PROTEIN PROTEIN INTERACTION NETWORK COMPARISON AND EMULATION

by

Fereydoun Hormozdiari

B.Sc. Sharif University of Technology, 2004

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
Fall 2006

APPROVAL

Name: Fereydoun Hormozdiari

Degree: MASTER OF SCIENCE

Title of thesis: Protein Protein Interaction Network Comparison And Emulation

Examining Committee: Dr. Jiangchuan Liu
Chair

Dr. S. Cenk Sahinalp, Senior Supervisor

Dr. Funda Ergun, Supervisor

Dr. Martin Ester, SFU Examiner

Date Approved: December 12, 2006

Abstract

The (asymptotic) degree distribution of the best known scale free network models are all similar and are independent of the seed graph used. Hence it has been tempting to assume that networks generated by these models are similar in general. In this thesis it is shown that several key topological features of such networks depend heavily on the specific model and seed graph used. Furthermore, it is shown that starting with right seed graph, the duplication model captures many topological features of publicly available PPI networks very well.

keywords: protein-protein interaction networks,topological properties,duplication model.

To my beloved parents.

"Animals, whom we have made our slaves, we do not like to consider our equal"

— CHARLES DARWIN

Acknowledgments

I feel most fortunate to have had the opportunity to study in Simon Fraser University and to carry out my Master's thesis at the Department of Computing Science.

The writing of this thesis would not have been possible without help of many people. It is a great pleasure to extend my sincere gratitude toward all of them.

First and foremost, I am highly indebted to my supervisor, Dr. S. Cenk Sahinalp, as this thesis borrows heavily from his unlimited help, stimulating effort and unrestricted patience. He guided me not to get lost during the development of this thesis and provided a motivating and enthusiastic atmosphere during the discussions we had. It was a great pleasure to do this thesis under his supervision.

I am deeply grateful to my other supervisors, Dr. Petra Berenbrink and Dr. Funda Ergun, excellent teachers, whose support and advise I have relied on during my masters program. I also would like to take the opportunity to thank Dr. Martin Ester for his encouragement and guidance during my Master program.

The collaboration with Dr. Natasa Pržulj on this topic was a pleasant experience and hope that it can be extended to future projects.

A special thanks goes to Shirin Hadizadeh for assisting me in typesetting and also being a main source of hope and encouragement. I also would like to thank Emre Karakoc for his motivating and intelligent discussions on this topic.

At last but not least, I would like to express my heartfelt and everlasting appreciation to my family for their endless love and support. I specially owe a great debt to my mother and father for sacrificing their own joy and comfort, and giving me moral support throughout all years of my study. Nothing would have ever been accomplished without their continuous love and encouragement. This thesis is dedicated to my parents.

Contents

List of Figures

Chapter 1

Introduction

Understanding the topology of protein-protein interactions (PPI) is an important problem in proteomics. It is believed that understanding the topology and dynamics of these networks can give deep insight into the inner working of a cell, which may lead to the development of potential drugs for complex diseases.

Protein-protein interactions play a central role in the execution of key biological functions of a cell. It is possible to represent interactions between pairs of proteins as binary relations which can be summarized as a undirected *graph* (network) in which each *node* represents a protein and each *edge* represents an interaction. A graph including all proteins and possible interactions between these proteins can be called the *proteome network* of an organism.

In the past few years protein-protein interaction (PPI) networks of several organisms have been derived and made publicly available.

1.1 PPI Databases

Using high throughput techniques a large amount of experimental protein-protein interaction data has been generated and made publicly available through various databases. Perhaps the best known PPI network database is DIP (Database of Interacting Proteins) [28] which includes the *S.Cerevisiae* (Yeast) PPI network (the best developed PPI network available with 4902 proteins and 17200 interactions), as well as the *C.Elegans* (Worm) network (with 2387 proteins and 3825 interactions). In this manuscript the main focus is on the Yeast PPI network from DIP but also there are results from other networks, such as Worm network from DIP as well as a more accurate but smaller CORE Yeast network(2345 proteins and

5609 interactions), also available through DIP [12]. Less developed PPI networks available through DIP [35] include those of the fruit fly, human, and mouse. Other PPI network databases include BIND [2], IntAct [16] and MINT [36].

1.2 Networks Interesting Properties

It is now well known that the structure of the Yeast proteome network seems to reveal two interesting graph theoretic properties [18, 31]:

(i) The degree distribution of nodes (i.e. the proportion of nodes with degree k as a function of k) approximates a *power-law* (i.e. is approximately ck^{-b} for some constants c, b).

(ii) The graph exhibits the *small world effect*: the shortest distance between a randomly selected pair of nodes is "small".

Small world phenomena and the power-law degree distributions have previously been observed in a number of naturally occurring graphs such as communication networks [13], web graphs [1, 3, 8, 10, 20, 21], research citation networks [26], human language graphs [14], neural nets [34], etc. These two properties cannot be observed in the classical random graph models studied by Erdös and Rényi [27] in which edges between pairs of nodes are determined independently.Erdös and Rényi random graph generation model starts with a set of vertices and add edges between each pair of nodes with constant probability $p = \frac{Avg.Degree}{n}$, where $Avg.Degree$ represents average degree of the graph we want to emulate, and also n represents number of nodes.

Since well known random graph models also have power-law degree distributions [3], [8], [33] it has been tempting to investigate whether these models agree with other topological features of the PPI networks.

There are two well known models that provide power law degree distributions (see [10, 9, 4]). The *preferential attachment* model [1, 8], was introduced to emulate the growth of naturally occurring networks such as the web graph; unfortunately, it is not biologically well motivated for modeling PPI networks. The *duplication model* on the other hand [7, 30, 23] is inspired by Ohno's hypothesis on genome growth by duplication [22]. Both models are iterative in the sense that they start with a *seed graph* and grow the network in a sequence of steps.

The degree distribution is commonly used to test whether two given networks are similar or not. However, networks with identical degree distributions can have very different

topologies.[1] Furthermore, it was observed in [15] that given two networks with substantially different initial degree distributions, a partial (random) sample from those networks may give subnetworks with very similar degree distributions. Thus the degree distribution can not be used as a sole measure of topological similarity.

In the recent literature two additional measures have been used to compare PPI networks with random network models. The first such measure is based on the *k-hop reachability.* The 1-hop reachability of a node is simply its degree (i.e. the number of its neighbors). The k-hop reachability of a node is the number of distinct nodes it can reach via a path of $\leq k$ edges. The k-hop reachability of all nodes whose degree is ℓ is the average k-hop reachability of these nodes. Thus the k-hop reachability (for $k = 2, 3, \ldots$) of nodes as a function of their degree can be a used to compare network topologies. An earlier comparison of the k-hop reachability of the Yeast network with networks generated by certain duplication models concluded that the two network topologies are quite different [5]. The second similarity measure is based on the *graphlet distribution.* Graphlets are small subgraphs such as triangles, stars or cliques. In [24] it was noted that certain "scale free" networks are quite different from the Yeast PPI network with respect to the *graphlet distribution.* This observation, in combination with that on the k-hop degree distribution seem to suggest that the known PPI networks may not be scale free and existing scale free network models may not capture the topological properties of the PPI networks.

There are other topological measures that have been commonly employed in comparing social networks etc., but not PPI networks. Two well known examples the *betweenness* distribution and the *closeness* distribution [32]. Betweenness of a vertex v is the number of shortest paths between any pair of vertices u and w that pass through v, normalized by the total number of such paths. Closeness of v is the inverse of the total distance of v to all other vertices u. Thus one can use betweenness and the closeness distributions, which respectively depict the number of vertices within a certain range of betweenness and closeness values can be used to compare network topologies.

[1] Consider, for example, an infinite two dimensional grid vs a collection of cliques of size 5; in both cases all nodes have degree 4.

Chapter 2

Measures for Comparing Networks

There are several topological features that can be used to test whether two networks are similar or not, starting at very rigorous measures like isomorphism, to very relaxed characteristics like degree distribution.

2.1 Graph Isomorphy

Two networks G and G' are called *isomorphic* if there exists a bijective mapping F from each node of G to a distinct node in G', such that two nodes v and w are connected in G if and only if $F(v)$ and $F(w)$ are connected. G and G' are called *approximately isomorphic* if by removing a "small" number of nodes and edges from G and G' they could be made isomorphic. Ideally a random graph model that aims to emulate the growth of a PPI network should produce a network that is approximately isomorphic to the PPI network under investigation.

Graph Isomorphism problem has been intensively studied, not only because of its many applications, but also in part because it is one of the few problems in NP that has resisted all attempts to be classified as NP-complete, or within P. Hence, it has been proposed by researchers to find some particular features and compare these features of the graphs instead of the graphs themselves.

2.2 Network Features

As explained in previous section, we do not know of a way of testing graph isomorphism in polynomial time. Hence, to fulfill the purpose of graph comparison we need to select a set of graph features and compare them. However, the selected features should have two main characteristics.

First, the feature's behavior for real PPI networks should be different from that of graphs generated using the Erdös-Rényi graph generation model [1]. Second, the selected feature should be robust, i.e. when minor changes happen it should not change behavior substantially. Furthermore, it should be sensitive, i.e. when a significant portion of edges are randomly changed the feature should be able to recognize the change.

2.2.1 Random Change of the Network

As explained in previous section, one of the properties that a feature should have in order to be acceptable for graph comparison is to be robust, i.e. not to change significantly when minor changes occur. In addition, the feature's behavior should be sensitive, i.e. when huge portion of edges have altered the features behavior should change significantly. To do this, a method is needed to randomly alter the edges of the given graph without changing the number of nodes and edges.

First we need to define the percentage of changes we are looking for, e.g. $p = 10\%, 20\%, ..., 100\%$. Then the algorithm with probability proportional to p deletes each edge in the graph, and adds one edge to the graph, by randomly selecting two unconnected nodes and connecting them (therefore the total amount of edges will not change). However, there is one exception, if the selected edge has an end point with degree 1 then we can only change the other end of the chosen edge to a random node (because if both ends of the edge are changed, it results in reduction of the number of nodes).

2.2.2 Degree Distribution

The first and foremost graph topology feature used in comparing two graphs (e.g. PPI networks and a graph generated by a particular random model) is *degree distribution.*

[1] In this classical random model all the nodes are created at once and then each pair of nodes get connected independently with fixed probability p

One very interesting property of some natural graphs such as the Internet or PPI networks is that their degree distribution seems to be in the form of power-law [18] [31], i.e. probability of a node having degree k is approximately ck^{-b} for some constants c, b. However, some recent works have challenged this behavior of PPI network, by attributing the power-law like behavior to sampling issues, experimental errors or statistical mistakes [15, 24, 29, 25, 11].

2.2.3 k-hop Reachability

Let $V(i)$ denote the set of nodes in V whose degree is i. Given a node v, its k-hop degree, i.e. the number of distinct nodes it can reach in at most k hops is denoted by $d(v,k)$. We Define $f(i,k)$, the k-hop reachability of $V(i)$ as

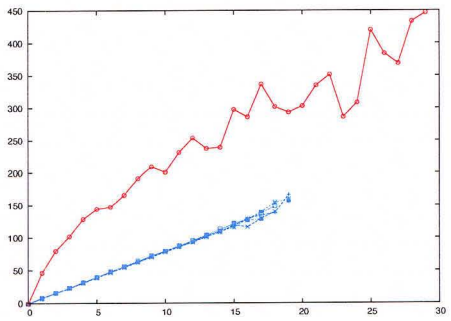
$$f(i,k) = \frac{1}{|V(i)|} \sum_{w \in V, d(w)=i} d(w,k)$$

Thus $f(i,k)$ is the "average" number of distinct nodes a node with with degree i can reach in k hops; e.g. $f(i,1) = i$ by definition.

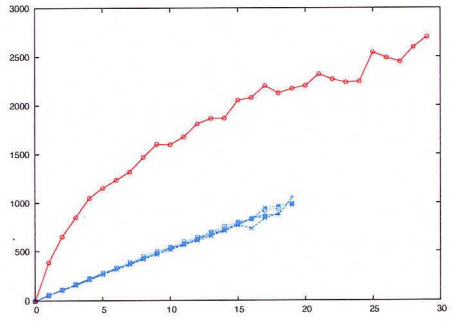
First, let us show that k-hop reachability is a useful feature for comparing PPI networks against graphs generated using random generation models. For this purpose, it is needed to compare the k-hop reachability of Yeast PPI network with the networks emulated using the simplest random graph generation model, i.e. Erdös-Rényi generation models. In Figure 2.1 it is shown that k-hop reachability of Yeast PPI network has a clear difference from that of Erdös-Rényi graphs, which means that k-hop reachability can be a good candidate for comparing different random model generations against real PPI networks.

The second second test is to see if k-hop reachability is robust (and sensitive) when yeast PPI network is considered or not.

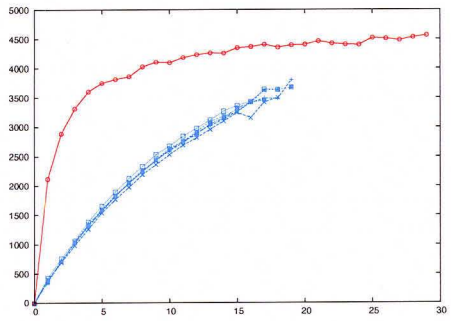
In figure 2.2, it is illustrated that k-hop reachability is fairly robust. The reason for this claim is that when a minor portion (10% or 20 %) of edges are changed the behavior of the feature does not change significantly (in figure 2.2, the red plots represent the Yeast PPI network, while green represents 10% and blue 20 % random changes in Yeast PPI network). In addition, in the same figure it is shown that k-hop reachability is also sensitive. Because when considerable portion (50% or 60%) of edges are changed it is reflected in features behavior. (in figure 2.2, the red plots represent the Yeast PPI network, yellow for 50% and black for 60% random changes).



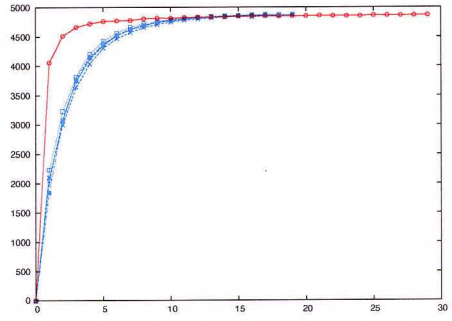
(a) 2-hop reachability



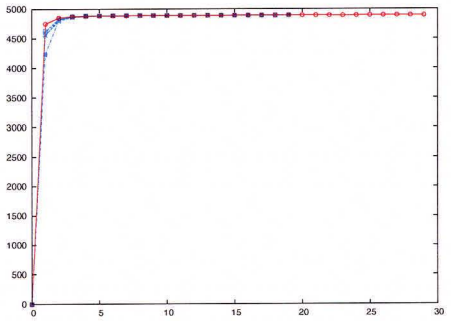
(b) 3-hop reachability



(c) 4-hop reachability

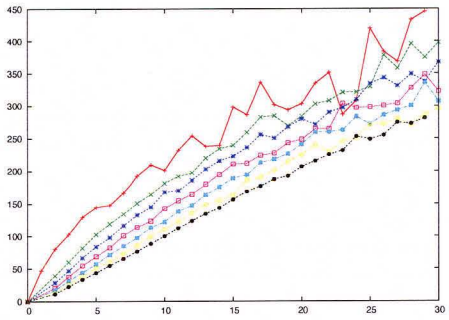


(d) 5-hop reachability

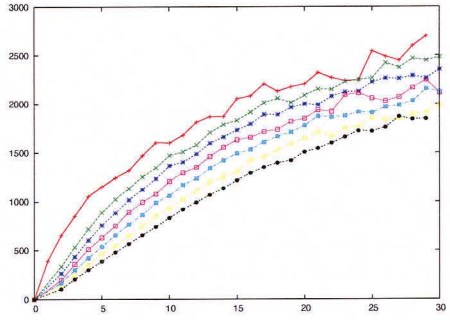


(e) 6-hop reachability

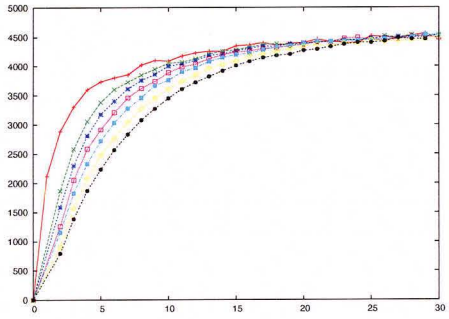
Figure 2.1: k-hop reachability ($k = 1 \ldots 6$) of five runs of Erdös-Rényi model (Blue) and Yeast PPI network (Red).



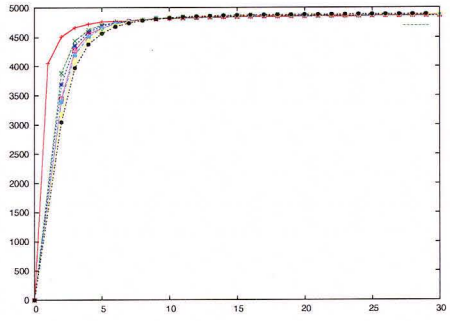
(a) 2-hop reachability



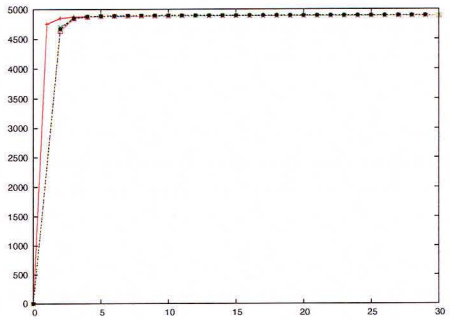
(b) 3-hop reachability



(c) 4-hop reachability



(d) 5-hop reachability



(e) 6-hop reachability

Figure 2.2: k-hop reachability of Yeast PPI network (red), 10% change in edges of Yeast PPI network (green), 20% change (blue), 30% change (purple), 40% change (light blue), 50% change (yellow) and 60% change (black)

2.2.4 Betweenness Distribution

The betweenness of a fixed node of a network measures the extent to which a particular point lies 'between' point pairs in the network $G = (V, E)$. The formal definition of betweenness is as the following. Let $\sigma_{x,y}$ be the number of shortest paths from $x \in V$ to $y \in V$ for all pairs of $x, y \in V$ (note that in undirected graphs $\sigma_{x,y} = \sigma_{y,x}$). Let $\sigma_{x,y}(v)$ be the number of shortest paths from $x \in V$ to $y \in V$ which go through node v. The betweenness $\mathrm{Bet}(v)$ of node v is now defined as:

$$\mathrm{Bet}(v) = \sum_{(i,j) \in V, i,j \neq v} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}}$$

In the same manner as k-hop reachability, betweenness feature must have two properties to be considered as an feature to compare PPI networks against graphs emulated using random graph generator models. First, we need to show that betweenness distribution of yeast PPI network has a different behavior from that of Erdös-Rényi networks Second, it must be shown that this feature is robust and sensitive to random change. In figure 2.3 we have illustrated both claims.

2.2.5 Closeness Distribution

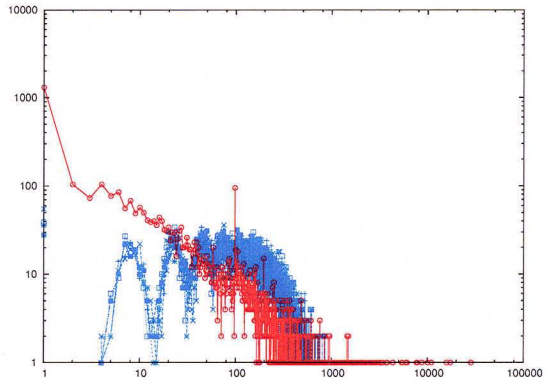
For all $x, y \in V$, we define $d_{x,y}$ as the length of the shortest path between x and y. The closeness of a node $v \in V$ is defined as

$$\mathrm{Cls}(v) = \frac{|V| - 1}{\sum_{i \in V} d_{v,i}}$$

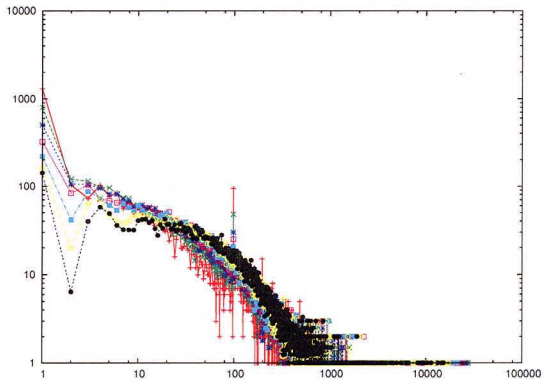
Here we will verify that closeness distribution also has the necessary properties to be considered as an feature for PPI network comparison. In figure 2.4, we show that Yeast PPI network has a different closeness distribution than Erdös-Rényi graphs. Moreover, we show that this feature is robust and sensitive to random changes.

2.2.6 Graphlet Frequency

The graphlet frequency has been introduced in [24] to compare the topological structure of networks. A graphlet is a small connected and induced subgraph of a large graph, for example a small triangle or a small clique. The *graphlet count (graphlet frequency)* of a

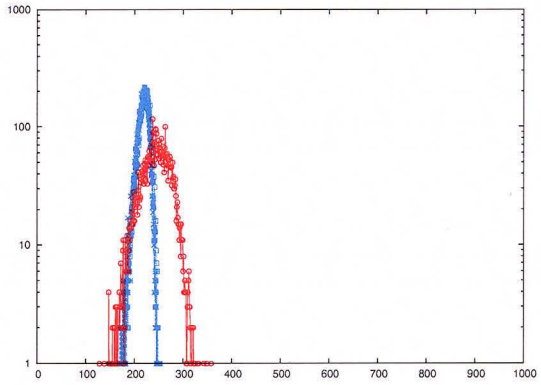


(a) Betweenness of Yeast vs. Erdös-Rényi

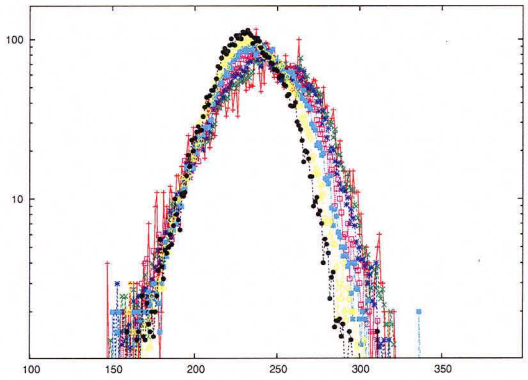


(b) Effect of Random Changes on Betweenness

Figure 2.3: (a) The comparison between betweenness distribution of Yeast PPI (red) and Eröds-Rényi random graph (blue) (b) The effect of random changes on betweenness distribution of Yeast (red), 10% change in edges of Yeast PPI network (green), 20% change (blue), 30% change(purple), 40% change (light blue), 50% change (yellow) and 60% change(black)



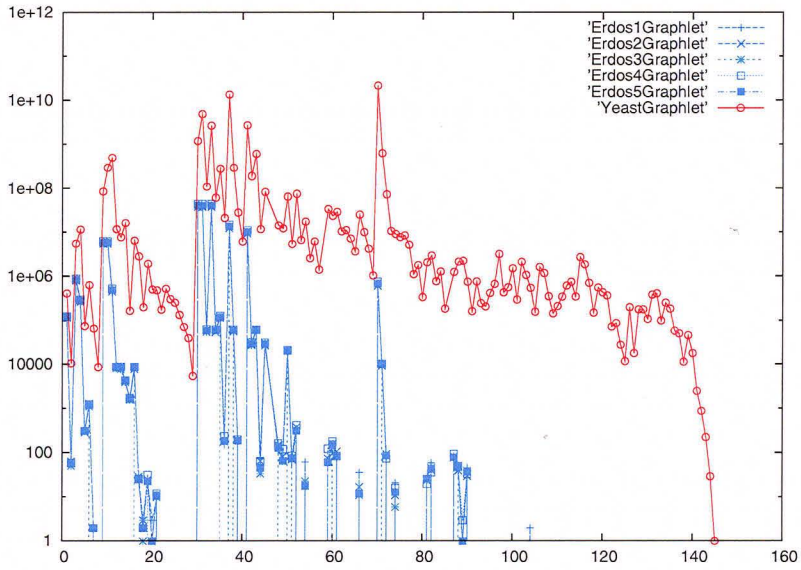
(a) Closeness of Yeast vs. Erdös-Rényi



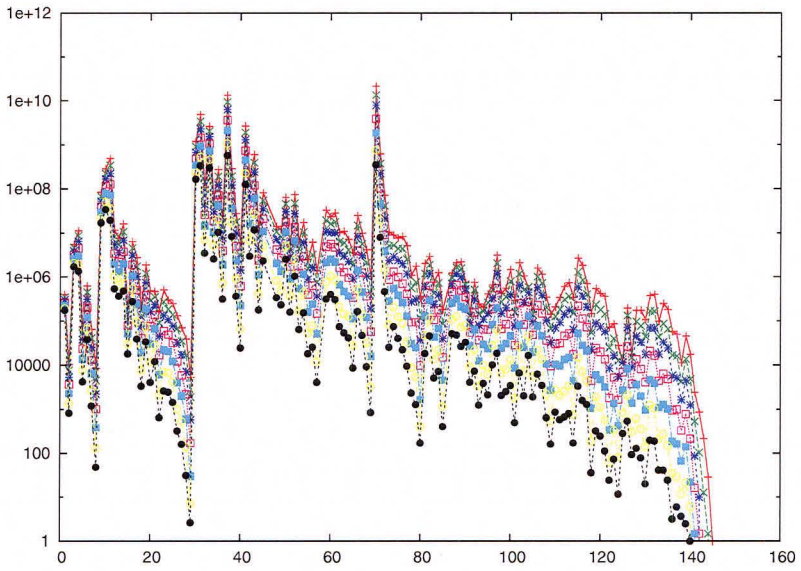
(b) Effect of Random Changes on Closeness

Figure 2.4: (a) The comparison between Closeness distribution of Yeast PPI (red) and Eröds-Rényi random graph (blue) (b) The effect of random changes on Closeness distribution of Yeast (red), 10% change in edges of Yeast PPI network (green), 20% change (blue), 30% change (purple), 40% change (light blue), 50% change (yellow) and 60% change (black)

given graphlet g with r nodes in a given graph $G = (V, E)$ is defined as the number of distinct subsets of V (with r nodes) whose induced subgraphs in G are isomorphic to g. In this paper we consider all 141 possible graphlets/subgraph topologies with 3, 4, 5 and 6 nodes. Furthermore, we consider cliques of sizes 7, 8, 9 and 10 (permutation of all graphlets is shown in A.1). Like all the previous features, graphlet frequency of Yeast PPI network must have different behavior form graphs generated using Erdös-Rényi graph generation model. In addition to being robust and sensitive to random changes. In Figure 2.5 we show that, Yeast PPI network has a different graphlet frequency than Erdös-Rényi graphs. In addition, it is shown that graphlet frequency is robust and sensitive to random change of network.



(a) Graphlet Frequency of Yeast vs. Erdös-Rényi



(b) Effect of Random Changes on Graphlet Frequency

Figure 2.5: (a) The comparison between Graphlet Frequency of Yeast PPI (red) and Eröds-Rényi random graph (blue) (b) The effect of random changes on Graphlet Frequency of Yeast (red), 10% change in edges of Yeast PPI network (green), 20% change (blue), 30% change (purple), 40% change (light blue), 50% change (yellow) and 60% change (black)

Chapter 3

Network Generation Models

There are many random network generator models, however, few of them achieve power-law degree distribution. Two of the most famous ones are *preferential attachment* and *duplication model*. Also, a new model has recently been proposed in [24] which claims to emulate the PPI networks accurately, called *random geometric* model. In the rest of these thesis we will focus on these three models.

The preferential attachment and duplication network models both start with a small seed graph and add one node to it in each iteration. In contrast, the random geometric model creates all the nodes in the first step, and then adds edges.

Let $G(t) = (V(t), E(t))$ be the graph at the end of time step t, where $V(t)$ is the set of nodes and $E(t)$ is the set of edges/connections. Let v_t be the node generated in time step t. Given a node v_τ, we denote its degree at the end of time step t by $d_t(v_\tau)$. In the coming sections we will explain each of these models in detail.

3.1 Preferential Attachment Model

As explained previously preferential attachment model is an iterative model, which achieves power-law degree distribution. The preferential attachment model was analyzed in [1], [6], [8] ,[10]. In step t it generates v_t and connects it to every other node v_τ independently with probability $c \cdot d_{t-1}(v_\tau)/2|E(t-1)|$, where c is the average degree of a node in G; i.e. v_t prefers to connect itself to high degree nodes.

3.2 Duplication Model

This model is based on Ohno's hypothesis of genome evolution [7], [23], [30]. In iteration t, a node v_τ of $G(t-1)$ is picked uniformly at random and "duplicated", i.e. an exact copy of v_τ as v_t is generated (including its edges). The model then updates v_t's edges, first by deleting each of its edges with probability $(1-p)$, then by connecting each node $v_{t'}$ (except the neighbors of v_τ) to v_t independently with probability $r/|V(t)|$. Here, p and r are user defined parameters. Much of the earlier work on the duplication model aim to maintain a constant average degree throughout the generation of the network; this is achieved by setting $r = 1 - 2p$.

As mentioned earlier, the degree distribution of the preferential attachment model as well as the duplication model asymptotically approaches a power law [1], [8], [10], [9]. More specifically, in the log-log scale, it forms a straight line (this is valid for only "high degree" nodes) whose slope is independent of the seed graph and a function of the values of p and r for the duplication model or c for the preferential attachment model. Thus, the two iterative models are equivalent with respect to the degree distribution.

Both the preferential attachment and the duplication model produce many *singletons* [1] [4]. Singletons are nodes which are not connected to any other node. Unfortunately there are no known bounds on the number of generated singletons in the duplication model. In the duplication model, for the special case $r = 0$, $p = 1/2$, the proportion of singletons asymptotically approaches 1. Also, we observed in our experiments that for other values of p and r, number of singletons generated by duplication model is much higher than number of singletons in known PPI networks (in PPI networks number of singletons are very small).

3.2.1 Modified Duplication Model.

It is well known that the number of singletons in PPI networks are quite limited. This does not come as a surprise as genes with no functionality are not conserved during evolution. Thus a slightly modified duplication model which deletes each singleton node as soon as it is generated may better emulate the growth of PPI networks. This model has also been shown to achieve a power law degree distribution [4].

[1] We also note that the known PPI networks have several self loops. Both the preferential attachment and the duplication models can be modified slightly to produce such self loops(homo dimers).

Unfortunately, similar to the number of singletons in duplication model, in modified model the total number of generated nodes is not known. Moreover, it is not known which values of p and r ensure that the expected average degree is constant through all iterations. In Section 3.2.2 we derive conditions on p and r that are necessary for having a constant expected degree. We later use the derived relationship between p and r so that the modified duplication model can well approximate the desired average degree as well as the degree distribution of the PPI networks under investigation.

3.2.2 The Parameters of the Modified Duplication Model

Here we show how to determine conditions on deletion probability $1-p$ and insertion probability r so that the expected average degree of the network can be set to any given value. [2] For this, we make the the assumption that the degree frequency distribution and the average degree of nodes are fixed asymptotically once the values of p and r are determined. Let $G(t) = (V(t), E(t))$ be the network generated by the modified duplication model and let $n(t) = |V(t)|$ and $e(t) = |E(t)|$. Also, let $n_k(t)$ be the number of nodes in time step t with degree k and $a(t)$ be the average degree of nodes in $G(t)$. Finally let $P_k(t) = n_k(t)/n(t)$, the frequency of nodes with degree k at time step t. We assume that $P_t(k)$ is asymptotically stable, i.e. $P_k(t) = P_k(t+1)$ for all $1 \leq k \leq t$ for sufficiently large values of t. In other words we assume that $P_k(t) = d_k$ for some fixed d_k. By definition

$$a(t) = \sum_{k=1}^{t} k \cdot \frac{n_k(t)}{n(t)} = \sum_{k=1}^{t} k \cdot P_k(t) = \sum_{k=1}^{t} k \cdot d_k.$$

Now we can calculate the average degree $a(t+1)$ under the condition that degree frequency distribution is stable and $a(t) = a$, a constant.

$$Exp[e(t+1)] = e(t) + \sum_{k=1}^{t} k \cdot P_k(t) \cdot p + r = \frac{n(t) \cdot a(t)}{2} + p \cdot a(t) + r.$$

Let $Pr_s(t)$ be the probability that v_{t+1} ends up as a singleton.

$$Pr_s(t) = \sum_{k=1}^{t} P_k(t) \cdot (1-p)^k \cdot \left(1 - \frac{r}{n(t)}\right)^{n(t)-k} \approx \sum_{k=1}^{t} d_k \cdot (1-p)^k \cdot \frac{1}{e^r}.$$

[2] There might be an alternative method which combines different biological factors to determine the values of p and r, such as mutation probabilities, probability of preserving interactions, etc.

Since this probability does not depend on t asymptotically, we can set $Pr_s(t) = Pr_s$. Now we can calculate the expected number of nodes and the expected number of edges in step $t+1$.

$$Exp[n(t+1)] = Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1).$$

$$Exp[e(t+1)] = Exp\left[\frac{n(t+1) \cdot a(t+1)}{2}\right] = \frac{a}{2} \cdot Exp[n(t+1)]$$

$$Exp[e(t+1)] = \frac{a}{2} \cdot (Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1)).$$

Comparing the above equation with the first equation for $Exp[e(t+1)]$ we get

$$\frac{a}{2} \cdot (Pr_s \cdot n(t) + (1 - Pr_s) \cdot (n(t) + 1)) = \frac{n(t) \cdot a(t)}{2} + p \cdot a(t) + r = \frac{n(t) \cdot a}{2} + p \cdot a + r.$$

Solving the above equation results in $a = 2r/(1 - Pr_s - 2p)$ where Pr_s is a function of p, r and d_k only.

The discussion above demonstrates that the two key parameters p and r of the (modified) duplication model are determined by the degree distribution (more specifically the slope of the degree distribution in the log-log scale) and the average degree of the PPI network we would like to emulate. Perhaps due to the strong evidence that the seed network does not have any effect on the asymptotic degree distribution [5], the role of the seed network (the only free parameter remaining) in determining other topological features of the duplication model has not been investigated.

3.3 Random Geometric Model

The random geometric graph [24] generation model generates the graph in a bounded ℓ-dimensional Euclidean space (ℓ is typically 3 or 4). It picks independently and uniformly at random points from the underlying space and assigns a node to every point picked. Two nodes are connected if the (Euclidean) distance between the nodes is at most some user defined value r. If we want the generated graph to have a particular average degree we need to pick the r value in a matter so that we can achieve the desired average degree.

Chapter 4

Scale-Free Models Comparison

The main results that we show is that, first, although Duplication Model and Preferential Attachment both generate graphs with power-law degree distribution, however other features of networks generated using these two models are different. Also, we show that the networks generated using duplication model is dependent to the initial "seed network" it starts with, although the degree distribution of networks generated is asymptotically independent of "seed network".

4.1 Duplication against Preferential Attachment Model

In this section we show that the modified duplication model and the preferential attachment model with similar degree distributions may have very different k-reachability, betweenness distribution, closeness distribution and graphlet frequency, thus considering only one of these models as a representative of "scale free" networks can be misleading.

Figure 4.1 depicts the degree distribution, k-hop reachability, closeness distribution, betweenness distribution and graphlet frequency of the duplication model and the preferential attachment model with 4902 nodes (as per they Yeast PPI network [28]). We set $r = 0.12$, $p = 0.365$ and $d = 7$ so that the average degree of nodes in both models is 7 (again as per the Yeast PPI network [28]). Figure 4.1 compares the k-hop reachability achieved by the two models for $k > 1$. As can be seen, the k-hop reachability is quite different especially for $k = 3, 4$. Figure 4.1 also shows how the graphlet distributions differ, especially for dense graphlets (e.g graphlets 17-29 and 85-145). For instance, as it can be seen in

Figure 4.1 duplication model generates around 500 times more $K_{2,4}$[1] subgraphs than preferential attachment. In terms of betweenness and closeness there are some differences as well.

4.2 The Effect of the Seed Network in Shaping the Topological Behavior of the Duplication model and Preferential Attachment Model

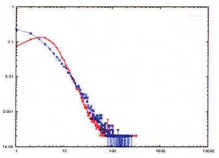
We now show that the seed network has a key role in characterizing the topology of the duplication model. Figure 4.2 depicts how various topological features of duplication models with identical parameters ($p = 0.365$ and $r = 0.12$) but different seed graphs vary. The first seed graph (red) is obtained by highly connecting two cliques of respective size 10 and 7 by several random edges. To reduce the average degree some additional nodes were generated and randomly connected to one of the cliques. The second seed graph (blue) is obtained by enriching a ring of 17 nodes by random connections so as to make the average degree match that of the first seed graph. The third seed graph (green) is formed by sparsely connecting two cliques of respective sizes 10 and 7 with some added nodes randomly connected to one of the cliques.

All three networks were grown until both had 4902 nodes as per the Yeast PPI network [28]. (We depict the average behavior of five independent runs of each of the models.) It can be observed that although all of them have very similar degree distributions, their graphlet distributions(Figure 4.2(i)) may be quite different, especially for dense graphlets. Note that the figure 4.2(i)and 4.2(g) are in logarithmic scale and seemingly small variations in the figure may imply several factors of magnitude of a difference between the two distributions. Figure 4.2 also compares the k-hop reachability, closeness and betweenness distributions. As can be seen the k-hop reachability and the closeness distribution can vary considerably.

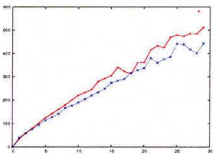
Now a very crucial and difficult question to answer is that how we should pick a "right" seed network to generate PPI networks using duplication model. We will return to this question in Section 5.3.1, and try to use a heuristic method to guide us in choosing the seed network.

For preferential attachment model, we have tried different seed networks and it seems

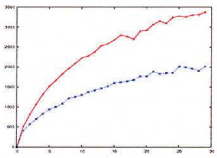
[1] $K_{2,4}$ has graphlet id of 115



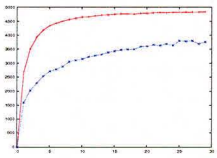
(a) Degree distribution



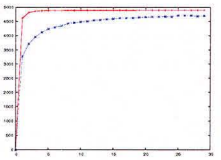
(b) 2-hop reachability



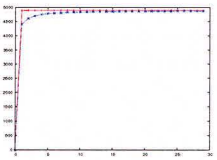
(c) 3-hop reachability



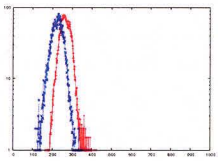
(d) 4-hop reachability



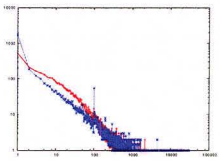
(e) 5-hop reachability



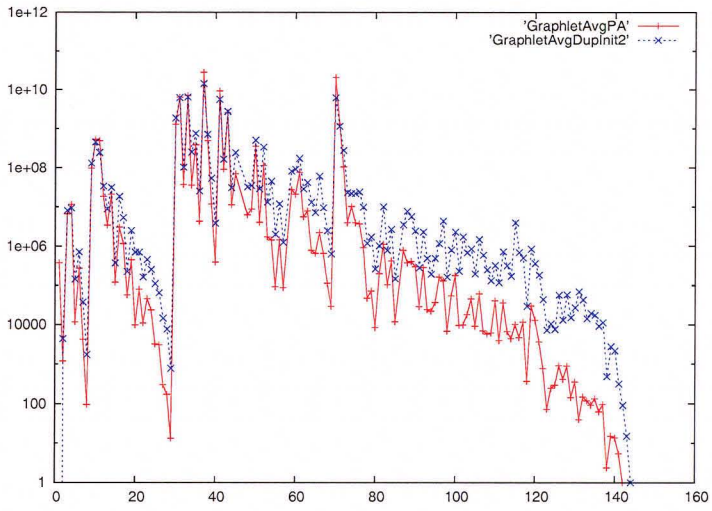
(f) 6-hop reachability



(g) Closeness

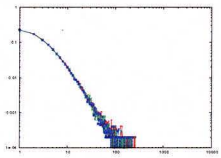


(h) Betweenness

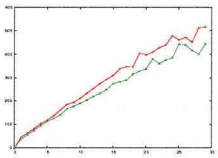


(i) Graphlet distribution

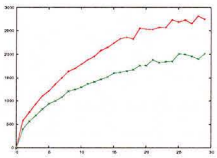
Figure 4.1: Degree distribution, k-hop reachability, graphlet, closeness and betweenness distributions of the preferential attachment model (red) and the duplication model (blue).



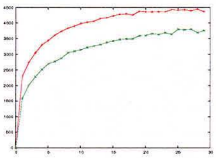
(a) Degree distribution



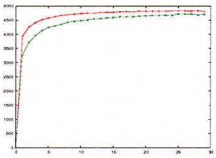
(b) 2-hop reachability



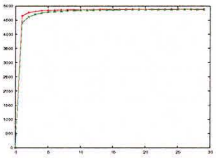
(c) 3-hop reachability



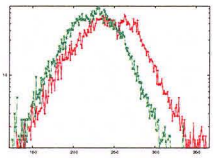
(d) 4-hop reachability



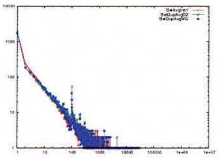
(e) 5-hop reachability



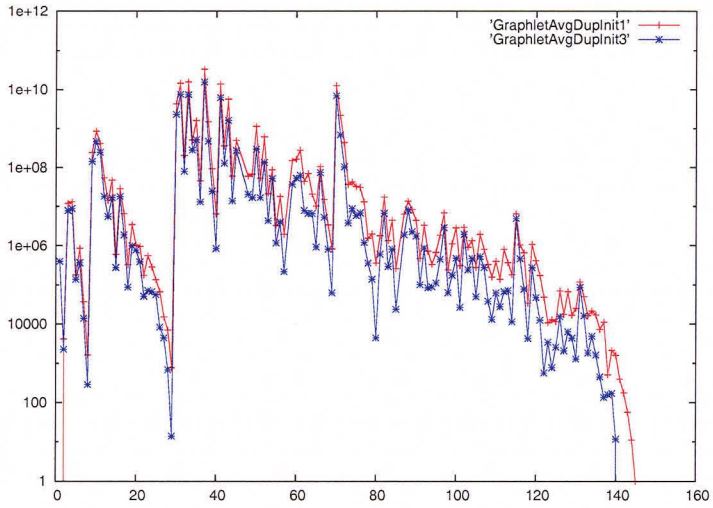
(f) 6-hop reachability



(g) Closeness



(h) Betweenness



(i) Graphlet distribution

Figure 4.2: The effect of the seed network on the degree distribution, k-hop reachability, graphlet, closeness and betweenness distributions. Each color (red, blue, green) depicts the behavior of a network with a particular seed graph. The parameters p and r are identical in all three models.

to be robust to seed network change. in another words, all the features we used were independent of the seed network for preferential attachment model.

Chapter 5

Emulating Yeast PPI Network

In this chapter we will look at how networks generated by different random network generation models compare with Yeast PPI network. The models we will consider here are preferal attachment, random geometric model and duplication model(modified).

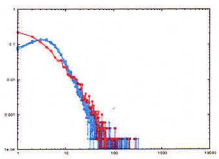
5.1 Preferal Attachment against Yeast PPI Network

In this section, we are going to compare the networks generated using preferal attachment with the Yeast PPI network available. The average degree of the Yeast PPI network is 7, hence we have to pick the preferal attachment parameters so that we can achieve the same average degree. As explained in subsection 3.2.2, the parameter that we need to decide on is $c = 7$.

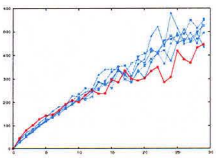
In Figure 5.1 we give the comparison of networks generated using preferal attachment with the Yeast PPI network.

5.2 Random Geometric Network against Yeast PPI Network

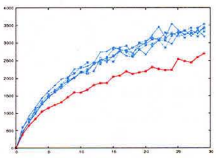
In this section we will compare the graphs generated using the random geometric model in euclidean 4-dimension space. The value r is chosen in such that it achieves average degree 7(the same as Yeast PPI network). In figure 5.2 comparison of graphs generated using random geometric model and Yeast PPI network is shown.



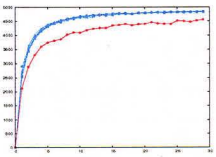
(a) Degree distribution



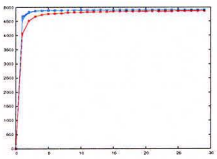
(b) 2-hop reachability



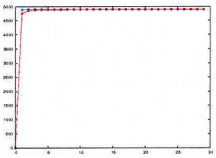
(c) 3-hop reachability



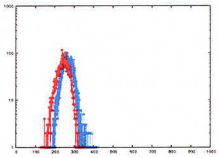
(d) 4-hop reachability



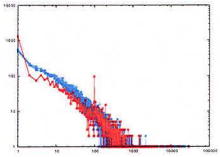
(e) 5-hop reachability



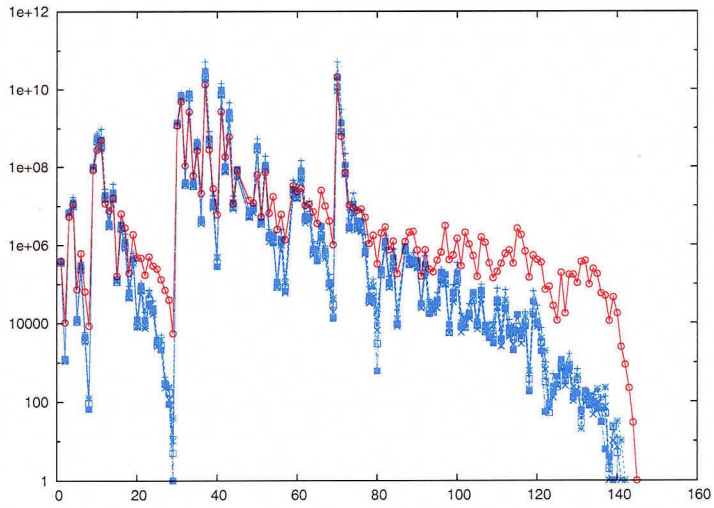
(f) 6-hop reachability



(g) Closeness

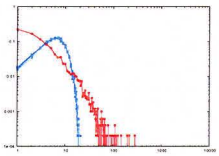


(h) Betweenness

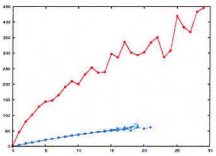


(i) Graphlet frequency

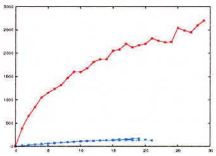
Figure 5.1: The degree distribution, the k-hop reachability, the graphlet, closeness and betweenness distributions of the Yeast PPI (red) network against five independent runs of the preferal attachment model (blue).



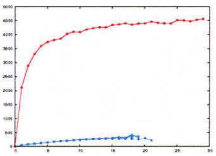
(a) Degree distribution



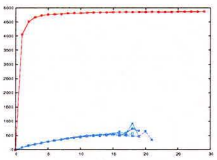
(b) 2-hop reachability



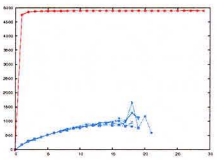
(c) 3-hop reachability



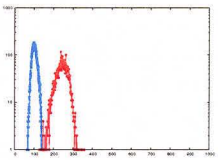
(d) 4-hop reachability



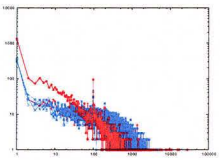
(e) 5-hop reachability



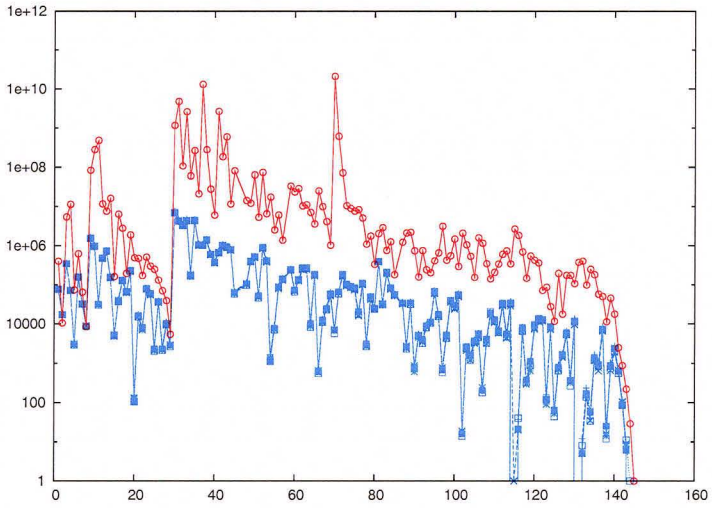
(f) 6-hop reachability



(g) Closeness



(h) Betweenness



(i) Graphlet frequency

Figure 5.2: The degree distribution, the k-hop reachability, the graphlet, closeness and betweenness distributions of the Yeast PPI (red) network against five independent runs of the random geometric model(blue).

5.3 Duplication Model (Modified) against Yeast PPI Network

As we have seen in previous chapter two networks generated by the same duplication model (and hence have very similar degree distributions) can differ substantially in terms of the other topological measures, if their seed networks are different.

If the seed selection makes such a difference in shaping the topology of the generated network, is it possible to select the "right" seed network [1] so that all interesting topological features of the PPI network in question can be captured [2]?

We answer this question positively by demonstrating that carefully chosen seeds can result in a network that is very similar to PPI networks we considered in terms of all of the above distributions.

The PPI networks we tested include (the largest connected component of) the complete Yeast PPI network [28] with 4902 proteins and 17200 edges (as of Jul 2006).

5.3.1 Seed Network Selection and Emulation of Yeast PPI Network via (Modified) Duplication Model

We were able to closely approximate the features of complete Yeast PPI network via the (modified) duplication model, using a specific seed graph that (to a great extent) exists in the Yeast PPI network as a subgraph. The choice of the seed graph was based on the observation that the duplication model is very unlikely to generate "large" cliques[3]. However, the Yeast PPI network includes a clique with 10 nodes, which, as a result, must be included in the seed network. There are other smaller cliques in Yeast PPI network which are represented in the seed graph as a single independent clique with 7 nodes. These two cliques were highly connected in the seed network, which also included a few additional nodes sparsely connected to the two cliques (the total number of nodes was 50) so that the normalized degree distribution of the Yeast PPI network was similar to that of the seed graph. This ensured that the (normalized) degree distribution of the Yeast PPI network as

[1] By "right" seed network we mean, a seed network, which partially exists in the PPI network we want to emulate, and the result of emulating a graph via duplication model using the chosen seed network to be close to real PPI network

[2] A combinatorial method to choose the best seed network has not yet been developed, and this can be one way that this project can be extended.

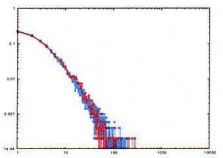
[3] By large cliques we mean its size should be bigger than 5 or 6 nodes

well as its clique frequency distribution (which turns out to be an important determinant of the overall graphlet distribution) were similar to that of the seed graph.

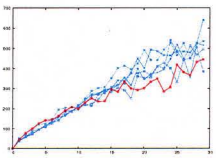
There are two additional parameters associated with the duplication model: p, the edge maintenance probability and r, the edge insertion probability (see Methods and Materials). These two parameters alone determine the (asymptotic) degree distribution and the average degree of the generated network. We chose $p = 0.365$ and $r = 0.12$ so that the degree distribution of the duplication model matches that of the Yeast PPI network.

We used the duplication model to generate 5 independent networks each with 4902 vertices. The resulting networks are compared to the Yeast PPI network in terms of the k-hop reachability, the graphlet , betweenness, and closeness distributions in Figure 5.3. Under all these measures, the Yeast network is very similar to those produced by the duplication model. In fact the duplication model we consider here provides much better fits to both the k-hop degree distribution and the graphlet distribution of the Yeast network than the random graph models described in of [5] and [24] - which were specifically devised to capture the respective features of PPI networks.

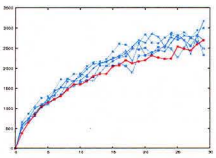
As it can be seen in Figures 5.1,5.2 and 5.3 it is obvious that *duplication model with right choice of initial seed network gives results much closer to the real Yeast PPI network.*



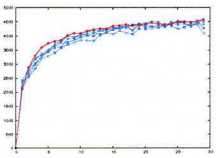
(a) Degree distribution



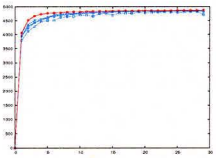
(b) 2-hop reachability



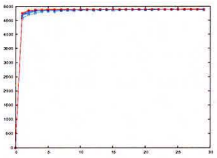
(c) 3-hop reachability



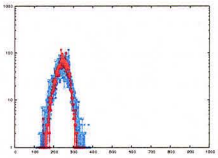
(d) 4-hop reachability



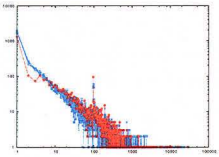
(e) 5-hop reachability



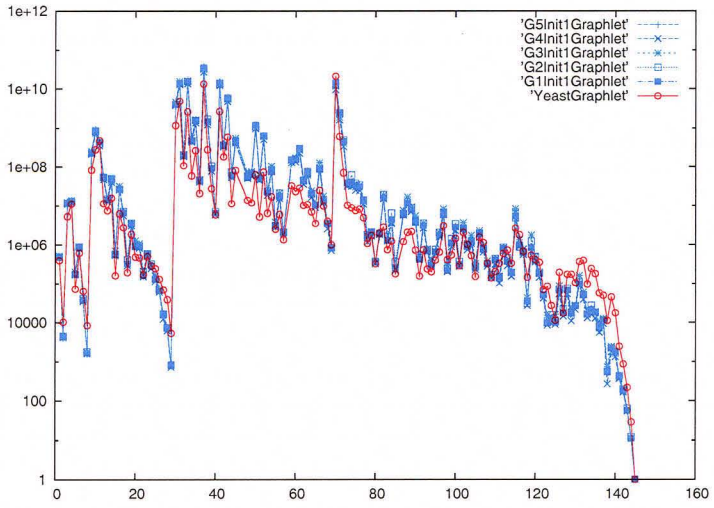
(f) 6-hop reachability



(g) Closeness



(h) Betweenness



(i) Graphlet frequency

Figure 5.3: The degree distribution, the k-hop reachability, the graphlet, closeness and betweenness distributions of the Yeast PPI (red) network against five independent runs of the duplication model (blue).

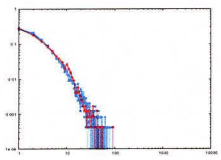
Chapter 6

Emulating PPI Networks via Duplication Model

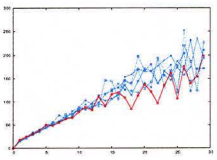
As seen in previous chapter duplication model was much better fit than other models considered when comparing their results against Yeast PPI netwrok [12]. In this chapter we want to compare results of duplication model against two other available PPI's. First, we compare the networks created using duplication model with Core Yeast PPI network and, second, we will compare the duplication networks against Worm network. In both experiments we see a very nice fit of the features, which further increases our confidence on duplication model.

6.1 Duplication Model against Core Yeast Network

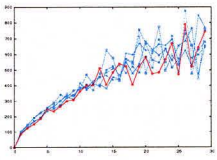
We provide some additional evidence on the power of the duplication model in capturing the topological features of available PPI networks. We first compare the duplication model with the main component of the CORE subset of Yeast network. The CORE subset contains the pairs of interacting proteins identified in Yeast that were validated according to the criteria described in [12]. It involves 2345 nodes and 5609 edges. The values of r and p were set to $r = 0.12$, $p = 0.322$ as prescribed by the average degree formula $a = 2r/(1 - P_s - 2p)$ and the fact that P_s is a function of r and p. The seed network we used was very similar to that used for the complete Yeast network. The results are shown in Figure 6.1.



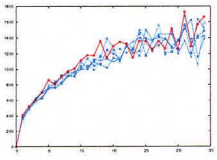
(a) Degree distribution



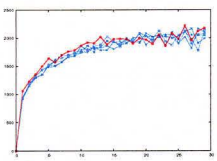
(b) 2-hop reachability



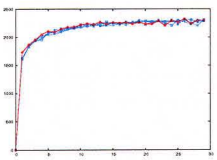
(c) 3-hop reachability



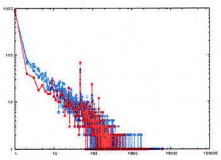
(d) 4-hop reachability



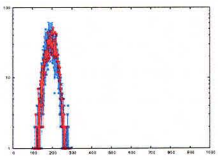
(e) 5-hop reachability



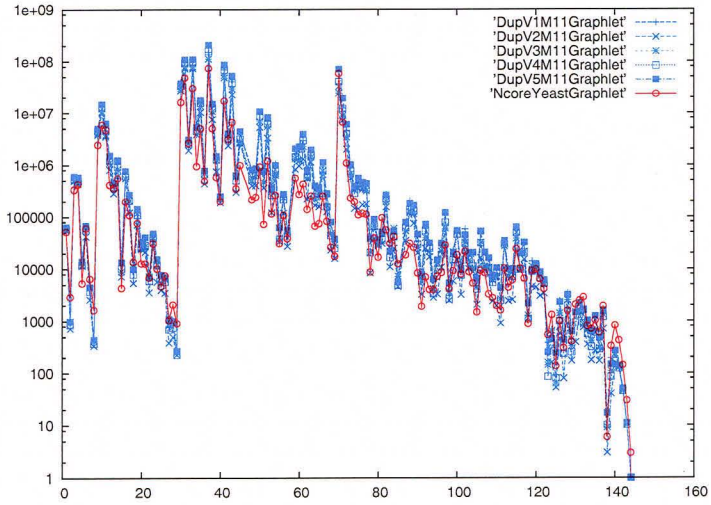
(f) 6-hop reachability



(g) Betweenness



(h) Closeness

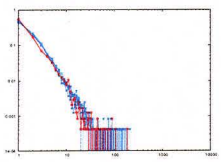


(i) Graphlet frequency

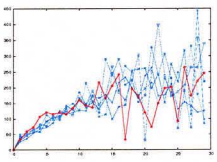
Figure 6.1: The topological properties of the duplication model (blue) compared to that of the CORE Yeast Network (red). The degree distribution, the k-hop reachability, graphlet, betweenness and closeness distributions of both networks are shown. The values obtained by five independent runs of the duplication model are given.

6.2 Duplication Model against Worm Network

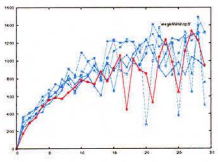
We compare the duplication model with the Worm PPI network [28] as well. This network is less developed than the Yeast network with only 2387 nodes and 3825 edges. The r and p values used for this network are $r = 0.12$, $p = 0.322$. The seed network used was again very similar to that used for the Yeast network. The comparative results are shown in Figure 6.2.



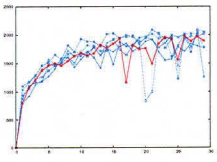
(a) Degree distribution



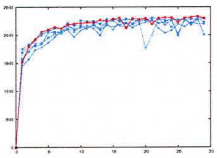
(b) 2-hop reachability



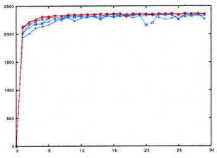
(c) 3-hop reachability



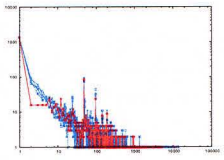
(d) 4-hop reachability



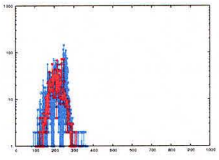
(e) 5-hop reachability



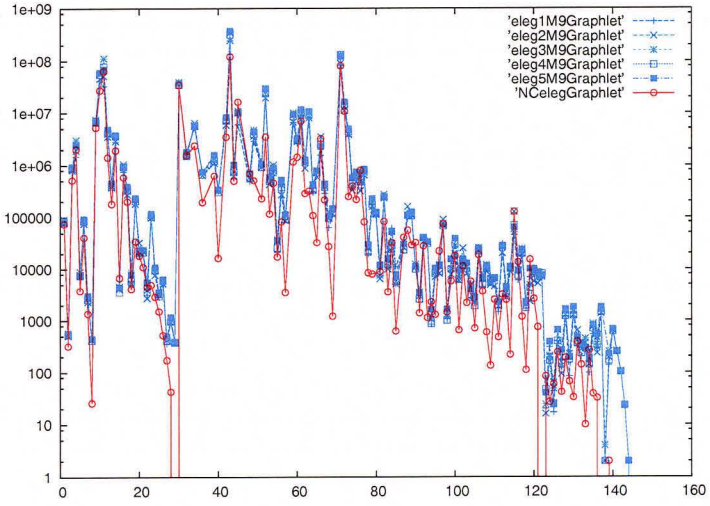
(f) 6-hop reachability



(g) Betweenness



(h) Closeness



(i) Graphlet frequency

Figure 6.2: The topological properties of the duplication model (blue) compared to that of the *C.Elegans* (Worm) network (red). The degree distribution, the k-hop reachability, graphlet, betweenness and closeness distributions of both networks are shown. The results of five independent runs of the duplication model are depicted.

Chapter 7

Conclusion

In this thesis comparison and emulation of protein protein interaction networks was studied. First problem that was addressed is how to compare graphs, knowing that no efficient algorithm for graph isomorphism or approximately graph isomorphism problems is known. In chapter 2, five different features (which each try to capture unique properties of PPI networks) for the purpose of comparing networks was explained.

In chapter 3 three different graph emulation models have been described. Two of them, preferential attachment and duplication model, produce graphs with power-law degree distribution. The third model is random geometric, which is not a scale-free emulator. However, it has been claimed in [24] to be a suitable emulator of PPI networks.

After showing the network generated using duplication model is dependent to initial seed network, we illustrated that duplication model can achieve a much closer network to yeast PPI (considering the features) than random geometric and preferential attachment models (chapter 5).

7.1 Contributions

The main contributions we made in this thesis are as follow: First, we selected a set of features from available literature on PPI network comparison and social network comparison and tested each feature (1) to capture a unique property of PPI networks,and (2) being robust to minor changes. Second, we choose a slightly modified version of duplication model (modified duplication model), and analyzed the relationship between p and r for it to achieve a desired stable average degree. Third, we showed that although preferential attachment

model and duplication model have almost the same degree distribution (power-law), the graph these models produce can be very different when other features are considered. Fourth, the most important contribution of this thesis is that we showed (modified) duplication model is dependent to the initial seed network it starts with (although its degree distribution is independent from the network chosen). Last but not least we were able to use (modified) duplication model with carefully chosen seed networks (see Section 5.3.1) to emulate Yeast PPI network, Core Yeast PPI network and Worm PPI network with high similarity to real networks [1].

7.2 Future work

There are several directions to which this work can be extended. One interesting path, is to analytically study the effect of the seed network on the graphs emulated using duplication model. We can also try to come up with a systematic way to generate the "correct" seed network for each given PPI network. Another challenging problem is to explore whether the duplication model is able to produce close networks to other available PPI networks. Finally it is challenging to find other features of PPI network that can be used to compare networks generated using network emulators against real PPI networks.

[1] similarity comparison is made using the selected features

Appendix A

Graphlets

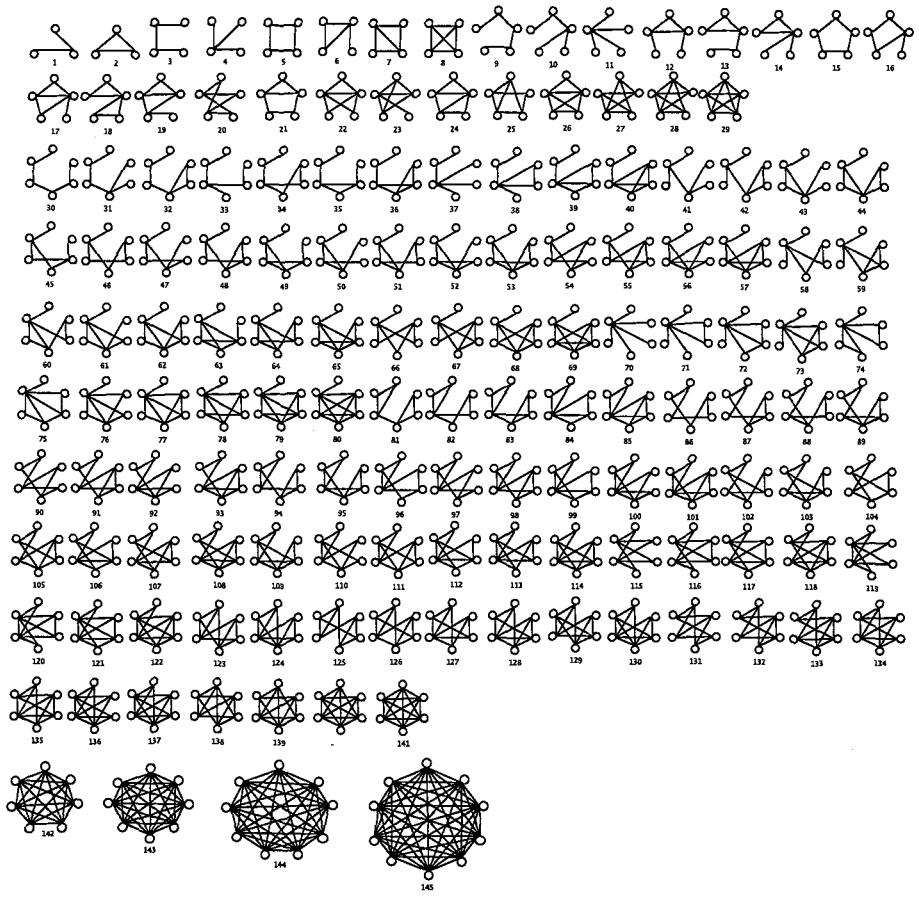


Figure A.1: Graphlets

Bibliography

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. In *Proceedings of ACM STOC*, pages 171–180, 2000.

[2] C. Alfarano and et al. The biomolecular interaction network database and related tools. *Nucleic Acids Research*, 33:418–424, 2005.

[3] A.-L. Barabási and R. A. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, and S.C Sahinalp. The degree distribution of the general duplication models. *Theoretical Computer Science.*

[5] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, and S.C Sahinalp. Topological properties of proteome networks. In *Proceedings of RECOMB satellite meeting on System Biology.* LNBI,Springer, 2005.

[6] N. Berger, B Ballobás, C. Borgs, J. Chayes, and O. Riordan. Degree distribution of the fkp network model. In *Proceedings of the ICALP, LNCS*, volume 2719, pages 725–738. Springer, 2003.

[7] A. Bhan, D. J. Galas, and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatis*, 18:1486–1493, 2002.

[8] B. Bollobás, O Riordan, J. Spencer, and G. Tusanády. The degree sequence of a scale-free random graph process. *Random Struct. Algorithms*, 18:279–290, 2001.

[9] F. Chung, L. Lu, and D.J. Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10:677–687, 2003.

[10] C. Cooper and A. Frieze. A general model of webgraphs. *Random Struct. Algorithms*, 22:311–335, 2003.

[11] E. De Silva and M.P.H. Stumpf. Complex networks and simple models in biology. *Journal of the Royal Society Interface*, 2:419–430, 2005.

[12] CM. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: Two methods for assessment of reliability of high-throughput observation. *Molecular and Cellular Proteomics*, 1:349–356, 2002.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[14] R. Ferrer i Cancho and C. Janssen. The small world of human language. In *Proceedings of Royal Society of London B*, volume 268, pages 2261–2266, 2001.

[15] J. Han, D. Dupuy, N. Bertin, M. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotech*, 23:839–844, 2005.

[16] H. Hermjakob and et al. Intact - an open source molecular interaction database. *Nucleic Acids Research*, 32:452–455, 2004.

[17] F. Hormozdiari, P. Berenbrink, N. Przulj, and S.C Sahinalp. Not all scale free networks are born equal: the role of the seed graph in ppi network emulation. In *Proceedings of RECOMB satellite meeting on System Biology*, 2006.

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

[19] I. Jurisica and D. Wigle. *Knowledge Discovery in Proteomics.* CRC press, 2006.

[20] J. Kleinberg, R. Kumar, PP. Raphavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models and methods. In *Proceedings of COCOON*, pages 1–17, 1999.

[21] R. Kumar, P. Raghavan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of FOCS*, pages 57–65, 2002.

[22] Ohno. *Evolution by gene duplication.* Springer, 1970.

[23] R. Pastor-Satorras, E. Smith, and R.V. Sole. Evolving protein interaction networks through gene duplication. *Journal of Theoretical biology*, 222:199–210, 2003.

[24] N. Przulj and et al. Modeling interactome: Scale-free or geometric?,. *Bioinformatics*, 150:216–231, 2005.

[25] T. Przytycka and Y.K. Yu. Scale-free networks versus evolutionary drift. *Computational Biology and Chemistry*, 28:257–264, 2004.

[26] S. Redner. How popular is your paper? an empirical study of the citations distribution. *European Physical journal B*, 4:131–134, 1998.

[27] Erdös & Rényi. On random graphsI. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[28] L. Salwinski and et al. The database of interacting proteins:2004 update. *Nucleic Acid Research*, 32:449–451, 2004.

[29] R. Tanaka and et al. Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579:5140–5144, 2005.

[30] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modelling of protein interaction networks. *Complexus*, 1:38–44, 2003.

[31] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18:1283–1292, 2001.

[32] S. Wasserman and K. Faust. *Social network analysis: methods and applications.* Cambridge University Press, 1999.

[33] D.J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness.* Princeton University Press, 1999.

[34] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[35] I. Xenarios and et al. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30:303–305, 2002.

[36] A. Zanzoni and et al. Mint: a molecular interaction database. *FEBS Letters*, 513:135–140, 2002.