# A Study of Random Duplication Graphs and Degree Distribution Pattern of Protein-Protein Interaction Networks

by

Zheng Ma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

# Abstract

The motivation of this thesis is to find the reason why protein-protein interaction networks present a unique degree distribution pattern, where the majority of the proteins are sparsely connected, while densely-connected proteins also exist.

Since the degree distribution pattern of protein-protein interaction networks arises through a long-time evolutionary process of gene duplication, we introduce the model of random duplication graph to depict protein-protein networks mathematically. Specifically, we intend to derive the degree distribution function of protein-protein interaction networks by modeling protein-protein interaction networks as a special case of random duplication graph.

The random duplication graph model mimics the gene duplication process. In a random duplication graph, one vertex is chosen uniformly at random to duplicate at every timestep $t$, and all the edges of the original vertex are preserved by the new vertex. We derive the expected degree distribution function of the model from the probability master function. Furthermore, we learn from the Erdös-Rényi random graph model that the degree distribution function does not necessarily converge in a single random duplication graph. In consequences, we define the $n$-fold of random duplication graphs, a combination of $n$ independent random duplication graphs, under which we are able to prove that the degree distribution function converges.

Furthermore, we model the protein-protein interaction networks as a special case of random duplication graph with sparse initial graph, and the degree distribution function of protein-protein interaction networks is derived. We compare this degree distribution function with degree distribution data of reconstructed protein-protein interaction networks, and we show that this degree distribution function indeed resembles the degree distribution pattern in protein-protein interaction networks.

Our model gives a theoretical analysis of the self-organization process of protein-protein interaction networks. Moreover, we have shown that it is the gene duplication process combined with the sparsely-connected initial condition that leads to the unique degree distribution pattern in protein-protein interaction networks. We can make a further prediction based on our analysis—as the gene duplication process proceeds, the percentage of densely-connected proteins will be higher.

# Acknowledgement

First and foremost, I would like to thank my advisor Professor Liang-Liang Xie for his invaluable guidance, generous support, and continuous encouragement during my master's program. He introduced me to the exciting area of network biology, and taught me not only the prerequisite knowledge for my research, but also the insights and inspirations to conduct further researches.

Second, I would like to thank the readers of this thesis, Professor Zhou Wang and Professor Pin-Han Ho, for taking their precious time to read my thesis and providing constructive advices.

Last but not least, I am deeply indebted to my friends and families, for their love and support.

# Table of Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Problems and Motivations

Protein-protein interaction network is the map of protein-protein interactions in a living organism. In the network, proteins are represented as vertices, and protein-protein interactions are represented as edges. An edge exists between two proteins if they can interact with each other.

To understand how cells and organisms are developed, a comprehensive analysis of the protein-protein interaction networks is of pivotal importance. In this regard, understanding the degree distribution pattern of protein-protein interaction networks has been a major interest for system biologists.

As we can see from the degree distribution data of reconstructed protein-protein interaction networks, those networks present a unique degree distribution pattern: the degree distribution function of the networks is monotonically decreasing, the majority of proteins are sparsely connected, but densely-connected proteins also exist.

It is our major interest to understand how this unique degree distribution pattern comes into being.

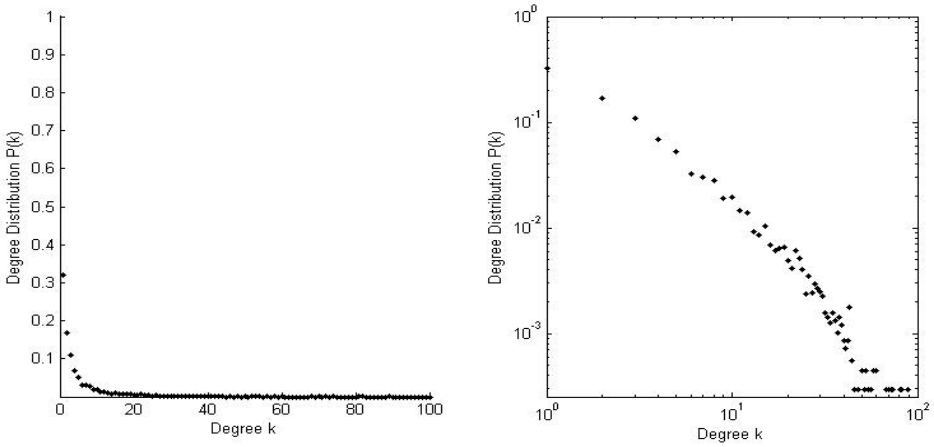


Figure 1.1: Degree distribution data of the reconstructed protein-protein interaction network of Drosophila melanogaster [2], through which we can see the unique degree distribution pattern.

Degree distribution pattern of protein-protein interaction networks is a direct result of the networks' self-organization process. The formation of the protein-protein interaction networks is a self-organization process, during which new proteins join the system over a long time period. In this process, new proteins are brought by gene duplication, a major mechanism through which new proteins are generated during molecular evolution.

Duplicated genes produce identical proteins that interact with the same set of protein partners. Therefore, each protein in contact with a duplicated protein gains an extra linkage. We suspect that the gene duplication process is the reason why protein-protein interaction networks present such unique degree distribution pattern.

Mathematically, the self-organization process of gene duplication can be modeled with random duplication graph, in which the gene duplication process is represented by random vertex duplication.

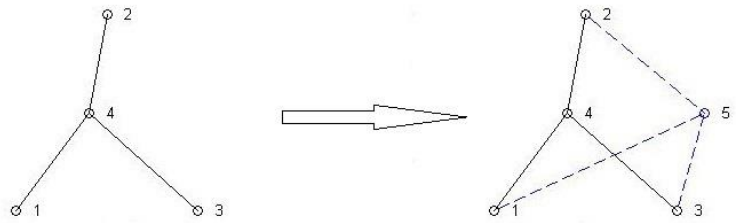


Figure 1.2: An example of random duplication graph.

Degree distribution function, $\mathbb{P}(k)$, is defined as the percentage of proteins with $k$ connections in the whole network [1]. The degree distribution function reveals the degree distribution pattern explicitly. In this thesis, we try to understand the degree distribution pattern of protein-protein interaction networks by studying the degree distribution function.

We are interested in the degree distribution function of the random duplication graph model. We hope to find the cause of the unique degree distribution pattern of protein-protein interaction networks, with the help of our random duplication graph model. Also, we intend to derive the degree distribution function of protein-protein interaction networks with the help of the random duplication graph model.

## 1.2 Contributions

- During our review of the Erdös-Rényi random graph model. We find that the degree distribution function of Erdös-Rényi random graph model is mistakenly derived. Previous researches used to assume that the degree distribution function, $\mathbb{P}(k)$, is equivalent to the probability that a randomly

chosen vertex has degree $k$. However, we point out that this claim is fallacious. As a result, the degree distribution function of the Erdös-Rényi random graph model is still unknown.

- The random duplication graph model is studied thoroughly. We derive the expected value of the degree distribution function through the probability master function. Similar to the case of Erdös-Rényi random graph model, it is difficult to discuss the convergence of degree distribution function of a single random duplication graph. So we define the $n$-fold of random duplication graphs, a combination of $n$ independent random duplication graphs, under which we prove that the degree distribution function converges.
- We model protein-protein interaction networks as a special case of the random duplication graph where the initial graph is sparse. The degree distribution function of protein-protein interaction networks is derived with the help of the random duplication graph model, and compared to degree distribution data of various reconstructed protein-protein interaction networks. The degree distribution function indeed resembles the degree distribution pattern in protein-protein interaction networks. Therefore, we show that it is the gene duplication process combined with the sparsely-connected initial condition that leads to the unique degree distribution pattern in protein-protein interaction networks.

## 1.3 Thesis Outline

This thesis is organized as follows,

In Chapter 2, we give a review of the preliminaries network biology and complex networks. As the prerequisite knowledge for this thesis, we introduce the study of

earlier random graph models: the Erdös-Rényi random graph model, and the Barabási-Albert Model. In particular, we discuss the study of the degree distribution function of Erdös-Rényi random graph model. We point out that the degree distribution function of the model was mistakenly derived in previous researches, since the assumption that the degree distribution function, $\mathbb{P}(k)$, is equivalent to the probability that a randomly chosen vertex has degree $k$ is fallacious.

In Chapter 3, we present the model of random duplication graph. We derive the expected degree distribution function of random duplication graph from the probability master function. Additionally, we propose the $n$-fold of random duplication graphs, a combination of $n$ independent random duplication graphs, under which we are able to prove that the degree distribution function converges in probability as $n \to \infty$.

In Chapter 4, we modeled protein-protein interaction networks as a special case of a random duplication graph where the initial graph is sparse, giving the degree distribution function of protein-protein interaction networks. Also, the properties of the acquired degree distribution function of protein-protein interaction network is analyzed, allowing us to predict the behavior of protein-protein interaction networks. We also give a comparison between our degree distribution function and degree distribution data of reconstructed protein-protein interaction networks, showing that our degree distribution function is valid.

Finally, we conclude this thesis and propose the possible future work to be done in Chapter 5.

# Chapter 2

# Preliminaries of Complex Networks and Network Biology

## 2.1 Erdös-Rényi Random Graph Model

In this section we introduce Erdös-Rényi random graph model. This model, proposed by Erdös and Rényi in 1959 [5], is the earliest study in random networks.

### 2.1.1 Definition and Properties of Erdös-Rényi Random Graph Model

The Erdös-Rényi random graph depicts a random graph $G(n,p)$ with $n$ given vertices $\{v_1, v_2, \cdots v_n\}$, and each pair of vertices connects with probability $p$, independent of every other pair of vertices.

To give a clear visualization of the Erdös-Rényi random graph and Poisson distribution, a realization of Erdös-Rényi random graph $G(100, 0.1)$ is given in Figure 2.1, and Figure 2.2 shows the degree distribution function of the Erdös-Rényi random graph $G(100, 0.1)$.

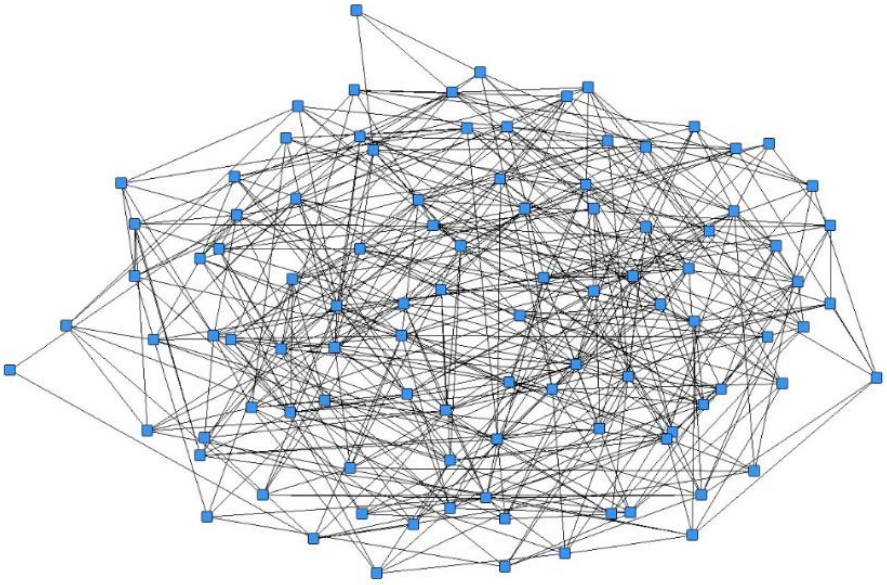

Figure 2.1: A realization of Erdös-Rényi random graph $G(100, 0.1)$.

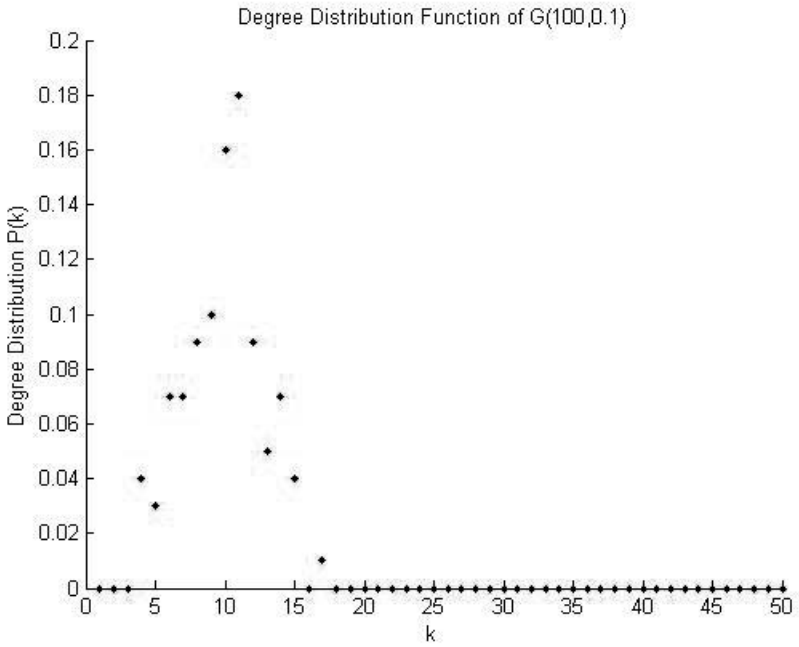


Figure 2.2: The degree distribution function of the Erdös-Rényi random graph $G(100, 0.1)$.

Our goal here is to find the degree distribution function of Erdös-Rényi random graph model. To discuss this, we need the formal definition of the degree distribution function.

***Definition 2.1.1.*** *The degree distribution function* $\mathbb{P}(k)$*, is defined as the number of degree-*$k$ *vertices divided by the total number of vertices for every* $k$*. Let* $F(k)$ *be the number of degree-*$k$ *vertices, and let* $n$ *be the total number of vertices,*

$$\mathbb{P}(k) = \frac{F(k)}{n} \tag{2.1}$$

The degree distribution function depicts important topological features, and the study of degree distribution function has been a major aspect in the study of random graphs.

It was stated in various papers that the degree distribution function of Erdös-Rényi random graph, $\mathbb{P}(k)$, is equivalent to the probability that a randomly chosen vertex has degree $k$ [1][3][4][6].

Furthermore, we hereby derive the degree distribution of a randomly chosen vertex in the Erdös-Rényi random graph model, or the probability that a randomly chosen vertex has degree $k$.

***Theorem 2.1.1.*** *In Erdös-Rényi random graph model, the degree distribution of a randomly chosen vertex follows a Poisson distribution, given the condition that* $p$ *is small, and* $n$ *is large.*

$$Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{2.2}$$

*where* $\lambda = (n-1)p$.

***Proof.*** *For* $\forall n$*, the probability that a randomly chosen vertex has degree* $k$ *is written as*

$$Prob_{n,p}(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2.3}$$

*The expected degree of a randomly chosen vertex,* $\lambda$*, is*

$$\lambda \equiv (n-1)p \tag{2.4}$$

*Considering the degree distribution as a function of the expected number of degrees, $\lambda$, we replace $p$ with $\lambda$. Equation 2.2 becomes*

$$Prob_{n,\lambda}(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = \binom{n-1}{k}\left(\frac{\lambda}{n-1}\right)^k\left(1-\frac{\lambda}{n-1}\right)^{n-1-k} \tag{2.5}$$

*Letting the number of vertices $n$ be sufficiently large, the degree distribution of a randomly chosen vertex then approaches*

$$Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) \tag{2.6}$$

$$= \lim_{n\to\infty} Prob_{n,\lambda}(\{a\ random\ chosen\ vertex\ has\ degree\ k\}) \tag{2.7}$$

$$= \lim_{n\to\infty} \binom{n-1}{k}\left(\frac{\lambda}{n-1}\right)^k\left(1-\frac{\lambda}{n-1}\right)^{n-1-k} \tag{2.8}$$

$$= \lim_{n\to\infty} \frac{(n-1)(n-2)\cdots(n-k)}{k!}\frac{\lambda^k}{(n-1)^k}\left(1-\frac{\lambda}{n-1}\right)^{n-1}\left(1-\frac{\lambda}{n-1}\right)^{-k} \tag{2.9}$$

$$= \lim_{n\to\infty} \frac{(n-1)(n-2)\cdots(n-k)}{(n-1)^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n-1}\right)^{n-1}\left(1-\frac{\lambda}{n-1}\right)^{-k} \tag{2.10}$$

$$= 1\cdot\frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1 \tag{2.11}$$

$$= \frac{e^{-\lambda}\lambda^k}{k!} \tag{2.12}$$

*and this completes the proof of Theorem 2.1.1.* □

Figure 2.3 shows the probability mass function of Poisson distribution with constant $\lambda = 10$.

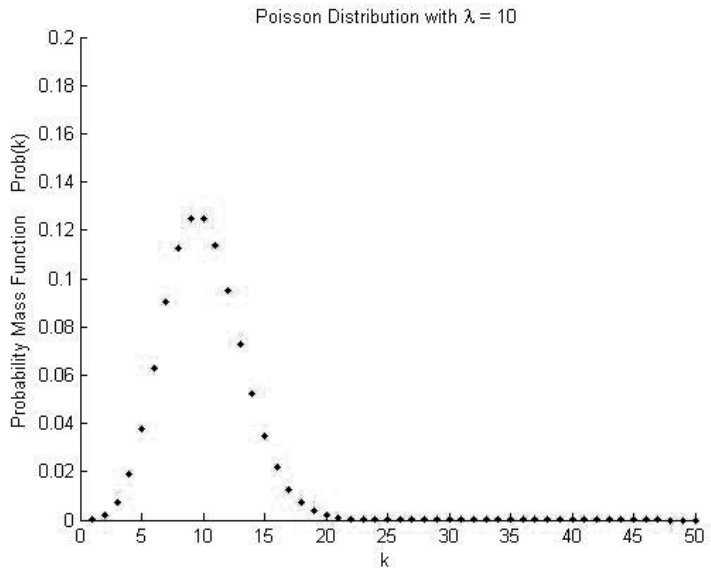


Figure 2.3: Poisson probability mass function.

If we accept the claim that the degree distribution function is equivalent to the probability that a randomly chosen vertex has degree $k$, the degree distribution function of the Erdös-Rényi random graph model will follow a Poisson distribution. In other terms, when $p$ is small, and as $n \to \infty$,

$$\mathbb{P}(k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{2.13}$$

## 2.1.2 Degree Distribution of Erdös-Renyi Random Graph Model

Based on the claim that the degree distribution function follows a Poisson distribution, Barabási and Albert gave a summary of the degree distribution pattern of Erdös-Rényi random graph [1]. Since the degree distribution of Erdös-Rényi random graph follows a Poisson distribution, the Erdös-Rényi random graph presents a feature of decentralization. Most vertices have approximately the same degree (close to the average degree $\lambda$). The tail (high $k$ region) of the degree distribution function

decreases exponentially, indicating that high-degree vertices are extremely rare. The number of low-degree vertices is small as well.

By comparing Figure 2.2 and Figure 2.3, we can see that the degree distribution function of the Erdös-Renyi random graph model is indeed similar to the Poisson distribution. In consequences, verifying Barabási and Albert's analysis of degree distribution pattern.

However, we question the claim that the degree distribution function is equivalent to the probability that a randomly chosen vertex has degree $k$.

In a matter of fact, if the degree distribution function of some random graph is given as $\mathbb{P}(k)$, it implies that the probability that a randomly chosen vertex has degree $k$ is $\mathbb{P}(k)$. On the contrary, if $Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\})$ is given, it does not necessarily imply that degree distribution function $\mathbb{P}(k)$ is equal to the probability that a randomly chosen vertex has degree $k$.

A simple example is given here to support this argument. Suppose some random graph reaches 3 distinct final states $\{G^1, G^2, G^3\}$ with probability $p_1$, $p_2$ and $p_3$ respectively, and each of the final state has a distinct degree distribution function $\{\mathbb{P}^1(k), \mathbb{P}^2(k), \mathbb{P}^3(k)\}$. Then the probability that a randomly chosen vertex has degree $k$ is written as

$$Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = p_1 \cdot \mathbb{P}^1(k) + p_2 \cdot \mathbb{P}^2(k) + p_3 \cdot \mathbb{P}^3(k) \qquad (2.14)$$

In this case, if we are given $Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\})$, the degree distribution function of the resulting random graph is still unknown. In fact, there will be 3 different possible degree distribution functions of the resulting random graph.

To give a clearer explanation of this argument, we take a closer look into the Erdös-Rényi random graph model. We notice that the degree distribution function of an Erdös-Rényi random graph of size $n$ can be written as

$$\mathbb{P}(k) = \frac{F(k)}{n} = \frac{\sum_{i=1}^{n} \mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}}{n} \tag{2.15}$$

where $\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}$ being the indicator random variable.

It is obvious that $\mathbb{E}[\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}] = Prob(\{a\ random\ chosen\ vertex\ has\ degree\ k\})$, since all vertices are created equally.

If we want $\mathbb{P}(k)$ to be equivalent to $Prob(\{a\ random\ chosen\ vertex\ has\ degree\ k\})$, it is necessary to have

$$\frac{\sum_{i=1}^{n} \mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}}{n} \xrightarrow{in\ probability} \mathbb{E}[\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}] \quad as\ n \to \infty \tag{2.16}$$

We can see that Equation 2.16 is in fact the result of weak law of large numbers.

To ensure that the weak law of large number holds, the condition that $\{\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}\}$ are independent random variables is sufficient. However, they are not independent. Consider an Erdös-Rényi random graph with size $n$, the probability that some vertex $v_i$ has degree $n-1$ is

$$Prob(\{vertex\ v_i\ has\ degree\ n-1\}) = p^{n-1} \tag{2.17}$$

However, given the condition that some vertex $v_i$ already has degree $n-1$, the conditional probability that some other vertex $v_j$ has degree $n-1$ is

$$Prob(\{vertex\ v_j\ has\ degree\ n-1\}|\{vertex\ v_i\ has\ degree\ n-1\}) = p^{n-2} \tag{2.18}$$

Thus we can see that $\{\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}\}$ are not independent random variables.

The weak law of large numbers can still hold for dependent random variables in some cases. Although there is no known necessary condition for the weak law of large numbers for dependent random variables to hold [7], the best known sufficient condition is the Bernstein's Theorem. However, $\{\mathbf{1}_{[vertex\ v_i\ has\ degree\ k]}\}$ do not satisfy the sufficient conditions given by Bernstein's Theorem either [8].

To sum up, it is not likely that the convergence in Equation 2.16 will hold. Thus, the claim that degree distribution function of Erdös-Rényi random graph model follows a Poisson distribution, $\mathbb{P}(k) = \frac{e^{-\lambda}\lambda^k}{k!}$, is fallacious. The behavior and degree distribution function of Erdös-Rényi random graph model need further study.

## 2.2 Complex Networks

Contemporary science has pointed out that systems in various disciplines ranging from molecular biology to computer science are composed of non-identical elements [9][10][11]. These systems have a particular topology feature: they form rather complex networks, whose vertices are the elements of the systems, and edges are the connections between the elements. For example, animals possess huge neural networks, whose vertices represent neurons, and edges represent the axon-dendrite connections between neurons. Also, living systems form huge protein-protein interaction networks, where vertices represent proteins, and edges represent the chemical interactions between proteins. Besides, complex networks present in social science and computer science. In a social network, vertices represent people, and edges represent the social interactions among people. In World Wide Web, vertices are HTML documents, and edges represent the links between HTML documents. Due to their large size and complexity, the topology of those complex networks remains largely unknown.

Driven by the automation of data acquisition, topological information of complex networks are become more and more available. Figure 2.4 shows the visualization of a small portion of the social network in Facebook, where vertices stand for users, and edges stand for the friendship relationships between users [12]. Figure 2.5 depicts the

degree distribution of the Facebook New Orleans network of 63,731 users and 817,035 friendships [13].

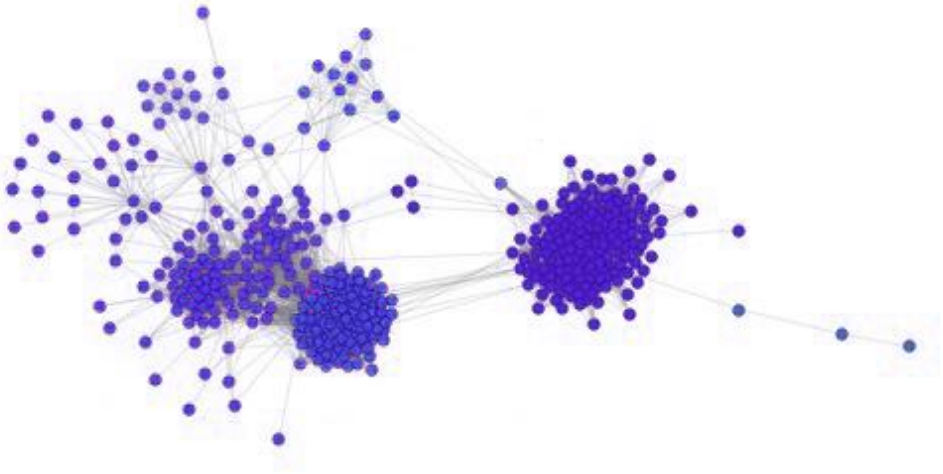

Figure 2.4: Social Network in Facebook.

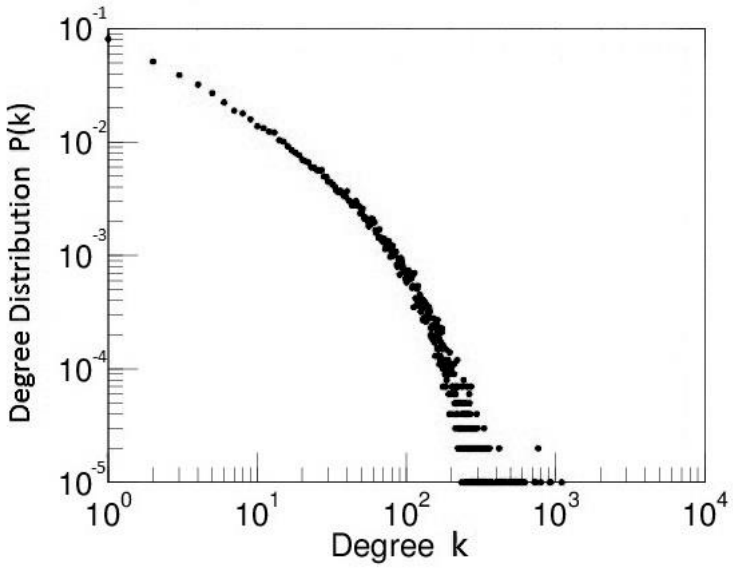


Figure 2.5: Degree distribution of the social network above.

From the Facebook social network, we can see that complex networks present a unique degree distribution pattern: the degree distribution function is monotonically decreasing, the majority of the vertices have relatively low degree, but high-degree

vertices also exist. The networks present some level of local clustering, where vertices are densely connected in small groups, but overall the network is still decentralized.

## 2.3 Scale-free Networks and Barabási-Albert Model

The Barabási-Albert model is considered a breakthrough in the study of complex networks. It is the first model depicting the self-organization process of certain real-world complex networks, and it explains why complex networks present unique degree distribution pattern. The degree distribution of Barabási-Albert model follows a scale-free distribution, so networks under the Barabási-Albert model are called scale-free networks.

### 2.3.1 Definition and Degree Distribution of Barabási-Albert Model

Traditionally, complex networks have been modeled using the Erdös-Rényi random graph model. But as we can see from acquired topological data of complex networks, the Erdös-Rényi random graph model cannot reveal the unique degree distribution pattern of complex networks.

In 1999, Barabási and Albert reported a model of self-organizing process of complex networks, namely the Barabási-Albert model [13].

Barabási and Albert pointed out that two generic aspects of real-world complex networks are absent in the Erdös-Rényi random graph model. First, the Erdös-Rényi random graphs are composed of a fixed number of vertices. On the contrary, real world complex networks are formed by continuous addition of new vertices to the network. For instance, in a social network, new members are introduced throughout the lifetime of the network, increasing the size of the size of the network. In World

Wide Web, the size grows exponentially in time by addition of new web pages. Second, Erdös-Rényi random graph model assumes that the probability that two vertices are connected is uniformly random. On the contrary, most real-life complex networks exhibit the behavior of preferential attachment, i.e. newly added vertices have a higher probability to be linked to a high-degree existing vertex than to a low-degree vertex. For example, in a social network, newly introduced members are more likely to be friend to those who are already popular in the social group, i.e. those who already have a lot of friends. In World Wide Web, if a webpage is already popular, it is more likely that it will be referred by newly created webpages, thus gaining more connections.

The Barabási-Albert model is built based on the two observations above. To depict the self-organization behavior of the network, scale-free networks are built from a small initial network with $n_0$ vertices. At every timestep, a new vertex with $m$ ($m \leq n_0$) edges is added to the network. The new vertex is linked to $m$ different vertices that already exist in the network. To depict the preferential attachment behavior, it is assumed that the probability $Prob(\{new\ vertex\ connects\ to\ vertex\ i\})$ that a new vertex will be connected to existing vertex $i$ depends on the degree $k_i$ of vertex $i$, such that $Prob(\{new\ vertex\ connects\ to\ vertex\ i\}) = k_i/\sum_j k_j$. After $t$ time steps, the model leads to a random network with $n = t + n_0$ vertices and $mt$ edges.

Simulation result with $m = n_0 = 5$ shows that the network under Barabási-Albert model evolves into an invariant state with degree distribution following a scale-free distribution with a constant $\gamma = 2.9 \pm 0.1$. More precisely,

$$\mathbb{P}(k) = ck^{-2.9\pm0.1} \tag{2.19}$$

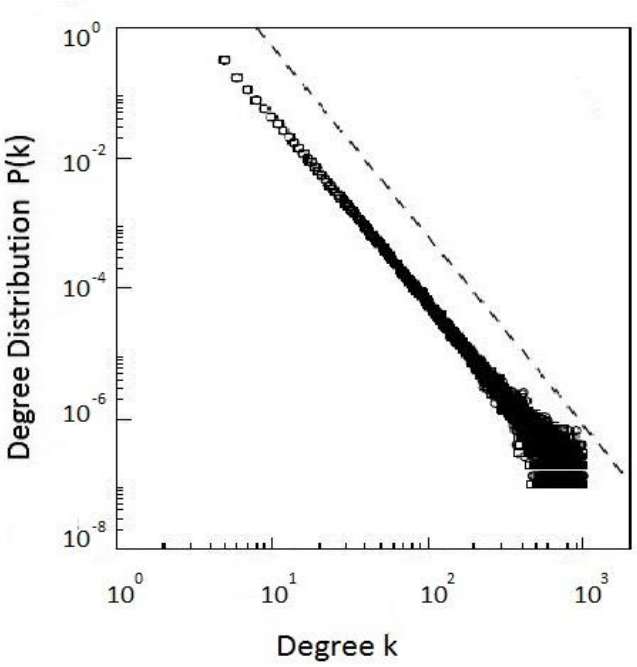


Figure 2.6: Simulation result of the degree distribution function of Barabási-Albert model, reused from [14] with permission.

The scale-free distribution is compared to degree distribution data from many real life complex networks. It is shown that the scale-free distribution function can indeed depict the degree distribution pattern of real-life complex networks [14][15][16].

In consequences, the Barabási-Albert model successfully reveals that the unique degree distribution pattern of various complex networks is a result of the preferential attachment behavior during these networks' self-organization process.

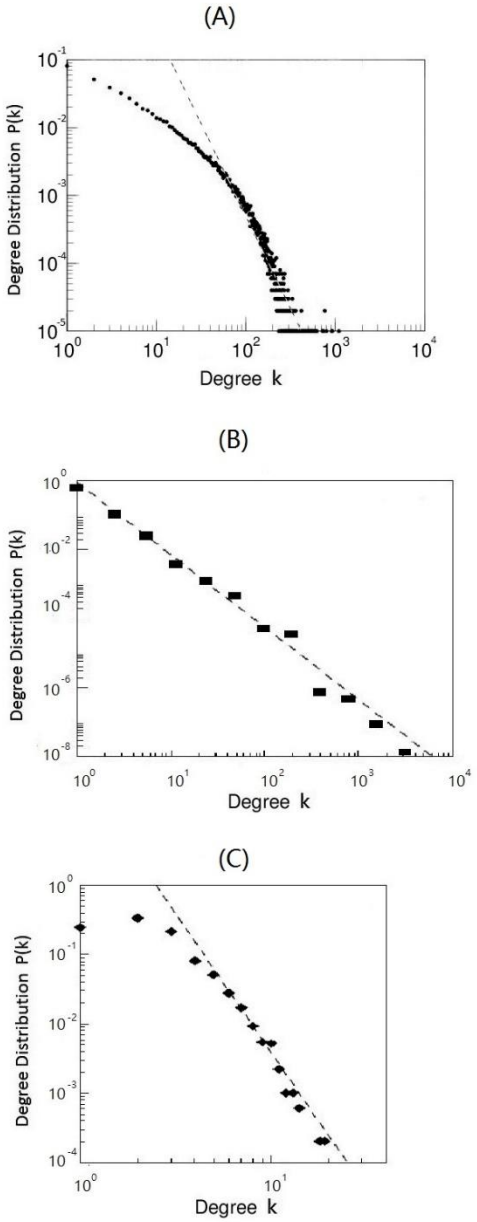


Figure 2.7: Degree distribution of some real-world complex networks. (A) Facebook New Orleans social network with 63,731 vertices; (B) World wide web with 325,729 vertices; (C) Power grid network with 4,941 vertices. Reused from [14] with permission.

### 2.3.2 Mean Field Theory for Barabási-Albert Model

To show that Barabási-Albert model results in a scale-free distribution, a mean-field method is developed to predict the growth dynamics of the Barabási-Albert model [4].

Recall that the Barabási-Albert model is a self-organization process with the behaviors of growth and preferential attachment. More precisely, the Barabási-Albert model starts with a small amount of $n_0$ vertices, at every timestep one new vertex with $m$ ($m \leq n_0$) edges is introduced to the network. The new vertex is linked to $m$ different vertices that already exist in the network. When choosing the links of the new vertex, the probability $Prob(\{new\ vertex\ connects\ to\ vertex\ i\})$ that a new vertex will be connected to existing vertex $i$ depends on the degree $k_i$ of vertex $i$, such that

$$Prob(\{new\ vertex\ connects\ to\ vertex\ i\}) = \frac{k_i}{\sum_j k_j} \tag{2.20}$$

As a result, after $t$ timesteps, the Barabási-Albert model produce a random network with $n = t + n_0$ vertices and $mt$ edges.

Figure 2.8 shows an example of the self-organization process of Barabási-Albert model with $n_0 = 5$ and $m = 5$.

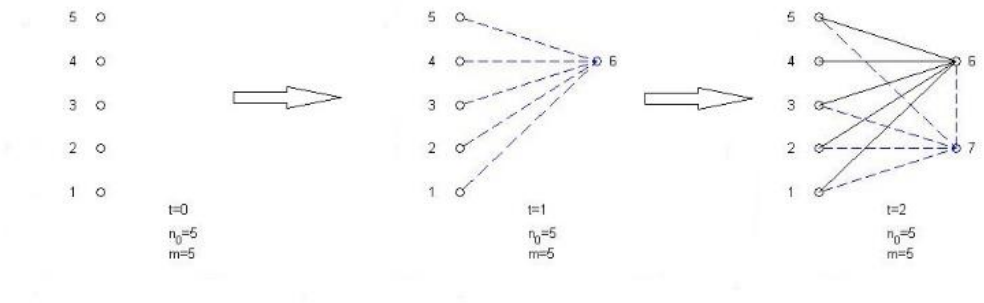


Figure 2.8: An example of the self-organization process of the Barabási-Albert model.

Assume that $k$ is continuous, then the probability that a new vertex will be connected to existing vertex $i$ becomes a continuous function of $k_i$. As a result, we have the following partial differential equation of the degree of vertex $i$.

$$\frac{\partial k_i}{\partial t} = C \cdot Prob(\{new\ vertex\ connects\ to\ vertex\ i\}) = C\frac{k_i}{\sum_{j=1}^{n_0+t-1} k_j} \tag{2.21}$$

Since the amount of edges at time $t$ is $mt$, we obtain that $\sum_{j=1}^{n_0+t-1} k_j = 2mt$. Also, since $m$ new edges are introduced at each timestep, we obtain that $C = m$. As a result, Equation 2.21 becomes

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t} \tag{2.22}$$

Take into consideration that vertex $i$ is introduced to the network at time $t_i$ with initial degree $k_i(t_i) = m$, the solution to Equation 2.22 is derived.

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{0.5} \tag{2.23}$$

Also, the probability that vertex $i$ has a degree $k_i(t)$ that is smaller than some $k$, i.e. $Prob(k_i(t) < k)$, is written as follows.

$$Prob(k_i(t) < k) = Prob\left(m\left(\frac{t}{t_i}\right)^{0.5} < k\right) = Prob\left(t_i > \frac{m^2 t}{k^2}\right) \tag{2.24}$$

Take into account that all vertices are introduced to the network at equal time intervals, the introduction time $t_i$ follows a uniform distribution.

$$Prob(t_i = c) = \frac{1}{n_0 + t} \quad for\ \forall c \le t \tag{2.25}$$

Combining Equation 2.24 and Equation 2.25, we obtain that

$$Prob\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - Prob\left(t_i \le \frac{m^2 t}{k^2}\right) = 1 - \frac{m^2 t}{k^2(t + n_0)} \tag{2.26}$$

Equation 2.26 is also the cumulative distribution function of some randomly chosen vertex $i$, i.e., the probability that a randomly chosen vertex $i$ has degree $k_i(t)$ that is smaller than $k$.

$$Prob(k_i(t) < k) = 1 - \frac{m^2 t}{k^2(t + n_0)} \tag{2.27}$$

The probability density function of $k_i(t)$, i.e. the degree distribution function of some randomly chosen vertex $i$, is derived by taking the derivative of the cumulative distribution function.

$$Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = \frac{\partial Prob(k_i(t) < \mathrm{k})}{\partial k} = \frac{2m^2 t}{n_0 + t}\frac{1}{k^3} \tag{2.28}$$

This indicates that the degree distribution function of some randomly chosen vertex $i$ follows a scale-free distribution.

$$Prob(\{a\ randomly\ chosen\ vertex\ has\ degree\ k\}) = ck^{-3} \tag{2.30}$$

With this result, Barabási and Albert claimed that the Barabási-Albert model result in a scale-free distribution with constant $\gamma = 3$, such that

$$\mathbb{P}(k) = ck^{-3} \tag{2.31}$$

However, the soundness of this claim is questionable.

The theory of Barabási-Albert model bares several weaknesses. First, the mean field theory for Barabási-Albert model is not a rigorous proof, the method of using a continuous distribution to approach a discrete distribution is questionable, and the validity of Equation 2.21 is unknown. Furthermore, even if the mean field theory is correct, the result from the mean field theory could just predict the probability that a randomly chosen vertex has degree $k$, which is not necessarily the degree distribution function of Barabási-Albert model. Additionally, not all complex networks follows the self-organization process of preferential attachment, leaving the Barabási-Albert model only suitable for a number of specific networks.

Nevertheless, the Barabási-Albert model is still recognized as a breakthrough in the theory of complex networks. First, it explains why complex networks present the special degree distribution pattern where most of the vertices have relatively low

degree, but high-degree vertices exist. Second, it introduces the scale-free degree distribution function which is valuable for understanding complex networks.

## 2.4 Network Biology and Protein-Protein Interaction Networks

Network biology is the study of complex networks in biological organisms. Every biological organism forms a large number of complex networks [17]. For instance, neural network is a series interconnected neurons, lined by synapses between axon terminals and dendrites of neurons. Gene regulatory network is a collection of DNA segments which interact with each other indirectly through their RNA and protein expression products to govern the gene expression levels. Among those biological complex networks, protein-protein interaction network is our major focus.

Protein-protein interactions are the physical contacts among proteins due to certain biochemical events. Multiple protein components organized by their protein-protein interactions form up the biological machines that carry out diverse essential biochemical processes. Thus, it is instrumental to understand protein-protein interactions in analyzing cellular functions.

Protein-protein interaction network is the map of protein-protein interactions in a given organism. In the network, proteins are represented as vertices, and an edge exists between two proteins if they can interact with each other. As systems biology advances, development of genome-scale protein-protein interaction networks became possible. To understand how cells and organisms are developed, a comprehensive analysis of the protein-protein interaction networks is of pivotal importance. In this regard, understanding the degree distribution pattern of protein-protein interaction networks has been a major interest for system biologists.

The protein-protein interaction networks falls into the category of complex networks due to various reasons.

First, protein-protein interaction networks have a relatively large scale. For instance, the published protein-protein interaction network of yeast Saccharomyces cerevisiae consists of 2,018 proteins and 2930 interactions. And the genome-scale protein-protein interaction networks of Drosophila melanogaster consists of 7,048 proteins and 20,405 interactions.

Also, the elements (proteins) in protein-protein interaction networks are non-identical, which is a key feature of complex networks.

Figure 2.9 shows a visualization of the protein-protein interaction network of Drosophila melanogaster [18]. From the visualization we can see that the reconstructed protein-protein interaction network presents the features of complex networks: the size is relatively large, and the network is composed of non-identical elements.

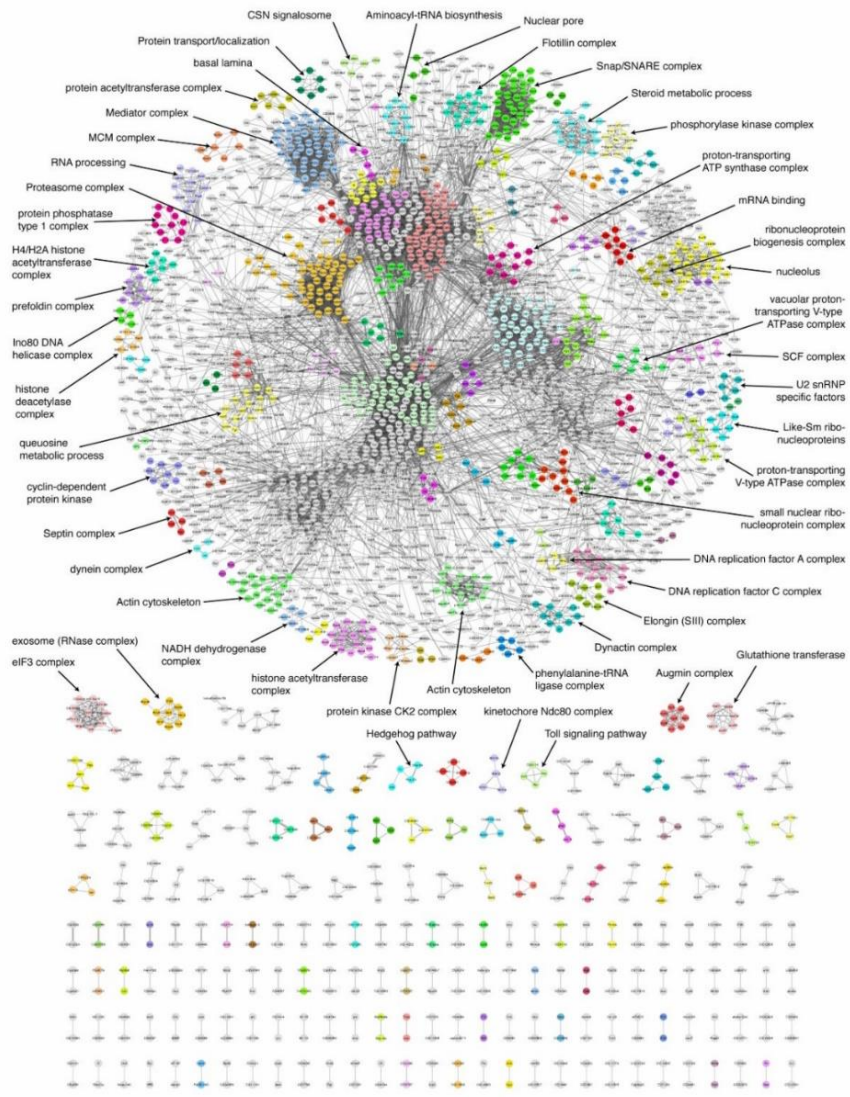


Figure 2.9: A graphical visualization of the protein-protein interaction network of Drosophila Melanogaster, reused from [18] with permission.

However, the Barabási-Albert model is not suitable for protein-protein interaction networks, since the self-organization process of protein-protein networks does not follow the preferential attachment property. As a result, we devise the random duplication graph model in an attempt to find out the reason why protein-protein interaction networks present such unique degree distribution pattern.

# Chapter 3

# Definition and Degree Distribution Function of Random Duplication Graph Model

## 3.1 Definition of Random Duplication Graph Model

We hereby give the formal mathematical definition of random duplication graph model.

The first thing we need to define is the initial graph. The initial graph, $G(t_0)$, is an undirected graph with $t_0$ vertices, (for simplicity let $t_0 \geq 2$).

Next we define the rule of random vertex duplication. At each discrete time step $t \geq t_0$, one vertex is chosen uniformly at random to duplicate itself, and all the existing edges of the chosen vertex are preserved by the new vertex. Since the duplicating vertex is chosen uniformly at random, the probability for any vertex to duplicate at time $t$ is $1/t$.

The random duplication graph at time $t$, $G(t)$, is the result of the random vertex duplication self-organization process, with some initial graph $G(t_0)$ as the initial state. As we can see, since exact one vertex is created during one timestep of the self-organization process, $G(t)$ contains $t$ vertices.

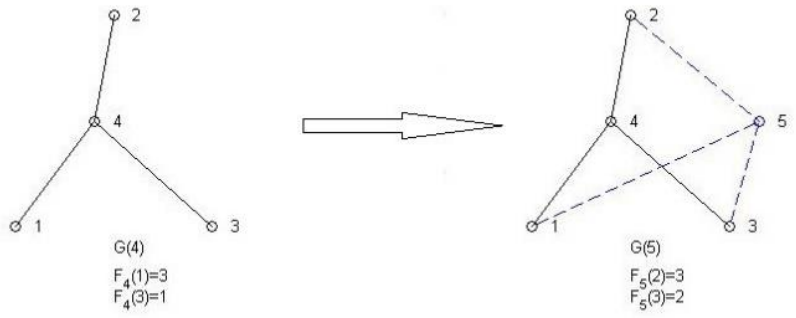


Figure 3.1: An example of the random vertex duplication process.

In Chapter 2, we learned that it is fallacious to assume that the degree distribution function $\mathbb{P}(k)$ is equivalent to the probability that a randomly chosen vertex has degree $k$. Therefore, here we try to derive the degree distribution function through its definition.

We define $F_t(k)$ as the number of vertices with $k$ connections at time $t$. Since there are $t$ vertices at time $t$, the degree distribution function at time *t* is written as $\mathbb{P}_t(k) = F_t(k)/t$. The information about the initial graph $G(t_0)$ is given as $\{F_{t_0}(i) \quad where\ 1 \le i \le t_0 - 1\}$ (obviously the maximum possible degree at time $t$ is $t - 1$). We are interested in the resulting degree distribution function of this self-organization process. More precisely, we are interested in $\{\mathbb{E}[\mathbb{P}_t(k)]\}$, the expected value of resulting degree distribution function of $G(t)$, in terms of the information about initial graph, $\{F_{t_0}(i) \quad where\ 1 \le i \le t_0 - 1\}$.

The reason why we only consider undirected graph in our model is because in the protein-protein interaction network, the connections between proteins are mutual. Two proteins being connected means that those two proteins can interact with each other.

The random vertex duplication mimics the duplication of genes. Gene duplication is defined as any duplication of a region of DNA that contains a gene. It is a major mechanism through which new proteins are generated during molecular evolution. Figure 3.2 illustrates the schematic of the process of a duplication event. We can see that after the duplication, two identical pieces of DNA fragments exist in the chromosome. If the duplicated area contains a gene, we expect to see an identical gene is introduced into the biological organism. These two genes will mutate independently from each other, over generations of the organism, becoming two different genes, providing different biological functions to the organism, which is called neofunctionalization. Although the two genes are different due to their independent mutation, they came from the same source. Thus they possess similar structure, making them interact the same set of other proteins. [19] This is the reason why after random vertex duplication, the edges of the duplicated vertex are copied as well. Since gene duplication is a purely random event, the duplicating vertex is chosen uniformly at random.

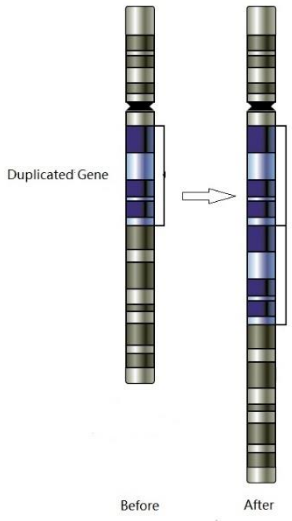


Figure 3.2: The schematic of gene duplication process.

## 3.2 Properties of Random Duplication Graph Model

To determine the degree distribution function of the random duplication graph $G(t)$, we analyzed the self-organization process of random vertex duplication. It is observed that the change of $F_t(k)$, the number of vertices with $k$ connections at time $t$, comes from three events. The event $\{some\ degree\text{-}k\ vertex\ duplicated\}$ will introduce a new vertex with degree $k$. Also, the event $\{some\ degree\text{-}k\ vertex's\ neighbor\ duplicated\}$ will change this existing degree-$k$ vertex to a degree-$(k+1)$ vertex. Similarly, the event $\{some\ degree\text{-}(k-1)\ vertex's\ neighbor\ duplicated\}$ will change this existing degree-$(k-1)$ vertex to a degree-$k$ vertex. Since for the existing vertices, the maximum possible degree change is 1, there is no other source of additional degree-$k$ vertices. Take Figure 3.3 as an example, the duplication of vertex 4 produces vertex 5. Vertex 4 is a degree-1 vertex itself, as well as a neighbor of a degree-3 vertex. As a result, duplication of vertex 4 causes the creation of a new degree-1 vertex (vertex 5). Also, vertex 1 is changed from a degree-3 vertex to a degree-4 vertex.

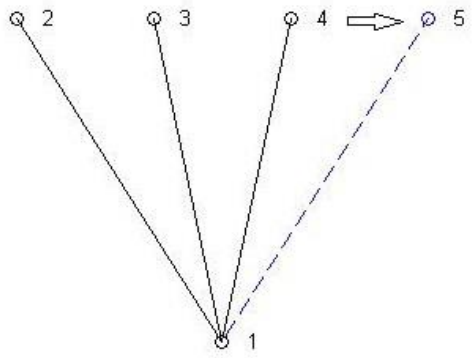


Figure 3.3: Vertex 4 duplicates to produce vertex 5. Vertex 4 is a degree-1 vertex itself, as well as a neighbor of a degree-3 vertex. As a result, duplication of vertex 4 causes the creation of a new degree-1 vertex (vertex 5). Also, vertex 1 is changed from a degree-3 vertex to a degree-4 vertex.

***Theorem 3.2.1.*** *The probability master function of random duplication graph model, or the relationship between* $G(t+1)$ *and* $G(t)$ *is given by*

$$\mathbb{E}[F_{t+1}(k)|\{F_t(i) \quad where\ 1 \le i \le t-1\}] = F_t(k) + \frac{1}{t}F_t(k) - \frac{k}{t}F_t(k) + \frac{k-1}{t}F_t(k-1) \tag{3.1}$$

***Proof.*** *Let us define some auxiliary random variables for convenience.*

*Define* $\theta_t(k) = \begin{cases} 1 & if\ the\ vertex\ chosen\ to\ duplicate\ at\ time\ t\ is\ degree\text{-}k \\ 0 & otherwise \end{cases}$*, i.e. the indicator random variable of event* $\{the\ vertex\ chosen\ to\ duplicate\ at\ time\ t\ is\ degree\text{-}k\}$*. Note that* $\theta_t$ *is also the indicator random variable of event* $\{the\ new\ vertex\ at\ time\ t+1\ is\ degree\text{-}k\}$.

*Define* $\lambda_t(k)$ *as the number of degree-k vertices at time* $t$ *that has the duplicating vertex as their neighbor. These degree-k vertices will be converted to degree-*$(k+1)$ *vertices at time* $t+1$.

*Define* $\mu_t(k)$ *as the number of degree-*$(k-1)$ *vertices at time* $t$ *that has the duplicating vertex as their neighbor. These degree-*$(k-1)$ *vertices will be converted to degree-k vertices at time* $t+1$.

*As a result, we could see that*

$$F_{t+1}(k) = F_t(k) + \theta_t(k) - \lambda_t(k) + \mu_t(k) \tag{3.2}$$

*and*

$$\mathbb{E}[F_{t+1}(k)|\{F_t(i) \quad where\ 1 \le i \le t-1\}] \tag{3.3}$$

$$= F_t(k) + \mathbb{E}[\theta_t(k)|\{F_t(i)\}] - \mathbb{E}[\lambda_t(k)|\{F_t(i)\}] + \mathbb{E}[\mu_t(k)|\{F_t(i)\}] \tag{3.4}$$

*Also, probability of the event* $\{the\ vertex\ chosen\ to\ duplicate\ at\ time\ t\ is\ degree\text{-}k\}$ *is* $F_t(k)/t$*, since the selection of the duplicating vertex is uniformly at random. As a result, we obtain that*

$$\mathbb{E}[\theta_t(k)|\{F_t(i)\}] = \frac{F_t(k)}{t} \tag{3.5}$$

*It is easily noted that* $\lambda_t(k)$ *is not binomial distributed, making it harder to compute* $\mathbb{E}[\lambda_t(k)|\{F_t(i)\}]$. *We observe that there are total* $k \cdot F_t(k)$ *neighbors of degree-k vertices, since each degree-k vertex has* $k$ *neighbors. However, some of the neighbors are shared by multiple degree-k vertices.*

*We assume that vertex* $v_1$ *is the neighbor of* $m_1$ *degree-k vertices, vertex* $v_2$ *is the neighbor of* $m_2$ *degree-k vertices, …, vertex* $v_j$ *is the neighbor of* $m_j$ *degree-k vertices. In consequence,* $k \cdot F_t(k) - m_1 - m_2 - \cdots - m_j$ *vertices are the neighbor of only one degree-k vertices separately.*

*As a result, we obtain that*

$$\mathbb{E}[\lambda_t(k)|\{F_t(i)\}] = 1 \cdot \frac{k \cdot F_t(k) - m_1 - m_2 - \cdots - m_j}{t} + m_1 \cdot \frac{1}{t} + m_2 \cdot \frac{1}{t} + \cdots + m_j \cdot \frac{1}{t} = \frac{k}{t} F_t(k) \tag{3.6}$$

*Although the expected value of* $\lambda_t(k)$ *resembles the expected value of a binomial distributed random variable, it is important to notice that* $\lambda_t(k)$ *is not binomial distributed, since the events* $\{some\ degree\text{-}k\ vertex's\ neighbor\ duplicated\}$ *and* $\{some\ other\ degree\text{-}k\ vertex's\ neighbor\ duplicated\}$ *are not independent.*

*Similarly, we have the following result for* $\mathbb{E}[\mu_t(k)|\{F_t(i)\}]$.

$$\mathbb{E}[\mu_t(k)|\{F_t(i)\}] = \frac{(k-1)F_t(k-1)}{t} \tag{3.7}$$

*To sum up, the relationship between* $G(t+1)$ *and* $G(t)$ *is derived as follows.*

$$\mathbb{E}[F_{t+1}(k)|\{F_t(i) \quad where\ 1 \le i \le t-1\}] = F_t(k) + \frac{1}{t}F_t(k) - \frac{k}{t}F_t(k) + \frac{k-1}{t}F_t(k-1) \tag{3.8}$$

*and this completes the proof of Theorem 3.2.1.* □

Equation 3.8 can be solved exactly by writing out each term on the right hand side in terms of earlier time steps.

***Corollary 3.2.1.*** *The solution to Equation 3.8 in terms of information about the initial graph,* $\{F_{t_0}(i) \quad where\ 1 \le i \le t_0 - 1\}$*, is given by*

$$\mathbb{E}[F_t(k)] = \sum_{j=max\{k-t+t_0;1\}}^{min\{k;t_0-1\}} \frac{\binom{k-1}{j-1}\binom{t-k}{t_0-j}}{\binom{t-1}{t_0-1}} F_{t_0}(j) \tag{3.9}$$

***Proof.*** *Write out the right hand side of Equation 3.8 in terms of earlier time steps, we obtain that*

$$\mathbb{E}[F_{t+1}(k)] \tag{3.10}$$

$$= \mathbb{E}[F_t(k)] + \tfrac{1}{t}\mathbb{E}[F_t(k)] - \tfrac{k}{t}\mathbb{E}[F_t(k)] + \tfrac{k-1}{t}\mathbb{E}[F_t(k-1)] \tag{3.11}$$

$$= \tfrac{t+1-k}{t}\mathbb{E}[F_t(k)] + \tfrac{k-1}{t}\mathbb{E}[F_t(k-1)] \tag{3.12}$$

$$= \tfrac{t+1-k}{t}\left[\tfrac{t-k}{t-1}\mathbb{E}[F_{t-1}(k)] + \tfrac{k-1}{t-1}\mathbb{E}[F_{t-1}(k-1)]\right] + \tfrac{k-1}{t}\left[\tfrac{t+1-k}{t-1}\mathbb{E}[F_{t-1}(k-1)] + \tfrac{k-2}{t-1}\mathbb{E}[F_{t-1}(k-2)]\right] \tag{3.13}$$

$$= \tfrac{(t+1-k)(t-k)}{t(t-1)}\mathbb{E}[F_{t-1}(k)] + 2\tfrac{(t+1-k)(k-1)}{t(t-1)}\mathbb{E}[F_{t-1}(k-1)] + \tfrac{(k-1)(k-2)}{t(t-1)}\mathbb{E}[F_{t-1}(k-2)] \tag{3.14}$$

$$= \tfrac{(t+1-k)(t-k)(t-1-k)}{t(t-1)(t-2)}\mathbb{E}[F_{t-2}(k)] + 3\tfrac{(t+1-k)(k-1)(k-2)}{t(t-1)(t-2)}\mathbb{E}[F_{t-2}(k-1)] + 3\tfrac{(t+1-k)(k-1)(k-2)}{t(t-1)(t-2)}\mathbb{E}[F_{t-2}(k-2)] + \tfrac{(k-1)(k-1)(k-3)}{t(t-1)(t-2)}\mathbb{E}[F_{t-2}(k-3)] \tag{3.15}$$

$$= \cdots \tag{3.16}$$

$$= \sum_{i=\max\{0;k-t_0+1\}}^{\min\{t+1-t_0;k-1\}} \frac{[(k-1)(k-2)\ldots(k-i)][(t+1-k)(t-k)\ldots(t_0+i-k+1)]}{t(t-1)(t-2)\ldots(t_0+1)t_0}\binom{t+1-t_0}{i}\mathbb{E}[F_{t_0}(k-i)] \tag{3.17}$$

*Since we are interested in the moment at time* $t$*, we adjust Equation 3.17 from time* $t+1$ *to time* $t$*, we obtain that*

$$\mathbb{E}[F_t(k)] \tag{3.18}$$

$$= \sum_{i=\max\{0;k-t_0+1\}}^{\min\{t-t_0;k-1\}} \frac{[(k-1)(k-2)\ldots(k-i)][(t-k)(t-k-1)\ldots(t_0+i-k+1)]}{(t-1)(t-2)\ldots(t_0+1)t_0} \binom{t-t_0}{i} \mathbb{E}[F_{t_0}(k-i)] \tag{3.19}$$

*Replacing* $k-i$ *with a single variable* $j$*, we obtain a simpler solution of* $\mathbb{E}[F_t(k)]$ *in terms of* $\{F_{t_0}(i) \quad where\ 1 \le i \le t_0 - 1\}$.

$$\mathbb{E}[F_t(k)] \tag{3.20}$$

$$= \sum_{j=\max\{k-t+t_0;\,1\}}^{\min\{k;t_0-1\}} \frac{[(k-1)(k-2)\ldots j][(t-k)(t-k-1)\ldots(t_0-j+1)]}{(t-1)(t-2)\ldots(t_0+1)t_0} F_{t_0}(j) \tag{3.21}$$

$$= \sum_j \frac{\frac{(k-1)!}{(j-1)!}\frac{(t-k)!}{(t_0-j)!}}{\frac{(t-1)!}{(t_0-1)!}} \frac{(t-t_0)!}{(k-j)!(t-t_0-k+j)!} F_{t_0}(j) \tag{3.22}$$

$$= \sum_j \frac{(k-1)!(t-k)!(t-t_0)!(t_0-1)!}{(j-1)!(t_0-j)!(t-1)!(k-j)!(t-t_0-k+j)!} F_{t_0}(j) \tag{3.23}$$

$$= \sum_{j=max\{k-t+t_0;1\}}^{min\{k;t_0-1\}} \frac{\binom{k-1}{j-1}\binom{t-k}{t_0-j}}{\binom{t-1}{t_0-1}} F_{t_0}(j) \tag{3.24}$$

*thus completing the proof of Corollary 3.2.1.* □

Furthermore, the expected value of the degree distribution function $\mathbb{E}[\mathbb{P}_t(k)]$, is a direct result by dividing $\mathbb{E}[F_t(k)]$ with the total number of vertices $t$.

$$\mathbb{E}[\mathbb{P}_t(k)] = \frac{1}{t} \sum_{j=max\{k-t+t_0;1\}}^{min\{k;t_0-1\}} \frac{\binom{k-1}{j-1}\binom{t-k}{t_0-j}}{\binom{t-1}{t_0-1}} F_{t_0}(j) \tag{3.25}$$

## 3.3 $N$-fold of Random Duplication Graphs and Convergence of Degree Distribution Function

As we can see from the case of Erdös-Rényi random graph model in Chapter 2, it is difficult for us to discuss the convergence of degree distribution function of a single random duplication graph, since a single random duplication graph could result in multiple possible degree distribution functions. In consequences, we seek out for other conditions under which we can discuss the convergence of the degree distribution function.

To discuss the convergence of the degree distribution function, we need to give the formal definition first. As we can see, $\{\mathbb{P}(k)\}$ are a series of random variables, the convergence of degree distribution function means that the series of random variables, $\{\mathbb{P}(k)\}$, converges simultaneously to their expected values, $\{\mathbb{E}[\mathbb{P}(k)]\}$. To guarantee that, we simply require that the supremum of $|\mathbb{P}(k) - \mathbb{E}[\mathbb{P}(k)]|$ converges to 0 in probability. In mathematical terms, the definition of convergence of degree distribution function is given below.

***Definition 3.3.1.*** *The degree distribution function* $\mathbb{P}(k)$*, converges to its expected value* $\mathbb{E}[\mathbb{P}(k)]$ *if and only if*

$$\sup_k\{|\mathbb{P}(k) - \mathbb{E}[\mathbb{P}(k)]|\} \xrightarrow{in\ probability} 0 \tag{3.26}$$

Now we define the $n$-fold of random duplication graphs, under which we can discuss the convergence of degree distribution function.

Consider some graph $G^*(t)$ composed of $n$ subgraphs $\{G^1(t), G^2(t), \cdots G^n(t)\}$. Each of the $n$ subgraphs itself is a random duplication graph, with an identical subgraph $G(t_0)$. During the self-organization process of random duplication, each subgraph evolves independently. $G^*(t)$ is named the $n$-fold of random duplication graphs.

An example is given below to illustrate the model of $n$-fold of random duplication graphs.

Suppose we start from the initial graph $G(4)$ in Figure 3.4.

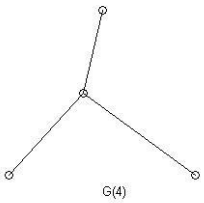


Figure 3.4: The initial graph G(4).

At the initial time $t_0 = 4$, we construct the 9-fold of random duplication graphs, $G^*(4)$, with 9 identical copies of this initial graph $G(4)$.

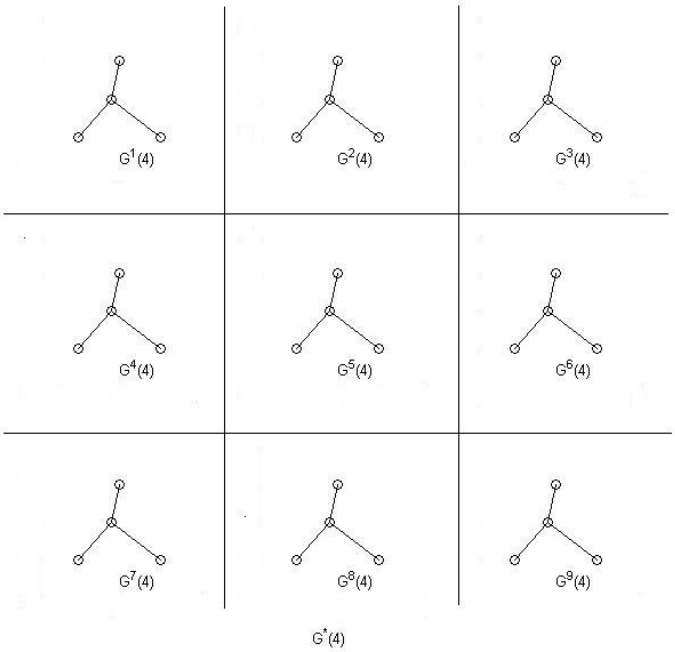


Figure 3.5: The 9-fold of random duplication graphs at time 4, $G^*(4)$.

After the initial time $t_0 = 4$, these 9 subgraphs evolves independently according to the rule of random vertex duplication, resulting in $G^*(5)$, the 9-fold of random duplication graphs at time 5.

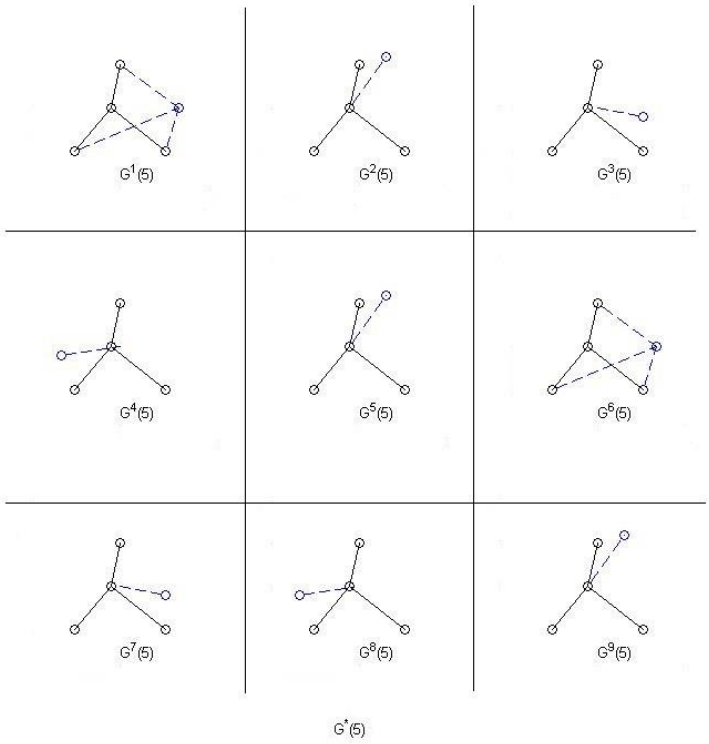


Figure 3.6: The 9-fold of random duplication graphs at time 5, $G^*(5)$.

The $n$-fold of random duplication graph could still depict very well the behavior of protein-protein interaction networks. In a matter of fact, the $n$-fold of random duplication graphs is just a partitioned random duplication graph model.

We define some random variables for our convenience. At time $t$, the number of degree $k$ vertices of subgraph $i$, $G^i(t)$, is denoted as $F_t^i(k)$, the degree distribution function of subgraph $i$, $G^i(t)$, is denoted as $\mathbb{P}_t^i(k)$. By definition,

$$\mathbb{P}_t^i(k) = \frac{F_t^i(k)}{t} \tag{3.27}$$

With the help of these auxiliary random variables, we hereby give the formal definition of the degree distribution function of the $n$-fold of random duplication graphs $G^*(t)$.

***Definition 3.3.2.*** *The degree distribution function of the $n$-fold of random duplication graphs $G^*(t)$ at time t is denoted as $\mathbb{P}_t^*(k)$, and written as*

$$\mathbb{P}_t^*(k) = \frac{\sum_{i=1}^{n} F_t^i(k)}{nt} = \frac{\sum_{i=1}^{n} \mathbb{P}_t^i(k)}{n} \tag{3.28}$$

Now we prove that the degree distribution function of the $n$-fold of random duplication graphs, $\mathbb{P}_t^*(k)$, converges to the expected value in Equation 3.25.

***Theorem 3.3.1.*** *The degree distribution function of the $n$-fold of random duplication graphs converges in probability as $n \to \infty$. More precisely, for $\forall t$ and $\forall \varepsilon$,*

$$\lim_{n\to\infty} Prob\left(\sup_k\{|\mathbb{P}_t^*(k) - \mathbb{E}[\mathbb{P}_t(k)]|\} > \varepsilon\right) = 0 \tag{3.29}$$

***Proof.*** *According to the weak law of large numbers for independent and identically distributed random variables, for $\forall t$, $\forall n$, and $\forall \varepsilon$,*

$$\lim_{n\to\infty} Prob\left(\left|\frac{\sum_{i=1}^{n} \mathbb{P}_t^i(k)}{n} - \mathbb{E}[\mathbb{P}_t(k)]\right| > \varepsilon\right) = 0 \tag{3.30}$$

$$\lim_{n\to\infty} Prob(|\mathbb{P}_t^*(k) - \mathbb{E}[\mathbb{P}_t(k)]| > \varepsilon) = 0 \tag{3.31}$$

*That is, for $\forall t$, $\forall \varepsilon$, and $\forall \xi$, and for every $k$, $\exists N_k$ such that if $n > N_k$,*

$$Prob(|\mathbb{P}_t^*(k) - \mathbb{E}[\mathbb{P}_t(k)]| > \varepsilon) < \xi \tag{3.32}$$

*Let $N = \max_k\{N_k\}$, for $\forall t$, $\forall \varepsilon$, and $\forall \xi$, if $n > N$,*

$$Prob\left(\sup_k\{|\mathbb{P}_t^*(k) - \mathbb{E}[\mathbb{P}_t(k)]|\} > \varepsilon\right) < \xi \tag{3.33}$$

*For $\forall t$ and $\forall \varepsilon$,*

$$\lim_{n\to\infty} Prob\left(\sup_{k}\{|\mathbb{P}_t^*(k) - \mathbb{E}[\mathbb{P}_t(k)]|\} > \varepsilon\right) = 0 \tag{3.34}$$

*thus completing the proof of Theorem 3.3.1.* □

This theorem indicates that if we start from some initial graph $G(t_0)$, with high probability as $n \to \infty$, the $n$-fold of random duplication graphs will present a degree distribution function of Equation 3.25, such that

$$\mathbb{P}_t(k) \xrightarrow{in\ probability} \frac{1}{t}\sum_{j=max\{k-t+t_0;1\}}^{min\{k;t_0-1\}} \frac{\binom{k-1}{j-1}\binom{t-k}{t_0-j}}{\binom{t-1}{t_0-1}} F_{t_0}(j) \quad as\ n \to \infty \tag{3.35}$$

# Chapter 4

# Protein-Protein Interaction Networks as a Special Case of Random Duplication Graph

## 4.1 Degree Distribution Function of Protein-Protein Interaction Networks

This model of random duplication graph explicitly depicts the self-organization process of protein-protein interaction networks.

In our model of protein-protein interaction networks, proteins are represented as vertices, protein-protein interactions are represented by edges, and gene duplications are represented as random vertex duplication.

Since duplicated genes produce identical proteins that interact with the exact same protein partners, all the edges are copied during random vertex duplication. Consequently, we can model protein-protein interaction networks as a special case of random duplication graph.

We take a valid assumption that, at the beginning of the gene duplication process, proteins are sparsely connected. As a result, in $G(t_0)$, we let $F_{t_0}(1) = t_0$ and $F_{t_0}(i) = 0$ for $\forall i \in [2, t_0 - 1]$, i.e., each protein only has one connection. This assumption complies with the situation of the beginning of the biological evolution—a few different proteins gathered together, not a lot of connections were formed.

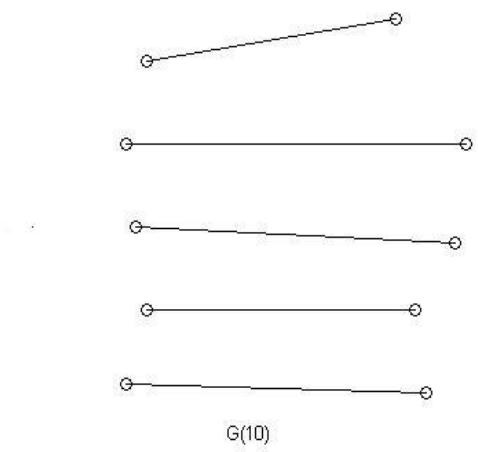


Figure 4.1: An example of the sparsely connected initial graph.

Based on this specific initial condition, solution to $\mathbb{E}[\mathbb{P}_t(k)]$ is derived by leaving only the term $j = 1$.

$$\mathbb{E}[\mathbb{P}_t(k)] = \frac{t_0}{t}\frac{\binom{t-k}{t_0-1}}{\binom{t-1}{t_0-1}} \quad for\ 1 \le k \le t - t_0 + 1 \tag{4.1}$$

To get a simpler expression of the degree distribution function, an upper bound and a lower bound are found for $\mathbb{E}[\mathbb{P}_t(k)]$.

***Corollary 4.1.1.*** *An upper bound and a lower bound exist for* $\mathbb{E}[\mathbb{P}_t(k)]$.

$$\frac{t_0}{t}\left(\frac{t-k-t_0+2}{t-1}\right)^{t_0-2} \le \mathbb{E}[\mathbb{P}_t(k)] \le \frac{t_0}{t}\left(\frac{t-k}{t-t_0+1}\right)^{t_0-2} \quad for\ 1 \le k \le t - t_0 + 1 \tag{4.2}$$

***Proof.*** *For* $1 \le k \le t - t_0 + 1$,

$$\mathbb{E}[\mathbb{P}_t(k)] \tag{4.3}$$

$$= \frac{t_0}{t}\frac{\binom{t-k}{t_0-1}}{\binom{t-1}{t_0-1}} \tag{4.4}$$

$$= \frac{t_0}{t}\frac{(t-k)(t-k-1)\cdots(t-k-t_0+2)}{(t-1)(t-2)\cdots(t-t_0+1)} \tag{4.5}$$

*It is easy to see from Equation 4.5 that*

$$\frac{t_0}{t}\left(\frac{t-k-t_0+2}{t-1}\right)^{t_0-2} \leq \mathbb{E}[\mathbb{P}_t(k)] \leq \frac{t_0}{t}\left(\frac{t-k}{t-t_0+1}\right)^{t_0-2} \quad for\ 1 \leq k \leq t-t_0+1 \tag{4.6}$$

*thus completing the proof of Corollary 4.1.1.* □

Furthermore, these two bounds are asymptotically tight as $t \to \infty$.

$$\lim_{t\to\infty}\frac{t_0}{t}\left(\frac{t-k-t_0+2}{t-1}\right)^{t_0-2} = \lim_{t\to\infty}\mathbb{E}[\mathbb{P}_t(k)] = \lim_{t\to\infty}\frac{t_0}{t}\left(\frac{t-k}{t-t_0+1}\right)^{t_0-2} \quad for\ 1 \leq k \leq t-t_0+1 \tag{4.7}$$

With the help of these bounds, a simple approximation to $\mathbb{E}[\mathbb{P}_t(k)]$ is obtained.

$$\mathbb{E}[\mathbb{P}_t(k)] \sim c\frac{t_0}{t}\left(\frac{t-k}{t-1}\right)^{t_0-2} \quad for\ some\ c, as\ n \to \infty \tag{4.8}$$

This approximation is sound because this approximation lies within the upper and lower bounds of $\mathbb{E}[\mathbb{P}_t(k)]$, and the bounds are asymptotically tight. In mathematical terms, we have the following results between the approximation and the bounds.

$$\frac{t_0}{t}\left(\frac{t-k-t_0+2}{t-1}\right)^{t_0-2} \leq \frac{t_0}{t}\left(\frac{t-k}{t-1}\right)^{t_0-2} \leq \frac{t_0}{t}\left(\frac{t-k}{t-t_0+1}\right)^{t_0-2} \tag{4.9}$$

$$\lim_{t\to\infty}\frac{t_0}{t}\left(\frac{t-k-t_0+2}{t-1}\right)^{t_0-2} = \lim_{t\to\infty}\frac{t_0}{t}\left(\frac{t-k}{t-1}\right)^{t_0-2} = \lim_{t\to\infty}\frac{t_0}{t}\left(\frac{t-k}{t-t_0+1}\right)^{t_0-2} \tag{4.10}$$

Figure 4.2 shows a comparison between $\mathbb{E}[\mathbb{P}_t(k)]$, the bounds, and the approximation, with parameters $t_0 = 25$ and $t = 200$. It is shown that $\mathbb{E}[\mathbb{P}_t(k)]$ is well approximated by Equation 4.8.

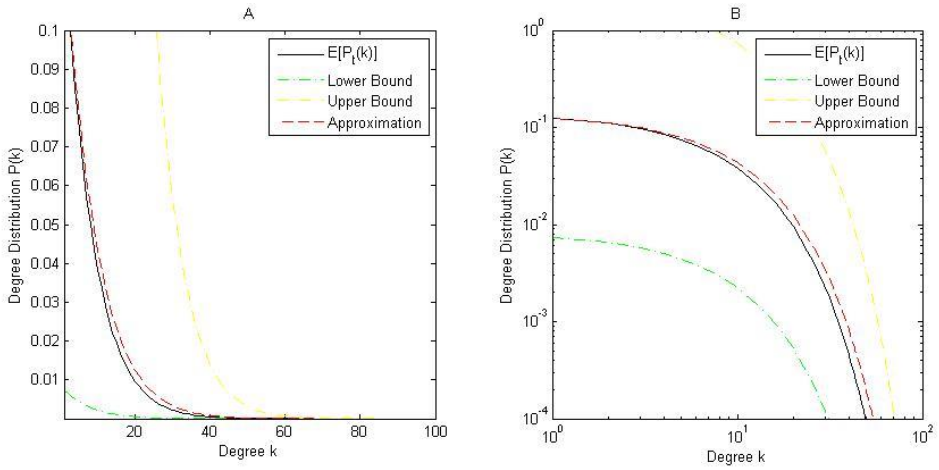


Figure 4.2: (A) The comparison between the degree distribution function, the bounds, and the approximation; (B) The comparison in log-log plot.

Replace the $t_0 - 2$ with a single parameter $t_1$, we propose a simple degree distribution function for protein-protein interaction networks, modeled as a special case of random duplication graph with a sparse initial graph.

$$\mathbb{E}[\mathbb{P}_t(k)] \sim c\left(\frac{t-k}{t-1}\right)^{t_1} \quad for\ some\ c \tag{4.11}$$

Note that Equation 4.11 is only the expected value of the degree distribution function, the convergence $\mathbb{P}_t(k) \to \mathbb{E}[\mathbb{P}_t(k)]$ is only valid under the $n$-fold.

In summary, we present our model of protein-protein interaction networks. We start from a sparsely connected initial graph, and devise a graph that contains n such subgraphs. As time advances, let the $n$ subgraphs evolves independently according to the rule of random vertex duplication. The resulting graph is our model of protein-protein interaction network. Our model of protein-protein interaction networks is shown in Figure 4.3.

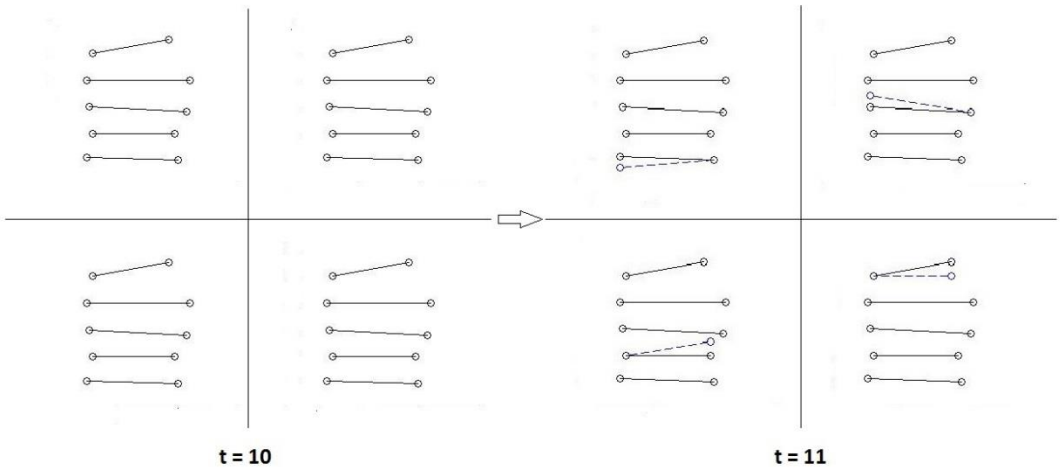


Figure 4.3: Our model of protein-protein interaction networks. We start from $n$ identical sparsely-connected initial graph, and let the $n$ subgraphs evolve independently according to the rule of random vertex duplication.

In addition, the degree distribution function of the protein-protein interaction networks is given in Equation 4.12.

$$\mathbb{P}_t^* \sim c\left(\frac{t-k}{t-1}\right)^{t_1} \quad for\ some\ c \tag{4.12}$$

As a result, we have hereby derived the degree distribution function of protein-protein interaction networks. Two parameters, $t_1$ and $t$, stand for the initial scale of the network, and the timesteps taken during the self-organization process respectively.

## 4.2 The Behaviors of the Degree Distribution Function

In this section we explore the behaviors of protein-protein interaction networks with the help of our degree distribution function in Equation 4.12.

First, Figure 4.4 shows the degree distribution function with parameters $t_1 = 27$ and $t = 200$. It is shown that our degree distribution function can indeed illustrate the degree distribution pattern of protein-protein interaction networks, where most vertices have relatively low degree, while high-degree vertices exist.

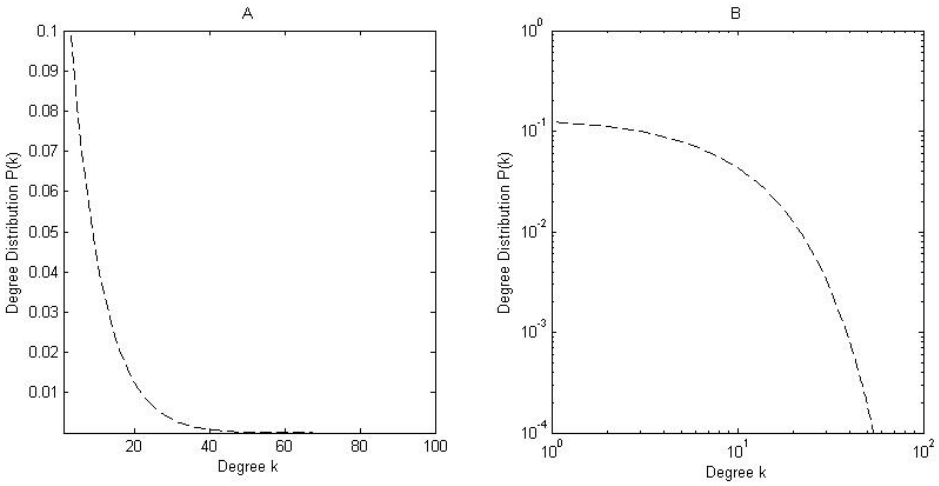


Figure 4.4: (A) Our degree distribution function with parameters $t_1 = 27$ and $t = 200$; (B) Log-log plot.

Furthermore, Figure 4.5 shows the comparison between the degree distribution functions in Equation 4.12 at time $t = 200$ and $t = 400$ (with $t_1 = 27$).

We can see that as time $t$ increases from 200 to 400, the percentage of high-degree vertices becomes larger.

We can make an important prediction based on the behavior of the degree distribution function—as the gene duplication process proceeds, the percentage of densely-connected proteins is higher. In other words, in ancient living organisms, the degree distribution pattern of their protein-protein interaction networks should be similar to the case $t = 200$ in Fig. 8. Whereas in modern living organisms, the pattern should be similar to the case $t = 400$.

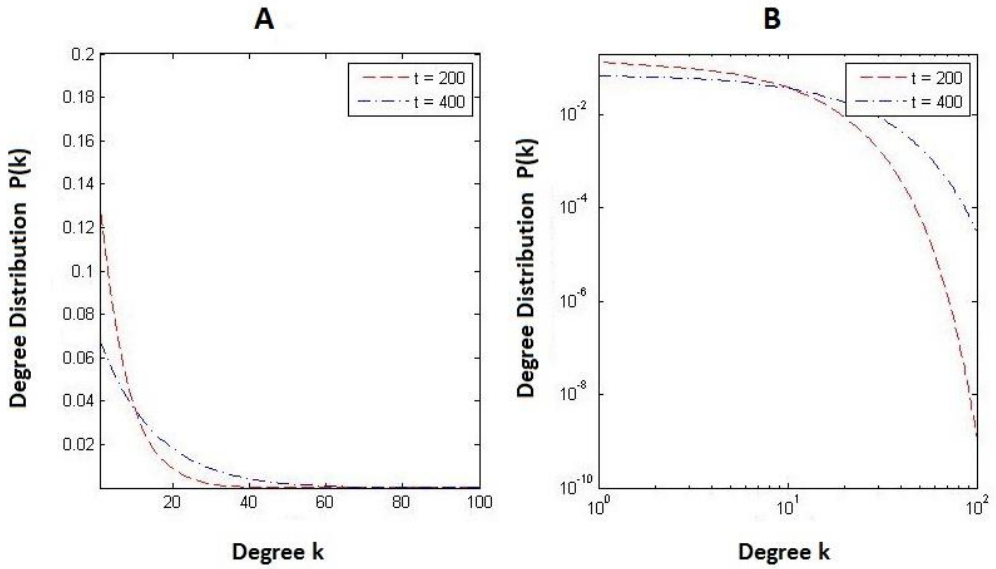


Figure 4.5: (A) The degree distribution function at time $t = 200$ and $t = 400$ respectively; (B) Log-log plot.

In addition, the two parameters, $t_1$ and $t$, in Equation 4.11 can be used to predict the initial scale of the network, and the timesteps taken during the self-organization process respectively.

## 4.3 Comparison with Degree Distribution Data

First, the degree distribution data of the protein-protein interaction network of Drosophila Melanogaster can be fitted by our degree distribution function with parameters $t_1 = 2600$, and $t = 7048$. The $r^2$ of the fit is greater than 0.96.

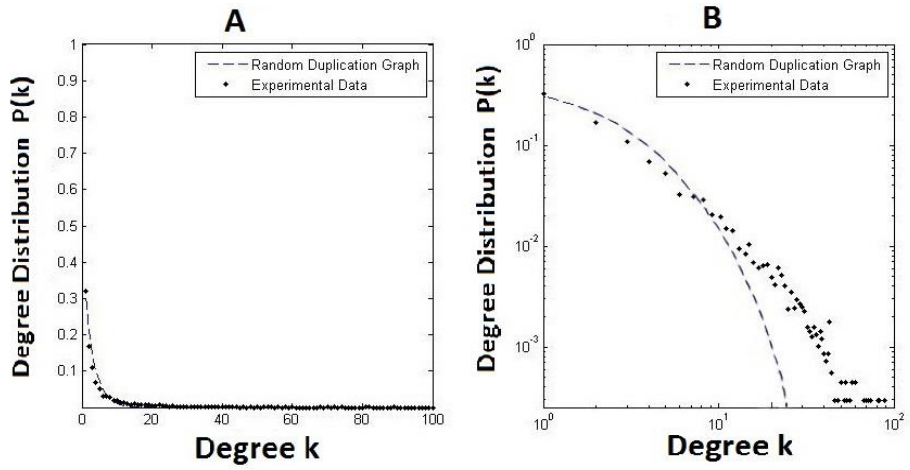


Figure 4.6: (A) Comparison between the degree distribution data of the protein-protein interaction network of Drosophila Melanogaster and our degree distribution function; (B) Log-log plot.

Next, we compare our degree distribution function with the degree distribution data of the protein-protein interaction network of Saccharomyces cerevisiae [20].

The parameters $t_1 = 410$, and $t = 2999$ provide the best fit, where $r^2$ is greater than 0.97.

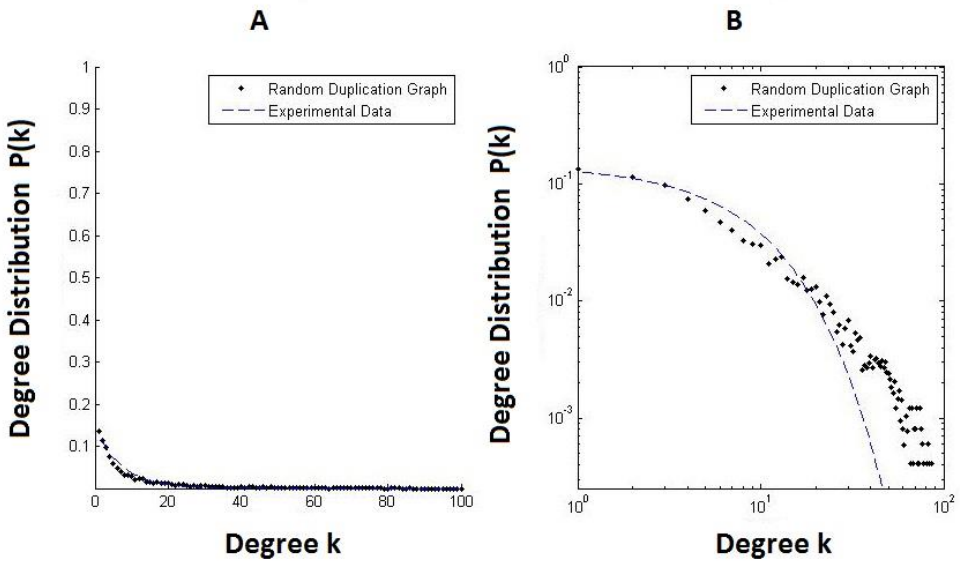


Figure 4.7: (A) Comparison between the degree distribution data of the protein-protein interaction network of Saccharomyces cerevisiae and our degree distribution function; (B) Log-log plot.

Then, we compare our degree distribution function with the degree distribution data of partial human protein-protein interaction network [21]. The parameters $t_1 = 3000$, and $t = 4825$ provide the best fit, where $r^2$ is greater than 0.96.

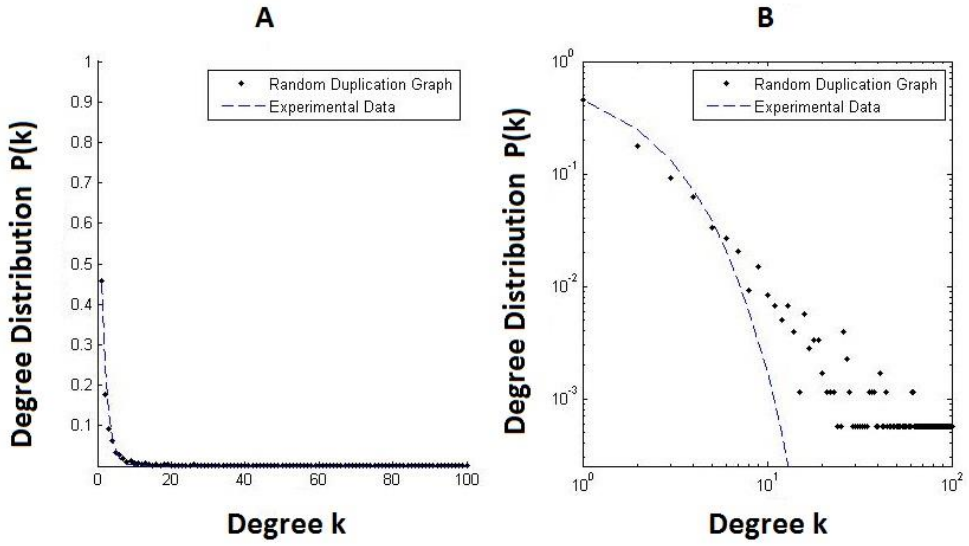


Figure 4.8: (A) Comparison between the degree distribution data of partial human protein-protein interaction network and our degree distribution function; (B) Log-log plot.

As we can see, the degree distribution data can be fitted adequately by our degree distribution function. Our degree distribution function indeed resembles the degree distribution pattern of protein-protein interaction networks, thus we can conclude that we have devised an appropriate model for protein-protein interaction networks. Furthermore, we have shown that it is the gene duplication process combined with the sparsely-connected initial condition that leads to the unique degree distribution pattern in protein-protein interaction networks.

# Chapter 5

# Conclusions and Future Work

## 5.1 Conclusions

Motivated by explaining the degree distribution pattern of protein-protein interaction networks, we try to model protein-protein interaction networks as random duplication graphs. We are particular interested in explaining why protein-protein interaction networks present a unique degree distribution pattern that most of the proteins are sparsely connected, while densely-connected proteins also exist. We intend to use random duplication graph model to find out the cause to such degree distribution pattern.

To find an approach of our research, we give a review of early models of complex networks, especially the Erdös-Rényi random graph model. We find that the previous approach to derive the degree distribution function of Erdös-Rényi random graph model is wrong. Also, it occurs to us that it is difficult to discuss the convergence of degree distribution function of a single Erdös-Rényi random graph.

When it comes to the random duplication graph model, we derive the expected degree distribution function through the probability master function. Instead of discussing the convergence of degree distribution function of a single random duplication graph, we propose the $n$-fold of random duplication graphs, a combination of $n$ independent random duplication graphs, under which we are able to prove the convergence of the degree distribution function.

Furthermore, we model the protein-protein interaction networks as a special case of random duplication graph with sparse initial graph. The degree distribution function of protein-protein interaction networks is derived under this model, and compared with degree distribution data of reconstructed protein-protein interaction networks. It is shown that our degree distribution function can provide a good fit with the degree distribution data. Moreover, we have shown that it is the gene duplication process combined with the sparsely-connected initial condition that leads to the unique degree distribution pattern in protein-protein interaction networks. One further prediction can be made based on our analysis—the longer the gene duplication process is, the more densely-connected proteins will be found in the protein-protein interaction network.

## 5.2 Future Work

During our review of the Erdös-Rényi random graph model, we find that the previous approach to derive the degree distribution function is fallacious. In a matter of fact, we cannot prove if Erdös-Rényi random graph model has a single converging degree distribution function or not, indicating that the Erdös-Rényi random graph model could result in multiple final states with different degree distribution function.

If we were to prove that Erdös-Rényi random graph model has a single converging degree distribution function, we would have to prove a stronger version of weak law of large numbers for dependent random variables. The weak law of large numbers we need must provide better conditions than the Bernstein's Theorem. Furthermore, much work remains to be done in the necessary and sufficient conditions for the weak law of large numbers of dependent random variables to hold.

If we were to accept the claim that the degree distribution function of Erdös-Rényi random graph model does not converge, it would be necessary to study the behavior of the degree distribution function beyond the law of large numbers. In this case, the Erdös-Rényi random graph model will result in multiple different degree distribution functions, we hope to understand what the probability distribution of the degree distribution function will be.

In addition, we need to study the Barabási-Albert model more rigorously. Till today there is no proof that Barabási-Albert model results in a scale-free distribution function. Similarly, the convergence of degree distribution function of the Barabási-Albert model need to be discussed.

When it comes to the random duplication graph model, we hope to invest into the behavior of the degree distribution function of a single random duplication graph, instead of the $n$-fold random duplication graphs, similar to the case of Erdös-Rényi random graph model.

# Letter of Copyright Permission

Figure 2.6 and Figure 2.7 are reused from [14], with copyright permission granted by AARS.

Figure 2.9 is reused from [18], with copyright permission granted by Cell Press. The Copyright License ID is 3675100171806.

# References


[1] A. L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nature Reviews*, vol. 5, pp. 101-113, 2004.

[2] L. Giot et al., “A Protein Interaction Map of Drosophila melanogaster,” *Science*, vol. 302, pp. 1727–1736, 2003.

[3] J. G. Gardeñes and L. Moreno, “From Scale-free to Erdos-Rényi Networks,” *Physics Review E*, vol. 73(5), pp. 056124, 2006.

[4] A. L. Barabási et al. “Mean-field theory for scale-free random networks,” *Physics A*, vol. 272, pp. 173-187, 1999.

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

[6] B. Bollobas. “Random Graphs,” *Academic Press*, London, 1985.

[7] E. Seneta, “A tricentenary history of the law of large numbers,” *Bernoulli*, vol. 19(4), pp. 1088-1121, 2013.

[8] W. L. Steiger, “Weak laws for dependent sums,” *Proceedings of the American Mathematical Society*, vol. 41, pp. 278-281, 1973.

[9] L. H. Hartwell et al., “From molecular to modular cell biology,” *Nature*, vol. 402, pp. C47–C52 , 1999.

[10] H. Kitano, “Computational systems biology,” *Nature*, vol. 420, pp. 206–210, 2002.

[11] R. Albert and A. L. Barabási, “Statistical mechanics of complex networks,” Rev. Mod. Phys., vol. 74, pp. 47–97, 2002.

[12] B. Viswanath et al., “On the evolution of user interaction in Facebook,” *Proc. Workshop on Online Social Networks*, pp. 37-42, 2009.

[13] “Facebook friendships network dataset,” *Konect*, 2015.

[14] A. L. Barabási and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, vol. 286(5439), pp. 509-512, 1999.

[15] S. H. Strogatz, “Exploring complex networks,” *Nature*, vol. 410(13), pp. 268–276, 2001.

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

[17] H. Kitano, “Computational systems biology,” *Nature*, vol. 420, pp. 206–210, 2002.

[18] K. G. Guruharsha et al. “A Protein Complex Network of Drosophila melanogaster,” *Cell*, vol. 147(3), pp. 690–703, 2011.

[19] E. V. Koonin et al., “The structure of the protein universe and genome evolution,” *Nature*, vol. 420, pp. 218–223, 2002.

[20] E.D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models, Boston*, MA: Springer, pp. 81-82, 2009.

[21] U. Stelzl et al., “A Human Protein-Protein Interaction Network: A Resource for Annotating Proteins,” *Cell*, vol. 122, pp. 957-968, 2005.