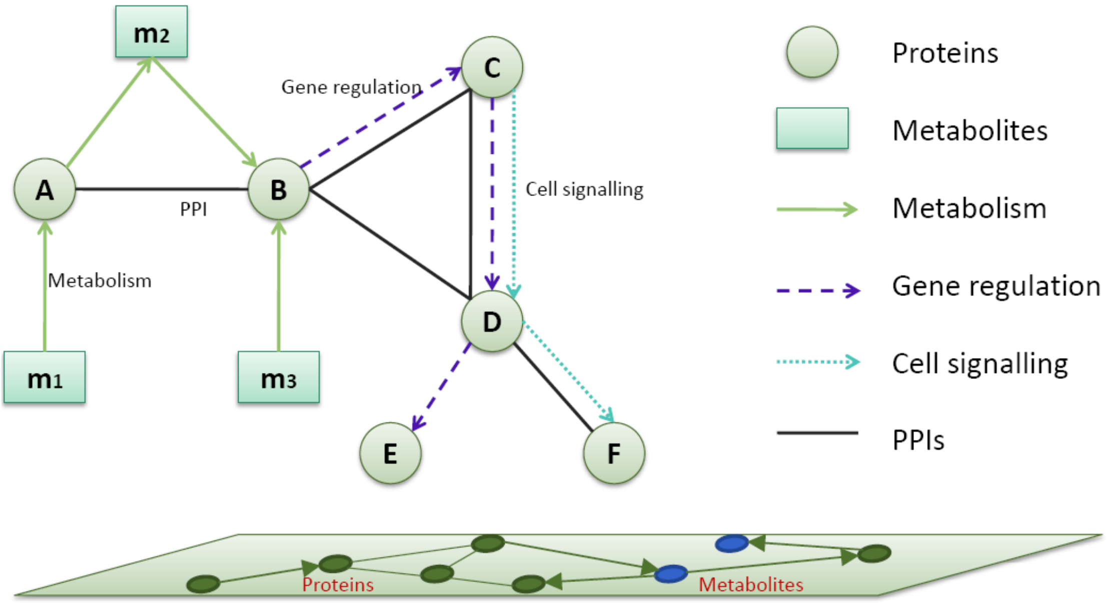


Figure 3.2.: Schematic representation of a possible integration of molecular interaction network. Here, circles are genes (or their protein products), and rectangles are metabolites. Different types of edge correspond to different interactions between genes (or genes and metabolites). Directed edges and undirected edges both appear in the network. This integrated molecular interaction network forms the bottom layer of the hybrid model.

We propose extend the graphlet-based methods to analyze this integrated molecular interaction network. Current graphlet-based methods have been only applied to undirected networks which contain only one type of nodes and one type os edges. Hence, new techniques need to be developed for analyzing complex networks (see section 4.1.1 for details).

3.1.3. Structure of Hybrid Network Model

As mentioned in section 2.1.2, the human disease network was built based on the association between disorders and disease genes. However, the knowledge of disease-gene association is not sufficient for understanding underlying mechanisms of human diseases, as human diseases are affected by a complex cellular system including genomic, proteomic, metabolomic, etc.

To gain system-level understanding of human diseases, we propose to map the disease network to the integrated molecular interaction network, as well as the drug network and electronic patient records. This leads to the design of the hybrid model, which is a multi-layer network, as shown in figure 3.3.

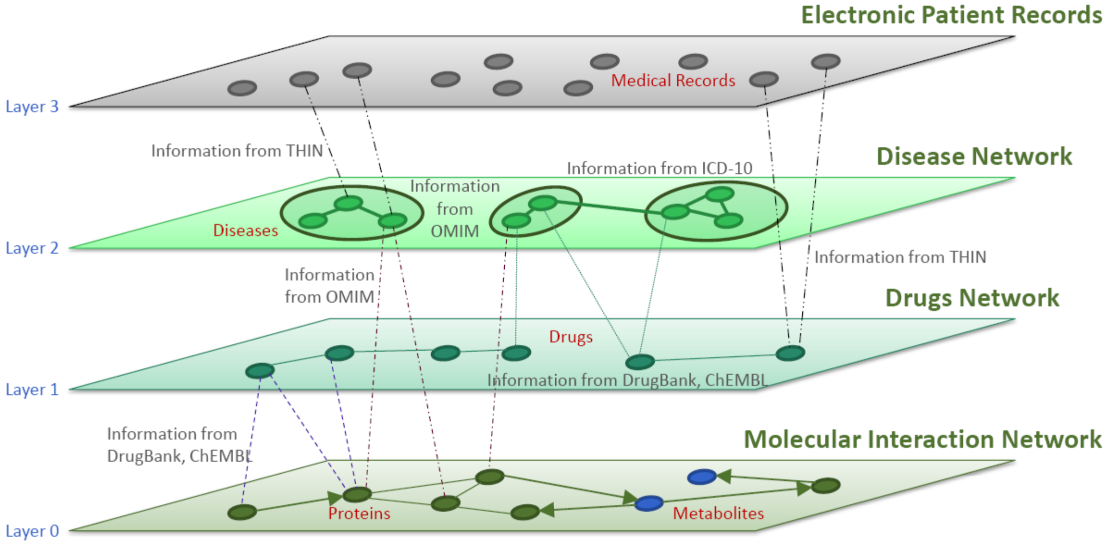


Figure 3.3.: Structure of the proposed hybrid model. Four layers will be contained in the hybrid model, namely molecular interaction network, drug network, disease network, and patient records (from bottom to up). Green nodes in layer 0 are genes (or their protein products), while blue nodes are metabolites. The black circles in layer 2 indicate the clustering of diseases, namely disease classification.

Layer 0 of the hybrid model is the integrated molecular interaction network discussed in section 3.1.2, in which nodes are proteins and metabolites, undirected edges stand for PPIs and directed edges stand for other interactions such as cell signalling or transcriptional regulation. Layer 1 is a drug network, in which nodes are drugs and two nodes are connected by an edge if these drugs share a common protein target. Nodes in the molecular interaction network and the drug network can be linked by edges across layers, based on the drug-target association information obtained from DrugBank and ChEMBL. Layer 2 of the hybrid model is the disease network, in which nodes are diseases and two diseases are connected by an edge if they are caused by a same gene. Similarly, nodes in the disease layer can be linked to the nodes in the molecular interaction network, based on the information obtained from OMIM and Orphanet. The upper layer of the hybrid model consists of numbers of electronic patient records. These patient records can be mapped to the disease network according to the medical records in THIN, as well as the drug network according to the therapy records.

Analysis of the hybrid model will bring us better understanding of the interrelationships among cellular, disease, drug and patient records. Each disease in the hybrid model is associated with a subnetwork that contains biological information from different layers. The similarity between two diseases will be calculated not only based on the topological neighborhood in the disease network, but also based on the integrated topological connectivity between network on different layers (see section 4.1.1 for details).

3.2. Network Analysis and Modeling

In the study [37], a systematic measure of structural similarity between large networks was introduced based on *graphlet degree distribution* (GDD, see section 2.2.1 for details). This measure is called GDD agreement (denoted by GDDA in the rest of this section), which can be either arithmetic or geometric. GDDA has been used to search network model that better fit the PPI networks. In [37], 14 PPI networks of the eukaryotic organisms were compared with random network models (ER, ER-DD, GEO, SF), and the GDDA scores suggested that GEO model better fit the PPI networks than other models.

This novel network comparison method has raised the interest of the public. In [74], Rito *et al.* provided a method for assessing the statistical significance of the fit between random graph models and biological networks based on non-parametric tests, and examined the use of GDDA. They concluded that the GDDA score was unstable in the graph density region between 0 and 0.01, which encompassed most of the PPI networks currently available. Furthermore, they found that none of the theoretical models considered in their study (ER, ER-DD and GEO-3D) fitted the PPI data according to their statistical method.

Since we propose to apply graphlet-based methods to analyze the integrated biological network (see section 3.1 for details), it is essential for us to validate the use of these methods, as well as understand the performance of them. We notice that the PPI networks analyzed in [74] were obtained from early studies, hence all these PPI networks were old, small and sparse. To examine whether Rito *et al.*'s conclusion still holds for latest PPI networks, we apply their methods to the latest PPI networks of yeast, fruitfly, nematode worm and human. Our results show that these latest PPI networks are not in the unstable region of GDDA, and we validate that GDDA is appropriate for analyzing these PPI networks.

In this section, we will first show how GDDA is used for searching network model that better fit the PPI networks. Then we will list the latest PPI networks we use in this project, comparing with the PPI networks analyzed in [74]. To identify the unstable region of GDDA, we calculate and visualize the empirical distribution of GDDA. Finally, we will assess the fit between model networks and latest PPI networks.

Note that in the rest of this section, GDDA is referred to arithmetic GDDA.

3.2.1. Using GDDA for Network Comparison

GDDA has been applied to many network comparison tasks. As a first step, we reproduce the experiment results in [37] to see how GDDA is used to search network model that better fit the real-world networks. Figure 3.4 shows the GDDA scores between 14

PPI networks of the eukaryotic organisms *Saccharomyces Cerevisiae* (yeast), *Drosophila Melanogaster* (fruitfly), *Caenorhabditis Elegans* (nematode worm), and *Homo Sapiens* (human), and five network models, namely ER, ER-DD, GEO, SF and STICKY (noted that STICKY model was not used in [37]). Points in the figure indicate the averages of GDD agreements between 25 model networks and the corresponding PPI networks, and the error bars represent one estimated SD below and above the average point. Our results are consistent with [40]. It shows the highest GDD agreement for the STICKY and GEO model, followed by SF, ER-DD and ER, which means STICKY and GEO model better fit the PPI data than other network models.

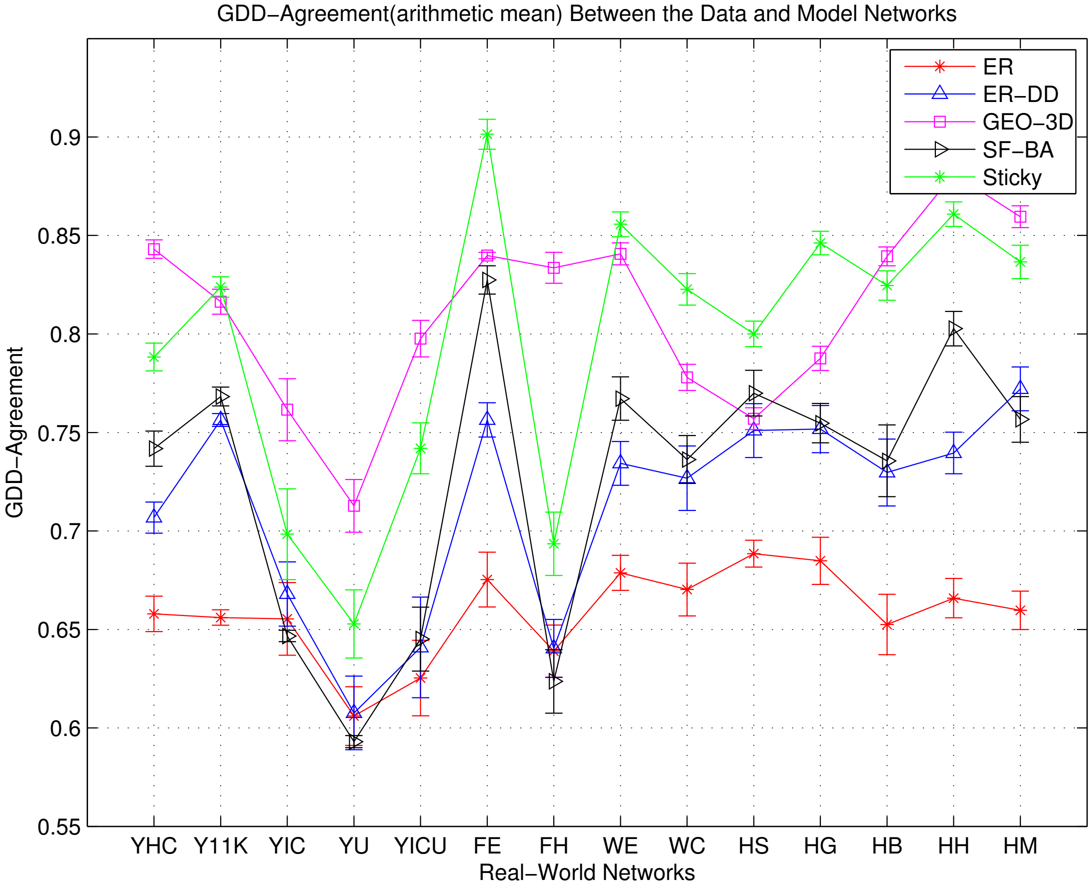


Figure 3.4.: Agreements between the 14 PPI networks and their corresponding model networks. The 14 PPI networks are (from left to right): high-confidence yeast PPI network obtained from [75] 'YHC', the top 11,000 PPIs obtained from [75] 'Y11K', core yeast PPI network obtained from [10] 'YIC', yeast PPI network obtained from [12] 'YU', the union PPI network of [10] and [75] 'YICU', fruitfly PPI network obtained from [76] 'FE', high-confidence fruitfly PPI network obtained from [76] 'FH', worm PPI network obtained from [77] 'WE', core worm PPI network obtained from [77] 'WC', human PPI network obtained from [8] 'HS', human PPI network obtained from [13] 'HG', human PPI network obtained from BIND [78] 'HB', human PPI network obtained from HPRD [65] 'HH' and human PPI network obtained from MINT [79] 'HM'. The points in the figure indicate the averages of GDDAs between 25 model networks and the corresponding PPI networks. GraphCrunch is used for the calculation of GDDA scores.

3.2.2. PPI networks considered

In [74], six PPI networks (two of yeast and four of human) were analyzed, as listed in table 3.5. These PPI networks were compared with three model networks, ER, ER-DD and GEO-3D, and it was shown the highest GDDA for GEO-3D, followed by ER-DD and ER models.

Name	# of nodes	# of edges	Density	Organism	Reference
YIC	796	841	0.00266	S. Cerevisiae	Ito *et al.* [10]
YHC	988	2,455	0.00503	S. Cerevisiae	Mering *et al.* [75]
HS	1,705	3,816	0.00219	H. sapiens	Stelzl et al. [8]
HG	3,134	6,725	0.00137	H. sapiens	Rual et al. [13]
BG-MS	1,923	3,866	0.00209	H. sapiens	BioGRID [64]
BG-Y2H	5,057	9,442	0.00074	H. sapiens	BioGRID [64]

Table 3.5.: PPIs analyzed in Rito *et al.*'s paper. BG-MS is the interaction data obtained from BioGRID filtered by key words 'Affinity Capture-MS', and BG-Y2H is the interaction data obtained from BioGRID filtered by key words 'Two-hybird'. This table is reproduced from [74].

It is notice that the PPI networks listed above are from early studies. For example, the interactions in YIC were detected by Ito *et al.* twelve years ago. Therefore, these PPI networks were old, small and sparse. To see whether the conclusion in [74] holds for the latest PPIs data, we analyzed the latest available PPI networks of yeast, fruitfly, nematode worm and human. Table 3.6 lists the details of the PPI networks we analyzed.

Name	# of nodes	# of edges	Density	Organism	Reference
HS	1,529	2,667	0.002283	H. sapiens	Stelzl et al. [8]
HG	1,873	3,463	0.001975	H. sapiens	Rual et al. [13]
HH	9,465	37,039	0.000827	H. sapiens	HPRD [65]
HR	9,141	41,456	0.000992	H. sapiens	Radivojac et al. [80]
HB	8,920	35,386	0.000890	H. sapiens	BioGRID [64]
WB	2,817	4,527	0.001141	C. Elegans	BioGRID [64]
FB	7,372	24,063	0.000886	D. Melanogaster	BioGRID [64]
YB	5,607	57,143	0.003636	S. Cerevisiae	BioGRID [64]

Table 3.6.: Details of latest PPIs we analyzed. Note that PPIs HS, HG are also analyzed in [74], but the number of nodes, edges and graph density are different, as we remove self-loops, reduplicate interactions and interspecies interactions from the PPIs data for more precise analysis.

In table 3.6, HS stands for the human PPIs obtained from [8], and the PPIs from [13] is denotes by HG. HH is the human PPIs download from HPRD [65] (version 9, released in April 2010). HR stands for the human PPIs collected from [80]. HB, WB, FB and YB are the PPIs of human, worm, fruitfly and yeast, obtained from BioGRID [64] (ver. 3.1.74, released in March 2011). We then compare these PPI networks with seven network models, namely ER, ER-DD, GEO-3D, GEO-GD, SF, SF-GD and STICKY model (details of these models are described in section 2.2.2), by calculating the GDDA scores between PPI networks and models (figure 3.5).

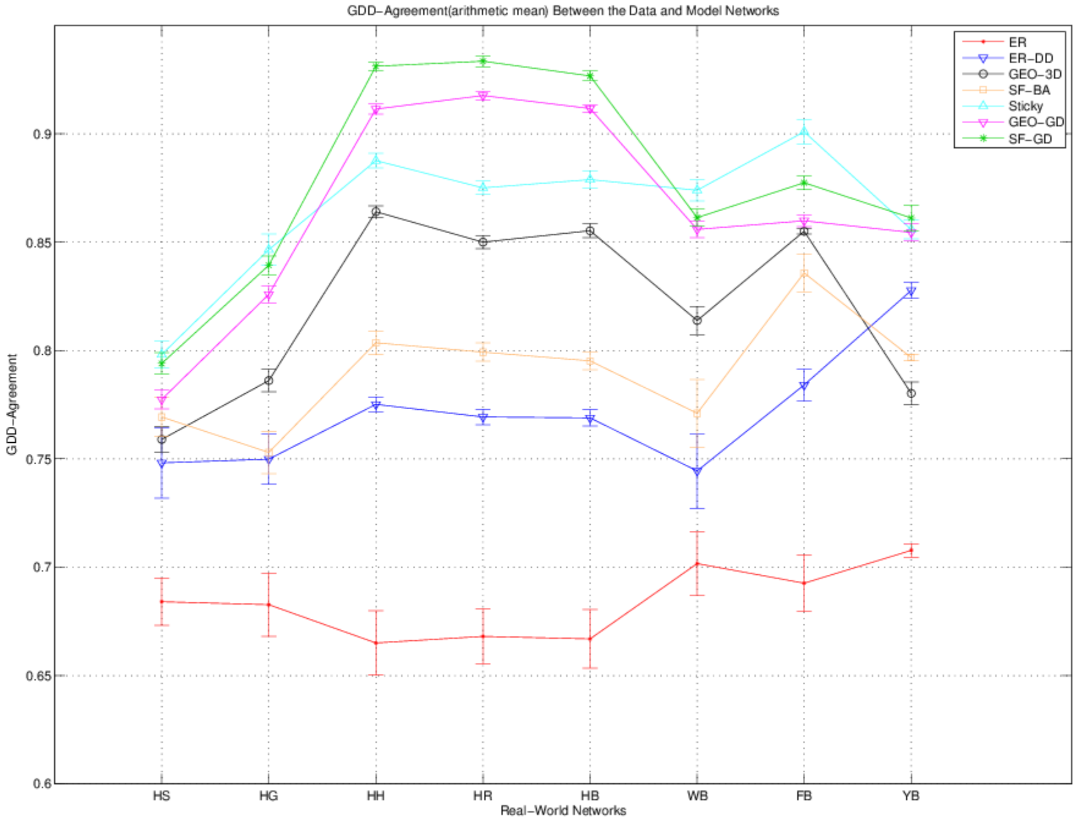


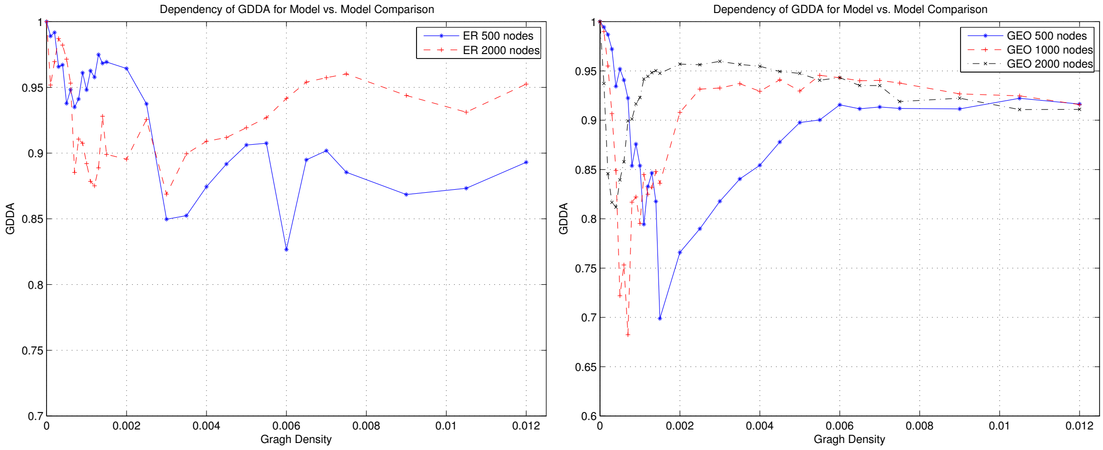
Figure 3.5.: GDDA between latest PPI networks and their corresponding random model networks.

Our results show the latest network models (sticky, GEO-GD and SF-GD) better fit the PPI networks than previous models.

3.2.3. Empirical Distributions of GDDA

To examine the use of GDDA for network comparison, Rito *et al.* generated networks of 500, 1000 and 2000 vertices with increasing graph density for both ER and GEO-3D model. There graphs were used as query networks and compared with 50 networks generated from the same model. We reproduce the results for comparing ER vs. ER and GEO vs. GEO networks with 500, 1000 and 2000 nodes across a range of graph densities [0, 0.0105], as shown in figure 3.6.

Our results are consistent with [74]. As shown in figure 3.6, the GDDA score is not stable in some graph density regions. However, it is partial to say that the GDDA score



(a) GDDA for ER vs. ER comparison

(b) GDDA for GEO vs. GEO comparison

Figure 3.6.: Dependency of GDDA of model vs. model comparisons on the number of vertices and edges of a network. GDDA of ER vs. ER with 500 and 2000 nodes, GDDA of GEO vs. GEO with 500, 1000 and 2000 nodes are plotted against graph density, and each value represents the average agreement of 50 networks. One may notice that figure 3.6 is not exactly the same as the one in Rito *et al.*'s paper. The reason for this is the randomness of the model networks, which is also mentioned in [74].

is unstable in a particular graph density region [0, 0.01], not only because the instability of GDDA is different for each model type, but also because the range of unstable region shrinks markedly as the increase of graph size (according to number of nodes and number of edges). For example, according to graph density, the GDDA volatile area for GEO-3D model with 500 nodes is around [0, 0.005], while for GEO-3D model with 2000 nodes is narrowed to [0, 0.0015]. To validate this, we also calculate the empirical distributions of GDDA for graphs with 5000 and 10000 nodes (results are listed in the appendix). Our results show that for a particular network model, the GDDA unstable region of networks with large number of nodes is much smaller than the one of networks with small number of nodes.

For better illustrating the volatility of GDDA scores, we construct a 3D view of empirical distributions for both ER and GEO-3D model, as shown in figure 3.7 and figure 3.8.

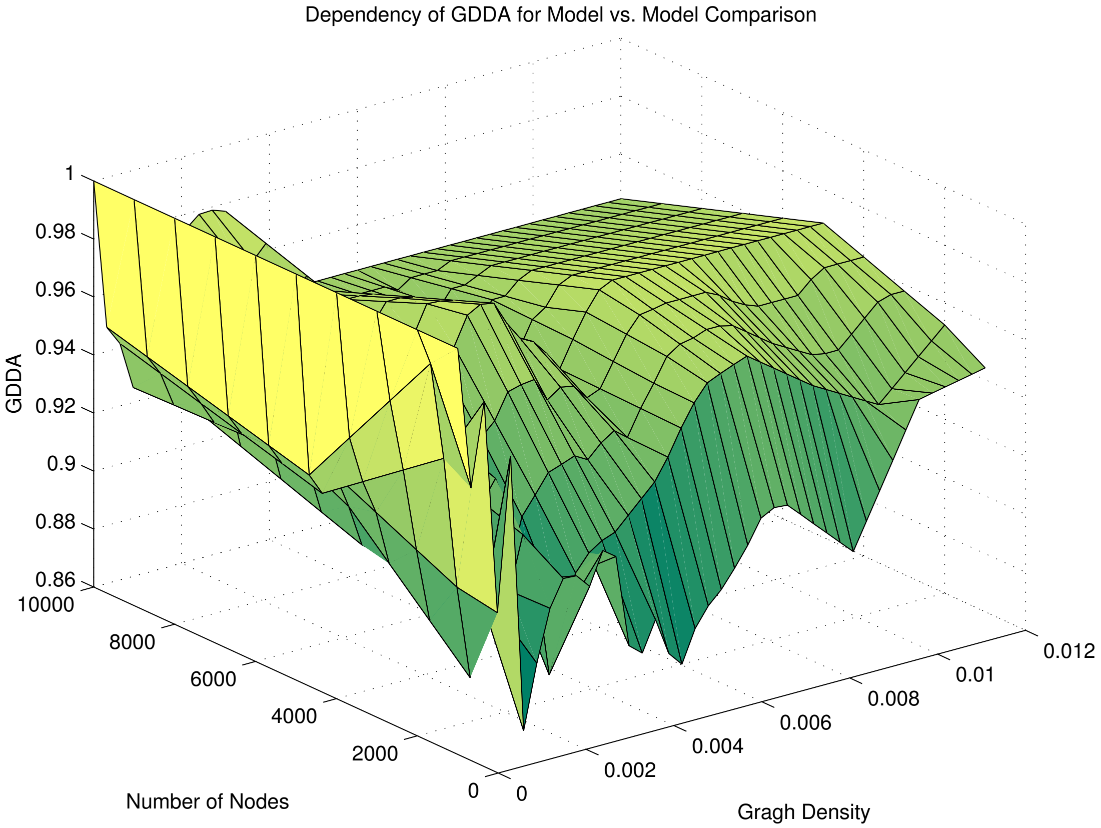


Figure 3.7.: 3D view of empirical distributions of GDDA for ER vs. ER comparisons

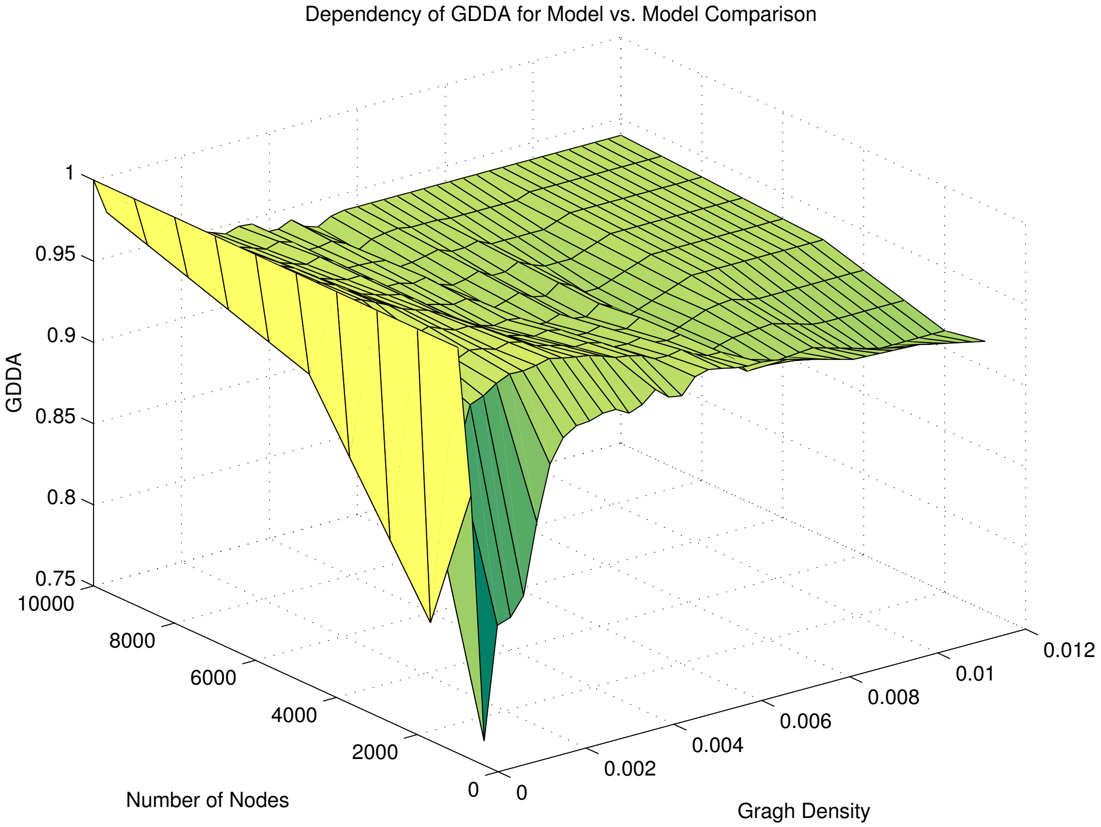


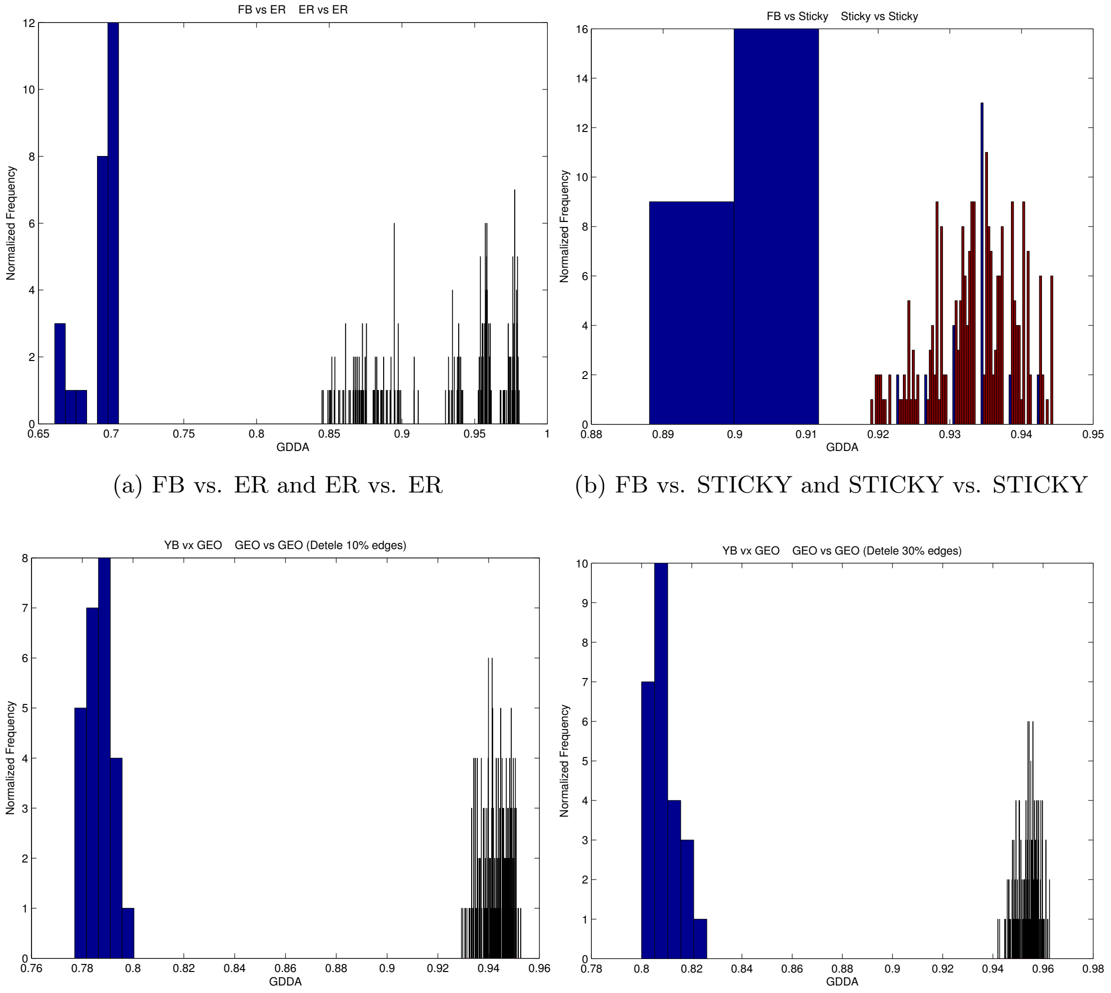
Figure 3.8.: 3D view of empirical distributions of GDDA for GEO vs. GEO comparisons

The PPI networks listed in table 3.6 can be plotted into figure 3.7 and figure 3.8 according to their size and graph density. The latest PPI networks of human, yeast and fruitfly, namely ‘HB’, ‘HH’, ‘YB’ and ‘FB’, as they all have more than 5,000 nodes and a graph density higher than 0.0008, obviously they are in the stable region of GDDA, for both ER model and GEO model. Smaller and earlier PPI networks of human and worm, namely ‘HS’, ‘HG’, and ‘WB’, are in the stable region for GEO model, but in the unstable region for ER model. As more and more PPIs have been identified, the size of PPI network is getting larger, which means GDDA is appropriate for analyzing these latest PPI networks.

3.2.4. Model Fitness

Rito *et al.* provided a method for assessing the statistical significance of the fit between random graph models and biological networks based on non-parametric tests [74]. In this method, several same model vs. model comparisons with roughly the same number of nodes and edges are carried out to assess the best obtainable score for this specific case (GDDA scores are recorded as sample A). GDDA scores are also calculated between the query network (PPI network) and graphs from the model network (sample B). Model fit are evaluated by gauging the differences between these two samples.

In [74], Rito *et al.* concluded that none of the theoretical models considered in their study (ER, ER-DD and GEO-3D) fitted the PPI data listed in table 3.5, according to their statistical method. Taking a PPI network as input, we compare the PPI network with ER, ER-DD, GEO, SF, STICKY models. The results show that there are no overlap between the GDDA scores recorded in sample A and sample B. It is not surprising, as none of these network models are supposed to fit the PPI networks perfectly. Recall that GDDA is proposed to search the network model that better fit the real-world data. Furthermore, we notice that though there are no overlap of the distribution of GDDA, distance between the distributions of sample A and sample B is smaller for the model which better fit the PPI network. Figure B.3a and figure B.3e illustrate that the histograms of GDDA values between PPI network FB vs. 25 ER model networks (left) and GDDA of 10 ER networks, each vs. 25 ER models. Figure B.3a illustrates the histograms of GDDA values between PPI network FB versus 25 ER model networks (left) and GDDA of 10 ER networks, each versus 25 ER models (right). Figure B.3e illustrates the histograms of GDDA values between PPI network FB versus 25 STICKY model networks (left) and GDDA of 10 ER networks, each versus 25 ER models (right). Obviously, the distance between two sample is much smaller for STICKY model than ER model.



(a) FB vs. ER and ER vs. ER

(b) FB vs. STICKY and STICKY vs. STICKY

(c) YB vs. GEO-3D and GEO-3D vs. GEO-3D (deleted 10% edges)

(d) YB vs. GEO-3D and GEO-3D vs. GEO-3D (deleted 30% edges)

Figure 3.9.: Normalized histograms of GDDA values. a) Histograms of GDDA values between PPI network FB versus 25 ER model networks (left) and GDDA of 10 ER networks, each versus 25 ER models (right). b) Histograms of GDDA values between PPI network FB versus 25 STICKY model networks (left) and GDDA of 10 ER networks, each versus 25 ER models (right). c) Histograms of GDDA values between PPI network YB versus 25 noisy GEO-3D model networks (left) and GDDA of 10 noisy GEO-3D networks, each versus 25 noisy GEO-3D models (right). The noisy GEO-3D models are generated by deleting 10% of edges from original GEO-3D models. d) Histograms of GDDA values between PPI network YB versus 25 noisy GEO-3D model networks (left) and GDDA of 10 noisy GEO-3D networks, each versus 25 noisy GEO-3D models (right). The noisy GEO-3D models are generated by deleting 30% of edges from original GEO-3D models.

As current PPI networks are noisy and incomplete, one idea is that the PPI networks may be better fitted by the model networks which contains noise rather than the original model. To validate this idea, we generate "noisy model networks", by adding, deleting, and rewiring 10%, 20% and 30% of edges in the model networks. Figure B.4d and figure B.4f show the histograms of GDDA values between PPI network YB versus 25 noisy GEO-3D model networks and GDDA of 10 noisy GEO-3D networks, each versus 25 noisy GEO-3D models. The noisy GEO-3D models are generated by deleting 10% (30%) of edges from original GEO-3D models. These results suggest that the noisy model network may better fit the PPI data. We are now working on this project, and hopefully we will submit a paper to *Bioinformatics* in two months.

4. Future Work

Our future plans will be presented in this chapter. The main research problems will be listed, along with detailed descriptions of proposed methodology for addressing them. These research problems include developing methods for analyzing the integrated network, classifying human diseases and implementing new software tools. A proposed project progress plan and an initial outline of the dissertation are also included in this chapter.

4.1. Research Problems and Proposed Methodology

4.1.1. Developing New Methods for Network Analysis

The structure of the integrated molecular interaction network (see section 3.1.2 for details) is complex, since it contains several different types of biological information. Currently, graphlet-based network analysis methods are mainly applied to undirected networks which contain only one type of nodes and one type of edges (figure 4.1a). Hence, we propose to heavily extend these methods for analyzing the integrated network.

As described in section 3.1.2,five types of biological network are proposed to be integrated into the molecular interaction network, namely transcriptional regulation network, metabolic network, cell signalling network, PPI network and genetic interaction network.

- There exist networks that have the same nodes, but different types of edges, for example, PPI network and genetic interaction network. Integration of these networks will introduce the problem of analyzing networks which contain different types of edges (figure 4.1b).

- Since both proteins and metabolites can be modeled as nodes in a metabolic network (see section 2.1.1 for details), the integrated network may also contains different types of nodes (figure 4.1c).

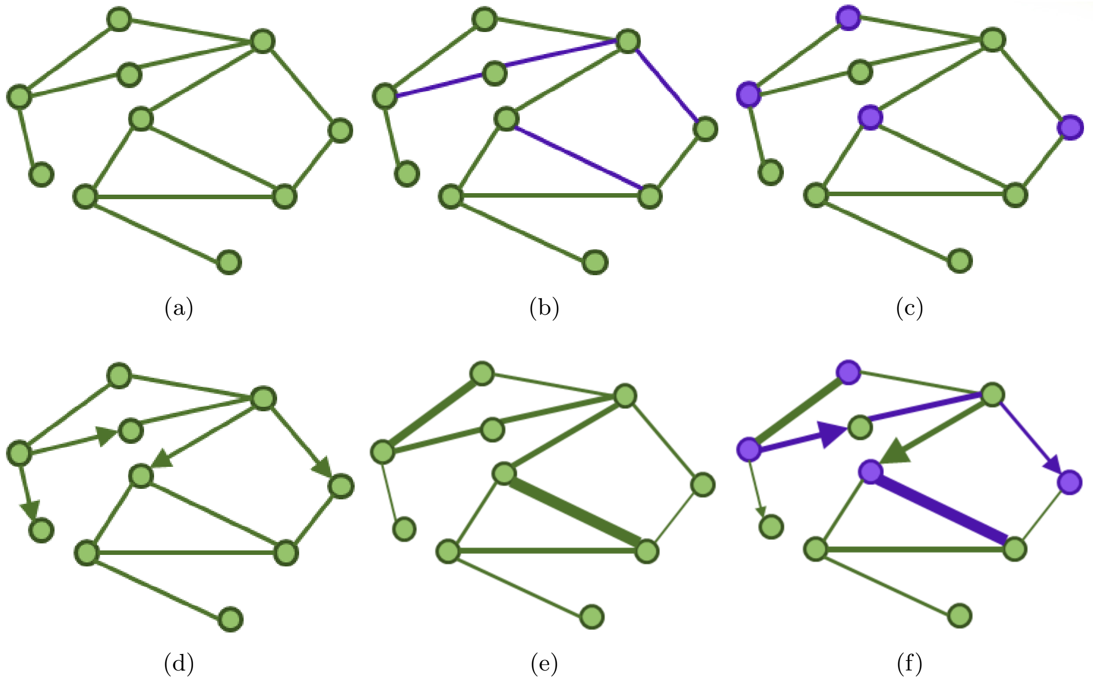


Figure 4.1.: A schematic representation of different types of network integration problems. (a) A network which contains only one type of nodes and one type of edges. (b) A network which contains different types of edges. (c) A network which contains different types of nodes. (d) A network which contains both undirected edges and directed edges. (e) A network which contains weighted edges. (f) A weighted network which contains different types of nodes and edges. Additionally, edges in this network can be directed or undirected.

- Moreover, some of these biological networks are naturally undirected (e.g., PPI network), while others are directed (e.g., transcriptional regulation network, cell signalling network). This require to extend graphlet-based methods for analyzing directed networks and networks which contains both directed and undirected edges (figure 4.1d).

- Furthermore, there may exist weights on edges. For example, high-confident links should have higher weights than low-confident links. This problem increases the need of developing analysis methods for weighted graphs (figure 4.1e).

- Finally, since all these different types of biological networks are proposed to be integrated into a molecular network, we need to develop new algorithms and tools to analyze complex networks which include all features described above (figure 4.1f).

The integration of molecular interaction network with disease network, drug network and patient record leads to the problem of analyzing "layered network". The layered network can be viewed as a model that contains several layers, and each of these layers is a network representing different knowledge. Figure 4.2 gives an example of a layered network consisting of three networks. Besides links within each layer, links between different layers will be constructed from biological data (see section 3.1.3 for details).

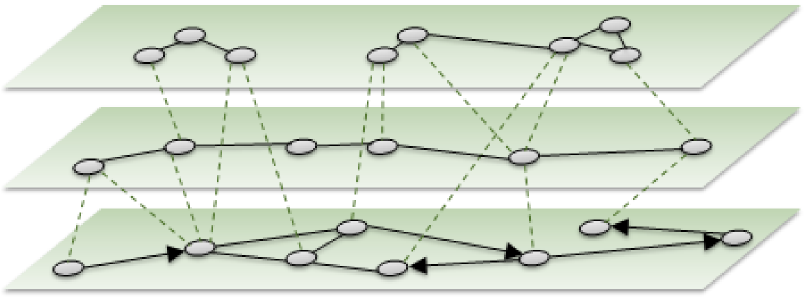


Figure 4.2.: An illustration of an integrated "layered network" consisting of three networks. Here, the solid lines represent the links within a layer, and the dashed lines represent the links across different layers.

We propose to extend the definition of graphlet degree vector (GDV, see section 2.2.1 for details) for analyzing these layered networks. While a node's GDV within the layer describes topological neighborhood of the node in that particular network, it's GDV that go along links between different layers of layered networks describe integrated topological connectivity between different networks. New methods will be developed to compare the similarity between nodes in the layered network.

4.1.2. Disease Re-classification

Our aim of this project to get new biological insight that would lead to better classification of human diseases. The strategy of disease classification may revelent to graph clustering problems. Diseases which are classified into the same catalogue can be viewed the nodes belong to a cluster in the disease network. The similarity measure may be design based on the "extended GDV" mentioned above.

4.1.3. Implementation of new methods

We propose to develop new tools for analyzing the integrated network. Since the size of the biological data are very large, the methods to analyze this data must be efficient. The time complexity of these new algorithms should be at most $O(n^2)$. Meanwhile, all these methods need to be robust to noise, since we already known that all of these biological networks are noisy and largely incomplete. New tools which implement these new network analysis methods should be accurate, stable and speedy. We prefer to use C++ as the coding language, and some C++ class libraries such as LEDA may be used in our implementation.

4.2. Project Progress Plan

According to chapter 1, the project includes 1) integrating different biological data, 2) designing a hybrid network to represent the data, 3) developing new algorithms and software to analysis the hybrid network and redefine disease classification, and 4) validating the results and writing up. A project progress plan is summarized in table 4.1.

Proposed timeline	Proposed research work
July 2011 to December 2011	1) Integrate transcriptional regulation network, metabolic network, cell signalling network, protein-protein interaction network and genetic interaction network into a molecular interaction network. 2) Develop new methods to analyze this integrated network. 3) Biological validation of the analysis results.
January 2012 to June 2012	4) Integrate molecular interaction network with disease network, drug network and patient records. 5) Design a hybrid network model to represent the data.
July 2012 to December 2012	6) Develop new methods to analyze the hybrid model. 7) Implement new biological network analysis software based on these new methods. 8) Evaluate the performance of the new software.
January 2013 to June 2013	9) Develop new methods for efficient and reliable disease classification. 10) Biological validation of the classification results.
July 2013 to October 2013	11) Finish the dissertation.

Table 4.1.: Project progress plan.

4.3. Outline of Dissertation

A proposed outline of the dissertation including chapters and expected section headings is listed in table 4.2.

Chapter 1 Introduction 1.1 Motivation 1.2 Introduction to biological networks 1.3 Graph theory for network analysis 1.4 Challenges in biological network research 1.5 Dissertation outline
Chapter 2 Biological Network Integration 2.1 Integration of molecular interaction networks 2.2 The hybrid network model 2.3 Analysis of the hybrid network model 2.4 Results and discussion
Chapter 3 Diseases Re-classification 3.1 Our approach 3.2 Biological validation of our results 3.3 Results and Discussion
Chapter 4 Conclusions 4.1 Summary of the dissertation 4.2 Future work

Table 4.2.: Proposed outline of dissertation.

The proposed chapters in the dissertation includes introduction, biological network integration, disease re-classification and conclusions. However, this outline is initial and may be modified. The final dissertation will be organized according to the real research works.

5. Conclusion

In this report, we give a statement of our research problem and proposed methods, as well as a review of the research progress so far. A literature survey which covers relevant topics on biological network modeling and analysis is presented.

We propose to re-define human disease classification via integration of biological networks. To do this, we will design a network-based mathematical model to represent the integrated biological data, and develop new computational algorithms and tools for its analysis.

So far, the biological information used for network integration have been collected from several public available databases. These biological data include molecular interactions, disease-gene association and drug-target association. The ideas of how to integrate these data into a hybrid network model and how this model can be used for disease re-classification, are also discussed in the report.

Our preliminary results include molecular and disease data collection, as well as evaluation of GDD agreement measure. We have applied the methods proposed in Rito et al.s paper to the latest PPI networks, to exam the use of GDD agreement for biological network comparison. Our results show that though the GDD agreement scores are not stable in some graph density region, this don't affect on the analysis of latest PPI data. We validate that GDD agreement is appropriate for analyzing these PPI networks.

The research problems we will address and the new techniques we will develop are stated in the future research section. The research project is scheduled, and a proposed outline of the dissertation is given at the end of the report.

Acknowledgement

I am very grateful to my supervisor Dr. Nataša Pržulj for her guidance and support. Dr. Pržulj has led me into the exciting world of bioinformatics, and provides me with valuable advices and inspiration.

Also, I would like to thank my assessment team: Dr. Nataša Pržulj, Prof. Duncan Fyfe Gillies and Prof. Yike Guo, for their accessibility and cooperation. Moreover, I thank Prof. Gillies for useful comments and suggestions on my research ideas and progress.

A special thank goes to my industrial supervisor Dr. Chris Larminie, and members of GlaxoSmithKline computational biology group, for their valuable discussion and feedback on the project.

Finally, I would like to thank GlaxoSmithKline (GSK) Research & Development Ltd for their financial support.

Bibliography

[1] J. Loscalzo, I. Kohane, and A.-L. Barabasi, "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.," *Molecular systems biology*, vol. 3, p. 124, Jan. 2007.

[2] W. H. Organization, *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Volume 2*, vol. 36. Geneva: World Health Organization, second edi ed., Apr. 2004.

[3] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization.," *Nature reviews. Genetics*, vol. 5, pp. 101–13, Feb. 2004.

[4] T. Milenkovic, *From Topological Network Analyses and Alignments to Biological Function, Disease, and Evolution.* PhD thesis, University of California, Irvine, 2010.

[5] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, p. 2, 2003.

[6] A. D. King, N. Pržulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, 2004.

[7] N. Pržulj, D. Wigle, and I. Jurisica, "Functional topology in a network of protein interactions," *Bioinformatics*, vol. 20, no. 3, pp. 340–348, 2004.

[8] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. Wanker, "A human protein-protein interaction network: A resource for annotating the proteome," *Cell*, vol. 122, pp. 957–968, 2005.

[9] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: New Features for Data Integration and Network Visualization.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 431–432, Dec. 2010.

[10] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto,

S. Kuhara, and Y. Sakaki, “Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins,” *Proc Natl Acad Sci U S A*, vol. 97, no. 3, pp. 1143–7, 2000.

[11] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proc Natl Acad Sci U S A*, vol. 98, no. 8, pp. 4569–4574, 2001.

[12] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, E. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleish, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, “A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae,” *Nature*, vol. 403, pp. 623–627, 2000.

[13] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, “Towards a proteome-scale map of the human protein-protein interaction network,” *Nature*, vol. 437, pp. 1173–78, 2005.

[14] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin, “A generic protein purification method for protein complex characterization and proteome exploration,” *Nature Biotechnol.*, vol. 17, pp. 1030–1032, 1999.

[15] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, “Functional organization of the yeast proteome by systematic analysis of protein complexes,” *Nature*, vol. 415, no. 6868, pp. 141–7, 2002.

[16] N. e. a. Krogan, “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae,” *Nature*, vol. 440, pp. 637–643, 2006.

[17] S. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. Holstege, J. Weissman, and N. Krogan, “Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae,” *Molecular and Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2008.

[18] N. Pržulj, *Biological networks uncover evolution , disease , and gene functions*, pp. 1–31.

[19] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, pp. 64–68, 2002.

[20] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824–827, 2002.

[21] J. D. Jordan, E. M. Landau, and R. Iyengar, "Signaling networks: the origins of cellular multitasking.," *Cell*, vol. 103, pp. 193–200, Oct. 2000.

[22] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Mnard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, no. 5659, pp. 808–813, 2004.

[23] N. Freimer and C. Sabatti, "The human phenome project.," *Nature genetics*, vol. 34, pp. 15–21, May 2003.

[24] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *PNAS*, vol. 104, no. 21, pp. 8685–8690, 2007.

[25] Y. Li and P. Agarwal, "A pathway-based view of human diseases and disease relationships.," *PloS one*, vol. 4, p. e4346, Jan. 2009.

[26] K. Lee, H.-Y. Chuang, A. Beyer, M.-K. Sung, W.-K. Huh, B. Lee, and T. Ideker, "Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species," *Nucl. Acids Res.*, vol. 6, p. e136, 2008.

[27] S. I. Berger and R. Iyengar, "Network analyses in systems pharmacology.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2466–72, Oct. 2009.

[28] M. a. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature biotechnology*, vol. 25, pp. 1119–26, Oct. 2007.

[29] S. Cook, "The complexity of theorem-proving procedures," in *Proc. 3rd Ann. ACM Symp. on Theory of Computing: 1971; New York*, pp. 151–158, Association for Computing Machinery, 1971.

[30] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[31] N. Pržulj, *Analyzing Large Biological Networks: Protein-Protein Interactions Example.* PhD thesis, University of Toronto, Canada, 2005.

[32] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[33] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–3, 2002.

[34] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *PNAS*, vol. 100, pp. 11980–11985, 2003.

[35] S. Mangan, A. Zaslaver, and U. Alon, "The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks," *JMB*, vol. 334/2, pp. 197–204, 2003.

[36] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, pp. 1538–1542, 2004.

[37] N. Pržulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, pp. e177–e183, 2007.

[38] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," *Cancer Informatics*, vol. 6, pp. 257–273, 2008.

[39] V. Memišević, T. Milenković, and N. Pržulj, "Complementarity of network and sequence structure in homologous proteins," 2009. in preparation.

[40] N. Przulj, "Erratum to Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 26, pp. 853–854, Mar. 2010.

[41] P. Erdös and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[42] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[43] M. Penrose, *Geometric Random Graphs.* Oxford University Press, 2003.

[44] N. Pržulj. and D. Higham, "Modelling protein-protein interaction networks via a stickiness index," *Journal of the Royal Society Interface*, vol. 3, no. 10, pp. 711–716, 2006.

[45] B. Bollobas, *Random Graphs.* Academic, London, 1985.

[46] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[47] D. Higham, M. Rasajski, and N. Pržulj, "Fitting a geometric graph to a protein-protein interaction network," *Bioinformatics*, vol. 24, no. 8, pp. 1093–1099, 2008.

[48] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *Journal of The Royal Society Interface*, 2010.

[49] N. Przulj, O. Kuchaiev, A. Stevanović, and W. Hayes, "Geometric evolutionary dynamics of protein interaction networks.," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 178–89, Jan. 2010.

[50] R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nature Biotechnology*, vol. 24, no. 4, pp. 427–433, 2006.

[51] B. P. Kelley, Y. Bingbing, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, "PathBLAST: a tool for alignment of protein interaction networks," *Nucl. Acids Res.*, vol. 32, pp. 83–88, 2004.

[52] R. e. a. Sharan, "Conserved patterns of protein interaction in multiple species," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 1974–1979, 2005.

[53] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, pp. 1746–1758, 2004.

[54] F. Schreiber and H. Schwobbermeyer, "MAVisto: a tool for the exploration of network motifs," *Bioinformatics*, vol. 21, pp. 3572–3574, 2005.

[55] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection.," *Bioinformatics (Oxford, England)*, vol. 22, pp. 1152–3, May 2006.

[56] V. Batagelj and A. Mrvar, "Pajek - program for analysis and visualization of large networks," *Timeshift - The World in Twenty-Five Years: Ars Electronica*, pp. 242–251, 2004.

[57] K. Yip, H. Yu, P. Kim, M. Schultz, and M. Gerstein, "The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks," *Bioinformatics*, vol. 22, pp. 2968–2970, 2006.

[58] Z. Liang, M. Xu, M. Teng, and L. Niu, "NetAlign: a web-based tool for comparison of protein interaction networks," *Bioinformatics*, vol. 22, no. 17, pp. 2175–2177, 2006.

[59] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks.," *Bioinformatics (Oxford, England)*, vol. 22, pp. 1021–3, Apr. 2006.

[60] T. Milenković, J. Lai, and N. Pržulj, "Graphcrunch: a tool for large network analyses," *BMC Bioinformatics*, vol. 9, no. 70, 2008.

[61] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: Scale-free or geometric?," *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[62] K. Oleksii, S. Aleksandar, and H. Wayne, "GraphCrunch 2: Software tool for network modeling, alignment and clustering," *BMC Bioinformatics*, vol. 12.

[63] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp, "Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data," in *Proceedings of the eighth annual international conference on Computational molecular biology (RECOMB'04)*, 2004.

[64] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, pp. D535–D539, 2006.

[65] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. N iranjan, H. C. Harsha, N. Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, M. Joy, H. N. S hivashankar, M. P. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. K. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey, "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Res*, vol. 32 Database issue, pp. D497–501, 2004. 1362-4962 Journal Article.

[66] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, pp. 27–30, 2000.

[67] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.," *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[68] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets.," *Nucleic acids research*, vol. 36, pp. D901–6, Jan. 2008.

[69] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.," *Nucleic acids research*,

vol. 39, pp. D1035–41, Jan. 2011.

[70] P. Agarwal and D. B. Searls, "Literature mining in support of drug discovery.," *Briefings in bioinformatics*, vol. 9, pp. 479–92, Nov. 2008.

[71] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs.," *Molecular systems biology*, vol. 6, p. 343, Jan. 2010.

[72] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes.," *Nucleic acids research*, vol. 37, pp. D619–22, Jan. 2009.

[73] A.-L. Barabási, "Network medicine–from obesity to the "diseasome".," *The New England journal of medicine*, vol. 357, pp. 404–7, July 2007.

[74] T. Rito, Z. Wang, C. M. Deane, and G. Reinert, "How threshold behaviour affects the use of subgraphs for network comparison.," *Bioinformatics (Oxford, England)*, vol. 26, pp. i611–i617, Sept. 2010.

[75] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[76] L. Giot, J. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. Stanyon, R. J. Finley, K. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. Shimkets, M. McKenna, J. Chant, and J. Rothberg, "A protein interaction map of drosophila melanogaster," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[77] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: definition, properties, and implications (extended version)," *arXiv:cond-mat/0501169*, 2005.

[78] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[79] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, H.-C. M., and G. Cesareni, "Mint: A molecular interaction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

[80] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and M. S. D., “An integrated approach to inferring gene-disease associations in humans,” *Proteins*, p. in press, 2008.

A. Supplemental Materials of Section 3.1

Sample of Patient Records

Patient records contain information on patient characteristics and registration details.

Sample data: 04XqA19810000014653120080530020000000000000000003N02002008021

1	Patient ID (04Xq)	2	Integrity of data (A for acceptable record)
3	Year of birth (1981)	4	Family ID (014653)
5	Sex (1 for male)	6	Registration date (2008/05/30)
7	Registration status (02 for permanent)	8	Date of transfer out (all 0s for no transfer)
9	Extended registration information (null)	10	Death date (all 0s for no death date)
11	Death information (null)	12	Acceptance type (3 for transfer-in)
13	Registration institute (N for unknown)	14	Marital (02 for married)
15	Dispensing (null)	16	Prescription exemption (00 for null record)
17	System date (2008/08/21)		

Table A.1.: A sample of patient records.

Sample of Medical Records

Medical records contain information on symptoms, diagnoses and interventions.

Sample data: 00?21995051100000000019192.00R000800000O0000000165YN00CD00CN19991012

1	Patient ID (00?2)	2	Event date (1995/05/11)
3	Event end date (all 0s for no date recorded)	4	Data type (01 for medical history)
5	Medical code (9192.00 for registered child surveillance)	6	Integrity flag (R for acceptable record)
7	Person entering record ID (0008)	8	Origin of record (all 0s for no record)
9	Episode type (all 0s for no record)	10	Secondary care speciality (0s for no record)
11	Location of consultation (O for others)	12	Text comment ID (00000001)
13	Medical entry (6 for administration)	14	Priority (lookups not yet available)
15	AIS extra information (null)	16	Event recorded in practices (Y for yes)
17	Private or NHS treatment (N for NHS)	18	Medical record ID (00CD)
19	Therapy AHD consultation ID (00CN)	20	System date (1999/10/12)
21	Edited by GP (N for no)		

Table A.2.: A sample of medical records.

Sample of Therapy Records

Therapy records contain information on details of prescriptions issued to patients.

Sample data: 00?21998071794794998Y000028100.0000000N0006100000.000501010300000000
0000008-1.00I5Y08Dp05Bj19991012N

1	Patient ID (00?2)	2	Prescription date (1998/07/17)
3	Drug code (94794998 for Amoxicillin)	4	Integrity flag (Y for acceptable record)
5	Dosage code (0000028 for one 5ml spoonsful to be taken three times a day)	6	Quantity prescribed (100)
7	Duration of the prescription (0 for null)	8	Private or NHS treatment (N for NHS)
9	Staff ID (0006)	10	Acute or repeat prescription (1 for acute)
11	Number of original packs ordered (0 for null)	12	BNF1 from DRUGCODES (05010103)
13	Repeat prescriptions' issue sequence number (0 for null)	14	Maximum number of repeat prescriptions' issue (0 for null)
15	Pack information (0000008 for ml)	16	Calculated daily dosage (null)
17	Location of consultation (I for surgery)	18	Source of drug (5 for in practice)
19	Event recorded in practices (Y for yes)	20	Therapy record ID (08DP)
21	Therapy AHD consultation ID (05Bj)	22	System date (1999/10/12)
23	Edited by GP (N for no)		

Table A.3.: A sample of therapy records.

Sample of AHD Records

Additional health data (AHD) records contain information on preventative health care immunizations and test results.

Sample data: 00?219950614100200010 0Y1IMM001DTPINP0046541.0000I00zc4NN00zc1UCz19991012N

1	Patient ID (00?2)	2	Event date (1995/06/14)
3	AHD code (1002000100 for 'Diphtheria')	4	Integrity flag (Y for acceptable record)
5	AHD specific data (1IMM001DTPINP004)	6	Read medical code (6541.00 for first diphtheria vaccination)
7	Origin of record (0 for no record)	8	Secondary care speciality (0s for no record)
9	Location of consultation (I for surgery)	10	Staff ID
11	Text comment ID (00zc)	12	Medical entry (4 for intervention)
13	AHD extra information	14	Event recorded in practices (N for no)
15	Private or NHS treatment (N for NHS)	16	AHD record ID (00zc)
17	Therapy AHD consultation ID (1UCz)	18	System date (1999/10/12)
19	Edited by GP (N for no)		

Table A.4.: A sample of AHD records.

Sample of Consult Records

Consult records contain information on consultation details.

Sample data: 3fyX000?2000?200306302003063012560501400100

1	Consultation ID (3fyX)	2	Patient ID (00?2)
3	Staff ID (000?)	4	Event date (2003/06/30)
5	System date (2003/06/30)	6	System time (12:56:05)
7	Type of consultation (014 for repeat issue)	8	Duration of consultation record open

Table A.5.: A sample of consult records.

Sample of Staff Records

Staff records contain information on staff (clinician, nurse, etc.) details.

Sample data: 00??0011

1	Staff ID (00??)	2	Sex (0 for male)
3	Role ID (011 for practice nurse)		

Table A.6.: A sample of staff records.

Sample of PVI Records

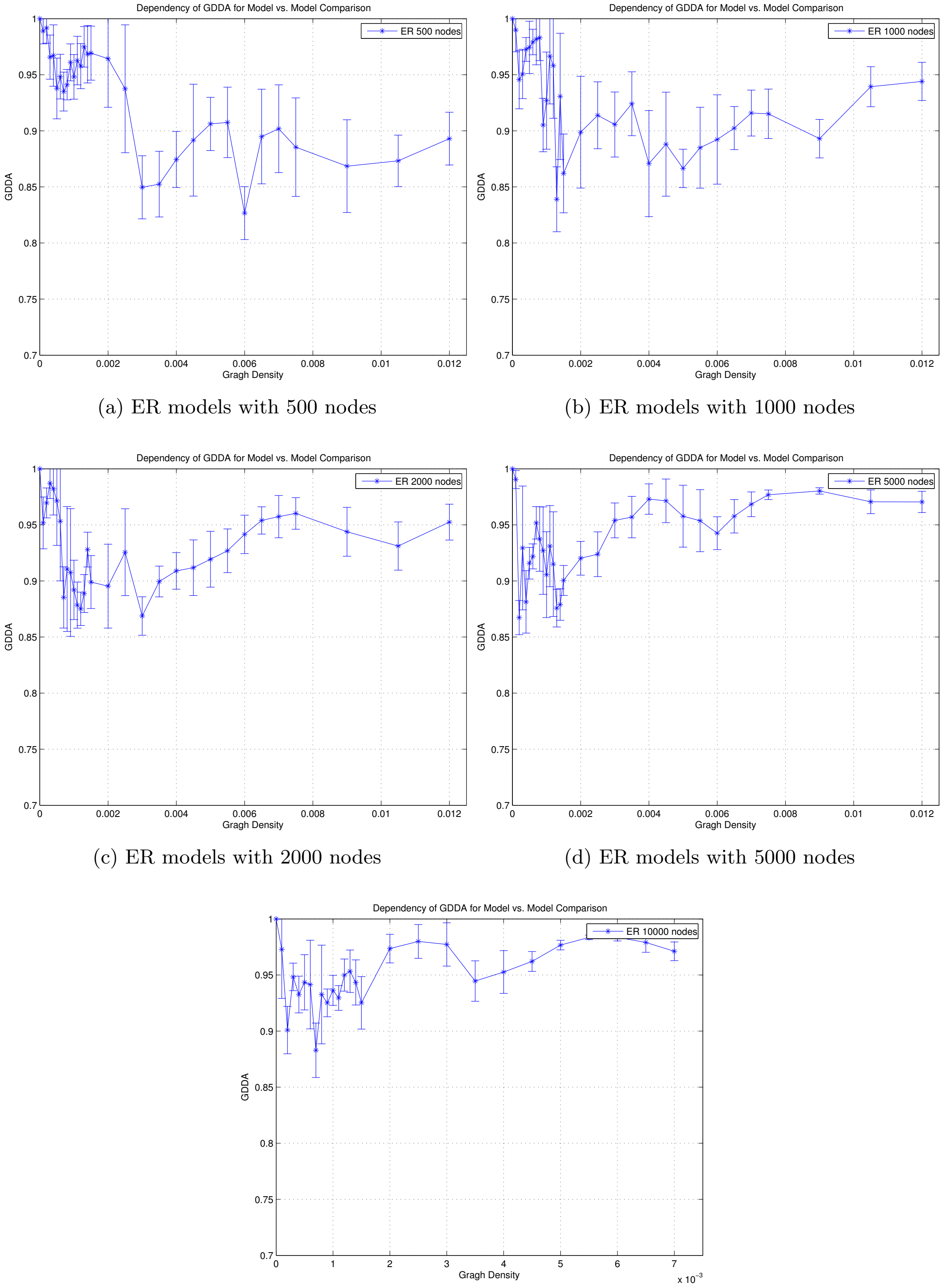
Postcode linked variables (PVI) records contain information on postcode-based socioeconomic, ethnicity and environmental indicators.

Sample data: 00?242445415435420081210

1	Patient ID (00?2)	2	Rural urban classification of wards (4 for urban ¿ 10k less sparse)
3	Quintile of proportion of ward population who define themselves as White etc.	4	Proportion of ward population with limiting long-term illness
5	Quintile of estimated level of NO_2	6	Quintile of estimated level of particulate matter
7	Quintile of estimated level of SO_2	8	Quintile of estimated level of NO_X
9	Quintile of Townsend score	10	Date of update (2008/12/10)

Table A.7.: A sample of PVI records.

B. Supplemental Materials of Section 3.2



(a) ER models with 500 nodes

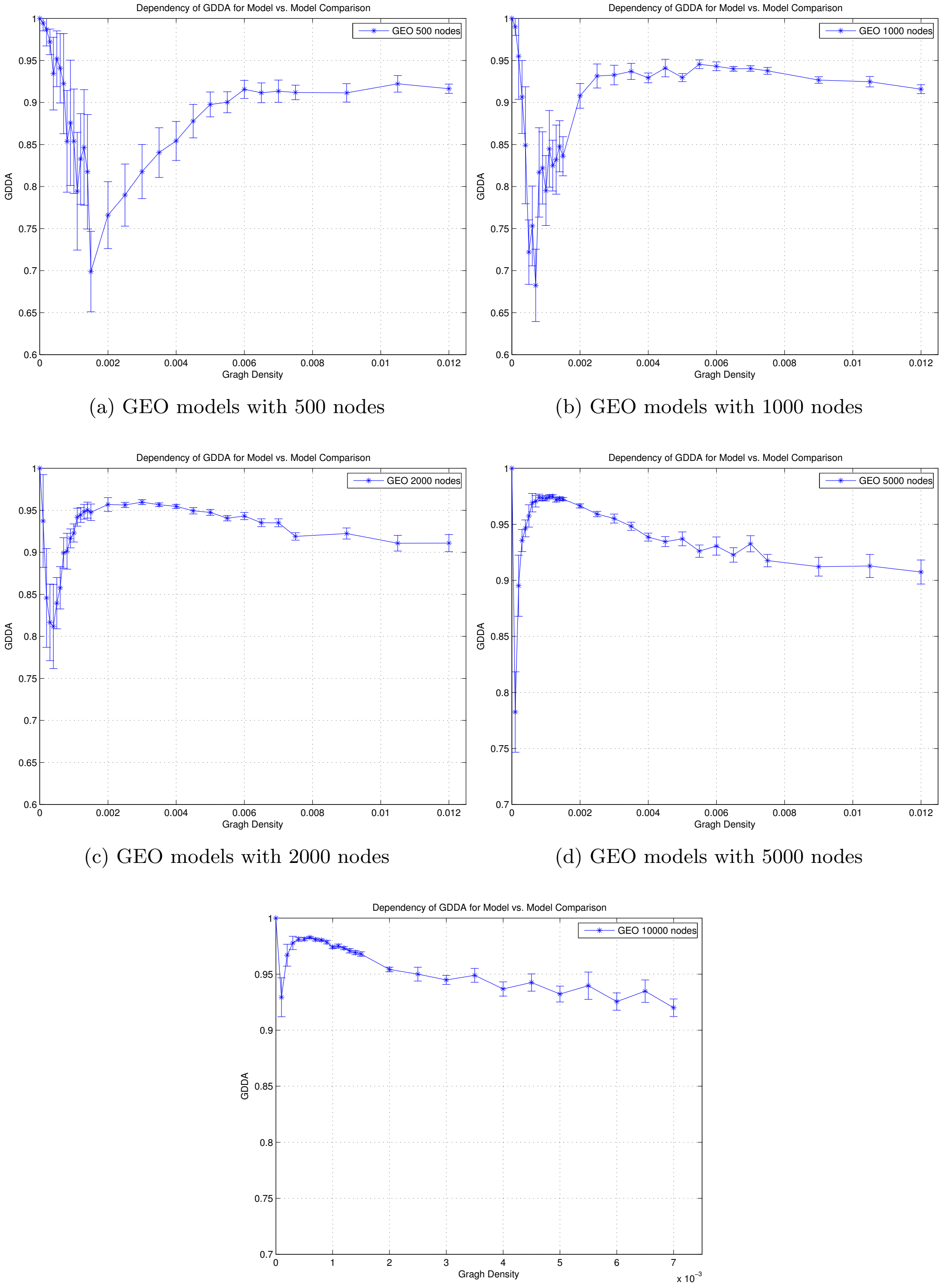
(b) ER models with 1000 nodes

(c) ER models with 2000 nodes

(d) ER models with 5000 nodes

(e) ER models with 10000 nodes

Figure B.1.: Empirical distribution of GDDA for ER vs. ER comparison.



(a) GEO models with 500 nodes

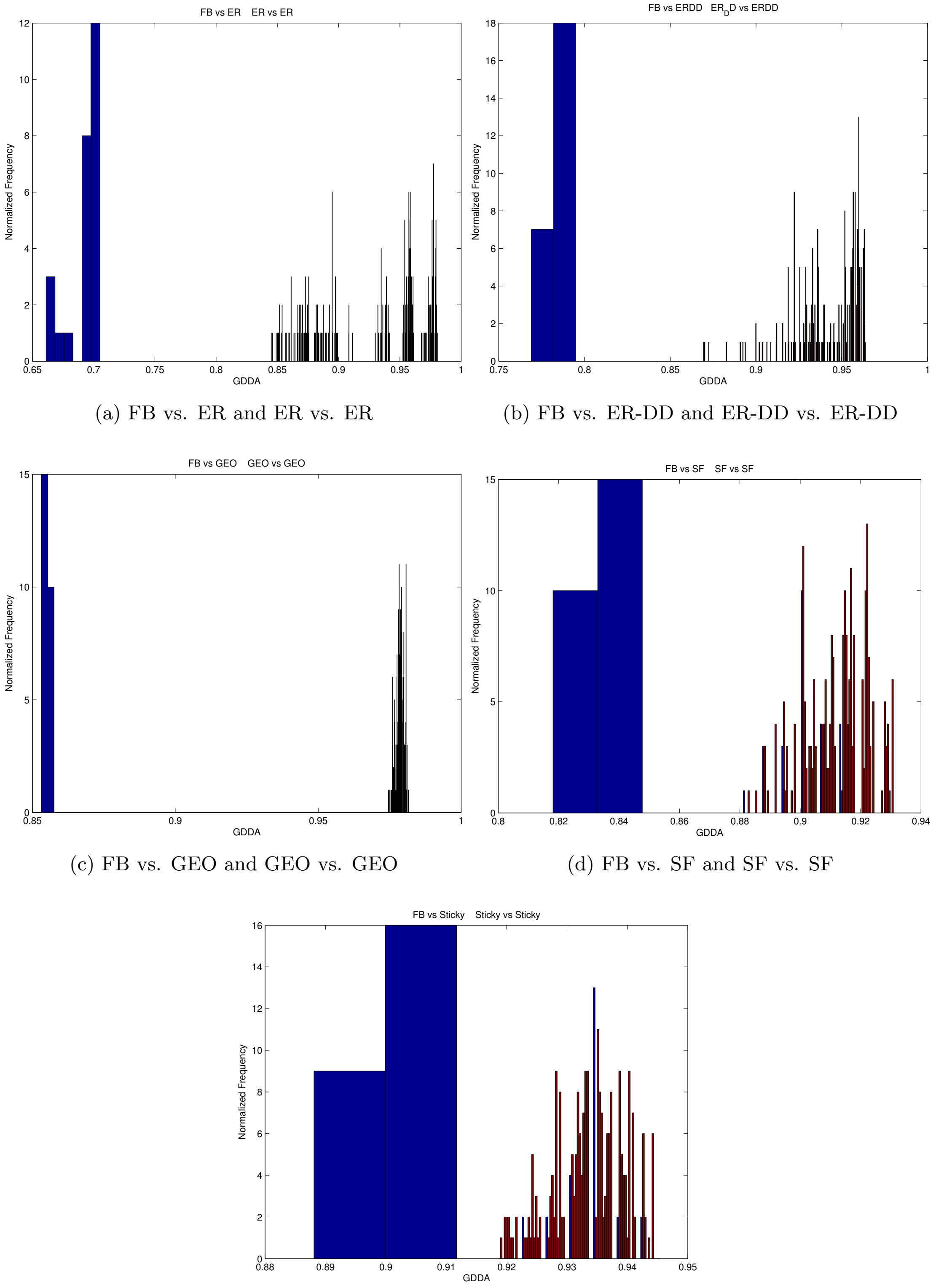
(b) GEO models with 1000 nodes

(c) GEO models with 2000 nodes

(d) GEO models with 5000 nodes

(e) GEO models with 10000 nodes

Figure B.2.: Empirical distribution of GDDA for GEO vs. GEO comparison.



(a) FB vs. ER and ER vs. ER

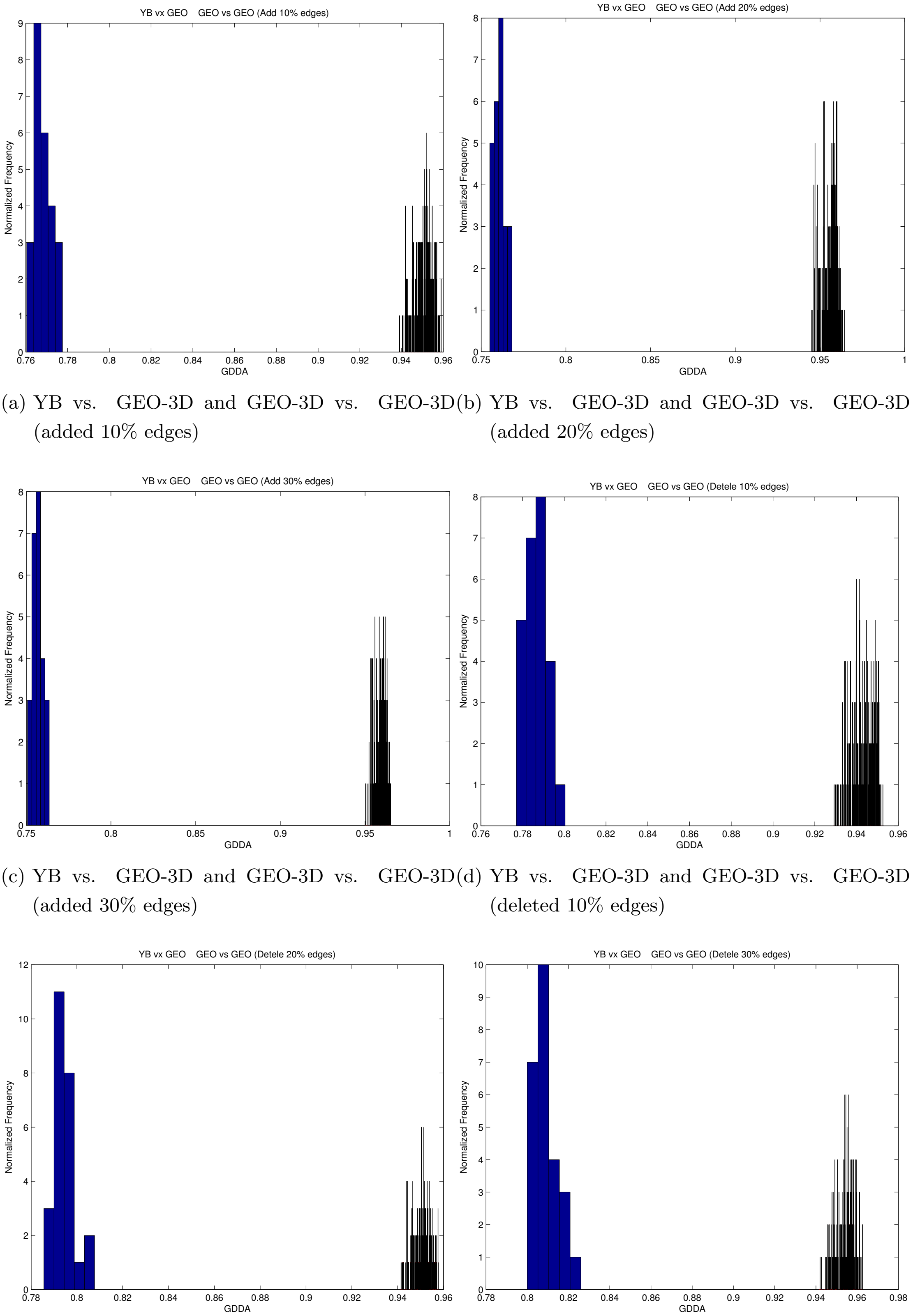
(b) FB vs. ER-DD and ER-DD vs. ER-DD

(c) FB vs. GEO and GEO vs. GEO

(d) FB vs. SF and SF vs. SF

(e) FB vs. STICKY and STICKY vs. STICKY

Figure B.3.: Normalized histograms of GDDA values (PPI network FB).



(a) YB vs. GEO-3D and GEO-3D vs. GEO-3D (added 10% edges)

(b) YB vs. GEO-3D and GEO-3D vs. GEO-3D (added 20% edges)

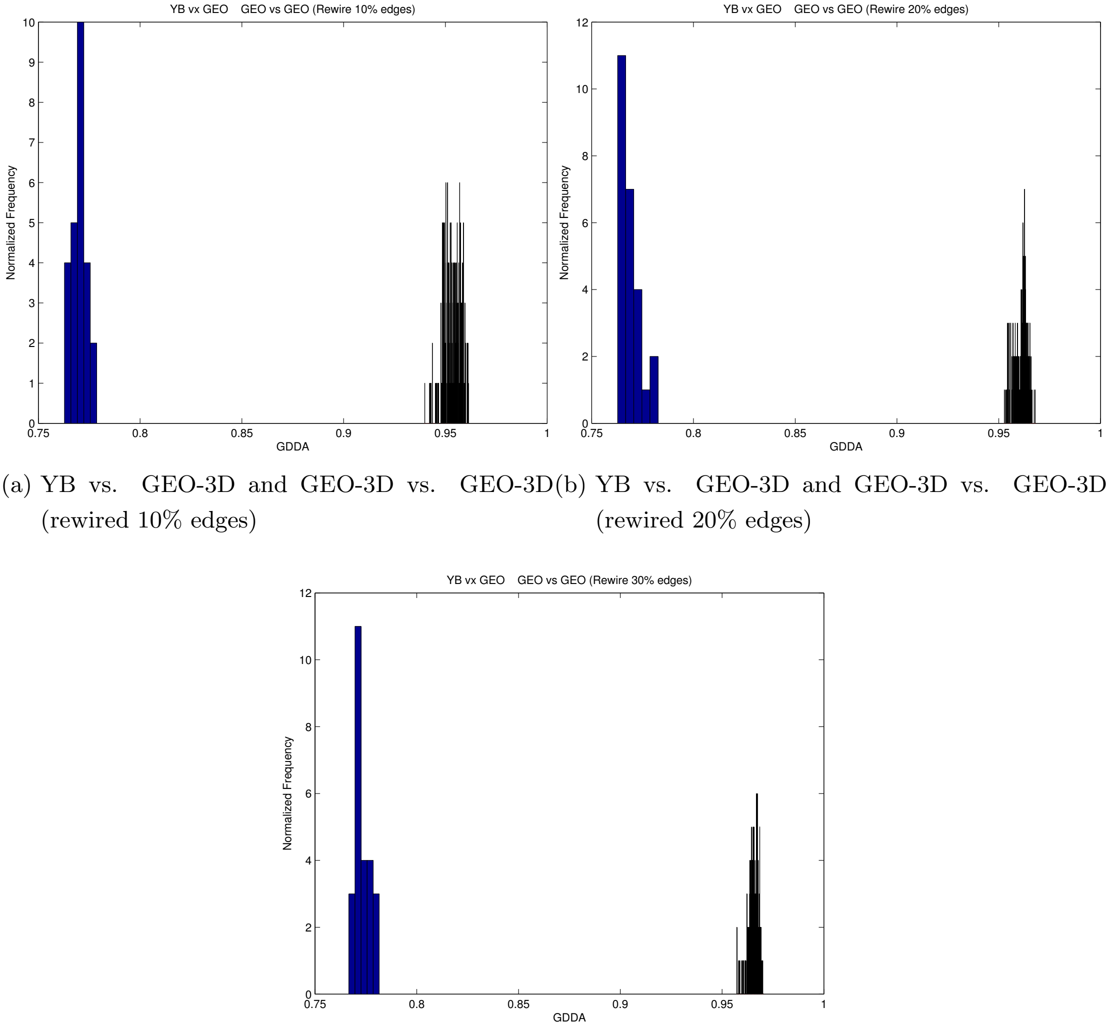
(c) YB vs. GEO-3D and GEO-3D vs. GEO-3D (added 30% edges)

(d) YB vs. GEO-3D and GEO-3D vs. GEO-3D (deleted 10% edges)

(e) YB vs. GEO-3D and GEO-3D vs. GEO-3D (deleted 20% edges)

(f) YB vs. GEO-3D and GEO-3D vs. GEO-3D (deleted 30% edges)

Figure B.4.: Normalized histograms of GDDA values (YB vs. noisy models).



(a) YB vs. GEO-3D and GEO-3D vs. GEO-3D (rewired 10% edges)

(b) YB vs. GEO-3D and GEO-3D vs. GEO-3D (rewired 20% edges)

(c) YB vs. GEO-3D and GEO-3D vs. GEO-3D (rewired 30% edges)

Figure B.5.: Normalized histograms of GDDA values (YB vs. noisy models).