SOME PROBLEMS IN PROTEIN-PROTEIN INTERACTION NETWORK GROWTH PROCESSES

LI SI

(B.Sc.(Hons.), SYSU)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

2013

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Li Si

12 July 2013

Acknowledgements

I would like to express my gratitude to my parents and my family. They have helped me throughout my education. Without them, this journey of pursuing my Ph.D degree would be impossible.

I would also like to thank my supervisor Associate Professor Choi Kowk Pui and my co-supervisor Associate Professor Zhang Louxin for their continuous encouragement, support and guidance during the past five years. Special thanks to Dr. Wu Taoyang for helpful suggestions and cooperation.

I also thank all the members in our computational biology group for useful presentations and idea sharing. Thanks to them, I have broadened my knowledge.

This list is by no means complete. I thank all the people who have helped me directly or indirectly.

Contents

Summary

The purpose of this thesis is to investigate the protein-protein interaction (PPI) networks via network growth modeling: The duplication models. The duplication models are biologically reasonable and have been proved to give good fit for real PPI networks. We have studied the evolutionary processes in two aspects: The forward and the backward. Specifically, for the forward, time increases and a network grows; for the backward, time decreases and a network is traced back.

We have studied one question in the backward aspect: What is the evolutionary history of an observed network? We answered this question by introducing a novel framework which incorporates the duplication forest to reconstruct the network evolutionary history. Under this framework, we reduced the searching space for reconstruction by simplifying the likelihood ratio between two histories. We proposed two algorithms: CherryGreedy (CG) and MinimumLossNumber (MLN) for reconstructing network evolutionary history. MLN is based on a more intuitive method and CG aims to provide more accurate results. Simulations show that our algorithms outperform others. Our algorithms were used to investigate the properties of real PPI networks from the view of evolution.

We have studied two questions in the forward aspect: (i) What is the degree distribution of a network when time is sufficiently large? and (ii) How does the seed graph affect the evolutionary process of a network? For (i), we have done rigorous mathematical analysis for the degree distribution of the partial duplication (PD) model. First the existence of the limiting degree distribution was established. A phase transition point for the PD model was showed. Moreover, the convergence rates and the connected components have also been analyzed. For (ii), we have run simulations to explore the topological statistics of four duplication models. Several features have been presented. This part provides an open direction for future work.

List of Figures

List of Tables

Chapter 1

Introduction

Functioning of a living cell is attributed to the interplay between its numerous components, such as DNA, RNA and proteins [9]. Despite their importance to biological systems, none of these molecules can individually execute the complex biological processes without collaboration with others. Therefore, understanding the interaction and regulation of molecules is crucial in modern biology [110]. In a conceptual and reductionism framework, there is a need to study the structure and the dynamics of biological networks.

A network is a mathematical object which consists of a set of nodes and a set of edges between them (see Subsection 1.1.1 for details). Depending on the molecules represented by nodes and the interactions by edges, molecular networks can be catalogued as metabolic networks, protein-protein interaction (PPI) networks and gene regulatory networks etc. [25, 97] (Fig. 1.1). For example, in a metabolic network, nodes correspond to biochemical metabolites and edges are chemical reactions that convert the reaction partners into substrates [25]. It should be kept in mind that all these biological networks overlap with each other and none of them stands alone in a living cell.

In the past decades, the advent of high-throughput experimental methods such

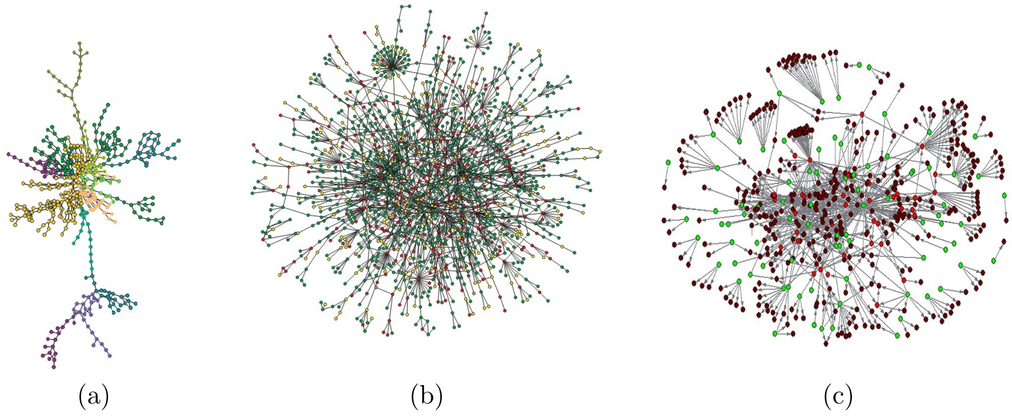

Figure 1.1: Examples of biological networks. (a) A metabolic network of *E. coli* with 574 interactions and 473 metabolites colored according to the KEGG pathway classification [38]. (b) Yeast PPI network. Color of a node indicates its lethality [47]. (c) *E. coli* transcriptional regulatory network with transcription factors colored with green and regulators colored with brown[39].

as yeast two-hybrid (Y2H) [30] and microarray [3] leads to the tremendous increase of biological interaction data, allowing studies attempting to reveal the design principles and evolutionary forces underlying biological networks [92]. Nonetheless, in spite of some progresses (reviewed in [9]), the properties and mechanisms of these biological networks are so far unknown.

1.1 PPI Networks

Among all the molecules in a living cell, proteins are essential parts of an organism and perform the most vast array of functions [55]. In the past, proteins were studied in isolation. Though remarkable knowledge on individual proteins has been gained [83], the functioning machinery of an organism cannot be comprehensively understood without investigation into the links between biological molecules, in particular, protein-protein interactions (PPI).

Protein-protein interactions are physical contacts between two or more proteins in a living cell or organism, often to carry out important biological processes. For example, G protein-coupled receptors interact with G proteins to transmit signals from stimuli outside a cell [84]. There are two main experimental approaches in wide use for detecting protein-protein interactions in large scale: Yeast two-hybrid (Y2H) [30] and tandem affinity purification coupled to mass spectrometry (TAP-MS) [81]. These high-throughput detection methods have led to the availability of large quantity of interaction data (Fig.1.2), which enable analysis of evolution and functionality of molecular and organisms. Large-scale experiments have been embarked on model-organisms, such as *S.cerevisiae* [45, 94], *C.elegans* [58, 99], *Helicobacter pylori* [78], *D.melanogaster* [36], and human [91]. These interaction data are collected and organized in databases, such as DIP [105], IntAct [49] and BioGRID [15], for easy reference.

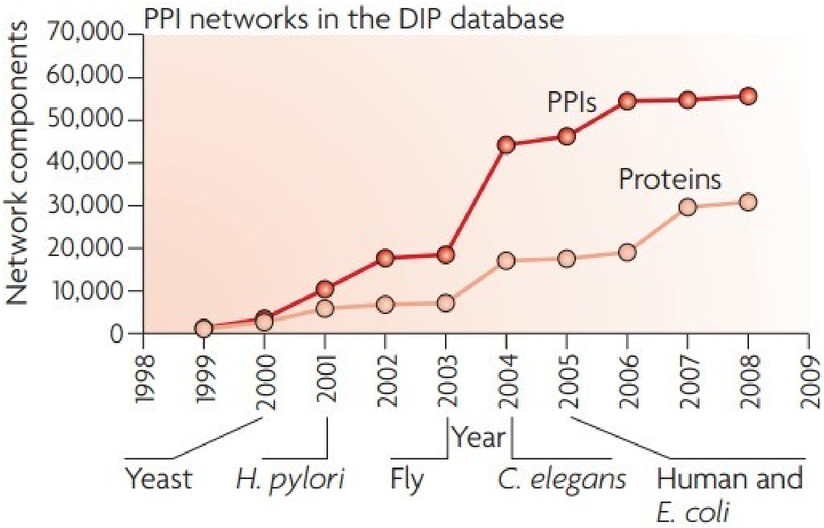

Figure 1.2: Accumulation of network components during the 10 years from 1999 to 2009. Image from [106].

1.1.1 Graph Representation and Properties

In mathematics, a network, which is also called a graph, consists of two components: Nodes and edges, where edges are an indicator function on the set of nodes. Specifically a set of nodes V and a set of indicator functions $E = \{e_{i,j}\}_{i,j \in V}$, define a graph $G(V, E)$, in which $e_{i,j} = 1$ if there is an edge between node i and j and $e_{i,j} = 0$ otherwise. If the pair of nodes (i, j) in the subscript of the indicator function $e_{i,j}$ are ordered (unordered), the graph G are called directed graph (undirected graph). Since we cannot say which protein binds with which one, protein-protein interactions are considered to be undirected. Hence in this thesis we focus on undirected networks, which means the order of the couple (i, j) does not matter and $e_{i,j} = e_{j,i}$.

Over the past decade, networks have been used to elucidate many complex systems in different disciplines, including computer science, biology, technology and social science. In biology, network provides a useful tool to represent and study interaction data of different types in cellular systems, such as protein-protein interaction, metabolic and gene regulation [9]. By investigating the interactions at a network level, new insights into the molecular mechanisms behind these systems can be discovered [97]. For example, a protein-protein interaction (PPI) network of the plant *Arabidopsis thaliana* containing about 6200 physical interactions between about 2700 proteins was constructed and reported in [4]. A study [65] based on it indicated how pathogens may exploit protein interactions to manipulate a plant's cellular machinery.

In PPI networks, nodes are proteins and edges are protein-protein interactions. Usually, a PPI network represents a collection of protein-protein interaction data in an organism. For example, by incorporating all the PPIs of the yeast obtained from a genome-scale study (such as [45]) we can generate a yeast PPI network. In order to understand the functioning and formation of a network, the first step should be to investigate its properties, which can be explored through the quantifiable tools of network theory. Network theory developed in other fields, such as Internet, physics, and sociology [18], can provide great help for the study of PPI networks. Several software tools have been introduced for network analysis. For example, the most commonly used software Cytoscape enables visualization and analysis of networks [87]. Even more powerful applications and extensions can be made via user-defined plug-ins. Another popular software tool GraphCrunch2 addresses network modeling, alignment and clustering [54].

If there is a link between node i and node j, we say i is a neighbor of j and vice versa. The number of neighbours of a node i is called its degree:

$$k_i = \sum_{j \in V} e_{i,j}.$$

It has been found that the degree of a protein has significant biological implications. The essential genes, whose malfunction would cause the death of an organism, are found to positively correlate with their degrees [47].

Probably the most basic quantity to investigate a network is the degree distribution $P(k)$, which can be defined as the proportion of nodes with degree k or, equivalently, the probability that a node, which is chosen uniformly at random, has degree k. Some interesting patterns of degree distribution have been realized in empirical networks. For example, scale-free is a widely observed characteristic in real networks, which means networks with a power-law degree distribution: $P(k) \sim k^{-\beta}$, where β is call the power-law exponent. In a scale-free network most nodes have a small number of interactions and a few nodes, the so-called hubs, interact with a large number of nodes. Owing to this property, scale-free networks are surprisingly robust against random external attack. Disabling a few number of nodes chosen at random would not cause fatal effect on a scale-free network. A scale-free network can tolerate up to 80% of its nodes to be disabled and still functions properly [77]. It is believed that scale-free property is shared by a wide range of real networks. Several non-biological networks, such as World Wide Web, social networks and citation networks, are scale-free, with power-law exponents greater than 2. The biological networks, such as yeast PPI network, E. coli metabolic network, yeast gene expression network and gene functional interactions, also follow a power-law, but with power-law exponents smaller than 2 (reviewed in [18]). A quantity relative to the degree distribution regards the average degree, which is defined to be the first moment of $P(k)$:

$$D = \sum_k kP(k) = 2e/n,$$

where $e = \sum_{i<j} e_{i,j}$ is the number of edges and $n = |V|$ is the number of nodes.

Other topological features commonly investigated include diameter, clustering coefficient and betweenness etc. Here we give a brief review on these three quantities. We first define the concept of path. Given two nodes, i and j, a path between i and j is a sequence of edges in which i and j as the two terminals and we can traverse from i to j by visiting each edge in the path exactly once. If there is no cycle in the path, we call it a simple path. The length of a path is the number of edges that the path contains. The shortest path between two nodes i and j is the path with the shortest length, which is called the distance between these two nodes, denoted by $l_{i,j}$. In a network, the maximum distances over all pairs of nodes is called diameter:

$$Diameter = \max_{i,j \in V} l_{i,j}.$$

A network with a small diameter is called a small-world network, in which a node can reach any other node by traversing a few number of connected nodes. This property allows efficient and prompt information transition in a network. Signal transduction and communication are tasks of many real networks. For instance, in PPI networks, signaling molecules from the exterior of an organism bind the receptor protein and signals are mediated through a sequence of protein-protein interactions to eventually activate the organism's reaction to the external signals [59]. The small-world effect has been found in many real networks, such as film actor corporation networks, power-grid networks and the yeast coexpression network [69, 101]. The emergence of small-world effect suggests that these real networks are likely to organize in such a way which facilitates signal and information transmission. Finally we introduce another important topological quantity: Clustering coefficient. Clustering coefficient, denoted by $c(u)$, of a given node u with degree k is defined as the proportion of pairs of this node's neighbors which are connected:

$$c(u) = \frac{\sum_{i,j \in N(u)} e_{i,j}}{\binom{k}{2}},$$

where $N(u)$ is the set of neighbors of node u. Equivalently, clustering coefficient is the probability that u and its two neighbors that are chosen uniformly at random from the set of the neighbors of u form a triangle. The average clustering coefficient is the mean of the clustering coefficient over all nodes: $\bar{c} = \frac{\sum_{u \in V} c(u)}{|V|}$. Clustering coefficient measures to what degree nodes tend to form a dense subgraph and it is often used an indicator for the modularity of a network [9]. High clustering coefficient has been observed in PPI networks, hinting at a high modularity. Given a node u, the betweenness of u, denoted by $b(u)$, is defined as the number of shortest paths from all vertices to all others that pass through u:

$$b(u) = \sum_{i,j} p_{ij}(u)/p_{ij},$$

where p_{ij} is the number of shortest paths between i and j, and $p_{ij}(u)$ is the number of shortest paths between i and j passing through u. Betweenness approximates the information flow that passes through a node and the essentiality of a node in the ability of a network to communicate [33].

Apart from the above quantities that describe the topology of a network, networks are often studied in terms of subgraphs, such as motifs and modules. Small subgraphs with statistical significance, which are termed motifs, have gained much attention in recent years. By applying methodologies for motif discovery, motifs of small sizes, such as triangles, are identified [48, 63, 104, 107]. Biomolecular network motifs are usually found to be associated with biological functions and considered to be basic building blocks for biological networks [63]. In [104], proteins in motifs are found to be conserved evolutionarily to a higher degree than those that are not members of motifs, indicating the biological importance of motifs in evolution. A module in a PPI network refers to a subgraph consisting a group of proteins and a group of interactions among them usually carry out important functions and may form a protein complex. Besides PPI networks, modules are also observed in networks of other fields such as World Wide Web and social networks [9]. Several techniques have been proposed to detect modules in PPI networks. For instance, Bader and Hogue [6] proposed the molecular complex detection algorithm (MCODE) which makes use of the so-called core clustering coefficient to predict molecular complexes. And Sharan et al. [88] developed a greedy likelihood algorithm called NetworkBlast to detect modules in protein interaction networks. Modules are evolutionary conserved parts in PPI networks.

1.2 Evolution of PPI Networks

Like other biological networks, PPI networks evolve with time. Only if we understand the evolutionary processes can we understand the network we observe today. However, due to the limited information and technology the evolutionary dynamics of PPI networks are still not well studied and the evolutionary mechanisms shaping the topology of PPI networks are not well understood. New techniques and methodologies are urged to explore the history of these networks.

1.2.1 The Central Dogma

Proteins are the “workhorses” that build up our body, but what monitor proteins are DNA, a polymer that contains genetic instruction. Francis Crick’s central dogma of molecular biology describes how the genetic information transfers between the three major information-carrying biopolymers: DNA, RNA and proteins[19]. The dogma emphasises the direction of the flow of information. In short, genetic information flow is formed by the following transfers: DNA→DNA transfer (DNA replication), DNA→RNA transfer (transcription) and RNA→proteins transfer (translation), known as the three general transfers (Fig.1.3). Other transfers are believed to be abnormal. In the process of transcription information contained in DNA is copied to a piece of messenger RNA (mRNA). Eventually mRNA is matched to transfer RNA (tRNA), thereby creating the corresponding amino acids, which are linked and folded to form proteins.

1.2.2 Nodes Addition and Deletion

Every protein is encoded by a stretch of DNA, namely a gene. By the central dogma, any mutation in the genome (the whole set of genes in an organism) may cause a change in its proteome (the whole set of proteins in an organism). It is observed that more than one third of genes in *E. coli* are orthologous to a human gene but few are conserved in more than 90% of sequenced bacteria [46]. This indicates that many genes are conserved across species and meanwhile the addition and deletion of genes play a fundamental role in the variety of protein functions. Gene loss, which is confirmed by the comparative analysis of sequences, is one of

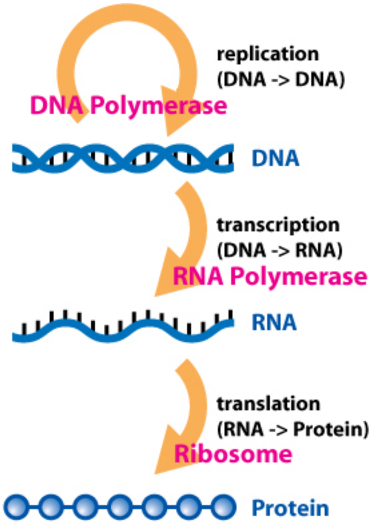

Figure 1.3: Illustration of the central dogma. Genetic information is transmitted from DNA to RNA and RNA makes the proteins via translation of the coded sequences. Image from "`http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology`".

the major evolutionary force [5, 64]. However, from the point of view of modeling a lost gene can be taken as a gene that never exists. Hence hereinafter we focus on the addition of nodes. The introduction of a new node into the genome can be either through horizontal gene transfer or gene duplication, which is the most frequent cases [106].

Gene duplication occurs in homologous recombination, which usually happens as unequal crossover [37](Fig.1.4), a retrotransposition event or duplication of an entire chromosome [109]. Gene duplication may happen in one single gene or a large-scale region in the genome and even the whole genome, in which case we call it the whole genome duplication (WGD). Gene duplication is widely observed in the genomes of various species. For example, it is believed that the yeast *S. cerevisiae* underwent a WGD about 150 million years ago [103]. The proportion of duplicate genes, which are usually detected by sequence alignment methods, is large and varies from more than 10% to over half [109]. Since the first reveal of gene duplications in 1930s and prevalence of this notion by Ohno's book in 1970, *Evolution by Gene Duplication* [72], gene duplication has been viewed as the main source of material for proteome evolution and play an an important role in developing novel functions. For instance, gene duplication is found to attribute to cold adaptation in Antarctic notothenioids [14, 16]. Immediately after a gene duplication event we can find two identical genes in the genome, which carry out exactly the same functions. The duplicate copy of a gene (or protein) is released from the pressure of natural selection at the time point of duplication and is likely to acquire a new, beneficial function that is preserved over time or lose the function its origin has. Specifically, the duplicate genes would be preserved via complementary or degenerate mutations. The functions carried out by the two identical duplicates would be partitioned by the pair, or one of them degenerates or acquires new functions [31] (Fig. 1.5). Genes that degenerate and do not function any more are called pseudogenes. Due to the functional redundancy, most duplicate genes become pseudogenes or lost. It is reported that there are more than 60% pseudogenes in human and 20% in mice [109]. However, the duplicate genes can be conserved if they differ in different functions. For example zebrafish *engrailed-1* and *engrailed-1b* are conserved duplicate genes that are expressed in different tissues of zebrafish [70].

1.2.3 Evolutionary Dynamics

Protein-protein interactions reflect the functions of proteins. The divergence of protein functions may cause loss or gain of interactions. Some hypotheses have been proposed for the evolution of PPI networks. For example, several authors emphasize the effect of domain shuffling on shaping the topology of PPI networks [13, 28, 34]. Among them, Evlampiev and Isambert proposed a model for PPI network evolution based on a refined version of whole genome duplication, in which

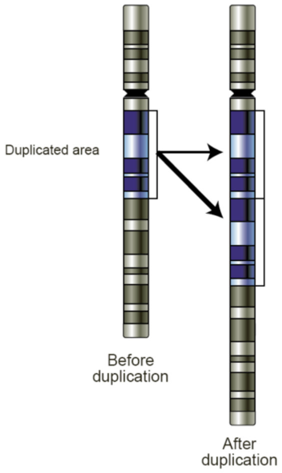

Figure 1.4: Illustration of gene duplication. Image from "http://en.wikipedia.org/wiki/Gene_duplication".

protein domains are introduced through different types of edges [28]. Preferential attachment of newcomers is also considered as a factor affecting the evolution of P-PI networks [20, 24]. For instance, based on the evolutionary conservation, Davids and Zhang [20] classified the *E. coli* genes into three categories: Core genes, Non-core genes and genes resulting from horizontal gene transfer (HGT). They claimed that the HGT genes link with Core genes in a preferential attachment manner. Some other authors focus on gene duplications (see [96, 98] for examples). By studying the relation between the fraction of duplicates with at least one common interacting neighbor and the fraction of synonymous substitutions per synonymous site [37], Wagner found that the higher the similarity between duplicates is the more interactions the duplicates share [98]. Based on this observations, the author proposed a model for the effect of gene duplications on the protein-protein interactions. In this model, the process of evolution by gene duplication and divergence is depicted as the rewiring of their adjacent links, including loss of adjacent edges and gain of new adjacent neighbors. This mechanism links the molecular evolution with the network evolution especially in the aspect of gene duplication.

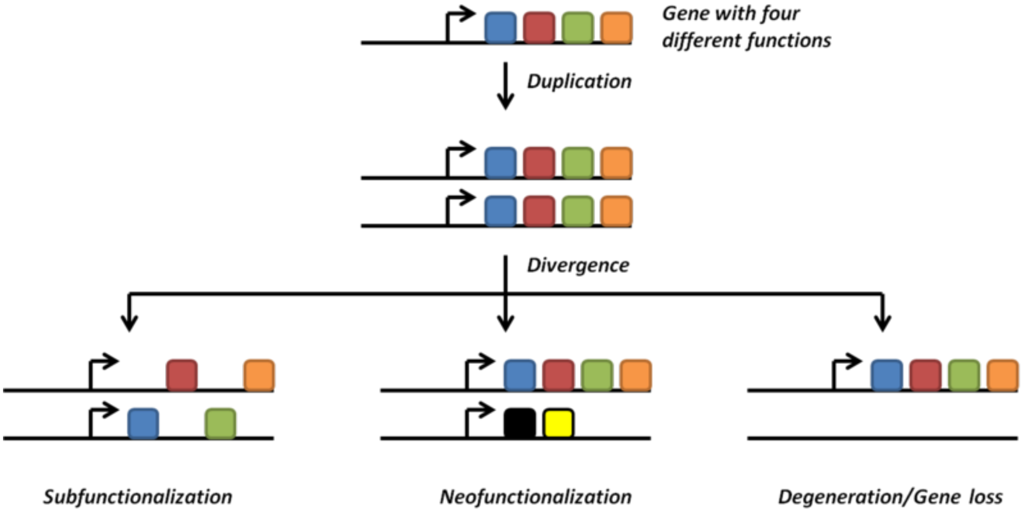

Figure 1.5: Evolutionary fate of duplicate genes. A gene with four functions is duplicated. In the divergence of the duplicate genes, four cases may happen: Subfunctionalization, neofunctionalization and degeneration. In subfunctionalization, functions are partitioned by the two duplicate genes. In this case, each carries out two of the four original functions. In neofunctionalization, a duplicate gene obtains new functions. Here one gene acquires two new functions. In degeneration, one of the duplicate genes loses its functions and become pseudogenes or unidentifiable. Image from "`http://en.wikipedia.org/wiki/Gene_duplication`".

1.3 Modelling PPI Networks

PPI networks that we observe today are results of millions of years of evolution. Not only the proteins themselves undergo mutations and natural selection, but also the interactions between them change with time. Even if the proteins remain unchanged, the interactions may still vary (examples can be found in the conserved modules in different species). Understanding how PPI networks evolve and how the properties of PPI networks emerge would shed light on the functioning machinery of a cell or organism and provide insight into human diseases at the molecular level [97]. Like in other disciplines, such as physics, a proper model in biology can provide a theoretical framework in the analysis of the dauntingly huge real data. With the help of computers, processes that cannot be realized in reality (such as the reconstruction of PPI network evolutionary history, see Chapter 2 for details) can be completed by embedding the models. A question should be asked beforehand: What is a "proper" model? To the best of our knowledge, there is no definite answer to it. However, the model should be simple enough to be mathematically tractable, and consistent with biological facts and fits the real data to some extent. Even if a model is not mathematically tractable and analytical results are difficult to be obtained, simulation studies can also provide valuable insights into the real networks of interest. Here we give a brief review on some interesting graph models which may be useful in our research.

1.3.1 Random Graph Models

Probably the best known random graph is the **Erdős-Rényi (ER)** model [26], which is named after Paul Erdős and Alfréd Rényi, who proposed the model in 1959. An ER model with n nodes and parameter p, denoted by $\mathcal{M}(n,p)$, generates networks by independently connecting each pair in the n nodes with probability p (Fig. 1.6). Note that there are $\binom{n}{2}$ edges in a complete graph with n nodes and under the ER model a network with n nodes and m edges, denoted by $G(n,m)$, is generated with probability $p^m(1-p)^{\binom{n}{2}-m}$. The degree distribution of ER model is binomial [67]:

$$P(\deg(v)=k)=\binom{n-1}{k}p^k(1-p)^{n-1-k},$$

which converges to a Poisson distribution when n is large and np is fixed. Further mathematical properties of ER model is described in [27]. There is another variant of the ER model $\mathcal{M}(n,m)$, where n is the number of nodes and m is the number of edges. In $\mathcal{M}(n,m)$, m edges are chosen uniformly at random from the $\binom{n}{2}$ potential edges. When $pn^2 \to \infty$, many graph properties in $\mathcal{M}(n,p)$ and $\mathcal{M}(n,m)$, with

$m = np$, are equivalent [27].

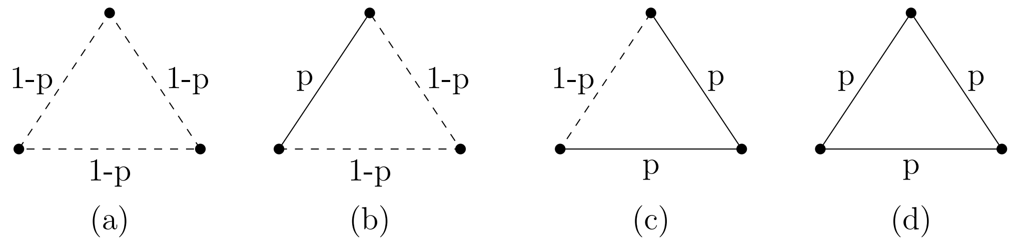

Figure 1.6: Four non-isomorphic samples of an ER model with $n = 3$. Given three nodes, every pair of nodes are linked independently with probability p. (a) None of the edges is present. Note that the probability for an absent edge is $1 - p$. Hence $P(G(3, 0)) = (1 - p)^3$. (b) In this sample, one edge is present and two are absent. So $P(G(3, 1)) = p(1 - p)^2$. (c) Two edges are present and one is absent. The probability is $P(G(3, 2)) = p^2(1 - p)$. (d) All edges are present: $P(G(3, 3)) = p^3$.

In order to obtain graphs similar to PPI networks, one has to compare the graphs generated by a model with PPI networks. Instead of identifying isomorphic graphs, whose computational complexity is still unknown, we compare properties of two networks such as degree distribution, which are feasible and efficient. We know that the yeast PPI network has a high average clustering coefficient and power-law degree distribution which has a fat tail, but the ER model has a bell-shaped binomial degree distribution and low clustering coefficients. Hence in terms of these two quantities ER model is not an ideal model for PPI networks.

The **Watts-Strogatz** model is another popular random graph model which generates networks with small-world property and high clustering coefficients, two important characteristics observed in various empirical networks [101]. The model starts with a regular ring lattice with n vertices and K degree per vertex, which can be defined by connecting each node on a ring to its K nearest neighbors ($K/2$ on each side, Fig. 1.7(a)). Each edge $e_{i,j}$ on the lattice, where $i < j$, is replaced by another edge $e_{i,k}$ with some probability p[1], where k is chosen uniformly

[1] With a slight abuse of notations, we use the same p as in the ER model when the context is clear. Similar cases occur occasionally in the following part of this thesis.

at random from the set of vertices, which are not the neighbors of node i, and $k \neq i$ (Fig. 1.7(b)). The model was designed by interpolating between regular and random networks tuning by parameter p. When p is 0, the model is definite; when p is 1, the model is complete disorder. Watts and his coauthor found that adjusting p from 0 to 1 the average length of the shortest paths decrease and meanwhile clustering coefficient decreases. Although the Watts-Strogatz model can generate high clustering coefficient and small average length of shortest paths, it fails in generating a scale-free network [10].

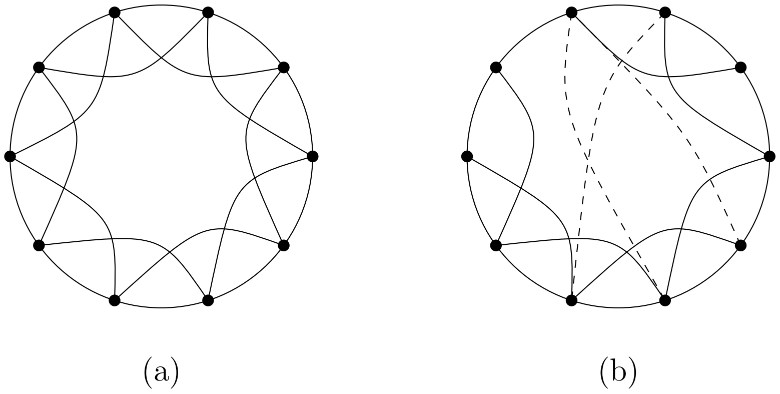

Figure 1.7: Illustration of the Watts-Strogatz model. A regular lattice is obtained by connecting each vertex on a ring with n vertices ($n = 10$ in this example) to its K ($K = 4$) nearest vertices. For each edge, with probability p one end is reconnected to another vertex, which is chosen uniformly at random from the set of nodes. Self-links and duplicate edges are forbidden. Three edges are rewired in this example.

1.3.2 Growing Graph Models

The ER model and the Watts-Strogatz model have successfully explained the emergence of some interesting properties of some real networks [67]. However, their limitations are: (1) As mentioned above, they fail to produce scale-free networks; (2) they generate networks on a fixed set of nodes. However, many real networks, especially biological networks are under processes of growth.

The **Barabási-Albert (BA)** model, which is also called the **preferential attachment (PA)** model, is a network growth model based on the preferential attachment mechanism [8]. A **network growth model** $\mathcal{M}(G_0, \Phi)$ can be recursively defined as follows: For each positive integer n, the network $G_n(V_n, E_n)$ generated by the model $\mathcal{M}(G_0, \Phi)$ is obtained from $G_{n-1}(V_{n-1}, E_{n-1})$ by adding a vertex, say v, into G_{n-1}: $V_n = V_{n-1} \cup \{v\}$ and deleting or adding edges according to some rule Φ: $E_n = \Phi(E_{n-1})$. In this thesis, we usually replace Φ by the parameters required by a model when the context is clear. For example, the PA model with initial graph G_0 and parameter m (see below) can be denoted by $\mathcal{M}(G_0, m)$. The PA model is the first graph model that incorporates the concept of growth. Following the PA model, many network growth models have been proposed. In the PA model, the description of Φ is preferential attachment. Specifically, at each time t, the new node v is connected to m nodes in the existing network with probability $\deg(u)/(2e)$, where $\deg(u)$ is the degree of node u and $e = |E_{t-1}|$, the number of edges in G_{t-1}. Note that the new node would have more chances to link with the nodes with high degrees. This phenomenon is usually termed as "the rich get richer". In the world wide web, it can be conceived as an analog of the phenomenon that new pages link preferentially to popular web pages. If we take it as a model of social networks, then a newcomer in a community is likely to befriend with popular people rather than the unpopular ones.

An important consequence of the PA model is that it generates networks with power-law degree distribution that is observed in many non-biological networks. However, how to explain the preferential attachment in PPI networks is not clear. Moreover, the power-law exponents generated by the PA model is different from those in PPI networks, which are smaller than 2 but the former ones are greater than 2 [18]. This may indicate that although both biological networks and some

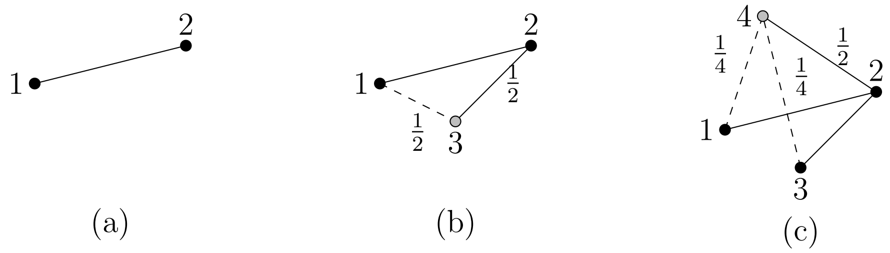

Figure 1.8: An example for the PA model with $m = 1$ and $G_0 = K_2$, i.e. the complete graph with 2 nodes. (a) The seed graph is given as K_2. (b) At time 3, a new node, namely node 3, is added into the graph and connected to node 2 with probability 1/2 since the number of edges $e = 1$ and $\deg(2) = 1$. Likewise the probability for node 3 to be linked with node 1 is 1/2 but the edge is not present in this sample. (c) At the next step, another new node, node 4, is added again and connected to node 2 with probability 1/2 since $e = 2$ and $\deg(2) = 2$ in the existing graph.

non-biological networks exhibit scale-free property, they undergo different growing mechanisms.

As reviewed in Subsections 1.2.2 and 1.2.3, gene duplications have a significant impact on the evolution of PPI networks. Duplication models are a more biologically relevant class of network models that incorporates gene duplications. At every time step a node in the existing network is chosen uniformly at random as the anchor node and duplicated. The anchor node and the duplicate node have the same neighbors after the duplication. And then edges adjacent to them are rewired [18, 95]. In some models, new edges linking the duplicate node and other existing nodes are allowed to be added [11, 17]. The duplication step is considered to be a major underlying mechanism in shaping the topology of P-PI networks[98] and duplication models are often used to investigate biological networks[52, 85, 102]. It has been found that some of the duplication models have a power-law degree distribution and fits biological networks well [18, 43].

The **full duplication (FD)** model is the simplest duplication model, in which only node duplications occur but no modification is made to the duplicate node after the duplication. Specifically, starting with a seed graph G_{t_0}, at each time point $t > t_0$, an anchor node, say u_t, is chosen uniformly at random from V_{t-1}, the set of nodes in G_{t-1}, and duplicated: The new node, usually denoted by v_t, is added into the network and copies all the edges adjacent to u_t (Fig. 1.9). Hence $V_t = V_{t-1} \cup \{v_t\}$ and $E_t = E_{t-1} \cup \{e_{v_t,v_i} | e_{v_t,v_i} = e_{u_t,v_i}, i = 1, \cdots, t-1\}$. We call this mechanism as the *duplication step.* If two nodes are duplicate nodes, we say they are in the same family. Note that we can classify all the nodes into $|V_{t_0}|$ different families. For example, in Fig. 1.9(c), there are 3 families: Node 1 itself is one, nodes 2 and 5 are in the same family and nodes 3 and 4 are in another. By such classification, we can model the FD model by a Polya urn, in which each family is represented by a color and the nodes in a family is the balls with the corresponding color. If there are nodes in two different families linking with each other, we call the two families are adjacent. Note that the adjacency relation is unchanged all the time. All the nodes in a family have the same neighbors which are all the nodes in the families adjacent to this family. We know that the number of nodes in each color would grow to infinity and thus the degree of each node will be infinitely large too. This unrealistic degree distribution generated by the FD model makes it difficult to be applied to real networks.

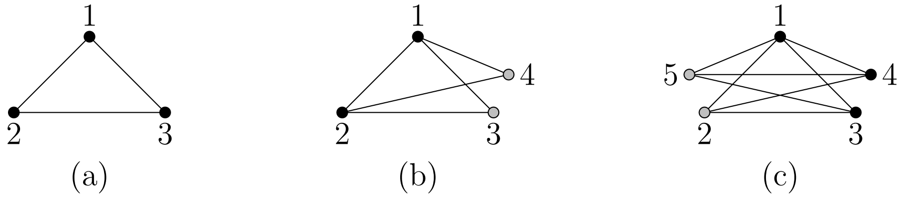

Figure 1.9: An example for the FD model with $t_0 = 3$ and $G_{t_0} = K_3$, i.e. the complete graph with 3 nodes. (a) The seed graph is given as K_3. (b) At time 4, node 3 is chosen as the anchor node (with probability 1/3). The new node 4 is added into the network and connected to all the neighbors of node 3. (c) At time 5, node 2 is chosen as the anchor node (with probability 1/4). The new node 5 is added into the network and copies all the edges adjacent to the anchor node.

The full duplication model captures the major driving force of PPI network evolution, i.e. gene duplication. However, the absence of gene divergence after duplication renders this model too ideal to mimic the real networks. The duplication and divergence evolutionary mechanism of gene duplication on PPI networks proposed by Wagner should be considered (Subsection 1.2.3). Despite its simplicity, the **partial duplication (PD)** model is not the first duplication model that incorporates the gene divergence. To the best of our knowledge, the first duplication model that makes use of Wagner's model is due to Vazquez et al. [95]. For the sake of easy understanding, the PD model will be introduced before other more complicated duplication models.

The partial duplication model is first depicted in [18] by Chung et al. to study its mathematical properties. The authors claimed that the networks generated by the PD model have a power-law degree distribution and derived a formula for the power-law coefficient. However later they stated that it is a wrong proof and modified the model by linking each duplicate node and its anchor node at each time, which results in a scale-free network [17]. Nonetheless their work has inspired other efforts in the mathematical properties of duplication models (see Chapter 3 for details). In the PD model $\mathcal{M}(G_{t_0}, p)$, where G_{t_0} is the seed graph and $0 < p \leq 1$ is the *selection probability* of the model, we start with G_{t_0} and at each time step t, the graph G_t is obtained from G_{t-1} by the following procedures: An anchor node u_t is chosen uniformly from the set of nodes in G_{t-1}, and a new node v_t is added and independently connected to each neighbor of u_t with probability p (see Fig. 1.10 for an illustration). From the point of view of duplication, the anchor node u_t is first duplicated as node v_t and each edge adjacent to v_t is independently lost with probability $1 - p$. The selection probability p is the probability that a duplicate node preserves one interaction (function). Defining $p > 0$ is to make sure that the trivial case, i.e. only singletons are generated, will not occur.

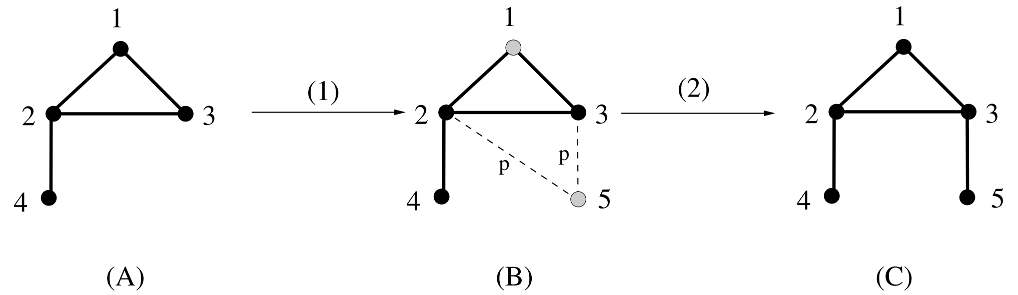

Figure 1.10: Illustration of one step of the PD model. (C) is obtained from (A) by one duplication step, in which node 1 is the anchor node and node 5 is the new node. The probability that node 1 is chosen as the anchor node is 1/4 because the network in (A) contains four nodes. Given that 1 is the anchor node and 5 is the new node, the probability that (C) is obtained is $p(1-p)$.

The **duplication-mutation with complementarity (DMC)** model proposed by Vazquez et al. in [95] is another popular duplication model [34], which is also the best model to fit the *D. melanogaster* PPI network according to a recent study by Middendorf et al. [62].

In the DMC model $\mathcal{M} := \mathcal{M}(p, p_c)$, where p and p_c are the parameters of the model, we start with an initial graph G_0, the so-called seed graph. At each time step t, the graph G_t is obtained from G_{t-1} by the following processes (see Fig. 1.11 for an illustration).

(1) (*Duplication*): A node u_t, the anchor node, is chosen uniformly at random from the set of nodes in G_{t-1}, and a new node v_t, the duplicate node, is added and connected to every neighbor of u_t.

(2) (*Mutation*): For each neighbor of u_t, say w, we choose one edge from (u_t, w) and (v_t, w) with equal probability, and this chosen edge is deleted with probability $1-p$.

(3) (*Homodimerization*): The nodes u_t and v_t are connected with probability p_c.

Step 1 reflects the idea that duplicate nodes have identical functions immediately after duplication and thus share the same interaction neighbors as anchor nodes [98]. As time goes on, mutation causes the disappearance of the interactions of the duplication pair, which is encoded in Step 2.

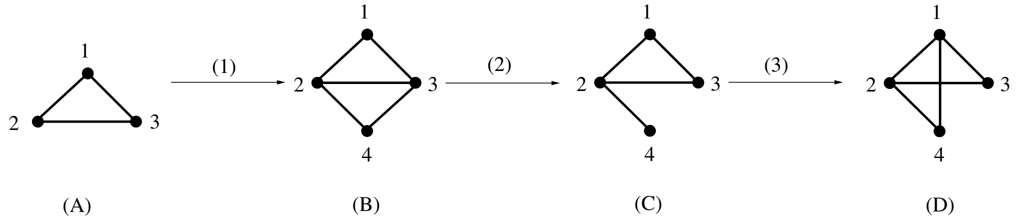

Figure 1.11: Illustration of the DMC model. (B) is obtained from (A) by one duplication step, with node 1 as the anchor node and node 4 as the duplicate node; the probability that node 1 is chosen as the anchor node is 1/3 because the network in (A) contains three nodes. (C) is obtained from (B) by the mutation step, which occurs with probability $p(1-p)/2$. (D) is obtained from (C) by the homodimerization step, which occurs with probability p_c.

The **duplication and divergence (DD)** model [73] is another duplication model we have also investigated in this thesis. As in the PD model, an anchor node u_t is chosen uniformly at random in G_t and the new node v_t copies each edge of u_t with probability p. After that, in the DD model the new node independently links with each existing node (except the neighbors of the anchor node) with probability $r/(t - \deg(u_t))$, where r is a parameter and $\deg(u_t)$ is the degree of anchor node u_t. We call this as the divergence. Note that r is the expected number of edges that v_t can get in the divergence step.

There are some other network growth models, such as the crystal growth (CG) model, the hierarchical networks [51, 80]. The modularity of biological networks is obtained by the crystal growth (CG) model, which mimics the incorporation of proteins into crystals in solution. It is shown that CG model fits the yeast PPI network well in terms of degree distribution, distribution of clustering coefficient and the age dependency of interaction density, which measures the connection

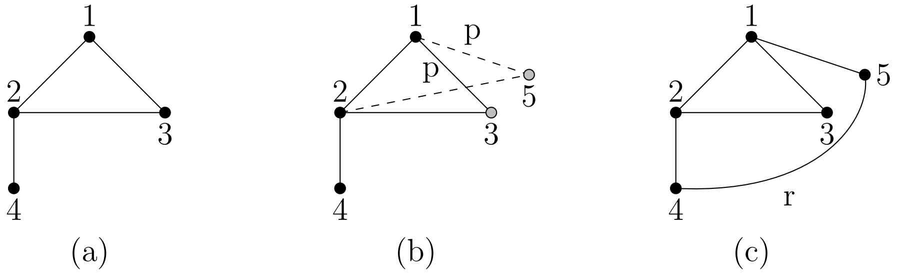

Figure 1.12: Illustration of a time step in the DD model. (a)At time $t = 4$, G_4 is given. (b) At time $t = 5$, node 3 is chosen as the anchor node (with probability 1/4) and the duplicate node 5 can copy each edge adjacent to node 3 with probability p. (c)Here the new node preserves one common neighbor of the anchor node, namely node 1, and links with node 4 which is not a neighbor of the anchor node with probability r since $t - \deg(3) = 3 - 2 = 1$.

between different age group of proteins[51]. The hierarchical networks are designed to capture the hierarchical modularity observed in biological networks. For a given k, we define $c(k)$ to be the average clustering coefficient of nodes with degree k. In the hierarchical networks, $c(k)$ is also power-law: $c(k) \sim k^{-\gamma}$ [9].

1.4 Objectives and Organization of Thesis

This thesis studies three mathematical issues about modelling PPI networks, which are presented in Chapters 2 to 4. Each chapter ends with a summary on the work and the possible extensions to the work presented in the chapter. Finally, Chapter 5 gives an overall summary on this thesis. The contents of each chapter are organized as follows.

Chapter 2 presents a novel gene-tree-based method for reconstructing the growth history of PPI network evolution. This method predicts the growing history of PPI networks by making use of the information of the duplication history of proteins and PPI network topology. Experiments are done to compare two proposed algorithms, namely MLN and CG, and a previously proposed algorithm by

Navlakha and Kingsford [66]. Applications to real PPI networks are also described.

Chapter 3 discusses the limiting behavior of the partial duplication model, a random network growth model in the duplication and divergence family. We show that for each non-negative integer k, the expected proportion of nodes of degree k approaches a limit as the network becomes large. This fills in a gap in previous studies. In addition, we prove that there is a phase transition point p_0 for the expected proportion of isolated nodes converging to 1, and hence provide hints to a question raised in [11]. We also obtain asymptotic bounds on the convergence rates of degree distribution. Since the observed networks typically do not contain isolated nodes, we study the subgraph consisting of all non-isolated nodes contained in the networks generated by the partial duplication model, and show that p_0 is again a phase transition point for the limiting behavior of its degree distribution.

Chapter 4 explores the effect of seed graphs on the growth of networks generated by duplication models. This chapter is presented as an open direction of future work. Simulations were run to investigate the topological features of the PD model, the DD model, the DMC model and the PA model: The clustering coefficient, the average degree, the average length of shortest paths and the degree distribution. Results show that the seed graphs have an impact on the network evolution but the impact is limited. For example, the clustering coefficient decreases with time for any chosen seed graph. The limiting degree distribution is determined by the parameters of the models and is not affected by the seed graphs.

Chapter 2

Reconstruction of Network Evolutionary History

2.1 Introduction

Over the past decade, it has become increasingly clear that in order to decipher the complex relationship between genotype and phenotype, we need to investigate protein-protein interaction (PPI), metabolic and gene regulation networks in addition to studying individual genes and their proteins [9, 71]. Since PPI networks are available for several model organisms, a natural but important next step will be to elucidate the evolutionary aspect of PPI networks [41, 66].

Evolutionary history of PPI and gene regulatory networks provides valuable insight into molecular mechanisms underlying network growth [97, 98]. It helps to understand some of the topological principles of these networks [89, 106], and even shed light on the unicellular-multicellular and invertebrate to vertebrate transitions [68].

Analogous to reconstructing evolutionary history at the level of the DNA or amino acid sequence, the starting point for our approach is to choose an evolution model. Unlike many networks studied in technology and sociology, the growth of PPI networks is mediated by gene duplication and divergence [98]. We have introduced the several duplication models in Chapter 1. A recent study by Middendorf et al. [62] showed that the duplication-mutation with complementarity (DMC) model, to be described in details in Section 2.2, fit the *D. melanogaster* (fruitfly) PPI network better than several other commonly used growth models. In this chapter, we shall focus on this DMC model.

In general, reconstructing the evolutionary history of an observed network under a given growth model includes inferring the relative order of the nodes according to which the network has evolved, and predicting edge arrival and loss events [75]. However, for the DMC model studied here it is sufficient to consider only the relative order, which will in turn determine the edge arrival and loss events (see Section 2.2 for details).

Several approaches have been proposed to address the problem of reconstructing network histories. Gibson and Goldberg introduced a merging algorithm to reconstruct the evolutionary history of PPI networks using gene trees reconciled against a species tree [35]. A novel likelihood-based framework for inferring histories was presented by Navlakha and Kingsford in [66]. Recently, Patro et al. [74, 75] proposed a maximum parsimony approach, in which the evolutionary history of network is coded by a graph.

Here we introduce a new history inferring framework based on the maximum likelihood principle. In contrast to the method in [66], our approach incorporates not only the topology of observed networks, but also the duplication history of the proteins in the networks. Indeed, duplication histories, which can be obtained from reconciled gene trees, have proven to be useful in understanding PPI network evolution. For example, Dutkowski and Tiuryn applied a Bayesian network framework to infer the posterior probability of interactions between ancestral nodes based on reconciled gene trees [23] for better prediction of protein modules. A similar approach was also used by Pinney et al. [76] to infer ancestral interactions between bZIP proteins. In these studies, the edge lengths are often assumed known and hence the internal nodes in the trees can be totally ordered. However, our approach only requires the topological information of the gene trees.

The rest of the chapter is organized as follows: In the following section, we review some basic definitions and background concerning network reconstruction. Section 2.3 presents some theoretical results that are key to our approach as they enable us to reduce the problem of finding a most probable history of a given network to a simpler optimization problem. Two efficient heuristic algorithms to solve the latter problem are proposed in Section 2.4. Based on simulation studies, we show in Section 2.5 that our method provides better inference than the one proposed by Navlakha and Kingsford [66]. We also apply our approach to the PPI networks of *S. cerevisiae* (budding yeast), *D. melanogaster* (fruitfly) and *C. elegans* (worm) to obtain a set of growth parameters, and study the change of the networks' clustering coefficient and the relationship between the number of duplications and the degree of nodes. We conclude in Section 2.6 with some future research directions.

2.2 Basic Definitions and Notations

In this section, we shall introduce some basic definitions and notations related to reconstructing network evolutionary history.

2.2.1 Modeling Protein-protein Interaction Networks

The vertex set and edge set of a network G will be respectively denoted by $V(G)$ and $E(G)$, and $|V(G)|$ is called the *size* (or *order*) of G. Given a vertex v in G, its neighborhood $N_G(v)$, or simply $N(v)$ when the context is clear, contains exactly those vertices that are adjacent to v in G. Note that by our definition v is not contained in $N(v)$.

Recall that the DMC model is based on three mechanisms: Duplication, mutation and homodimerization, and two parameters: the selection probability p and the homodimerization rate p_c. The DMC model is Markovian, that is, $\mathbb{P}(G_t \,|\, G_s, s < t, \mathcal{M}) = \mathbb{P}(G_t \,|\, G_{t-1}, \mathcal{M})$, which depends on p and p_c, the parameters of $\mathcal{M}$. For example, denoting the network (A) and (D) in Fig. 1.11 by G_{t-1} and G_t, respectively, then the probability $\mathbb{P}(G_t|G_{t-1}, \mathcal{M})$ that G_t evolves from G_{t-1} in one step under the model $\mathcal{M}$ is $p(1-p)p_c/6$.

2.2.2 Network History and its Reconstruction

Given an observed network G, a *growth history* $\mathcal{H}$ of G is a graph sequence $(G_0, G_1, \cdots, G_n)$ such that $G_n = G$ and for $1 \leq t \leq n$, graph G_t can be obtained from G_{t-1} in one step under the DMC model $\mathcal{M}$. The first graph G_0 and the number n are called respectively the seed graph and the *span* of the history. Clearly, a history $\mathcal{H}$ induces a unique sequence $\theta := \theta(\mathcal{H})$ of duplicate nodes. More precisely, we have $\theta(\mathcal{H}) = (v_1, \cdots, v_n)$ in which for each t, node v_t is the duplicate node at time t, that is, the unique node in $V(G_t) \setminus V(G_{t-1})$. For example, a growth history $\mathcal{H}$ with span 3 is depicted in Fig. 2.1(A), in which the seed graph consists of two connected nodes, and we have $\theta(\mathcal{H}) = (2, 4, 5)$.

Given a network G, let $\mathcal{H}$ be the growth history we hope to infer. The probability of G being evolved according to history $\mathcal{H}$, when viewed as a function of the unknown history $\mathcal{H}$, is called the *likelihood function* $L(\mathcal{H} \,|\, G, \mathcal{M})$. Since the DMC model is Markovian, the likelihood function can be simplified as

$$L(\mathcal{H} \,|\, G, \mathcal{M}) = \prod_{t=1}^{n} \mathbb{P}(G_t | G_{t-1}, \mathcal{M}).$$

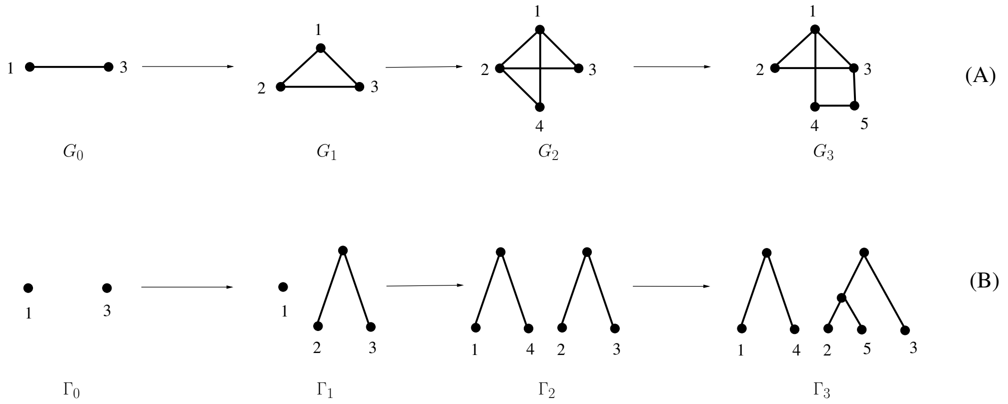

Figure 2.1: An example of growth history (A) and duplication history (B). Here the seed graph is an edge; the duplicate sequence is $(2, 4, 5)$ and the anchor list is $\{3, 1, 2\}$.

Following [66], we adopt a maximum likelihood criterion to infer the history of G as below.

Problem 1. Given a network G together with a natural number n and model $\mathcal{M}$, construct a growth history $\mathcal{H}$ that maximizes the likelihood $L(\mathcal{H} \,|\, G, \mathcal{M})$ among all histories with span n.

Typical (in the sense of highest probability, as commonly understood) histories correspond to histories with maximum probability. Maximum likelihood principle corresponds to choosing the parameters which best explain the observed data. We shall adopt this approach in inferring the network history. This problem is difficult since the number of possible histories grows exponentially. It is not known whether Problem 1 is polynomial-time solvable. In [66], a greedy algorithm called NetArch is introduced, in which a history is recursively constructed from G_n to

G_{n-1} by choosing a pair of anchor and duplicate nodes that maximizes the likelihood function. Since protein duplication relationship can be obtained from the gene duplication history for gene families, we propose an alternative approach which integrates the duplication forest (to be introduced below) to address this history reconstruction problem.

2.2.3 Duplication History

A tree T is a connected graph that contains no cycle, and all trees considered here are rooted. Node u is a child of v if they are adjacent, and the path from the root to u contains v. A tree is called binary if each internal node has exactly two children. A binary forest consists of a collection of binary trees; it is *trivial* if each tree in this forest has exactly one node.

For later use, we describe a scheme that encodes the duplication history in a growth history by a binary forest, called *duplication forest*. We start with a trivial forest Γ_0 with isolated nodes corresponding to the nodes in the seed graph. At each step t, the forest Γ_t is obtained from Γ_{t-1} by replacing the anchor node u_t with a cherry $\{u_t, v_t\}$, where v_t is the duplicate node at step t. Here a *cherry* $\{u, v\}$ means a subtree consisting of two leaves u and v and the internal node adjacent to them. For example, the forest Γ_3 in Fig. 2.1(B) is the duplication forest of the growth history depicted in Fig. 2.1(A). Note that this duplication forest corresponds to the anchor list $\{3, 1, 2\}$, that is, the first three anchor nodes used are $3, 1, 2$. A different choice of anchor nodes may lead to a different duplication forest. In other words, a duplication forest is uniquely determined by the growth history and a list of anchor nodes.

The idea of encoding duplication history by a binary forest can be traced back at least to the work by Chung and Lu [17]. One key observation used in our study is that the duplication forest of a PPI network can be inferred independently without using the network growth history. For instance, such a forest can be reconstructed by using the phylogenetic relationships among the genes that specify the proteins in the network [76]. Indeed, in a different paradigm a maximum parsimony approach to reconstruct the network history from a duplication forest is proposed in a recent study by Patro et al. [74].

2.2.4 Backward Operator

Consider one particular step in a growth history, that is, graph G_t obtained from G_{t-1} by using anchor node u_t and duplicate node v_t. We want to define a backward operator $\mathcal{R}$ so that G_{t-1} can be reconstructed by knowing the triplet (G_t, u_t, v_t). To this end, let $\mathcal{R}_{v_t}^{u_t}(G_t)$ be the graph obtained from G_t by merging the two nodes u_t and v_t in G_t. More precisely,

(i) for each node w in $N(v_t) \setminus \big(N(u_t) \cup \{u_t\}\big)$, add an edge (w, u_t);

(ii) delete the node v_t and all edges incident to it.

For instance, for the graphs in Fig. 2.1(A), we have $G_2 = \mathcal{R}_5^2(G_3)$ and $G_1 = \mathcal{R}_4^1(G_2)$.

Similarly, the backward operator can be applied to the duplication forest, that is, $\mathcal{R}_{v_t}^{u_t}(\Gamma_t)$ is the forest obtained from Γ_t by replacing the cherry $\{u_t, v_t\}$ with the leaf u_t. Note that this definition is consistent with the backward operator defined on graphs in the following sense: If Γ_t is the duplication forest corresponding to the network G_t, then $\mathcal{R}_{v_t}^{u_t}(\Gamma_t)$ is the duplication forest associated with $\mathcal{R}_{v_t}^{u_t}(G_t)$. For example, in Fig. 2.1, we have $G_2 = \mathcal{R}_5^2(G_3)$ and $\Gamma_2 = \mathcal{R}_5^2(\Gamma_3)$, in which Γ_i is the duplication forest associated with G_i for $i = 2, 3$. When the anchor node u_t is clear from the context, we simply write $\mathcal{R}_{v_t}$ for $\mathcal{R}_{v_t}^{u_t}$.

2.3 Reconstruction with Known Duplication History

In this section, we shall study the problem of reconstructing network growth history when the duplication forest is known, a simplification of Problem 1. We first show that a growth history with known duplication forest is determined by its duplicate sequence. We adopt the convention that a node sequence consists of distinct nodes, whereas a node list may contain repeated nodes.

A node sequence $\theta = (v_1, \cdots, v_n)$ and a duplication forest Γ are said to be *compatible* if there exists a (necessarily unique) sequence $(\Gamma^\theta_0, \cdots, \Gamma^\theta_n)$ of forests such that $\Gamma^\theta_n = \Gamma$, Γ_0 is trivial, and $\Gamma^\theta_{t-1} = \mathcal{R}_{v_t}(\Gamma_t)$ holds for each $t \in \{1, \cdots, n\}$. Note that a necessary condition for θ and Γ being compatible is that v_t belongs to a cherry in Γ^θ_t for each t. Denote the sibling of v_t in Γ^θ_t by u_t for $1 \leq t \leq n$. The list $\pi = \{u_1, \cdots, u_n\}$ is called the *anchor* list determined by Γ and θ.

As mentioned above, a growth history $\mathcal{H}$ specifies a duplicate sequence θ. Together with a list of anchor nodes, such growth history also determines a duplication forest Γ. In this case, the sequence θ and the forest Γ must be compatible. On the other hand, given a duplication forest Γ associated with a network G and a sequence θ that is compatible with Γ, then there exists a unique growth history $\mathcal{H}$ such that θ is induced from $\mathcal{H}$. In other words, when the duplication forest Γ is fixed, a growth history $\mathcal{H} = (G^\theta_0, \cdots, G^\theta_n)$ is uniquely determined by a duplicate sequence $\theta = (v_1, \cdots, v_n)$ compatible with Γ. That is, we have $G^\theta_n = G$, and $G^\theta_{t-1} = \mathcal{R}^{u_t}_{v_t}(G^\theta_t)$ for $1 \leq t \leq n$, in which u_t is the unique leaf in Γ^θ_t that forms a cherry with v_t. In this context, the likelihood function is defined as

$$L(\theta \,|\, G, \Gamma, \mathcal{M}) := \prod_{i=1}^{n} \mathbb{P}(G^\theta_i \,|\, G^\theta_{i-1}, \Gamma, \mathcal{M}),$$

where $\mathbb{P}(G_i^\theta|G_{i-1}^\theta,\Gamma,\mathcal{M})$ is the probability that G_i^θ evolves from G_{i-1}^θ in one step under the DMC model $\mathcal{M}$ and using the anchor node u_t specified by θ and Γ. Note that in general the probability $\mathbb{P}(G_i^\theta|G_{i-1}^\theta,\Gamma,\mathcal{M})$ is different from $\mathbb{P}(G_i^\theta|G_{i-1}^\theta,\mathcal{M})$. Indeed, the latter can be regarded as the "average" of the former over all possible anchor nodes.

The problem of inferring growth history with given duplication forest, a variant of Problem 1, can be formally stated as below.

Problem 2. Given a network G together with a duplication forest Γ and a growth model $\mathcal{M}$, construct a duplicate sequence θ such that the likelihood $L(\theta\,|\,G,\Gamma,\mathcal{M})$ is maximized.

In the above problem, the parameters p and p_c in the DMC model $\mathcal{M}$ are assumed explicitly known. However, our reconstruction methods do not require to know the parameters of $\mathcal{M}$ in advance thanks to Theorem 2.3.3. Moreover, our methods provide natural estimators for the parameters in the DMC model, which is more computationally efficient than the estimators proposed in [66]. Before stating our algorithms to solve the above problem in the next section, we present here some theoretical results. The first one shows that when a network is given, the seed graph is uniquely determined by the duplication forest.

Lemma 2.3.1. *Given a network G with duplication forest Γ, for any two node sequences θ_1 and θ_2 that are compatible with Γ, graph $G_0^{\theta_1}$ is isomorphic to $G_0^{\theta_2}$.*

Proof. Assume that Γ consists of k binary trees $T_1,\cdots,T_k$ for some integer $k\geq 1$, and θ is a duplicate sequence compatible with Γ. For each graph G' in the graph sequence $(G_0^\theta,\cdots,G_n^\theta)$, we can associate it with a graph $\Pi(G')$ as follows. The vertex set of $\Pi(G')$ is $\{1,\cdots,k\}$ and two distinct vertices i and j are adjacent if and only if there exist some adjacent nodes g_i and g_j in G' such that g_i is a leaf in the tree T_i and g_j is a leaf in T_j.

Fix $t \in \{1, \cdots, n\}$. Denote the anchor node and duplicate node used to obtain G_t^θ from G_{t-1}^θ in the DMC model by u_t and v_t, respectively. Then v_t is the t-th element contained in θ. Since θ is compatible with Γ, u_t and v_t are leaves in the same tree in Γ. Note that for any vertex g that is distinct from u_t and v_t, by the definition of backward operator $\mathcal{R}$ we know that g is adjacent to u_t or v_t in G_t^θ if and only if g is adjacent to u_t in $G_{t-1}^\theta = \mathcal{R}_{v_t}^{u_t}(G_t^\theta)$. Therefore, we can conclude $\Pi(G_t^\theta) = \Pi(\mathcal{R}_{v_t}^{u_t}(G_t^\theta))$. Because t is arbitrary, we must have $\Pi(G_0^\theta) = \Pi(G_n^\theta)$. On the other hand, from the construction we know that $\Pi(G_0^\theta)$ is isomorphic to G_0^θ. In consequence, for two compatible duplicate sequences θ_1 and θ_2, since $G_n^{\theta_1} = G_n = G_n^{\theta_2}$, we can conclude that $G_0^{\theta_1}$ and $G_0^{\theta_2}$ are isomorphic, as required. □

Fix a pair of graph G and duplication forest Γ. Given a duplicate sequence $\theta = (v_1, v_2, \cdots, v_n)$, we shall associate it with three numbers that are crucial to our analysis. To this end, for each duplicate node v_i in θ, let $\delta(v_i)$ be the indicator function that takes value 1 if v_i is connected to its anchor node u_i in G_i^θ, and 0 otherwise; $\alpha(v_i)$ the number of the common neighbors of v_i and u_i, and $\beta(v_i) := \beta(v_i, G_i^\theta)$ the number of nodes adjacent only to v_i or u_i in G_i^θ. That is, we have $\alpha(v_i) = |N(v_i) \cap N(u_i)|$ and

$$\beta(v_i) = \Big|\Big(N(v_i) \setminus \big(N(u_i) \cup \{u_i\}\big)\Big) \bigcup \Big(N(u_i) \setminus \big(N(v_i) \cup \{v_i\}\big)\Big)\Big|.$$

Note that $2\delta(v_i) + 2\alpha(v_i) + \beta(v_i)$ is equal to the sum of the degree of v_i and that of u_i in G_i^θ. Finally, the sums

$$\delta(\theta) := \sum_{i=1}^n \delta(v_i), \quad \alpha(\theta) := \sum_{i=1}^n \alpha(v_i) \quad \text{and} \quad \beta(\theta) := \sum_{i=1}^n \beta(v_i)$$

are called the *homodimerization number*, *extension number* and *loss number* of θ, respectively.

The example below illustrates these definitions.

Example: Consider G_i and Γ_i in Fig. 2.1 and let $G = G_3$ and $\Gamma = \Gamma_3$. Then $\theta = (v_1, v_2, v_3)$ with $v_1 = 2, v_2 = 4, v_3 = 5$ is compatible with Γ. In addition, the anchor list determined by Γ and θ is $\pi = (3, 1, 2)$. It is easy to check that $\Gamma_i^\theta = \Gamma_i$ and $G_i^\theta = G_i$ for $0 \leq i \leq 3$. Furthermore, we have

$$\delta(v_1) = \delta(v_2) = 1, \delta(v_3) = 0;$$

$$\alpha(v_1) = \alpha(v_2) = \alpha(v_3) = 1;$$

$$\beta(v_1) = 0, \beta(v_2) = 1, \beta(v_3) = 2.$$

This implies $\delta(\theta) = 2$, $\alpha(\theta) = \beta(\theta) = 3$.

The theorem below says that the homodimerization number and the extension number are constant over all compatible duplicate sequences.

Theorem 2.3.2. *Given a network G with duplication forest Γ and two compatible duplicate sequences θ_1 and θ_2, we have $\delta(\theta_1) = \delta(\theta_2)$ and $\alpha(\theta_1) = \alpha(\theta_2)$.*

Proof. We shall establish the theorem by induction on the number of cherries in Γ. The base case that Γ is trivial, that is, it contains no cherry, is clear because this implies $\theta_1 = \theta_2$ as both of them contain no elements.

Now assume that Γ contains m cherries, and that the theorem holds when the number of cherries in the duplication forest is at most $m - 1$. Fix a cherry $\{u, v\}$ in Γ and choose a label g that is not used before. Consider the network G^* that is obtained from $\mathcal{R}_v^u(G)$ by relabeling u with g, and the duplication forest Γ^* obtained from Γ by replacing the cherry $\{u, v\}$ with a leaf labeled as g. Note that either node u or v (possible both) must appear in the duplicate sequence of θ_1; we replace them with g and denote the sequence with the first g removed by θ_1^*. Then θ_1^* is a duplicate sequence that is compatible with Γ^*. In addition, we have $\delta(\theta_1) = \delta(\theta_1^*) + 1$ if u and v are adjacent in G, and $\delta(\theta_1) = \delta(\theta_1^*)$ otherwise.

Similarly, the sequence θ_2^* obtained from θ_2 in the same way is also compatible with Γ^*. Now the induction assumption implies $\delta(\theta_1^*) = \delta(\theta_2^*)$. Together with

$$\delta(\theta_1) - \delta(\theta_1^*) = \delta(\theta_2) - \delta(\theta_2^*),$$

we have $\delta(\theta_1) = \delta(\theta_2)$, as required.

On the other hand, the number of edges increased from G_{i-1}^{θ} to G_i^{θ} is given by $\delta(v_i)$ and $\alpha(v_i)$, where v_i is the duplicate node. Together with Lemma 2.3.1, this implies

$$\delta(\theta_1) + \alpha(\theta_1) = |E(G_n)| - |E(G_0^{\theta_1})| = |E(G_n)| - |E(G_0^{\theta_2})| = \delta(\theta_2) + \alpha(\theta_2).$$

Since $\delta(\theta_1) = \delta(\theta_2)$, we have $\alpha(\theta_1) = \alpha(\theta_2)$. □

We can now establish the main result in this section, which relates the likelihood ratio of two compatible duplicate sequences to their loss numbers.

Theorem 2.3.3. *Given a network G with duplication history Γ, the likelihood ratio of two compatible duplicate sequences θ_1 and θ_2 is given by*

$$\frac{L(\theta_1 \,|\, G, \mathcal{M}, \Gamma)}{L(\theta_2 \,|\, G, \mathcal{M}, \Gamma)} = \Big(\frac{1-p}{2}\Big)^{\beta(\theta_1)-\beta(\theta_2)}.$$

In particular, $L(\theta_1 \,|\, G, \mathcal{M}, \Gamma) \geq L(\theta_2 \,|\, G, \mathcal{M}, \Gamma)$ if and only if $\beta(\theta_1) \leq \beta(\theta_2)$.

Proof. Let $\theta = (v_1, \cdots, v_n)$ be a duplicate sequence that is compatible with the duplication forest Γ. By Lemma 2.3.1 and Theorem 2.3.2, it is sufficient to note that

$$L(\theta \,|\, G, \mathcal{M}, \Gamma) = (1 - p_c)^{n-\delta(\theta)} p_c^{\delta(\theta)} p^{\alpha(\theta)} q^{\beta(\theta)}$$

holds with $q := (1 - p)/2$, which follows from

$$\mathbb{P}(G_i^\theta \,|G_{i-1}^\theta, \Gamma, \mathcal{M}) = (1 - p_c)^{1-\delta(v_i)} p_c^{\delta(v_i)} p^{\alpha(v_i)} q^{\beta(v_i)} \quad 1 \leq i \leq n.$$

□

One important consequence of Theorem 2.3.3 is that Problem 2 is equivalent to the following problem, which is computationally more tractable.

Problem 3. Given a network G and its duplication forest Γ, construct a duplicate sequence θ such that the loss number $\beta(\theta)$ is minimized among all sequences compatible with Γ.

2.4 Reconstruction Algorithms

In this section, we present two heuristic algorithms to solve Problem 3, and hence Problem 2. Moreover, these algorithms lead to natural estimators for the DMC parameters.

Before stating our reconstruction algorithms, we need some further notations and results. Two duplicate sequences $\theta_1 = (v_1, \cdots, v_n)$ and $\theta_2 = (v'_1, \cdots, v'_n)$ are said to be *adjacent* at position m for some $1 \leq m \leq n-1$ if we have $v'_m = v_{m+1}$, $v'_{m+1} = v_m$, and $v'_i = v_i$ for all other indices i.

Lemma 2.4.1. *Given a network G with duplication forest Γ, if θ_1 and θ_2 are two compatible duplicate sequences that are adjacent at position m, then we have $G_i^{\theta_1} = G_i^{\theta_2}$ for each $i \in \{0, \cdots, n\}$ with $i \neq m$.*

Proof. Let $\theta_1 = (v_1, \cdots, v_{m-1}, v_m, v_{m+1}, v_{m+2}, \cdots, v_n)$. So

$$\theta_2 = (v_1, \cdots, v_{m-1}, v_{m+1}, v_m, v_{m+2}, \cdots, v_n).$$

Clearly, we have $G_i^{\theta_1} = G_i^{\theta_2}$ for $i > m$. Hence we can set $G_{m+1} = G_{m+1}^{\theta_1} = G_{m+1}^{\theta_2}$.

To show $G_i^{\theta_1} = G_i^{\theta_2}$ for $i < m$, it suffices to show $G_{m-1}^{\theta_1} = G_{m-1}^{\theta_2}$. For $i \in \{m, m+1\}$, let u_i be the anchor node of v_i. Since θ_1 and θ_2 are both compatible with Γ, we know that $\{u_m, v_m\}$ and $\{u_{m+1}, v_{m+1}\}$ are two distinct cherries in $\Gamma_{m+1}^{\theta_1} = \Gamma_{m+1}^{\theta_2}$. In particular, the four nodes u_m, v_m, u_{m+1} and v_{m+1} are distinct in G_{m+1}. Therefore, by the definition of $\mathcal{R}$ we have

$$\mathcal{R}_{v_m}^{u_m}\big(\mathcal{R}_{v_{m+1}}^{u_{m+1}}(G_{m+1})\big) = \mathcal{R}_{v_{m+1}}^{u_{m+1}}\big(\mathcal{R}_{v_m}^{u_m}(G_{m+1})\big),$$

as required. □

Let θ_1 and θ_2 be two compatible duplicate sequences that are adjacent at position m. By Theorem 2.3.3, we know that $L(\theta_1 \,|\, G, \Gamma, \mathcal{M}) \geq L(\theta_2 \,|\, G, \Gamma, \mathcal{M})$ if and only if $\beta(\theta_1) \leq \beta(\theta_2)$ holds. On the other hand, Lemma 2.4.1 implies $\beta(\theta_1) \leq \beta(\theta_2)$ if and only if for $G_{m+1} = G_{m+1}^{\theta_1} = G_{m+1}^{\theta_2}$, we have

$$\beta(v_{m+1}, G_{m+1}) + \beta\big(v_m, \mathcal{R}_{v_{m+1}}(G_{m+1})\big) \leq \beta(v_m, G_{m+1}) + \beta\big(v_{m+1}, \mathcal{R}_{v_m}(G_{m+1})\big). \tag{2.1}$$

Motivated by the above observations, for two cherries $\{u, v\}$ and $\{u', v'\}$ in a duplication history Γ_t associated with network G_t, we say $\{u, v\}$ is more *favorable* than $\{u', v'\}$, denoted by $\{u, v\} \succ \{u', v'\}$, if

$$\beta(v, G_t) + \beta\big(v', \mathcal{R}_v^u(G_t)\big) < \beta(v', G_t) + \beta\big(v, \mathcal{R}_{v'}^{u'}(G_t)\big) \tag{2.2}$$

holds. Note that in general the relation $\succ$ is not transitive, that is, $\{u, v\} \succ \{u', v'\}$

and $\{u', v'\} \succ \{u^*, v^*\}$ do not imply $\{u, v\} \succ \{u^*, v^*\}$. In addition, we present a characterization of the favorability introduced above, which is computationally more efficient.

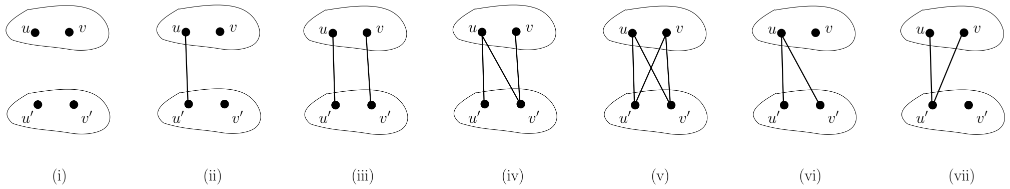

Figure 2.2: A schematic representation of the graph types used in the proof of Proposition 2.4.2. This classification is designed according to the edges between $\{u, v\}$ and $\{u', v'\}$, in which u and v, as well as u' and v', are interchangeable.

Proposition 2.4.2. *For two cherries $\{u, v\}$ and $\{u', v'\}$ in a duplication history Γ associated with network G, $\{u, v\} \succ \{u', v'\}$ if and only if either $\{u, v\} \subseteq N(u') \setminus N(v')$ or $\{u, v\} \subseteq N(v') \setminus N(u')$ holds.*

Proof. By the assumption of the proposition, we know that u, v, u', v' are four distinct nodes in G. For simplicity, one edge is said to between $\{u, v\}$ and $\{u', v'\}$ if it connects a node in $\{u, v\}$ and a node in $\{u', v'\}$. By swapping the labeling of u and v, and those of u' and v' if necessary, graph G can be classified into one of the seven types in Fig. 2.2, according to the edges between $\{u, v\}$ and $\{u', v'\}$. For instance, Type (i) means there is no edge between $\{u, v\}$ and $\{u', v'\}$ while Type (v) means there are four edges between them.

"$\Leftarrow$" Without loss of generality, we may assume $\{u, v\} \subseteq N(u') \setminus N(v')$, that is, graph G belongs to Type (vii) in Fig. 2.2. Note that for two nodes $x \in \{u', v'\}$ and $y \in V(G) \setminus \{u, v, u', v'\}$, x and y are adjacent in G if and only if they are adjacent in $\mathcal{R}_v^u(G)$. This implies

$$\beta(v', G) - \beta\big(v', \mathcal{R}_v^u(G)\big) = 1$$

because $\{u, v\} \subseteq N(u') \setminus \big(N(v') \cup \{v'\}\big)$, and u and v are merged to form $\mathcal{R}_v^u(G)$. On the other hand, we have

$$\beta(v, G) = \beta\big(v, \mathcal{R}_{v'}^{u'}(G)\big)$$

because neither u' nor v' contributes to $\beta(v, G)$ or $\beta\big(v, \mathcal{R}_{v'}^{u'}(G)\big)$. Therefore, we can conclude

$$\beta(v, G) + \beta\big(v', \mathcal{R}_v^u(G)\big) < \beta(v', G) + \beta\big(v, \mathcal{R}_{v'}^{u'}(G)\big),$$

as required.

"⇒" To establish this direction, assuming $\{u, v\} \succ \{u', v'\}$, then we need to show that graph G must belong to Type (vii) in Fig. 2.2. Indeed, if graph G belongs to Type (i)-(v), then we have

$$\beta(v, G) + \beta\big(v', \mathcal{R}_v^u(G)\big) = \beta(v', G) + \beta\big(v, \mathcal{R}_{v'}^{u'}(G)\big),$$

a contradiction to $\{u, v\} \succ \{u', v'\}$. On the other hand, if G belongs to Type (vi), then we have $\{u', v'\} \succ \{u, v\}$, contradicting $\{u, v\} \succ \{u', v'\}$. □

Now we present our main inference algorithm called cherry greedy (CG), which runs as follows: At every backward reconstruction step, we choose a node from the most favorable cherry C, that is, the number of cherries C' with $C \succ C'$ is maximized. If several cherries are equally favorable, we randomly choose one of them. More precisely, starting from $G_t := G$ and $\Gamma_t := \Gamma$, we choose a most favorable cherry $\{u, v\}$ from Γ_t and randomly choose one node from the cherry, say v_t, as the duplicate node at this step. Then we construct G_{t-1} as $\mathcal{R}_{v_t}(G_t)$ and $\Gamma_{t-1} = \mathcal{R}_{v_t}(\Gamma_t)$. This process continues until G_0 is obtained. Note that Proposition 2.4.2 provides an efficient way to find the most favorable cherry.

Besides algorithm CG, we also introduce another greedy algorithm called minimum loss number (MLN), which is different from CG in that at each backward step a pair of duplicate and anchor nodes having the smallest loss number is chosen among all cherries in the duplicate forest. Let n be the number of the vertices in the input PPI network. Since $\beta(v)$ for each vertex v can be computed in time $O(n)$, and the backward operator for the duplication forest and network can be done in $O(n)$, we know that each backward step in MLN has running time $O(n^2)$, and hence the running time for MLN is $O(n^3)$. On the other hand, a similar analysis shows that the theoretical running time for CG is $O(n^4)$. Algorithm MLN is conceptually simpler than CG and typically runs faster in our experimental studies. However, CG is more accurate (see Section 2.5 for more details). We have run some greedy algorithms in an aim to obtain optimal solutions. The optimal solutions, in the sense of likelihood, have likelihood larger than those obtained by the algorithms by several times. The Kendall's tau is slightly larger than those obtained by CG and MLN by no more than two times.

From the results in Section 2.3 and the two algorithms presented above, it is clear that the parameters of the DMC model are not used in our inference framework. Moreover, here we will present a method by which the parameters can be estimated after a growth history being inferred.

To this end, assume that a growth history $\mathcal{H} = (G_0, \cdots, G_n)$, together with the duplicate sequence $(v_1, \cdots, v_n)$ and anchor list $\{u_1, \cdots, u_n\}$, is given. Note that for each neighbor w of node u_i in G_{i-1}, the probability that w is adjacent to both u_i and v_i in G_i is p. In other words, the extension number $\alpha(v_i)$ at i-th step, that is, the number of the common neighbors shared by u_i and v_i in G_i, has the binomial distribution with parameters p and $\beta(u_i) + \alpha(v_i)$, where $\beta(u_i) + \alpha(v_i)$ is the number of neighbors that u_i has in G_{i-1}. On the other hand, the random variable $\delta(v_i)$ has Bernoulli distribution with parameter p_c. Therefore, we are led to propose the estimators

$$\hat{p} = \frac{\alpha(\theta)}{\beta(\theta) + \alpha(\theta)} \qquad \text{and} \qquad \hat{p}_c = \frac{\delta(\theta)}{n} \tag{2.3}$$

to estimate the parameters p and p_c respectively.

2.5 Experimental Results

Our reconstruction algorithms, minimum loss number (MLN) and cherry greedy (CG), have been implemented and are available upon request. Given a network G and duplication forest Γ, each outputs a hypothetical duplicate sequence θ that approximates the one with the minimum loss number.

2.5.1 Simulation Studies

To compare and validate our algorithms, we generated 100 random networks for each DMC model $\mathcal{M}(p, p_c)$, where the parameters p and p_c are chosen respectively from $\{0.05, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Each network contains 100 nodes and is generated from the same seed graph K_2 (i.e., the graph with two nodes and one edge). For each simulated network G, its duplication forest Γ and duplicate sequence θ_{real} were recorded. Next, we reconstructed duplicate sequences using our algorithms. The one output from MLN is denoted by θ_{MLN}, and the one from CG by θ_{CG}.

To compare the performance of MLN and that of CG, we calculated the average loss number for θ_{MLN} and θ_{CG} for the simulated data set. The results are summarized in Table 2.1, from which it is clear that on average CG has smaller loss number than MLN does. Therefore, CG performs better in terms of solving Problem 3 and we recommend it for accuracy. However, MLN is much faster and we recommend it when the underlying network is large.

To further assess their performance, we measured the difference between the

p_c p	0.05	0.1	0.3	0.5	0.7	0.9	
0.05	**18.25**	**28.43**	**65.98**	**101.24**	139.48	**173.93**	MLN
	18.53	28.79	66.40	101.72	**139.31**	174.33	CG
0.1	**21.45**	31.07	67.12	107.94	144.57	183.64	MLN
	22.13	**31.03**	**67.00**	**107.55**	**144.02**	**182.92**	CG
0.3	**30.85**	45.62	88.97	138.09	188.25	233.74	MLN
	32.50	**44.63**	**87.51**	**136.05**	**184.84**	**228.16**	CG
0.5	**63.03**	76.55	128.56	197.13	258.98	317.77	MLN
	65.25	**73.46**	**122.98**	**189.94**	**252.11**	**306.20**	CG
0.7	**110.53**	125.90	191.22	265.66	332.09	391.03	MLN
	112.52	**119.70**	**181.89**	**252.50**	**314.70**	**374.29**	CG
0.9	**114.76**	128.09	167.59	219.05	267.88	303.71	MLN
	117.61	**123.88**	**162.70**	**211.43**	**257.47**	**293.07**	CG

Table 2.1: Comparing the performance of the two algorithms: minimum loss number (MLN) and cherry greedy (CG). Columns 2 to 7 correspond to $p_c = 0.05, 0.1, \ldots, 0.9$; and rows 2 to 7 for $p = 0.05, 0.1, \ldots, 0.9$. For each pair of parameters, p and p_c, 100 simulated networks were generated using the DMC model $\mathcal{M}(p, p_c)$. Each entry in the table consists of two numbers: The top one (respectively, the bottom one) is the average of the loss numbers of the reconstructed histories by MLN (respectively, by CG). A smaller loss number corresponds to a higher likelihood of the reconstructed history, and hence a better reconstruction. Smaller averages are highlighted in bold face.

inferred duplicate sequence and the 'real' one. One popular index for this purpose is Kendall's tau K_τ [7, 66]. Formally, for two sequences $\theta_1 = (v_1, \cdots, v_n)$ and $\theta_2 = (v'_1, \cdots, v'_n)$ on a set of nodes, $K_\tau(\theta_1, \theta_2)$ is defined as

$$K_\tau(\theta_1, \theta_2) = \frac{2(n_c - n_d)}{n(n-1)},$$

where n_c is the number of concordant pairs, and n_d the number of discordant pairs. For example, considering $\theta_1 = (1, 2, 3, 4)$ and $\theta_2 = (4, 2, 1, 3)$, then we have $n = 4$, $n_c = 2$ and $n_d = 4$, and hence $K_\tau(\theta_1, \theta_2) = -1/3$. Note that $K_\tau(\theta_1, \theta_2) = 1$ if and only if the sequences are identical, and $K_\tau(\theta_1, \theta_2) = -1$ if and only if they are exactly opposite.

For comparison, we reconstructed duplicate sequence θ_{NetArch} using NetArch [66]. Moreover, we computed $K_\tau(\theta_{\text{real}}, \theta)$ for $\theta \in \{\theta_{\text{MLN}}, \theta_{\text{CG}}, \theta_{\text{NetArch}}\}$ and calculated the average K_τ for each pair of parameters. The results are summarized in Fig. 2.3. In order to obtain a more detailed comparison between NetArch and CG, we counted how many times one method outperformed the other. More precisely, for each of the 100 networks with a given pair of parameters, the algorithm by which the sequence reconstructed has higher Kendall's tau received one vote (when there is a tie, we split the vote). Entries in Table 2.2 represent the total number of votes received for the given p, p_c and algorithm.

From these results, we can see that compared to NetArch, our algorithms substantially increase the values of Kendall's τ. This agrees with the intuition that incorporating more information often leads to good reconstruction methods.

2.5.2 Parameters Estimation

As discussed in Section 2.4, after a growth history of G being inferred, the parameters p and p_c in the DMC model $\mathcal{M}(p, p_c)$ that generates G can be estimated

p_c p	0.05	0.1	0.3	0.5	0.7	0.9	
0.05	39	24	13	12	16	16	NetArch
	61	**76**	**87**	**88**	**84**	**84**	CG
0.1	28	27	28	28	28	26	NetArch
	72	**73**	**72**	**72**	**72**	**74**	CG
0.3	36	31.5	29	28	24	28	NetArch
	64	**68.5**	**71**	**72**	**76**	**72**	CG
0.5	23	21	22.5	18	20	21	NetArch
	77	**79**	**77.5**	**82**	**80**	**79**	CG
0.7	32	16	15	23.5	14	13.5	NetArch
	68	**84**	**85**	**76.5**	**86**	**86.5**	CG
0.9	11	16	10	10	4	6	NetArch
	89	**84**	**90**	**90**	**96**	**94**	CG

Table 2.2: Detailed comparison of two reconstruction methods: NetArch and cherry greedy (CG). For each specified p_c and p in the table, we generated a network under the DMC model $\mathcal{M}(p, p_c)$ from an edge (seed graph) until it contains 100 nodes. We then ran algorithms NetArch and CG on this network. The algorithm with higher Kendall's tau received one vote, the other zero vote. When there was a tie, we split the vote. This procedure repeated 100 times. Entries in the table represent the total number of votes received for the given p_c, p and algorithm. A higher number of total votes, highlighted in bold face, corresponds to a better reconstruction.

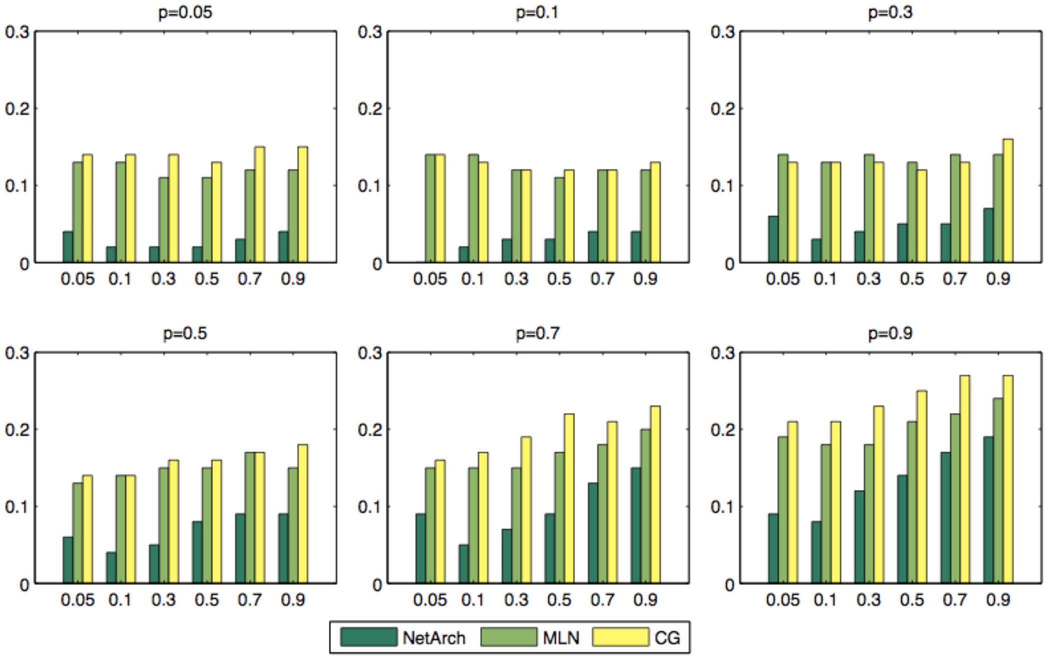

Figure 2.3: Average accuracy of three reconstruction methods. The x-axes show the DMC parameter p_c used to grow the network, and the y-axes show the average Kendall's tau for three reconstruction methods. A higher Kendall's tau indicates that the history reconstructed is closer to the real one, and hence a better reconstruction.

using the estimators $\hat{p}$ and $\hat{p}_c$ defined in Eq. (2.3). Recall that in [66] a pair of estimators, denoted by p^{best} and p_c^{best}, is also proposed.

To compare the performance of these two sets of estimators, we generated 100 networks using DMC models with random parameters. For each simulation, we first generated a pair of parameters p and p_c uniformly from the interval $(0, 1)$, and then obtained one graph G with 30 nodes from the seed graph K_2 using the DMC model $\mathcal{M}(p, p_c)$, as well as the associated duplication forest Γ. Next, we estimated the parameters using estimators $\hat{p}$ and $\hat{p}_c$, as well as p^{best} and p_c^{best}. Now the accuracy of the estimator $\hat{p}$ can be measure by $|p - \hat{p}|$: The closer this difference to 0, the better the estimation is. Similarly, we can measure the accuracy of the other three estimators. The box plots for these four differences over 100 simulations are presented in Fig.2.4, from which we can see that our method has smaller means of errors and smaller confidence intervals for estimating p and p_c.

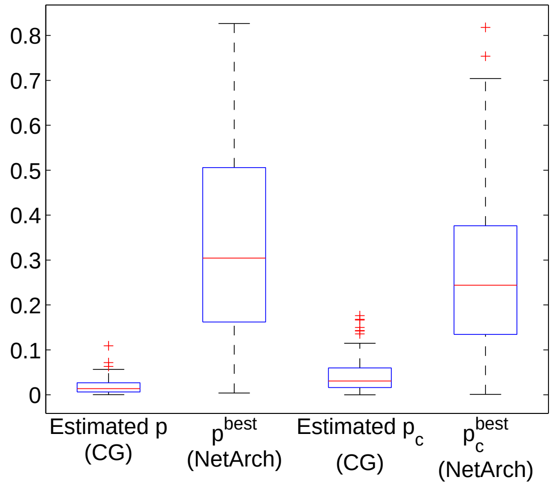

Figure 2.4: Box plot for errors of parameter estimation. Here 100 pairs of parameters (p, p_c) were generated uniformly from the interval $(0, 1)$. For each pair of parameters p and p_c, one network with 30 nodes was generated using the DMC model $\mathcal{M}(p, p_c)$. Then the four estimators $\hat{p}$, $\hat{p}_c$, p^{best} and p_c^{best} were computed, and the four error numbers, $|p - \hat{p}|$, $|p_c - \hat{p}_c|$, $|p - p^{\text{best}}|$ and $|p_c - p_c^{\text{best}}|$ were calculated.

2.5.3 Application to Real PPI Networks

Note that the methods developed in this chapter are based on the assumption that the observed network is generated by the DMC model. The adequacy of this assumption can be checked by comparing the topological characteristics of the DMC model and the real network. This work has been done by such as [43, 62]. In the cases that the assumption is violated but the gene tree is believable, we can still get positive Kendall's tau, which in general greater than those obtained by NetArch. If the gene tree is untrue, the results depend.

We downloaded 460 gene trees from [23]: These trees were inferred from protein sequences extracted from DIP (http://dip.doe-mbi.ucla.edu/dip/Main.cgi) and reconciled using NOTUNG [22]. The gene trees contain genes found in *S. cerevisiae* (budding yeast), *D. melanogaster* (fruitfly) and *C. elegans* (worm). We derived a family of species-specific gene trees by projecting the downloaded gene trees on the three species respectively. For each species, these species-specific gene trees were collected to form a duplication forest for our purpose of reconstructing network duplication history. Although the original gene trees in [23] were timed, we only made use of their topological information in this experiment. In addition, we downloaded the PPI networks from the database DIP for the three species. The size of these networks and the number of trees in the corresponding duplication forests are given in Table 2.3.

For each G of these PPI networks, we inferred a duplicate sequence θ using our algorithm CG. We then constructed the anchor list π from the duplication forest and θ. Finally, we obtained the growth history $\mathcal{H}$ of G from θ and π.

Using Eq. (2.3) we estimated the growth parameters p and p_c for each PPI network (Table 2.3). Our estimation of p_c is in line with the assertion that p_c is smaller than 0.1 [29, 98]. In contrast, the parameters p_c and p estimated for the *S. cerevisiae* network by NetArch are respectively 0.7 and 0.6 [66]. Moreover, when the growth parameters were estimated by NetArch in [66], further information on protein ages was used. Therefore, we demonstrated again the advantage of incorporating duplication history in growth history reconstruction.

The reconstructed growth history enables us to further analyze two features in the growth of these PPI networks: change of modularity and the relation between the number of duplications and the degree of nodes in the extant network.

To measure modularity, we use the *clustering coefficient* that is defined as the ratio of the number of edges between the vertices in $N(v)$ to $|N(v)|(|N(v)| - 1)/2$ for each vertex v [9]. Then the clustering coefficient of a graph is the average of

	S.cerevisiae	*C. elegans*	*D. melanogaster*
Number of vertices	1361	2624	7027
Number of duplication trees	213	1912	5033
$\hat{p}$	0.061	0.021	0.026
$\hat{p}_c$	0.053	0.048	0.024

Table 2.3: Parameters and estimated parameters for three PPI networks downloaded from DIP. The corresponding duplication histories were obtained from the reconciled gene trees reported in [23]. $\hat{p}$ and $\hat{p}_c$ are the estimated parameters in the DMC model.

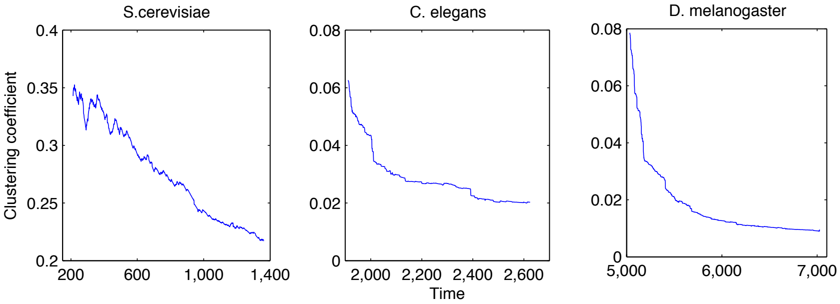

Figure 2.5: Change in clustering coefficients over time in three PPI networks. Here the growth histories were constructed by CG. The x-axes show the number of vertices in the networks in the histories while the y-axes show the values of clustering coefficient. An overall trend of clustering coefficient decreasing was revealed.

clustering coefficients over all vertices. For each PPI network G, a growth history $\mathcal{H} = (G_0, G_1, \ldots, G_n)$ was obtained, where G_n is the extant network G and the number of vertices in the seed network G_0 equals to the number of trees in the corresponding duplication forest. The clustering coefficients of these intermediate networks for each of the three PPI networks were computed and presented in Fig. 2.5. Note that for each PPI network, clustering coefficients decrease as the network evolved over time, a trend only reported in [66] for the *S.cerevisiae* PPI network and in [53] for metabolic networks.

Given a growth history $\mathcal{H}$, the number of duplications of a node v in the extant

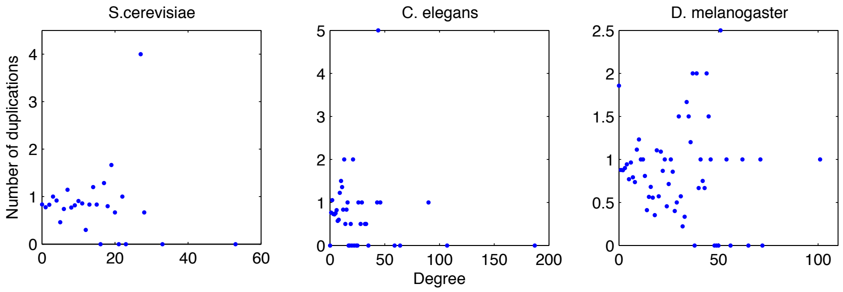

Figure 2.6: Relationship between degree and number of duplications in three PPI networks. The x-axes show the values of degrees in the extant networks while the y-axes show the average number of duplications for the nodes with given degree. No significant monotone relation between these two quantities has been found. However, inverse relation is suggested in [32, 60].

network is defined as the number of times v was duplicated in the history, that is, the number of v contained in the anchor list determined by $\mathcal{H}$. It has been suggested (for examples, [32] and [60]) that the larger the degree of a node, the smaller the number of duplications the node has. However, our results on these three PPI networks show no significant relation between them (see Fig.2.6), which agree with the findings in [61].

2.6 Conclusion

Assuming the PPI networks evolve according to the DMC model, we have presented a likelihood-based approach for recovering the most probable network evolutionary history by exploiting the known duplication history trees of paralogs in the network. Through a series of reduction of the search space of all histories to (i) compatible duplicate sequences and then (ii) the set of favored duplicate nodes, we have provided a computationally efficient framework. Our approach successfully retraces the network evolution especially in the scenario that the labels of ancestor nodes are not necessarily to be one of the duplicates. As a useful by-product of our reconstruction, estimators for the model parameters are proposed.

The reconstruction framework presented in this chapter is described in the context of the DMC model, and it would be interesting to see how it can be generalized to other network evolutionary models. Another possible extension to this work is to investigate network evolutions across different species, which remains a challenging problem (see [75] for a parsimony approach). Finally, the complexity of solving Problem 3 requires further research, to yield more insights into the performance of the algorithms proposed here.

Chapter 3

Degree Distribution of Large Networks Generated by The Partial Duplication Model

3.1 Introduction

Arguably, one of the most fundamental models in the class of duplication models is the partial duplication (PD) model studied by [18]. In this model, at each step an anchor node is chosen uniformly from the current network and a new node is added and independently this new node is connected to each neighbor of the anchor node with selection probability p (see Subsection 1.3.2). This model is particularly attractive for two reasons: it captures the basic principles behind PPI evolution, and its simplicity enables us to conduct rigorous mathematical analysis. By studying this model we can gain insights into other more sophisticated DD models.

Here we focus on the degree distribution of the PD model. By degree distribution we mean the sequence $\{f_t(k)\}_{k\geq 0}$, where $f_t(k)$ denotes the *expected proportion* of nodes of degree k at time t. Note that the PD model is studied at the ensemble level in this chapter, that is, we are mainly interested in the average behavior over many different realizations. One general tool to study the degree distribution of random networks is the master equation of $f_t(k)$ (see [21] and the references therein). However, despite the simplicity of the PD model, its master equation is still too complicated to be solved analytically and no analytic solution is known yet, except for the full duplication model, the special case when $p = 1$ [79]. Instead, the attention has been centered on the limiting degree distribution, which provides valuable information on the long run behavior of the model [11, 18, 52].

Since isolated nodes are generally irrelevant to the observed PPI networks, here we also study the subgraph consisting of all non-isolated nodes in the PD model. If $f(0) < 1$, the limiting degree distribution in the connected components does exist and it is $(0, 0, \cdots)$, that is, the expected fraction of degree k in this subgraph tends to 0 for all $k \geq 1$. Therefore, the limiting degree distribution does not follow a power law in this region. For the case when $f(0) = 1$, we assume that the limiting degree distribution exists and then prove that the entries in this limiting distribution must be strictly positive, and they satisfy a system of equations. In addition, the limiting degree distribution in this region also follows a power law. Our results are then applied to three real PPI networks to obtain the power law exponent and selection probability for each network.

An important property of the PD model is that it may produce graphs containing a large proportion of isolated nodes, that is, $f(0)$ is typically large when p is small. Therefore, it is of interest to know the behavior of $f(0)$ relative to selection probability p. Indeed, one central problem for the PD model, as stated in Section 3.1 of [11], is to characterize the values of p for which $f_t(0)$ tends to 1. Here we attempt to answer this question by showing that there is a phase transition point $p_0 \in [\frac{1}{2}, \frac{1}{\sqrt{2}}]$ for the expected proportion of isolated nodes converging to 1. More precisely, $f(0) < 1$ for $p_0 < p \leq 1$, and $f(0) = 1$ for $0 < p < p_0$. In addition, we also obtain upper and lower asymptotic bounds on the convergence rate of $\{f_t(0)\}_{t\geq 0}$, as well as a uniform upper bound on the convergence rate of $\{f_t(k)\}_{t\geq 0}$ for all $k \geq 1$.

Prior to studying limiting degree distribution, we need to establish its existence, that is, whether the limit of $f_t(k)$ for a given k exists as t approaches infinity. For the special case $k = 0$, the existence of $f(0) = \lim_{t\to\infty} f_t(0)$ was proved by [11] by showing $\{f_t(0)\}_{t\geq 0}$ is indeed a non-decreasing sequence. However, the other cases remained open and it was often *assumed* that they do exist in previous studies. For example, Lemma 2 in [11] states that for $k \geq 1$, if $f_t(k)$ tends to a limit, then this limit must be 0. We close this gap by showing that the limit of $f_t(k)$ *does* exist for each $k \geq 0$, and hence the sequence $(f_t(0), f_t(1), f_t(2), \cdots)$ converges pointwise to $(f(0), 0, 0, \cdots)$ as t approaches infinity.

The structure of the rest of this chapter is as follows. In the next section, we describe the PD model and the master equation for the expected degree sequence. In Section 3.3, we present some preliminary results. Section 3.4 is devoted to the bounds on rates of convergence. In Section 3.5 we study the limiting degree distribution of the subgraph with all isolated nodes removed, and apply the results to three real PPI networks. In Section 3.6, we establish the existence of limiting degree distribution and show a possible interval for the phase transition point of the expected fraction of isolated nodes converging to 1. Finally, we end with Section 3.7 for some concluding comments and possible directions for further study.

3.2 The Model

Let $\mathbf{F}_t(k)$ denote the number of nodes of degree k in G_t and let $\mathbf{F}_t = (\mathbf{F}_t(0), \mathbf{F}_t(1), \cdots)$ be the corresponding degree sequence. In addition, set $F_t(k) := \mathbb{E}\left[\mathbf{F}_t(k)\right]$ and let

$f_t(k) = F_t(k)/t$ be the expected proportion of nodes with degree k in G_t. Since the number of nodes in G_t is always t, we know $\mathbf{F}_t(k) = 0$, and hence also $F_t(k) = f_t(k) = 0$, for all $k \geq t$. Here we also use the convention $F_t(-1) = 0$ for all $t \geq t_0$.

The expected degree sequence satisfies the recursion equation

$$F_{t+1}(k) = \Big(1 - \frac{pk}{t}\Big)F_t(k) + \frac{p(k-1)}{t}F_t(k-1) + \frac{1}{t}\sum_{j \geq k}\binom{j}{k}p^k q^{j-k}F_t(j) \qquad (3.1)$$

for all $k \geq 0$ and $t \geq t_0$, which is often referred to as the *master equation* for the expected degree sequence (see [21] for a general discussion on master equation). The master equation for the PD model was first studied by [18], and its complete form as above was presented by [11], and also [52]. The correctness of Eq. (3.1) can be seen in the following way. The first term on the right-hand side describes the contribution of nodes of degree k in G_t; the second term corresponds to the case in which a node of degree $k-1$ in G_t is connected to the new node in G_{t+1}, while the last term represents the probability that the new node at step $t+1$ has degree k.

When the selection probability p and the seed graph G_{t_0} are given, it is clear that the degree sequence $(F_t(0), F_t(1), \cdots)$ is uniquely determined by Eq. (3.1). Therefore, much information concerning the long run behavior of the model can be obtained from the master equation. As an example, we present the solution of Eq. (3.1) for the special case in which $p = 1$ and the seed graph is K_2, that is, the graph contains exactly one edge.

Example: Considering the PD model $\mathcal{M}(K_2, 1)$, then we have

$$F_t(k) = \frac{2(t-k)}{(t-1)} \qquad \text{and} \qquad f_t(k) = \frac{2(t-k)}{t(t-1)}$$

for $t \geq 2$ and $1 \leq k \leq t$. In particular, we have $\lim_{t\to\infty} f_t(k) = 0$, and $\lim_{t\to\infty} f_t(k+$

$1)/f_t(k) = 1$, for all $k \geq 1$.

The analytic solution in the above example was given in [50], which can also be easily verified by using Eq. (3.1) and the boundary condition that $F_2(1) = 2$ and $F_2(k) = 0$ for $k \neq 2$. For the special case when $p = 1$, a general solution of Eq. (3.1) for any seed graph was obtained by [79]. However, no analytic solution for other cases are known to us, and in this chapter we will study the long run behavior of $f_t(k)$ without using its analytic form.

Since G_t may contain a large portion of isolated nodes and isolated nodes do not correspond to nodes in observed PPI network, so we are led to study the *non-isolated* subgraph G_t^+ obtained from G_t by removing all isolated nodes. Clearly, the number of isolated nodes contained in G_t is $F_t(0)$. By Eq. (3.1), we have

$$F_{t+1}(0) = F_t(0) + \frac{1}{t}\sum_{k\geq 0} F_t(k)q^k \tag{3.2}$$

for $t \geq t_0$. To study the non-isolated subgraph, let $\mathbf{F}_t^+$ denote the number of nodes contained in G_t^+, set $F_t^+ := \mathbb{E}\left[\mathbf{F}_t^+\right]$ and let $f_t^+(k) := F_t(k)/F_t^+$ be the expected proportion of nodes with degree k in G_t^+. Then clearly we have $\mathbf{F}_t^+ = t - \mathbf{F}_t(0)$, $F_t^+ = t - F_t(0)$ and $f_t^+ = 1 - f_t(0)$ for all $t \geq t_0$. Together with Eq. (3.2), this implies

$$F_{t+1}^+ = F_t^+ + \frac{1}{t}\sum_{j\geq 1} F_t(j)(1-q^j) \tag{3.3}$$

for $t \geq t_0$.

3.3 Preliminary Results and Notations

In this section, we introduce some notations and present several preliminary results that will be used later in the chapter. To begin with, we recall some standard asymptotic notations. For two functions $a(x)$ and $b(x)$ of a real variable x, $a(x) =$

$O(b(x))$ (as $x \to \infty$) means there exists a constant $\beta > 0$ such that $a(x) < \beta b(x)$ for all large x. In addition, we write $a(x) = O(b(x))$ if $b(x) = \Omega(a(x))$ holds. If both $a(x) = O(b(x))$ and $a(x) = \Omega(b(x))$ hold, then we write it as $a(x) = \Theta(b(x))$. Finally, $a(x) = o(b(x))$ means $\lim_{x\to\infty} a(x)/b(x) = 0$. Note that similar notations are used for real sequences $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$.

The lemma below is elementary; it is included here for completeness.

Lemma 3.3.1. *For a constant $c > 0$, as $t \to \infty$, we have*

$$\Gamma(t+c) = \Gamma(t)\Big(1 + O(t^{-1})\Big)t^c \tag{3.4}$$

for Gamma function Γ, and

$$\prod_{s=1}^{t}\Big(1 + \frac{c}{s}\Big) = \Theta(t^c). \tag{3.5}$$

Proof. Eq. (3.4) follows immediately from Stirling's formula for Gamma function (see, for example, Lemma 1 in [18]). To establish Eq. (3.5) , let r be the smallest integer larger than c, and put $\kappa(t) := \prod_{s=r}^{t}(1 + (c/s))$. By the Taylor series for $\ln(1+x)$, we have $x(1-x/2) < \ln(1+x) < x$ for $0 < x < 1$, which leads to

$$\sum_{s=r}^{t}\Big(\frac{c}{s} - \frac{c^2}{2s^2}\Big) \leq \ln\kappa(t) \leq \sum_{s=r}^{t}\frac{c}{s}$$

for $t \geq r$. Since $\sum_{s=1}^{t} 1/s = \Theta(\ln t)$ and $\sum_{s=1}^{t} 1/s^2 \leq 2$, Eq. (3.5) follows from the above inequalities. □

In order to study degree distribution, it is often instructive to consider the average degree first. To this end, let $\mathbf{e}_t$ be the number of edges in G_t and set $e_t := \mathbb{E}[\mathbf{e}_t]$. Then $\mathbf{D}_t$, defined as the average degree of nodes in G_t, is equal to

$2\mathbf{e}_t/t$. In addition, setting $D_t := \mathbb{E}[\mathbf{D}_t]$ then we have

$$2e_t = \sum_{k\geq 0} kF_t(k) = tD_t. \tag{3.6}$$

As a generalization of Lemma 1 in [11], the following result shows that the expected average degree in G_t is determined by D_{t_0}, the average degree of the seed graph, and selection probability p.

Proposition 3.3.2. *The expected average degree D_t is given by*

$$D_t = D_{t_0}\frac{\Gamma(t+2p)\Gamma(t_0+1)}{\Gamma(t+1)\Gamma(t_0+2p)} \tag{3.7}$$

for $t \geq t_0$. In particular, as $t \to \infty$ we have

$$D_t = D_{t_0}\frac{\Gamma(t_0+1)}{\Gamma(t_0+2p)}t^{2p-1}\Big(1+O(t^{-1})\Big) \tag{3.8}$$

and hence $D_t = \Theta(t^{2p-1})$.

Proof. As stated in the proof of Lemma 1 in [11], e_{t+1} is given by

$$e_{t+1} = \Big(1+\frac{2p}{t}\Big)e_t$$

for $t \geq t_0$. Together with Eq. (3.6) and $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$, the above formula leads to

$$D_t = D_{t_0}\frac{t_0}{t}\prod_{i=t_0}^{t-1}\frac{i+2p}{i} = D_{t_0}\frac{\Gamma(t+2p)\Gamma(t_0+1)}{\Gamma(t+1)\Gamma(t_0+2p)},$$

which establishes (3.7). Finally, Eq. (3.8) follows from Eq. (3.7) by Eq. (3.4) in Lemma 3.3.1. □

From the above proposition, it is clear that if p_0 is a phase transition point for the growth pattern of the average degree. More specifically, as t goes to infinity, the average degree strictly decreases to 0 when $0 < p < p_0$, strictly increases to infinity when $p_0 < p \leq 1$. As we shall see later, p_0 is also the phase transition point for several other properties of the PD model.

We end this section with the following two technical results concerning the long run behavior of $F_t(k)$, which will also be used in Section 3.5.

Lemma 3.3.3. *Let $0 < p < 1$. For each $k \geq 0$, there exists an integer $\tau_k \in [t_0, t_0 + k]$ such that $F_t(k) > 0$ for all $t \geq \tau_k$.*

Proof. By Eq. (3.2), the lemma clearly holds for $k = 0$. We shall establish the other cases by induction on k. For the base case $k = 1$, by Eq. (3.1) we have

$$F_{t+1}(1) = \Big(1 - \frac{p}{t}\Big)F_t(1) + \frac{1}{t}\sum_{j\geq 1} jpq^{j-1}F_t(j) = F_t(1) + \frac{1}{t}\sum_{j\geq 2} jpq^{j-1}F_t(j), \quad (3.9)$$

which implies that $\{F_t(1)\}_{t\geq t_0}$ is non-decreasing as $F_t(j) \geq 0$ for all j. Since G_{t_0} contains at least one edge, $F_{t_0}(j) > 0$ holds for some $j \geq 1$. Therefore, we have $F_{t_0+1}(1) > 0$, and hence $F_t(1) > 0$ for $t \geq t_0 + 1$.

For the induction step, fix $k \geq 1$ and assume there exists a number $\tau_k \in [t_0, t_0 + k]$ so that $F_t(k) > 0$ for $t \geq \tau_k$. Let $\tau_{k+1} := \max\{\tau_k, p(k+1)\}$; then $\tau_{k+1} \in [t_0, t_0 + k]$ and hence it suffices to show $F_t(k+1) > 0$ for $t \geq \tau_{k+1}$. Indeed, by Eq. (3.1) and the choice of τ_{k+1} we have

$$F_{t+1}(k+1) \geq \Big(1 - \frac{p(k+1)}{t}\Big)F_t(k+1) + \frac{pk}{t}F_t(k) \geq \frac{pk}{t}F_t(k) > 0$$

for all $t \geq \tau_{k+1}$, which completes the proof of the induction step, and hence also the lemma. □

Proposition 3.3.4. *Let $0 < p < 1/2$. The sequence $\{F_t(1)\}_{t\geq t_0+1}$ strictly increases to infinity as $t \to \infty$. Moreover, we have $F_t(1) = \Omega(\ln t)$.*

Proof. Denoting $F_{t_0+1}(1)$ by α, by Lemma 3.3.3 and its proof, we know that $\{F_t(1)\}_{t\geq t_0+1}$ is strictly increasing and bounded below by $\alpha > 0$. We next show $F_t(2) \geq p\alpha/2$ for $t \geq 2t_0+1$. To this end, by $2p < 1$ and Eq. (3.1) we have

$$F_{t+1}(2) = F_t(2) - \frac{2p}{t}F_t(2) + \frac{p}{t}F_t(1) + \frac{1}{t}\sum_{j\geq 2}\binom{j}{2}p^2q^{j-2}F_t(j) \geq \Big(1-\frac{1}{t}\Big)F_t(2) + \frac{p\alpha}{t}$$

for $t \geq t_0+1$. Consider the sequence $\{\beta_t\}_{t\geq t_0+1}$ defined as

$$\beta_t = \frac{t_0}{t-1}\beta_{t_0+1} + \Big(1 - \frac{t_0}{t-1}\Big)p\alpha$$

with $\beta_{t_0+1} := F_{t_0+1}(2)$. Since

$$\beta_{t+1} = \Big(1-\frac{1}{t}\Big)\beta_t + \frac{p\alpha}{t}$$

holds for $t \geq t_0+1$, we have $F_t(2) \geq \beta_t$ for $t \geq t_0+1$, which implies

$$F_t(2) \geq \beta_t \geq p\alpha/2$$

for $t \geq 2t_0+1$. Together with Eq. (3.9), this leads to

$$F_{t+1}(1) \geq F_t(1) + \frac{2pq}{t}\frac{p\alpha}{2} = F_t(1) + \frac{p^2q\alpha}{t} \geq p^2q\alpha\sum_{s=2t_0+1}^{t}\frac{1}{s}$$

for $t \geq 2t_0+1$, from which we can conclude $F_t(1) = \Omega(\ln t)$, and in particular $F_t(1) \to \infty$ as $t \to \infty$. □

Remark: It would be interesting to see whether $F_t(k) = \Omega(\ln t)$ holds for all $k \geq 1$.

3.4 Rates of Convergence

To this end, let $f(k) = \lim_{t\to\infty} f_t(k)$ provided that the limit exists. In addition, the limiting distribution $(f(0), f(1), \cdots)$ is said to follow a power law if there exist a number $k_{\min}$, constant $c > 0$ and γ such that $f(k) = c(1 + o(1/k))k^{\gamma}$ for all $k \geq k_{\min}$, in which γ is referred to as the *exponent* of the power law.

In this section, we will study the rates of the convergence of $f_t(k)$, that is, $f_t(k)$ converges to a number $f(k)$ as t approaches infinity (we will establish the existence of $\lim_{t\to\infty} f_t(k)$ for $k \geq 0$ in Section 3.6. Note that the analysis in this section does not rely on the results presented in Section 3.6). To begin with, we have the following results concerning F_t^+, the expected number of non-isolated nodes in G_t.

Proposition 3.4.1. *For a partial duplication model $\mathcal{M}(G_{t_0}, p)$, the following statements hold:*

(i) *If $f(0) < 1$, setting $c = 1 - f(0)$ then we have $c > 0$ and $ct \leq F_t^+ \leq t$, that is, $F_t^+ = \Theta(t)$. In particular, $F_t^+ = t(1 - f_{t_0}(0))$ for $p = 1$.*

(ii) *If $f(0) = 1$, we have*

$$c_1 t^p(1 + O(t^{-1})) \leq F_t^+ \leq c_2 t^{2p}(1 + O(t^{-1}))$$

with $c_1 = F_{t_0}^+ \frac{\Gamma(t_0)}{\Gamma(t_0+p)}$ and $c_2 = 2D_{t_0} \frac{\Gamma(t_0+1)}{\Gamma(t_0+2p)}$.

Proof. (i) From Eq. (3.2), it is clear that $f_t(0)$ is non-decreasing. We have

$$t \geq F_t^+ = t \sum_{k\geq 1} f_t(k) = t(1 - f_t(0)) \geq t(1 - f(0)).$$

(ii) By Eq. (3.6) and Proposition 3.3.2, we have

$$F_t^+ = \sum_{k\geq 1} F_t(k) \leq \sum_{k\geq 1} k\, F_t(k) = 2tD_t = c_2 t^{2p}(1 + O(t^{-1})),$$

which establishes the upper bound.

Now we proceed to prove the lower bound. To begin with, by $0 < q = 1-p < 1$ we have

$$\sum_{k\geq 1} F_t(k)q^k \leq q \sum_{k\geq 1} F_t(k) = qF_t^+ \tag{3.10}$$

for $k \geq 1$. Since $F_t^+ = t - F_t(0)$, from Eq. (3.2) we have

$$t + 1 - F_{t+1}^+ = t - F_t^+ + \frac{1}{t}\sum_{k\geq 0} F_t(k)q^k.$$

Together with Eq. (3.10), this implies

$$\begin{aligned} F_{t+1}^+ &= F_t^+ + 1 - \frac{1}{t}\sum_{k\geq 0} F_t(k)q^k = F_t^+ + 1 - \frac{F_t(0)}{t} - \frac{1}{t}\sum_{k\geq 1} F_t(k)q^k. \\ &= F_t^+ + \frac{F_t^+}{t} - \frac{1}{t}\sum_{k\geq 1} F_t(k)q^k \\ &\geq F_t^+(1 + \frac{1}{t} - \frac{q}{t}) = F_t^+(1 + \frac{p}{t}). \end{aligned}$$

Using Eq. (3.4) in Lemma 3.3.1, we can conclude

$$F_{t+1}^+ \geq F_{t_0}^+ \frac{\Gamma(t_0)\Gamma(t+p+1)}{\Gamma(t_0+p)\Gamma(t+1)} = F_{t_0}^+ \frac{\Gamma(t_0)}{\Gamma(t_0+p)} t^p \Big(1 + O(t^{-1})\Big).$$

□

With the above proposition, we can establish the main result of this section.

Theorem 3.4.2. *Let $0 < p < 1$; then the following assertions hold:*

(i) *We have*

$$1 - c_2\, t^{2p-1}(1 + O(t^{-1})) \le f_t(0) \le 1 - c_1\, t^{p-1}(1 + O(t^{-1}))$$

with $c_1 = F_{t_0}^{+}\Gamma(t_0)/\Gamma(t_0 + p)$ and $c_2 = 2D_{t_0}\Gamma(t_0 + 1)/\Gamma(t_0 + 2p)$.

(ii) *For $k \ge 1$, we have*

$$f_t(k) \le c_3\, t^{\frac{1}{p}-p-1}(1 + O(t^{-1})),$$

with $c_3 = \Gamma(t_0 + 2)/\Gamma(t_0 + q + p^{-1})$.

Remark: Since $f_t(k) \le 1$ by definition, the upper bound in Part (ii) is non-trivial only if $1 - p - p^2 < 0$, that is, $p > (\sqrt{5} - 1)/2$. On the other hand, recall that the example in Section 3.2 shows that for $1 \le k \le t$, $f_t(k) = \frac{2(t-k)}{t(t-1)}$ for the PD model $\mathcal{M}(K_2, 1)$, and Part (ii) in the above result implies $f_t(k) \le \frac{3}{t}(1 + O(t^{-1}))$. This indicates the upper bound in Part (ii) is good when p is close to 1.

Proof. Part (i) follows directly from Proposition 3.4.1 and $f_t(0) = F_t(0)/t$.

To establish Part (ii), we first note

$$\sum_{j \ge k} \binom{j}{k} p^k q^{j-k} = \frac{1}{p} \tag{3.11}$$

because

$$\sum_{k\ge 0}\sum_{j \ge k} \binom{j}{k} p^k q^{j-k} x^k = \sum_{j\ge 0}\sum_{k=0}^{j} \binom{j}{k} (px)^k q^{j-k} = \sum_{j \ge 0}(px + q)^j = \frac{1}{p(1-x)}$$

holds for any real number x with $|x| < 1$.

Considering the constant $c := pt_0(t_0 + 1)/[pt_0 + (1 + p)q]$ and $\kappa := \frac{1}{p} - p$, then it suffices to show that

$$F_t(k) \le c\prod_{i=t_0}^{t-1}\Big(1+\frac{\kappa}{i}\Big) \tag{3.12}$$

for $k \ge 1$ and $t > t_0$, because together with Eq. (3.4) in Lemma 3.3.1, this implies

$$F_t(k) \le c\frac{\Gamma(t_0)\Gamma(t+\kappa)}{\Gamma(t_0+\kappa)\Gamma(t)} = c\frac{\Gamma(t_0)}{\Gamma(t_0+\kappa)}t^{\kappa}\Big(1+O(t^{-1})\Big) = c_3t^{\kappa}\Big(1+O(t^{-1})\Big),$$

from which the conclusion clearly follows.

In the rest of the proof, we shall establish Inequality (3.12) by induction on t. The base case $t = t_0 + 1$ is clear, because for $k \ge 1$ we have

$$c\Big(1+\frac{\kappa}{t_0}\Big) = t_0 + 1 \ge F_{t_0+1}(k).$$

For the induction step, assuming Eq. (3.12) holds for some $t > t_0$ and we shall show that it also holds for $t+1$. To this end, we can further assume $k \le t$ as otherwise we have $F_{t+1}(k) = 0$. Since $k \le t$ implies $1 - \frac{pk}{t} > 0$, substituting the induction assumption $F_t(k) \le c\prod_{i=t_0}^{t-1}(1+\frac{\kappa}{i})$ into Eq. (3.1) leads to

$$F_{t+1}(k) \le c\prod_{i=t_0}^{t-1}\Big(1+\frac{\kappa}{i}\Big)\left(1-\frac{pk}{t}+\frac{p(k-1)}{t}+\frac{1}{t}\sum_{j\ge k}\binom{j}{k}p^kq^{j-k}\right) = c\prod_{i=t_0}^{t}\Big(1+\frac{\kappa}{i}\Big),$$

where the last equality follows from Eq. (3.11). By Eq. (3.5) in Lemma 3.3.1, this completes the proof of the induction step, and hence also the theorem. □

3.5 The Non-isolated Subgraph

In this section, we study the degree distribution of G_t^+, the non-isolated subgraph obtained from G_t by removing all isolated nodes. Such subgraph is useful to model real PPI networks as isolated nodes are typically discarded in the observed networks. We start with the following technical result that will be used later.

Lemma 3.5.1. *Let $\{a_t\}_{t\geq 0}$ and $\{b_t\}_{t\geq 0}$ be two sequences of real numbers such that b_t strictly increases to infinity as $t \to \infty$, and $\lim\limits_{t\to\infty} a_t/b_t$ exists. If $\lim\limits_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t}$ also exists, then we have*

$$\lim_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t} = \lim_{t\to\infty} \frac{a_t}{b_t}.$$

Proof. Let $\beta := \lim_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t}$, where the notation := means to define; then it suffices to show $\beta = \lim_{t\to\infty} a_t/b_t$. Here we only prove the lemma for the case when $\beta \in (-\infty, \infty)$, as the cases in which $\beta = \infty$ or $\beta = -\infty$ can be established by a similar argument. Without loss of generality, we may also assume b_t is positive.

Fix an arbitrary number $\varepsilon \in (0, 1)$; by definition, there exists a number t' such that

$$\beta(1-\varepsilon)(b_{t+1}-b_t) < a_{t+1}-a_t < \beta(1+\varepsilon)(b_{t+1}-b_t)$$

holds for all $t \geq t'$. Summing up the above inequalities over t and canceling terms, we have

$$\beta(1-\varepsilon)(b_{t+1}-b_{t'}) < a_{t+1}-a_{t'} < \beta(1+\varepsilon)(b_{t+1}-b_{t'}).$$

Divide each side of the above inequalities by b_{t+1} and let $t \to \infty$; we obtain

$$\beta(1-\varepsilon) \leq \lim_{t\to\infty} \frac{a_t}{b_t} \leq \beta(1+\varepsilon),$$

which implies $\beta = \lim_{t\to\infty} a_t/b_t$ as ε is an arbitrary number in $(0, 1)$. □

Remark: Note that the condition that $\lim_{t\to\infty} \frac{a_{t+1}-a_t}{b_{t+1}-b_t}$ exists is required for the above lemma. For example, considering $a_t = t$ for t odd and $a_t = t-1$ for t even, and $b_t = t$, then $\lim_{t\to\infty} \frac{a_t}{b_t} = 1$ but $\frac{a_{t+1}-a_t}{b_{t+1}-b_t}$ is divergent.

Recall that $f_t^+(k) = F_t(k)/F_t^+$ is the expected proportion of nodes with degree k in G_t^+. For $k \geq 1$, denote $\lim_{t\to\infty} f_t^+(k)$ by $f^+(k)$ if this limit exists. For simplicity, we will also use the convention $f^+(0) = 0$. In addition, let

$$\lambda = 1 - \frac{\sum_{j\geq 1} f^+(j)q^j}{\sum_{k\geq 1} f^+(k)}. \tag{3.13}$$

Below is the main result of this section.

Theorem 3.5.2. *The following assertions hold for the partial duplication model $\mathcal{M}(G_{t_0}, p)$:*

(i) *If $f(0) < 1$, then we have $f^+(k) = \lim_{t\to\infty} f_t^+(k) = 0$ for $k \geq 1$.*

(ii) *If $f(0) = 1$ and $f^+(k) = \lim_{t\to\infty} f_t^+(k)$ exists for all $k \geq 1$, then*

(a) *we have $f^+(k) > 0$ for all $k \geq 1$;*

(b) *we have*

$$\lambda = \lim_{t\to\infty} t\,\frac{F_{t+1}^+ - F_t^+}{F_t^+} > 0,$$

and $f^+(k)$ satisfies the following equation

$$-pkf^+(k) + p(k-1)f^+(k-1) + \sum_{j\geq k} \binom{j}{k} p^k q^{j-k} f^+(j) = \lambda f^+(k)$$

for all $k \geq 1$.

Proof. (i) We know that $f(k) = 0$ if it exists [11]. Together with $F_t^+ = \Omega(t)$ from Proposition 3.4.1, this implies $f^+(k) = \lim_{t\to\infty} f_t^+(k) = 0$, as required.

(ii) For simplicity, set $b_k(j) := \binom{j}{k} p^k q^{j-k}$, and

$$\varphi_t(k) := t\frac{F_{t+1}(k) - F_t(k)}{F_t^+}$$

for $t \geq t_0$. Multiplying both sides of Eq. (3.1) by t/F_t^+ leads to

$$\varphi_t(k) = -pkf_t^+(k) + p(k-1)f_t^+(k-1) + \sum_{j\geq k} b_k(j) f_t^+(j). \tag{3.14}$$

Given that $\lim_{t\to\infty} f_t^+(k)$ exists and the limit is bounded above by 1 for each k, from $\sum_{j\geq k} b_k(j) = 1/p$ (see Eq. (3.11)) we know that $\varphi(k) := \lim_{t\to\infty} \varphi_t(k)$ exists and is finite.

Part (ii-a) follows immediately from the two claims below:

Claim 1. If there exists a number $k^* \geq 1$ such that $f^+(k^*) = 0$, then $f^+(k) = 0$ for all $k \geq 1$.

Claim 2. There exists some k with $f^+(k) > 0$.

To establish Claim 1, we consider the smallest positive number $k_0 \geq 1$ such that $f^+(k_0) = 0$. Together with Eq. (3.14), the choice of k_0 implies $\varphi(k_0) = \lim_{t\to\infty} \varphi_t(k_0) \geq 0$.

We shall show $\varphi(k_0) = 0$. If not, there exist $\varepsilon > 0$ and a number $t_1 \geq t_0$ so that for all $t \geq t_1$, we have $\varphi_t(k_0) > \varepsilon$, that is,

$$F_{t+1}(k_0) - F_t(k_0) > \varepsilon F_t^+/t. \tag{3.15}$$

On the other hand, from Eq. (3.3) we have

$$\frac{F_t^+}{t} = F_{t+1}^+ - F_t^+ + \sum_{j\geq 1} f_t(j)q^j.$$

for all $t \geq t_0$. Substituting the above equation into the inequality (3.15) leads to

$$F_{t+1}(k_0) - F_t(k_0) > \varepsilon \frac{F_t^+}{t} = \varepsilon\Big(F_{t+1}^+ - F_t^+ + \sum_{j\geq 1} f_t(j)q^j\Big) \geq \varepsilon(F_{t+1}^+ - F_t^+).$$

Summing up the above equation over t and canceling terms, we have

$$F_{t+1}(k_0) - F_{t_1}(k_0) > \varepsilon(F_{t+1}^+ - F_{t_1}^+).$$

Dividing both sides by F_{t+1}^+ and noting that Proposition 3.3.4 implies $F_{t+1}^+ \to \infty$

as $t \to \infty$, we obtain

$$f^+(k_0) = \lim_{t\to\infty} f_t^+(k_0) \geq \varepsilon > 0,$$

a contradiction to the assumption $f^+(k_0) = 0$. Hence we establish $\varphi(k_0) = 0$.

Since $f^+(k_0) = 0$ and $\varphi(k_0) = 0$, from Eq. (3.14) we have

$$p(k_0 - 1)f^+(k_0 - 1) + \sum_{j\geq k_0} b_{k_0}(j)f^+(j) = 0.$$

Noting that $p > 0$, we must have $f^+(k_0 - 1) = 0$. Because k_0 is the smallest positive number such that $k_0 \geq 1$ and $f^+(k_0) = 0$, we know $k_0 = 1$ and hence $f^+(k) = 0$ for all $k \geq 1$. This completes the proof of Claim 1.

We proceed to establish Claim 2. For later use, we fix $\delta > 0$ so that $2p + \delta < 1$, and $k' \geq 1$ so that $1 - q^{k'} > 2p + \delta$. To obtain a contradiction, we assume $f^+(k) = 0$ for all $k \geq 1$. This implies

$$\lim_{t\to\infty} \sum_{k=1}^{k'} \frac{F_t(k)}{F_t^+} = \sum_{k=1}^{k'} \lim_{t\to\infty} \frac{F_t(k)}{F_t^+} = \sum_{k=1}^{k'} f^+(k) = 0.$$

Therefore, by setting $\delta_t := \sum_{k>k'} F_t(k)/F_t^+$, we have $\delta_t \to 1$ as $t \to \infty$. From Eq. (3.3) and $0 < q = 1 - p < 1$ we have

$$\begin{aligned} F_{t+1}^+ &= F_t^+ + \frac{1}{t}\sum_{k=1}^{k'} F_t(k)(1 - q^k) + \frac{1}{t}\sum_{k>k'} F_t(k)(1 - q^k) \\ &\geq F_t^+ + \frac{1}{t}\sum_{k=1}^{k'} F_t(k)(1 - q^k) + \frac{1 - q^{k'}}{t}\sum_{k>k'} F_t(k). \end{aligned}$$

Taking out the factor F_t^+, this implies

$$F_{t+1}^+ \geq \Big(1 + \frac{1 - q^{k'}}{t}\delta_t\Big)F_t^+ \geq F_{t_0}^+ \prod_{s=t_0}^{t} \Big(1 + \frac{1 - q^{k'}}{s}\delta_s\Big).$$

Since $\delta_t \to 1$, there exists a number t' with $\delta_s > 1-\delta$ for all $s > t'$. Therefore, for $t > t'$ we have

$$F^+_{t+1} \geq F^+_{t_0} \prod_{s=t_0}^{t'} \Big(1+\frac{1-q^{k'}}{s}\delta_s\Big) \prod_{s=t'+1}^{t} \Big(1+\frac{1-q^{k'}}{s}(1-\delta)\Big) = \Omega(t^{(1-q^{k'})(1-\delta)}), \quad (3.16)$$

where the equality follows from Eq. (3.5) in Lemma 3.3.1. This contradicts $F^+_t = O(t^{2p})$ as $(1-q^{k'})(1-\delta) > 1-q^{k'}-\delta > 2p$, which completes the proof of Claim 2, and hence also Part (ii-a).

We proceed to establish Part (ii-b). To begin with, recall that by Lemma 3.3.3, for each $k \geq 1$ there exists $\tau_k \geq t_0$ so that $F_t(k) > 0$ for all $t \geq \tau_k$. Therefore, for $k \geq 1$ we can define

$$\psi_t(k) := t\frac{F_{t+1}(k) - F_t(k)}{F_t(k)}$$

for $t \geq \tau_k$. In other words, we have $\varphi_t(k) = f^+_t(k)\psi_t(k)$ for $t \geq \tau_k$. Since $\varphi(k) = \lim_{t\to\infty}\varphi_t(k)$ exists and is finite, $f^+(k) > 0$ from Part (i) implies that $\psi(k) := \lim_{t\to\infty}\psi_t(k)$ exists and is finite for $k \geq 1$.

We first show $\psi(k) \neq 0$ by contradiction. If this is not the case, there exist $0 < r < p$ and a number t_1 so that $\psi_t(k) < r$ for all $t \geq t_1$. By Eq. (3.5) in Lemma 3.3.1, this implies

$$F_{t+1}(k) < \Big(1+\frac{r}{t}\Big)F_t(k) < \prod_{s=t_1}^{t}\Big(1+\frac{r}{s}\Big)F_s(k) = O(t^r).$$

and hence $F_t(k) = O(t^r)$. On the other hand, we have $F_t(k) = f^+(k)F^+_t(1+o(1))$, where $f^+(k) > 0$, and $F^+_t = \Omega(t^p)$ by Proposition 3.4.1. This implies $F_t(k) = \Omega(t^p)$, contradicting $F_t(k) = O(t^r)$ as $r < p$. Thus we must have $\psi(k) \neq 0$.

We next show $\psi(k) = \psi(1)$ for all $k > 1$. Indeed, we have

$$\frac{\psi(k)}{\psi(1)} = \lim_{t\to\infty}\frac{F_t(1)}{F_t(k)}\frac{F_{t+1}(k)-F_t(k)}{F_{t+1}(1)-F_t(1)} = \frac{f^+(1)}{f^+(k)}\lim_{t\to\infty}\frac{F_{t+1}(k)-F_t(k)}{F_{t+1}(1)-F_t(1)},$$

and hence $\lim_{t\to\infty}\frac{F_{t+1}(k)-F_t(k)}{F_{t+1}(1)-F_t(1)}$ must exist. Using Proposition 3.3.4 and Lemma 3.5.1, we have

$$\lim_{t\to\infty}\frac{F_{t+1}(k)-F_t(k)}{F_{t+1}(1)-F_t(1)}=\lim_{t\to\infty}\frac{F_t(k)}{F_t(1)}=\frac{f^+(k)}{f^+(1)}$$

and hence $\psi(k)=\psi(1)$.

Since $\psi(k)$ is a constant for all $k\geq 1$, we denote it by ψ. By Proposition 3.3.4, we have $\psi(1)>0$ and hence also $\psi>0$. Now from Eq. (3.14) we know that $\{f^+(k)\}_{k\geq 1}$ satisfies the following equation

$$-pkf^+(k)+p(k-1)f^+(k-1)+\sum_{j\geq k}b_k(j)f^+(j)=\psi f^+(k),$$

where $f^+(0)=0$. By summing the above equation for all $k\geq 1$ and canceling the first two summations of the left-hand side, we obtain

$$\psi\sum_{k\geq 1}f^+(k)=\sum_{k\geq 1}\sum_{j\geq k}b_k(j)f^+(j)=\sum_{j\geq 1}\sum_{j\geq k\geq 1}b_k(j)f^+(j)=\sum_{j\geq 1}f^+(j)(1-q^j),$$

where the last equality follows from $\sum_{j\geq k\geq 1}b_k(j)=1-q^j$. Hence, we have

$$\psi=1-\frac{\sum_{j\geq 1}f^+(j)q^j}{\sum_{k\geq 1}f^+(k)}=\lambda,$$

which completes the proof of Part (ii). □

Remark: Intuitively the frequencies for degree $k\geq 1$ should be larger when p is larger. However Thm. 3.5.2 concerns the non-isolated components and the divisor of $f_t^+(k)$ is F_t^+. As p increases, F_t^+ increases at a larger rate than $F_t(k)$.

The above result shows that the limiting degree distribution $(f^+(1),f^+(2),\cdots)$ does not follow a power law when $f(0)<1$ because $f^+(k)=0$ for all $k\geq 1$. On the other hand, Fig. 3.1 indicates that a power law may exist for small p. Indeed,

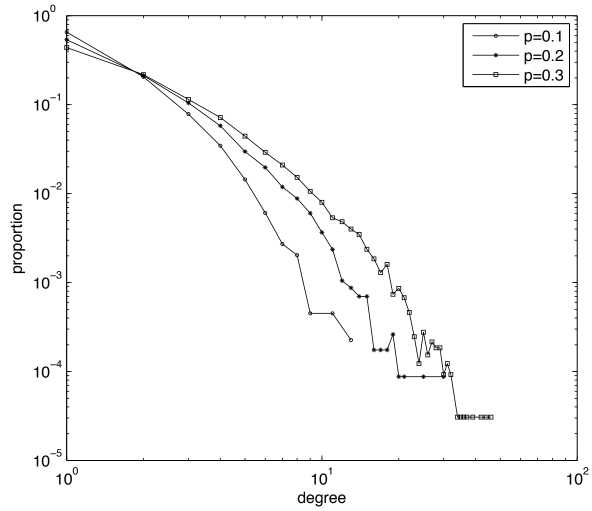

Figure 3.1: Log-Log plot of degree distribution of the PDM model $\mathcal{M}(K_2, p)$ with $p \in \{0.1, 0.2, 0.3\}$. For each p, 1000 graphs with 1000 nodes were generated using the PDM model, and the average degree distribution is depicted.

let γ be the solution of the equation

$$\gamma p - p + p^{\gamma-1} = \lambda, \tag{3.17}$$

where λ is the constant defined in Eq. (3.13). The following result states that when $f(0) = 1$, if the limiting degree distribution $(f^+(1), f^+(2), \cdots)$ follows a power law, then the exponent γ is defined above.

Corollary 3.5.3. *Suppose $f(0) = 1$. If $f^+(k) = \lim_{t\to\infty} f_t^+(k)$ exists for all $k \geq 1$, the power law whose exponent γ is given in Eq. (3.17) is a solution to the limiting degree distribution $(f^+(1), f^+(2), \cdots)$.*

Proof. The proof is similar to that of Theorem 1 in [18]. From Theorem 3.5.2, $f^+(k)$ satisfies the following recursion

$$-pkf^+(k) + p(k-1)f^+(k-1) + \sum_{j\geq k} b_k(j)f^+(j) = \lambda f^+(k),$$

for $k \geq 1$, where $b_k(j) = \binom{j}{k} p^k q^{j-k}$. Substituting $f^+(k) = c(1+o(1/k))k^{-\beta}$, where $c > 0$ is a positive constant and β is to be determined later, and multiplying both sides by k^β, we obtain

$$\underbrace{-pk\Big(1+o\Big(\frac{1}{k}\Big)\Big))}_{\text{I}} + \underbrace{p(k-1)\Big(1+o\Big(\frac{1}{k-1}\Big)\Big)\Big(\frac{k}{k-1}\Big)^\beta}_{\text{II}} + \underbrace{\sum_{j\geq k} b_k(j)\Big(1+o\Big(\frac{1}{j}\Big)\Big)\Big(\frac{k}{j}\Big)^\beta}_{\text{III}}$$
$$= \lambda\Big(1+o\Big(\frac{1}{k}\Big)\Big).$$

Since $(\frac{k}{k-1})^\beta = 1 + \frac{\beta}{k} + O(\frac{1}{k^2})$, we have

$$\begin{aligned} \text{II} &= p(k-1)\Big(1+o\Big(\frac{1}{k-1}\Big)\Big)\Big(1+\frac{\beta}{k}+O\Big(\frac{1}{k^2}\Big)\Big) \\ &= p(k-1)\Big(1+o\Big(\frac{1}{k-1}\Big)\Big)\Big(1+\frac{\beta}{k}\Big)+O\Big(\frac{1}{k}\Big). \end{aligned}$$

Because $p(k-1)(1+o(\frac{1}{k-1}))(1+\frac{\beta}{k}) = p(k-1)(1+\frac{\beta}{k})+o(1)$ and $p(k-1)(1+\frac{\beta}{k}) = pk(1+\frac{\beta}{k}) - p + o(1)$, we have $\text{II} = pk + p\beta - p + o(1)$ and hence

$$\text{I} + \text{II} = p\beta - p + o(1).$$

From Lemma 1 in [18], we have

$$\binom{j}{j-k}\Big(\frac{k}{j}\Big)^\beta = \Big(1+O\Big(\frac{1}{k}\Big)\Big)\binom{j-\beta}{j-k}.$$

Applying this formula to III leads to

$$\begin{aligned} \text{III} &= \sum_{j\geq k} p^k q^{j-k}\Big(1+o\Big(\frac{1}{j}\Big)\Big)\binom{j-\beta}{j-k}\Big(1+O\Big(\frac{1}{k}\Big)\Big) \\ &= \Big(1+O\Big(\frac{1}{k}\Big)\Big)\sum_{j\geq k} p^k q^{j-k}\binom{j-\beta}{j-k}. \end{aligned}$$

Let $l = j - \beta$. The above equation can be simplified as

$$\begin{aligned} \mathrm{III} &= \Big(1 + O\Big(\frac{1}{k}\Big)\Big) \sum_{l \geq k-\beta} p^k q^{l+\beta-k} \binom{l}{k-\beta} \\ &= p^{\beta}\Big(1 + O\Big(\frac{1}{k}\Big)\Big) \sum_{l \geq k-\beta} \binom{l}{k-\beta} p^{k-\beta} q^{l-(k-\beta)} \\ &= p^{\beta-1}\Big(1 + O\Big(\frac{1}{k}\Big)\Big). \end{aligned}$$

Therefore, we see that β satisfies the equation $\beta p - p + p^{\beta-1} = \lambda$ as $k \to \infty$, which completes the proof. □

As an application, we applied the above results to three real PPI networks, *S. cerevisiae* (budding yeast), *D. melanogaster* (fruitfly) and *C. elegans* (worm), downloaded in August 2012 from DIP (http://dip.doe-mbi.ucla.edu/dip/Main.cgi). Since a protein is collected in the database only if it is involved in an observed interaction, these networks can be better modeled by the non-isolated graph generated by the PD model. Corollary 3.5.3 states that the long-run degree distribution of the non-isolated graphs may follow a power law distribution. Indeed, we estimated the power law exponent γ for each network using linear regression. In addition, by Eq. (3.17) we inferred the selection probability p for each network using the degree distribution and estimated γ. The results are presented in Table 3.1.

	C. elegans	*S.cerevisiae*	*D. melanogaster*
γ	1.6	1.7	2.0
p	0.01	0.4	0.3

Table 3.1: Estimated power law exponent γ and selection probability p for three PPI networks. The networks were downloaded from DIP. For each network, the degree distribution was computed, from which the power law exponent and selection probability were estimated.

3.6 Limiting Behavior of Degree Distribution

In this section, we shall establish the existence of the limiting degree distribution for the PD model. We first recall the following results by [11]: The sequence $\{f_t(0)\}_{t \geq t_0}$ is non-decreasing, and hence $\lim_{t \to \infty} f_t(0)$ exists with $f(0) \leq 1$. In addition, for each $k \geq 1$, if $\lim_{t \to \infty} f_t(k)$ exists, then $f(k) = \lim_{t \to \infty} f_t(k) = 0$. One consequence of their results is that the limiting degree distribution cannot follow a power law. However, one important problem remained unsettled is whether $\lim_{t \to \infty} f_t(k)$ exists for $k \geq 1$, which is the subject of the following theorem, where we also show that a phase transition exists for the expected proportion of isolated nodes converging to 1 (see Fig. 3.2 for some numeric results), and hence give some hint to a question raised in [11].

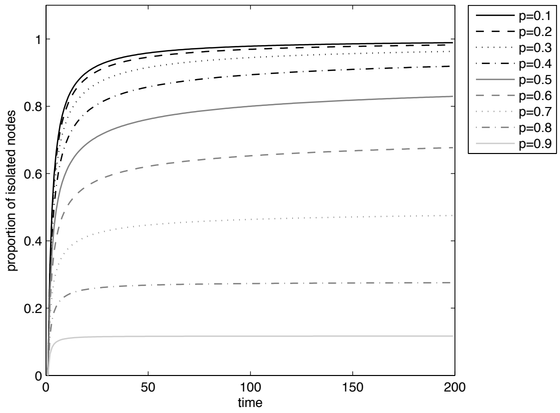

Figure 3.2: Expected proportion of isolated nodes in the PDM model $\mathcal{M}(K_2, p)$. Here the result is obtained by numerically solving Eq. (3.1) with boundary condition $F_2(1) = 2$ and $F_2(k) = 0$ for $k \neq 2$.

Theorem 3.6.1. *The following assertions hold for the partial duplication model $\mathcal{M}(G_{t_0}, p)$:*

(i) *For* $k \geq 1$, $\lim_{t \to \infty} f_t(k) = 0$.

(ii) *For* $0 < p < 1/2$, *we have* $f(0)$=*1; for* $1/\sqrt{2} < p < 1$ *we have* $f(0) < 1$.

Proof. (i) When $p = 1$, $\mathcal{M}(G_{t_0}, p)$ is reduced to the full duplication model, and the statement is well known to hold (see, for example, [79]). So we assume $p < 1$, and hence also $q = 1 - p > 0$. For each $k \geq 1$, we introduce a function c_k on non-negative intergers defined as

$$c_k(j) = \begin{cases} \binom{j}{k} p^k q^{j-k} & j > k \\ p^k - pk & j = k \\ p(k-1) & j = k-1 \\ 0 & 0 \leq j < k-1 \end{cases} .$$

Note that we have $c_k(j) \geq 0$ for all $j \neq k$, and $c_k(0) = 0$ for all $k \geq 1$. Now Eq. (3.1) can be rewritten as

$$F_{s+1}(k) = F_s(k) + \sum_{j \geq 0} c_k(j) \frac{F_s(j)}{s},$$

where $s \geq t_0$. By definition, this is equivalent to

$$(s+1) f_{s+1}(k) = s\, f_s(k) + \sum_{j \geq 0} c_k(j) f_s(j).$$

By summing the above equation over s up to t, we get

$$\sum_{s=t_0}^{t} (s+1) f_{s+1}(k) = \sum_{s=t_0}^{t} s f_s(k) + \sum_{j \geq 0} \Big(c_k(j) \sum_{s=t_0}^{t} f_s(j) \Big).$$

Therefore, canceling terms and dividing by $t+1$ on both sides leads to

$$f_{t+1}(k) = \frac{t_0 f_{t_0}(k)}{t+1} + \sum_{j \geq 0} \Big(c_k(j) \sum_{s=t_0}^{t} \frac{f_s(j)}{t+1} \Big). \tag{3.18}$$

On the other hand, by Eq. (3.2) we have

$$(s+1)\Big(f_{s+1}(0) - f_s(0)\Big) = \sum_{j\geq 1} f_s(j)q^j$$

for $s \geq t_0$. Summing the above equation over s up to t leads to

$$(t+1)f_{t+1}(0) - \sum_{s=t_0}^{t} f_s(0) - t_0 f_{t_0}(0) = \sum_{j\geq 1}\sum_{s=t_0}^{t} f_s(j)q^j.$$

Since $q > 0$, for any $j \geq 1$ the above equation implies

$$\frac{\sum_{s=t_0}^{t} f_s(j)}{t+1} \leq q^{-j}\Big(f_{t+1}(0) - \frac{\sum_{s=t_0}^{t} f_s(0)}{t+1} - t_0\frac{f_{t_0}(0)}{t+1}\Big).$$

Since $\lim_{t\to\infty} f_t(0)$ always exists and is necessarily finite, the right-hand side of the above inequality converges to 0. In other words, for each $j \geq 1$ we have

$$\lim_{t\to\infty} \frac{\sum_{s=t_0}^{t} f_s(j)}{t+1} = 0. \tag{3.19}$$

Therefore, from Eq. (3.18) and (3.19) we have

$$\lim_{t\to\infty} f_t(k) = \lim_{t\to\infty} \sum_{j\neq 0} c_k(j) \sum_{s=t_0}^{t} \frac{f_s(j)}{t+1} = \sum_{j\neq 0} c_k(j) \lim_{t\to\infty} \sum_{s=t_0}^{t} \frac{f_s(j)}{t+1} = 0$$

for each $k \geq 1$. The interchange of the summation and the limit follows from the dominated convergence theorem: $\sum_{s=t_0}^{t} \frac{f_s(j)}{t+1} \leq 1$ and $\sum_{j\neq 0} c_k(j) < \infty$. This completes the proof of Part (i).

(ii) Since Proposition 3.3.2 implies that D_t, the expected average degree in G_t, is asymptotically $ct^{2p-1} + o(1)$ for some constant c, we have

$$f_t(0) = 1 - \sum_{k\geq 1} f_t(k) \geq 1 - \sum_{k\geq 1} k\, f_t(k) = 1 - D_t = 1 - 2ct^{2p-1} + o(1)$$

for $0 < p < \frac{1}{2}$. This implies $f(0) = \lim_{t\to\infty} f_t(0) = 1$ for $0 < p < \frac{1}{2}$, as required.

Recall that we have $f^+(k) > 0$ if $F_t^+ = o(t)$ by Thm. 3.5.2. Together with Thm. 3.4.2, we have

$$0 < f_t^+(k) \leq \frac{t^{1/p-p}}{F_t^+}.$$

If $p > \frac{\sqrt{5}-1}{2}$, which means $1/p - p < 1$, F_t^+ should not have a higher order than $t^{1/p-p}$, i.e. $F_t^+ = O(t^{1/p-p})$. On the other hand, we have $F_t^+ = \Omega(t^p)$. It follows that $p \leq 1/p - p$, i.e. $p \leq 1/\sqrt{2}$. In another word, if $p > 1/\sqrt{2}$, then $F_t^+ = \Theta(t)$ and $f(0) < 1$.

□

Motivated by [11], a model $\mathcal{M}(G_{t_0}, p)$ is called *defective* if $\sum_{k\geq 0} f(k) < 1$. For instance, the PD model $\mathcal{M}(K_2, 1)$, the example studied in Section 3.2, is defective. Note that defective model is usually identified with the existence of a giant component, and it is observed in [11] that $\mathcal{M}(G_{t_0}, p)$ is defective for $p = 1$, and not defective for $p < 1/2$, and the problem remained open is at what value of p the model becomes defective. The following result provides an possible interval for the phase transition point of $\mathcal{M}(G_{t_0}, p)$.

Corollary 3.6.2. *There is a phase transition point* $p_0 \in [1/2, 1/\sqrt{2}]$ *for the partial duplication model* $\mathcal{M}(G_{t_0}, p)$.

Proof. We first show that $f(0)$ is a decreasing function of p. Suppose $p_1 < p_2$, $F_t(0)$ and $\tilde{F}_t(0)$ correspond to the number of singletons in $\mathcal{M}(G_{t_0}, p_1)$ and $\mathcal{M}(G_{t_0}, p_2)$ respectively. Recall that $F_{t+1}(0) = F_t(0) + \frac{1}{t}\sum_{k\geq 0} F_t(k)q^k$. Obviously $F_{t_0+1}(0) \geq$

$\tilde{F}_{t_0+1}(0)$. Suppose $F_t(0) \geq \tilde{F}_t(0)$ for t. The difference at time $t+1$ is

$$\begin{aligned}
&F_{t+1}(0) - \tilde{F}_{t+1}(0) \\
=&(1+\frac{1}{t})(F_t(0) - \tilde{F}_t(0)) + \frac{1}{t}\sum_{k\geq 1}(F_t(k) - \tilde{F}_t(k))q^k \\
=&(1+\frac{1}{t})\sum_{k\geq 1}(\tilde{F}_t(k) - F_t(k)) + \frac{1}{t}\sum_{k\geq 1}(F_t(k) - \tilde{F}_t(k))q^k \\
=&\sum_{k\geq 1}(\tilde{F}_t(k) - F_t(k))(1+\frac{1}{t} - \frac{q^k}{t}) \\
\geq&\sum_{k\geq 1}(\tilde{F}_t(k) - F_t(k)) \\
=&\Big(t - \tilde{F}_t(0)\Big) - \Big(t - F_t(0)\Big) \geq 0.
\end{aligned}$$

Hence $F_t(0)$ is a decreasing function of p and so is $f_t(0)$. Taking limits we have $f(0)$ is a decreasing function of p. Together with Thm. 3.6.1 we have the results as claimed. □

3.7 Discussion

This chapter presents a rigorous analysis on the degree distribution of the partial duplication (PD) model, as a step toward understanding the long run behavior of more sophisticated network growth models in the duplication and divergence family that have been developed to model protein-protein networks and other biological networks.

Although the main focus in this chapter is the mathematical properties of the PD model, the results obtained here are biologically relevant. For example, Theorem 3.6.1 shows that in terms of degree distribution, a popular summary statistic used in describing biological networks, the network generated under the PD model stabilizes at the ensemble level as it grows. In other words, when the network is sufficiently large, adding new vertices will not change the overall degree distribution of the network.

Our results also clarify the existence or the lack of power-law degree distributions under the PD model. [11] proved that degree distribution of the networks generated under the PD model does not follow a power-law distribution. This corrects a claim in [18] and leads to a further question: whether the subgraph consisting of all non-isolated nodes in the PD model follows a power-law distribution? Theorem 3.5.2 and Corollary 3.5.3 show that, for this subgraph, a power-law distribution possibly exist only when the graph is defective.

In addition, our results provide further insights into the simulation study of biological networks. For instance, in applying the PD model to simulate biological networks, one wants to know which feasible values of the parameter p will generate reasonably realistic networks. Theorem 3.6.1 shows that one should restrict the choice of p to be in $(1/\sqrt{2}, 1)$ if the expected network contains a relatively small proportion of singletons. On the other hand, to generate biological networks with a power-law distribution, Theorem 3.5.2 and Corollary 3.5.3 indicate that one should choose p in $(0, 1/2)$ and consider the subgraph consisting of all non-isolated nodes. Finally, Theorem 3.4.2 on convergence rate can be used to determine the bounds on the size of the simulated networks when the expected degree distribution is known.

It is also worthy to note that the results obtain in this chapter show that many features related to the long-run degree distribution, such as the existence of limiting distribution and the phase transition point for the expected proportion of isolated nodes converging to 1, are dependent on the selection probability p and independent of the seed graph. This agrees well with the observation made in [43] through simulation: The degree distribution of large-scale networks generated by many duplication models is solely determined by the model parameters, and not by the initial 'seed' graph.

Several problems remain open from this study. The first one is about the rates of convergence. Some rates are established in Theorem 3.4.2, but they may not be best possible. In addition, we have shown that the expected fraction of nodes with degree one grows as $\Omega(\ln t/t)$, and it remains to see whether similar bounds hold for nodes with higher degree. The second one concerns the limiting behavior of the non-isolated subgraph in the region $f(0) = 1$. In particular, a proof of the existence (or lack) of the limiting degree distribution in this region is required. Although a range is given, the exact value of the phase transition point has not been obtained in our study.

Many extensions of the PD model have been proposed in the literature. A natural extension is to allow connecting the anchor node and new node in each step of the PD model with a probability p_c. When $p_c = 0$, this extension reduces back to the PD model studied in this chapter. The special case when $p_c = 1$ was studied by Chapter 4 of [17] and it was shown that the limiting degree distribution in this case exists. Some further analysis of this extended model for general p_c was conducted by [52], and the tools developed in this study could be applied to study this model, as well as several others, such as the duplication-mutation with complementarity (DMC) model studied by [62], and the model proposed by [73].

The PD model has been studied at the ensemble level in this chapter, that is, the average behavior over many different realizations is considered. However, [50] presented an example to show that the behavior of a single realization of the PD model could be very different from the average one. As pointed out by [40], a statement about convergence of the expected proportions does not imply a similar statement about the proportions in a single realization. Therefore, one interesting direction for future research is to see whether the results obtained in this chapter are also valid at the level of individual realizations. For instance, we have shown there is a phase transition point for several properties of the PD model at the ensemble level, and it remains to see whether this is also the case for some properties at the individual level, such as the emergence of giant components.

Chapter 4

Effect of Seed Graphs on The Evolution of Network Topology

4.1 Introduction

The structure of PPI networks has been extensively studied [9, 106]. Properties, such as power-law [1], high clustering coefficient [108] and modularity [36] etc., are observed in PPI networks (reviewed in [9]). On the other hand, evolutionary mechanisms shaping the topology of networks have been proposed [13, 98, 100], which aim to explain the emergence of some topological features of PPI networks [18, 43]. Based on the evolutionary mechanisms several graph models for PPI networks are developed [95, 100], such as the duplication models and hierarchical networks. The validity of a graph model is usually affirmed by comparing the topology of the networks generated by the graph model with that of the empirical networks. The more topological features they share, the more similar they are. For example, 5 graph models were compared with the yeast PPI network in terms of 7 topological measures in [34] leading to a conclusion that the iSite model, which was proposed by the authors, gives the best fit. However, since a PPI network is only a snapshot of the network history, this strategy of validating PPI network models is limited in the context of evolution, in which the topology of a PPI network may also be evolving. Exploring how the topology of a PPI network changes with time can shed further light on the formation of the extant PPI networks and understanding the evolution of PPI networks.

A potential factor that may have significant impact on the formation of the topology of an observed network is the network it started with, called seed network or graph. In [43], the effect of seed graphs on shaping the topology of networks generated by the preferential attachment (PA) model and the duplication and divergent (DD) model was studied. Hormozdiari et al. [43] demonstrated that different seed graphs may lead to different topology in the observed network. The study of the effect of seed graphs on the topology of networks can guide us in selecting seed graphs to generate networks for modeling real networks. In [86] the choice of seed graphs by Hormozdiari et al. [43] was applied to produce families of PPI networks in a network synthetic model, i.e. a model of selecting proper models for input networks. Intuitively, seed graphs affect not only the topology of the extant networks but also the evolutionary processes. Therefore, we are interested to ask "How do networks evolve from different seed graphs?" In other words, we are not only interested in the final resulting network but the whole process in which the network evolves.

The models we shall investigate are the partial duplication (PD) model, the duplication and divergent (DD) model, the duplication-mutation with complementarity (DMC) model and the preferential attachment (PA) model (see Section 4.2 for definitions). Analogous to comparing networks, the evolutionary processes can be studied in terms of the network characteristics such as degree distribution and clustering coefficient. In our study of network history reconstruction [57] we found that clustering coefficient generally decreases as time increases. Similar observations were made in [66]. Here we explore the conditions under which such a pattern would exist. Other topological features will also be investigated, see Section 4.3 for further details.

4.2 Network Models and Parameters

Scale-free property is widely observed in many empirical networks, such as the yeast PPI network, world wide web and citation networks (reviewed in [18]). The four graph models investigated in this chapter are all aimed to capture this property. Besides, many real networks are under a process of growth. In another word, the number of nodes and edges in the networks increase with time. We have defined the definition of network growth model in Subsection 1.3.2 and we will briefly recall the terminology below. In a network growth model, the model starts with a seed network. At every time step, a new node is added into the existing network and with some probability the topology of the network may be rewired according to some rules, which are defined by the model. We have also introduced the definitions of the PD model, the DD model, the DMC model and the PA model in Subsection 1.3.2. For convenience, we denote the selection probability in the PD model by p_{PD} , that in the DD model by p_{DD} and that in the DMC model by p_{DMC}. Note that there is one more parameter for each of the DD model and the DMC model, namely the divergence rate r for the DD model and the homodimerization rate p_c for the DMC model. The PD model, the DD model and the DMC model all belong to the class of duplication models, a biologically relevant class of network models [12, 18, 44, 90, 93, 95], which are based on the duplication step. The PA model is based on another mechanism: The preferential attachment, in which the new node v connects to each existing node, say u, with a probability proportional to its degree:

$$P(e_{u,v} = 1) = \min\{c\frac{\deg(u)}{2e}, 1\},$$

where e is the number of edges and c is a parameter of the model.

With different choice of the parameters, a model may generate networks with different topology. How the parameters of the models should be chosen is still not settled [34]. Parameters can be either estimated by fitting the topology of an empirical network [43, 95] or calculated in the aspect of evolutionary studies [34, 57]. In [57] we inferred the parameters of the duplication and mutation with complementarity model (DMC) in the process of reconstructing the evolutionary history of networks. For the DMC model, we chose the same parameters as those estimated for the yeast PPI network in Chapter 2: $p_{\text{DMC}} = 0.061$ and $p_c = 0.053$ (see Table 2.3). We set $p_{\text{PD}} = p_{\text{DMC}} = 0.061$ in the PD model. For the DD model, we applied the same parameters used in [43], i.e. $r = 0.12$ and $p_{\text{DD}} = 0.365$, which is also used by Rito et al. in [82] to construct gene duplication network in investigating the relation between protein age and their degree. Recall that in the PA model given a node u, the probability that it is connected to the new node is $c\frac{\deg(u)}{2e}$. Let I_u to be the indicator function of the edge between node u and the new node and X to be the degree of the new node. We have $X = \sum_{u \in V_{t-1}} I_u$. Hence the expected number of edges the new node can get is

$$\mathbb{E}(X) = \sum_{u \in V_{t-1}} \mathbb{E}(I_u) = \sum_{u \in V} c\frac{\deg(u)}{2e} = c.$$

In [73], it is reported that $c = 1.83$ is the average degree of a yeast PPI network. In our experiments, we chose the same c.

4.3 Topological Statistics

Networks are characterized by some commonly used statistics. We have introduced some topological statistics in Subsection 1.1.1. Here we give a review on three commonly used quantities that are used in our experiments. The connectivity of a network is usually measured by clustering coefficient, which can be defined as follows. Given a node v, let $T(v)$ be the number of triangles that v is involved in as a vertex of a triangle. Then the clustering coefficient $C(v)$ of v is calculated as $c(v) = 2 * T(v)/(\deg(v) * (\deg(v) - 1))$. The clustering coefficient is usually applied to estimate the existence of an inherent modularity. Recall that $c(k)$ is the average clustering coefficient of nodes with degree k. In hierarchical networks the clustering coefficient as a function of degree follows a power-law: $c(k) \propto k^{-1}$. It is shown that in the PA model for all k, $c(k)$ is fixed [9]. To the best of our knowledge, no theoretical results about the clustering coefficient is known for the DD and PD models. Another frequently studied feature of a network is the degree distribution. Given a non-negative integer k, $P(\deg(v) = k)$ is the proportion of nodes with degree k. It is shown in [56] that the PD model produces networks with trivial limiting degree distribution for p_{PD} less than 0.5, i.e. the fraction of nodes with positive degree asymptotically approaches to 0. The limiting degree distribution of the PA model follows a power-law: $P(k) \propto k^{-3}$ [8]. For the DD model, a power-law degree distribution is also demonstrated in [11], where the power-law exponent is associated with the duplication parameter p_{DD}. The average degree is defined as $D = \sum_k kP(\deg(v) = k)$, a quantity we experimented. As discussed above, the average degree of the PA model has an expectation of c [8]. The average degree of the PD model converges to 0 as the order of the network is large when p_{PD} is smaller than 0.5 [56]. In [42], the author showed that the expected number of edges in the DD model satisfies the recursive relation: $e(t+1) = e(t)(1+2p/t)+r$. Solving a corresponding ODE $e' = \frac{2p}{t}e + r$ we have $e(t) = at^{2p} + \frac{r}{1-2p}t$, where a

is a constant dependent on the initial condition and the average degree $D(t) = 2e(t)/t = 2at^{2p-1} + \frac{2r}{1-2p}$, converging to $\frac{2r}{1-2p}$ as $t \to \infty$ for $p < 0.5$. Note that all experiments were run for connected components, so the expected average degree should be not smaller than that in the whole graphs. Another commonly observed property in empirical networks is the small-world property, i.e. a network with small diameter. The average length of the shortest paths for the PA model $l \propto \frac{\ln t}{\ln \ln t}$ as time is sufficiently large [2]. For the DD model and the PD model, to the best of our knowledge, there is no analytical results for the average length of shortest paths.

4.4 Experiments and Results

Since all nodes in a PPI network has degree of at least 1, we only considered connected components in our experiments. Specifically, at the end of each time step, we remove the singletons if there are any. For each seed graph, every model was run until the order of the network, i.e. the number of nodes, reached 1000. We have run experiments on the topological statistics described above to explore the effect of seed graphs on the network topology. For each feature, we selected 9 seed graphs which were classified into three groups. The three seed graphs in each group have different topology but the same feature that is under investigation. By such choice of seed graphs, we can test whether the initial value of the feature affects the growing behavior of the network. If it does, then we can further look into that under the same setting of the feature, whether the topology of the network has an impact on the network evolution or not. The topological statistics of seed graphs are summarized in Table 4.1.

Figure 4.1 depicts how the clustering coefficient varies with time for 9 different seed graphs. At each time step, clustering coefficient of every network was calculated. The plots are based on the average over 100 runs. The seed graph used in each plot is included in the title. Notice that the PD model will not generate any triangles if there is no triangle in the seed graph. Hence in the first row of Fig. 4.1, where the seed graphs have a clustering coefficient of 0, the clustering coefficient for the PD model is always 0. It can be observed that even if they start with the same seed graphs the DD model, the PA model and the PD model may generate networks with different clustering coefficients. This may suggest that the initial clustering coefficient may be a determining factor for its growing curve.

Figure 4.2 plots how the average degree changes with time. The first row of the seed graphs have average degree of 2, the second have average degree of 3 and the third have average degree of 4. We can see that for all the four models the average degree tends to a limit as the number of nodes gets larger and larger. All the observed average degrees are larger than the theoretical ones for the whole networks. For the DD model, an expected average degree of $\frac{2r}{1-2p} \approx 0.92$ was obtained above and is smaller than 4.7 which is the observed average degree of the connected components with order 1000. The average degrees of the PD model and the DMC model are very close. This suggests that under our choices of parameters, the selection probability p plays a major role in shaping the average degree of networks and the homodimerization rate p_c only has a minor effect. The expected average degree of the PA model is also larger than the theoretical one for the whole graph, which is $c = 1.83$.

Figure 4.3 plots the average length of shortest paths (ALSP) at each time point from the initial time to time point 1000. Seed graphs in the first row have ALSP of 1, the second have ALSP of 1.5 and the ALSP in the third row is 5/3. We can see that all the four models generate networks with ALSP no more than 10, which implies the small-world property of these networks. For all the three models, the ALSP increases as the networks expand. For the same model, all the curves have no significant difference when t is large. This indicates that the ALSP may be an inherent property of the model and its parameters.

Figures 4.4, 4.5, 4.6 and 4.7 describe the degree distribution at 5 different time points: t_0, $t_0 + 2$, $t_0 + 5$, $t_0 + 10$ and $t_0 + 900$. All nodes in the 4 seed graphs in the first row have degree 2. In the second row, 2/3 nodes have degree 2 and 1/3 nodes have degree 3. In the third row, 2/5 nodes have degree 1, 2/5 nodes have degree 2 and 1/5 have degree 4. We can see that the initial degree distribution determines the plots of the degree distribution, while the topology of the seed graph does not affect the degree distribution a lot when the initial degree distribution is fixed.

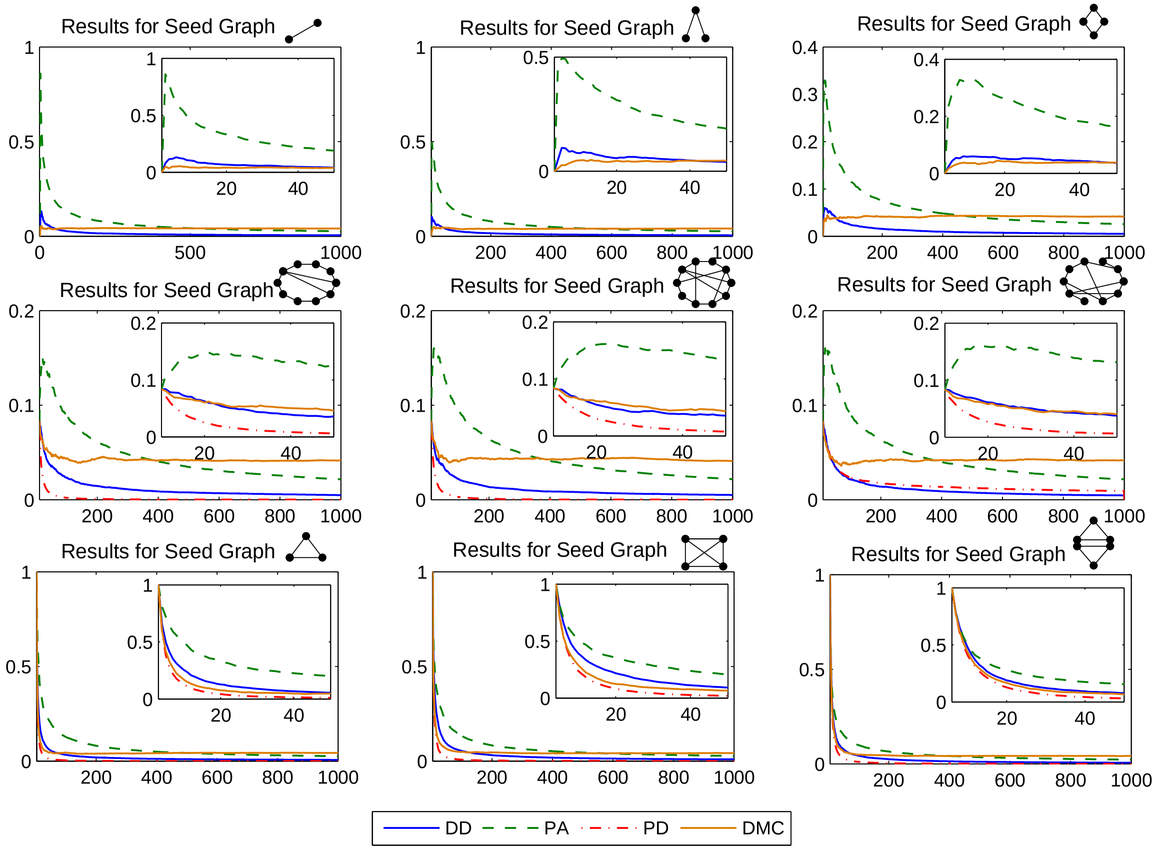

Figure 4.1: Plots for clustering coefficients of connected components in networks generated by the duplication and divergence model (blue), the preferential attachment model (green), the partial duplication model (red) and the duplication-mutation with complementarity model (orange) for 9 seed graphs. The x-axis is time, or equivalently the number of nodes of the network. The y-axis is clustering coefficient. The plots in the small boxes are enlarged figures for the curves in the time range between the initial time and time point 50.

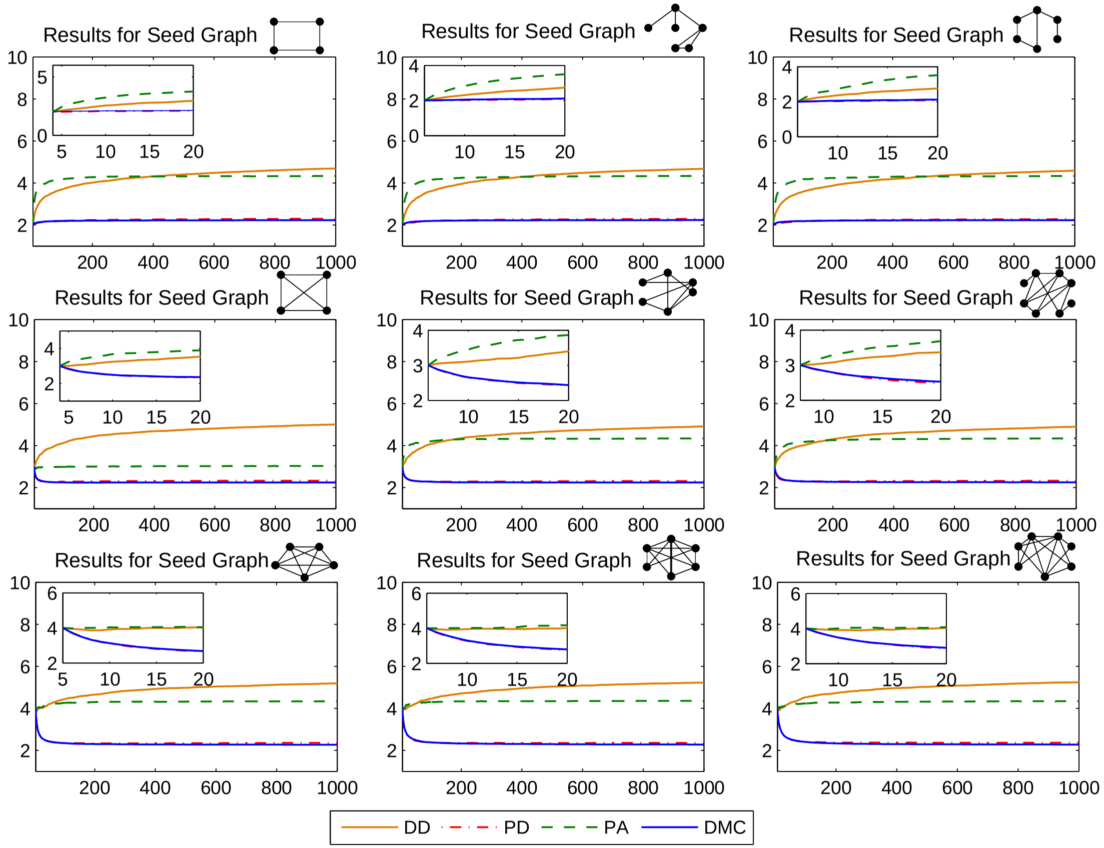

Figure 4.2: Plots for average degrees of connected components in networks generated by the duplication and divergence model (orange), the preferential attachment model (green), the partial duplication model (red) and the duplication-mutation with complementarity model (blue) for 9 seed graphs. The x-axis is time, or equivalently the number of nodes of the network. The y-axis is average degree. The plots in the small boxes are enlarged figures for the curves in the time range between the initial time and time point 20.

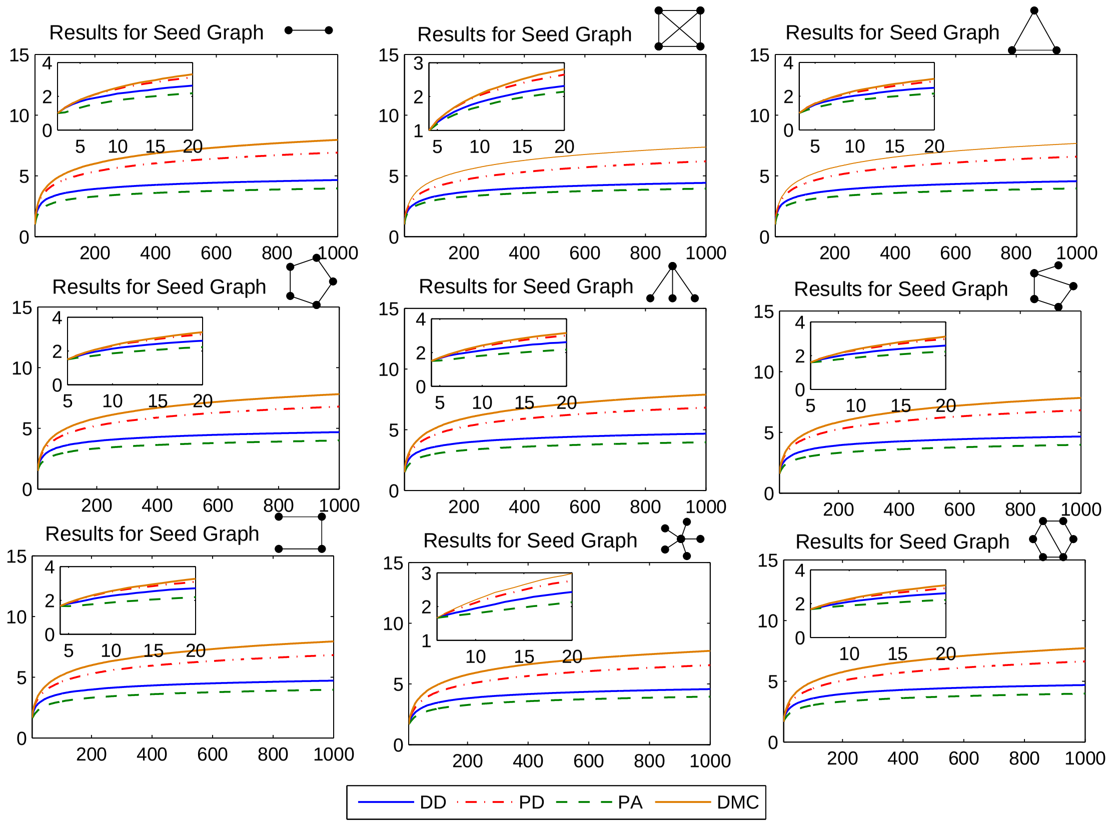

Figure 4.3: Change of average length of the shortest paths in networks generated by the duplication and divergence model (orange), the preferential attachment model (green), the partial duplication model (red) and the duplication-mutation with complementarity model (blue) for 9 seed graphs. The x-axis is time, or equivalently the number of nodes of the network. The y-axis is average length of the shortest paths. The plots in the small boxes are enlarged figures for the corresponding curves in the time range between the initial time and time point 20. Small length of shortest paths indicates that these models generate small-world networks.

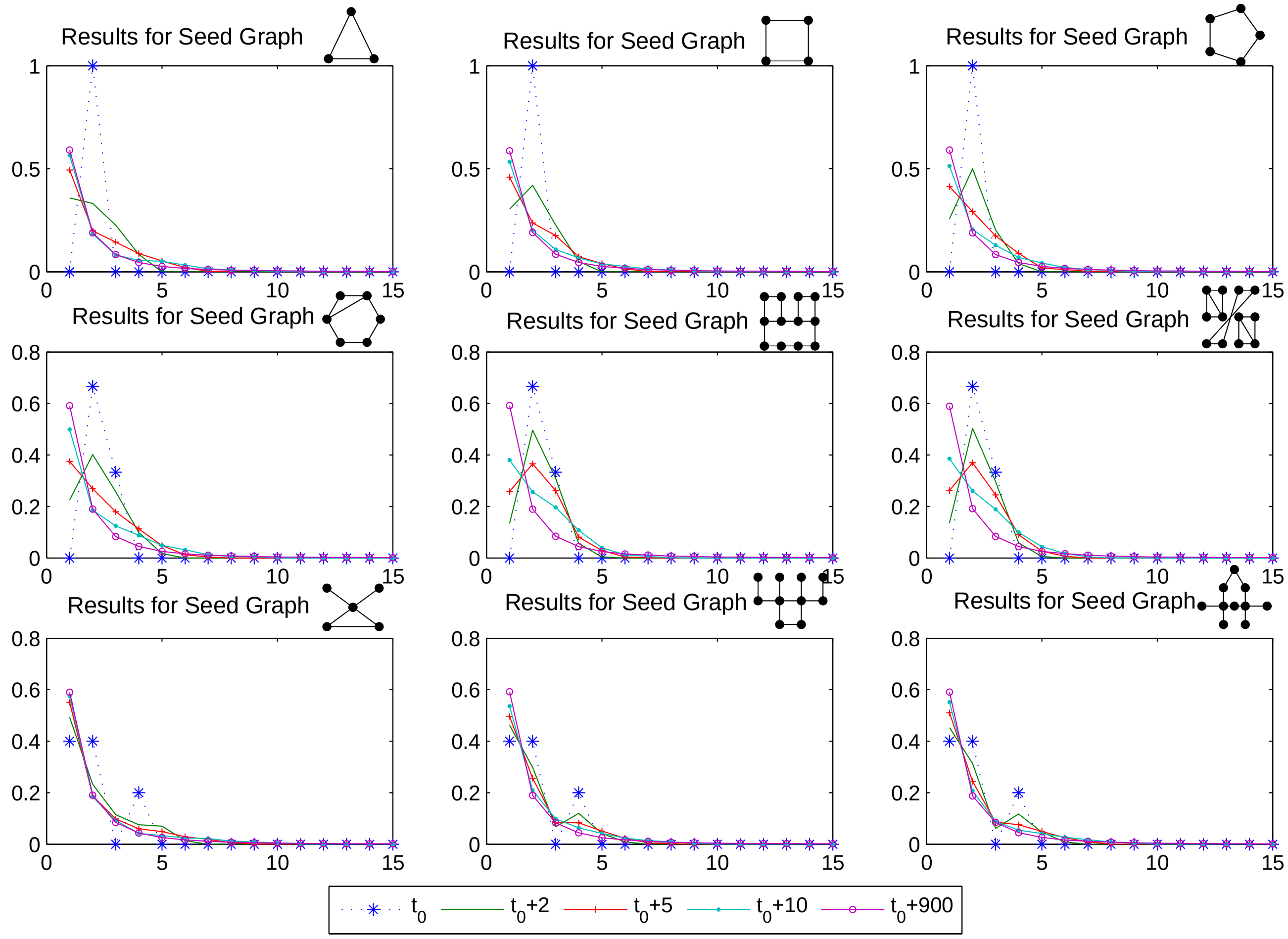

Figure 4.4: Plots for degree distribution of networks generated by the PD model for 9 seed graphs. The x-axis is the degree. The y-axis is the proportion of nodes with a certain degree. Plots are obtained at five time points: t_0 (blue), $t_0 + 2$ (green), $t_0 + 5$ (red), $t_0 + 10$ (aqua) and $t_0 + 900$ (purple).

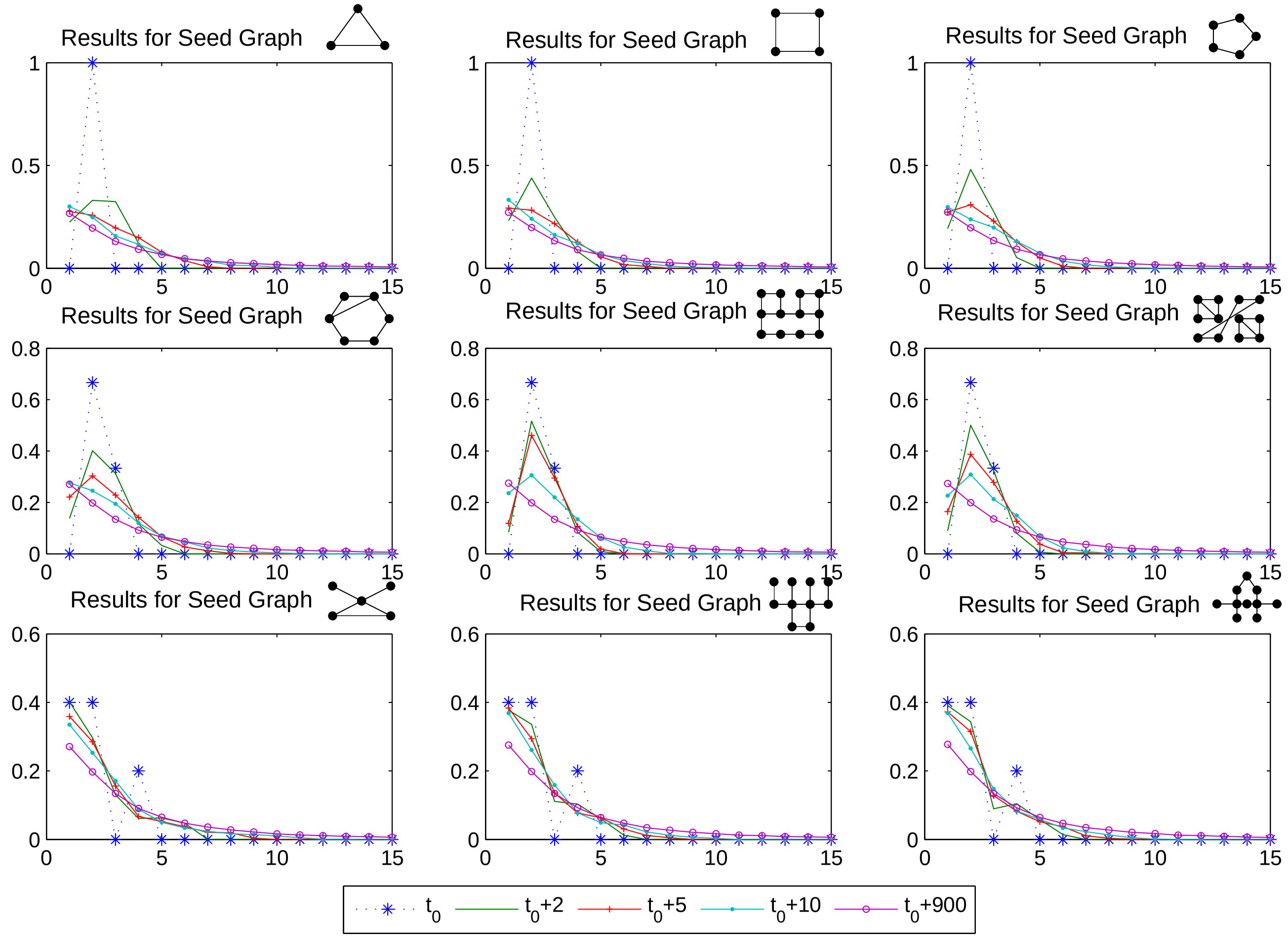

Figure 4.5: Plots for degree distribution of networks generated by the DD model for 9 seed graphs. The x-axis is the degree. The y-axis is the proportion of nodes with a certain degree. Plots are obtained at five time points: t_0 (blue), $t_0 + 2$ (green), $t_0 + 5$ (red), $t_0 + 10$ (aqua) and $t_0 + 900$ (purple).

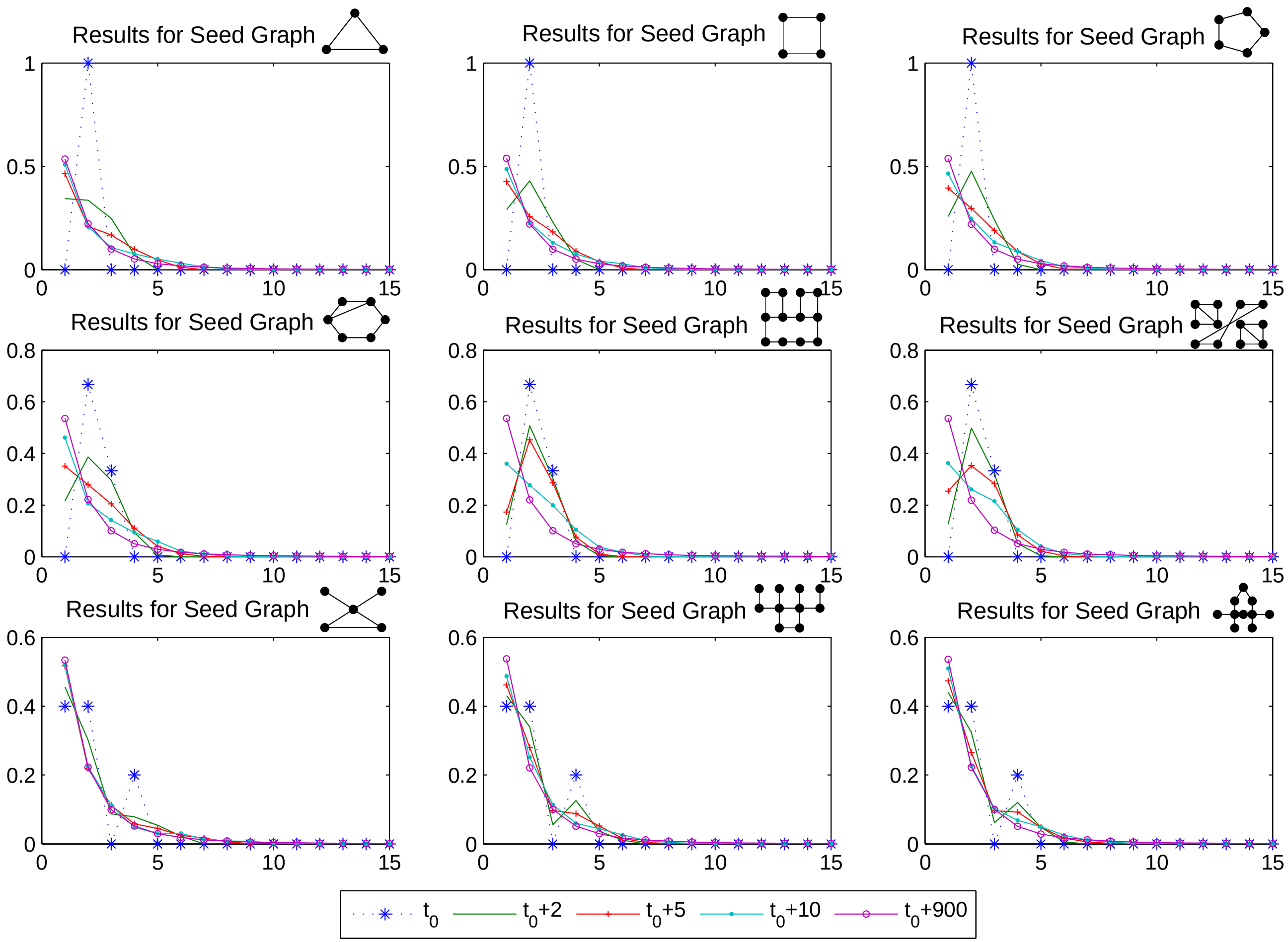

Figure 4.6: Plots for degree distribution of networks generated by the DMC model for 9 seed graphs. The x-axis is the degree. The y-axis is the proportion of nodes with a certain degree. Plots are obtained at five time points: t_0 (blue), $t_0 + 2$ (green), $t_0 + 5$ (red), $t_0 + 10$ (aqua) and $t_0 + 900$ (purple).

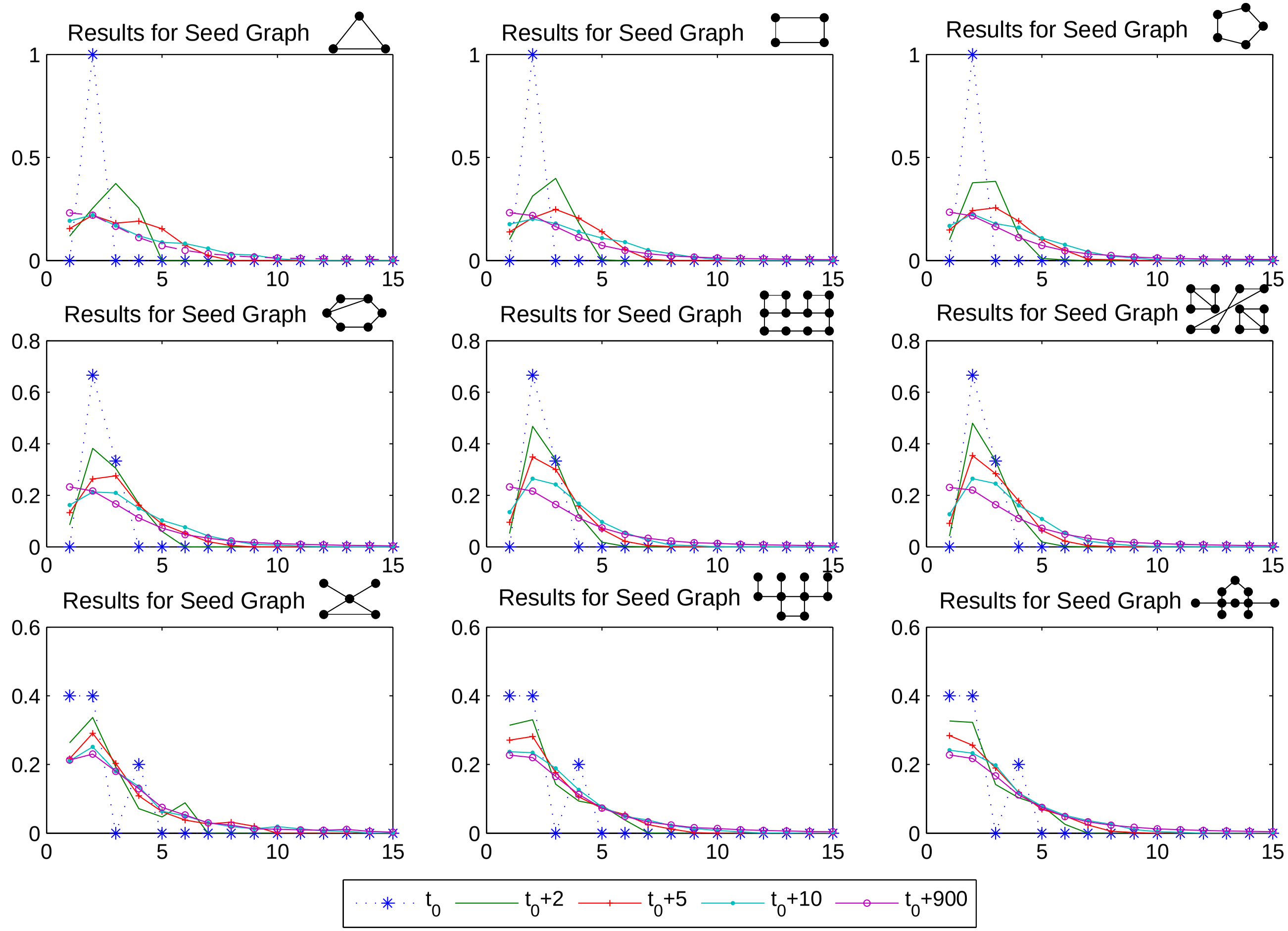

Figure 4.7: Plots for degree distribution of networks generated by the PA model for 9 seed graphs. The y-axis is the proportion of nodes with a certain degree. Plots are obtained at five time points: t_0 (blue), $t_0 + 2$ (green), $t_0 + 5$ (red), $t_0 + 10$ (aqua) and $t_0 + 900$ (purple).

	Clustering coefficient	Average degree	Average length of shortest paths	Degree distribution
Group1	0	2	1	$P(k = 2) = 1$
Group2	0.0833	3	1.5	$P(k = 2) = 2/3$ $P(k = 3) = 1/3$
Group3	1	4	5/3	$P(k = 1) = 2/5$ $P(k = 2) = 2/5$ $P(k = 4) = 1/5$

Table 4.1: Topological statics of seed graphs.

4.5 Discussion

We have done simulation studies on the duplication and divergence model, the preferential attachment model, the duplication-mutation with complementarity and the partial duplication model to investigate how the seed graphs affect the evolution of networks generated by these four models. The topological statistics we explored include clustering coefficient, average degree, average length of shortest paths and degree distribution. We found that in all the four models the clustering coefficient decreases as time is sufficiently large. The average degree of the DD model and the PA model approximately approach to a limit while the average degree of the PD model increase to infinity. These models all produce networks with small-world property, i.e. networks with small average length of shortest paths. We also find that the degree distribution of networks generated by these three models converge fast to a limit and the convergence rate depends on the degree distribution of the seed graph.

Chapter 5

Conclusion and Future Work

In summary, this thesis is devoted to modelling biological networks, especially the protein-protein interaction (PPI) networks, focusing on both the forward and backward properties of the network growth models.

For the backward issue of reconstructing the evolutionary history of PPI networks, we introduced a novel framework, based on the duplication-mutation with complementarity (DMC) model, to incorporate the information of the duplication history of its proteins. In earlier works of other authors, this problem was either studied by inference solely on networks [66] or methods combining the gene trees and PPI networks. The definition of duplication forest was introduced to represent the duplication history of the proteins in a PPI network [35]. The difficulty is that despite restricting histories to be compatible with a given duplication forest, the space of the network evolutionary history is still large, let alone the cases without duplication histories, in which the number of all possible histories are 2^n (n is the number of nodes). We observed that the seed graphs of two histories which are compatible with a given duplication history forest are isomorphic (Lemma 2.3.1). Based on this observation, the likelihood ratio between two histories has been proved to depend on only one parameter, the so-called loss number: The likelihood of one history is bigger than another if and only if its loss number is smaller than another (Theorem 2.3.3). This simplification allows us to formulate two efficient heuristic algorithms: MLN and CG. Simulation studies showed that MLN is faster than CG, but CG gives better results than MLN. Comparisons between our algorithm and an existing algorithm NetArch were done. Our methods outperformed NetArch in both speed and accuracy. Applications to the PPI networks of the baker's yeast, the worm and the fly were presented and analyzed. Our methods are based on the DMC model. Methods based on other models can be explored under the same framework.

The second issue deals with the degree distribution of networks generated by the partial duplication (PD) model. The PD model, just like the DMC model, belongs to the class of duplication models. The existence of the limiting degree distribution was established. Starting with the master equation Eq. 3.1, we proved that there is a phase transition point $p_0 \in [1/2, 1/\sqrt{2}]$ in the sense that the model generates networks with almost all nodes being singletons for $p < p_0$. Convergence rates were also derived. The existence of the limiting degree distribution for the connected components was also established. In contrast to the whole graph, the connected components were showed to be highly dense for $p < p_0$ when time is large. Furthermore for $p > p_0$ the connected components of the PD model were shown to follow a power-law degree distribution with the power-law exponent satisfying Eq. 3.17. The degree distribution of other duplication models can be investigated via the corresponding master equation too. Limiting analysis may also provide insight into other topological statistics.

The final part of the thesis explored the effect of seed graphs on the evolution of network models. Simulations to calculate the properties as a function of time were done for the DMC model, the duplication and divergent (DD) model, the PD model, and the preferential attachment (PA) model. Results have shown that the seed graphs have an impact on the evolution of the network models but this impact is not significant but limited. For instance, the decreasing tendency of the clustering coefficient is independent of the seed graphs. Extension of this part can be made to compare the topological features revealed by different methods for reconstructing evolutionary history which were considered in the first part of the thesis. Moreover, the seed graphs under consideration were all small graphs (with the number of nodes smaller than 20). However, the ancient networks obtained from many methods such as network comparisons and our two reconstruction algorithms are usually far larger than the seed networks we selected. Hence experiments can be designed for sufficiently large networks (such as networks with several hundred nodes) to see how the size of the seed graphs affects the network evolution.

Bibliography

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *Proc. 32nd Annual ACM Symposium on Theory of Computing*, pages 171–180, 2000.

[2] R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, 2002.

[3] P. Angenendt. Progress in protein and antibody microarray technology. *Drug Discov. Today*, 10:503–511, 2005.

[4] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–607, 2011.

[5] L. Aravind, H. Watanabe, D. J. Lipman, and E. V. Koonin. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci.*, 97:11319–11324, 2000.

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

[7] J. Bar-Ilan, M. Mat-Hassan, and M. Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, 2006.

[8] A. L. Barabasi and R. Albert. Emergence of scaling in random network. *Science*, 286(5439):509–512, 1999.

[9] A. L. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

[10] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter*, 13(3):547–560, 2000.

[11] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, and S. C. Sahinalp. The degree distribution of the generalized duplication model. *Theor. Comp. Sci.*, 369(1–3):239–249, 2006.

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

[13] D. Cancherini, G. Franca, and S. de Souza. The role of exon shuffling in shaping protein-protein interaction networks. *BMC Genomics*, 11(Suppl 5): S11, 2010.

[14] V. Carginalea, F. Trinchellab, C. Capassoa, R. Scudierob, and E. Parisi. Gene amplification and cold adaptation of pepsin in antarctic fish. a possible strategy for food digestion at low temperature. *Gene*, 336(2):195–205, 2004.

[15] A. Chatr-aryamontri et al. The biogrid interaction database: 2013 update. *Nucl. Acids. Res.*, 41(D1):D816–D823, 2013.

[16] C. H. C. Cheng, L. Chen, T. J. Near, and Y. Jin. Functional antifreeze glycoprotein genes in temperate-water new zealand nototheniid fish infer an antarctic evolutionary origin. *Mol. Biol. Evol.*, 20(11):1897–1908, 2003.

[17] F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[18] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. Duplication models for biological networks. *J. Comput. Biol.*, 10(5):677–687, 2003.

[19] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[20] W. Davids and Z. Zhang. The impact of horizontal gene transfer in shaping operons and protein interaction networkscdirect evidence of preferential attachment. *BMC Evol. Biol.*, 8:23, 2008.

[21] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[22] D. Durand, B. V. Halldorsson, and B. Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.*, 13(2):320–335, 2006.

[23] J. Dutkowski and J. Tiuryn. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 23(13):i149–i158, 2007.

[24] E. Eisenberg and E. Y. Levanon. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.*, 91(13):138701, 2003.

[25] F. Emmert-Streib and G. Glazko. Network biology: A direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 3:379–391, 2011.

[26] P. Erdos and A. Renyi. On random graphs I. *Publ. Math. Debreccen*, 6: 290–297, 1959.

[27] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–C61, 1960.

[28] K. Evlampiev and H. Isambert. Modeling protein network evolution under genome duplication and domain shuffling. *BMC Systems Biology*, 1:49, 2007.

[29] N. Farid and K. Christensen. Evolving networks through deletion and duplication. *New J. Phys.*, 8:212, 2006.

[30] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340:245–246, 1989.

[31] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[32] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568): 750–752, 2002.

[33] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[34] T. A. Gibson and D. C. Goldberg. Improving evolutionary models of protein interaction networks. *Bioinformatics*, 27(3):376–382, 2011.

[35] T. A. Gibson and D. S. Goldberg. Reverse engineering the evolution of protein interaction networks. *In Proc. of Pac. Symp. Biocomput*, pages 190–202, 2009.

[36] L. Giot et al. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[37] D. Graur and W. H. Li. *Fundamentals of Molecular Evolution.* Sinauer Associates, 2000.

[38] R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[39] L. Guzman-Vargas and M. Santillan. Comparative analysis of the transcription-factor gene regulatory networks of e. coli and s. cerevisiae. *BMC Systems Biology*, 2:13, 2008.

[40] O. Hagberg and C. Wiuf. Convergence properties of the degree distribution of some growing network models. *Bull. Math. Biol.*, 68(6):1275–1291, 2006.

[41] L. Hakes, J. W. Pinney, D. L. Robertson, and S. Lovell. Protein-protein interaction networks and biology–what's the connection? *Nat. Biotech.*, 26: 69–72, 2008.

[42] F. Hormozdiari. Protein protein interaction network comparison and emulation, 2006.

[43] 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 evolution. *PLoS Comp. Biol.*, 3:e118, 2007.

[44] I. Ispolatov, P. L. Krapivsky, and A. Yuryev. Duplication-divergence model of protein interaction network. *Phys. Rev. E*, 71(6):061911, 2005.

[45] T. Ito et al. 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. USA*, 97(3): 1143–1147, 2000.

[46] L. J. Jensen and P. Bork. Not comparable, but complementary. *Science*, 322 (5898):56–57, 2008.

[47] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[48] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[49] S. Kerrien et al. The intact molecular interaction database in 2012. *Nucl. Acids. Res.*, 40(D1):D841–CD846, 2012.

[50] J. Kim, P. L. Krapivsky, B. Kahng, and S. Redner. Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E*, 66(5): 055101, 2002.

[51] W. K. Kim and E. M. Marcotte. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS Comput. Biol.*, 4(11):e1000232, 2008.

[52] M. Knudsen and C. Wiuf. A markov chain approach to randomly grown graphs. *Journal of Applied Mathematics*, 2008:14, 2008.

[53] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci. USA*, 105 (19):6976–6981, 2008.

[54] O. Kuchaiev, A. Stevanovic, W. Hayes, and N. Przulj. Graphcrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, 12:24, 2011.

[55] E. S. Lander. *Calculating the Secrets of Life: Applications of the Mathematical Sciences in Molecular Biology.* Natl. Academy Pr., 1995.

[56] S. Li, K. P. Choi, and T. Wu. Degree distribution of large networks generated by the partial duplication model. *Theor. Comput. Sci.*, 476:94–108, 2013.

[57] S. Li, K.P. Choi, T. Wu, and L. Zhang. Maximum likelihood inference of the evolutionary history of a ppi network from the duplication history of its proteins. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, PP(99), 2013.

[58] S. Li et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

[59] H. Lodish et al. *Molecular Cell Biology.* W. H. Freeman, 4th edition, 2000.

[60] T. Makino, Y. Suzuki, and T. Gojobori. Differential evolutionary rates of duplicated genes in protein interaction network. *Gene*, 385:57–63, 2006.

[61] B. Manna, T. Bhattacharya, B. Kahali, and T. C. Ghosh. Evolutionary constraints on hub and non-hub proteins in human protein interaction network: Insight from protein connectivity and intrinsic disorder. *Gene*, 434:50–55, 2009.

[62] M. Middendorf, E. Ziv, and C. H. Wiggins. Inferring network mechanisms: The *drosophila melanogaster* protein interaction network. *Proc. Natl. Acad. Sci. USA*, 102(9):3192–3197, 2005.

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

[64] N. A. Moran. Microbial minimalism: Genome reduction in bacterial pathogens. *Cell*, 108:583–586, 2002.

[65] M. S. Mukhtar et al. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, 333(6042):596–601, 2011.

[66] S. Navlakha and C. Kingsford. Network archaeology: Uncovering ancient networks from present-day interactions. *PLoS Comput. Biol.*, 7:e1001119, 2011.

[67] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2): 026118, 2001.

[68] S. A. Nichols, W. Dirks, J. S. Pearse, and N. King. Early evolution of animal cell signaling and adhesion genes. *Proc. Natl. Acad. Sci. USA*, 103 (33):12451–12456, 2006.

[69] V. Van Noort, B. Snel, and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.*, 5(3):280C–284, 2004.

[70] M. A. Nowak, M. C. Boerlijst, J. Cooke, and J. M. Smith. Evolution of genetic redundancy. *Nature*, 388:167–171, 1997.

[71] P. Nurse and J. Hayles. The cell in an era of systems biology. *Cell*, 144(6): 850–854, 2011.

[72] S. Ohno. *Evolution by Gene Duplication.* Springer, 1970.

[73] R. Pastor-Satorras, E. Smith, and R. V. Sole. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.*, 222:199–210, 2003.

[74] R. Patro, E. Sefer, J. Malin, G. Marcais, S. Navlakha, and C. Kingsford. Parsimonious reconstruction of network evolution. *In Proc. of WABI'11, LNCS 6833.*

[75] R. Patro, E. Sefer, J. Malin, G. Marcais, S. Navlakha, and C. Kingsford. Parsimonious reconstruction of network evolution. *Algorithms Mol. Biol.*, 7: 25, 2012.

[76] J. W. Pinney, G. D. Amoutzias, M. Rattray, and D. L. Robertson. Reconstruction of ancestral protein interaction networks for the bzip transcription factors. *Proc. Natl. Acad. Sci. USA*, 104(51):20449–20453, 2007.

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

[78] J. C. Rain et al. The proteincprotein interaction map of *helicobacter pylori. Nature*, 409:211–215, 2001.

[79] A. Raval. Some asymptotic properties of duplication graphs. *Phys. Rev. E*, 68(6):066119, 2003.

[80] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297 (5586):1551–1555, 2002.

[81] 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. *Nat. Biotechnol.*, 17(10):1030–1032, 1999.

[82] T. Rito, C. M. Deane, and G. Reinert. The importance of age and high degree, in protein-protein interaction networks. *J. Comput. Biol.*, 19(6):785–795, 2012.

[83] J. De Las Rivas and C. Fontanillo. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, 6(6):e1000807, 2010.

[84] M. Rodbell. The role of hormone receptors and gtp-regulatory proteins in membrane transduction. *Nature*, 284(5751):17–22, 1980.

[85] S. A. Romano and M. C. Egui. Characterization of degree frequency distribution in protein interaction networks. *Phys. Rev. E*, 71(3):31901, 2005.

[86] S. M. E. Sahraeian and B. J. Yoon. A network synthesis model for generating protein interaction network families. *PLoS One*, 7(8):e41474, 2012.

[87] P. Shannon et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[88] R. Sharan, T. Ideker, B. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol.*, 12(6):835–846, 2005.

[89] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation networks of *Escherichia Coli*. *Nat. Genet.*, 31: 64–68, 2002.

[90] R. V. Sole, R. Pastor-Satorras, E. Smith, and T. Kepler. A model of large-scale proteome evolution. *Adv. Complex Syst.*, 5:43–54, 2002.

[91] U. Stelzl et al. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122:1173–1178, 2005.

[92] M. G. Sun and P. M. Kim. Evolution of biological interaction networks: From models to real data. *Genome Biol.*, 12:235–245, 2011.

[93] S. A. Teichmann and M. M. Bab. Gene regulatory network growth by duplication. *Nat. Genet.*, 36:492–496, 2004.

[94] P. Uetz et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[95] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1(1):38–44, 2003.

[96] A. S. Veron, K. Kaufmann, and E. Bornberg-Bauer. Evidence of interaction network evolution by whole-genome duplications: a case study in mads-box proteins. *Mol. Biol. Evol.*, 24(3):670–678, 2007.

[97] M. Vidal, M. E. Cusick, and A. L. Barabasi. Interactome networks and human disease. *Cell*, 144:986–998, 2011.

[98] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, 18(7):1283–1292, 2001.

[99] A. J. M. Walhout et al. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.

[100] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, 30(2):159–164, 2012.

[101] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440C–442, 1998.

[102] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *PNAS*, 103(20):7566–7570, 2006.

[103] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.

[104] S. Wuchty, Z. N. Oltvai, and A. L. Barabasi. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genet.*, 35:176–179, 2003.

[105] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. Dip: The database of interacting proteins. *Nucl. Acids. Res.*, 28(1):289–291, 2000.

[106] T. Yamada and P. Bork. Evolution of biomolecular networks–lessons from metabolic and protein interactions. *Nat. Rev. Mol. Cell Biol.*, 10:791–803, 2009.

[107] E. Yeger-Lotem et al. Network motifs in integrated cellular networks of transcriptioncregulation and proteincprotein interaction. *Proc. Natl. Acad. Sci. USA*, 101(16):5934–5939, 2004.

[108] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[109] J. Zhang. Evolution by gene duplication: An update. *Trends Ecol. Evol.*, 18 (6):292–298, 2003.

[110] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: Analysis and principles of biological networks. *Genes Dev.*, 21:1010–1024, 2007.