**İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY**

# PROPERTIES OF CONTENT-BASED NETWORKS

**Ph.D. Thesis by**

**Duygu BALCAN, M.Sc.**

**Department : Physics**

**Programme: Physics Engineering**

**MARCH 2007**

**İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY**

# PROPERTIES OF CONTENT-BASED NETWORKS

**Ph.D. Thesis by**
**Duygu BALCAN, M.Sc.**

**(509032101)**

**Date of submission : 8 February 2007**

**Date of defence examination: 2 March 2007**

**Supervisor (Chairman): Prof. Dr. Ayşe ERZAN**

**Members of the Examining Committee Prof. Dr. Ayşe Hümeyra BİLGE (İTÜ.)**

**Assoc. Prof. Dr. Nazmi POSTACIOĞLU (İTÜ.)**

**Assoc. Prof. Dr. Canan ATILGAN (SÜ.)**

**Asst. Prof. Dr. Muhittin MUNGAN (BÜ.)**

**MARCH 2007**

**İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ**

# İÇERİK-TEMELLİ AĞLARIN ÖZELLİKLERİ

**DOKTORA TEZİ**

**Y. Müh. Duygu BALCAN**
**(509032101)**

**Tezin Enstitüye Verildiği Tarih : 8 Şubat 2007**
**Tezin Savunulduğu Tarih : 2 Mart 2007**

**Tez Danışmanı : Prof. Dr. Ayşe ERZAN**
**Diğer Jüri Üyeleri Prof. Dr. Ayşe Hümeyra BİLGE (İ.T.Ü.)**
**Doç. Dr. Nazmi POSTACIOĞLU (İ.T.Ü.)**
**Doç. Dr. Canan ATILGAN (S.Ü.)**
**Yrd. Doç. Dr. Muhittin MUNGAN (B.Ü.)**

**MART 2007**

## DEDICATION

**To my grandmother** I would hereby like to dedicate my thesis to my grandmother. Since I am neither the best daughter nor the easiest person to live with, I wish that she forgives me in the case that I have ever hurt her. I also wish that she will be together with us for many more years. Babaanneciğim seni seviyorum.

**ACKNOWLEDGEMENTS**

If you have been working with the same supervisor for many years, she becomes your third mother. So it is not possible to summarize how grateful I am in one line. I would hereby like to thank to my supervisor Prof. Dr. Ayşe ERZAN for her very many contributions to my life, both in the scientific sense and personally. She has always guided me toward the directions where science becomes more interesting and integrated into actual life. Working with her has been a great luck and pleasure for me. I wish that she will be doing research and sharing her beautiful ideas and knowledge with young people for many more years.

I would also like to thank to Asst. Prof. Dr. Muhittin MUNGAN and Asst. Prof. Dr. Alkan KABAKÇIOĞLU for their collaboration in most of the research outlined in this thesis.

My brother (whom I love a lot much) has recently helped me a lot in the emotional sense and made my life easier. I would also like to thank to my parents for those they have/haven't been doing and for always being with me. I am very lucky that I have such a beautiful family. Annem, babam ve kardeşim sizi seviyorum.

MARCH, 2007 Duygu BALCAN

CONTENT

## ABBREVIATIONS

| | | |
|---|---|---|
| DNA | : | Deoxyribonucleic Acid |
| RNA | : | Ribonucleic Acid |
| mRNA | : | Messenger Ribonucleic Acid |
| TF | : | Transcription Factor |
| RS | : | Regulatory Sequence |
| PR | : | Promoter Region |
| GRN | : | Genetic Regulatory Network |
| TRN | : | Transcriptional Regulatory Network |
| RBD | : | Random Boolean Dynamics |
| RBN | : | Random Boolean Network |

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF SYMBOLS

| | | |
|---|---|---|
| $P_{\mathrm{in}}(d)$ | : | Probability of finding a node with $d$ incoming edges |
| $P_{\mathrm{out}}(d)$ | : | Probability of finding a node with $d$ outgoing edges |
| $P(d)$ | : | Probability of finding a node with $d$ nearest neighbors |
| $P_{\mathrm{c}}(1)$ | : | Probability of finding a realization with a single connected cluster |
| $d_{\mathrm{nn}}(d)$ | : | Average degree of nearest neighbors of nodes with degree $d$ |
| $c(d)$ | : | Average clustering coefficient of nodes with degree $d$ |
| $r(d)$ | : | Rich-club coefficient of nodes with degrees greater than $d$ |
| $\langle k_{\mathrm{max}} \rangle$ | : | Average number of $k$-shells |
| $n(k)$ | : | Number of nodes in $k$-shell |
| $e(k, k')$ | : | Number of links between $k$- and $k'$-shell |
| $p_{\mathrm{RS}}(l)$ | : | Probability of finding an RS of length $l$ |
| $p_{\mathrm{PR}}(l)$ | : | Probability of finding a PR of length $l$ |
| $d_{\mathrm{o},\,l}$ | : | Average out-degree of nodes with RSs of length $l$ |
| $d_{\mathrm{i},\,l}$ | : | Average in-degree of nodes with PRs of length $l$ |
| $P_{\mathrm{a}}(n_{\mathrm{a}})$ | : | Probability of finding a realization with $n_{\mathrm{a}}$ attractors |
| $P_{\mathrm{l}}(l_{\mathrm{a}})$ | : | Probability of finding an attractor of length $l_{\mathrm{a}}$ |
| $P_{\mathrm{s}}(s)$ | : | Probability of finding a basin of attraction of size $s$ |
| $P_{\mathrm{p}}(n_{\mathrm{p}})$ | : | Probability of finding a configuration with $n_{\mathrm{p}}$ precursors |
| $P_{\tau}(\tau)$ | : | Probability of finding a configuration with $\tau$ transient time |
| $x(t)$ | : | Average overlap between configurations at time $t$ |

## PROPERTIES OF CONTENT-BASED NETWORKS


### SUMMARY

The research we present in this thesis has been devoted to the modelling and understanding of transcriptional gene regulatory networks, on the basis of an information theoretical approach. Transcriptional gene regulation involves special proteins, namely the transcription factors, which bind to the DNA by recognizing specific subsequences, namely the transcription factor binding sites, embedded in them. We have modelled the transcriptional regulation network of yeast within this approach by associating random linear codes with the genes of the organism represented by nodes in our content-based network, and establishing edges between the nodes if and only if they share a certain amount of information, which has been realized via a sequence-matching rule. The distribution of the amount of shared information, which has been represented by the bitwise Shannon information of the random linear codes associated with the binding sequences and the promoter regions, are the most important biological inputs to our content-based model. We have made a very careful analysis of the transcriptional regulation networks of yeast, and compared their topological features with those of the ensemble of our content-based networks. We have observed that our content-based model is able to reproduce all the global topological features of these networks, which provides us with an understanding of their emergent nature. We conclude that the complex networks of gene regulation can arise spontaneously even with the random codes, so they do not need to be constructed from scratch by evolutionary mechanisms. We have also introduced the hidden-variable version of our content-based model involving only the pairwise connection probabilities as a function of the string lengths and observed that this model is able capture the main properties of our double-string model. So the analytical calculation on the hidden-variable model can provide us with making some predictions on the further properties of real networks. Very close topological similarities between the content-based models and genetic regulatory networks have led us to consider a modified random Boolean dynamics on our content-based networks, which we believe will help us with the understanding of the relationship between the architecture of the underlying network and the function of these systems. Our results point to further promising research problems in biological systems, where interactions between different components require the fulfillment of a series of constraints, which means the exchange of a certain amount of information. Examples are immune systems and protein interactions.

# İÇERİK-TEMELLİ AĞLARIN ÖZELLİKLERİ

## ÖZET

Burada sunulan tez çalısmasının ana teması transkripsiyon gen regülasyonu (düzenleme) çizgelerinin oluşumuna katkıda bulunan unsurların ve bu çizgelerin yapısal (topolojik) özelliklerinin enformasyon teorisi yaklaşımı ile modellenmesidir. Transkripsiyonel gen kontrolünde, transkripsiyon faktörleri olarak isimlendirilen proteinler DNA üzerinde özel alt dizilere bağlanarak, gen ifadesinin düzenlenmesine katkıda bulunmaktadırlar. Böyle bir proteinin tanıyıp bağlanabildiği DNA motiflerinin bilgi içeriğini başka bir alfabede ifade etmek mümkün olabilir. Bu yaklaşımla mayanın transkripsiyonel gen düzenleme ağını, içerik temelli ağın her bir düğümü bir gene karşılık gelmek üzere, her bir düğümüne gelişigüzel ikilik sistemde içerikleri olan diziler atayarak ve düğümler arasına, onlara atanan dizilerin birbirleri içerisinde tekrarlanma durumlarına göre, belli koşulları sağlamaları sonunda kenarlar yerleştirerek modelledik. Paylaşılan bilgi miktarının dağılımı modelimizin en önemli girdisi olup, ortaya çıkacak olan çizgenin özelliklerini tamamen belirlemektedir. Mayanının etkileşim ağını ayrıntılı biçimde inceleyerek, çizgenin yapısal özelliklerini içerik temelli modelimizin istatistiksel topluluğunun üyeleriyle karşılaştırdık. Gördük ki, içerik temelli modelimiz maya çizgesinin bütün özelliklerini barındırmakta ve bu tür ağların yapısal özelliklerinin anlaşılmasına imkan sağlamaktadır. Tamamen gelişigüzel dizilerden oluşturduğumuz içerik temelli çizgenin mayanın kontrol ağına yakınlığı, bu tür karmaşık ağ yapılarının evrim altında ereksel biçimde yoktan var edilmeleri gerekmedikleri sonucuna varmamıza neden olmaktadır. İçerik temelli modelimizin kabalaştırılması sonunda elde ettiğimiz ve (sadece dizi uzunluklarına bağlı) gizli-değişkenli olarak isimlendirilen modelin bizim içerik temelli modelimizi ve gerçek maya çizgesini yakından izleyen yapısal özellikleri nedeniyle, bu kaba model üzerinde yapılacak analitik hesapların düzenleme ağlarının yapılarıyla ilgili öngörülerde bulunabileceğini göstermektedir. İçerik temelli çizgelerin gen kontrol ağlarına yakınlıkları, gelişigüzel Boolean dinamiğini içerik temelli ağlara uyarlamamızı özendirmiştir. Bu yolla gen ifadesinin kontrol çizgelerinin topolojilerinin gen ifadesi dinamiği üzerindeki etkilerini anlamak mükün olabilir. Sonuçlarımız içerik temelli ağların bağışıklık sistemi yada protein etkileşimleri gibi çok sayıda koşulun yerine gelmesi sayesinde oluşan etkileşim ağlarının modellenmesi için elverişli olanaklar sunduğunu göstermektedir.

# 1 INTRODUCTION

Networks have become essential tools of researchers devoting themselves to the understanding of complex systems. Ecosystems, the brain, metabolic pathways, regulatory networks and immune systems, the internet and world wide web, economic systems, epidemics and social networks are among the numerous examples of complex systems. The features common to all these systems have found the possibility of exploration with the rise of network science which has brought a new global view into the study of complex systems.

Complex systems are organizations consisting of many heterogenous parts interacting locally and exhibiting emergent global behavior without any central organizing principle or control [1]. The emergence arises from the fact that the components of the system interact. The whole is more than the sum of its parts. The collective behavior arising from the interactions among the components, and the mapping from individual actions (which are relatively easy to describe) to the collective behavior is non-trivial [2]. Genetic regulatory networks might be the best examples of complex systems, where the expression profile of a gene is not determined by its genetic makeup but its interactions.

The main theme of the research presented in this thesis is that the topological features of networks based on information sharing are determined by the statistics of the shared information. The fact that certain biological networks, among them gene regulatory networks, operate on this principle has led us to make a detailed comparison of available data on the transcriptional regulatory network (TRN) of yeast, and the network which results from our model, given the relevant biological input consisting of the distribution of shared information. The strong similarity between the ensemble of various realizations of our model network and the yeast TRN confirms our hypothesis that complexity embodied in biological systems may arise simply due to the physical, chemical, etc., properties shared by the

constituent elements, and that complex interaction networks do not have to be fashioned from scratch by evolution. This view is strongly shared by a number of workers in the field. It has been forcefully and eloquently put forward by Richard Dawkins in *The Blind Watchmaker* [3] and by Stuart Kauffman in the *The Origins of Order* [4].

We have used the static structure of our content-based model to motivate a somewhat modified Random Boolean Network (RBN), whose dynamics we have investigated. We find that RBD on such networks possesses both the required properties of robustness and versatility needed to model gene regulation as a mechanism for phenotypic diversity at the cellular level.

In the next sections we supply some introductory material on the subjects we have tackled in this thesis. We summarize some of random network models and topological measures used to characterize complex networks in Section 1.1. The mechanisms of gene expression have been briefly discussed in Section 1.2, followed by a review of some earlier work on genetic regulatory networks in Section 1.3.

In Section 2, we introduce our content-based networks [5, 6] and summarize some of their topological properties. The results on the single-string model was published in [6], done in collaboration with Dr. Muhittin Mungan and Dr. Alkan Kabakçıoğlu. The first example of the double-string models, where the analytical calculation of degree distributions are carried out, was published [7] among the student papers of the Complex Systems Summer School at the Santa Fe Institute, done in collaboration with Dr. Brett Calcott and Dr. Paul Hohenlohe. The analytical calculations on the second example of the double-string models was guided by the research [8] done in collaboration with Prof. Ayşe Hümeyra Bilge.

We present our content-based model of the transcriptional regulatory network of yeast in Section 3, where we use the bitwise information content of binding motifs and the power-law form of intergenic regions as biologic input. The research we present in this section was done in collaboration with Dr. Muhittin Mungan and Dr. Alkan Kabakçıoğlu, and has been submitted to PLoSONE [9] for publication.

In Section 4, we introduce the hidden-variable version of our content-based model

networks and calculate some of the topological features of the networks analytically and compare them with simulation results. The research has been submitted to Chaos [10] for publication.

We provide an introduction to random Boolean dynamics, and present our content-based random Boolean dynamics that we have proposed on our content-based networks in Section 5. Some of the results presented here have been published in [11, 12].

We end up with a discussion in Section 6.

### 1.1 Networks: An Overview

Networks are collections of items represented by nodes (vertices) connected among themselves by edges (links) signifying interactions or physical contacts between these items. Recently, network science has found an indispensable place in the study of complex systems with the developments in mathematics, technology and computer sciences which have enabled researchers to collect, store, analyze and manipulate huge amount of data [2, 13, 14, 15, 16, 17]. However network theory goes back to the 18th century, attributed to Euler's solution of the Königsberg bridge puzzle [13]. The formulation of another social puzzle (so called, the six-degree separation) by Kochen and Pool in the 1950's, where the classical random graphs are defined, triggered two mathematicians Erdös and Rényi [18] to identify the properties of classical random graphs which are known by the names of these two mathematicians.

Topological properties of discrete objects such as graphs refer to the compactness and connectivity of a graph deducible from its adjacency matrix. For example, the number of connected components and the number of loops of a graph are topological invariants which are not affected by stretching or shrinking the links. In the context of network theory, topological properties have come to mean the degree distributions, the degree-degree correlation of nearest neighbors, the clustering coefficient, the rich-club coefficient, the $k$-core structure, etc. [14, 15, 16, 17] In this section we aim to summarize some quantifiers of network structures which

we will be using throughout this thesis.

### 1.1.1 Degree distributions

The degree $d$ of a node is defined as the number of nodes having an interaction with this node, i.e., the number of edges attached to it. The degree distribution $P(d)$ is the probability of encountering a node with degree $d$ if we pick a node at random. If the network is directed, then one distinguishes the out-degree $d_{\mathrm{o}}$ and in-degree $d_{\mathrm{i}}$ of a node (corresponding to the number of its out-going and in-coming edges) with their corresponding distributions. In this case, we may define the (total) degree of a node as the number of edges connecting this node with distinct nodes, i.e., $d = d_{\mathrm{o}} + d_{\mathrm{i}} - d_{\mathrm{b}}$ where $d_{\mathrm{b}}$ is the number of (bidirectional) edges pointing in both directions. In such networks the joint probability $P(d_{\mathrm{o}}, d_{\mathrm{i}})$, that a randomly chosen node has out-degree $d_{\mathrm{o}}$ and in-degree $d_{\mathrm{i}}$, completely determines the topological properties of the network in the absence of correlations [15].

The degree distributions have received a lot of interest after the discovery that many real-world networks representing a diverse class of systems deviate from classical random graphs in their degree distributions [16]. In classical random graphs, the nodes are connected to each other randomly and independently with a constant probability, thus they have binomial, or Poisson, degree distributions in the limit of large network sizes. We may characterize such networks with the average degree $\langle d \rangle$ of nodes, which is almost the degree of all the nodes in the network.

### 1.1.2 Deviations in degree distributions from classical random graphs

Very nice examples of this deviation from a Poisson distribution mentioned above are those networks whose degree distributions follow power-laws, $P(d) \propto d^{-\gamma}$. Such networks have been called *scale-free networks* [14], although in most cases it is only their degree distributions which are scale-free [16]. Other common forms of degree distributions are exponentials and power-laws with exponential cut-offs [16]. Another class of networks as we have posed recently, the content-based networks [5, 6, 9, 19], have also very distinct degree distributions with their broad

tails, although we have demonstrated that they can be thought of superpositions of Erdös-Rényi random graphs. In the case of scale-free networks, the data occurring in the tails of the distributions is very noisy. A common technique used here is plotting cumulative degree distribution $Q(d) = \sum_{d' \geq d} P(d')$, where one obtains another power-law, $Q(d) \propto d^{-(\gamma-1)}$.

### 1.1.3 Degree correlations

Assortative mixing [16] is the tendency of nodes with similar properties to be connected to each other. A special case of this tendency may be probed for the degrees if one thinks of them as the properties of nodes. If the nodes with similar degrees are connected to each other, then the networks are called *assortative*, and *disassortative* if not. Degree correlations [20] of nearest neighbors (connected pairs of nodes) may be measured by the conditional probability $p(d'|d)$ that randomly selected nearest neighbors of nodes with degree $d$ have degree $d'$ in an undirected network. Another measure [21] of the same property is the average degree $d_{\mathrm{nn}}(d)$ of nearest neighbors of nodes with degree $d$, $d_{\mathrm{nn}}(d) = \sum_{d'} d' p(d'|d)$. Since the latter quantity is much easier to compute via simulations and to display, it has found more use in the literature. One may easily generalize this concept for directed networks [15], where one may ask the variations of the question whether nodes with large out-degrees are preferentially connected to nodes with high in-degrees, etc.

### 1.1.4 Clustering coefficient

The average local density of edges between nearest neighbors of a node is called the clustering coefficient [22] of a network. The clustering coefficient $c_i$ of a node $i$ can be calculated as $c_i = 2\Delta_i / d_i(d_i - 1)$ where $d_i$ is the degree of the node and $\Delta_i$ is the number of those triangles containing this node and its nearest neighbors. If the degree of a node is less than two, then its clustering coefficient is equivalently zero. Then the average clustering coefficient $\langle c \rangle$ of the network is given by $\langle c \rangle = \sum_i c_i / N$, where $N$ is the total number nodes. We could also define the clustering coefficient [16] of a network by $\langle c \rangle = 3\Delta / N_\Delta$ where $\Delta$ is the number of triangles and $N_\Delta$ is the number of connected triples of nodes (those

nodes which are separated from each other by two edges) in the network. The difference between two definitions is that the first one is the average of ratios whereas the second one is the ratio of averages, so the former definition may give rise to a larger clustering coefficient. The latter quantity is easier to evaluate analytically whereas the first one is easier to calculate via simulations. We may as well determine the spectrum of the average clustering coefficient $c(d)$ as a function of degree [23, 24], $c(d) = \sum_i c_i \delta_{d_i,\, d}/N(d)$ where $N(d)$ is the number of nodes with degree $d$. Again we may generalize these definitions for the directed networks [5], where one can calculate the fraction of triangles with respect to the out-going and in-coming edges of nodes.

### 1.1.5 Deviations in correlations from generalized random graphs

It has been the custom to compare the topological properties of the network under consideration with those of the random graphs whose nodes follow the same degree distribution as the "target network". The randomness of the "control graphs" comes from the fact that the edges between pairs of nodes are established randomly and independently without respecting any properties of the nodes.

In random graphs, the probability $p(d'|d)$ of finding a node with degree $d'$ among the nearest neighbors of nodes with degree $d$ is independent from $d$, just depending on $d'$ and the average degree $\langle d \rangle$ of the nodes, viz., $p(d'|d) = d'P(d')/\langle d \rangle$. Thus, the average degree [25] of such nodes is $d_{\rm nn}(d) = \langle d^2 \rangle / \langle d \rangle$. A similar observation [25] is valid for the clustering coefficient $c(d)$, which, in the case of random graphs, has no dependence on the degree of the nodes, and is given by $c(d) = (\langle d^2 \rangle - \langle d \rangle)^2 / N \langle d \rangle^3$. By contrast, those of most real networks [14, 15, 16] display different dependencies on $d$.

### 1.1.6 Small-world effect

Imagine an undirected network, where we may define the geodesic distance $\ell_{ij}$ between a pair of nodes $i$ and $j$, as the smallest number of edges to be crossed to reach from one node to the other. Then the average shortest path length $\langle \ell \rangle$ of the network is calculated over all pairs of nodes, as $\langle \ell \rangle = 2 \sum_{i,\, j>i} \ell_{ij}/N(N-1)$

where $N$ is the size of the network and we have assumed that the network contains a single cluster. If the network contains more than one cluster, then one may calculate the inverse of the shortest path length, $\langle \ell^{-1} \rangle = 2 \sum_{i,\, j>i} \ell_{ij}^{-1} / N(N-1)$. If the average shortest path length scales with the logarithm of network size or slower, then it is said that the network exhibits the *small-world effect* [15, 16]. If the network is directed, then $\ell_{ij} \neq \ell_{ji}$, in general.

### 1.1.7 Robustness of networks with respect to damage

A network may contain disconnected parts, called the clusters or connected components of the network. If the relative size of the largest cluster stays finite as the network size increases, then it is said that the network is above the percolation threshold and this largest cluster is called the "giant connected component" of the network. If the network is directed then one distinguishes strongly and weakly connected components [15]; the latter are obtained by ignoring the directionality of edges. The resilience of networks against random removal of their nodes has gained a lot of interest, especially since this is important for the dynamical processes taking place on them. Although the removal of nodes has been extensively used as the main strategy here, other types of attacks have been also studied [16], such as removal of edges.

### 1.1.8 Rich-club ordering

The nodes with high degrees (i.e, a large number of edges) may be referred to as "rich," and the subgraph composed of such nodes with their interconnecting edges as the "rich-club". The rich-club coefficient [26, 27] is intended as a measure of well connectedness of "rich guys" among themselves. Denoting the number of nodes with degrees greater than $d$ by $N_{>d}$, and the number of edges between such nodes by $E_{>d}$, the rich-club coefficient [26] is given by $r(d) = 2E_{>d}/N_{>d}(N_{>d} - 1)$. The rich-club coefficient goes beyond the mixing property in a network; for example, a network displaying disassortative mixing can exhibit the rich-club property as well. For uncorrelated random graphs it has been shown [27] in the limit of infinite network size where the maximum degree tends to infinity, that $r(d) \sim d^2/\langle d \rangle N$ in the limit $d \to \infty$, where $\langle d \rangle$ denotes the average degree of nodes. The increase

observed for the rich-club coefficient even for random graphs made it necessary to compare the coefficient of the network at hand with that $r_{\text{rand}}(d)$ of the random version of the network. If $r(d) > r_{\text{rand}}(d)$ then the network is said to be exhibiting the rich-club property.

### 1.1.9 $k$-core structure

Nodes of a network may be classified with respect to some local or global properties. A global classification can be done via the $k$-core decomposition [28]. One can obtain the $k$-core by successively removing the nodes with degrees less than $k$, until the remaining nodes have degrees at least $k$. Let us note here that the $k'$-core with $k' > k$ is a subgraph of the $k$-core. The nodes belonging to the $k$-core but not to the $(k+1)$-core constitutes the $k$-shell. Thus, shells are distinct (containing different nodes). The last definition we want to give here is the $k$-crust, which is the subgraph containing all the shells with $k' \leq k$. Thus, the $k$-crust is the complement of the $(k+1)$-core. Recently, $k$-core decomposition has been used as an algorithm for the visualization of large scale networks [29] by Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat and Alessandro Vespignani. Their visualization can be used to distinguish between networks having very different organizational principles although the visualization by itself is not sufficient for the complete description of the network. The quantitative analysis [30, 31] of the $k$-core structure has been studied extensively and seems to be a promising way to understand the hierarchical organization of complex networks.

## 1.2 Transcriptional Gene Regulation in Eukaryotes

Regulation of gene expression in eukaryotes involves a diverse set of mechanisms including initiation of transcription, alternative splicing of RNA, mRNA stability, several forms of post-transcriptional modification, translational control, and protein degradation [32]. Among all, transcriptional initiation is the primary mechanism of gene expression, since it is the first check point of protein synthesis in a cell.

There are three main components of transcriptional regulation, $i$) DNA segments,

namely the promoter regions, usually occurring upstream of coding regions and acting as controlling elements in the expression of genes, *ii*) proteins, namely the transcription factors (gene regulatory proteins), which recognize and bind to specific sequences on the DNA and regulate the initiation of transcription, and finally, *iii*) the binding sites which are short DNA sequences where the regulatory proteins bind preferentially. [32, 33]

In eukaryotes, operons (sets of coding regions –loci– controlled by the same promoters) are not usual [33], thus we may assume that genes are regulated independently, in the sense that they are controlled by different promoter regions. Promoter regions can be thought as the computers of genes, collecting and analyzing the data about the status of the cell and altering the initiation of transcription. This data reaches promoter regions through transcription factors. The nucleotide sequences of transcription factor binding sites determine the transcription factors to be associated with the promoter region including these binding motifs. Therefore, the expression profile of a gene is determined by its promoter region as well as the expression of those genes which code the transcription factors recognized by the binding sites embedded in its promoter region.

Although the number of binding sites in a promoter region is not known exactly, there are between 10-50 binding sites according to well-studied eukaryotic promoters [33]. Most transcription factors may bind to several distinct sequences with different affinities. Differences in binding affinities may be more important if a binding motif (site) is recognized by more than one transcription factor, or if two binding sites are located nearby or overlap. Most binding motifs influence the expression of a single gene. However there can be cases where the same binding site regulates the expression of paralogous loci located on the opposite strands of DNA [33].

Transcription factors have several distinct domains including DNA-binding, protein interaction and ligand binding domains. DNA-binding domains are typically short sequences (roughly up to 20 base pairs) and are highly conserved evolutionarily [33]. There may be several DNA-binding domains in a transcription factor. As well as the transcription factors, the cofactors which are proteins interacting

with transcription factors, are also important in the regulation of gene expression. Ligands can also bind to transcription factors and alter their activity. It is also common in eukaryotes that regulatory proteins can bind to DNA at very distant locations from the promoter regions of genes and regulate their expression by looping out the intervening DNA [32].

## 1.3 Genetic Regulatory Networks

Genetic regulatory networks are directed graphs, where each node represents a gene and the directed link from Gene A to Gene B signifies that regulatory interaction in which the expression of Gene A controls the expression of Gene B. The development of efficient experimental techniques [34] has made a large amount of data on gene interactions [32, 33] available [35, 36, 37, 38], which reveals a complex and highly specific network. The organizational principles underlying these genetic regulatory networks are of great experimental [35, 39, 40, 41] and theoretical [42, 43, 44, 45, 46, 47] interest.

The degree distributions [39, 40] in genetic regulatory networks have been the main object of both empirical and network theoretical approaches. Barabasi and co-workers [48] have claimed that the global properties of genetic regulatory networks of *Saccharomyces cerevisiae* and *Escherichia coli*, as well as protein-protein interaction and metabolic networks, can be understood in terms of the growth mechanism [44] of these networks and can be modelled by the preferential attachment [43] rule, thus they are scale free, with the degree distribution having a scaling exponent $\gamma \sim 2$, which they claim to find from experimental results [48]. Smaller exponents, in the vicinity of 1.5 have been reported in the literature [35, 40]. It has been suggested that the degree distribution might in fact have a universal scale-free behavior independent of any particular organism [49]. Guelzim et al. [39] have made a careful analysis of the transcriptional regulatory network of yeast, revealing that the in- and out-degree distributions are rather different, with the former having an exponential-like decay and being confined to a much narrower range.

It should be also mentioned that the idea of using linear codes to model a broad

set of requirements for the binding of proteins to other molecules, as embodied in our sequence matching rule, has a quite long standing history. Complementarity of binary sequences of fixed uniform length representing anticores and the antigens which "recognize" them have been employed in modelling immune networks in the early 1990's [50], although the emphasis at this stage was more on the dynamics of small networks constructed in this way, rather than on their topological features. There have also been several earlier studies of models of gene regulatory networks on rather elaborate "Artificial Genomes" (AG) [51] based on various alphabets and matching rules [52, 53, 54, 55], some of them coupled with the duplication and divergence model introduced by Wagner [56, 57, 58]. The results are not uniform and depend on the detailed assumptions made in the models.

# 2 A BRIEF HISTORY OF CONTENT-BASED NETWORKS

The term "content-based" refers to the fact that the nodes of the model networks contain information represented by linear codes and the interactions between them are established conditional to the sharing of a certain amount of information. In this section we summarize the single-string models and then introduce a model where two different strings, with specialized functions, are associated with each node. We introduce and summarize global topological properties of content-based networks [5, 6, 9, 19] proposed as null models of regulatory interactions. This is followed by a discussion on the validity of effective-medium type of analytical calculations of the connection probabilities and topological properties. We also provide a section on our information theoretical approach to interaction networks, and end up with our calculations on the bitwise information contents of linear codes represented in an arbitrary alphabet.

## 2.1 Single-String Models

In our original content-based model [5, 6] first proposed as a toy model of RNA interference [59, 60, 61, 62], an artificial chromosome of fixed length $L$ is constructed randomly whose characters are chosen from an alphabet of $r+1$ letters according to the distribution

$$P(x) = (1-q)\,\delta_{x,\,r} + \frac{q}{r}\,\sum_{a=0}^{r-1}\delta_{x,\,a} \quad , \tag{2.1.1}$$

where the character "$r$" represents the delimiters and $1-q$ the probability of finding a delimiter along this linear code. The sequences between successive occurrence of the delimiter are associated with genes corresponding to the nodes of our content-based network. Thus in fact, the linear codes associated with the genes are chosen from an alphabet of size $r$ whose letters have an equal chance $1/r$ to occur in a random sequence. The directed interactions between pairs of nodes/genes are established with respect to the sequence-matching rule. If the se-

quence $G_i$ associated with the $i$th node occurs as an uninterrupted subsequence in the linear code $G_j$ associated with the $j$th node, then a directed link from the $i$th node to the $j$th node is drawn. Setting $w_{ii} = 0$, we may write the element $w_{ij}$ of adjacency matrix as

$$w_{ij} = \begin{cases} 1 & \text{if } G_i \subset G_j \\ 0 & \text{otherwise} \end{cases} , \tag{2.1.2}$$

where one should note that the length $l_i$ of the first sequence has to be smaller than or equal to the length $l_j$ of the second sequence. Thus, if $l_i > l_j$ then $w_{ij} = 0$ identically. We should also note that $w_{ij} \neq w_{ji}$, in general. If $w_{ij} = 1$ then one may easily predict that $w_{ji} = 0$ unless the sequences are identical; in this case, $w_{ij} = w_{ji} = 1$. Another property following from the definition in Eq. 2.1.2 is the transitivity property that if $G_i$ is embedded in $G_{i'}$ and if $G_{i'}$ is embedded in $G_j$, then we know for sure that $G_i$ is also embedded in $G_j$. So in terms of the elements of the interaction matrix, if $w_{ii'}w_{i'j} = 1$, then $w_{ij} = 1$ identically.

With the definition in Eq. 2.1.1 the length distribution $p(l)$ of sequences associated with nodes along the artificial chromosome is of exponential form $p(l) \propto q^l$. It is possible to obtain an ensemble of sequences following a predetermined length distribution [19] by realizing a chromosome with successive assignments of lengths of the sequences from the desired length distribution and choosing the characters of the sequences from an alphabet of size $r$ with identically distributed letters, then placing a delimiter just next to the position of the last letter of the previously generated sequence on the chromosome. One may easily observe that although the number of nodes (the sequences of nonzero length) fluctuates from one realization of the chromosome to the other, the construction of an artificial chromosome affords more possibilities to employ evolutionary procedures, such as transposition as well as duplication and divergence [19]. We may also construct our content-based network by considering a fixed number $N$ of nodes where we associate a linear code with each node whose content and length are chosen from the desired distributions. The interactions between the nodes of the network is again established with respect to the sequence-matching rule (see Eq. 2.1.2).

The ensemble of networks constructed as defined above, even with null assumptions for the length distributions, exhibits very distinct topological properties

common to some real complex networks such as being of small-world type, having long tailed out-degree distributions, and displaying high resilience to random removal of nodes [5]. Moreover the networks are tractable analytically [6, 19] under some assumptions leading to the calculation [6] of the connection probability $p(l,k)$,

$$p(l,k) = 1 - \left(1 - \frac{1}{r^l}\right)^{k-l+1} \quad , \tag{2.1.3}$$

that an exact match occurs between randomly chosen pairs of sequences of lengths $l$ and $k \geq l$. This result should be considered as a zeroth order approximation because it has been obtained by assuming that all the sequences of same length are equivalent in their sequence-matching probabilities (effective-medium approximation) and ignoring the correlations between subsequences in the linear code forming the search space (which we can think of as a mean-field approach). Under these simplifying assumptions one may write

$$p(l,k) = \sum_{n=1}^{k-l+1} \binom{k-l+1}{n} \left(\frac{1}{r^l}\right)^n \left(1 - \frac{1}{r^l}\right)^{k-l+1-n} \quad , \tag{2.1.4}$$

where each of $n$ trials of the sequence-matching condition is assumed to have the same chance $1/r^l$ to be satisfied without taking into account the overlapping subsequences of length $l$ in the sequence of length $k$. The result is Eq. 2.1.3.

The out- and in-degree distributions are superpositions of binomial distributions which may be approximated [6] by Gaussian distributions in the limit of very large number of nodes,

$$P_{\rm out}(d) = \sum_l p(l) P_l^{\rm out}(d) \quad , \tag{2.1.5}$$

$$P_{\rm in}(d) = \sum_l p(l) P_l^{\rm in}(d) \quad , \tag{2.1.6}$$

where $P_l^{\rm out}(d)$ and $P_l^{\rm in}(d)$ are the out- and in-degree distributions of nodes with sequences of length $l$. They can be approximated by Gaussians with the means $d_{{\rm o},\,l}$ and $d_{{\rm i},\,l}$,

$$d_{{\rm o},\,l} = N \sum_{k \geq l} p(k) p(l,k) \quad , \tag{2.1.7}$$

$$d_{{\rm i},\,l} = N \sum_{k \leq l} p(k) p(k,l) \quad , \tag{2.1.8}$$

and the variances $\sigma^2_{\mathrm{o},\,l}$ and $\sigma^2_{\mathrm{i},\,l}$,

$$\sigma^2_{\mathrm{o},\,l} = N \sum_{k \geq l} p(k)p(l,k)[1 - p(k)p(l,k)] \ , \tag{2.1.9}$$

$$\sigma^2_{\mathrm{i},\,l} = N \sum_{k \leq l} p(k)p(k,l)[1 - p(k)p(k,l)] \ . \tag{2.1.10}$$

One should note here the differences in the probabilities and the sets of sequences over which the summations are performed. In the calculation of the average out-degree $d_{\mathrm{o},\,l}$ and its variance $\sigma^2_{\mathrm{o},\,l}$ we sum over all the nodes with length $k \geq l$, whereas in the calculation of the average in-degree $d_{\mathrm{i},\,l}$ and its variance $\sigma^2_{\mathrm{i},\,l}$ we consider all the nodes with length $k \leq l$.

We display in Fig. 2.1, the out- and in-degree distributions obtained via the simulations of an artificial chromosome and the distributions given in Eqs. (2.1.5, 2.1.6) [6] to give an insight into the global topological properties of the ensemble of content-based networks. We observe that although the theoretical curves capture the main characteristics of the distributions, the analytical solutions deviate from the simulation results in the large degree region of the out-degree distribution (see Fig. 2.1a, and Fig. 2.2 for better comparison) and in the small degree region of the in-degree distribution (see Fig. 2.1b). The differences come from the "mean-field" approximations used in the calculation of the sequence-matching probability (see Eq. 2.1.3), which leads also to the assumption that all the nodes of equal length follow the same out- and in-degree distributions (see Eqs. (2.1.5, 2.1.6)). It turns out that the fine structure [7, 8, 63, 64] due to the contents of the sequences should be taken into account for better approximations. We postpone this discussion to Section 2.3 where the fine splitting in degree distributions is demonstrated via naive examples. We should note here that since both types of interactions of a node are determined by the same linear code in this model, the out- and in-degrees of nodes are anti-correlated. If the number of out-going edges of a node is very large then one may easily predict that the number of its in-coming edges is small.

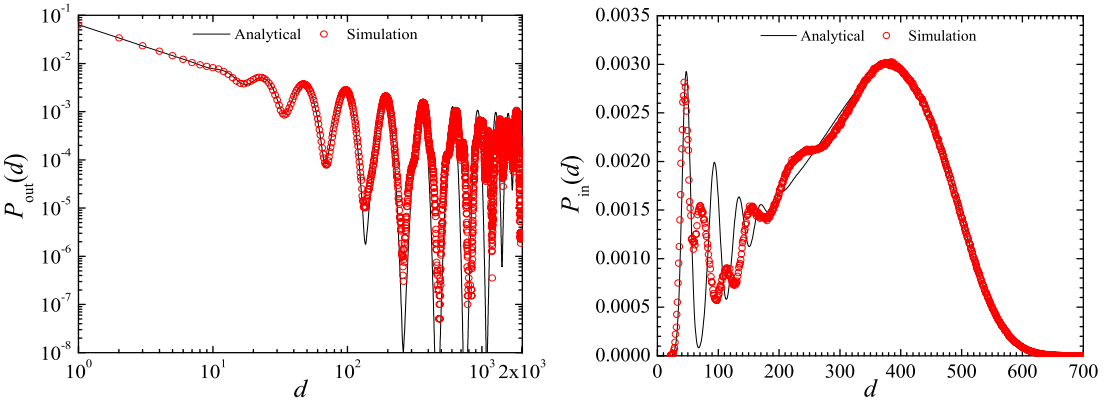


Figure 2.1: The directed degree distributions as obtained by our analytical solutions and simulations (red circles). The data points coming from our simulations have been obtained for the ensemble of content-based networks by averaging over $2 \times 10^4$ realizations of an artificial chromosome of length $4 \times 10^4$. The sequences between delimiters are random binary linear codes (thus, $r = 2$) following an exponential length distribution $p(l) \propto q^l$ with $q = 0.95$, within the interval $1 \leq l \leq 351$. The analytical results come from superpositions of Gaussian distributions centered around average degrees of sequences of different lengths (see Eqs. (2.1.7, 2.1.8)). (a) The out-degree distribution displays a continuous regime followed by well separated peaks corresponding to sequences of small lengths. (b) The in-degree distribution is much more localized comparing to the out-degree distribution.

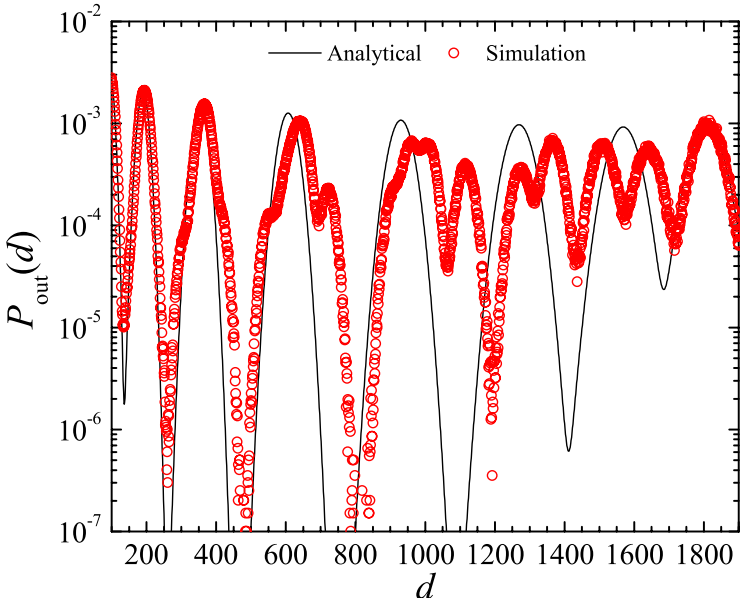


Figure 2.2: The large degree region of the out-degree distribution displayed in Fig. 2.1a has been re-plotted in the log-linear scale to allow better comparison of the results of simulations and analytical calculations. The deviations observed here are due to the fine structures of the sequences ignored in the calculation of the connection probability in Eq. 2.1.3.

## 2.2 Double-String Models

In the double-string model [7, 9], we associate two random sequences $G_i^{\text{key}}$ and $G_i^{\text{lock}}$ with each node $i$ of the content-based network of size $N$. The lengths of these sequences are chosen from different length distributions $p_{\text{key}}(l)$ and $p_{\text{lock}}(k)$, in general, whereas their contents are constructed randomly and independently from a common alphabet with identically distributed $r$ letters. The directed edges between pairs of nodes are established according to the sequence-matching rule. If the sequence $G_i^{\text{key}}$ associated with the node $i$ exactly matches a subsequence in $G_j^{\text{lock}}$ associated with the node $j$ then a directed edge from the first node to the second is drawn. Then the element $w_{ij}$ of adjacency matrix is given by

$$w_{ij} = \begin{cases} 1 & \text{if } G_i^{\text{key}} \subset G_j^{\text{lock}} \\ 0 & \text{otherwise} \end{cases} . \tag{2.2.11}$$

Note here that self-interactions ($w_{ii} = 1$) are also possible, as distinct from the single-string model. Another important difference coming with the double-string association is that the transitivity property exhibited in the single-string model has been lost, $w_{ii'}w_{i'j} = 1$ does not imply $w_{ij} = 1$ any more.

The tags "key" and "lock" have been used to distinguish the two specialized sequences associated with each node. This signifies that the content-based model discussed here is intended to model networks of regulatory interactions where each node "recognizes" nodes via its key-sequence and is "recognized" by other nodes through its lock-sequence. In the case of transcriptional regulatory networks, the key-sequences correspond to the binding motifs of the transcription factors and the lock-sequences to the promoter regions. The length distributions of these sequences totally determine the topological properties of the content-based network.

### 2.2.1 Simulation results for generic length distributions

We demonstrate some topological features of the ensemble of content-based networks assuming generic length distributions used for the random Boolean dynamics we have employed on these networks presented in Section 5. The binary key- and lock-sequences have been assumed to follow the same length distribution $p(l)$

confined within the interval $1 \leq l \leq 25$, either an exponential $p(l) \propto q^l$ with $q = 0.9$ or Gaussian $p(l) \propto \exp[-(l - \langle l \rangle)^2/2\sigma^2]$ with $\langle l \rangle = 13$ and $\sigma^2 = 50$.

In Fig. 2.3, we display the probability of finding a single connected cluster in a random realization of the network, as a function of system size. The networks almost certainly consist of a single cluster for $N \geq 180$ for the exponential length distribution and for $N \geq 1400$ for the Gaussian case. According to our observations we can say that the model networks are very resilient to random removal of nodes. Although the total degree distributions (see Fig. 2.6) of the model networks do not follow power-law forms, they have no percolation threshold, as in the case of scale-free networks [14] with exponent $\gamma \leq 3$.

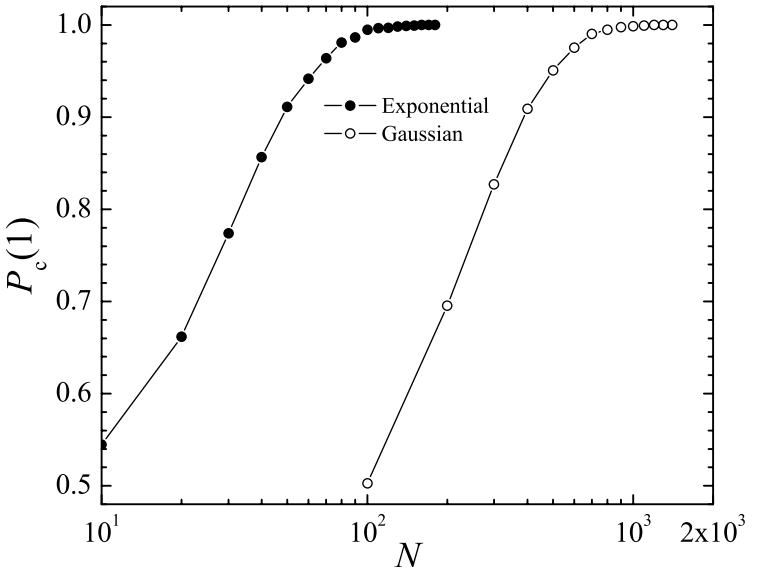


Figure 2.3: The probability $P_c(1)$ of finding a single connected cluster in a random realization of the model either with an exponential or a Gaussian length distribution, as the system size is increased. The data points have been obtained by generating $10^4$ realizations of the sequences.

In Figs. (2.4-2.6) we display the out-, in- and total degree distributions. Although the out-degree distributions exhibit very similar characteristics in both cases, having a continuous regime, followed by well separated peaks corresponding to key-sequences of small lengths, we observe very fine differences in the large degree regions (see Fig. 2.4). The differences due to the forms of the length distributions become more visible in the in- (and consequently the total) degree distributions. The in-degrees are distributed in much narrower intervals compared to the out-

degrees (see Fig. 2.5). In Fig. 2.6 we show the total degree distribution which, in general, is not the superposition of in- and out-degree distributions.

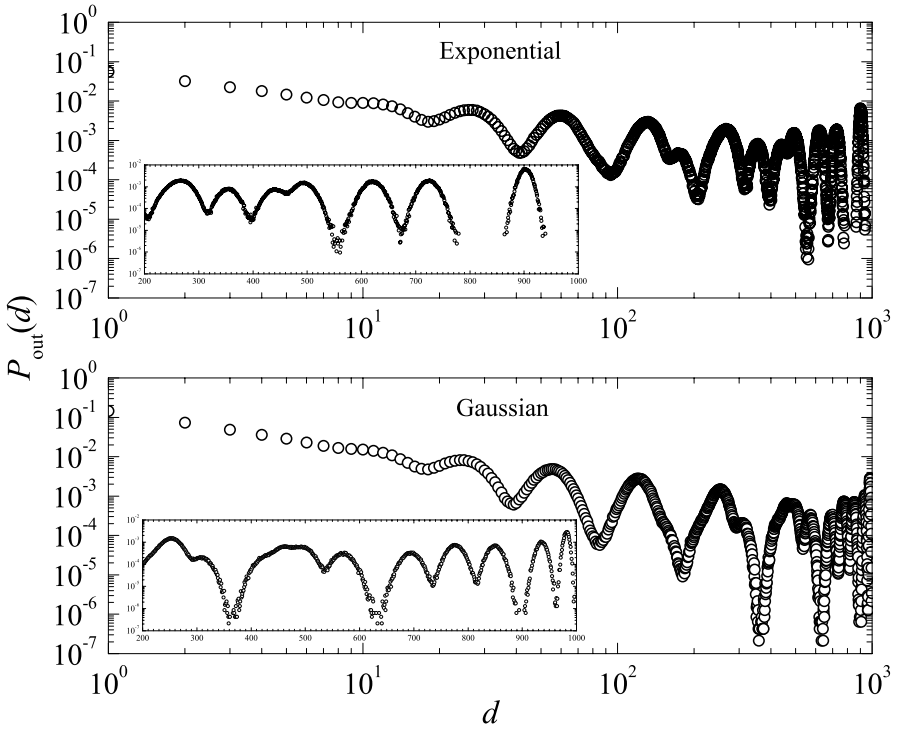


Figure 2.4: The out-degree distributions of the ensemble of content-based networks of sizes $N = 10^3$ averaged over $10^4$ realizations, with the associated strings obeying either exponential (above) or Gaussian (below) length distributions. The insets exhibit the large degree regions plotted in semi-logarithmic scale.

The average clustering coefficients of the model networks are very close to each other, $\langle c\rangle = 0.781$ and $\langle c\rangle = 0.777$, larger than those of the random versions of the networks with the same total degree distributions, $\langle c_{\rm rand}\rangle \approx \langle d\rangle/N = 0.417$ and $\langle c_{\rm rand}\rangle \approx \langle d\rangle/N = 0.145$ for the exponential and Gaussian length distributions, respectively. Their average shortest path lengths are also very small and close to each other, being $\langle \ell\rangle = 1.586$ and $\langle \ell\rangle = 1.855$. Thus, we may say that the model networks are of the smallest-world type [5] where "smallest-world" refers to the fact that the average shortest path length is independent of the network size above a certain threshold which here corresponds to the size above which the network consists of one connected cluster. Actually we can interpret this result for any given length distribution of key-sequences which is confined within an interval where the minimum sequence length is unity. Requiring that there are at least two key-sequences of unit length (i.e, 1 and 0), we can show that $\ell \leq 4$. Consider

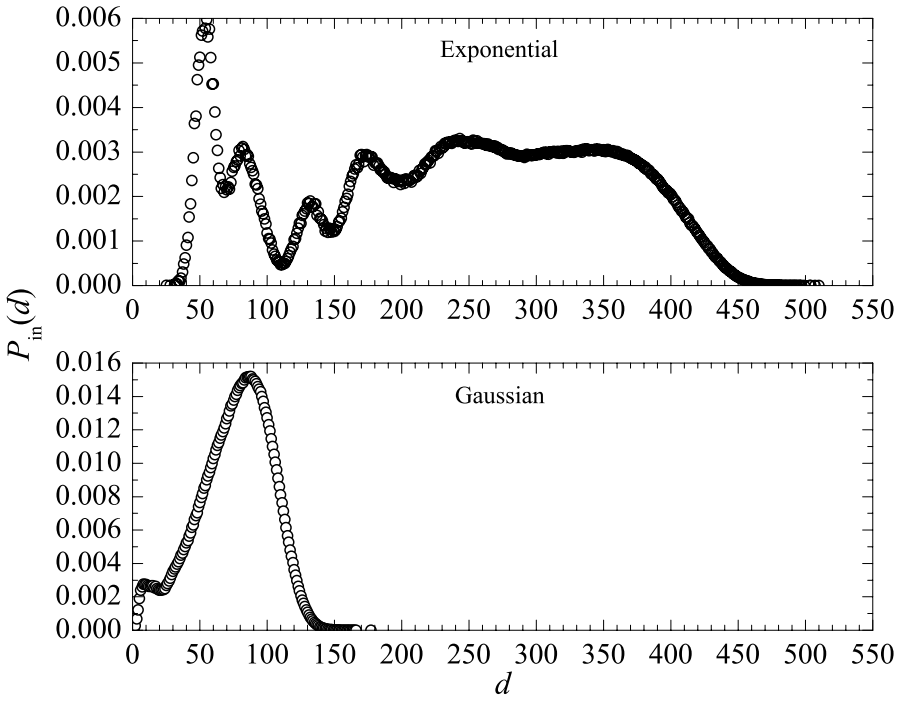


Figure 2.5: The in-degree distributions of the ensemble of content-based networks of sizes $N = 10^3$ averaged over $10^4$ realizations of sequences following either exponential (above) or Gaussian (below) forms.

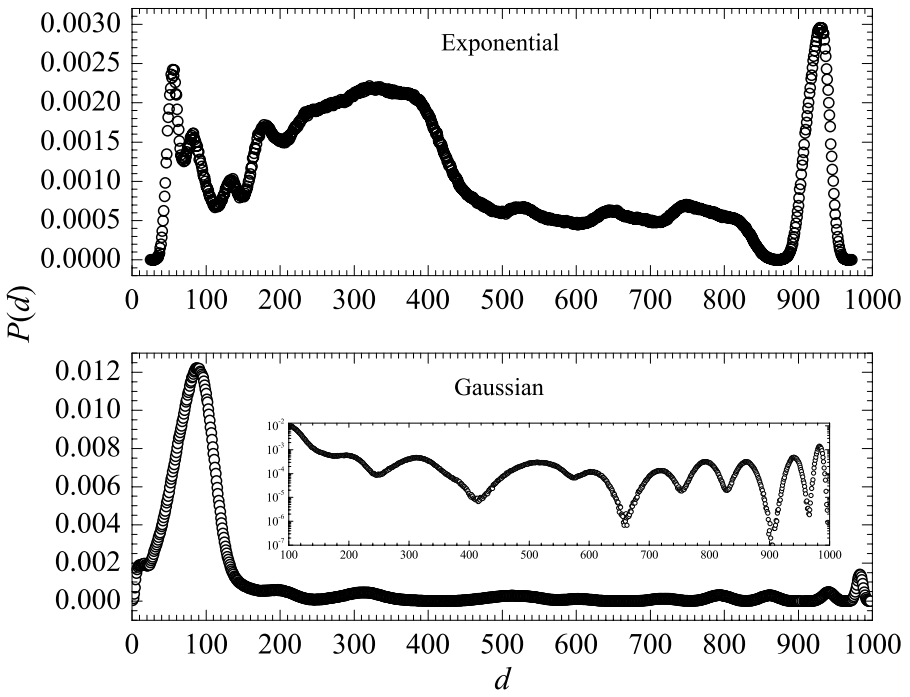


Figure 2.6: The total degree distributions of the ensemble of content-based networks of sizes $N = 10^3$ averaged over $10^4$ realizations of sequences following either exponential (above) or Gaussian (below) form. We have also re-plotted the large degree region of the distribution for the Gaussian case, in semi-logarithmic for better visibility.

the "extreme case" where one is searching for a path between a node whose key- and lock-sequence are of all ones (i.e, $\ldots 1111 \ldots$) and a node whose key- and lock-sequence are of all zeros (i.e, $\ldots 0000 \ldots$). The shortest path between these nodes needs to pass through key-sequences of unit length and a hybrid lock-sequence containing both ones and zeros (i.e, $\ldots 1 \ldots 0 \ldots$). Thus the minimal path between these two extreme nodes ($\ldots 1111 \ldots \leftarrow 1 \rightarrow \ldots 1 \ldots 0 \ldots \leftarrow 0 \rightarrow \ldots 0000 \ldots$) gives the $N$-independent upper bound $\max(\ell) = 4$.

### 2.3 Fine Structure due to Contents

We have shown in Fig. 2.1 that although our analytical calculations in the effective-medium or mean-field approximation for the out- and in-degree distributions can capture their behavior qualitatively, we have also observed that the theoretical curves deviate from the simulation results in large out-degree and small in-degree regions. We should state here that the degree of agreement between the analytical approximations and simulations is totally determined by the length distributions of sequences. We want to start with a simple but instructive example [7] where our approximation totally misses the in-degree distribution, and we will find the solution only by considering the different contents of the lock-sequences.

Imagine that we have an ensemble of networks of size $N$ where the lengths of the key-sequences are fixed at $l = 1$ and those of the lock-sequences at $k$. In this case, the matching probability in Eq. 2.1.3 is exact, $p(1, k) = 1 - (1 - 1/r)^k$ without recourse to any mean-field approximation. In the limit of very large number of nodes (such that, all the sequences of length $k$ are realized), the degree distributions are binomials. The out-degree distribution is given by

$$P_{\text{out}}(d) = \binom{N}{d} \; [p(1,k)]^d \; [1 - p(1,k)]^{N-d} \quad . \tag{2.3.12}$$

The in-degree distribution would, in the naive effective-medium approach, be given by Eq. 2.1.6. However, a more careful analysis shows that it is in fact a superposition of binomial distributions each for a different number $I$ of letters occurring in the lock-sequences with a mean and variance depending upon $I$. Let us denote the total number of different configurations of sequences of length $k$ by

$\omega = r^k$ and the number of those sequences containing $1 \leq I \leq \min(k, r)$ different letters by $\omega(I)$, $\omega = \sum_I \omega(I)$. For the lock-sequences with $I$ different letters, the distribution of in-coming edges is given by

$$P_I^{\rm in}(d) = \binom{N}{d} \left(\frac{I}{r}\right)^d \left(1 - \frac{I}{r}\right)^{N-d} \quad , \tag{2.3.13}$$

where $I/r$ is the probability that a randomly selected key-sequence consisting of only one letter is one of the $I$ different letters contained in the lock-sequence. Then the in-degree distribution may be written as

$$P_{\rm in}(d) = \sum_{I=1}^{\min(k,\, r)} \frac{\omega(I)}{\omega}\, P_I^{\rm in}(d) \quad , \tag{2.3.14}$$

where $\omega(I)/\omega$ is the probability of encountering a randomly selected lock-sequence with $I$ different letters. Now we will calculate the number of configurations of lock-sequences constituted by $I$ different letters. Let us denote the multiplicity of letter "$a_i$" in a sequence of length $k$ by $n_{a_i}$. Given the $I$ and $k$ we have two constrains,

$$k = \sum_{i=1}^{I} n_{a_i} \quad , \quad 1 \leq n_{a_i} \leq k - I + 1 \quad . \tag{2.3.15}$$

Fixing the set of $I$ different letters and their multiplicities $\{n_{a_i}\}$, the number of configurations $\omega(I|\{n_{a_i}\})$ of such sequences is a multinomial coefficient,

$$\omega(I|\{n_{a_i}\}) = \binom{k}{n_{a_1}} \binom{k - n_{a_1}}{n_{a_2}} \ldots \binom{k - \sum_{j=1}^{I-2} n_{a_j}}{n_{a_{I-1}}} \quad . \tag{2.3.16}$$

Using the constraints in Eq. 2.3.15 we write the number of sequences containing $I$ different letters as

$$\begin{aligned} \omega(I) &= \binom{r}{I} \sum_{\{n_{a_i}\}} \omega(I|\{n_{a_i}\}) \\ &= \binom{r}{I} \sum_{n_{a_1}=1}^{k-I+1} \sum_{n_{a_2}=1}^{k-n_{a_1}-I+2} \cdots \sum_{n_{a_{I-1}}=1}^{k-(n_{a_1}+\ldots+n_{a_{I-2}})-1} \omega(I|\{n_{a_i}\}) \quad , \end{aligned} \tag{2.3.17}$$

where $\binom{r}{I}$ is the total number of ways $I$ different letters can be chosen from an alphabet of $r$ letters. If we successively sum over the multiplicities $n_{a_i}$ appearing in Eq. 2.3.17, starting with the last one, we get

$$\omega(I) = \binom{r}{I} \sum_{n=0}^{I-1} \binom{I}{n} (I - n)^k \, (-1)^n \quad . \tag{2.3.18}$$

In Fig. 2.7, we compare our analytical results with those of the simulations obtained by generating $10^6$ realizations of the model networks of size $10^3$ with sequences constructed from an alphabet of $r = 10$ letters where the lengths of lock-sequences are fixed at $k = 3$. We see that the theoretical curves, where we have plotted the exact binomial forms (see Eqs. (2.3.12, 2.3.14)) as well as their Gaussian and Poisson approximations, are in excellent agreement with simulations. We should note here that, had we not taken into account the contents of the lock-sequences (thus, the fine structure of each sequence), we would have obtained the same result for the in-degree distribution as the one we got for the out-degree distribution. Because the out-degree distribution $P_{\rm out}(d) = P_1^{\rm out}(d)$ is the binomial with the mean $d_{{\rm o},\,1} = Np(1,3)$ (Eqs. (2.1.5, 2.1.7)) and the in-degree distribution $P_{\rm in}(d) = P_3^{\rm in}(d)$ is the binomial with the mean $d_{{\rm i},\,3} = Np(1,3)$ (Eqs. (2.1.6, 2.1.8)), $P_{\rm out}(d) = P_{\rm in}(d)$ for this model in the naive effective-medium approach. In contrast to the previous content-based networks, the in-degree distribution has wider support than the out-degree distribution.

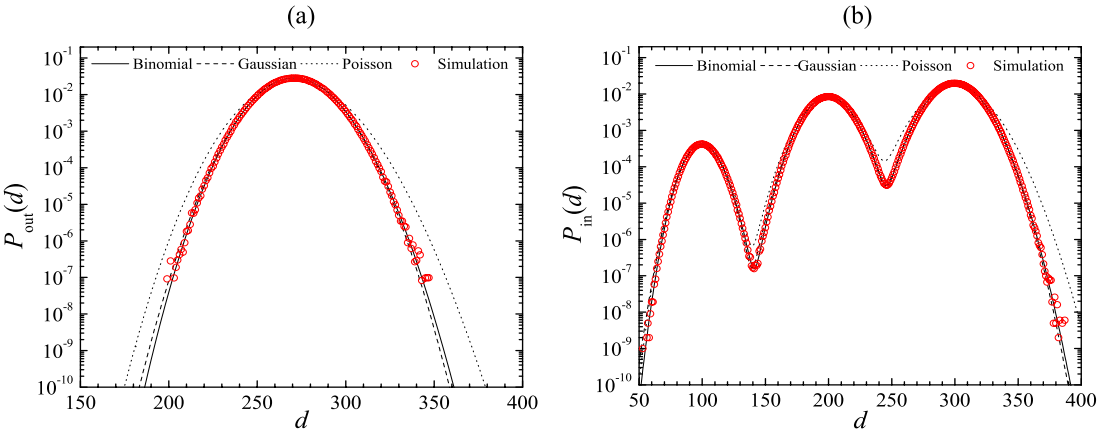


Figure 2.7: The directed degree distributions as obtained by our analytical solutions and simulations (red circles). (a) The out-degree distribution is a binomial with the average out-degree $d_{\rm o} = 271$ and the variance $\sigma_{\rm o}^2 = 197.559$ (see Eq. 2.3.12). (b) The in-degree distribution is a superposition of binomials with the average in-degree depending upon $I$, the number of different characters contained in the lock-sequence associated with the node, viz., $d_{{\rm i},\,1} = 100$, $d_{{\rm i},\,2} = 200$ and $d_{{\rm i},\,3} = 300$, and the variances $\sigma_{{\rm i},\,1}^2 = 90$, $\sigma_{{\rm i},\,2}^2 = 160$ and $\sigma_{{\rm i},\,3}^2 = 210$ (see Eq. 2.3.14). In both cases, we have also plotted the Gaussian and Poisson approximations to the degree distributions to allow a better comparison.

The lock-sequences of length $k$, as has been illustrated above, can be grouped with respect to the number $I_l$ of different subsequences of length $l$ embedded in them.

For arbitrary values of $l > 1$ and $k > l$, it is very hard to calculate the number $\omega_k(I_l)$ of sequences of length $k$ containing $1 \leq I_l \leq \min(k-l+1, r^l)$ different subsequences of length $l$. In this case (i.e., $l > 1$ and $k > l$), the connection probability in Eq. 2.1.3 is not valid for the key-sequences either. But now the key-sequences of length $l$ can be grouped into equivalence classes with respect to their shift-match numbers [8] or binary vectors [63, 64] which measure the auto-correlations within sequences. Following the notation of [8], the shift-match number $s(a)$ of a sequence $a$ of length $l$ (say, $a = a_1 a_2 a_3 \ldots a_l$) is defined as the binary sequence of the same length $l$, whose $i$th digit $s_i$ is $s_i = \prod_{j=i}^{l} \delta_{a_{1-i+j},\, a_j}$. For example, if $a = 110$ then $s(a) = 100$. We will demonstrate here the situation, via a simple example, where the out-degree distribution of the key-sequences of the same length $l$ splits with respect to their shift-match numbers.

Let us consider an ensemble of model networks of size $N$ where the lengths of the key-sequences are fixed at an arbitrary value $l$ and those of the lock-sequences at $k = l + 1$. The out- and in-degree distributions are binomials in the limit of very large network size. So we assume that all the configurations of the key- and lock-sequences are realized. In the case we consider here, we can easily write the number $\omega_{l+1}(I_l)$ of configurations of sequences of length $k = l + 1$ containing $1 \leq I_l \leq \min(2, r^l)$ different subsequences of length $l$. The number of the lock-sequences containing only the identical subsequences (thus, $I_l = 1$) of length $l$ is $\omega_{l+1}(1) = r$, and the rest of the configurations of the lock-sequences contain two different subsequences (i.e., $I_l = 2$) of length $l$, thus $\omega_{l+1}(2) = r^{l+1} - r$. The in-degree distribution is a superposition of binomials in the limit of very large system size. Generalizing the expression for the in-degree distribution in Eq. 2.3.14,

$$P_{\rm in}(d) = \sum_{I_l=1}^{\min(k-l+1,\, r^l)} \frac{\omega_k(I_l)}{\omega_k}\, P_{I_l}^{\rm in}(d) \quad , \tag{2.3.19}$$

and using the results of Eq. 2.3.13, we get

$$P_{\rm in}(d) = \frac{1}{r^{k-1}} \binom{N}{d} \left[ \left(\frac{1}{r^l}\right)^d \left(1 - \frac{1}{r^l}\right)^{N-d} + (r^{k-1} - 1) \left(\frac{2}{r^l}\right)^d \left(1 - \frac{2}{r^l}\right)^{N-d} \right] . \tag{2.3.20}$$

The out-degree distribution is also a superposition of binomials each centered around the mean values according to the shift-match numbers of the key-

sequences,

$$P_{\rm out}(d) = \sum_s \frac{\widetilde{\omega}_l(s)}{\omega_l} \; P_s^{\rm out}(d) \;\; , \tag{2.3.21}$$

where $\widetilde{\omega}_l(s)$ is the number of configurations of the key-sequences of length $l$ with shift-match number $s$, and $P_s^{\rm out}(d)$ is the out-degree distribution of such sequences. (Note here that $s$ depends on $l$.) Let us consider the key-sequences of length $l$ and all the lock-sequences of length $l+1$ we can generate from these key-sequences. In this way, we will calculate the number $n(s,k)$ of configurations of those lock-sequences of length $k = l+1$ containing a given key-sequence with shift-match number $s$. $(i)$ The key-sequences with the highest shift-match number $s^*$ are the ones containing $l$ identical letters. The number of configurations of such sequences is obviously $\widetilde{\omega}_l(s^*) = r$. Consider a given sequence of this kind and add a letter out of $r-1$ letters to the right or left most side of this sequence. By this process, for each different letter added one obtains a different sequence of length $l+1$ for the given key-sequence of length $l$. Thus the number $n(s^*,k)$ of configurations of the lock-sequences of length $k = l+1$ containing such a sequence is $n(s^*,k) = 2(r-1)+1$, where $(+1)$ comes from the addition of a letter which is identical to the already existing ones. $(ii)$ Now consider the key-sequences with shift-match numbers $s \neq s^*$. For a given sequence of this kind, again we can add a letter, now out of $r$ letters, to the right or left most side of this sequence which yields a different sequence of length $l+1$ for each different letter we add. Thus, the number $n(s \neq s^*,k)$ of configurations of the lock-sequences containing this key-sequence is $n(s \neq s^*,k) = 2r$. Note here that we have reached this result without knowing the specific value of the shift-match number $s$, all we know is that it is different from $s^*$ . Now we can write down the expressions for the out-degree distributions of nodes with respect to their shift-match numbers as

$$P_{s^*}^{\rm out}(d) = \binom{N}{d} \left(\frac{2r-1}{r^{l+1}}\right)^d \left(1 - \frac{2r-1}{r^{l+1}}\right)^{N-d} \;\; , \tag{2.3.22}$$

and

$$P_{s\neq s^*}^{\rm out}(d) = \binom{N}{d} \left(\frac{2r}{r^{l+1}}\right)^d \left(1 - \frac{2r}{r^{l+1}}\right)^{N-d} \;\; . \tag{2.3.23}$$

By substituting these results into Eq. 2.3.21 we obtain

$$\begin{aligned} P_{\rm out}(d) &= \frac{\widetilde{\omega}_l(s^*)}{\omega_l}\, P^{\rm out}_{s^*}(d) + \sum_{s\neq s^*} \frac{\widetilde{\omega}_l(s)}{\omega_l}\, P^{\rm out}_{s}(d) \ , \\ &= \frac{\widetilde{\omega}_l(s^*)}{\omega_l}\, P^{\rm out}_{s^*}(d) + \frac{\omega_l - \widetilde{\omega}_l(s^*)}{\omega_l}\, P^{\rm out}_{s\neq s^*}(d) \ . \end{aligned} \tag{2.3.24}$$

In Fig. 2.8 we compare our analytical calculations, including the exact binomial distributions (see Eqs. (2.3.20, 2.3.24)) as well as their Gaussian and Poisson approximations, with simulation results which have been obtained by averaging over $10^5$ realizations of the model networks of size 1280. The key- and lock-sequences are random binary strings whose lengths have been fixed at $l = 2$ and $k = 3$, respectively. We see that the theoretical curves agree very well with the simulation results, where the out- and in-degree distributions split into two distributions. As one may recognize, the number of peaks in the degree distribution is totally determined by the lengths of the sequences (compare Fig. 2.8 with Fig. 2.7). The number of peaks in the in-degree distribution is determined by $\max(I_l) = \min(k - l + 1, r^l)$. On the other hand the observation of splitting in the out-degree distribution is determined by the shift-match numbers of the key-sequences as well as the difference between the lengths of the lock- and key-sequences. Thus, some of the shift-match numbers are degenerate in the sequence-matching probabilities depending on $k - l$ as have been shown in [8], where it has been rediscovered that the sequence-matching probabilities of the key-sequences depend on their shift-match numbers and $k - l$. The work has not been published, as it has turned out that the same problem had been studied much earlier [63, 64] and the role played by the auto-correlations in the string-matching problem elucidated.

## 2.4 Information Theoretic Approach to Interaction Networks

The information-theoretic approach we would like to present here is quite generic and promises to be widely applicable to systems which can be described in terms of networks of interacting nodes. In this approach, an interaction, represented as an edge connecting a pair of nodes, is established if and only if a number of more or less stringent constraints are fulfilled. The number and strictness of the constraints may be quantified as a certain amount of information, or code, that

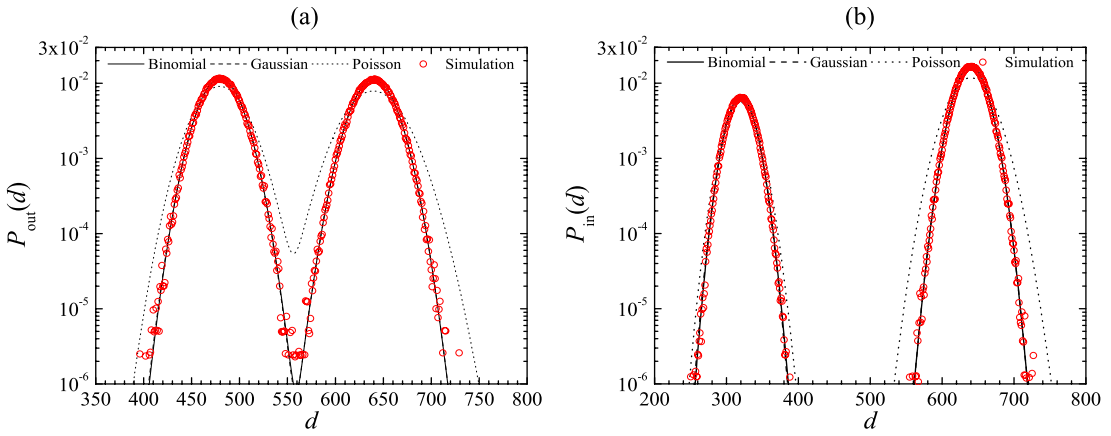


Figure 2.8: The directed degree distributions as obtained by our analytical solutions and simulations (red circles). (a) The out-degree distribution is a superposition of binomials with the respective average out-degrees depending upon the shift-match numbers of the key-sequences, viz., $d_{\mathrm{o},\,s^*} = 480$ and $d_{\mathrm{o},\,s\neq s^*} = 640$, and the variances $\sigma^2_{\mathrm{o},\,s^*} = 800$ and $\sigma^2_{\mathrm{o},\,s\neq s^*} = 640$ (see Eq. 2.3.24). (b) The in-degree distribution is a superposition of binomials with the average in-degrees according to the number of different subsequences of length $l$ contained, $d_{\mathrm{i},\,1} = 320$ and $d_{\mathrm{i},\,2} = 640$, and the variances $\sigma^2_{\mathrm{i},\,1} = 960$ and $\sigma^2_{\mathrm{i},\,2} = 640$ (see Eq. 2.3.20).

has to be shared between the two nodes. The topology of the interaction network is then determined by the distribution function of the required amount of shared information between the interacting nodes.

The way in which we model the shared information, corresponding to a set of constraints, is via a string-matching condition we have introduced earlier [5, 6] and discussed in detail in the previous sections. In this content-based model, the condition for establishing a connection is that a code, represented by a string associated with one node, match, letter for letter, a substring of the code associated with another node. In this model, matching each successive letter will correspond to satisfying an additional constraint. The number of constraints can thus be mapped to the length of a string to be matched. The chance satisfaction of the constraints is smaller, the longer the alphabet.

To use a "lock and key" analogy to illustrate the idea of simultaneously satisfying a number of constraints, the first string may be regarded as the "key" combination that opens the "lock," which in this case may be opened by more than one key. The probability of a chance hit on one of the right combinations decreases exponentially with the length $l$ of the sequence, as $\exp(-l \ln r)$ (see Eq. 2.1.3).

Note that $-l \ln r$ is in fact the so called "Shannon information" [65] of a random "key," selected from an alphabet of $r$ letters.

This approach seems to be particulary well-suited to the description of genetic regulatory networks, which operate on a cognate/cognate responsive element basis. Transcription factors (TF) are the cognates which bind the cognate responsive elements, i.e., the binding sites (regulatory sequences) within the promoter regions (PR) of different genes. The "key" to the promoter region, so to speak, is the binding motif. We have modelled the transcriptional regulation network of yeast presented in Section 3, by using the bitwise information contents of the binding motifs that we have extracted from the data reported by Harbison et al. [41] as described below.

### 2.5 Bitwise Information Content

If we assume that the letters in a string are not correlated, although their relative frequencies of occurrence may depend on their position, then the information content of a sequence can be computed as the sum of the information content of each letter in the sequence. Let us also assume that positions within the sequence have equal significance, i.e., the maximum amount of information which can be contained in any position within the sequence is uniform.

In a given sequence of length $L$, with letters chosen from an alphabet of length $r$, the information content, which is the negative of the Shannon entropy [65], is given by $I = \sum_{j=1}^{L} I_j$ with

$$I_j = \sum_{i=1}^{r} f_{ij} \ln f_{ij} \quad , \tag{2.5.25}$$

where $f_{ij}$ with $i = 1, \ldots, r$ are the relative frequencies of the different letters at each position $j$ in the sequence. Note that $I_j = 0$ if we know for sure that a certain (e.g., $i$th) letter and no other, will appear (in which case the relative frequency $f_{ij} = 1$, and $f_{i'j} = 0$ for $i' \neq i$). Thus, Shannon entropy is the amount of information which we receive from a signal over and above what we already knew about the system. Let us define a relative Shannon information, $R = \sum_j R_j \equiv \sum_j I_j + L \ln r$, which is the difference between $I$ and the Shannon information

communicated by a signal composed from an alphabet with equi-probable letters. This relative quantity is the definition of information content which we will use.

The bitwise information content of a sequence is the number of binary digits {0,1}, needed to code the same amount of information. The bitwise information content of the $j$th member of the sequence, namely the length increment which it will contribute, is,

$$\delta l_j = \frac{R_j}{\ln 2} \ , \tag{2.5.26}$$

since the number of bits, i.e., binary digits, needed to specify a character from an alphabet of length $r$ is (at most) $n = \ln r / \ln 2$. However, $\delta l_j$ is not, in general, an integer. Therefore a coarse graining, which entails a certain amount of arbitrariness, is called for. This coarse graining may be guided by the real system where the values of $\delta l_j$ fall, of themselves, into a number of distinct clusters.

# 3 MODELLING THE TOPOLOGICAL PROPERTIES OF TRANSCRIPTIONAL REGULATORY NETWORKS: A COMPARISON WITH YEAST

We hereby would like to present a content-based model [9] which is able to capture all the considered global topological properties of the transcriptional regulatory network of yeast when the distribution of the amount of shared information is used as the biological input. We believe that this approach provides an understanding of how interactions based on shared information might arise spontaneously between subsequences of any sufficiently long linear code, even when this code is completely random, and how a complex network emerges as a result.

We construct a null model of a transcriptional regulatory network (TRN) by adapting the sequence-matching rule we have introduced earlier [5, 6]. The biological input to the model consists of the effective length distribution of the binding sequences recognized by the transcription factors of yeast [35, 41] and the form of the length distribution of the intergenic regions [66], in the absence of more specific information regarding the lengths of the promoter regions. By "effective length" we mean the bitwise information content (site specificity) of a binding motif with variations and uncertainties, such as rTCAytnnnnAcg. The model is null in the sense that it does not take into account all the complications and processes taking place at the level of transcriptional initiation discussed in Section 1.2.

We make a very detailed analysis of the topological features of the TRN of yeast *Saccharomyces cerevisiae*, using the available data [35, 36, 37, 38], in order to test the predictions of this model (see Table 3.1 for the databases used). We are able to demonstrate that our model is able to capture with convincing precision all the global topological features discussed in the literature, such as in-, out- and total degree distribution [14, 15, 16], the degree-degree correlation [20, 21], the clustering coefficient [22, 23, 24] the rich-club coefficient [26, 27] and the

$k$-core structure [28, 30, 31], i.e., the hierarchical organization of the links. This thorough topological characterization allows us to discriminate between different, more restricted null-null models which capture some but not all of the features of the yeast network. We thus show that our model is in a sense a minimal null model.

## 3.1 Sequence Matching Model for the Transcriptional Regulatory Networks

The genes of the organism under consideration are represented by the nodes in our content-based model network, only a small percentage of which code transcription factors (TFs). We associate a sequence with each node representing the promoter region (PR) of the corresponding gene through which the gene may be regulated. With those genes coding TFs we also associate a second sequence, uncorrelated with the first, representing the binding motifs recognized by the TF through which the corresponding gene may regulate the expression of other genes or the gene itself. For simplicity we assume that there is only one transcription factor that is coded by each regulatory gene. (See Fig. 3.1)

In our model the binding motifs and the PRs are represented as random binary sequences (thus, the size of the common alphabet is 2), whose lengths obey different probability distributions. The TF binding motifs are typically short sequences with a narrow length distribution [35, 41], since a TF selectively binds 5-10 bases and not much more. A single TF can bind a number of similar sequences, and we have used the information content of the binding motif representing these sequences in order to obtain a distribution of effective lengths for the randomly generated binary sequences representing our TF binding motifs. The details of this calculation are discussed in Section 2.5.

We assume that the lengths of the PRs are distributed in the same way as the lengths of the intergenic regions, obeying long tailed power-law distribution [66] whose exponent is the only free parameter in the model, and will be determined from a comparison of the topological features of the model and the experimental regulatory networks, as described in the next section.

As has been already shown in the previous sections, the length distributions of the sequences associated with the nodes of a content-based network determines its topological properties. The amount of information coded in these randomly generated binding motifs and promoter regions thus constitutes the essential biological ingredient of our model and dictates the overall topology of the resultant networks.

The mechanism for establishing connections between nodes of the genetic regulatory network is given by the string matching condition [5, 6, 9] between the binding motifs of the TFs and all possible uninterrupted subsequences of the PRs. The directed network of regulatory gene interactions is obtained by drawing a directed link from each TF-producing node A to all those nodes B, B′, B″, ... , whose PRs contain the binding motif associated with the TF coded by node A (see Fig. 3.1).

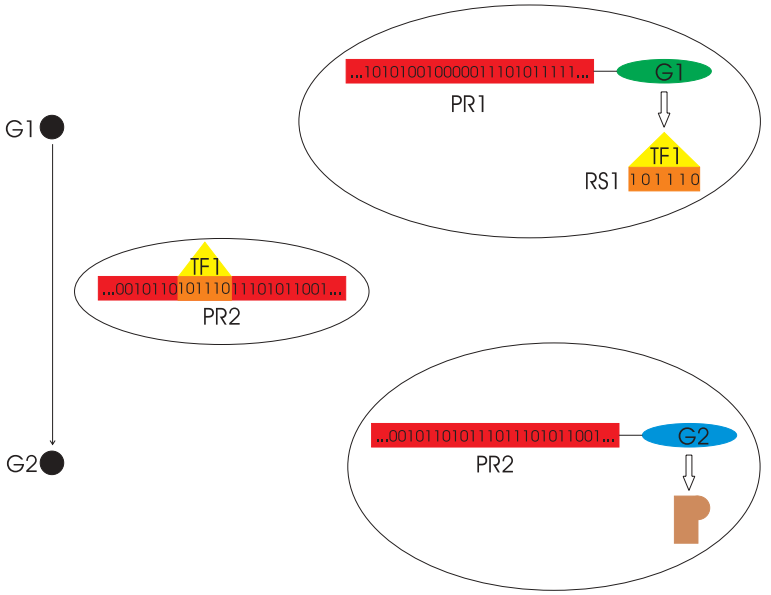


Figure 3.1: Our content-based model of transcriptional regulation networks. A pair of genes with their associated regions and the mechanism of interaction between them, is displayed. Each gene has a promoter region (red boxes, PR1 and PR2) through which the corresponding gene is regulated. One of the genes (green ellipse, G1) codes a transcription factor (yellow triangle, TF1) with the regulatory sequence it recognizes (orange box, RS1), whereas the other gene (blue ellipse, G2) codes a structural protein ("P" shape in brown). Since the RS1 occurs as a subsequence in the PR2, a directed link from G1 to G2 is established signifying a regulatory interaction between them that the expression (production of TF1) of G1 may regulate the expression (production of the structural protein coded by) of G2.

Table 3.1: The summary of databases for the TRN of yeast. The number of interacting genes, regulated genes, regulatory genes (coding TFs), and interacting pairs that appear in the yeast regulatory network as obtained from different sources [35, 36, 37, 38], and the average values, obtained from one hundred realizations of our model ($\pm$ the standard deviations) with $\mu = 0.1$.

| Source | Genes | Regulated | Regulatory | Interacting Pairs |
|---|---|---|---|---|
| Yeastract[†] [37] | 4252 | 4229 | 146 | 12530 |
| Lee et al.[‡] [35] | 2884 | 2850 | 102 | 6441 |
| Luscombe et al.[§] [36] | 3459 | 3420 | 142 | 7071 |
| Kımıkoğlu et al.[♭] [38] | 3763 | 3709 | 180 | 9135 |
| Model | $4167 \pm 177$ | $4103 \pm 181$ | $202 \pm 14$ | $14365 \pm 2067$ |

[†] `http://www.yeastract.com`
[‡] `http://fraenkel.mit.edu/Harbison/release_v24/bound_by_factor/`
[§] `http://sandy.topnet.gersteinlab.org/index2.html`
[♭] *private communication*

### 3.2 Modelling the Transcriptional Regulatory Network of Yeast

We choose the total number of genes and the proportion of those genes coding transcription factors in conformity with the largest data set, Yeastract [37], to make a quantitative comparison with the transcriptional regulatory network of yeast possible. (See Table 3.1.)

The length distribution of the binding motifs in the model genome was derived from the yeast data provided by Harbison et al. [41], where the motifs were reported as letter sequences with upper case letters {A,T,G,C} (high preference), lowercase letters (a weaker preference) and "ambiguity codes" S = C or G, W = A or T, R = A or G, Y = C or T, K = G or T, M = A or C, the letters {H,B,V,D} each correspond to preferences for different triplets out of the four letters, and N indicates "no preference." The bitwise information content of a sequence is the number of binary digits {0,1}, needed to code the same amount of information. The bitwise information content $\delta l_j$ of the $j$th member of the sequence is the length increment which it will contribute. Plotting $\delta l_j$ against the largest of the relative frequencies encountered for that site, one finds that these labels fall into distinct clusters that are determined by $\delta l_j$ alone. The following choice leading to an integer-valued $\Delta l_j$,

$$\Delta l_j = \begin{cases} 2 & \text{for } \delta l_j > 1.04 \\ 1 & \text{for } 0.3 < \delta l_j \leq 1.04 \\ 0 & \text{for } \delta l_j \leq 0.3 \end{cases} , \tag{3.2.1}$$

with the bitwise length of a sequence being finally given by $\sum_j \Delta l_j$, then effectively corresponds to the coarse graining where the upper case letters {A,T,G,C} contribute two bits, the letters {S,W,R,Y,K,M} and their lower case versions contribute one bit, and the letters {H,B,V,D,N} or their lower case versions zero bits to the length of the bit-strings representing the binding sequences in our model. The length of the binary sequence obtained in this way roughly corresponds to the amount of shared information, measured by the Shannon entropy [65], required for the binding of the TF. Performing this calculation (see Section 2.5) for each TF [41], we obtain the length distribution shown in Fig. 3.2.

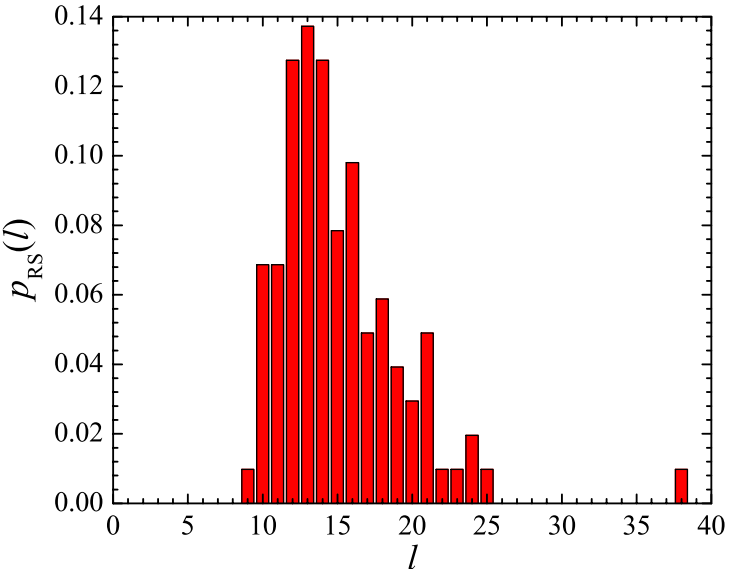


Figure 3.2: Distribution of the amount of bitwise information coded by each regulatory sequence recognized and bound by the 102 TFs in the yeast genome, compiled from the recently published data by Harbison et al. [41]. This distribution is adopted as the length distribution of the random regulatory sequences (binding motifs) in our model.

We assume that the lengths of the PRs follow a power law distribution similar to that of the intergenic regions [66], with

$$p_{\mathrm{PR}}(l) \propto l^{-1-\mu} \quad , \tag{3.2.2}$$

where $0 \leq \mu \leq 2$. For $\mu > 2$ the distribution would not really be fat tailed anymore. In our case, where the range of the PR lengths is finite, it is immaterial whether $\mu < 0$, i.e., the length distribution would still be normalizable; one only needs $\mu > -1$, for the distribution to be decaying power law. It is interesting that we find a borderline value for $\mu$, namely $\mu = 0.1$. We also stipulate that

$l$ is restricted to the interval $l_{\min} \leq l \leq l_{\max}$, where $l_{\min}$ coincides with the peak of the length distribution of the binding motifs shown in Fig. 3.2, while $l_{\max} - l_{\min} + 1 = 250$. In this choice we are guided by the finding [41] that most of the probability for encountering a TF binding site is contained within a window of 250 base pairs (bps) located approximately 100 bps upstream of a gene. In moving from 4-letter alphabet to the binary one the 250 bps window does not double, because the number of edges in the network is required to remain invariant under this transformation. This may be understood by using the approximate result for the sequence-matching probability $p(l, k)$ in Eq. 2.1.3, which determines the average number of edges in the content-based networks. If the sequences to be matched, with large number $r^l$ of configurations of such sequences, are much shorter than the sequences configuring the search space, the matching probability $p(l, k)$ becomes $k/r^l$. Now if one wants to move from an alphabet of $r$ letters to another with $r'$ letters while keeping the number of edges invariant, then it needs to satisfy the condition $p(l', k') = k'/(r')^{l'} = p(l, k)$ where $l'$ and $k'$ are the lengths of the sequences in this new representation. In our case (i.e., $r = 4$ and $r' = 2 = r^{1/2}$), the lengths $l$ of regulatory sequences would double, $l' = 2l$, when we represent them as binary sequences. Thus the matching probability becomes $p(l', k') = k'/r^l$, and $k'$ has to be fixed at $k$ to obtain the same number of edges.

Once the shape of the length distribution of the binding sequences and the functional form, as well as the width, of the PR length distribution have been fixed through the available biological data, the only remaining adjustable parameter in our model is the exponent $\mu$ of the power law distribution of PR lengths, $p_{\rm PR}(l)$.

Clearly, the length distribution for the PRs must be tested against null assumptions, and this we do in Section 3.3. We find that, once the form of the distribution has been chosen as in Eq. 3.2.2, any value of $\mu$ within the interval [0,2] performs reasonably well, while, say, fixing all the PR lengths to be identical gives markedly different results.

In order to optimize the value of $\mu$, we could compare all the available topological characterizations of randomly generated model networks obeying the constraints on the number of nodes and the length distribution of the binding sequences,

for different values of $0 \leq \mu \leq 2$, with those of the yeast TRN. It is obviously desirable, however, to find one number to compare with experiment, rather than, say, the whole degree distribution or the degree-degree correlation function $d_{nn}(d)$. In fact, once $p_{\rm PR}(l)$ is chosen to be of the power-law form given in Eq. 3.2.2, then choosing $\mu$ to make the maximum number $k_{\rm max}$, of $k$-cores of the model and the network obtained from the Yeastract [37] data source coincide is sufficient for the rest of the topological features of the respective networks to fall right on top of each other, as shown in Figs. (3.3, 3.4) and Figs. (3.6–3.11). Note that we have chosen the Yeastract [37] data to tune the parameter $\mu$ and to compare with our results, because it is the largest available data set (see Table 3.1). But we also make comparison with the networks obtained from other data sets as well in Section 3.5.

We should point out that the ensemble of our model networks contains 4167 nodes with nonzero interaction (out of 6000 in total) with a few clusters of size 2 (6% of clusters containing interacting nodes). Thus, almost all the nodes with nonzero degree belongs to the giant component. The analysis for the degree distributions, the rich-club coefficient and the $k$-core structure have been done for all the interacting nodes whereas the clustering coefficient and the degree-degree correlation of nearest neighbors have been evaluated on the giant component.

In Figs. (3.3, 3.4), we show the $k$-core visualizations [29] of one realization of the model network and the Yeastract [37] data. Here $\mu$ has been fixed to 0.1, making the mean and the mode of $k_{\rm max}$ for the model ensemble to coincide with the value we compute from the Yeastract [37] database, at $k_{\rm max} = 9$. Both the model and the experimental network exhibit a highly hierarchical structure with a nested sequence of $k$-shells and an almost exclusively radial arrangement of the edges. The distinct hierarchical organization of the edges is not very sensitive to the precise value of $\mu$, while the total number of shells decreases as $\mu$ increases. (See Section 3.3 for details.). Fig. 3.5 showing the $k$-core visualization [29] of the preferential attachment model of Barabasi-Albert (BA) [14], which has also been claimed to provide a good description of the gene regulatory network of yeast, should be compared with Fig. 3.4. Note the absence of a well-defined hierarchical structure.

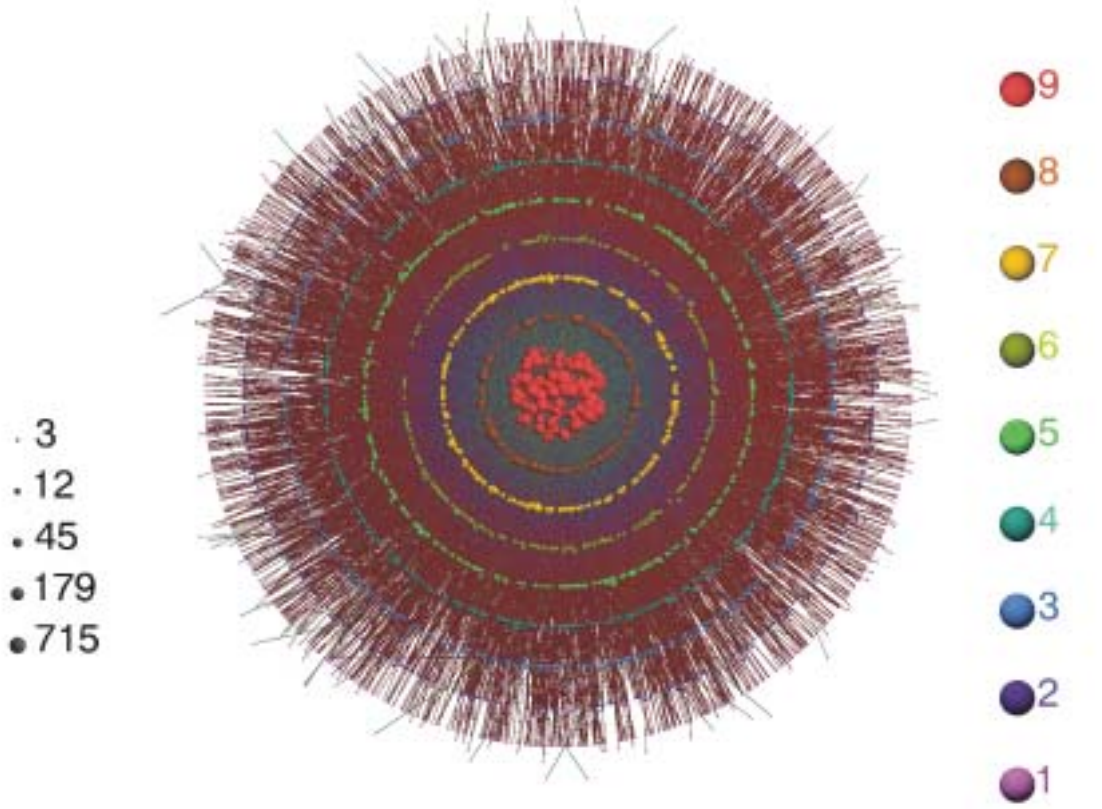


Figure 3.3: The $k$-core visualization of a single realization of our model network obtained by the visualization tool LaNet-vi (http://xavier.informatics.indiana.edu/lanet-vi/). The length distribution exponent of the PRs has been adjusted to $\mu = 0.1$ to match the number of $k$-cores to that obtained from Yeastract [37] data. Dots represent the nodes of the network, while edges between nodes depict connections. Nodes belonging to different $k$-shells are indicated by different colors (on the right hand side) and are arranged around concentric circles, whose average radius decreases with $k$. In particular, a node of a given shell is placed just inside (outside) the corresponding circle, if it is preferentially connected to lower (higher) $k$-shells. The sizes of dots indicate the degrees of the respective nodes (see legend on the left hand side of the figure for representative sizes).

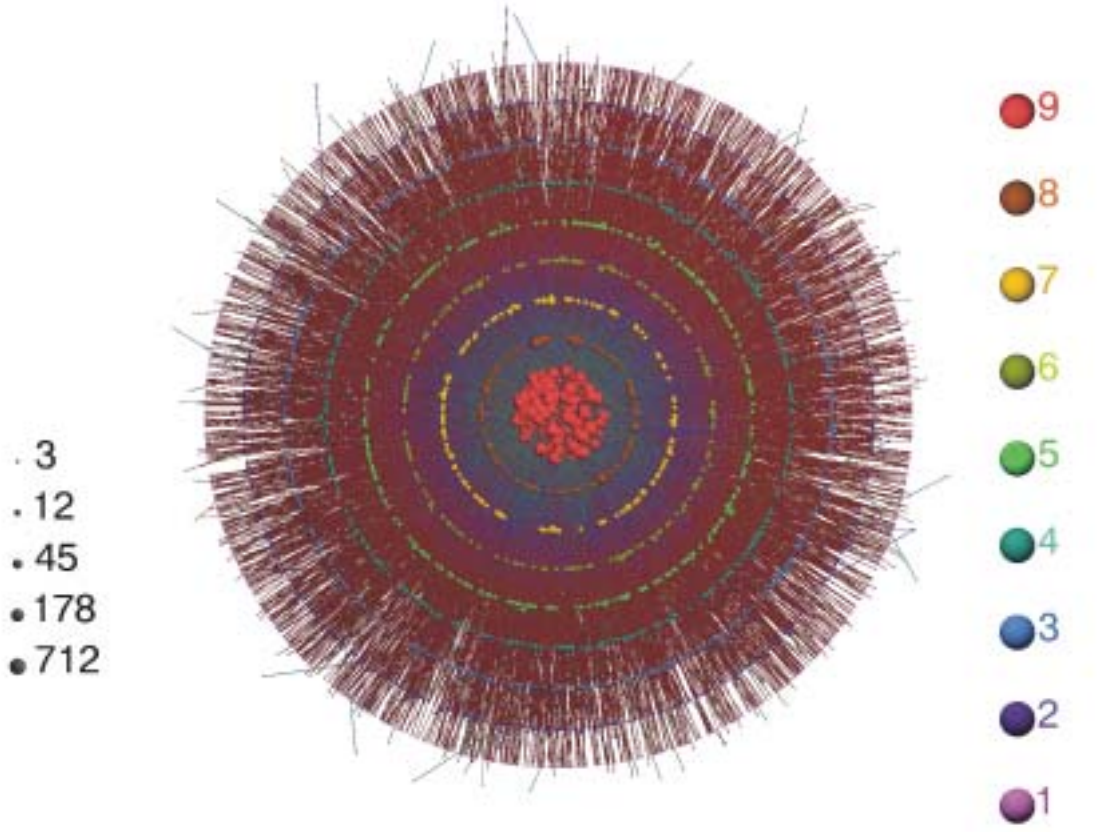


Figure 3.4: The $k$-core visualization of the network extracted from Yeastract [37] data obtained by the visualization tool LaNet-vi (http://xavier.informatics.indiana.edu/lanet-vi/).

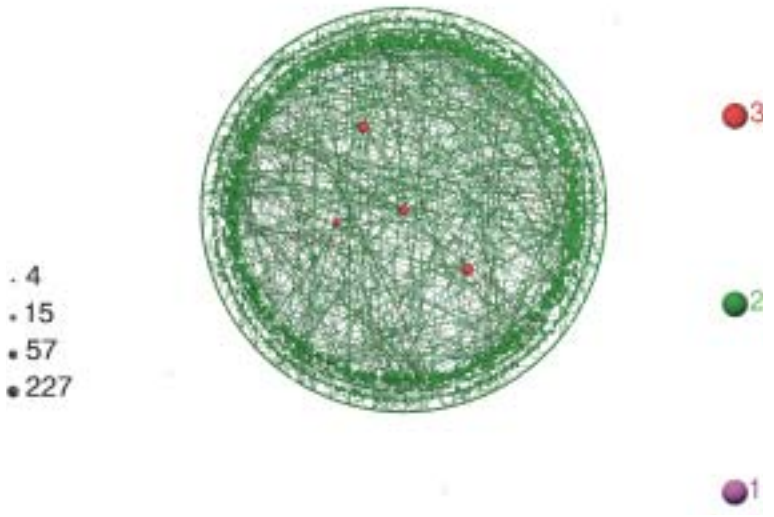


Figure 3.5: The $k$-core visualization of the Barabasi-Albert (BA) model [14] obtained by the visualization tool LaNet-vi (http://xavier.informatics.indiana.edu/lanet-vi/). The network has 5000 nodes, and is built by starting from a fully connected four-cluster and adding nodes with two edges at a time. The number of edges is 9998. Only % 5 of the edges are shown for better visibility. Quantitative analysis of the $k$-core structure has shown that there are only three shells, with 99.9 % of the nodes in the second shell, and the third shell being just the four completely connected set of nodes from which the network is grown.

In Figs. (3.6–3.11), we report our results for the in-, out- and total degree distribution [14, 15, 16], the degree-degree correlation [20, 21], the clustering coefficient [22, 23, 24] and the rich-club coefficient [26, 27], with the choice of $\mu = 0.1$. Results for the yeast TRN, which we have extracted from the Yeastract [37] data have been superposed on the scatter plots of one hundred independent realizations of randomly generated model networks with identical parameters.

The total degree distribution is obtained by ignoring the directionality of the interactions and is generally different from the superposition of in- and out-degree distributions. In Fig. 3.6, Yeastract [37] data for the degree distribution is shown on top of a scatter plot obtained by superposing the results of the ensemble of model networks. The total degree distribution displays a crossover behavior at around $d = 2d_{\rm av}$, before and after which it has an exponential decay and a rather large scatter of points, respectively.

In Fig. 3.7, we exhibit the in-degree distribution obtained from the Yeastract [37] data, and the corresponding scatter plot. The in-degrees are distributed in a very

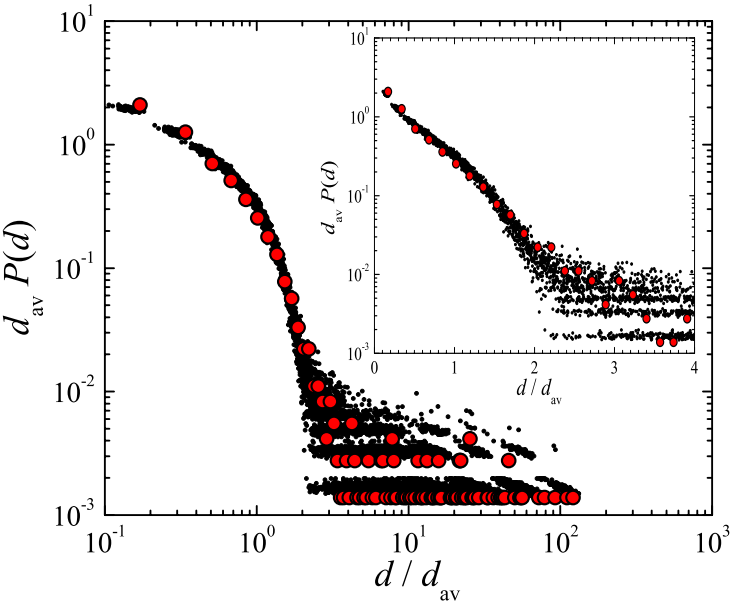


Figure 3.6: Total degree distribution extracted from the Yeastract [37] data (red circles), superposed on the corresponding degree distributions of one hundred realizations of the model network (black dots). The total degree distribution with an inset showing a log-linear plot for $d/d_{\rm av} \leq 4$, where one may observe that both the model and the data points almost fall on a straight line. The axes are scaled by the appropriate average total degree in order to factor out fluctuations in the network size (the number of nodes with nonzero interactions).

narrow interval with an exponential decay.

The out-degree distribution of the yeast and those of model networks have rather large scatter of points due to the relatively small number of TFs. Comparing with the scatter plot obtained from one hundred realizations, we find again that the actual yeast data falls within the boundaries set by the model ensemble (Fig. 3.8).

In Figs. (3.9–3.11), we report the three topological coefficients, the degree-degree correlation, the clustering coefficient and the rich-club coefficient, that go beyond degree distributions in characterizing the network. The agreement is extremely good, in particular, the shoulder observed in the rich-club coefficient in Fig. 3.11, a feature common to both gene regulation and protein-protein interaction networks [27], is captured accurately in our model. Note that the topological features displayed in Figs. (3.6–3.11) are obtained without any further adjustment of $\mu$.

In the degree-degree correlation function (see Fig. 3.9), two regions are distinguished as in the total degree distribution. The small degree region is dominated

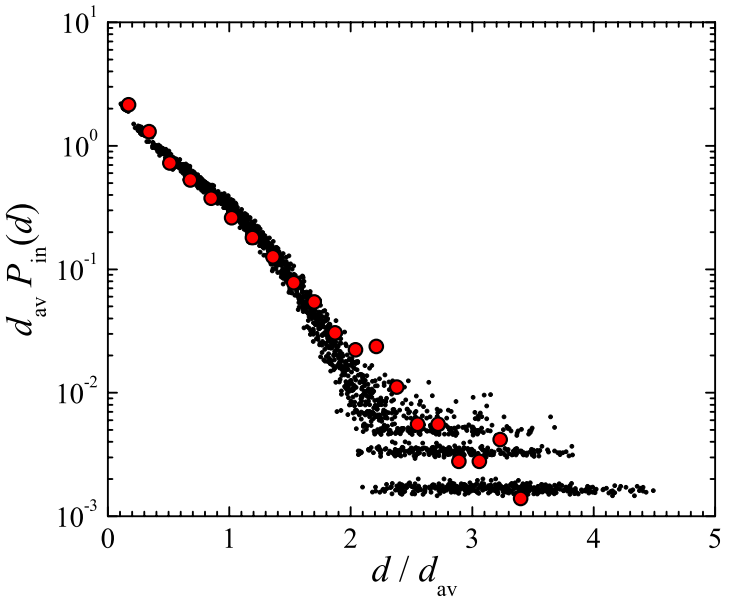


Figure 3.7: In-degree distribution extracted from the Yeastract [37] data (red circles), superposed on the corresponding degree distributions of one hundred realizations of the model network (black dots). The in-degree distribution plotted on a semi-logarithmic scale.

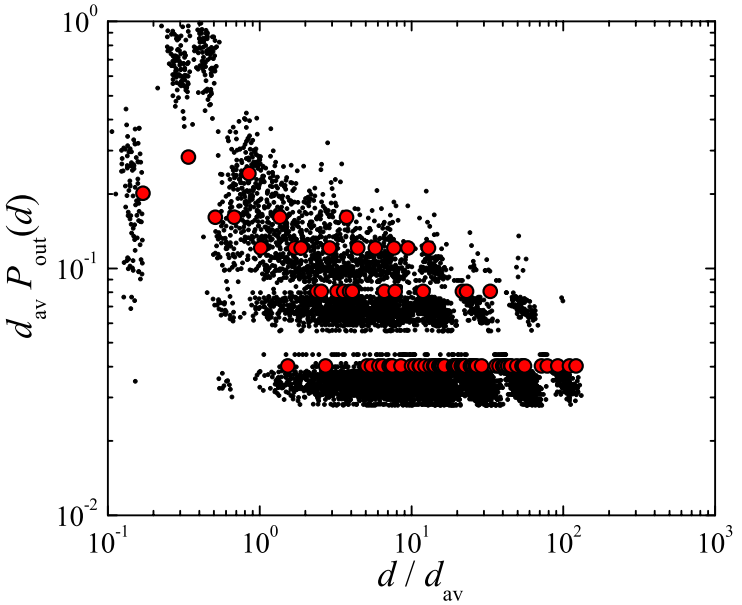


Figure 3.8: Out-degree distribution extracted from the Yeastract [37] data (red circles), superposed on the corresponding degree distributions of one hundred realizations of the model network (black dots). The out-degree distribution plotted on a log-log scale.

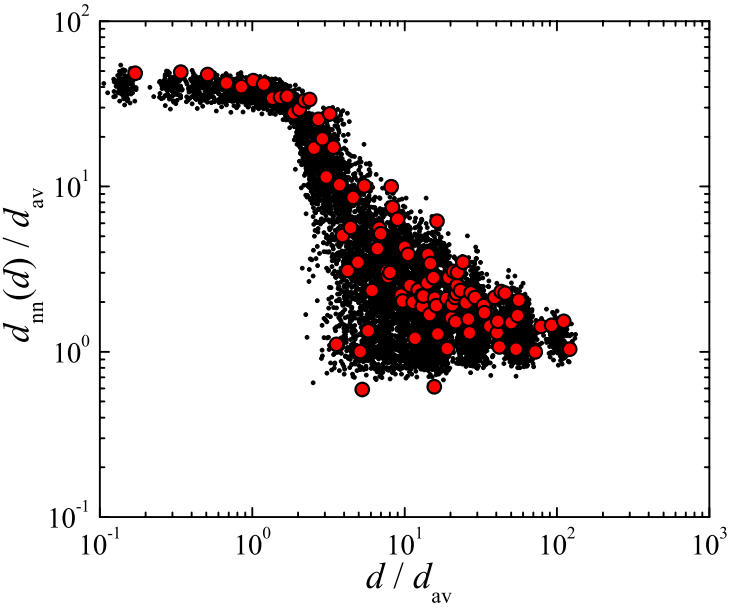


Figure 3.9: Comparison of the degree-degree correlations between neighboring nodes $d_{nn}(d)$ for one hundred realizations of the model (black dots) and the Yeastract [37] data (red circles).

by genes coding structural proteins (thus, the total degree of such a node is equal to its in-degree). The nodes of this type are connected to TF-coding genes, which have relatively large degrees (coming from their out-going links). On the other hand, the large degree region is determined by the TF-coding genes. Most of their nearest neighbors are those nodes coding structural proteins, which have small degrees (coming from their in-coming links). The networks have disassortative characteristics; nodes with large degrees are connected to nodes with small degrees on average, i.e., $d_{\rm nn}(d)$ decreases as $d$ increases.

The behavior of the clustering coefficient $c(d)$, follows from the same principle as above (compare Figs. (3.9, 3.10)). Note here that one needs at least two TF-coding genes to obtain a three-clique [67]. The nearest neighbors of nodes in the first region are TF-coding genes, which are also interconnected with each other, whereas the nearest neighbors of nodes in the second region are mostly those nodes coding structural proteins, which do not interact with each other. Contribution to the clustering coefficient of a node in this region comes from the interactions of the relatively small number of TF-coding nearest neighbors of this node.

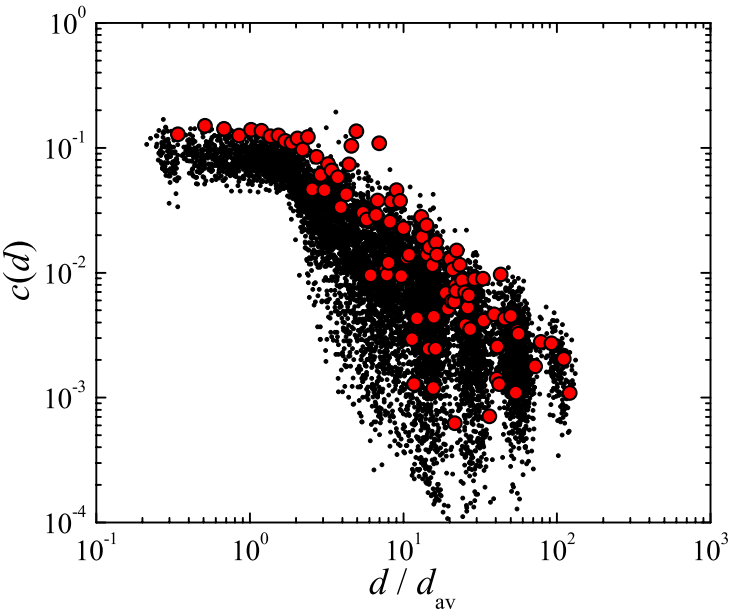


Figure 3.10: Comparison of the clustering coefficient $c(d)$ for one hundred realizations of the model (black dots) and the Yeastract [37] data (red circles).

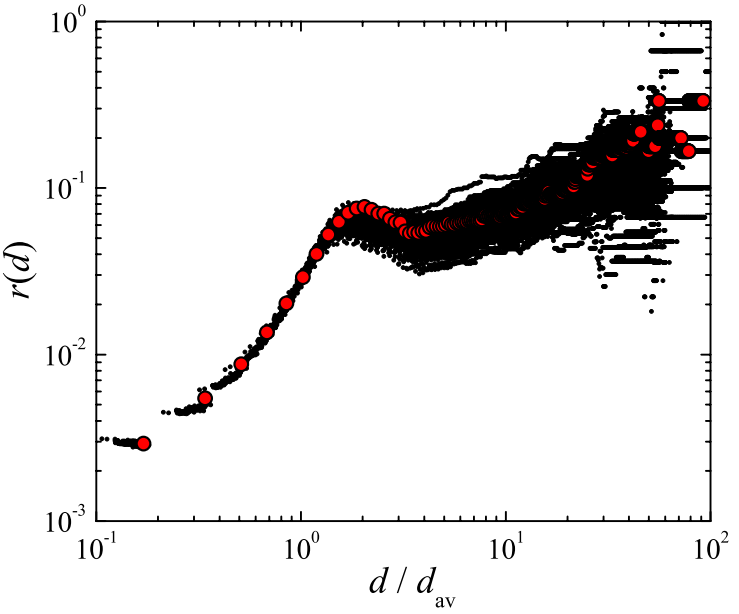


Figure 3.11: Comparison of the rich-club coefficient $r(d)$ for one hundred realizations of the model (black dots) and the Yeastract [37] data (red circles).

## 3.3 Qualitative and Quantitative Aspects of the $k$-core Structure: Choosing the Length Distribution of the Promoter Regions

In this section we would like to discuss certain qualitative and quantitative aspects of the $k$-core structure, how these are related to the other global topological features of the network and how we have used them in deciding upon a length distribution for the PR sequences. We would also like to warn here against an undue emphasis on the visual appearance of the $k$-core structure, which may obscure more subtle differences between networks with respect to such properties as the distribution of nodes over different $k$-shells.

The $k$-core pictures provide a necessary but not sufficient input in modelling a real network. The $k$-core visualization may aid in a quick elimination of certain candidate models, and may amplify certain subtle differences, but in the case of our content-based model, the choice of the appropriate length distribution for the target strings cannot be made on the basis of the $k$-core visualization alone.

This section is organized as follows. First we present the detailed considerations leading to our choice of the exponent $\mu$ in the power law distribution for the PR lengths, the quantitative analysis of the $k$-core structure of the resulting network, and comparison to that of the TRN of yeast. For comparison, we discuss the effect of varying $\mu$ over its whole range, and we show the results for the global topological features of the model networks, again superposed with those of Yeastract [37], for $\mu = 2$. Then, we pose the question whether a different length distribution altogether, such as fixing the PR lengths at some relatively large value, could not have yielded similar agreement with yeast data. As a worst-case example we present a $k$-core visualization which agrees very closely with that of yeast, but which sharply diverges in its quantitative $k$-core structure and other topological features.

### 3.3.1 Determining the value of $\mu$

We find that employing a power law form, $p_{\mathrm{PR}}(l) \propto l^{-1-\mu}$ for the length distribution [66] of the promoter regions in the present model, leads to the hierarchical organization that can be seen in Fig. 3.3, with the connectivities essentially being

between the innermost core and the outer $k$-shells. This feature is not sensitive to the precise value of $\mu$.

To choose the value of $\mu$ to be used in our simulations, we proceed as follows: For different values of $\mu$, we generate an ensemble of one hundred realizations of our model network, with $N = 6000$ genes in total, as in the yeast genome. The value of $\mu = 0.1$ yields the greatest proportion of model networks (36 out of 100) with $k_{\max} = 9$, coinciding with that of yeast. Moreover, for $\mu = 0.1$, the distribution of $k_{\max}$ for different realizations of the model network is symmetric about the mode.

Without any further adjustment of the value of $\mu$, we find that the size of the connected network relative to the size of the whole genome (in number of nodes) is in the right ballpark. On the average, 4167 genes out of the total contribute to the $\mu = 0.1$ model regulatory network, to be compared with 4252 for yeast (see Table 3.1). Out of these 4167 genes, we choose 4.8 %, namely 202 genes, to be TF-coding genes. They end up taking part in a total of 14365 interactions, again on the average. The corresponding values for the yeast regulatory networks reported in the available databases are given in Table 3.1.

Quantitative analysis of the $k$-core structure provides a highly detailed topological characterization of the network, with the total number of shells, the distribution of the nodes over the shells and inter- and intra-shell connectivity [31]. There is detailed quantitative agreement between the $k$-core organizations of the Yeastract network and our model, for $\mu = 0.1$, as can be seen from Figs. (3.12–3.14) where we show the population of each shell and the distribution of the links among different shells. Both for the yeast data and the model, the great majority of the links connect the innermost shell with the others and the growth of the connection probability between shells $k$ and shells with $k' > k$, is like $\exp[ak']$ with $a \approx 3$, as can be seen from Figs. (3.13, 3.14).

The results reported in Figs. (3.12–3.14) may be compared with those of Carmi et al. [31] for the internet. The shell population dependence on $k$, namely $\sim k^{-1}$ is much weaker than that found for the internet Autonomous Systems. The general trend of the proportion of links out to the crust, into the core and in the same

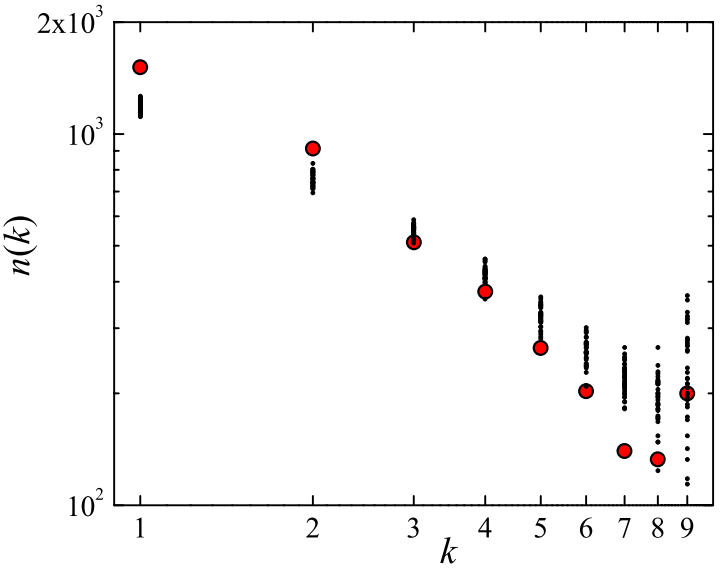


Figure 3.12: Sizes $n(k)$, of $k$-shells of the networks extracted from the Yeastract [37] data (red circles) and the model realizations (black dots). We have taken only the 36 realizations (out of 100) with $k_{\max} = 9$, for ease of comparison. Note that in both the model and the experimental yeast network, the innermost core is over-represented with respect to the common trend, which is approximately proportional to $k^{-1}$.

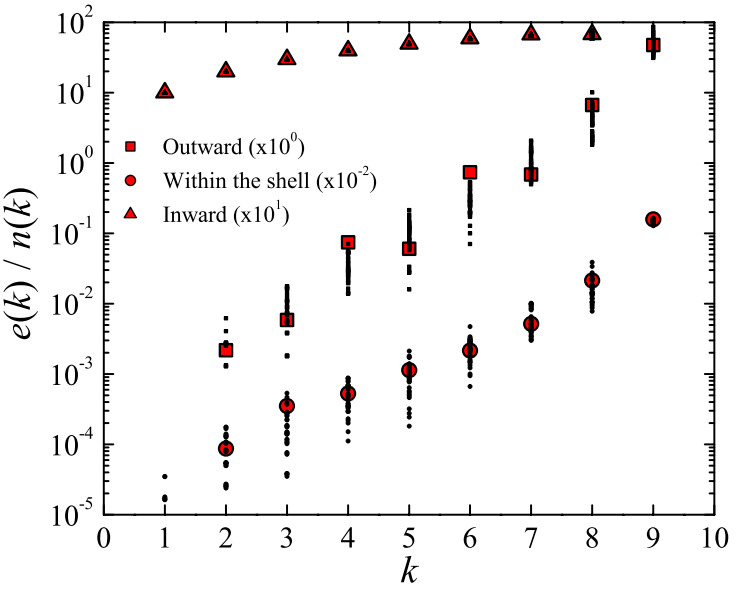


Figure 3.13: Average number $e(k)/n(k)$, of links per node that are radially outward (connecting to nodes in shells with $k' < k$), inward ($k' > k$), and within the shell ($k' = k$), as a function of the shell-number $k$. The labels "outward" and "inward" refer only to the circular arrangement of nodes placed at a greater or smaller distance from the center of the figures in the $k$-core visualizations employed in Figs. (3.3, 3.4), where the directionality of the edges is ignored. The red symbols pertain to Yeastract [37] data whereas the black symbols represent 36 model realizations with $k_{\max} = 9$. Note that the different sets of model values and data points have been shifted with respect to each other for greater clarity, the "inward" connections upwards by one decade, and the same-shell connections, downwards by two decades.

shell, are similar. However the identical exponential growth in the number of links connecting any given shell to shells deeper and deeper in the core-structure, is a strongly distinguishing feature of yeast and the model network which we are considering here.

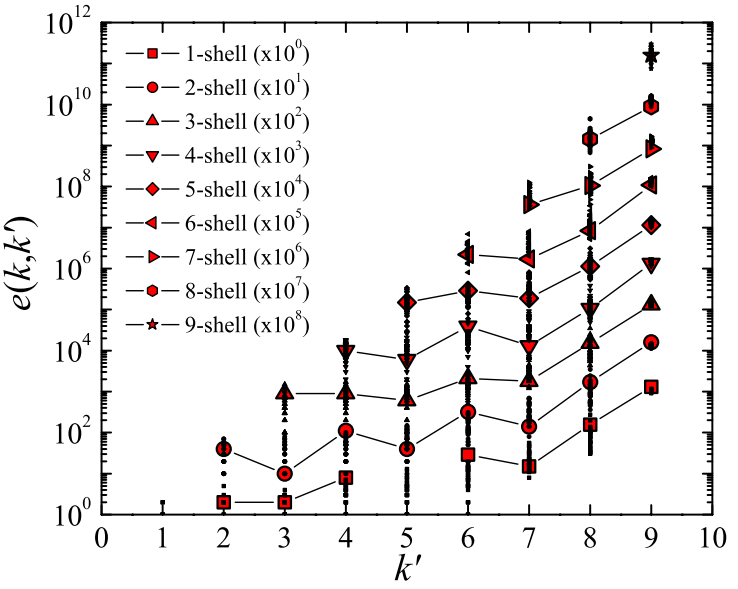


Figure 3.14: Distribution of the number $e(k, k')$, of links connecting nodes in various $k$-shells (different symbols) with nodes in $k' \geq k$ shells. The red symbols pertain to Yeastract [37] data whereas the black symbols represent 36 model realizations with $k_{\max} = 9$. The different $k$-series have been offset with respect to each other for greater visibility, and the experimental points of each series connected as a guide to the eye.

### 3.3.2 Comparison with other values of $\mu$

We have varied $\mu$ within the range $-0.5 \leq \mu \leq 2.5$, to see how far this affects the topological features of the model network qualitatively and quantitatively (see Table 3.2 for a summary of ensemble averages of our model). For $\mu > 2$ the distribution is not fat tailed (Almirantis and Provata [66] suggest $0 \leq \mu \leq 2$) and for smaller values the fluctuations would totaly dominate. (Note that the distribution stays normalizable since the range of $l$ is finite.) We find that the relatively larger $\mu$ values result in fewer $k$-shells, with the average value of $k_{\max}$ ranging from 10.61 to 3.27 for $-0.5 \leq \mu \leq 2.5$ (see Fig. 3.15). The plots corresponding to Fig. 3.14 are successively truncated from the left as $\mu$ is increased, so that the hierarchical structure of the $k$-cores is, nevertheless, preserved throughout this range of $\mu$ values.

Table 3.2: The summary of our model ensemble with the power law distribution of PR lengths. The average number of interacting genes, regulated genes, regulatory genes (coding TFs), and interacting pairs obtained from one hundred realizations of our model ($\pm$ the standard deviations) with $\mu$ in the range $-0.5 \leq \mu \leq 2.5$.

| $\mu$ | Genes | Regulated | Regulatory | Interacting Pairs |
|---|---|---|---|---|
| -0.5 | $4962 \pm 130$ | $4927 \pm 135$ | $203 \pm 14$ | $21727 \pm 2798$ |
| 0.0 | $4290 \pm 163$ | $4231 \pm 168$ | $201 \pm 14$ | $15356 \pm 1996$ |
| 0.5 | $3554 \pm 210$ | $3472 \pm 212$ | $193 \pm 15$ | $10085 \pm 1407$ |
| 1.0 | $2864 \pm 198$ | $2769 \pm 199$ | $179 \pm 13$ | $6310 \pm 862$ |
| 1.5 | $2365 \pm 220$ | $2261 \pm 222$ | $168 \pm 13$ | $4218 \pm 625$ |
| 2.0 | $1985 \pm 212$ | $1877 \pm 215$ | $156 \pm 12$ | $2974 \pm 452$ |
| 2.5 | $1714 \pm 206$ | $1619 \pm 206$ | $130 \pm 11$ | $2285 \pm 398$ |

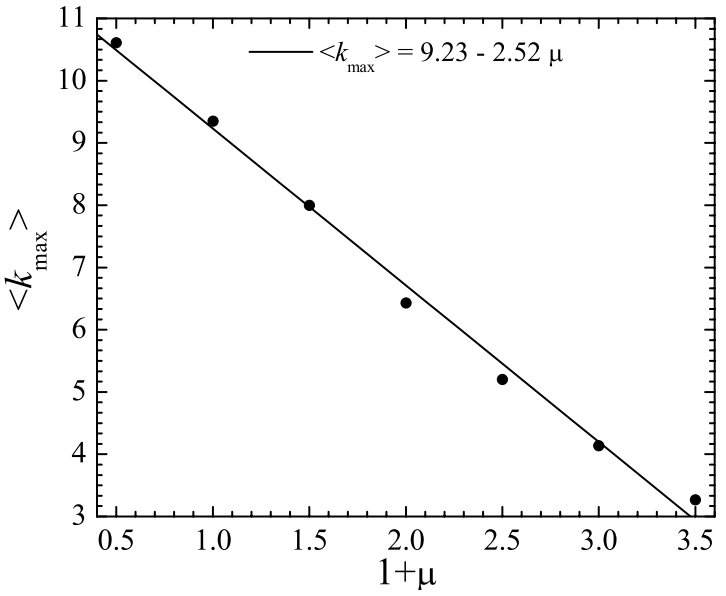


Figure 3.15: Average number of $k$-shells as a function of $1 + \mu$, the exponent of the PR length distribution, $p_{\mathrm{PR}}(l) \propto l^{-1-\mu}$.

In Fig. 3.16 we show the topological features of the model network, computed for the relatively large value of $\mu = 2$. The figures here are qualitatively somewhat similar to Figs. (3.6–3.11), although the quantitative agreement achieved for $\mu = 0.1$ is lost. For the lower value of $\mu = 0$, the behavior is very similar to that for $\mu = 0.1$, so we have not displayed it here. It should be remarked that the degree distribution is much less sensitive to the change in $\mu$ as compared to the clustering coefficient, the degree-degree correlations and the rich-club coefficient.

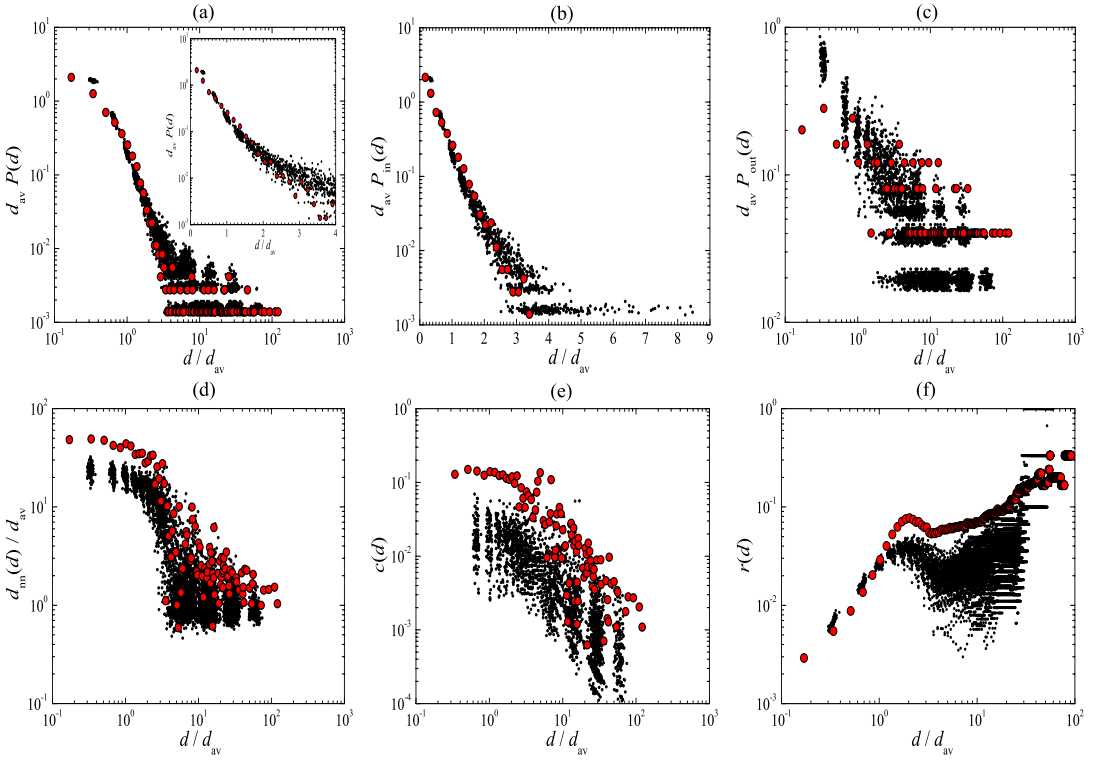


Figure 3.16: The topological features of the model network computed for $\mu = 2$, compared with those of the Yeastract [37] data. For larger values, the length distribution would cease to be fat tailed as suggested by Almirantis and Provata [66].

### 3.3.3 A null-hypothesis for the length distribution of the target sequences

There is a degree of arbitrariness in the way we have assumed that the lengths of the promoter regions should follow a power law distribution as do the intergenic regions [66]. Therefore we decided to test the null-hypothesis that they are all of the same length, $L$.

Trying out different $L$ values shows that e.g., for $L = 50$, the degree distributions

are qualitatively similar to those shown in Figs.(3.6–3.8), with a slightly faster decay of the in-degree distribution. The $k$-core structure has the same hierarchical appearance as in Fig. 3.3, but with $\langle k_{\max} \rangle = 4.35$. Nevertheless, the distribution of the nodes over the different $k$-shells, follows approximately the same $k^{-1}$ decay as found before (see Fig. 3.12).

For larger $L$, we find that a peak forms in the in-degree distribution and moves to the right as $L \geq 100$, while for $L = 200$ one may indeed observe two humps. The average number of $k$-shells at this point is $\langle k_{\max} \rangle = 13.5$. The hierarchical organization of the connectivity is preserved, although the distribution of the nodes over the different $k$-shells becomes non-monotone for $L \geq 100$.

Plotting the number of $k$-cores against $L$, it is easy to find that for $L = 127$, the $k$-core plot shown in Fig. 3.17 almost coincides exactly with Figs. (3.3, 3.4), with $k_{\max} = 9$. However, the detailed qualitative and quantitative agreement with the topological features of the yeast network is lost, as shown in Fig. 3.18 and Fig. 3.19. The distribution of the number of nodes over different $k$-shells may be compared with that of the internet in [31]. We may note here that although the average number of nodes taking part in the connected networks is 5897 (out of 6000) with 25046 interactions, again on average, it is interesting to see that the average number of $k$-shells just goes up to 9.

## 3.4 Randomization Procedures and Null-Null Models

In this section we briefly discuss comparing the topological features of our model with randomized versions thereof, and with "null-null" models which incorporate only a few elements of our model, selected to mimic certain phenomenological properties of the target network.

### 3.4.1 Randomizing the edges of the model network and the yeast network

To check the significance of our results, we compared the $k$-core structures, the clustering coefficients, the degree-degree correlations, and the rich-club coefficients of the Yeastract data with those obtained after randomly reconnecting the

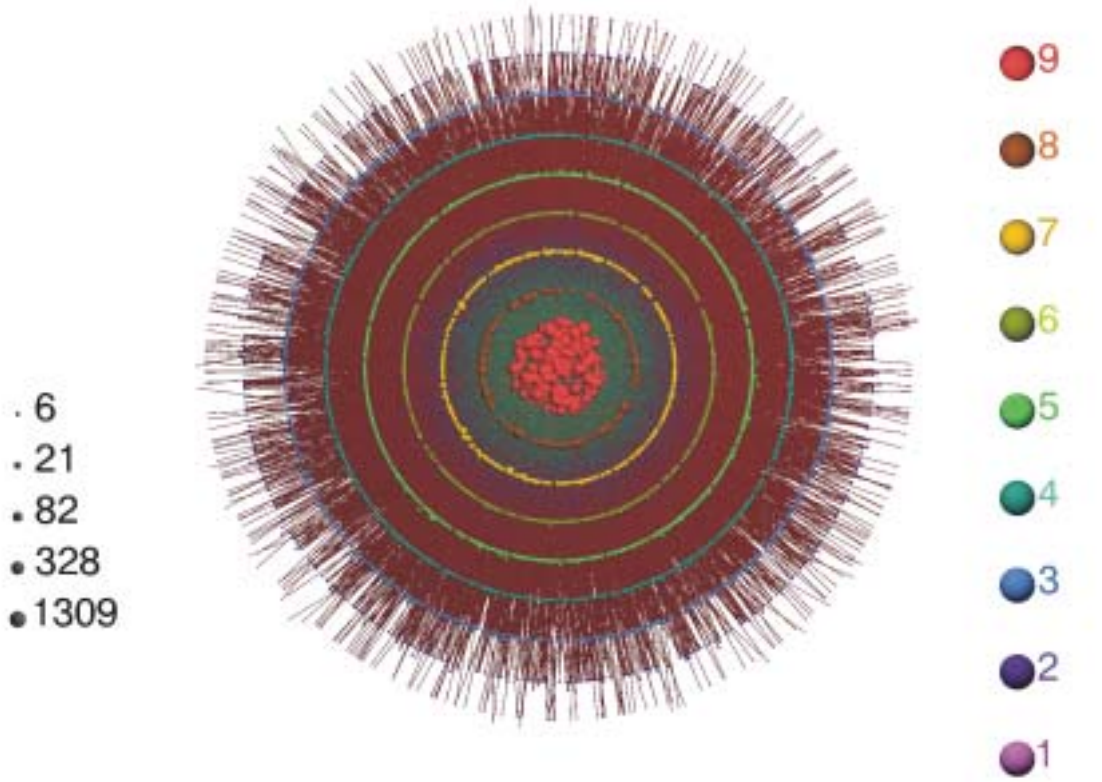


Figure 3.17: The $k$-core visualization of one realization of the model network with the lengths of the PR sequences fixed at $L = 127$. The value of $L$ was chosen to make $\langle k_{\max} \rangle = 9$, to coincide with the corresponding value for the Yeastract [37] data.

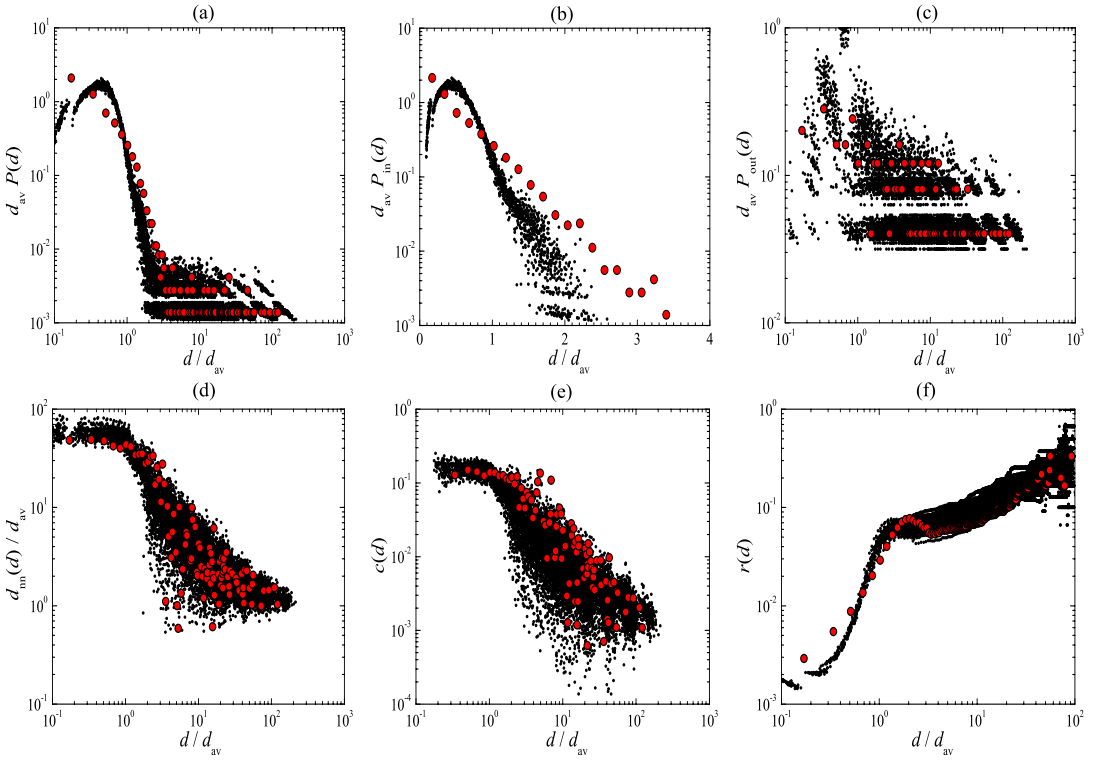


Figure 3.18: Superposition of Yeastract [37] results and the scatter plots obtained from one hundred realizations of the model networks, with the lengths of the PRs fixed at $L = 127$. The value of $L$ was chosen to make $\langle k_{\max} \rangle = 9$. Although the plots look superficially like the ones reported in Figs. (3.6–3.11), note, in particular, that the exponential decay of the in-degree distribution and the resulting total degree distribution have been modified. The curve has acquired a maximum at around $d/d_{\rm av} = 0.5$ and a faster than exponential decay. The rich-club coefficient has lost the distinctive shoulder.

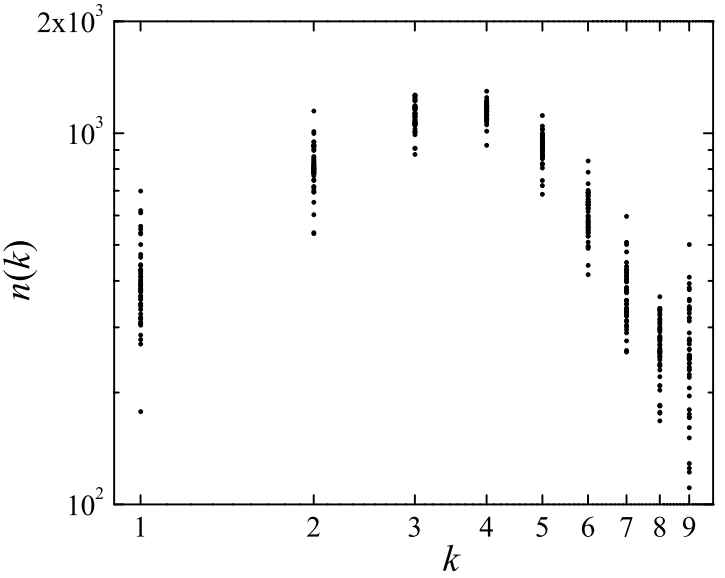


Figure 3.19: The distribution of nodes over different $k$-shells, for the null hypothesis for the lengths of the promoter regions, namely a fixed PR length of $L = 127$. Only the 49 realizations with $k_{\max} = 9$ (out of 100) of the respective model have been shown. At this value of the PR length, one finds a superficial similarity of the $k$-core visualization of this reduced model network with that of yeast. This figure should be contrasted with the distribution for the null-model shown in Fig. 3.14.

edges of the real and the model networks while *i)* keeping the in- and out-degree of each node fixed separately and *ii)* only keeping the degree of each node fixed, irrespective of the directionality.

In Fig. 3.20a we display the randomized $k$-core plots for yeast and the same realization of the model network whose $k$-core plot was displayed in Fig. 3.3. We randomly choose pairs of edges and exchange their ends, either the in with in, or the out with out, by avoiding the occurrence of a particular interaction twice. We repeat this procedure $2 \times E$ times, where $E$ is the total number of edges. This preserves the directionality of the edges, as well as the in- and out-degree of each node separately. In Fig. 3.20b we show parallel results obtained when we ignore the directionality of the edges under the random rewiring procedure, but conserve the total degree of each node.

We see that the randomization procedure which preserves directionality essentially leaves the $k$-core structure invariant, while randomizing without respecting directionality of the edges produces a strikingly different picture, both for the yeast and the model network; $k_{\max}$ becomes $29\pm1$ rather than 9. While in the

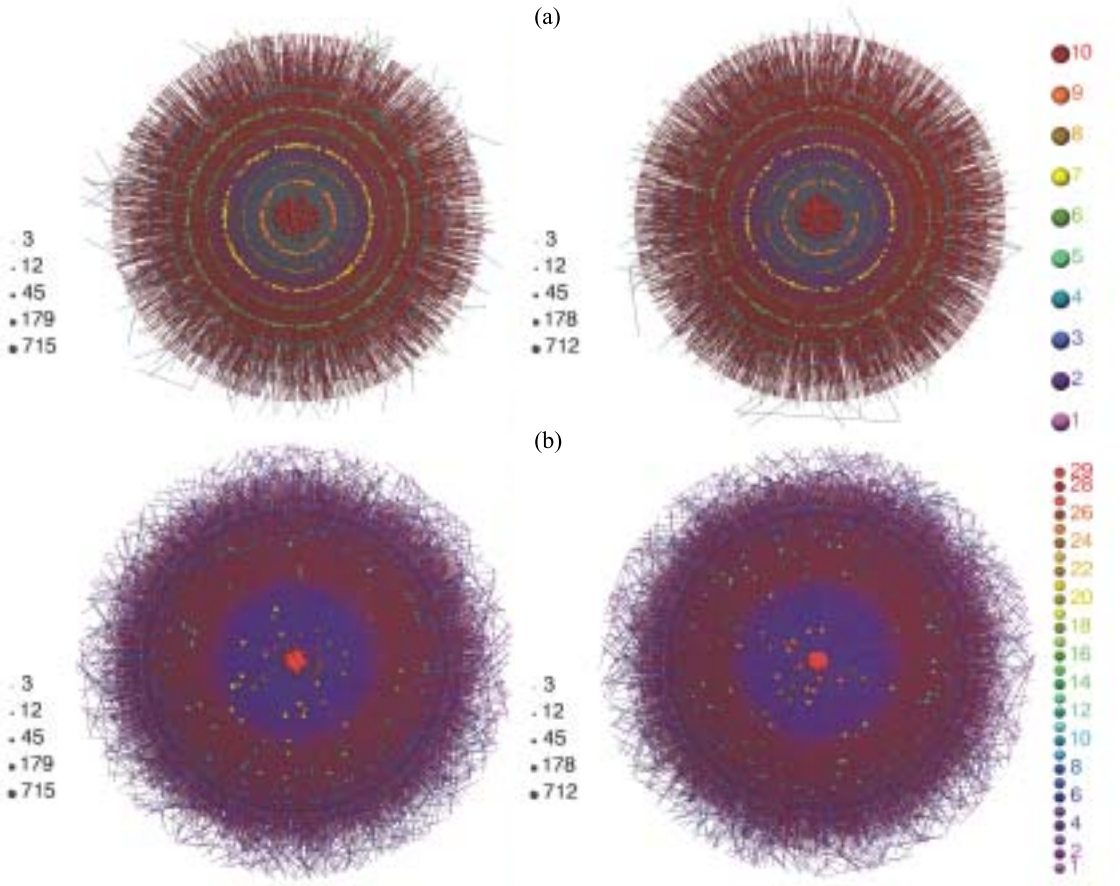


Figure 3.20: The $k$-core visualizations of the randomized versions of one realization of the model (left panel) and Yeastract (right panel) networks, preserving separately the in- and out-degrees (a) or the total degree (b) of each node. The first set of $k$-core plots (a) are essentially indistinguishable from Figs. (3.3, 3.4). In contrast, those obtained by the randomization procedure ignoring the directionality of the edges are strikingly different. The number of shells have gone up to 29 from 9, and the much higher intra-shell rather than inter-shell connectivity (as can be seen by following the edges) indicates that the hierarchical nature of the yeast network, which is faithfully reproduced by the model, is destroyed by the nondirectional randomization process.

yeast and model networks, the largest fraction of connections is to the innermost shell, the $k_{\max}$-core (see Figs. (3.3, 3.4) and Fig. 3.14), in the randomized networks without direction conservation, there is a high degree of intra-shell connectivity. The topological coefficients are also completely altered in this case, as displayed in Fig. 3.21.

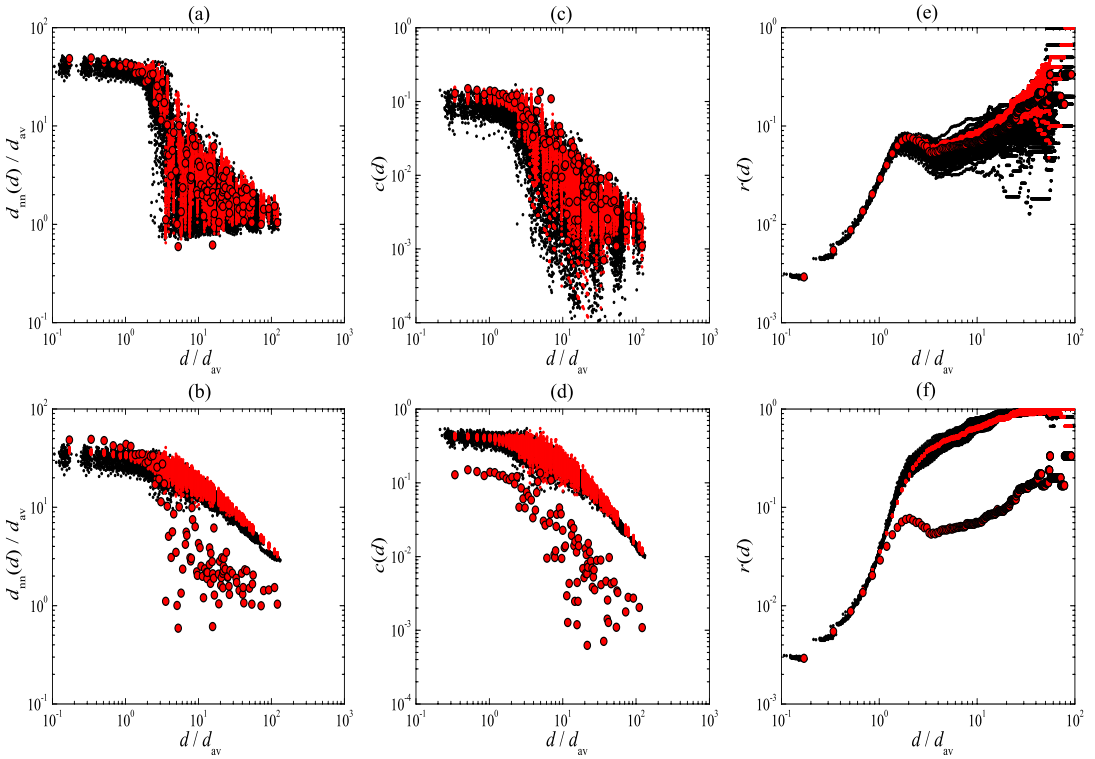


Figure 3.21: The effect of the same randomization procedure in as in Fig. 3.20 on the degree-degree correlations between neighboring nodes (a,b), the clustering coefficient (c,d), and the rich-club coefficient (e,f) of the Yeastract data (red circles) and one hundred realizations of the model network. In the top row, we show the results of keeping in- and out-degrees of each node fixed, while the bottom row only preserves the total degree of each node. The black dots correspond to one randomization of each of one hundred realizations of the model network and the red dots to an ensemble of one hundred independent randomizations of the Yeast network.

It should be noted that, in our model the lengths (and the contents) of the sequences representing the binding motifs and the PR associated with the same node are assigned independently of each other. Thus there is no correlation between the in- and out-degree of a given node. Our model is, therefore, a null model in this respect. Invariance of the topological features under a random rewiring which conserves the in- and out- degree distributions, suggests that the in- and out-degree distributions together are able to determine *all* of the global topolog-

ical features of the network in question. The achievement of our model is that it does not have to import these degree distributions from empirical data; it is able to capture them by means of the string-matching rule and the length distributions of the regulatory and PR sequences appropriate to the organism under study.

### 3.4.2 The configuration model

An ensemble of networks generated from one model realization by randomizing the edges while keeping the directionality fixed, as is done above, is in fact equivalent to a configuration model [68]. To double check our randomization procedure, we have simulated an ensemble of configuration model networks by taking the in- and out-degree sequence of each realization of the content-based network, removing the edges, exchanging the in- and the out-degree assignments between randomly chosen pairs of nodes to remove any possible residual correlations between these quantities, attaching corresponding numbers of arrows to the nodes, and then randomly connecting pairs of in and out arrows.

The in- and out-degree distributions of resulting networks are identical to the content-based model, and the rest of the topological features are indistinguishable from each other, to the extent that they are determined by the in- and out-degree distributions. The results agree very closely with the topological properties of the networks as computed from Yeastract data (see Figs. (3.9–3.11) and the top row of Fig. 3.21). The crucial fact to keep in mind is that the in- and out-degree distributions of our model are *not* imported from any data set, but independently generated by the information sharing mechanism embodied in the string-matching condition underlying our model, given the effective length distributions for the binding motifs and the PRs extracted from the yeast data.

### 3.4.3 A modified Erdös-Rényi model

To see whether we could reproduce certain features of the yeast TRN using only the fact that there are two types of nodes in this network, those coding for TFs, and others that do not, we constructed an Erdös-Rényi type of null-null model by picking a subset (of size $N_{\mathrm{TF}}$) of all the nodes ($N$), and allowing only these

(i.e., TF-coding) nodes to have out-going edges to randomly picked nodes in the whole network, with a probability $p$. This probability is to be fixed by the density of edges on the real yeast TRN, i.e., $p = E/E_{\max}$, where $E$ is the number of edges, and $E_{\max} = NN_{\rm TF}$ the total number of possible edges in this network. The resulting network (see Fig. 3.22) has a bi-modal degree distribution consisting of the superposition of two well-separated Poissonian peaks, centered over the mean in- and out-degree contributions, $E/N$ and $E/N_{\rm TF}$.

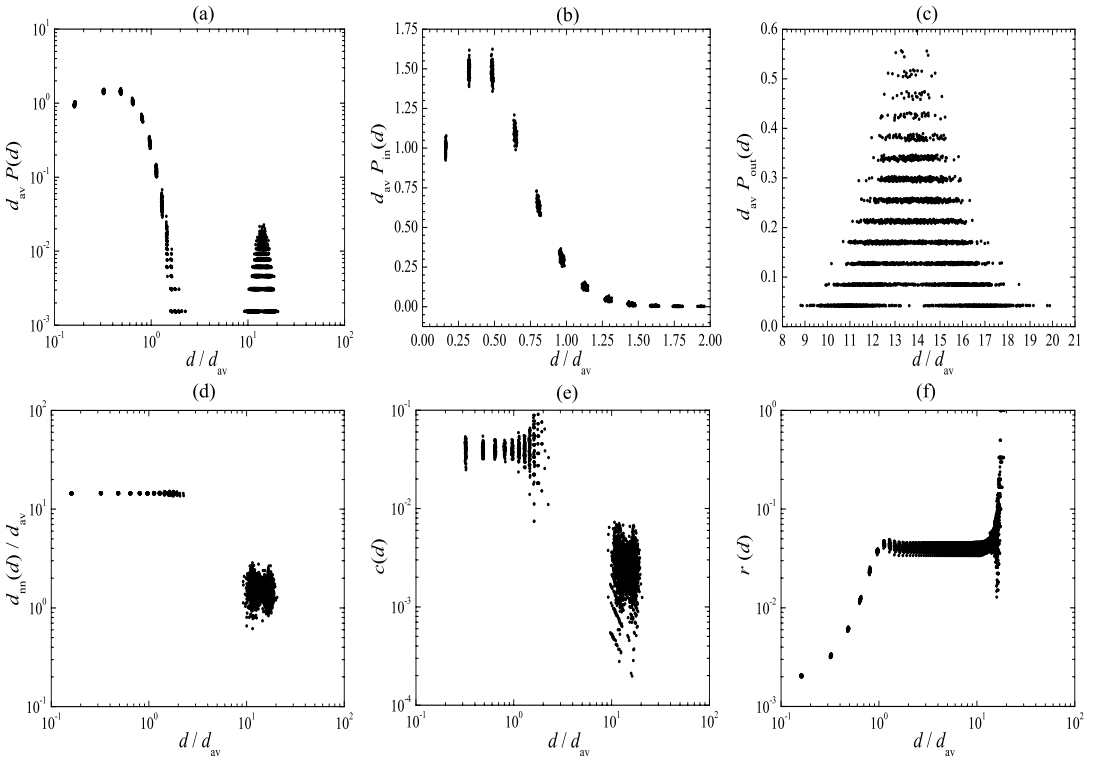


Figure 3.22: The topological features of the Erdös-Rényi random network version of the TRN, with the connection probability determined only from the density of edges of the empirical network (Yeastract [37] data). The figures have been obtained from one hundred independent realizations of the adapted Erdös-Rényi random model.

The topological features are qualitatively different from our null-model and the yeast TRN in all respects. In particular, the bi-modal degree distribution gets reflected in disconnected plots of the degree-degree correlation and the clustering coefficient, with essentially degree-independent small-degree region (coming from the relatively small in-degrees) and a disconnected large degree part coming from the out-degrees of the TF-coding nodes. The $k$-core decomposition, on the other hand, indicates a highly hierarchical structure and looks indistinguishable from Figs. (3.3, 3.4), with $k_{\max} = 7$. A closer inspection reveals, however, that the

distribution of the number of nodes over the different shells is in fact qualitatively different, resembling that found when the lengths of all the PR sequences are set to a large number well separated from the lengths of regulatory sequences (see Section 3.3 and Fig. 3.19).

### 3.4.4 Comparison with a hidden-variable model

A coarse-grained, or mean-field, version of our model is obtained if, instead of the fluctuations coming from the chance coincidence of individual strings, one takes the ensemble averaged probability for a matching to occur between strings of given lengths, as in Eq. 2.1.3. This can be thought of as a hidden-variable model [69], where, instead of just the two types of nodes considered above, one has a superposition of a whole spectrum of Erdös-Rényi networks, with the connection probabilities $p(l, k)$.

A mean-field version of our model can be constructed by assigning to each node $i$, two random variables $(l_i,\ k_i)$, distributed in the same way as the lengths of the regulatory and the PR strings associated with the organism under consideration. (To account for the fact that only a fraction $\eta = 4.8\%$ of the nodes code TFs, in practice, $1 - \eta$ of the nodes are assigned binding motif lengths exceeding the maximum PR lengths.) These hidden-variables take the place of the regulatory and PR strings associated with the nodes. To simulate this effective model, we use the ensemble averaged probabilities $p(l_i, k_j)$ and $p(l_j, k_i)$ derived from the string matching condition [6], for inserting directed edges between the nodes $(i, j)$ (see Eq. 2.1.3).

We show in Fig. 3.23 the simulation results for the topological features of the hidden-variable model, averaged over one hundred realizations and superposed on the ensemble averages of our content-based model. The ensemble averaged results of the hidden-variable model and the content-based model are very close to each other, except that the content-based model has an in-degree distribution with a longer tail (in the range $15 \leq d \leq 32$) than the hidden-variable model, albeit with very small probabilities. This gives rise to small differences in the other topological features in this degree range. The subtle difference in the rare

event range of the in-degree distribution may arise from an underestimation of the connection probabilities $p(l, k)$ by the mean field approximation [6].

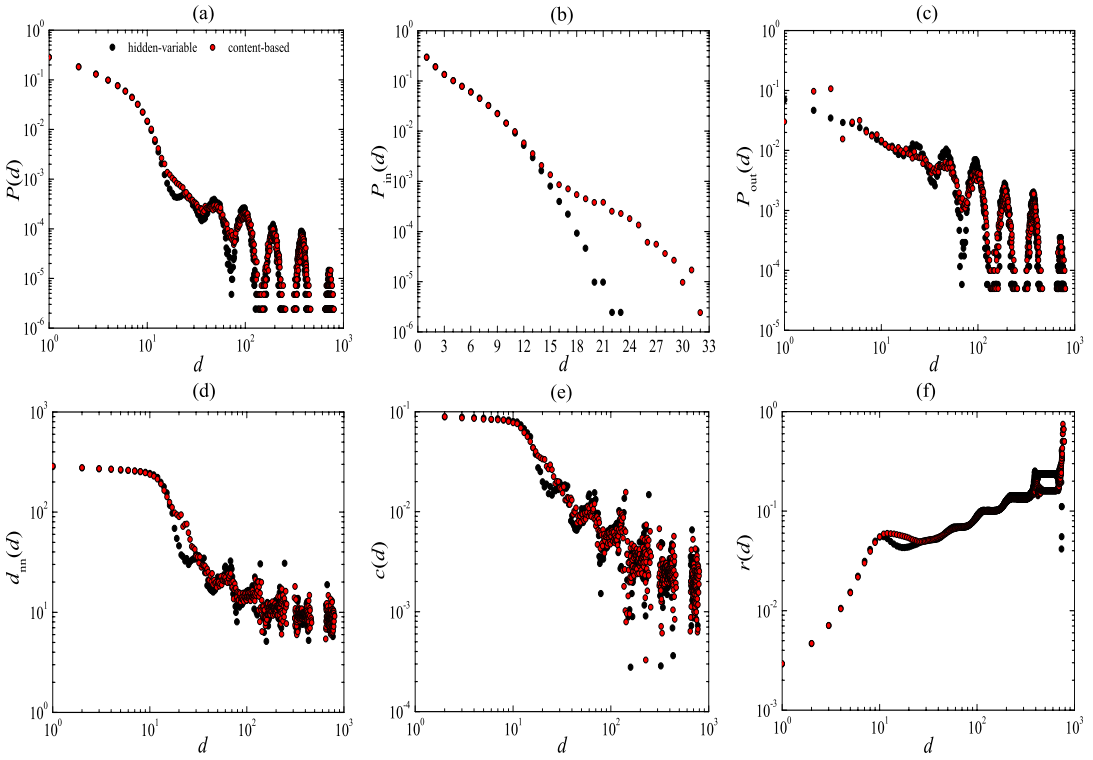


Figure 3.23: The topological features of the hidden-variable model, with $N = 6000$ nodes, ensemble averaged over one hundred realizations, superposed on the ensemble averaged results for our null-model whose scatter plots have been displayed in Figs. (3.6–3.11).

The fidelity of the mean-field version to the yeast TRN is indistinguishable from that of the full content-based model. This gives us confidence that analytical calculations of ensemble averaged properties are quite meaningful. It should be remembered that *i)* the length distributions of the PR or regulatory strings have been extracted from empirical data using our information-theoretical approach to the binding specificities of the binding motifs, and *ii)* that the connection probabilities $p(l, k)$ were derived from the string-matching condition.

## 3.5 Comparison with Other Databases

The agreement observed with the Yeastract [37] data is not source-specific, as can be seen from a comparison of the topological properties of our model networks, with those obtained from the different sources listed in Table 3.1. We emphasize

that the agreement of all four data sets with the model ensemble is achieved while the parameter $\mu$ has only been optimized with respect to the Yeastract data set.

We display in Fig. 3.24, the topological coefficients as computed from different data sources for the yeast TRN, superposed on the scatter plots for the ensemble of model networks. We see that although there are partial differences between the data sets, all are compatible with the model results. Since the different data sets encompass slightly different sets of genes, this goes further to show that the model captures the essential building principles of the networks.

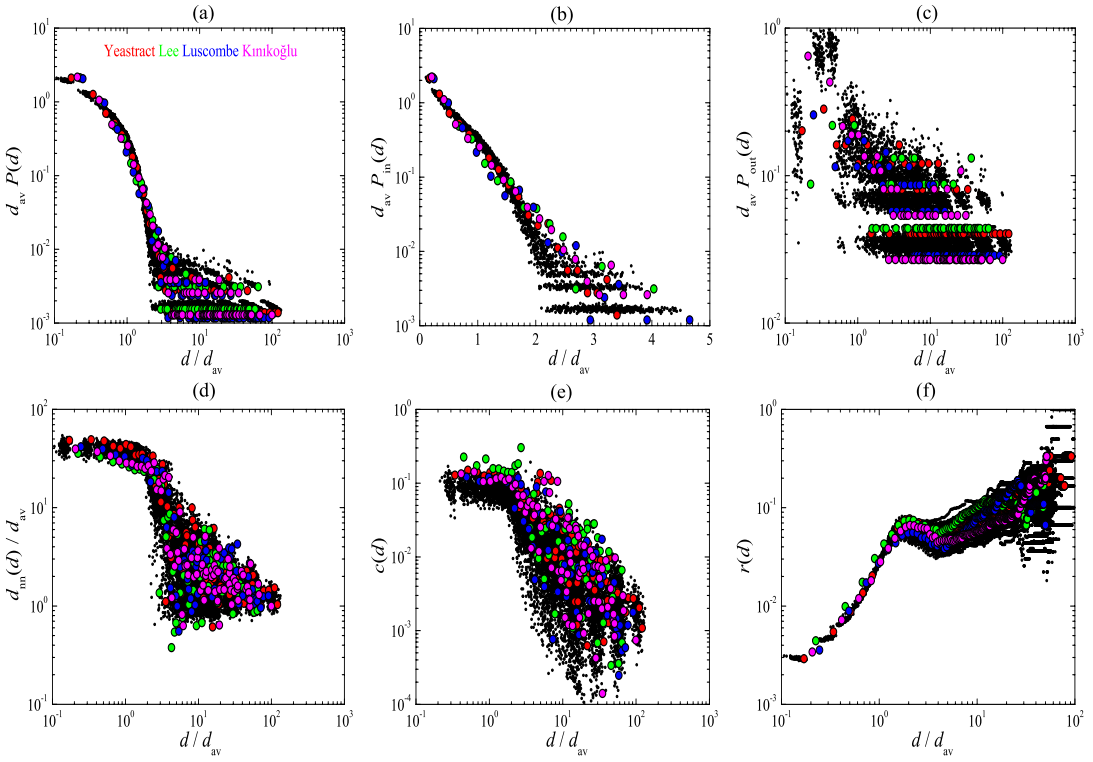


Figure 3.24: The network statistics extracted from the sources listed in Table 3.1 superposed on the simulation results corresponding to one hundred realizations of the model network (black dots). The agreement is extremely good with all of these sets of data, which almost completely cover, but do not exceed the phase space of our model. (Red, green, blue, and magenta correspond to the Yeastract [37], Lee [35], Luscombe [36], and Kınıkoğlu [38] data respectively.)

## 3.6 Discussion

Our results support our hypothesis that the topology of the TRN is predominantly determined by the sequence matching rule, which schematizes the shared information involved in the interaction between the genes. The close structural

similarity between the model and the real yeast transcriptional regulatory network, with respect to a diverse set of criteria shows that they are part of the same statistical ensemble of networks [70].

The sequence matching rule should be viewed as an information-theoretical constraint, where the interaction between two genes requires the fulfillment of a set of conditions which we symbolically represent as the matching of two random sequences. The more stringent the prerequisites of the interaction, the longer is the random binding motif that is to be matched. The length of the PR establishes the size of the phase space in which the motif is to be sought. The properties of the network are then determined by the distributions obeyed by the lengths of the binding motifs as well as the promoter regions.

Interpreted within an information-theoretical framework, our model has sufficient generality to accommodate other interactions based on constraint satisfaction mechanisms, such as protein-protein interaction networks, where the interactions are dictated by certain steric and chemical conditions.

The topological features of the networks investigated here and shown to be shared by the yeast transcriptional regulatory network strongly point to the possibility that these networks did not have to be assembled from scratch, but rather emerged spontaneously, given any sufficiently long, complex linear code, and a mechanism for the transcription of some of its subsequences into molecules (proteins) that in their turn have an affinity for parts of this code and bind it. This proposition by no means minimizes the role of evolutionary pressures on such networks; instead, it suggests that a network with essentially the current topology could have provided a starting point for further fine-tuning.

# 4 ANALYTICAL CALCULATIONS ON THE HIDDEN-VARIABLE MODEL

Here we present the hidden-variable or mean-field version of the content-based approach [9] proposed as a null model of the transcriptional regulation network of yeast (see Section 3). We carry out the analytical calculations for the degree distributions [14, 15, 16], the degree-degree correlation [20, 21] of nearest neighbors, the clustering coefficient [22, 23, 24], and the rich-club coefficient [26, 27] on the ensemble of hidden-variable networks. We drive the analytical calculations as far as we can and evaluate the expressions numerically to display the results. We also provide the comparison of our analytical results with those of the simulations of the mean-field version of our content-based model. The simulation results we display here for the ensemble averages of the quantities under consideration have been computed over one hundred realizations of the hidden-variable model.

The hidden-variable or mean-field version of our content-based model may be constructed by taking $N$ nodes and assigning two random variables, $l$ and $k$, to each of them, then establishing the directed edge originating from node $i$ and terminating at node $j$ with respect to their random variables, $l_i$ and $k_j$. The directed edges of the content-based network are drawn between the regulatory sequences and the promoter regions of pairs of nodes if the amount of information contained in the regulatory sequence of the first node is shared by the promoter region of the second. This is made concrete by requiring that the regulatory sequence of the first node occurs as a subsequence in the promoter region of the second node. In the mean-field version, the first and second variable, $l$ and $k$, assigned to each node represent the lengths of the regulatory sequence and the promoter region, respectively. If $l_i \leq k_j$ the directed link from node $i$ to node $j$ is drawn with the probability $p(l_i, k_j)$,

$$p(l_i, k_j) = 1 - \left(1 - \frac{1}{r^{l_i}}\right)^{k_j - l_i + 1} \quad , \tag{4.0.1}$$

which may be seen as a zeroth order approximation to the probability of occurrence of a randomly selected sequence of length $l_i$ in a randomly selected sequence of length $k_j$ if the sequences are chosen from a common alphabet of $r$ letters where each of them has equal chance, $1/r$, to occur in a random sequence [6]. Thus, the element of the adjacency matrix, $w_{ij}$ may be written as

$$w_{ij} = \begin{cases} 1 & \text{with probability } p(l_i, k_j) \text{ if } l_i \leq k_j \\ 0 & \text{otherwise} \end{cases} \quad , \tag{4.0.2}$$

and implies that the nodes having an RS of length exceeding the lengths of PRs will not contribute to the number of out-going edges, and the ones with a PR of length less than the lengths of RSs will not contribute to the in-coming edges. The distribution $\widetilde{p}_{\rm RS}(l)$, of the first random variable $l$,

$$\widetilde{p}_{\rm RS}(l) = \eta_{\rm RS} p_{\rm RS}(l) + (1 - \eta_{\rm RS}) q_{\rm RS}(l) \quad , \tag{4.0.3}$$

where $p_{\rm RS}(l)$ is defined in the interval $\Lambda_{\rm RS} = [l_{\rm min}, l_{\rm max}]$, is the biological input to the model and may change according to the organism under consideration. On the other hand, the distribution $\widetilde{p}_{\rm PR}(k)$, of the second variable $k$,

$$\widetilde{p}_{\rm PR}(k) = \eta_{\rm PR} p_{\rm PR}(k) + (1 - \eta_{\rm PR}) q_{\rm PR}(k) \quad , \tag{4.0.4}$$

where $p_{\rm PR}(k)$ is defined in the interval $\Lambda_{\rm PR} = [k_{\rm min}, k_{\rm max}]$, is more flexible, and we have also introduced the parameter $\eta_{\rm PR}$ to drive more general results. (This parameter does not exist in the content-based model presented in Section 3, where $\eta_{\rm PR} = 1$.) These two distributions, $p_{\rm RS}(l)$ and $p_{\rm PR}(k)$, will totally determine the topological properties of the ensemble of networks in question. The distributions $q_{\rm RS}(l)$ and $q_{\rm PR}(k)$ are defined in the regions $l > k_{\rm max}$ and $k < l_{\rm min}$, respectively. Note that the variables (RS and PR lengths) attached to those nodes having potential to contribute to the interactions in the network obey the distributions $p_{\rm RS}(l)$ and $p_{\rm PR}(k)$, not the $\widetilde{p}_{\rm RS}(l)$ and $\widetilde{p}_{\rm PR}(k)$. Defining the distributions of random variables by $\widetilde{p}_{\rm RS}(l)$ and $\widetilde{p}_{\rm PR}(k)$ is just a way of saying that only the $\eta_{\rm RS}$ of nodes may have an RS and the $\eta_{\rm PR}$ of nodes may have a PR. Note again that all the distributions are normalized in the given intervals.

Below we calculate the distributions of number of nodes which may contribute to the interactions in the network and the degree distributions (out-, in-, and total),

to be used for the calculations of the two and three point correlations, namely the degree-degree correlation of nearest neighbors and the clustering coefficient, and the rich-club coefficient.

### 4.1 Fluctuations in Node and Edge Properties

Because of the probabilistic nature of the model the number of nodes with a given variable as well as the number of edges between the nodes with given variables fluctuates from one realization to another. Here we will define and calculate the corresponding distributions with their means and variances, to be used for the calculations of degree distributions.

Denoting the total number of nodes by $N$, the number $n_{\mathrm{RS}}(l)$ of nodes with RSs of length $l \in \Lambda_{\mathrm{RS}}$, and the number $n_{\mathrm{PR}}(k)$ of nodes with PRs of length $k \in \Lambda_{\mathrm{PR}}$ are binomially distributed,

$$P(n_{\mathrm{X}}(x)) = \binom{N}{n_{\mathrm{X}}(x)} \left[\eta_{\mathrm{X}} p_{\mathrm{X}}(x)\right]^{n_{\mathrm{X}}(x)} \left[1 - \eta_{\mathrm{X}} p_{\mathrm{X}}(x)\right]^{N - n_{\mathrm{X}}(x)} \quad , \tag{4.1.5}$$

with the mean $\langle n_{\mathrm{X}}(x) \rangle = N \eta_{\mathrm{X}} p_{\mathrm{X}}(x)$ and variance $\sigma^2_{n_{\mathrm{X}}(x)} = N \eta_{\mathrm{X}} p_{\mathrm{X}}(x) \left[1 - \eta_{\mathrm{X}} p_{\mathrm{X}}(x)\right]$, where X stands either for RS or PR; $p_{\mathrm{X}}(x)$ is the probability of finding a string X of length $x$ in a random realization of the model.

The total number of directed edges is given by the sum of the elements of the adjacency matrix, $e = \sum_{i,j} w_{ij}$ where it should be recalled that $w_{ij}$ is a random variable taking the value of 0 or 1. Note that the interaction matrix is not symmetric, $w_{ij} \neq w_{ji}$ in general. We may rewrite the number of edges as $e = \sum_l \sum_{i \in G_{\mathrm{RS}}(l)} \left[ \sum_{k \geq l} \sum_{j \in G_{\mathrm{PR}}(k)} w_{ij} \right] = \sum_k \sum_{j \in G_{\mathrm{PR}}(k)} \left[ \sum_{l \leq k} \sum_{i \in G_{\mathrm{RS}}(l)} w_{ij} \right]$ where the nodes have been grouped into sets labelled with their hidden variables [6], $G_{\mathrm{RS}}(l)$ and $G_{\mathrm{PR}}(k)$. Now let us define $e^{(i)}_{lk}$,

$$e^{(i)}_{lk} = \sum_{j \in G_{\mathrm{PR}}(k)} w_{ij} \quad , \tag{4.1.6}$$

as the number of edges originating from a randomly selected node $i$ with an RS of length $l$ and terminating at nodes with PRs of length $k$. The distribution of

the number of such edges is binomial,

$$P(e_{lk}^{(i)} = d_{lk}|n_{\rm PR}(k)) = \binom{n_{\rm PR}(k)}{d_{lk}} \ [p(l,k)]^{d_{lk}} \ [1-p(l,k)]^{n_{\rm PR}(k)-d_{lk}} \ ,$$

$$P(e_{lk}^{(i)} = d_{lk}) = \sum_{n_{\rm PR}(k)=d_{lk}}^{N} P(n_{\rm PR}(k)) \ P(e_{lk}^{(i)} = d_{lk}|n_{\rm PR}(k)) \ ,$$

$$P(e_{lk}^{(i)} = d_{lk}) = \binom{N}{d_{lk}} \left[\eta_{\rm PR} p_{\rm PR}(k) p(l,k)\right]^{d_{lk}} \left[1-\eta_{\rm PR} p_{\rm PR}(k) p(l,k)\right]^{N-d_{lk}} \ , \quad (4.1.7)$$

with the mean

$$\langle e_{lk}^{(i)} \rangle = N \eta_{\rm PR} p_{\rm PR}(k) p(l,k) \ , \quad (4.1.8)$$

and variance

$$\sigma^2_{e_{lk}^{(i)}} = N \eta_{\rm PR} p_{\rm PR}(k) p(l,k) \left[1-\eta_{\rm PR} p_{\rm PR}(k) p(l,k)\right] \ , \quad (4.1.9)$$

where we have used Eq. 4.1.5 for $P(n_{\rm PR}(k))$. We may similarly define $\widetilde{e}_{lk}^{(j)}$ as the number of edges originating from nodes with RSs of length $l$ and terminating at a randomly selected node $j$ with a PR of length $k$,

$$\widetilde{e}_{lk}^{(j)} = \sum_{i \in G_{\rm RS}(l)} w_{ij} \ . \quad (4.1.10)$$

The distribution of the number of such edges is again binomial,

$$P(\widetilde{e}_{lk}^{(j)} = \widetilde{d}_{lk}) = \binom{N}{\widetilde{d}_{lk}} \left[\eta_{\rm RS} p_{\rm RS}(l) p(l,k)\right]^{\widetilde{d}_{lk}} \left[1-\eta_{\rm RS} p_{\rm RS}(l) p(l,k)\right]^{N-\widetilde{d}_{lk}} \ , \quad (4.1.11)$$

with the mean

$$\langle \widetilde{e}_{lk}^{(j)} \rangle = N \eta_{\rm RS} p_{\rm RS}(l) p(l,k) \ , \quad (4.1.12)$$

and variance

$$\sigma^2_{\widetilde{e}_{lk}^{(j)}} = N \eta_{\rm RS} p_{\rm RS}(l) p(l,k) \left[1-\eta_{\rm RS} p_{\rm RS}(l) p(l,k)\right] \ . \quad (4.1.13)$$

## 4.2 Degree Distributions

Here we calculate the probabilities of finding a node with given out-, in-, or total degree if we randomly pick a node.

Let us start with the out-degree distribution. The total number of edges $e_l^{(i)}$, originating from a randomly chosen node $i$ with an RS of length $l$, is nothing but the sum of random variables,

$$e_l^{(i)} = \sum_{k \geq l} \sum_{j \in G_{\rm PR}(k)} w_{ij} = \sum_{k \geq l} e_{lk}^{(i)} \ . \quad (4.2.14)$$

In the limit of very large network size, the distribution $P_l^{\rm out}(e_l^{(i)} = d)$ of the sum in Eq. 4.2.14, which is the probability of finding a node with out-degree $d$ if we randomly choose a node among the nodes with RSs of length $l$, is binomial with the mean $d_{{\rm o},\,l}$,

$$d_{{\rm o},\,l} = \sum_{k \geq l} \langle e_{lk}^{(i)} \rangle = \sum_{k=\max(l,\,k_{\rm min})}^{k_{\rm max}} N \eta_{\rm PR} p_{\rm PR}(k) p(l,k) \quad , \tag{4.2.15}$$

and variance $\sigma^2_{{\rm o},\,l}$,

$$\sigma^2_{{\rm o},\,l} = \sum_{k \geq l} \sigma^2_{e_{lk}^{(i)}} = \sum_{k=\max(l,\,k_{\rm min})}^{k_{\rm max}} N \eta_{\rm PR} p_{\rm PR}(k) p(l,k) \left[1 - \eta_{\rm PR} p_{\rm PR}(k) p(l,k)\right] \quad . \tag{4.2.16}$$

We display in Fig. 4.1a, the relationship between the average out-degree and RS-length, where we have used the parameters as used during the simulations of the hidden-variable model (see Fig. 4.4 for the length distributions used). We find that the average out-degree decreases exponentially with RS-length. We also exhibit in Fig. 4.1b for comparison, the differences between the means and variances of individual out-degree distributions and see that the variances are very close to the mean values for the parameters used.

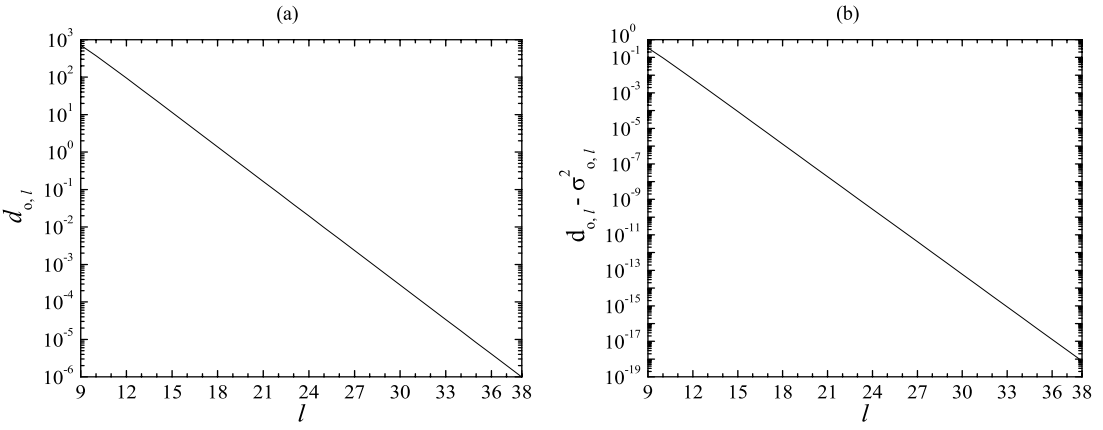


Figure 4.1: The means (a) and variances (b) of out-degree distributions. The length distribution of PRs have been assumed to be of power-law form, $p_{\rm PR}(k) \propto k^{-(1+\mu)}$ with $\mu = 0.1$, confined to the interval $\Lambda_{\rm PR} = [13, 262]$. The size of the network is $N = 6000$ and $\eta_{\rm PR} = 1$. The average out-degree $d_{{\rm o},\,l}$ of nodes with RSs of length $l$ decrease exponentially with $l$. The means and the variances are almost equal to each other for the parameters used. (See Section 3 for the details of the parameter choices.)

Remembering that if a node has an RS of length exceeding the lengths of PRs then its out-degree is zero, it is very easy to write the expression for the out-degree

distribution $P_{\rm out}(d)$,

$$P_{\rm out}(d) = \sum_l \widetilde{p}_{\rm RS}(l) P_{\rm out}(d|l) = \eta_{\rm RS} \sum_{l \in \Lambda_{\rm RS}} p_{\rm RS}(l) P_l^{\rm out}(d) + (1 - \eta_{\rm RS})\delta(d) \quad . \quad (4.2.17)$$

Recall that the sum in Eq. 4.2.14 follows a binomial distribution with the mean and variance given by Eqs. (4.2.15, 4.2.16). In the limit of large number of nodes and small probabilities (thus, we can neglect the higher order terms of the probabilities $\eta_{\rm PR} p_{\rm PR}(k) p(l,k)$) we may approximate the out-degree distribution of the nodes with RSs of length $l$ by a Poisson distribution with the mean $d_{{\rm o},\,l}$ (see Eq. 4.2.15),

$$P_l^{\rm out}(d) = \exp[-d_{{\rm o},\,l}] \; \frac{(d_{{\rm o},\,l})^d}{d!} \quad , \quad (4.2.18)$$

which proves to be more convenient to treat analytically, or for that matter, numerically.

Not all of the nodes contributes to the interactions in the network with outgoing edges even if they have RSs of lengths in the appropriate range, $\Lambda_{\rm RS}$. The probability of finding a node with nonzero out-degree among nodes with RSs of length $l$ is $1 - P_l^{\rm out}(0)$. Now we may calculate the probability of finding a node with an RS of length $l$ and having nonzero out-degree among the nodes with RSs of lengths in the range $\Lambda_{\rm RS}$, as $p_{\rm RS}(l)[1 - P_l^{\rm out}(0)]$. Thus, the effective length distribution of RSs, $p_{\rm RS}^{\rm eff}(l)$ is given by

$$p_{\rm RS}^{\rm eff}(l) = \frac{p_{\rm RS}(l)[1 - P_l^{\rm out}(0)]}{1 - \sum_l p_{\rm RS}(l) P_l^{\rm out}(0)} \quad , \quad (4.2.19)$$

the probability of finding a node with an RS of length $l$ among the nodes having nonzero out-degree (see Fig. 4.4). However, the renormalization of this distribution always cancels in the properly normalized quantities, and therefore is not of great use. This is also true for the renormalized length distribution of the PRs.

In Fig. 4.2 we compare our analytical result for the out-degree distribution (see Eq. 4.2.17) with that of the simulations. We find that the out-degree distribution has a continues regime in the small degree region ($d \leq 10$) followed by well separated peaks corresponding to relatively small RS-lengths (see also [6]), with a broad support. The agreement between the simulation and analytical results are extremely good except the fine differences in the amplitudes in relatively large

out-degree region. The observed discrepancies might be caused by the Poisson approximation we have made to the full binomial distributions.

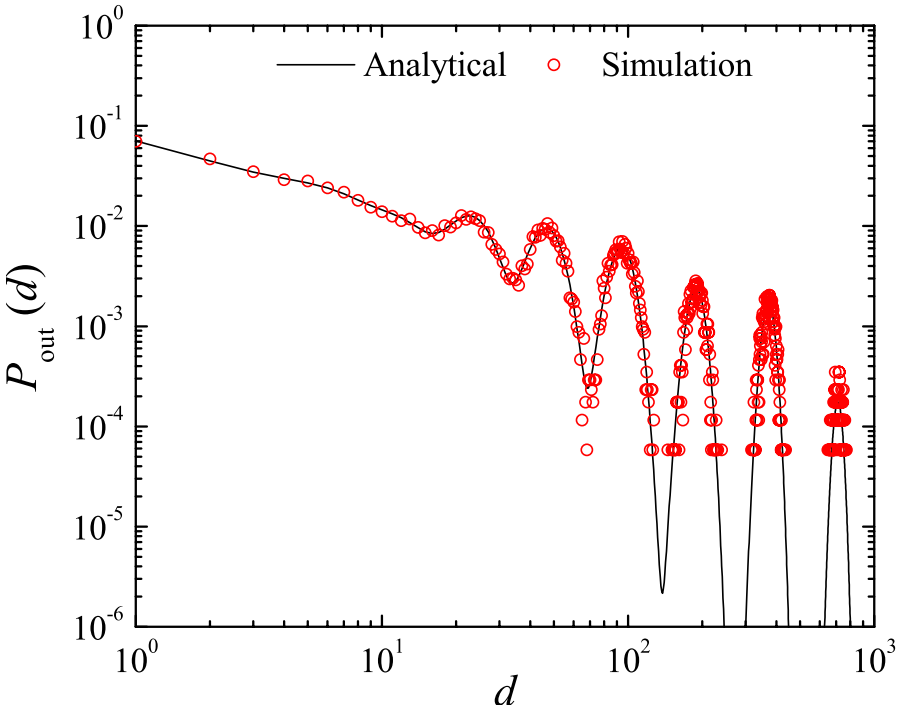


Figure 4.2: The comparison of out-degree distributions obtained by numeric evaluations of the analytical expressions and simulation results (red circles) which have been obtained by averaging over 100 realizations of the hidden-variable model. The discrete part of the spectrum arises from individual Poisson peaks contributed by nodes with small RS-lengths $l$. As one progresses to smaller degrees (longer $l$), the variances shrink less fast than the spacings between the peaks, and there is a crossover to a continuous regime [5, 6, 9, 19].

By following a similar approach to that used above we may easily calculate the in-degree distribution $P_{\rm in}(d)$, the probability of finding a node with $d$ in-coming edges if we randomly choose a node. The total number of edges $\widetilde{e}_k^{(j)}$, terminating at a randomly selected node $j$ with a PR of length $k$, is the sum of the random variables,

$$\widetilde{e}_k^{(j)} = \sum_{l \le k} \widetilde{e}_{lk}^{(j)} \quad , \tag{4.2.20}$$

and binomially distributed with the mean $d_{{\rm i},\,k}$,

$$d_{{\rm i},\,k} = \sum_{l \le k} \langle \widetilde{e}_{lk}^{(j)} \rangle = \sum_{l=l_{\min}}^{\min(k,\,l_{\max})} N \eta_{\rm RS} p_{\rm RS}(l) p(l,k) \quad , \tag{4.2.21}$$

and variance $\sigma^2_{\mathrm{i},\,k}$,

$$\sigma^2_{\mathrm{i},\,k} = \sum_{l \le k} \sigma^2_{\widetilde{e}^{(j)}_{lk}} = \sum_{l=l_{\min}}^{\min(k,\,l_{\max})} N\eta_{\mathrm{RS}} p_{\mathrm{RS}}(l) p(l,k) \left[1 - \eta_{\mathrm{RS}} p_{\mathrm{RS}}(l) p(l,k)\right] \quad . \qquad (4.2.22)$$

In Fig. 4.3a, the relationship between the average in-degree and PR-length is displayed, where we have used the parameters as used during the simulations of the hidden-variable model (see Fig. 4.4 for the length distributions used). We find that the average in-degree increases almost linearly with PR-length. We also exhibit in Fig. 4.3b for comparison, the differences between the means and variances of individual in-degree distributions and see that the variances are very close to the mean values for the parameters used.

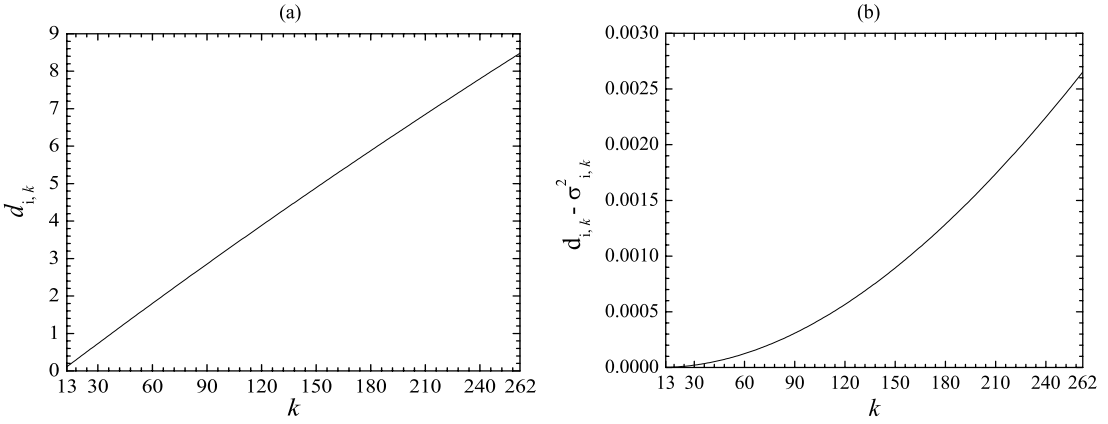


Figure 4.3: The means (a) and variances (b) of in-degree distributions. The length distribution of RSs, $p_{\mathrm{RS}}(l)$, is confined to the interval $\Lambda_{\mathrm{RS}} = [9, 38]$, and comes from the bitwise information content of binding motifs of 102 transcription factors of yeast (see Fig. 4.4a for the shape of the length distribution). The fraction $\eta_{\mathrm{RS}}$ of nodes having RS strings is $\eta_{\mathrm{RS}} = 4.8\%$. The average in-degree $d_{\mathrm{i},\,k}$ of nodes with PRs of length $k$ increases almost linearly with $k$. The means and the variances are almost equal to each other for the parameters used. (See Section 3 for the details of the parameter choices.)

In the limit of large number of nodes and small probabilities, as which is the case here, the probability $P^{\mathrm{in}}_k(d)$ of finding a node with in-degree $d$ among all the nodes with PRs of length $k$ may be approximated by a Poisson distribution with the mean $d_{\mathrm{i},\,k}$ (see Eq. 4.2.21),

$$P^{\mathrm{in}}_k(d) = \exp[-d_{\mathrm{i},\,k}] \, \frac{(d_{\mathrm{i},\,k})^d}{d!} \quad . \qquad (4.2.23)$$

Remembering that the nodes with PRs of length less than the lengths of the RSs will have zero in-degree, we can easily write the expression for the in-degree

distribution $P_{\rm in}(d)$,

$$P_{\rm in}(d) = \sum_k \widetilde{p}_{\rm PR}(k) P_{\rm in}(d|k) = \eta_{\rm PR} \sum_{k\in\Lambda_{\rm PR}} p_{\rm PR}(k) P_k^{\rm in}(d) + (1-\eta_{\rm PR})\delta(d) \quad . \quad (4.2.24)$$

The nodes with PRs of length $k$ bigger than the lengths of RSs may have zero in-degree with probability $P_k^{\rm in}(0)$. The probability of finding a node with a PR of length $k$ and also having nonzero in-degree among the nodes with PRs of lengths in the range $\Lambda_{\rm PR}$ is given by $p_{\rm PR}(k)[1-P_k^{\rm in}(0)]$. Then, it follows that the effective length distribution of PRs, $p_{\rm PR}^{\rm eff}(k)$ can be written as

$$p_{\rm PR}^{\rm eff}(k) = \frac{p_{\rm PR}(k)[1-P_k^{\rm in}(0)]}{1-\sum_k p_{\rm PR}(k) P_k^{\rm in}(0)} \quad , \quad (4.2.25)$$

which is the probability of finding a node with a PR of length $k$ among the nodes having nonzero in-degree. We show in Fig. 4.4, the effective length distributions of PRs and RSs for the parameters used during the simulations of the hidden-variable model. One may easily observe that the relatively small RS-lengths are amplified whereas the relatively large RS-lengths are suppressed (see Fig. 4.4a). The situation is reverse for the PR-lengths, the large PR-lengths are amplified and the small PR-lengths are suppressed (see Fig. 4.4b). Again note that the renormalization of these distributions always cancel in the properly normalized quantities, and therefore are not of great use.

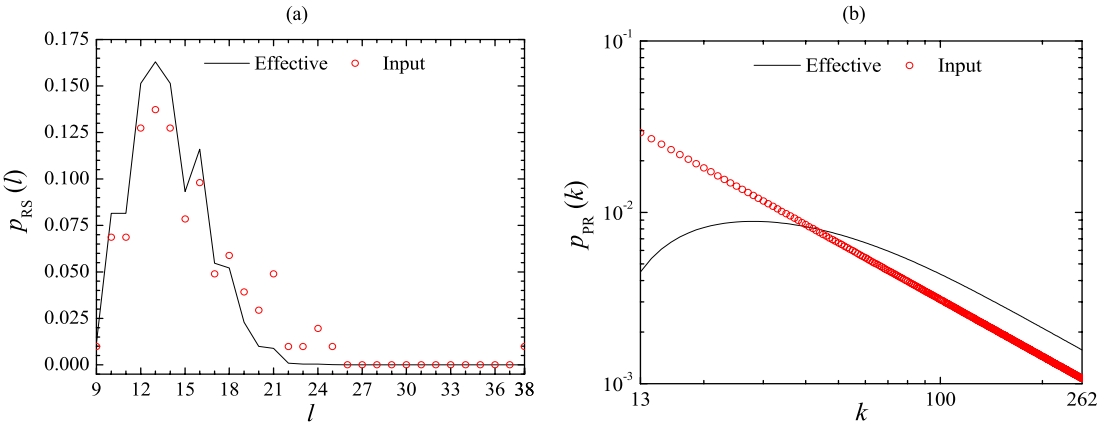


Figure 4.4: The comparison of the input length distributions (red circles) with the effective length distributions of RSs (a) and PRs (b) obtained by numeric evaluations of the expressions in Eqs. (4.2.19, 4.2.25).

In Fig. 4.5 we compare our analytical result for the in-degree distribution (see Eq. 4.2.24) with that of the simulations. The in-degree distribution has a very

narrow support by comparison with the out-degree distribution (Fig. 4.2). The agreement between the simulation and analytical results are extremely good except the fine differences in the amplitudes in relatively large in-degree regions. The observed discrepancies might be again caused by the Poisson approximation we have made to the full binomial distributions.

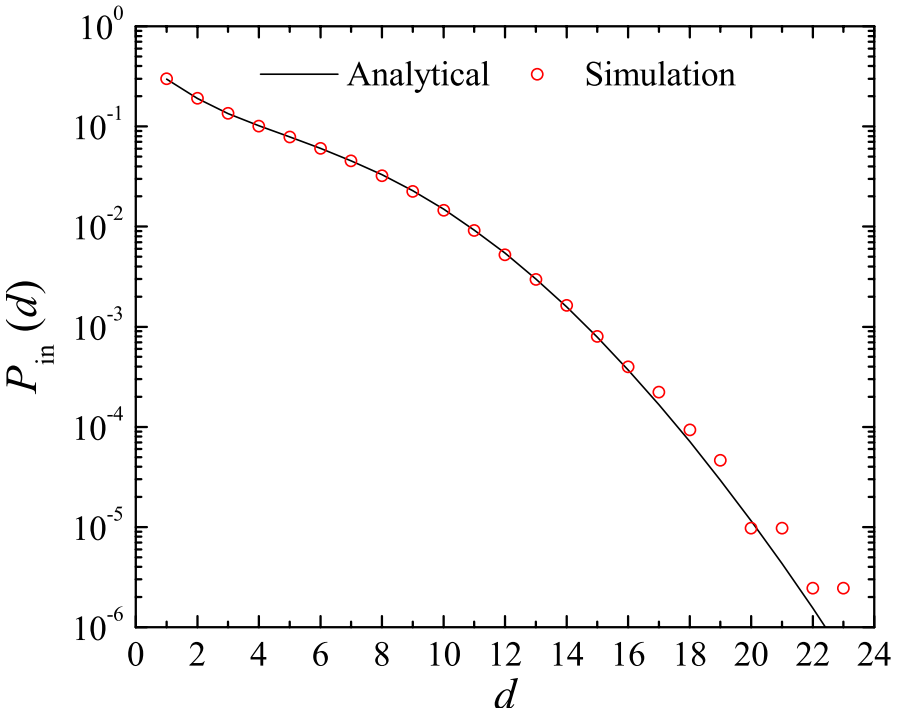


Figure 4.5: The comparison of in-degree distributions obtained by numeric evaluations of the analytical expressions and simulation results (red circles online). The in-degree is approximately exponential as can be seen from this log-linear plot.

Now we are ready to calculate the total degree distribution $P(d)$, the probability of finding a node interacting with $d$ nodes. Here we will assume that the total degree of a node is the sum of its out- and in-degree (ignoring the bidirectional edges between pairs of nodes), $d = d_{\rm o} + d_{\rm i}$,

$$P(d) = \sum_{d_{\rm o},d_{\rm i}} P(d_{\rm o}, d_{\rm i})\delta(d - d_{\rm o} - d_{\rm i}) = \sum_{d_{\rm o}=0}^{d} P_{\rm out}(d_{\rm o})P_{\rm in}(d - d_{\rm o}) \quad , \tag{4.2.26}$$

where we have used the fact that the out- and in-degree of a node are independent from each other in this double-string model, so the joint probability is separable. By substituting the expressions in Eqs. (4.2.17, 4.2.24) into the above equation

we get

$$
\begin{aligned}
P(d) &= \Bigg\{ \eta_{\rm RS}\eta_{\rm PR} \sum_{l,\,k} p_{\rm RS}(l) p_{\rm PR}(k) P_{l,\,k}(d) \\
&+ \eta_{\rm RS}(1-\eta_{\rm PR}) \sum_{l} p_{\rm RS}(l) P_l^{\rm out}(d) + (1-\eta_{\rm RS})\eta_{\rm PR} \sum_{k} p_{\rm PR}(k) P_k^{\rm in}(d) \\
&+ (1-\eta_{\rm RS})(1-\eta_{\rm PR})\delta(d) \Bigg\} \ , \qquad (4.2.27)
\end{aligned}
$$

where $P_{l,\,k}(d)$ is defined as the degree distribution of nodes with RSs of length $l$ and PRs of length $k$,

$$
P_{l,\,k}(d) = \sum_{d_{\rm o}=0}^{d} P_l^{\rm out}(d_{\rm o}) P_k^{\rm in}(d - d_{\rm o}) \ . \qquad (4.2.28)
$$

As you have already recognized in Eq. 4.2.27, different terms come from different types of nodes with respect to the random variables attached to them, for example, the second term is the contribution of the nodes having RSs of lengths in the range $\Lambda_{\rm RS}$ but with PRs of lengths less than the values of RSs ($k < l_{\rm min}$). In Eq. 4.2.28, it is easy to perform the summation over $d_{\rm o}$ in the case of Poisson out- and in-degree distributions (see Eqs. 4.2.18, 4.2.23) and this gives rise to another Poisson distribution with the mean $d_{{\rm o},\,l} + d_{{\rm i},\,k}$ for each pair of variables $l$ and $k$, $P_{l,\,k}(d) = \exp[-(d_{{\rm o},\,l} + d_{{\rm i},\,k})]\ (d_{{\rm o},\,l} + d_{{\rm i},\,k})^d / d!$.

We exhibit in Fig. 4.6, the total degree distributions obtained analytically and via simulations. For the theoretical curve we have evaluated Eq. 4.2.27. The total degree distribution in the small degree region is essentially determined by the in-degree distribution of the nodes having only PRs, who are in a majority. The large degree region is dominated by nodes with RSs.

## 4.3 Degree-Degree Correlation of Nearest Neighbors

This section has been devoted to the calculation of degree-degree correlation [20, 21] of nearest neighbors (connected pairs of nodes). We will be calculating the probability of finding a node with degree $d'$ among the nearest neighbors of nodes of degree $d$.

For the purposes of this section, it is useful to note that the ensemble of networks in question consist of four types of nodes:

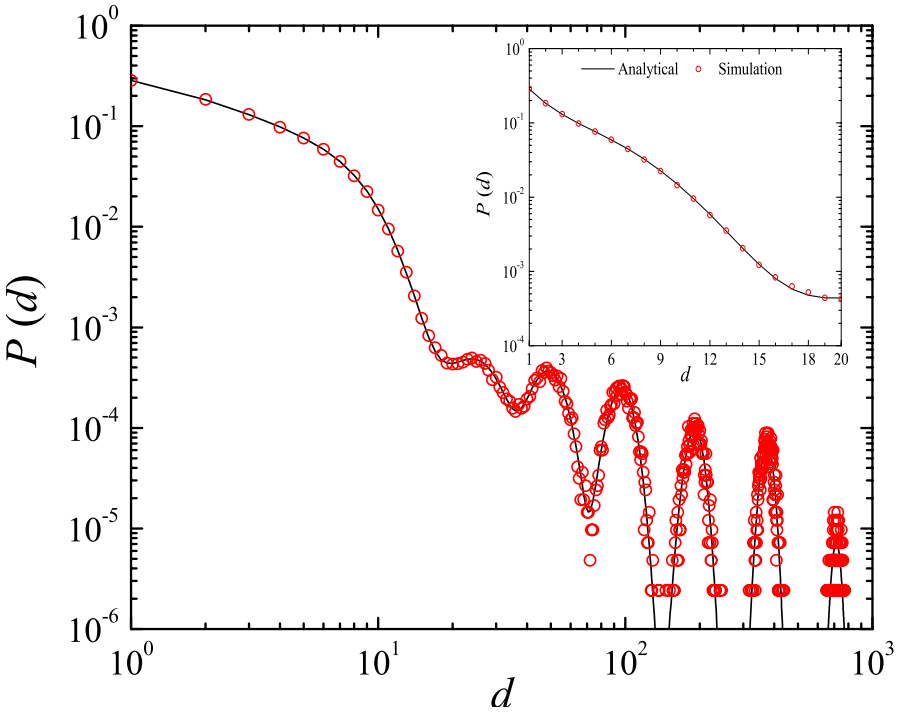


Figure 4.6: The total degree distribution with an inset showing a log-linear plot for $d \leq 20$. Theoretical curve has been obtained by numerical evaluations of the expressions in Eq. 4.2.27.

- **Type I**$(l, k)$**:** Consists of the nodes each having an RS of length $l \in \Lambda_{\text{RS}}$ and a PR of length $k \in \Lambda_{\text{PR}}$. The probability $\eta_{\text{I}}$ of finding a node of this type is $\eta_{\text{I}} = \eta_{\text{RS}}\eta_{\text{PR}}$. The nodes of Type I may have out-going and in-coming edges. We will refer to a node of Type I with an RS of length $l$ and a PR of length $k$ by I$(l, k)$.

- **Type II**$(k)$**:** Contains the nodes having only PRs, meaning that $k \in \Lambda_{\text{PR}}$ and their RSs have the value $l > k_{\max}$. The probability $\eta_{\text{II}}$ of finding a node of this type is $\eta_{\text{II}} = (1 - \eta_{\text{RS}})\eta_{\text{PR}}$. Type II nodes may only have in-coming links. We will refer to a node of Type II with a PR of length $k$ by II$(k)$.

- **Type III**$(l)$**:** Collection of those nodes having only RSs, meaning $l \in \Lambda_{\text{RS}}$, with PRs that are too short, i.e., $k < l_{\min}$. The probability $\eta_{\text{III}}$ of finding a node of this type is $\eta_{\text{III}} = \eta_{\text{RS}}(1 - \eta_{\text{PR}})$. Type III nodes may only have out-going links. We will refer to a node of Type III with an RS of length $l$ by III$(l)$.

- **Type IV:** The nodes having neither an RS nor a PR in the appropriate ranges $\Lambda_{\text{RS}}$ and $\Lambda_{\text{PR}}$, $l > k_{\max}$ and $k < l_{\min}$. The probability $\eta_{\text{IV}}$ of finding

a node of this type is $\eta_{\rm IV} = (1-\eta_{\rm RS})(1-\eta_{\rm PR})$. The nodes of this type have neither an out-going nor an in-coming edge, so they do not contribute to the interactions in the network.

In Fig. 4.7 we display all the possible configurations of directed pairwise connection. Let us note that we will be using the averages of the quantities rather than their distributions. The average number $E_{lk}$, of edges originating from nodes with RSs of length $l$ and terminating at nodes with PRs of length $k$, is $E_{lk} = N_{\rm RS}(l)N_{\rm PR}(k)p(l,k) = N^2\eta_{\rm RS}\eta_{\rm PR}p_{\rm RS}(l)p_{\rm PR}(k)p(l,k)$, where $N_{\rm RS}(l)$ and $N_{\rm PR}(k)$ are defined as the average number of nodes with RSs of length $l$ and PRs of length $k$, respectively, given in Eq. 4.1.5. We may group the edges into the sets with respect to the types of nodes at their initial and terminal ends,

$$E_{lk}^{\rm T-T'} = \frac{\eta_T\eta_{\rm T'}}{\eta_{RS}\eta_{\rm PR}}\, E_{lk} \ , \tag{4.3.29}$$

where $E_{lk}^{\rm T-T'}$ corresponds to the average number of edges originating from nodes of Type T with RSs of length $l$ and terminating at nodes of Type T$'$ with PRs of length $k$. Defining $E$ as the average of the total number of edges, we may write the probability $\mathcal{P}(l,k)$, that a randomly chosen edge is originating from an RS of length $l$ and terminating at a PR of length $k$, as

$$\mathcal{P}(l,k) = \frac{E_{lk}}{E} = \frac{N_{\rm RS}(l)N_{\rm PR}(k)p(l,k)}{\sum_{l',\,k'\geq l'} N_{\rm RS}(l')N_{\rm PR}(k')p(l',k')} \ . \tag{4.3.30}$$

Then the probability $\mathcal{P}_{\rm T-T'}(l,k)$, of finding an edge from a node of Type T with an RS of length $l$ to a node of Type T$'$ with a PR of length $k$ if we randomly pick an edge, is given by

$$\mathcal{P}_{\rm T-T'}(l,k) = \frac{E_{lk}^{\rm T-T'}}{E} = \frac{\eta_{\rm T}\eta_{\rm T'}}{\eta_{\rm RS}\eta_{PR}}\,\mathcal{P}(l,k) \ . \tag{4.3.31}$$

Now we are ready to write down the expression for $p(d,d')$, the probability of finding a pair of nodes of (total) degrees $d$ and $d'$ at the two ends of a randomly selected edge. This probability can be written in terms of contributions coming from different types of pairs of nodes,

$$p(d,d') = \sum_{\rm T,\,T'} p_{\rm T-T'}(d,d') \ , \tag{4.3.32}$$

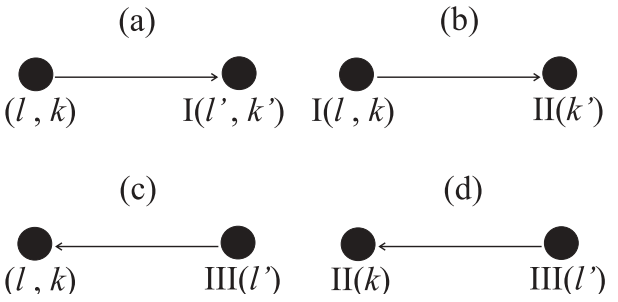


Figure 4.7: Possible configurations of directed pairwise connection. Nodes of Type I with hidden variables $(l, k)$ may have connections with nodes of Type I, II, and III. Nodes of Type II with hidden variable $(k)$ may only have connections with nodes of Type I$(l', k')$ and III$(l')$ with $l' \leq k$. Nodes of Type III with hidden variable $(l)$ may only have connections with nodes of Type I$(l', k')$ and II$(k')$ with $k' \geq l$.

where $p_{\mathrm{T}-\mathrm{T}'}(d, d')$ is the probability of finding a pair of nodes of Type T and Type T$'$ either of them with degree $d$ or $d'$ at the ends of a randomly selected edge. To give a better understanding we will calculate the contributions to this probability term by term, coming from different types of pairs of nodes.

Let us start with the easiest term, the contribution coming from the nodes of Type II and Type III, $p_{\mathrm{III}-\mathrm{II}}(d, d')$. This probability is given by

$$\begin{aligned} p_{\mathrm{III}-\mathrm{II}}(d, d') &= \frac{1}{2} \sum_{l,\, k \geq l} \mathcal{P}_{\mathrm{III}-\mathrm{II}}(l, k) \left[ \frac{d P_l^{\mathrm{out}}(d)}{d_{\mathrm{o},\, l}} \, \frac{d' P_k^{\mathrm{in}}(d')}{d_{\mathrm{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_l^{\mathrm{out}}(d')}{d_{\mathrm{o},\, l}} \, \frac{d P_k^{\mathrm{in}}(d)}{d_{\mathrm{i},\, k}} \right] , \\ &= (1 - \eta_{\mathrm{RS}})(1 - \eta_{\mathrm{PR}}) \frac{1}{2E} \sum_{l,\, k \geq l} E_{lk} \left[ \frac{d P_l^{\mathrm{out}}(d)}{d_{\mathrm{o},\, l}} \, \frac{d' P_k^{\mathrm{in}}(d')}{d_{\mathrm{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_l^{\mathrm{out}}(d')}{d_{\mathrm{o},\, l}} \, \frac{d P_k^{\mathrm{in}}(d)}{d_{\mathrm{i},\, k}} \right] , \end{aligned} \tag{4.3.33}$$

where $d'' P_l^{\mathrm{out}}(d'')/d_{\mathrm{o},\, l}$ is the probability of finding a node with out-degree $d''$ at the initial end of a randomly chosen edge among all the edges originating from nodes with RSs of length $l$, and $d'' P_k^{\mathrm{in}}(d'')/d_{\mathrm{i},\, k}$ is the probability of finding a node with in-degree $d''$ at the terminal end of a randomly chosen edge among all the edges terminating at nodes with PRs of length $k$. Note here that the normalization $d_{\mathrm{o},\, l} = \sum_{d''} d'' P_l^{\mathrm{out}}(d'')$ as given in Eq. 4.2.15, is the average out-degree of the nodes with RSs of length $l$, and the normalization $d_{\mathrm{i},\, k} = \sum_{d''} d'' P_k^{\mathrm{in}}(d'')$ as given in Eq. 4.2.21, is the average in-degree of the nodes with PRs of length $k$. Again note that the nodes of Type III have zero in-degree, so the total degree of such a node is equal to its out-degree, and the nodes of Type II have zero out-degree,

so the total degree of such a node is equal to its in-degree. The summation in the [...] with a factor of 1/2 out front comes because it is not specified which of the nodes has degree $d$ or $d'$.

The second contribution comes from pairs of nodes of Type I and Type III, $p_{\text{III}-\text{I}}(d, d')$. The probability of finding a node of Type I and a node of Type III, either of them with degree $d$ or $d'$, at the two ends of an edge which is chosen at random is given by

$$\begin{aligned} p_{\text{III}-\text{I}}(d, d') &= \frac{1}{2} \sum_{l,\, k \geq l} \mathcal{P}_{\text{III}-\text{I}}(l, k) \sum_{l'} p_{\text{RS}}(l') \left[ \frac{dP_l^{\text{out}}(d)}{d_{\text{o},\, l}} \, \frac{d' P_{l',\, k}(d')}{d_{\text{o},\, l'} + d_{\text{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_l^{\text{out}}(d')}{d_{\text{o},\, l}} \, \frac{dP_{l',\, k}(d)}{d_{\text{o},\, l'} + d_{\text{i},\, k}} \right] , \\ &= \eta_{\text{RS}}(1 - \eta_{\text{PR}}) \frac{1}{2E} \sum_{l,\, k \geq l} E_{lk} \sum_{l'} p_{\text{RS}}(l') \left[ \frac{dP_l^{\text{out}}(d)}{d_{\text{o},\, l}} \, \frac{d' P_{l',\, k}(d')}{d_{\text{o},\, l'} + d_{\text{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_l^{\text{out}}(d')}{d_{\text{o},\, l}} \, \frac{dP_{l',\, k}(d)}{d_{\text{o},\, l'} + d_{\text{i},\, k}} \right] , \end{aligned} \tag{4.3.34}$$

where $\sum_{l'} p_{\text{RS}}(l') d'' P_{l',\, k}(d'') / (d_{\text{o},\, l'} + d_{\text{i},\, k})$ is the probability of finding a node with total degree $d''$, if we follow a randomly chosen edge among all the edges terminating at the nodes of Type I with PRs of length $k$. Note here that the normalization $d_{\text{o},\, l'} + d_{\text{i},\, k} = \sum_{d''} d'' P_{l',\, k}(d'')$, is the average total degree of nodes with RSs of length $l'$ and PRs of length $k$ (see Eqs. (4.2.15, 4.2.21, 4.2.27)).

The third part comes from the pairs of nodes of Type I and Type II, $p_{\text{I}-\text{II}}(d, d')$. This probability is given by

$$\begin{aligned} p_{\text{I}-\text{II}}(d, d') &= \frac{1}{2} \sum_{l,\, k \geq l} \mathcal{P}_{\text{I}-\text{II}}(l, k) \sum_{k'} p_{\text{PR}}(k') \left[ \frac{dP_{l,\, k'}(d)}{d_{\text{o},\, l} + d_{\text{i},\, k'}} \, \frac{d' P_k^{\text{in}}(d')}{d_{\text{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_{l,\, k'}(d')}{d_{\text{o},\, l} + d_{\text{i},\, k'}} \, \frac{dP_k^{\text{in}}(d)}{d_{\text{i},\, k}} \right] , \\ &= (1 - \eta_{\text{RS}}) \eta_{\text{PR}} \frac{1}{2E} \sum_{l,\, k \geq l} E_{lk} \sum_{k'} p_{\text{PR}}(k') \left[ \frac{dP_{l,\, k'}(d)}{d_{\text{o},\, l} + d_{\text{i},\, k'}} \, \frac{d' P_k^{\text{in}}(d')}{d_{\text{i},\, k}} \right. \\ &\qquad \left. + \frac{d' P_{l,\, k'}(d')}{d_{\text{o},\, l} + d_{\text{i},\, k'}} \, \frac{dP_k^{\text{in}}(d)}{d_{\text{i},\, k}} \right] , \end{aligned} \tag{4.3.35}$$

where $\sum_{k'} p_{\text{PR}}(k') d'' P_{l,\, k'}(d'') / (d_{\text{o},\, l} + d_{\text{i},\, k'})$ is the probability of finding a node with total degree $d''$ if we follow a randomly chosen edge among all the edges originating from the nodes of Type I with RSs of length $l$.

Finally the last term we have to consider here is the contribution of the pairs of nodes, both of which are of Type I, $p_{\mathrm{I-I}}(d,d')$. The probability of finding a pair of nodes which are both of Type I, either of them with degree $d$ or $d'$, at the two ends of a randomly chosen edge is given by

$$\begin{aligned} p_{\mathrm{I-I}}(d,d') &= \frac{1}{2}\sum_{l,\,k\geq l}\mathcal{P}_{\mathrm{I-I}}(l,k)\sum_{l',\,k'}p_{\mathrm{RS}}(l')p_{\mathrm{PR}}(k')\left[\frac{dP_{l,\,k'}(d)}{d_{\mathrm{o},\,l}+d_{\mathrm{i},\,k'}}\,\frac{d'P_{l',\,k}(d')}{d_{\mathrm{o},\,l'}+d_{\mathrm{i},\,k}}\right. \\ &\qquad\left.+\frac{d'P_{l,\,k'}(d')}{d_{\mathrm{o},\,l}+d_{\mathrm{i},\,k'}}\,\frac{dP_{l',\,k}(d)}{d_{\mathrm{o},\,l'}+d_{\mathrm{i},\,k}}\right]\ , \\ &= \eta_{\mathrm{RS}}\eta_{\mathrm{PR}}\frac{1}{2E}\sum_{l,\,k\geq l}E_{lk}\sum_{l',\,k'}p_{\mathrm{RS}}(l')p_{\mathrm{PR}}(k')\left[\frac{dP_{l,\,k'}(d)}{d_{\mathrm{o},\,l}+d_{\mathrm{i},\,k'}}\,\frac{d'P_{l',\,k}(d')}{d_{\mathrm{o},\,l'}+d_{\mathrm{i},\,k}}\right. \\ &\qquad\left.+\frac{d'P_{l,\,k'}(d')}{d_{\mathrm{o},\,l}+d_{\mathrm{i},\,k'}}\,\frac{dP_{l',\,k}(d)}{d_{\mathrm{o},\,l'}+d_{\mathrm{i},\,k}}\right]\ . \end{aligned} \tag{4.3.36}$$

Note that the average number of edges is $E=\sum_{l,\,k\geq l}E_{lk}=\sum_{k,\,l\leq k}E_{lk}$, equivalently,

$$E=\eta_{\mathrm{RS}}N\sum_{l}p_{\mathrm{RS}}(l)d_{\mathrm{o},\,l}=\eta_{\mathrm{PR}}N\sum_{k}p_{\mathrm{PR}}(k)d_{\mathrm{i},\,k}\ . \tag{4.3.37}$$

This is simply the statement that the total number of out-going edges is equal to the total number of in-coming edges. If we sum over $d$ and $d'$ in Eqs. (4.3.33, 4.3.34, 4.3.35, 4.3.36) we get

$$\sum_{d,\,d'}p_{\mathrm{T-T'}}(d,d')=\frac{\eta_{\mathrm{T}}\eta_{\mathrm{T'}}}{\eta_{\mathrm{RS}}\eta_{\mathrm{PR}}}\ , \tag{4.3.38}$$

thus, the probability $p(d,d')$ is normalized, $\sum_{d,d'}p(d,d')=1$.

The degree-degree correlation of nearest neighbors is measured by the probability of finding a node with degree $d'$ among the nearest neighbors of nodes with degree $d$, $p(d'|d)$. This conditional probability is given in terms of $p(d,d')$ and the probability of finding a node of degree $d$ at the end of a randomly selected edge, $p(d)=dNP(d)/2E$,

$$p(d'|d)=\frac{p(d,d')}{p(d)}\ . \tag{4.3.39}$$

Note here that the probability $p(d)$, of ending up at a node of degree $d$ if we follow a randomly picked edge, is proportional to $dP(d)$, and not just to the

degree distribution itself. Thus, $p(d'|d)$ is written as

$$
\begin{aligned}
p(d'|d) = \frac{\eta_{\rm RS}\eta_{\rm PR}N}{P(d)} \Bigg\{ & (1-\eta_{\rm RS})(1-\eta_{\rm PR}) \sum_{l,\, k\geq l} p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ \frac{P_l^{\rm out}(d)}{d_{{\rm o},\, l}} \, \frac{d' P_k^{\rm in}(d')}{d_{{\rm i},\, k}} \frac{d' P_l^{\rm out}(d')}{d_{{\rm o},\, l}} \, \frac{P_k^{\rm in}(d)}{d_{{\rm i},\, k}} \right] \\
& + \eta_{\rm RS}(1-\eta_{\rm PR}) \sum_{l',\, l,\, k\geq l} p_{\rm RS}(l')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ \frac{P_l^{\rm out}(d)}{d_{{\rm o},\, l}} \, \frac{d' P_{l',\, k}(d')}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} + \frac{d' P_l^{\rm out}(d')}{d_{{\rm o},\, l}} \, \frac{P_{l',\, k}(d)}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} \right] \\
& + (1-\eta_{\rm RS})\eta_{\rm PR} \sum_{k',\, l,\, k\geq l} p_{\rm PR}(k')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ \frac{P_{l,\, k'}(d)}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} \, \frac{d' P_k^{\rm in}(d')}{d_{{\rm i},\, k}} + \frac{d' P_{l,\, k'}(d')}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} \, \frac{P_k^{\rm in}(d)}{d_{{\rm i},\, k}} \right] \\
& + \eta_{\rm RS}\eta_{\rm PR} \sum_{l',\, k',\, l,\, k\geq l} p_{\rm RS}(l')p_{\rm PR}(k')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ \frac{P_{l,\, k'}(d)}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} \, \frac{d' P_{l',\, k}(d')}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} + \frac{d' P_{l,\, k'}(d')}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} \, \frac{P_{l',\, k}(d)}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} \right] \Bigg\} \quad . \qquad (4.3.40)
\end{aligned}
$$

Again note that $p(d'|d)$ is normalized, $\sum_{d'} p(d'|d) = 1$. Since one needs a three dimensional graph to display the conditional probability, instead, mostly the average degree of nodes connected to nodes with degree $d$, $d_{\rm nn}(d) = \sum_{d'} p(d'|d)d'$ is displayed. The latter quantity has a much simpler expression. We get

$$
\begin{aligned}
d_{\rm nn}(d) = \frac{\eta_{\rm RS}\eta_{\rm PR}N}{P(d)} \Bigg\{ & (1-\eta_{\rm RS})(1-\eta_{\rm PR}) \sum_{l,\, k\geq l} p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ (1+d_{{\rm i},\, k}) \, \frac{P_l^{\rm out}(d)}{d_{{\rm o},\, l}} + (1+d_{{\rm o},\, l}) \, \frac{P_k^{\rm in}(d)}{d_{{\rm i},\, k}} \right] \\
& + \eta_{\rm RS}(1-\eta_{\rm PR}) \sum_{l',\, l,\, k\geq l} p_{\rm RS}(l')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ (1+d_{{\rm o},\, l'} + d_{{\rm i},\, k}) \, \frac{P_l^{\rm out}(d)}{d_{{\rm o},\, l}} + (1+d_{{\rm o},\, l}) \, \frac{P_{l',\, k}(d)}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} \right] \\
& + (1-\eta_{\rm RS})\eta_{\rm PR} \sum_{k',\, l,\, k\geq l} p_{\rm PR}(k')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ (1+d_{{\rm i},\, k}) \, \frac{P_{l,\, k'}(d)}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} + (1+d_{{\rm o},\, l} + d_{{\rm i},\, k'}) \, \frac{P_k^{\rm in}(d)}{d_{{\rm i},\, k}} \right] \\
& + \eta_{\rm RS}\eta_{\rm PR} \sum_{l',\, k',\, l,\, k\geq l} p_{\rm RS}(l')p_{\rm PR}(k')p_{\rm RS}(l)p_{\rm PR}(k)p(l,k) \\
& \cdot \left[ (1+d_{{\rm o},\, l'} + d_{{\rm i},\, k}) \, \frac{P_{l,\, k'}(d)}{d_{{\rm o},\, l} + d_{{\rm i},\, k'}} + (1+d_{{\rm o},\, l} + d_{{\rm i},\, k'}) \, \frac{P_{l',\, k}(d)}{d_{{\rm o},\, l'} + d_{{\rm i},\, k}} \right] \Bigg\} \quad , \qquad (4.3.41)
\end{aligned}
$$

where we have used the equality of the means and the variances of Poisson dis-

tributed quantities, for example, if $d''$ follows a Poisson distribution with the mean $\langle d'' \rangle$, then $\langle d''^2 \rangle = \langle d'' \rangle + \langle d'' \rangle^2$. We can easily carry out the summations over the independent variables in the above equation,

$$\sum_{l'} p_{\mathrm{RS}}(l')(1 + d_{\mathrm{o},\,l'} + d_{\mathrm{i},\,k}) = 1 + \langle d_{\mathrm{o}} \rangle + d_{\mathrm{i},\,k} \quad , \tag{4.3.42}$$

$$\sum_{k'} p_{\mathrm{PR}}(k')(1 + d_{\mathrm{o},\,l} + d_{\mathrm{i},\,k'}) = 1 + \langle d_{\mathrm{i}} \rangle + d_{\mathrm{o},\,l} \quad , \tag{4.3.43}$$

where we have made the definitions $\langle d_{\mathrm{o}} \rangle = \sum_{l'} p_{\mathrm{RS}}(l') d_{\mathrm{o},\,l'}$, the average out-degree of nodes with RS-lengths $\in \Lambda_{\mathrm{RS}}$, and $\langle d_{\mathrm{i}} \rangle = \sum_{k'} p_{\mathrm{PR}}(k') d_{\mathrm{i},\,k'}$, the average in-degree of nodes with PR-lengths $\in \Lambda_{\mathrm{PR}}$. If we substitute these equalities into $d_{\mathrm{nn}}(d)$ and make the necessary simplifications we get

$$\begin{aligned}
d_{\mathrm{nn}}(d) &= \frac{\eta_{\mathrm{RS}} \eta_{\mathrm{PR}} N}{P(d)} \\
\times \quad & \Bigg\{ (1 - \eta_{\mathrm{PR}}) \sum_{l,\, k \geq l} p_{\mathrm{RS}}(l) p_{\mathrm{PR}}(k) p(l,k) (1 + \eta_{\mathrm{RS}} \langle d_{\mathrm{o}} \rangle + d_{\mathrm{i},\,k}) \, \frac{P_l^{\mathrm{out}}(d)}{d_{\mathrm{o},\,l}} \\
+ \quad & (1 - \eta_{\mathrm{RS}}) \sum_{l,\, k \geq l} p_{\mathrm{RS}}(l) p_{\mathrm{PR}}(k) p(l,k) (1 + \eta_{\mathrm{PR}} \langle d_{\mathrm{i}} \rangle + d_{\mathrm{o},\,l}) \, \frac{P_k^{\mathrm{in}}(d)}{d_{\mathrm{i},\,k}} \\
+ \quad & \eta_{\mathrm{RS}} \sum_{l',\, l,\, k \geq l} p_{\mathrm{RS}}(l') p_{\mathrm{RS}}(l) p_{\mathrm{PR}}(k) p(l,k) (1 + \eta_{\mathrm{PR}} \langle d_{\mathrm{i}} \rangle + d_{\mathrm{o},\,l}) \, \frac{P_{l',\,k}(d)}{d_{\mathrm{o},\,l'} + d_{\mathrm{i},\,k}} \\
+ \quad & \eta_{\mathrm{PR}} \sum_{k',\, l,\, k \geq l} p_{\mathrm{PR}}(k') p_{\mathrm{RS}}(l) p_{\mathrm{PR}}(k) p(l,k) (1 + \eta_{\mathrm{RS}} \langle d_{\mathrm{o}} \rangle + d_{\mathrm{i},\,k}) \, \frac{P_{l,\,k'}(d)}{d_{\mathrm{o},\,l} + d_{\mathrm{i},\,k'}} \Bigg\} \quad .
\end{aligned} \tag{4.3.44}$$

The degree-degree correlation of nearest neighbors measures the tendency of nodes with similar degrees to be connected to each other. The disassortative behavior observed in Fig. 4.8 for the simulation results and captured accurately by our analytical calculations which have been obtained by evaluating the summations in Eq. 4.3.44 numerically, is very easy to explain. The nodes in the small degree region of the function have only PRs and connections with those nodes having RSs. The lower the degree of such a node is, the smaller the PR-length associated with this node. This indicates that if a node with an RS happens to be connected with such a node its RS must be very small, thus its out-degree is very high. The behavior in the large degree region is also quite easy to understand. The nodes dominating this region have RSs and are mostly connected to those nodes

having only PRs, and therefore, whose degree is relatively small. Oscillations are observed in the same degree range as they appear in the degree distribution. The peaks coincide with the dips in $P(d)$ which normalizes $d_{\rm nn}(d)$ (Eq. 4.3.44).

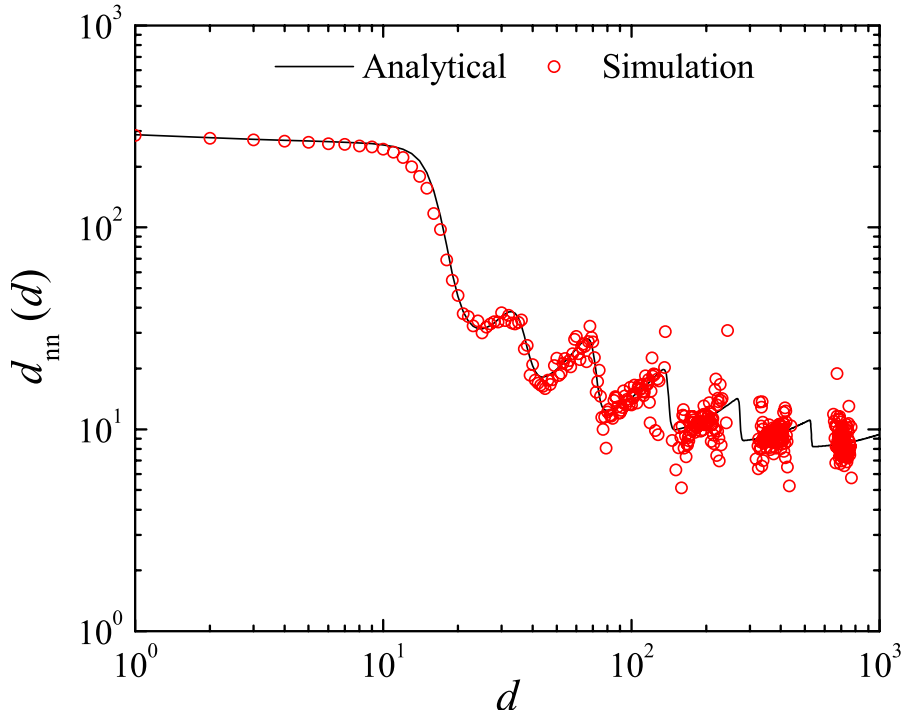


Figure 4.8: Average degree of nearest neighbors of nodes with degree $d$ for the analytical solutions obtained by evaluating the expressions in Eq. 4.3.44 numerically, and simulations (red circles).

## 4.4 Clustering Coefficient

Another important and well studied quantity giving information about the organization of a network is the three point correlations of nodes, namely the clustering coefficient $c(d)$ [22, 23, 24]. The clustering coefficient is the probability that a pair of nodes chosen at random among the nearest neighbors of a randomly chosen node with degree $d$, are connected. Pick a node with degree $d$ and consider the subgraph $\mathcal{G}$ containing all the nearest neighbors of this node which we will call the "root" of $\mathcal{G}$. The probability that any given pair of nodes in $\mathcal{G}$ are connected, depends on the type to which the root belongs, and on its hidden variables. Our strategy for calculating $c(d)$ will be to compute these probabilities, namely

$\Delta_{\rm I}(l,k)$, $\Delta_{\rm II}(k)$ and $\Delta_{\rm III}(l)$, from which we obtain,

$$c(d) = \sum_{l,\,k} P_{\rm I}(l,k|d)\ \Delta_{\rm I}(l,k) + \sum_{k} P_{\rm II}(k|d)\ \Delta_{\rm II}(k) + \sum_{l} P_{\rm III}(l|d)\ \Delta_{\rm III}(l)\ \ , \quad (4.4.45)$$

where $P_{\rm T}(.|d)$ are the probabilities of encountering nodes of Type T with hidden variables (.) and degree $d$. The probability $P_{\rm I}(l,k|d)$, that a randomly chosen node among nodes of degree $d$ is of Type I$(l,k)$, may be written as

$$P_{\rm I}(l,k|d) = \eta_{\rm RS}\eta_{\rm PR} p_{\rm RS}(l) p_{\rm PR}(k)\ \frac{P_{l,\,k}(d)}{P(d)}\ \ , \quad (4.4.46)$$

where $P(d)$ is the total degree distribution and $P_{l,\,k}(d)$ is the degree distribution of nodes of Type I$(l,k)$ (see Eq. 4.2.27). The probability $P_{\rm II}(k|d)$, of finding a node of Type II$(k)$, is given by

$$P_{\rm II}(k|d) = (1-\eta_{\rm RS})\eta_{\rm PR} p_{\rm PR}(k)\ \frac{P^{\rm in}_{k}(d)}{P(d)}\ \ , \quad (4.4.47)$$

where $P^{\rm in}_{k}(d)$ is the degree distribution of nodes of Type II$(k)$ (see Eq. 4.2.24). The probability of encountering a node of Type III$(l)$ among nodes of given degree $d$, $P_{\rm III}(l|d)$ is given in a similar way,

$$P_{\rm III}(l|d) = \eta_{\rm RS}(1-\eta_{\rm PR}) p_{\rm RS}(l)\ \frac{P^{\rm out}_{l}(d)}{P(d)}\ \ , \quad (4.4.48)$$

where $P^{\rm out}_{l}(d)$ is the degree distribution of nodes of Type III$(l)$ (see Eqs. 4.2.17).

The probabilities $\Delta_{\rm T}(.)$ can be further decomposed into probabilities of connection between different pairs of types of nodes within $\mathcal{G}$ as shown in Fig. 4.9. The connection probabilities between different nodes in $\mathcal{G}$ of given types, are in fact the probabilities of occurrence of the triangles shown in Fig. 4.9, where the apex should be identified with the root. Let us start by giving some definitions, to be used for the calculations of these probabilities. The average number of edges $E_l$, originating from nodes with RSs of length $l$, is given by

$$\begin{aligned} E_l &= \sum_{k' \geq l} E_{lk'} = N_{\rm RS}(l) \sum_{k' \geq l} N_{\rm PR}(k') p(l,k') = N_{\rm RS}(l)\ d_{{\rm o},\,l}\ \ , \\ &= \varphi_l + \eta_{\rm PR} N_{\rm RS}(l)\ d_{{\rm o},\,l}\ \ , \end{aligned} \quad (4.4.49)$$

where $\varphi_l$ is defined as the average number of nearest neighbors of nodes of Type III$(l)$,

$$\varphi_l = (1-\eta_{\rm PR}) N_{\rm RS}(l)\ d_{{\rm o},\,l} = \eta_{\rm RS}(1-\eta_{\rm PR}) N p_{\rm RS}(l)\ d_{{\rm o},\,l}\ \ . \quad (4.4.50)$$

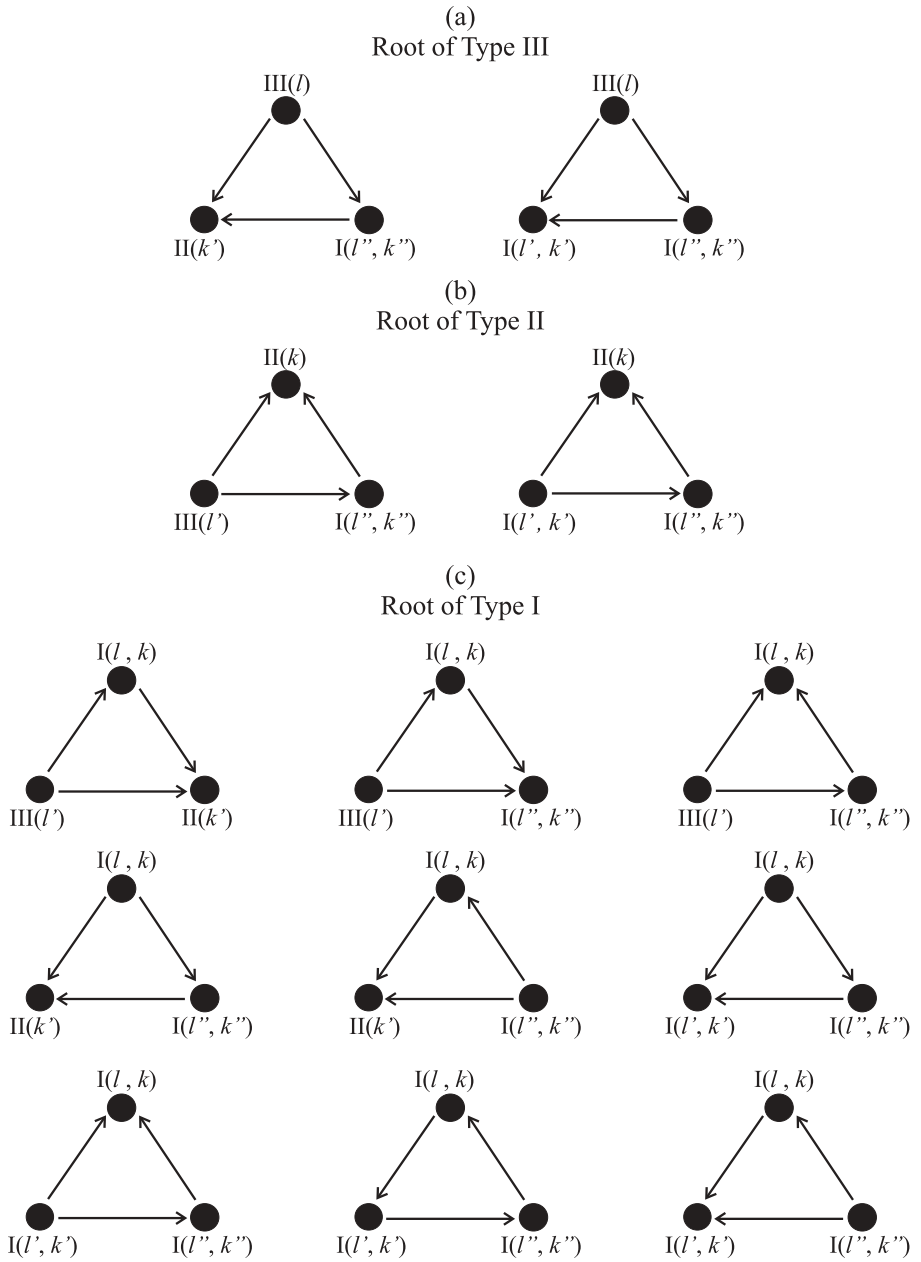


Figure 4.9: Possible configurations of triangles. (a) Triangles with a root of Type III($l$) are constituted by nodes of Type II($k'$) and I($l'', k''$), or of Type I($l', k'$) and I($l'', k''$). (b) Triangles with a root of Type II($k$) may consist of nodes of Type III($l'$) and I($l'', k''$), or of Type I($l', k'$) and I($l'', k''$). (c) Triangles with a root of Type I($l, k$) may contain nodes of Type III($l'$) Type II($k'$), of Type III($l'$) and I($l'', k''$), of Type II($k'$) and I($l'', k''$), or of Type I($l', k'$) and I($l'', k''$).

The latter term in Eq. 4.4.49 comes from nodes of Type I with RSs of length $l$. In the same way, we may write down the expression for the average of the total number of edges $\widetilde{E}_k$, terminating at nodes with PRs of length $k$,

$$\begin{aligned} \widetilde{E}_k &= \sum_{l' \leq k} E_{l'k} = N_{\rm PR}(k) \sum_{l' \leq k} N_{\rm RS}(l') p(l', k) = N_{\rm PR}(k)\ d_{{\rm i},\,k}\ , \\ &= \psi_k + \eta_{\rm RS} N_{\rm PR}(k)\ d_{{\rm i},\,k}\ , \end{aligned} \tag{4.4.51}$$

where $\psi_k$ is defined as the average number of nearest neighbors of nodes of Type II($k$),

$$\psi_k = (1-\eta_{\rm RS}) N_{\rm PR}(k)\ d_{{\rm i},\,k} = (1-\eta_{\rm RS})\eta_{\rm PR} N p_{\rm PR}(k)\ d_{{\rm i},\,k}\ . \tag{4.4.52}$$

Again, the latter term in Eq. 4.4.51 comes from nodes of Type I with PRs of length $k$. One may easily see that $\lambda_{l,\,k}$, the average number of edges incident on nodes of Type I($l, k$), is

$$\lambda_{l,\,k} = \eta_{\rm PR} p_{\rm PR}(k) E_l + \eta_{\rm RS} p_{\rm RS}(l) \widetilde{E}_k = \eta_{\rm RS}\eta_{\rm PR} N p_{\rm RS}(l) p_{\rm PR}(k) (d_{{\rm o},\,l} + d_{{\rm i},\,k})\ . \tag{4.4.53}$$

This is also the average number of nearest neighbors of nodes of Type I($l, k$). Note here that if we sum all these contributions, $\varphi_l$, $\psi_k$ and $\lambda_{l,\,k}$, over all possible values of $l$ and $k$, we will obtain the total number of nearest neighbors of all the nodes, which is equal to twice the total number of edges, $\sum_d dN(d) = \sum_l \varphi_l + \sum_k \psi_k + \sum_{l,\,k} \lambda_{l,\,k} = 2E$.

Nodes of Type III may have connections with nodes of Type I or II (see Fig. 4.7). The probability $\varphi_l^{\rm I}(l', k')$, of finding a node of Type I($l', k'$) among all the nearest neighbors of nodes of Type III($l$), is given by

$$\begin{aligned} \varphi_l^{\rm I}(l', k') &= \frac{(1-\eta_{\rm PR}) N_{\rm RS}(l) \eta_{\rm RS} p_{\rm RS}(l') N_{\rm PR}(k') p(l, k')}{\varphi_l}\ , \\ &= \frac{\eta_{\rm RS}\eta_{\rm PR} N}{d_{{\rm o},\,l}}\ p_{\rm RS}(l') p_{\rm PR}(k') p(l, k')\ . \end{aligned} \tag{4.4.54}$$

The probability $\varphi_l^{\rm II}(k')$, of finding a node of Type II($k'$) among all the nearest neighbors of nodes of Type III($l$), is given by

$$\begin{aligned} \varphi_l^{\rm II}(k') &= \frac{(1-\eta_{\rm PR}) N_{\rm RS}(l) (1-\eta_{\rm RS}) N_{\rm PR}(k') p(l, k')}{\varphi_l}\ , \\ &= \frac{(1-\eta_{\rm RS})\eta_{\rm PR} N}{d_{{\rm o},\,l}}\ p_{\rm PR}(k') p(l, k')\ . \end{aligned} \tag{4.4.55}$$

Nodes of Type II may have connections with nodes of Type I or III (see Fig. 4.7). The probability $\psi_k^{\rm I}(l',k')$, of finding a node of Type I$(l',k')$ among all the nearest neighbors of nodes of Type II$(k)$, is given by

$$\begin{aligned}\psi_k^{\rm I}(l',k') &= \frac{(1-\eta_{\rm RS})N_{\rm PR}(k)\eta_{\rm PR}p_{\rm PR}(k')N_{\rm RS}(l')p(l',k)}{\psi_k} \ , \\ &= \frac{\eta_{\rm RS}\eta_{\rm PR}N}{d_{{\rm i},\,k}}\, p_{\rm RS}(l')p_{\rm PR}(k')p(l',k) \ . \end{aligned} \tag{4.4.56}$$

The probability $\psi_k^{\rm III}(l')$, of finding a node of Type III$(l')$ among all the nearest neighbors of nodes of Type II$(k)$, is given by

$$\begin{aligned}\psi_k^{\rm III}(l') &= \frac{(1-\eta_{\rm RS})N_{\rm PR}(k)(1-\eta_{\rm PR})N_{\rm RS}(l')p(l',k)}{\psi_k} \ , \\ &= \frac{\eta_{\rm RS}(1-\eta_{\rm PR})N}{d_{{\rm i},\,k}}\, p_{\rm RS}(l')p(l',k) \ . \end{aligned} \tag{4.4.57}$$

Finally we may write down the probabilities for the nearest neighbors of nodes of Type I$(l,k)$, to be of Type I, II, or III (see Fig. 4.7). The probability $\lambda_{l,\,k}^{\rm I}(l',k')$, of encountering a node of Type I$(l',k')$ among all the nearest neighbors of nodes of Type I$(l,k)$ is given by

$$\begin{aligned}\lambda_{l,\,k}^{\rm I}(l',k') &= \frac{\eta_{\rm RS}\eta_{\rm PR}Np_{\rm RS}(l)p_{\rm PR}(k)\eta_{\rm RS}\eta_{\rm PR}Np_{\rm RS}(l')p_{\rm PR}(k')[p(l,k')+p(l',k)]}{\lambda_{l,\,k}} \ , \\ &= \lambda_{l,\,k}^{\rm I(i)}(l',k') + \lambda_{l,\,k}^{\rm I(o)}(l',k') \ , \end{aligned} \tag{4.4.58}$$

where $\lambda_{l,\,k}^{\rm I(i)}(l',k')$ is defined as the probability that a randomly chosen node among the nearest neighbors of nodes of Type I$(l,k)$ is of Type I$(l',k')$, and connected to the root by an in-coming edge,

$$\lambda_{l,\,k}^{\rm I(i)}(l',k') = \frac{\eta_{\rm RS}\eta_{\rm PR}N}{d_{{\rm o},\,l}+d_{{\rm i},\,k}}\, p_{\rm RS}(l')p_{\rm PR}(k')p(l,k') \ . \tag{4.4.59}$$

In the similar way, $\lambda_{l,\,k}^{\rm I(o)}(l',k')$ is defined as the probability that a randomly chosen node among the nearest neighbors of nodes of Type I$(l,k)$ is of Type I$(l',k')$ and connected to the root by an out-going edge,

$$\lambda_{l,\,k}^{\rm I(o)}(l',k') = \frac{\eta_{\rm RS}\eta_{\rm PR}N}{d_{{\rm o},\,l}+d_{{\rm i},\,k}}\, p_{\rm RS}(l')p_{\rm PR}(k')p(l',k) \ . \tag{4.4.60}$$

The probability $\lambda_{l,\,k}^{\rm II}(k')$, of finding a node of Type II$(k')$ among all the nearest neighbors of nodes of Type I$(l,k)$ is given by

$$\begin{aligned}\lambda_{l,\,k}^{\rm II}(k') &= \frac{\eta_{\rm RS}\eta_{\rm PR}Np_{\rm RS}(l)p_{\rm PR}(k)(1-\eta_{\rm RS})N_{\rm PR}(k')p(l,k')}{\lambda_{l,\,k}} \ , \\ &= \frac{(1-\eta_{\rm RS})\eta_{\rm PR}N}{d_{{\rm o},\,l}+d_{{\rm i},\,k}}\, p_{\rm PR}(k')p(l,k') \ . \end{aligned} \tag{4.4.61}$$

The probability $\lambda^{\rm III}_{l,\,k}(l')$, of finding a node of Type III($l'$) among all the nearest neighbors of nodes of Type I($l,k$) is given by

$$\begin{aligned}\lambda^{\rm III}_{l,\,k}(l') &= \frac{\eta_{\rm RS}\eta_{\rm PR}Np_{\rm RS}(l)p_{\rm PR}(k)(1-\eta_{\rm PR})N_{\rm RS}(l')p(l',k)}{\lambda_{l,\,k}} \ , \\ &= \frac{\eta_{\rm RS}(1-\eta_{\rm PR})N}{d_{{\rm o},\,l}+d_{{\rm i},\,k}}\ p_{\rm RS}(l')p(l',k) \ . \end{aligned} \tag{4.4.62}$$

In Fig. 4.9 we display all the possible configurations of triangles, grouped with respect to the types of their roots. These configurations are also all the possible ways of obtaining three completely connected nodes, i.e., a 3-clique [67]. We will calculate the connection probabilities starting with the easiest ones. We may write $\Delta_{\rm III}(l)$, the probability for an edge to exist between the nearest neighbors of a randomly selected root of Type III($l$), as a sum

$$\Delta_{\rm III}(l) = \Delta^{\rm I-I}_{\rm III}(l) + \Delta^{\rm I-II}_{\rm III}(l) \ , \tag{4.4.63}$$

where $\Delta^{\rm I-I}_{\rm III}(l)$ comes from the triangles with nodes which are both of Type I and $\Delta^{\rm I-II}_{\rm III}(l)$ from the triangles with nodes of Type I and Type II, with their apex which is a root of Type III($l$) (see Fig. 4.9a for better visualization). Thus, $\Delta^{\rm I-I}_{\rm III}(l)$ is the probability for an edge to exist between two randomly selected Type I neighbors of a randomly selected root of Type III($l$), etc. These quantities are calculated to be

$$\Delta^{\rm I-I}_{\rm III}(l) = 2\sum_{l'}\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}\varphi^{\rm I}_l(l',k')\ \varphi^{\rm I}_l(l'',k'')\ p(l'',k') \ , \tag{4.4.64}$$

$$= \ 2\frac{\eta^2_{\rm RS}\eta^2_{\rm PR}N^2}{d^2_{{\rm o},\,l}}\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}p_{\rm PR}(k')p_{\rm RS}(l'')p_{\rm PR}(k'')p(l,k')p(l,k'')p(l'',k') \ ,$$

and

$$\Delta^{\rm I-II}_{\rm III}(l) = 2\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}\varphi^{\rm II}_l(k')\ \varphi^{\rm I}_l(l'',k'')\ p(l'',k') \ , \tag{4.4.65}$$

$$\begin{aligned} = \ 2\frac{(1-\eta_{\rm RS})\eta_{\rm RS}\eta^2_{\rm PR}N^2}{d^2_{{\rm o},\,l}}\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}&[p_{\rm PR}(k')p_{\rm RS}(l'')p_{\rm PR}(k'') \\ &\cdot p(l,k')p(l,k'')p(l'',k')] \ , \end{aligned}$$

where $2\varphi^{\rm T}_l(.)\ \varphi^{\rm T'}_l(.)$ is the probability of encountering a pair of nodes of Type T(.) and Type T$'$(.) among the nearest neighbors of nodes of Type III($l$). Since the interaction terms in $\Delta^{\rm I-I}_{\rm III}(l)$ are independent from the variable $l'$ we have carried

out the summation over it, $\sum_{l'} p_{\rm RS}(l') = 1$. Now if we perform the summations over the variables each of which occurs once in the interaction terms, namely $l''$ and $k''$, we get

$$\Delta_{\rm III}(l) = 2\frac{\eta_{\rm PR}}{d_{{\rm o},\,l}} \sum_{k' \geq l} p_{\rm PR}(k')\; d_{{\rm i},\,k'}\; p(l,k') \quad . \tag{4.4.66}$$

In a similar way we may easily calculate $\Delta_{\rm II}(k)$, the probability for an edge to exist between the nearest neighbors of a randomly chosen node of Type II($k$),

$$\Delta_{\rm II}(k) = \Delta_{\rm II}^{\rm I-I}(k) + \Delta_{\rm II}^{\rm III-I}(k) \quad , \tag{4.4.67}$$

where $\Delta_{\rm II}^{\rm I-I}(k)$ comes from the triangles with nodes which are both of Type I and $\Delta_{\rm II}^{\rm III-I}(k)$ from the triangles with nodes of types I and III, rooted at a randomly selected node of Type II($k$) (see Fig. 4.9b). These quantities are given by

$$\Delta_{\rm II}^{\rm I-I}(k) = 2\sum_{l' \leq k}\sum_{k'} \sum_{l'' \leq k}\sum_{k'' \geq l'} \psi_k^{\rm I}(l',k')\; \psi_k^{\rm I}(l'',k'')\; p(l',k'') \quad , \tag{4.4.68}$$

$$= \; 2\frac{\eta_{\rm RS}^2 \eta_{\rm PR}^2 N^2}{d_{{\rm i},\,k}^2} \sum_{l' \leq k} \sum_{l'' \leq k} \sum_{k'' \geq l'} p_{\rm RS}(l') p_{\rm RS}(l'') p_{\rm PR}(k'') p(l',k) p(l'',k) p(l',k'') \quad ,$$

and

$$\Delta_{\rm II}^{\rm III-I}(k) = 2\sum_{l' \leq k} \sum_{l'' \leq k}\sum_{k'' \geq l'} \psi_k^{\rm III}(l')\; \psi_k^{\rm I}(l'',k'')\; p(l',k'') \quad , \tag{4.4.69}$$

$$= \; 2\frac{\eta_{\rm RS}^2 \eta_{\rm PR}(1-\eta_{\rm PR}) N^2}{d_{{\rm i},\,k}^2} \sum_{l' \leq k} \sum_{l'' \leq k} \sum_{k'' \geq l'} [p_{\rm RS}(l') p_{\rm RS}(l'') p_{\rm PR}(k'') \cdot p(l',k) p(l'',k) p(l',k'')] \quad ,$$

where $2\psi_k^{\rm T}(.)\; \psi_k^{\rm T'}(.)$ is the probability of encountering a pair of nodes of Type T(.) and Type T$'$(.) among the nearest neighbors of nodes of Type II($k$). Since the interaction terms in $\Delta_{\rm II}^{\rm I-I}(k)$ are independent from the variable $k'$ we have carried out the summation over it, $\sum_{k'} p_{\rm PR}(k') = 1$. If we perform the summations over the variables each of which occurs once in the interaction terms, namely $l''$ and $k''$, we get

$$\Delta_{\rm II}(k) = 2\frac{\eta_{\rm RS}}{d_{{\rm i},\,k}} \sum_{l' \leq k} p_{\rm RS}(l')\; d_{{\rm o},\,l'}\; p(l',k) \quad . \tag{4.4.70}$$

Now we will calculate $\Delta_{\rm I}(l,k)$, the probability for an edge to exist between the nearest neighbors of a randomly selected node of Type I($l,k$), term by term coming from the different types of nearest neighbors of such nodes (see Fig. 4.9c).

The first possible configuration contributing to the number of triangles comes from nodes of Type III and Type II, $\Delta_{\rm I}^{\rm III-II}(l,k)$,

$$\Delta_{\rm I}^{\rm III-II}(l,k) = 2\sum_{l'\le k}\sum_{k'\ge\max(l,\,l')} \lambda^{\rm III}_{l,\,k}(l')\;\lambda^{\rm II}_{l,\,k}(k')\;p(l',k') \;\;, \tag{4.4.71}$$

$$= \; 2\frac{\eta_{\rm RS}(1-\eta_{\rm RS})\eta_{\rm PR}(1-\eta_{\rm PR})N^2}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{l'\le k}\sum_{k'\ge\max(l,\,l')} [p_{\rm RS}(l')p_{\rm PR}(k') \cdot p(l',k)p(l,k')p(l',k')] \;\;.$$

The second configuration comes from nodes of Type III and Type I, $\Delta_{\rm I}^{\rm III-I}(l,k)$ which can be written as a sum

$$\Delta_{\rm I}^{\rm III-I}(l,k) = \Delta_{\rm I}^{\rm III-I(i)}(l,k) + \Delta_{\rm I}^{\rm III-I(o)}(l,k) \;\;, \tag{4.4.72}$$

where $\Delta_{\rm I}^{\rm III-I(i)}(l,k)$ comes from the triangles with nodes of Type III and Type I which is connected to the root by an in-coming edge and $\Delta_{\rm I}^{\rm III-I(o)}(l,k)$ from the triangles with nodes of Type III and Type I which is connected to the root by an out-going edge. These quantities are calculated to be

$$\Delta_{\rm I}^{\rm III-I(i)}(l,k) = 2\sum_{l'\le k}\sum_{l''}\sum_{k''\ge\max(l,\,l')} \lambda^{\rm III}_{l,\,k}(l')\;\lambda^{\rm I(i)}_{l,\,k}(l'',k'')\;p(l',k'') \;\;, \tag{4.4.73}$$

$$= \; 2\frac{\eta_{\rm RS}^2\eta_{\rm PR}(1-\eta_{\rm PR})N^2}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{l'\le k}\sum_{k''\ge\max(l,\,l')} p_{\rm RS}(l')p_{\rm PR}(k'')p(l',k)p(l,k'')p(l',k'') \;\;,$$

and

$$\Delta_{\rm I}^{\rm III-I(o)}(l,k) = 2\sum_{l'\le k}\sum_{l''\le k}\sum_{k''\ge l'} \lambda^{\rm III}_{l,\,k}(l')\;\lambda^{\rm I(o)}_{l,\,k}(l'',k'')\;p(l',k'') \;\;, \tag{4.4.74}$$

$$= \; 2\frac{\eta_{\rm RS}(1-\eta_{\rm PR})\;d_{{\rm i},\,k}}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{l'\le k} p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;p(l',k) \;\;.$$

The third contribution $\Delta_{\rm I}^{\rm II-I}(l,k)$, comes from the nodes of Type II and Type I and can be written as a sum

$$\Delta_{\rm I}^{\rm II-I}(l,k) = \Delta_{\rm I}^{\rm II-I(i)}(l,k) + \Delta_{\rm I}^{\rm II-I(o)}(l,k) \;\;, \tag{4.4.75}$$

where $\Delta_{\rm I}^{\rm II-I(i)}(l,k)$ comes from the triangles with nodes of Type II and Type I which is connected to the root by an in-coming edge and $\Delta_{\rm I}^{\rm II-I(o)}(l,k)$ from the triangles with nodes of Type II and Type I which is connected to the root by an

out-going edge. These quantities are calculated to be

$$\Delta_{\rm I}^{\rm II-I(i)}(l,k) = 2\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}\lambda^{\rm II}_{l,\,k}(k')\;\lambda^{\rm I(i)}_{l,\,k}(l'',k'')\;p(l'',k')\;\;, \quad (4.4.76)$$

$$= \;\; 2\frac{(1-\eta_{\rm RS})\eta_{\rm PR}\;d_{{\rm o},\,l}}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{k'\geq l}p_{\rm PR}(k')\;d_{{\rm i},\,k'}\;p(l,k')\;\;,$$

and

$$\Delta_{\rm I}^{\rm II-I(o)}(l,k) = 2\sum_{k'\geq l}\sum_{l''\leq \min(k,\,k')}\sum_{k''}\lambda^{\rm II}_{l,\,k}(k')\;\lambda^{\rm I(o)}_{l,\,k}(l'',k'')\;p(l'',k')\;\;, \quad (4.4.77)$$

$$= \;\; 2\frac{\eta_{\rm RS}(1-\eta_{\rm RS})\eta_{\rm PR}^2N^2}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{l''\leq k}\sum_{k'\geq\max(l,\,l'')}p_{\rm RS}(l'')p_{\rm PR}(k')p(l'',k)p(l,k')p(l'',k')\;\;,$$

note here that $\sum_{k'\geq l}\;\sum_{l''\leq\min(k,\,k')}=\sum_{l''\leq k}\;\sum_{k'\geq\max(l,\,l'')}$. The fourth contribution $\Delta_{\rm I}^{\rm I(i)-I(i)}(l,k)$, comes from nodes which are both of Type I, here the root is connected to them by out-going edges,

$$\Delta_{\rm I}^{\rm I(i)-I(i)}(l,k) = 2\sum_{l'}\sum_{k'\geq l}\sum_{l''\leq k'}\sum_{k''\geq l}\lambda^{\rm I(i)}_{l,\,k}(l',k')\;\lambda^{\rm I(i)}_{l,\,k}(l'',k'')\;p(l'',k')\;\;, \quad (4.4.78)$$

$$= 2\frac{\eta_{\rm RS}\eta_{\rm PR}\;d_{{\rm o},\,l}}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\;\sum_{k'\geq l}p_{\rm PR}(k')\;d_{{\rm i},\,k'}\;p(l,k')\;\;.$$

The fifth contribution $\Delta_{\rm I}^{\rm I(o)-I(o)}(l,k)$, comes from the pairs of nodes each of which is of Type I where the root is connected to them by in-coming edges,

$$\Delta_{\rm I}^{\rm I(o)-I(o)}(l,k) = 2\sum_{l'\leq k}\sum_{k'}\sum_{l''\leq k}\sum_{k''\geq l'}\lambda^{\rm I(o)}_{l,\,k}(l',k')\;\lambda^{\rm I(o)}_{l,\,k}(l'',k'')\;p(l',k'')\;\;, \quad (4.4.79)$$

$$= 2\frac{\eta_{\rm RS}\eta_{\rm PR}\;d_{{\rm i},\,k}}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\;\sum_{l'\leq k}p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;p(l',k)\;\;.$$

The sixth contribution $\Delta_{\rm I}^{\rm I(i)\to I(o)}(l,k)$, comes from the pairs of nodes which are both of Type I where the root is connected to the first node by an out-going and to the second one by an in-coming edge, and the first one is connected to the second one by an out-going edge,

$$\Delta_{\rm I}^{\rm I(i)\to I(o)}(l,k) = 2\sum_{l'}\sum_{k'\geq l}\sum_{l''\leq k}\sum_{k''\geq l'}\lambda^{\rm I(i)}_{l,\,k}(l',k')\;\lambda^{\rm I(o)}_{l,\,k}(l'',k'')\;p(l',k'')\;\;, \quad (4.4.80)$$

$$= 2\frac{\eta_{\rm RS}\;d_{{\rm o},\,l}\;d_{{\rm i},\,k}}{N(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\;\sum_{l'}p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;\;,$$

note here that $\eta_{\rm RS}\sum_{l'}p_{\rm RS}(l')\;d_{{\rm o},\,l'}=\eta_{\rm PR}\sum_{k'}p_{\rm PR}(k')\;d_{{\rm i},\,k'}$ (see Eq. 4.3.37). The last and the seventh contribution $\Delta_{\rm I}^{\rm I(i)\leftarrow I(o)}(l,k)$, comes from the connection configuration of the nodes which are both of Type I where the root is connected to

the first node by an out-going and to the second one by an in-coming edge, and in this case the first one is connected to the second one by an in-coming edge,

$$\Delta_{\rm I}^{\rm I(i)\leftarrow I(o)}(l,k) = 2\sum_{l'}\sum_{k'\geq l}\sum_{l''\leq \min(k,\,k')}\sum_{k''}\lambda_{l,\,k}^{\rm I(i)}(l',k')\;\lambda_{l,\,k}^{\rm I(o)}(l'',k'')\;p(l'',k')\;\;, \tag{4.4.81}$$

$$= 2\frac{\eta_{\rm RS}^2\eta_{\rm PR}^2N^2}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\sum_{l''\leq k}\sum_{k'\geq \max(l,\,l'')}p_{\rm RS}(l'')p_{\rm PR}(k')p(l'',k)p(l,k')p(l'',k')\;\;.$$

Now if we sum all the contributions by making necessary simplifications we get $\Delta_{\rm I}(l,k)$, the probability for an edge to exist between the nearest neighbors of a randomly selected node of Type I$(l,k)$, as

$$\Delta_{\rm I}(l,k) = \frac{2}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\;\Bigg\{\eta_{\rm RS}\;\frac{d_{{\rm o},\,l}\;d_{{\rm i},\,k}}{N}\;\sum_{l'}p_{\rm RS}(l')\;d_{{\rm o},\,l'}$$
$$+\;\eta_{\rm RS}\;d_{{\rm i},\,k}\;\sum_{l'\leq k}p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;p(l',k)+\eta_{\rm PR}\;d_{{\rm o},\,l}\sum_{k'\geq l}p_{\rm PR}(k')\;d_{{\rm i},\,k'}\;p(l,k')$$
$$+\;\;\eta_{\rm RS}\eta_{\rm PR}N^2\sum_{l''\leq k}\sum_{k'\geq\max(l,\,l'')}p_{\rm RS}(l'')p(l'',k)p_{\rm PR}(k')p(l,k')p(l'',k')\Bigg\}\;\;. \tag{4.4.82}$$

If we substitute all the terms into Eq. 4.4.45 we obtain $c(d)$, the probability that a randomly chosen pair of nodes rooted at a randomly selected node of degree $d$ is connected, as

$$c(d) = 2\frac{\eta_{\rm RS}\eta_{\rm PR}}{P(d)}\;\Bigg\{(1-\eta_{\rm PR})\sum_{l}p_{\rm RS}(l)\frac{P_l^{\rm out}(d)}{d_{{\rm o},\,l}}\;\sum_{k'\geq l}p_{\rm PR}(k')\;d_{{\rm i},\,k'}\;p(l,k')$$
$$+\quad(1-\eta_{\rm RS})\sum_{k}p_{\rm PR}(k)\frac{P_k^{\rm in}(d)}{d_{{\rm i},\,k}}\sum_{l'\leq k}p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;p(l',k)$$
$$+\quad\sum_{l,\,k}p_{\rm RS}(l)p_{\rm PR}(k)\;\frac{P_{l,\,k}(d)}{(d_{{\rm o},\,l}+d_{{\rm i},\,k})^2}\;\Bigg[\eta_{\rm RS}\;\frac{d_{{\rm o},\,l}\;d_{{\rm i},\,k}}{N}\;\sum_{l'}p_{\rm RS}(l')\;d_{{\rm o},\,l'}$$
$$+\quad\eta_{\rm RS}\;d_{{\rm i},\,k}\;\sum_{l'\leq k}p_{\rm RS}(l')\;d_{{\rm o},\,l'}\;p(l',k)+\eta_{\rm PR}\;d_{{\rm o},\,l}\sum_{k'\geq l}p_{\rm PR}(k')\;d_{{\rm i},\,k'}\;p(l,k')$$
$$+\quad\eta_{\rm RS}\eta_{\rm PR}N^2\sum_{l''\leq k}\sum_{k'\geq\max(l,\,l'')}p_{\rm RS}(l'')p_{\rm PR}(k')p(l'',k)p(l,k')p(l'',k')\Bigg]\Bigg\}\;\;. \tag{4.4.83}$$

The comparison of simulation results for the average clustering coefficient spectrum with our analytical results is displayed in Fig. 4.10. One may easily note the

close similarity between Figs. (4.8, 4.10). The nearest neighbors of those nodes dominating the small degree region, are nodes having RSs and which are relatively well connected to each other. On the other hand, most of the nearest neighbors of those nodes in the large degree region, are nodes having only PRs, who are unable to make connections with each other. The nonzero clustering coefficient in the large degree region is due to the interactions of the relatively few nodes that also have RSs.

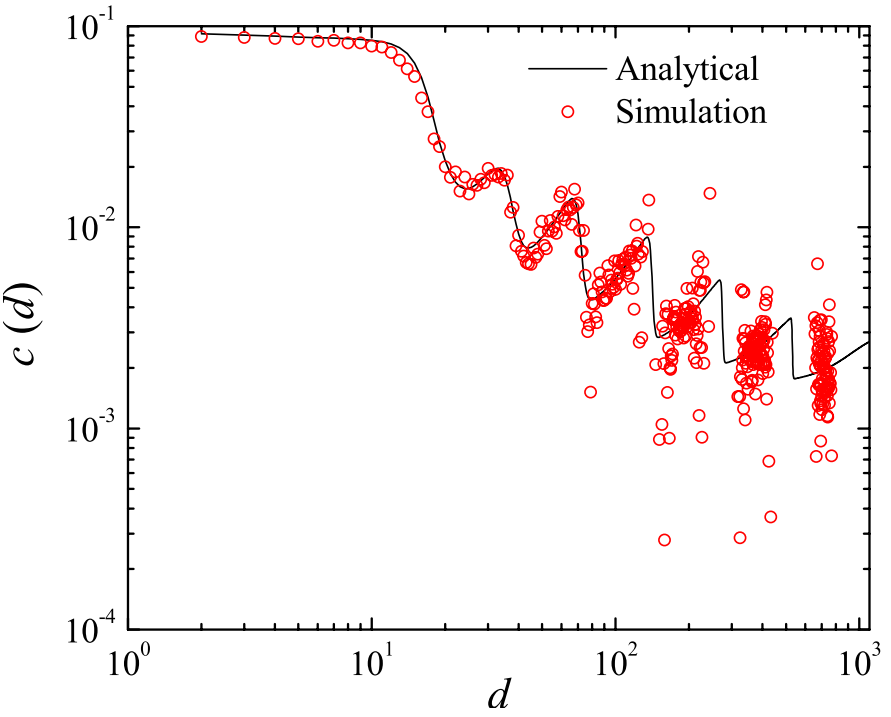


Figure 4.10: The probability of finding an edge between a randomly selected pair of nearest neighbors of nodes with degree $d$ for the analytical solution (see Eq. 4.4.83) and simulations (red circles).

## 4.5 Rich-club Coefficient

We may group the nodes in our network with respect to their total degrees, for example, the subgraph containing the nodes with degrees greater than a given value $d$, and the edges connecting these nodes may give us an idea on how the "rich guys" in the network are connected among themselves. The quantity measuring the well-connectedness of nodes with degrees greater than $d$, namely the rich-club coefficient $r(d)$, is the probability that a randomly selected pair of nodes with degrees greater than $d$ are connected [26, 27]. In practice, one may calculate

this probability by counting the number of edges among the nodes with degrees greater than $d$, $E_{>d}$, and then dividing this number by its possible maximum value, $N_{>d}(N_{>d}-1)/2$ where $N_{>d}$ is the number of such nodes. In this section we will calculate the rich-club coefficient for the ensemble of networks in question.

We may define $Q(d)$ as the probability that a randomly chosen node has degree greater than $d$. This probability is obviously given by

$$Q(d) = \sum_{d'>d} P(d') = 1 - \sum_{d'=0}^{d} P(d') \ , \tag{4.5.84}$$

where $P(d')$ is the total degree distribution. If we use the expression for the degree distribution in Eq. 4.2.27 and make the necessary simplification we get,

$$\begin{aligned} Q(d) &= \Big\{ \eta_{\rm RS}(1-\eta_{\rm PR}) \sum_l p_{\rm RS}(l) Q_l^{\rm out}(d) + (1-\eta_{\rm RS})\eta_{\rm PR} \sum_k p_{\rm PR}(k) Q_k^{\rm in}(d) \\ &+ \eta_{\rm RS}\eta_{\rm PR} \sum_{l,k} p_{\rm RS}(l) p_{\rm PR}(k) Q_{l,\,k}(d) \Big\} \ . \end{aligned} \tag{4.5.85}$$

One may easily observe that $Q_l^{\rm out}(d) = 1 - \sum_{d'=0}^{d} P_l^{\rm out}(d')$ is the probability of finding a node with out-degree greater than $d$ among all the nodes with RSs of length $l$, and $Q_k^{\rm in}(d) = 1 - \sum_{d'=0}^{d} P_k^{\rm in}(d')$ is the probability of finding a node with in-degree greater than $d$ among all the nodes with PRs of length $k$. The last term in the above equation $Q_{l,\,k}(d) = 1 - \sum_{d'=0}^{d} P_{l,\,k}(d')$, gives the the probability of finding a node with total degree greater than $d$ among all the nodes with RSs of length $l$ and PRs of length $k$. Now, in a similar way to what we have done in the previous section, we may write down the probabilities of finding nodes of given types and hidden-variables in the set of nodes with degrees greater than $d$. The probability of finding a node of Type I$(l,k)$ among all the nodes with degrees greater than $d$, $Q_{\rm I}(l,k|d)$, is given by

$$Q_{\rm I}(l,k|d) = \eta_{\rm RS}\eta_{\rm PR} p_{\rm RS}(l) p_{\rm PR}(k) \, \frac{Q_{l,\,k}(d)}{Q(d)} \ . \tag{4.5.86}$$

The probability, $Q_{\rm II}(k|d)$ that a randomly selected node among all the nodes with degrees greater than $d$ is of Type II$(k)$ is

$$Q_{\rm II}(k|d) = (1-\eta_{\rm RS})\eta_{\rm PR} p_{\rm PR}(k) \, \frac{Q_k^{\rm in}(d)}{Q(d)} \ . \tag{4.5.87}$$

The probability of finding a node of Type III($l$) among all the nodes with degrees greater than $d$, $Q_{\rm III}(l|d)$, is given by

$$Q_{\rm III}(l|d) = \eta_{\rm RS}(1-\eta_{\rm PR})p_{\rm RS}(l)\ \frac{Q_l^{\rm out}(d)}{Q(d)}\ \ . \tag{4.5.88}$$

We may write the rich-club coefficient $r(d)$, as a sum of the connection probabilities between different types of nodes $r_{\rm T-T'}(d)$. Let us randomly chose a pair of nodes from the set of all the nodes with degrees greater than $d$. The probability that this pair contains one node of Type T and a second of Type T$'$ and that these two nodes are connected is defined as $r_{\rm T-T'}(d)$. Then, $r(d)$ is given by

$$r(d) = \sum_{\rm T,\,T'} r_{\rm T-T'}(d)\ \ . \tag{4.5.89}$$

We find $r_{\rm I-I}(d)$ as

$$r_{\rm I-I}(d) = 2\sum_{l,\,k}\ \sum_{l',\,k'\geq l} Q_{\rm I}(l,k|d)\ Q_{\rm I}(l',k'|d)\ p(l,k')\ \ , \tag{4.5.90}$$

$$= \ 2\left(\frac{\eta_{\rm RS}\eta_{\rm PR}}{Q(d)}\right)^2 \sum_{l,\,k}\ \sum_{l',\,k'\geq l} p_{\rm RS}(l)p_{\rm PR}(k)p_{\rm RS}(l')p_{\rm PR}(k')p(l,k')Q_{l,\,k}(d)Q_{l',\,k'}(d)\ \ ,$$

where $2Q_{\rm T}(.|d)Q_{\rm T'}(.|d)$ is the probability that a pair of nodes chosen at random among the nodes with degrees greater than $d$ are of Type T(.) and Type T$'$(.). One may observe in the above equation that the terms are coupled, so it is very hard to lead the analytical calculations further. But one may also recognize that

$$Q_{l,\,k}(d)Q_{l',\,k'}(d) = \left(1-\sum_{d'\leq d} P_{l,\,k}(d')\right)\left(1-\sum_{d''\leq d} P_{l',\,k'}(d'')\right)\ \ ,$$

$$= \ 1-\sum_{d'\leq d} P_{l,\,k}(d') - \sum_{d''\leq d} P_{l',\,k'}(d'') + \sum_{d'\leq d}\sum_{d''\leq d} P_{l,\,k}(d')P_{l',\,k'}(d'') \tag{4.5.91}$$

If we substitute this equality into $r_{\rm I-I}(d)$ we obtain,

$$\begin{aligned} r_{\rm I-I}(d) = {} & 2\frac{\eta_{\rm RS}^2\eta_{\rm PR}^2}{Q(d)^2}\ \left\{\frac{E}{\eta_{\rm RS}\eta_{\rm PR}N^2}\right. \\ - \ & \frac{1}{\eta_{\rm PR}N}\ \sum_{d'\leq d}\ \sum_{l,\,k} p_{\rm RS}(l)p_{\rm PR}(k)\ d_{{\rm o},\,l}\ P_{l,\,k}(d') \\ - \ & \frac{1}{\eta_{\rm RS}N}\ \sum_{d''\leq d}\ \sum_{l',\,k'} p_{\rm RS}(l')p_{\rm PR}(k')\ d_{{\rm i},\,k'}\ P_{l',\,k'}(d'') \\ + \ & \left.\sum_{d'\leq d}\ \sum_{d''\leq d}\ \sum_{l,\,k}\ \sum_{l',\,k'\geq l} p_{\rm RS}(l)p_{\rm PR}(k)p_{\rm RS}(l')p_{\rm PR}(k')p(l,k')P_{l,\,k}(d')P_{l',\,k'}(d'')\right\}\ \ , \end{aligned} \tag{4.5.92}$$

where we have performed the summations over the variables which do not appear in the degree distributions (see Eqs. 4.2.15, 4.2.21, and 4.3.37 for the expressions of the average out-degree $d_{\mathrm{o},\,l}$, the average in-degree $d_{\mathrm{i},\,k'}$, and the average of the total number of edges $E$, respectively). Although the last summation is still not possible to do in closed form, this expression is easier to evaluate numerically than Eq. 4.5.90, at least for small $d$. The probability $r_{\mathrm{I-II}}(d)$, that two randomly chosen nodes with degrees greater than $d$ are of Type I and Type II and connected, may be given in a similar way. This is,

$$
\begin{aligned}
r_{\mathrm{I-II}}(d) &= 2\sum_{l,\,k}\sum_{k'\geq l} Q_{\mathrm{I}}(l,k|d)\; Q_{\mathrm{II}}(k'|d)\; p(l,k') \;\;, \qquad (4.5.93)\\
&= 2\frac{\eta_{\mathrm{RS}}(1-\eta_{\mathrm{RS}})\eta_{\mathrm{PR}}^2}{Q(d)^2}\;\left\{\frac{E}{\eta_{\mathrm{RS}}\eta_{\mathrm{PR}}N^2}\right.\\
&- \frac{1}{\eta_{\mathrm{PR}}N}\sum_{d'\leq d}\sum_{l,\,k} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)\; d_{\mathrm{o},\,l}\; P_{l,\,k}(d')\\
&- \frac{1}{\eta_{\mathrm{RS}}N}\sum_{d''\leq d}\sum_{k'} p_{\mathrm{PR}}(k')\; d_{\mathrm{i},\,k'}\; P^{\mathrm{in}}_{k'}(d'')\\
&+ \left.\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l,\,k}\sum_{k'\geq l} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p_{\mathrm{PR}}(k')p(l,k')P_{l,\,k}(d')P^{\mathrm{in}}_{k'}(d'')\right\} \;\;.
\end{aligned}
$$

The probability $r_{\mathrm{I-III}}(d)$, that two randomly chosen nodes with degree greater than $d$ are of Type I and Type III and connected, is also given by

$$
\begin{aligned}
r_{\mathrm{I-III}}(d) &= 2\sum_{l,\,k}\sum_{l'\leq k} Q_{\mathrm{I}}(l,k|d)\; Q_{\mathrm{III}}(l'|d)\; p(l',k) \;\;, \qquad (4.5.94)\\
&= 2\frac{\eta_{\mathrm{RS}}^2\eta_{\mathrm{PR}}(1-\eta_{\mathrm{PR}})}{Q(d)^2}\;\left\{\frac{E}{\eta_{\mathrm{RS}}\eta_{\mathrm{PR}}N^2}\right.\\
&- \frac{1}{\eta_{\mathrm{RS}}N}\sum_{d'\leq d}\sum_{l,\,k} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)\; d_{\mathrm{i},\,k}\; P_{l,\,k}(d')\\
&- \frac{1}{\eta_{\mathrm{PR}}N}\sum_{d''\leq d}\sum_{l'} p_{\mathrm{RS}}(l')\; d_{\mathrm{o},\,l'}\; P^{\mathrm{out}}_{l'}(d'')\\
&+ \left.\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l,\,k}\sum_{l'\leq k} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p_{\mathrm{RS}}(l')p(l',k)P_{l,\,k}(d')P^{\mathrm{out}}_{l'}(d'')\right\} \;\;.
\end{aligned}
$$

The last term in $r(d)$ is the probability $r_{\mathrm{II-III}}(d)$, that two randomly chosen nodes with degree greater than $d$ are of Type II and Type III and connected, is given

by

$$r_{\mathrm{II-III}}(d) = 2\sum_{l}\sum_{k\geq l} Q_{\mathrm{II}}(k'|d)\; Q_{\mathrm{III}}(l|d)\; p(l,k) \;\;, \tag{4.5.95}$$

$$\begin{aligned}
&= 2\frac{\eta_{\mathrm{RS}}(1-\eta_{\mathrm{RS}})\eta_{\mathrm{PR}}(1-\eta_{\mathrm{PR}})}{Q(d)^2}\;\left\{\frac{E}{\eta_{\mathrm{RS}}\eta_{\mathrm{PR}}N^2}\right.\\
&- \frac{1}{\eta_{\mathrm{PR}}N}\sum_{d'\leq d}\sum_{l} p_{\mathrm{RS}}(l)\; d_{\mathrm{o},\,l}\; P_l^{\mathrm{out}}(d')\\
&- \frac{1}{\eta_{\mathrm{RS}}N}\sum_{d''\leq d}\sum_{k} p_{\mathrm{PR}}(k)\; d_{\mathrm{i},\,k}\; P_k^{\mathrm{in}}(d'')\\
&+ \left.\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l}\sum_{k\geq l} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p(l,k)P_l^{\mathrm{out}}(d')P_k^{\mathrm{in}}(d'')\right\} \;\;.
\end{aligned}$$

Now we can write down the expression for the rich-club coefficient by grouping the similar terms together and remembering the definition of $\eta_{\mathrm{T}}$, that the probability of finding a node of Type T,

$$\begin{aligned}
r(d) &= \frac{2}{N^2Q(d)^2}\;\left\{E\right.\\
&- \eta_{\mathrm{III}}N\sum_{d'\leq d}\sum_{l} p_{\mathrm{RS}}(l)\; d_{\mathrm{o},\,l}\; P_l^{\mathrm{out}}(d')\\
&- \eta_{\mathrm{II}}N\sum_{d'\leq d}\sum_{k} p_{\mathrm{PR}}(k)\; d_{\mathrm{i},\,k}\; P_k^{\mathrm{in}}(d')\\
&- \eta_{\mathrm{I}}N\sum_{d'\leq d}\sum_{l,\,k} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)\;(d_{\mathrm{o},\,l}+d_{\mathrm{i},\,k})\; P_{l,\,k}(d')\\
&+ \eta_{\mathrm{II}}\eta_{\mathrm{III}}N^2\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l}\sum_{k\geq l} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p(l,k)P_l^{\mathrm{out}}(d')P_k^{\mathrm{in}}(d'')\\
&+ \eta_{\mathrm{I}}\eta_{\mathrm{III}}N^2\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l,\,k}\sum_{l'\leq k} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p_{\mathrm{RS}}(l')p(l',k)P_{l,\,k}(d')P_{l'}^{\mathrm{out}}(d'')\\
&+ \eta_{\mathrm{I}}\eta_{\mathrm{II}}N^2\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l,\,k}\sum_{k'\geq l} p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p_{\mathrm{PR}}(k')p(l,k')P_{l,\,k}(d')P_{k'}^{\mathrm{in}}(d'')\\
&+ \eta_{\mathrm{I}}^2N^2\sum_{d'\leq d}\sum_{d''\leq d}\sum_{l,\,k}\sum_{l',\,k'\geq l} [p_{\mathrm{RS}}(l)p_{\mathrm{PR}}(k)p_{\mathrm{RS}}(l')p_{\mathrm{PR}}(k')p(l,k')\\
&\qquad\qquad \cdot P_{l,\,k}(d')P_{l',\,k'}(d'')]\left.\right\} \;\;.
\end{aligned} \tag{4.5.96}$$

Let us note here that the expression obtained above is equivalent to the expres-

sion,

$$
\begin{aligned}
r(d) = \frac{2}{Q(d)^2} \Bigg\{ & \eta_{\rm II}\eta_{\rm III} \sum_{d'>d} \sum_{d''>d} \sum_{l} \sum_{k\geq l} p_{\rm RS}(l) p_{\rm PR}(k) p(l,k) P_l^{\rm out}(d') P_k^{\rm in}(d'') \\
+ \; & \eta_{\rm I}\eta_{\rm III} \sum_{d'>d} \sum_{d''>d} \sum_{l,\,k} \sum_{l'\leq k} p_{\rm RS}(l) p_{\rm PR}(k) p_{\rm RS}(l') p(l',k) P_{l,\,k}(d') P_{l'}^{\rm out}(d'') \\
+ \; & \eta_{\rm I}\eta_{\rm II} \sum_{d'>d} \sum_{d''>d} \sum_{l,\,k} \sum_{k'\geq l} p_{\rm RS}(l) p_{\rm PR}(k) p_{\rm PR}(k') p(l,k') P_{l,\,k}(d') P_{k'}^{\rm in}(d'') \\
+ \; & \eta_{\rm I}^2 \sum_{d'>d} \sum_{d''>d} \sum_{l,\,k} \sum_{l',\,k'\geq l} p_{\rm RS}(l) p_{\rm PR}(k) p_{\rm RS}(l') p_{\rm PR}(k') p(l,k') P_{l,\,k}(d') P_{l',\,k'}(d'') \Bigg\} \; .
\end{aligned}
\tag{4.5.97}
$$

We display in Fig. 4.11, the rich-club coefficient, as obtained by our analytical treatment and via the simulations. Although our theoretical curve is also able demonstrate a non-monotonic behavior, it remains below the simulation results up to the crossover region. The plateaus of the curve correspond to the local minima in the total degree distribution (see Fig. 4.6), where the probabilities of finding nodes with these degrees are very small. Thus, if we successively increase the degree $d$ to search for the nodes with degrees $d' > d$ we do not obtain a new set of nodes till we cross these barriers.

Note that the crossover behavior in all three topological coefficients shown in Figs. (4.8, 4.10, 4.11) essentially arises from the fact that in this network there are two kinds of nodes, namely those that have RSs and those that do not. The discrepancy between the analytic and simulation results in Figs. (4.8, 4.10, 4.11) all fall within the interval where we slightly under estimate the in-degree distribution (see Fig. 4.5) for relatively large degrees. We have argued above that this is due to the approximation of the binomial distributions with Poissonians for greater ease in computation.

## 4.6 Remarks on the Hidden-Variable Approximation

The aim of the section was approaching the content-based networks problem analytically. The discrepancies between the content-based and hidden-variable models arise from two sources. One source is the approximations that go into the derivation of the pairwise connection probability in Eq. 4.0.1. The second is

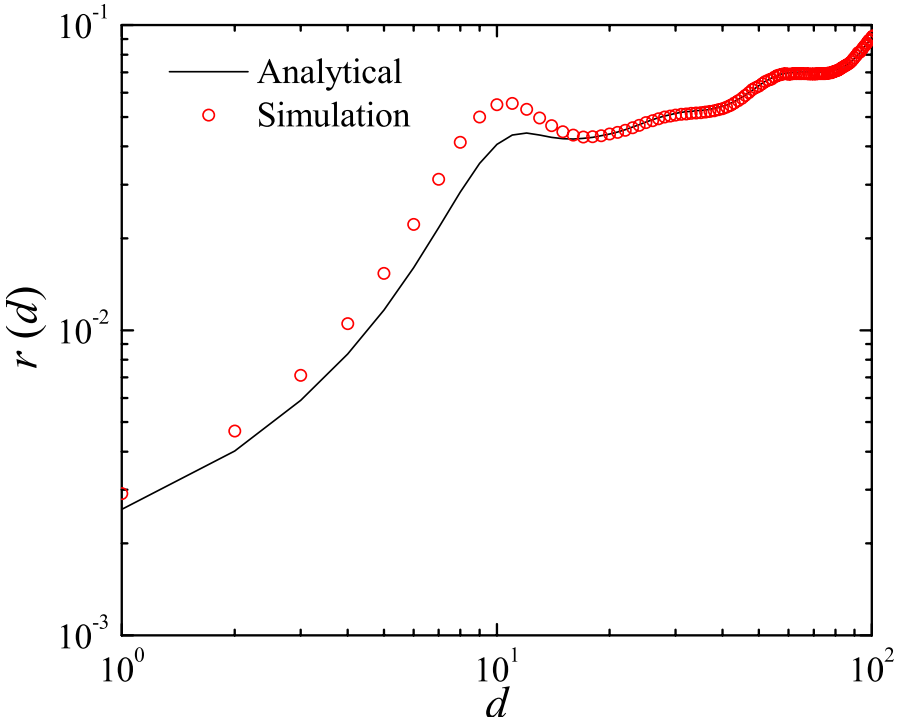


Figure 4.11: The probability of finding an edge between a randomly selected pair of nodes with degree $d' > d$ for the analytical solution (see Eq. 4.5.96) and simulations (red circles).

the fact that we have only used pairwise connection probabilities in computing the topological quantities discussed above, which amounts to neglecting higher correlations.

The approximate connection probability [6] in Eq. 4.0.1 assumes that all the sequences of same length are equivalent in their string-matchings, i.e., they have an equal chance to be reproduced in longer strings of given length and to contain shorter strings of same length (see Section 2 for examples). This is a coarse-grained, or effective-medium approach where one ignores the precise content of the sequences, and assumes that they are maximally randomized. Then, besides the length of the alphabet from which the letters are chosen, the only relevant quantity characterizing a string is its length. Another simplification which goes into Eq. 4.0.1 can be thought of as a mean-field approximation, where all the consecutive overlapping subsequences of a given length have been treated as if they were all independent, which is obviously not true.

We have already pointed out that the joint probabilities of edges converging

upon, or going out from a given node do not factorize. Nevertheless, we have neglected the correlations between sequences reproduced in the same string, as well as between those sequences containing the same string, and used only pairwise connection probabilities in the foregoing discussion. For a better approximation to the clustering coefficient and the rich-club coefficient, one should also consider multi-point connection probabilities. This mean-field type of approximation is in fact similar to treating all successive subsequences of a given string as independent from each other, and is only valid if the key-sequences are much shorter than the lock-sequences. Therefore the quality of the agreement between the hidden-variable models (using Eq. 4.0.1) and content-based models are totally determined by the length distributions of the lock- and key-sequences (see Section 2 and Section 3 for examples).

# 5 THE RANDOM BOOLEAN DYNAMICS ON CONTENT-BASED NETWORKS

We have studied the properties of the random Boolean dynamics on small content-based networks obtained via generic string length distributions, whose topological features in the large system size limit have been discussed in Section 2.2.1. The aim of the research was establishing a starting point on modelling the dynamical properties of gene regulation within our information-theoretical approach.

We modify random Boolean dynamics within our content-based approach and outline our results. We have focused on the number and length distributions of attractors, the size distribution of the basins of attraction, the distribution of precursor numbers and transient times, as well as the propagation of information. The aim was classifying our networks with respect to their dynamical properties.

We start with an introduction on the random Boolean dynamics and summarize some earlier results. Then we introduce the content-based random Boolean dynamics we have proposed on the content-based networks and demonstrate the properties of the dynamical phase space via some examples. In this content-based version of random Boolean dynamics, beside the topological properties of the underlying network, the assignment of random Boolean functions are also different.

## 5.1 Random Boolean Networks: NK Models of Gene Regulation

Random Boolean networks, so called the $N$-$K$ models, were introduced by Stuart Kauffman [42] in the context of regulation of gene activations and fitness landscapes in 1969. The model has gained sufficient interest and found application in different fields ranging from biology, mostly in the context of gene expression and cell differentiation [71, 72, 73, 74], to physics in the study of chaos as well as the glassy and disordered materials [75, 76, 77], and social sciences. A huge literature is available on the topic as reviewed in Refs. [4, 78, 79, 80]. We here summarize

essential ingredients of the model and its variations with some earlier results.

In random Boolean networks consisting of $N$ nodes, each of the nodes corresponds to a random variable $\sigma_i$ which takes its value from the set $\{0, 1\}$ according to the values of other nodes (variables). The node $i$ is coupled with $k_i$ controlling elements (thus, its in-degree is $k_i$) according to a predetermined ensemble of wiring diagrams (adjacency matrices). These variables $(\sigma_{n_1(i)}, \sigma_{n_2(i)}, \ldots, \sigma_{n_{k_i}(i)})$ constitutes the inputs of the random Boolean function $F_i$ assigned to the variable $\sigma_i$. Again the choices of the random Boolean functions are made from an ensemble of a predetermined subset of Boolean functions. The value of each variable at time $t + 1$ is determined by the values of its controlling elements at time $t$,

$$\sigma_i(t+1) = F_i(\sigma_{n_1(i)}(t), \sigma_{n_2(i)}(t), \ldots, \sigma_{n_{k_i}(i)}(t)) \ , \quad i = 1, 2, \ldots, N \ . \tag{5.1.1}$$

One realization of the wiring diagram (for each node $i$, the number $k_i$ and then the set of the controlling elements which can be denoted by their indices $n_{1(i)}, n_{2(i)}, \ldots, n_{k_i(i)}$) and the assignments of the Boolean functions (for each node $i$, the random Boolean function $F_i$) is called one realization of the model. Given a realization of the model, the dynamics of the systems is totally determined by Eq. 5.1.1. We may represent the state (configuration) $\Sigma(t)$ of the system at time $t$ by a list of its variables,

$$\Sigma(t) = (\sigma_1(t), \sigma_2(t), \ldots, \sigma_N(t)) \ , \tag{5.1.2}$$

as well as by an integer number in base 10,

$$\Sigma(t) = \sum_{i=1}^{N} 2^{N-i} \sigma_i(t) \ , \tag{5.1.3}$$

which provides a sufficient ease in numerical calculations (simulations) at least for small values of system size $N$. In the context of gene regulation, the nodes of random Boolean network correspond to genes and random variables to the activation states (1 if gene is "on", and 0 if it is "off") of genes. The directed edge from node $i$ to node $j$ represents the regulatory interaction where the expression of $i$th gene controls the expression of the $j$th gene.

In his original model, Kauffman assigned the same number $K$ of controlling elements (inputs) to each variable, where the inputs were chosen randomly with uniform probability, $1/N$. Thus, the in-degree distribution is given by $\delta_{k,K}$, whereas

the out-degree distribution is a Poisson with the same mean $K$. Since then, other ensembles of network architectures have been introduced, such as the scale-free networks where the in-degree [81] or out-degree [82] distributions follow power-law form.

The argument of the Boolean function $F_i(\sigma_{n_1(i)}, \sigma_{n_2(i)}, \ldots, \sigma_{n_{k_i}(i)})$ can take $2^{k_i}$ values because each of its inputs is a binary variable. Thus, the total number of Boolean functions is $2^{2^{k_i}}$. An extensively used ensemble of Boolean functions assumes that all the functions have equal chance to appear. Alternatively, Boolean functions can be weighted by introducing a bias $p$, where the function $F_i$ takes the value 1 with the probability $p$ or 0 with the probability $1-p$ for each configuration of its inputs. So the first one is a special case of the latter scheme where $p = 1/2$. Another ensemble can be obtained by only considering canalizing functions in which the value of the function is determined when just one of its inputs is given a specific value. For example, the Boolean function AND is a canalizing function, where if one of its inputs is 0 then we now for sure that the value of the function is 0.

Once the network has been established and the Boolean functions have been assigned, one may update the states of all the nodes in the same time step according to Eq. 5.1.1. This is called the synchronous update. One may as well choose a set of variables randomly to update in each time step, which is called asynchronous update.

If the wiring diagram and the Boolean functions are fixed during one realization of the system, then it is said that the system is quenched. One could assume an annealed approach, where both the wiring diagram and the assignments of Boolean functions are changed in each time step. Or one could follow an intermediate approach to achieve predetermined task [7].

If the system is quenched, all one needs to do is iterating the state vector $\Sigma(0)$ of the system according to Eq. 5.1.1 to obtain one of its trajectories in the phase (state) space, which will eventually end up at a fixed point or in a cyclic orbit. Since the dynamics is totally deterministic, starting from all the initial config-

urations $\{\Sigma(0)\}$ of the variables one can fully explore the state space. Let us give some definitions used in the characterization of the phase space in the next sections.

### 5.1.1 Transfer of information in Kauffman networks

The $N$-$K$ model assumes a quenched system, where all the nodes have the same number $K$ of controlling elements; the random Boolean functions are weighted with the probability $p$; and the variables are updated synchronously. These networks have become prototypes of the dynamical systems exhibiting "ordered" and "chaotic" phases with a separator regime (so called, the "critical" phase) determined by the two parameters $K$ and $p$. These regimes have been defined with respect to the propagation of the differences in the information coded in the initial states. If the system remembers small changes in its initial configurations in long time limit and propagate them to a finite fraction of variables in the large networks size limit, then the system is said to be in the chaotical phase. If small differences in the initial configurations die out over time, then the system is said to be in the ordered (frozen) phase.

Let us consider a pair of randomly chosen initial configurations, $\Sigma(0) = (\sigma_1(0), \ldots, \sigma_N(0))$ and $\widetilde{\Sigma}(0) = (\widetilde{\sigma}_1(0), \ldots, \widetilde{\sigma}_N(0))$, and follow their trajectories (time evolution of these configurations) under the same dynamics (Eq. 5.1.1) to determine the Hamming distance $H(t)$ between them at time $t$,

$$H(t) = \sum_{i=1}^{N} \left(\sigma_i(t) - \widetilde{\sigma}_i(t)\right)^2 \quad . \tag{5.1.4}$$

If the system is in the frozen regime the small differences $H(0)$ in the initial states will not grow in time, whereas they will propagate over the entire system in the case of chaotic phase. Another quantity proposed to probe the same property is the normalized overlap $x(t)$ between configurations,

$$x(t) = 1 - \frac{H(t)}{N} \quad , \tag{5.1.5}$$

which is the fraction of nodes having the same state in the configurations $\Sigma(t)$ and $\widetilde{\Sigma}(t)$. Here one looks for the conditions (parameters) for which $x$ goes to unity in the long time limit. In the chaotic regime the Hamming distance increases

exponentially and then saturates over time, whereas in the ordered regime it decreases and again achieves an asymptotic value. On the other hand if the system is in the critical regime, the Hamming distance first decreases and then starts to increase to saturate eventually [78].

Now we will reproduce the result of Derrida and Pomeau [76] to give an idea how the phase diagram of such systems are obtained. We want to drive $x(t+1)$ as a function of $x(t)$. In the limit of large system size, $x(t)$ corresponds to the probability that a randomly chosen variable has the same value in the configurations $\Sigma(t)$ and $\widetilde{\Sigma}(t)$. The node will have the same value at time $t+1$, if the values of its inputs in $\Sigma(t)$ are equal to those in $\widetilde{\Sigma}(t)$ at time $t$, whose probability is $[x(t)]^K$. If at least one of its controlling elements is in a different state in these configurations (the probability of this event is $1-[x(t)]^K$), then the node will be in the same state if and only if the Boolean function assigned to this node takes the same values for these different arguments, where the probability is $p^2+(1-p)^2$. Then $x(t+1)$ the probability of finding a randomly chosen node with the same value in the configurations $\Sigma(t+1)$ and $\widetilde{\Sigma}(t+1)$ may be written as

$$\begin{aligned} x(t+1) &= [x(t)]^K + (1-[x(t)]^K)\,(p^2+(1-p)^2) \ , \\ &= 1-2p(1-p)(1-[x(t)]^K) = \mathcal{F}(x(t)) \ , \end{aligned} \tag{5.1.6}$$

where $\mathcal{F}(x(t)) = x(t+1)$ has been obtained as a monotone increasing function of $x(t)$. If the slope of the curve is smaller than 1 as we approach unity from below, then the system is said to be in the ordered regime, if larger, then in the chaotic regime where the fixed point at unity becomes unstable and another fixed point $x^* \neq 1$, such that $x^* = F(x^*)$, appears. The condition where the corresponding slope is 1 gives the critical regime. Then one gets,

$$\begin{aligned} K &> 1/2p(1-p) \ \text{ chaotic phase} \ , \\ K &= 1/2p(1-p) \ \text{ critical phase} \ , \\ K &< 1/2p(1-p) \ \text{ ordered phase} \ . \end{aligned} \tag{5.1.7}$$

For $N$-$K$ models the system has very narrow intervals of its parameters to exhibit an ordered or critical regime. For large values $K$ of connectivity the bias of the system has to be adjusted to a very high or low value ($p \approx 1$ or $p \approx 0$, thus

almost constant Boolean functions) as can be followed from Eq. 5.1.7. Recognize that the phase diagram is symmetric around $p = 1/2$. It has been shown [81] by generalizing the same mean-field approach that the scale-free networks, with power-law in-degree distributions of exponent $\gamma$, display ordered behavior for $\gamma > 2.5$ and exhibit chaotic phase for $\gamma < 2$ without respecting the precise value of $p$.

This approximate calculation based on a mean-field approach can only be meaningful if the system size is very large such that the whole ensemble of the wiring diagrams and, more importantly, all the Boolean functions are realized [78]. For finite systems, even in the ordered regime, small Hamming distances will always persist. This is what we will be discussing in Section 5.3.2.

## 5.2 Content-Based Random Boolean Dynamics: CB Models of Gene Regulation

The random Boolean dynamics we employ here is totally deterministic once the Boolean functions are assigned, and the systems is quenched, neither the wiring diagram (the adjacency matrix) nor the Boolean functions are changed during one realization. The updating of the activation states of the nodes is synchronous, the states of the nodes at time $t + 1$ are updated all together according to their Boolean functions and the states of the nodes at time $t$. It is also worth noting that the dynamics is discrete, the states of the nodes as well as the time steps we operate the dynamical process are discrete.

In our content-based network, as discussed in detail in the previous sections (see Section 2.2 and Section 3), the nodes represent genes with two regions specifying their promoter and coding regions. Each promoter region is associated with a random linear code through which the expression level of the corresponding gene is regulated. If a gene codes a transcription factor we also associate a second random sequence representing the binding motifs (regulatory sequences) recognized by the TF, through which the corresponding gene regulates the expression level of other nodes. (See Fig. 5.1.)

Uninterrupted subsequences of length $k'$ in a promoter region of length $k$ consti-

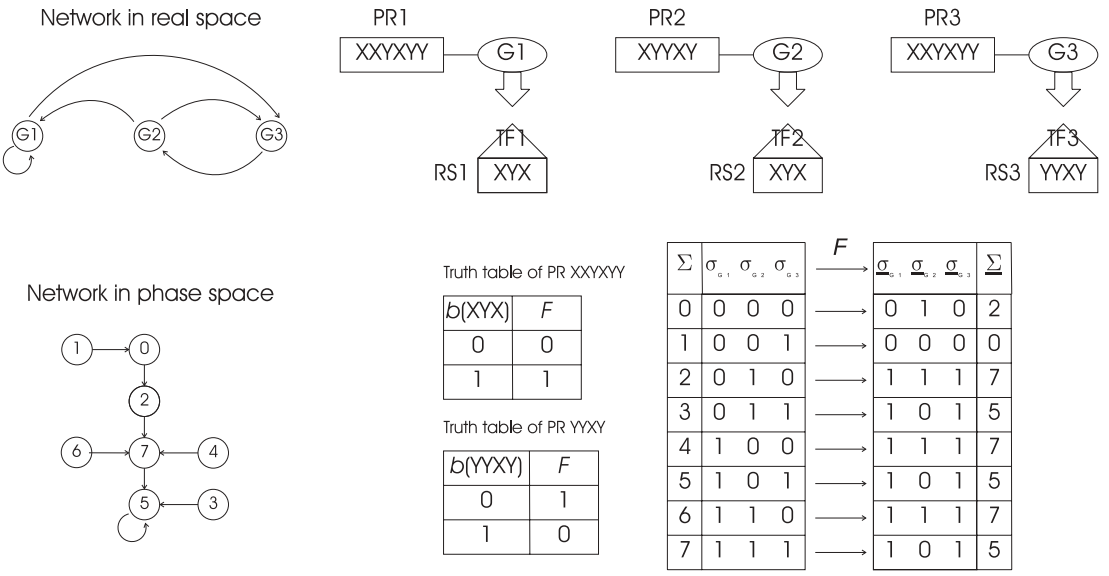


Figure 5.1: Demonstration of our content-based random Boolean dynamics on a sample network. In the upper panel, we display a network of 3 genes (G1, G2, G3) with their associated PR and TF strings. The random sequences associated with the nodes have been chosen from the alphabet {X,Y}. Two of the nodes (G1 and G3) share the same promoter region (PR1 and PR3 are identical), and the TFs coded by two nodes (G1 and G2) recognize the same recognition site (RS1 and RS2 are identical). In the lower panel, we display the Boolean functions associated with the PRs whose truth tables are constructed with respect to the binding states of their recognition sites. In the tables, not the binding states of all the recognition sites are displayed because there are no TFs recognizing them (their binding states are identically zero). The phase space may be considered as a directed graph where each node represents a configuration and edge the evolution of a configuration under a dynamical step. Given an initial configuration $\Sigma = \sum_{i=1}^{N} 2^{N-i}\sigma_{\mathrm{G}_i}$, representation as a number in base 10, we determine the state $\underline{\Sigma}$ of the system under a successive step of the dynamics according to the truth tables. The sample system has 1 fixed point at configuration $\Sigma = 5$, where G1 and G3 are on (TF1 and TF3 are produced), whereas G2 is off.

tutes its possible "recognition sites" for the regulatory sequences of this length $l = k'$. Let us denote the position of the first letter of a subsequence of length $k'$ in a promoter region of length $k$ by $\nu(k', k)$. The number of "relevant" recognition sites in a promoter region depends on the length distribution of TF binding motifs $p_{\mathrm{RS}}(l)$ as well as that of promoter regions $p_{\mathrm{PR}}(k)$, defined within the intervals $\Lambda_{\mathrm{RS}} = [l_{\min}, l_{\max}]$ and $\Lambda_{\mathrm{PR}} = [k_{\min}, k_{\max}]$. Because, for example, the recognition sites of lengths $k' < l_{\min}$ are not recognized by the TFs and do not have any effect by themselves on the regulation of the corresponding gene according to our sequence-matching rule. Thus we may specify the set of relevant recognition sites by $1 \leq \nu(k', k) \leq k - k' + 1$ for the possible values of $k'$ where only the $l_{\min} \leq k' \leq \min(l_{\max}, k)$ are relevant. If a regulatory sequence is identical to the subsequence (recognition site) specified by $\nu = \nu(k', k)$, then we assume that the promoter region is recognized (and bound) by the corresponding TF at this site $\nu$.

Let us reformulate the elements of the adjacency matrix by defining $w_{ij}^{\nu}$ as the characteristic function of the event that an exact match occurs between the regulatory sequence of length $l_i$ associated with the $i$th node and the promoter region of length $k_j$ associated with the $j$th node at the recognition site $\nu = \nu(l_i, k_j)$. Then the element of the adjacency matrix $w_{ij}$ is given by $w_{ij} = 1 - \prod_{\nu=1}^{k_j - l_i + 1}(1 - w_{ij}^{\nu})$ which takes the value 1 if and only if there exists at least one recognition site $\nu'$ in the promoter region of the $j$th node identical to the regulatory sequence recognized by the TF coded by the $i$th node ($w_{ij}^{\nu'} = 1$), 0 otherwise.

We define the random Boolean dynamics on our content-based network by assigning a value $\sigma_i(0)$, from the set $\{1, 0\}$ at time zero to each node $i$, indicating the activation state (on or off, respectively) of the corresponding gene at that time, then following the trajectory of the states evolving under random Boolean functions at successive time steps. The random Boolean functions (RBFs) are assigned to the promoter regions of the genes with the inputs being the "binding states" of their recognition sites. As we have already stated above, we assume that if a recognition site is identical to a regulatory sequence, then the corresponding transcription factor recognizes and binds the corresponding promoter region at this site. Obviously this event is realized only if the TF is available at that time

(the gene coding the TF has to be active; we assume that the TFs do not survive from one time step to another). Thus, we may define the binding state $b_j^\nu(t)$ of the recognition site specified by $\nu = \nu(k', k_j)$ in the promoter region associated with the $j$th node, by $b_j^\nu(t) = 1 - \prod_i (1 - w_{ij}^\nu \sigma_i(t))$. Then, it follows that $b_j^\nu(t)$ takes the value 1 if and only if there exists at least one active gene coding the TF whose regulatory sequence is identical to the recognition site ($w_{i'j}^\nu = 1$ and $\sigma_{i'}(t) = 1$), 0 otherwise. The random Boolean function $F_j(\{b_j^\nu(t)\})$, associated with the promoter region of the $j$th node determines the activation state of the corresponding gene at the next time, $t + 1$, via $\sigma_j(t + 1) = F_j(\{b_j^\nu(t)\})$ where $\{b_j^\nu(t)\}$ denotes the list of binding states of the relevant recognition sites in the promoter region. (See Fig. 5.1.)

The truth tables for the Boolean functions associated with the promoter regions are constructed randomly. For each different realization of the list $\{b^\nu(t)\}$, $F$ is assigned the value 1 with probability $p$ and 0 with probability $1 - p$. Once the Boolean functions are assigned then they are fixed once and for all.

One should note here that since the random Boolean functions are associated with the promoter regions and take values with respect to the binding states of their recognition sites, if some of the genes have identical promoter regions they are operated by the same Boolean function, i.e., their expression profiles are also identical (they are expressed or depressed together). Another difference from the $N$-$K$ models coming with our modification is that if some of the transcription factors coded by different genes recognize the same binding motif (so the regulatory sequences associated with these nodes are identical), then the number of inputs of the Boolean functions associated with the promoter regions containing the recognition sites recognized by these TFs reduces. Because the promoter regions do not care about the identities of the nodes but the proteins coded by them. Thus, the activation states of those genes whose TFs recognizing the same binding motif are degenerate in the sense that the binding state of such a recognition site is 0 if non of these nodes are active and 1 if at least one of them is active. (See Fig. 5.1.)

For systems having a finite number $N$ of such nodes with finite number of possible

states (in this case, two), the volume $\Omega$ of the phase space constituted by all the possible initial states $\{\Sigma(0) = (\sigma_1(0), \sigma_2(0), \ldots, \sigma_N(0))\}$ of the nodes is also finite, $\Omega = 2^N$. If the system starts from an initial configuration $\Sigma(0)$, it will eventually fall into a fixed point or a cyclic orbit (will start revisiting some already visited states). The flow diagram (time evolutions of configurations) of the phase space can be thought as a directed network, where the vertices correspond to the initial configurations of the node (gene) activations and each link to one time step in which we employ the dynamics. Thus, in this network, a directed link from vertex $\Sigma(0)$ to vertex $\widetilde{\Sigma}(0)$ appears if $\Sigma(1) = \widetilde{\Sigma}(0)$. It is sufficient to analyze the "adjacency matrix" of the phase space to study the properties of the configuration space. For example, in this network the out-degree of each vertex is one, and the in-degree is called the *precursor number* of the vertex. Each cluster in the network correspond to a *basin of attraction*, and the number of vertices in the cluster to the *basin size*. Each loop in the network is called an *attractor*, and the number of vertices in the loop as the *length of attractor*. Each path between any pair of vertices is unique. The path length (number of directed links) starting from a vertex (initial configuration) and ending in the attractor is called the *transient time* of the initial configuration. In Fig. 5.2 we display an example for a system of size 8 (thus, the size of the phase space is 256), where the phase space is partitioned into 10 basins of attraction with attractors of lengths 1 and 2.

The properties of the phase space are determined by the Boolean functions as well as the topology of the underlying network (do not confuse with the flow diagrams of the phase space). It is of great interest to determine the effect of the topological properties on the asymptotic behavior of the dynamics. This is what we try to demonstrate in the subsequent sections.

### 5.3 Simulations on Small Content-Based Networks

We have simulated the random Boolean dynamics defined above on small content-based networks which have been constructed with identical toy length distributions for the PRs and RSs, of either truncated exponential or Gaussian form. The choice of an exponential length distribution was motivated by the fact that it lends itself to analytical treatment [6]. We have adjusted the parameters of the

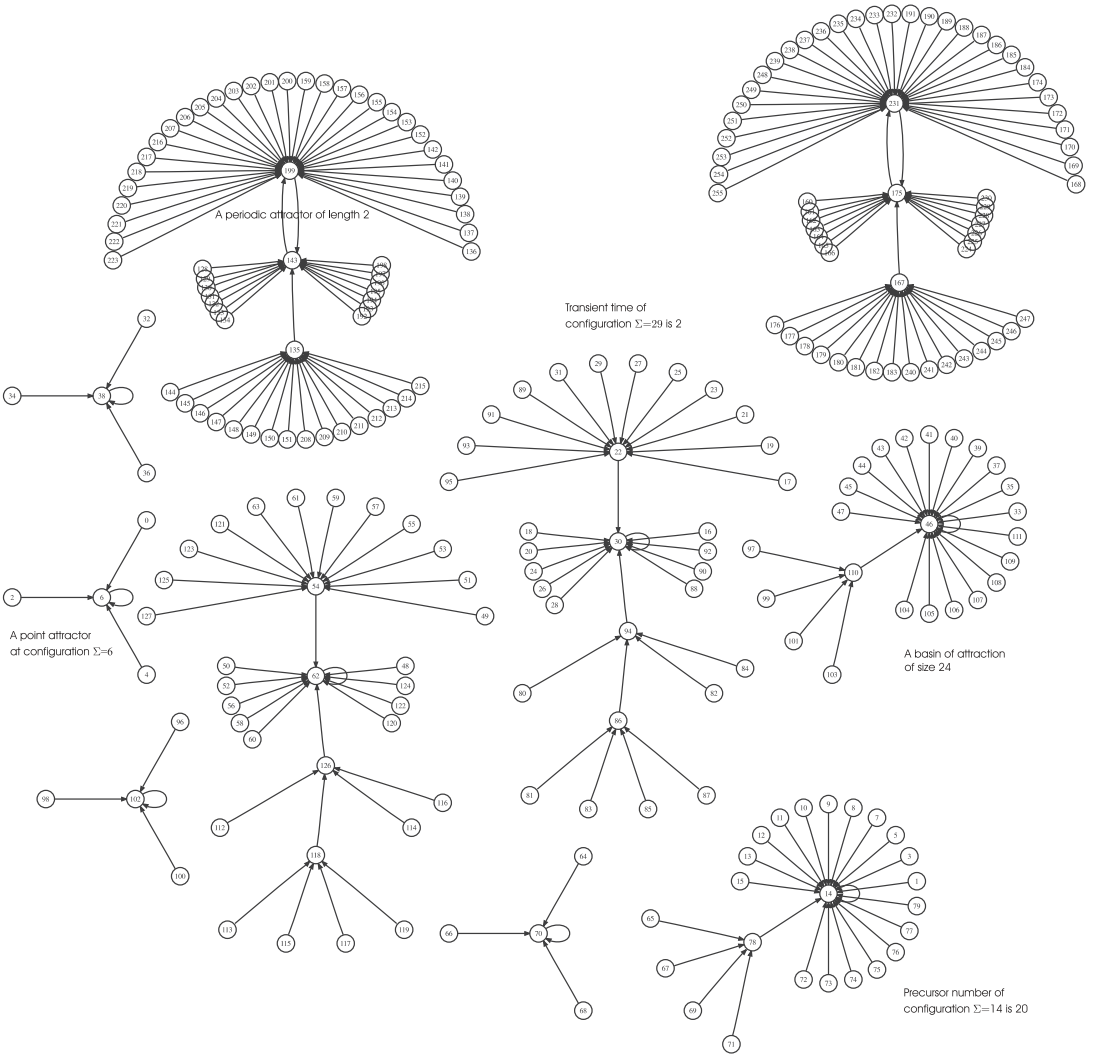


Figure 5.2: The flow diagram of phase space for one realization of random Boolean functions has been represented by a directed network of size $\Omega = 2^8$ for one realization of a content-based network of size $N = 8$. The vertices corresponding to initial configurations of activation states of the nodes have been denoted by numbers in base 10. For example, $\Sigma(0) = 128$ denotes the configuration where only the first node is on, etc. The directed link from $\Sigma$ to $\widetilde{\Sigma}$ represents the dynamical evolution of $\Sigma$ in one time step. The sample system consists of 10 basins of attraction, four of which of size 4, two of which of size 24, two of which of size 32 and two of which of size 64. There are 8 point attractors and 2 periodic attractors of length 2. (Network has been drawn by the Graphviz visualization tool freely available at http://www.graphviz.org.)

Gaussian distribution so as to give networks that are not very sparse. Moreover, we have assumed that all the genes code TFs. The reason was again, that we did not want to obtain very sparse graphs. We should note here that it makes sense to operate the dynamics just for the subset of the nodes which are TF-coding genes, because the others coding structural proteins do not have any effect on the regulation of the genes, so on the asymptotic behavior of dynamics. Because of the forms and the intervals of the length distributions used here we do not have any claim on the modelling of real regulatory networks. But by themselves they are very interesting to work on because the topological properties of the ensemble of networks obtained in this way share similar characteristics with those of real complex networks as shown in Section 2.2.1.

We have generated an ensemble of content-based networks of size $N$, $5 \leq N \leq 16$ with given length distributions. For each realization, $10^4$ in total, of the model network we have realized the assignments of the random Boolean functions once with $p = 0.5$ (no bias). The lengths $l$ of the random binary sequences associated with nodes have been confined to the interval $1 \leq l \leq 25$ both for the exponential ($p(l) \propto q^{-l}$ with $q = 0.9$) and Gaussian ($p(l) \propto \exp[(l - \langle l \rangle)^2/2\sigma^2]$ with $\langle l \rangle = 13$ and $\sigma^2 = 50$) distributions. We have explored the phase space fully, which is only possible in practice for systems of small sizes (roughly up to $N = 20$), by starting from all the initial configurations of the gene activities.

### 5.3.1 Properties of phase space

We have considered here the scaling relations of the average number and length of attractors as well as the average transient time of configurations and basin size of attraction with system size (Fig. 5.3). Moreover we have determined the number and length distribution of attractors, the basin size distribution of attraction, the distribution of precursor numbers and transient times, as shown in Figs. (5.4–5.9). We display our results for the largest system size $N = 16$ considered here if not stated otherwise.

The mean values of $n_{\mathrm{a}}$ and $l_{\mathrm{a}}$, the number and length of the attractors, signifies the stability and versatility of the system. We find (see Figs. (5.3a, 5.3b, 5.3d))

that $\langle n_a \rangle$, $\langle l_a \rangle$ as well as the average transient time $\langle \tau \rangle$ increase linearly with system size $N$, for both the Gaussian and exponential string length distributions. However in all cases, the exponential has higher growth rates with $N$ for the above quantities, whereas the Gaussian length distribution gives more stable results. On the other hand, the average basin size $\langle s \rangle$ increases exponentially with system size (see Fig. 5.3c), in each case. But now, the growth rate obtained for the Gaussian length distribution of sequences is higher than that found for the exponential one, as one would expect from the observation for the average number of attractors (recognize here that $\langle s \rangle = \Omega / n_a$ for a given realization). It was believed for long time that the average number of attractors of the Kauffman networks grew as $N^{1/2}$ [78]. But recent numerical studies [83] have shown that $\langle n_a \rangle$ increases linearly with system size. On random scale-free networks (of sizes up to 20), Aldana [81] has shown numerically that $\langle n_a \rangle$, $\langle l_a \rangle$ and $\langle \tau \rangle$ grows linearly in the ordered and critical regimes, whereas in the chaotic regime of the dynamics $\langle l_a \rangle$ and $\langle \tau \rangle$ increases exponentially with system size.

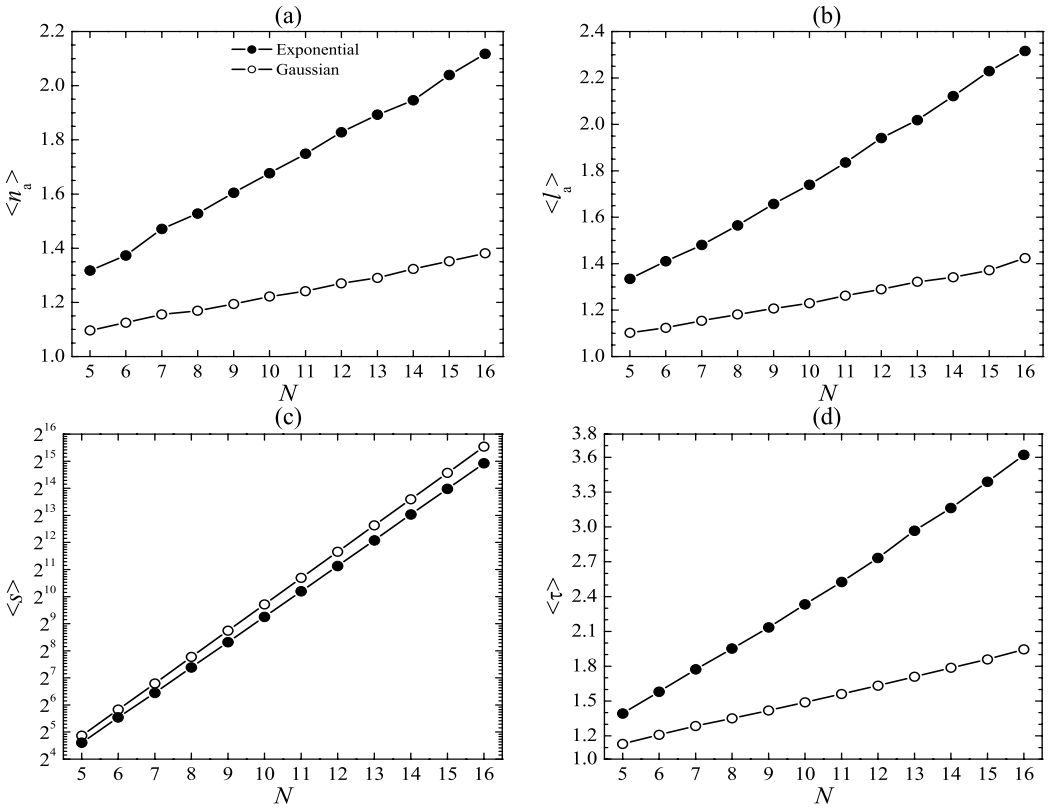


Figure 5.3: Average values of the numbers (a), lengths (b), basin sizes (c) and transient times (d) with respect to system size $N$ ($5 \leq N \leq 16$) for the exponential (•) and Gaussian (∘) string length distributions.

In Figs. (5.4, 5.5) we display the probability of finding $n_{\rm a}$ attractors in a realization of the system, and the probability of finding an attractor of length $l_{\rm a}$. Although the number of attractors is very small compared to the size of the phase space, the exponential length distribution gives rise to a fatter tailed distribution than the Gaussian. The distributions of attractor lengths behaves in a similar way as before, but with much broader intervals.

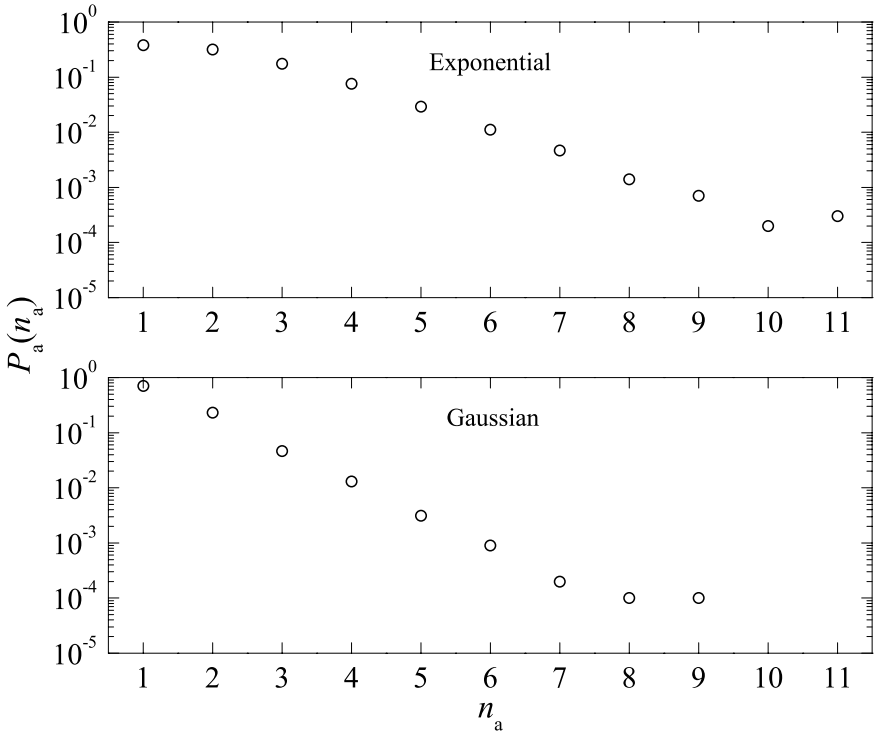


Figure 5.4: Distributions of number of attractors have been plotted in semi-logarithmic scales for better visibility of data points for the exponential (above) and Gaussian (below) length distributions. The distribution has a broader tail (goes up to 11) for the exponential length distribution comparing with the Gaussian case which goes up to 9. Although the difference in the maximum numbers is very small, one should note the difference in the frequencies.

In Figs. 5.6 we exhibit the probability of finding an attractor of basin size $s$. We find that the configuration space is more frequently partitioned into basins of attraction of sizes $s = \Omega/2^n$, where $n \geq 0$ is an integer number as found [81] for the scale-free networks. These numbers are harmonics of the size of the phase space $\Omega$. It is also very interesting to note that the nonzero frequencies show up at even basin sizes. The finer structures starting to occur between these harmonics correspond to the increasing complexity of the phase space (basins of attraction in any scale).

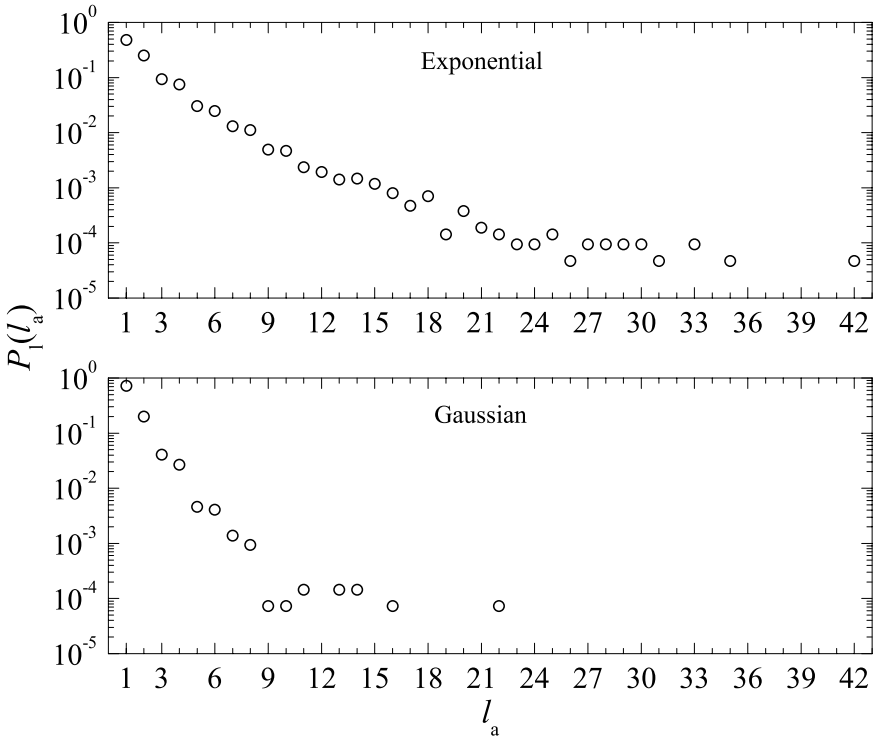


Figure 5.5: Distributions of attractor lengths plotted in semi-logarithmic scales have been calculated for ensembles of the model either with the exponential (above) or the Gaussian (below) distribution of string-lengths. The distribution is fat tailed for the exponential length distribution, going up to 42, comparing with the Gaussian case (up to 22). Again note the difference in the frequencies.

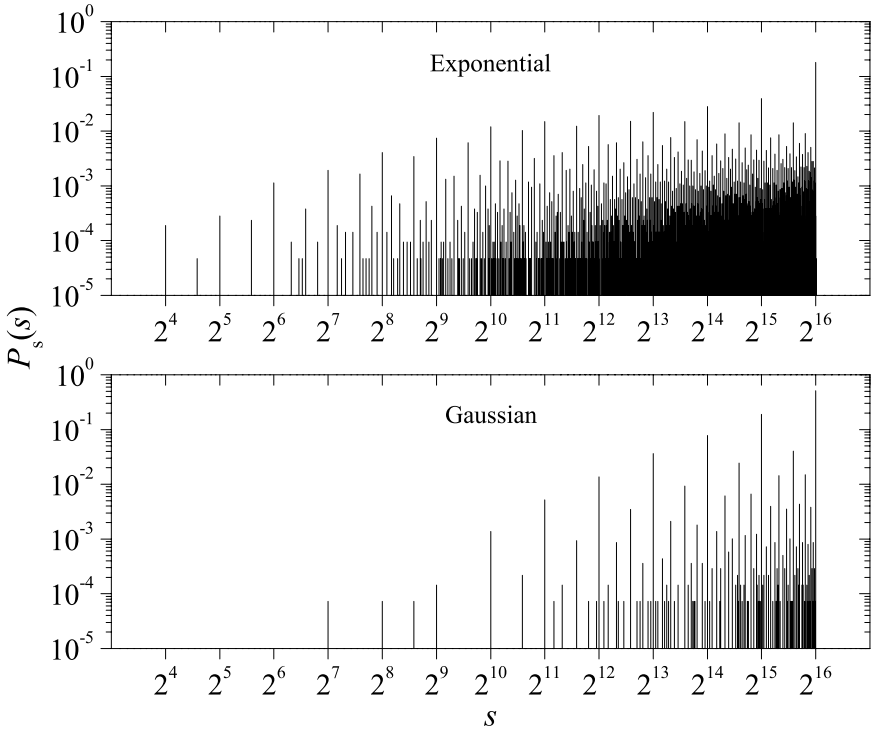


Figure 5.6: Size distributions of basins of attraction have been plotted in the log 2-log 10 scales for better visibility, for the exponential (above) and Gaussian (below) length distribution. The distributions have expressed picks at basin sizes $s = \Omega/2^n = 2^{N-n}$ ($n = 0, 1, 2, \ldots$). It is interesting to note that the nonzero frequencies occur at even basin sizes. The fine structures much more dominated in the upper panel signify the increasing complexity of the phase space, that there are many attractors with basin sizes of any value.

We display in Fig. 5.7, the probability of finding a configuration with $n_{\rm p}$ precursors. The probabilities of finding configurations with zero precursors have been suppressed for clarity, which we exhibit in Fig. 5.8 for different system sizes. Again one should note here that the relatively high frequencies are realized at the harmonics of the size of the phase space, and the nonzero probabilities appear at even precursor numbers. The increasing probabilities in the small precursor number region of the distribution observed for the model ensemble with exponential length distribution may be related with the increasing complexity of the dynamics [81], accordingly the transient times increases.

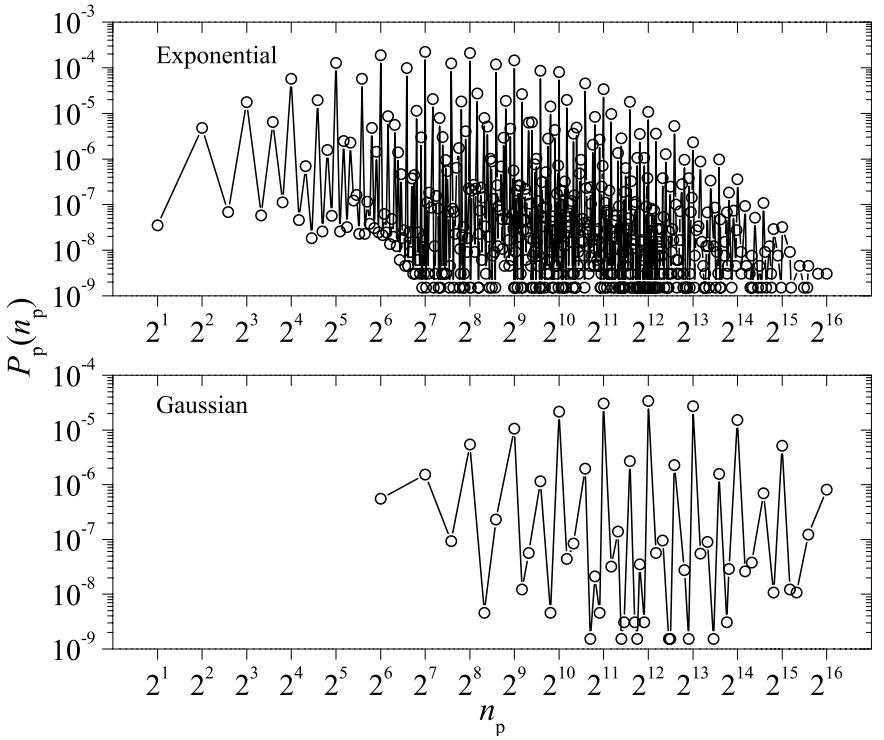


Figure 5.7: Distributions of number of configurations with $n_{\rm p}$ precursors have been plotted in the log 2-log 10 scales for better visibility, for the exponential (above) and Gaussian (below) length distributions. The data points at $n_{\rm p} = 0$ have been suppressed for clarity in both panels. The relatively high frequencies again show up at $n_{\rm p} = 2^{N-n}$ with integer values of $n$, and the nonzero data points only have been observed at even precursor numbers. Again the points starting to occur toward the left side of the spectrum in the upper panel may show that the phase space of the system is getting "chaotic".

In Fig. 5.9, we display the probability of finding a configuration with transient time $\tau$. The exponential length distribution gives rise to broader distribution comparing with the Gaussian case.

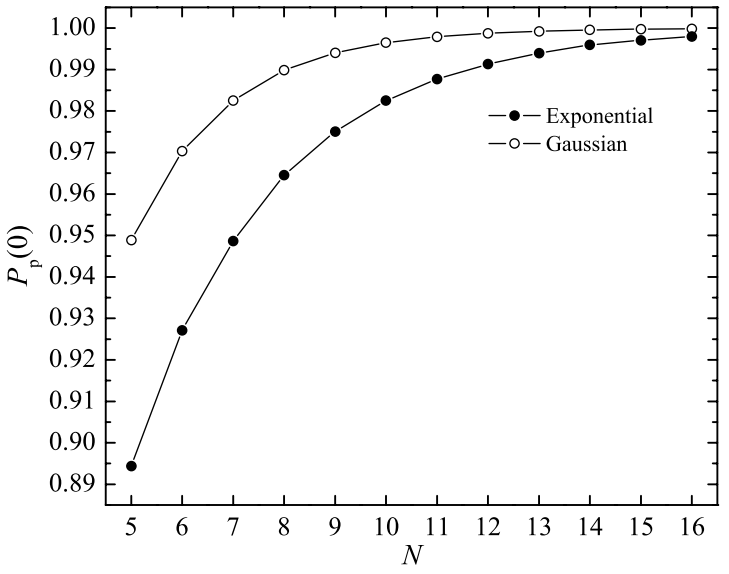


Figure 5.8: Probabilities of finding configurations with zero precursors for different system sizes, $5 \leq N \leq 16$ with the length distributions of exponential or Gaussian form. Such configurations may be thought as constituting the boundary of the phase space.

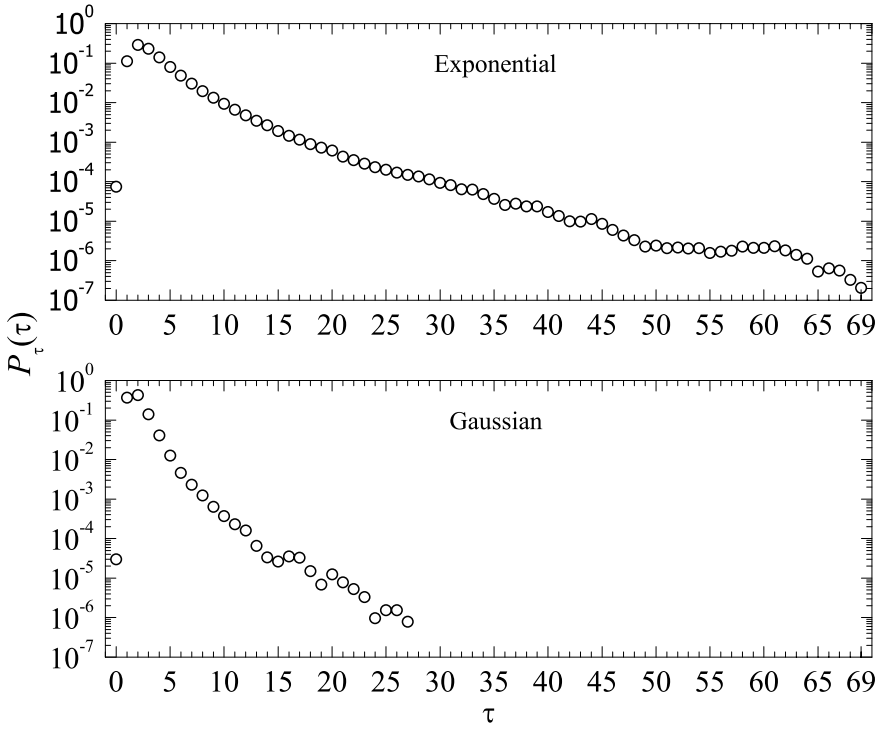


Figure 5.9: Distributions of number of configurations with transient time $\tau$ are plotted in semi-logarithmic scales. The distribution has much longer tail (up to 69) obtained by the exponential length distribution (above) than the one (up to 27) obtained by the Gaussian length distribution (below).

### 5.3.2 Stability and description of attractors

The stability of random Boolean networks as discussed briefly in Section 5.1.1, is defined with respect to the propagation of small differences in initial configurations over time. With the aim of determining the dynamical regimes of the systems under consideration, we have determined the evolution of the overlap $x(t)$ in one time step, for each possible values of $x(t)$, to obtain the curve of $x(t+1) = \mathcal{F}(x(t))$. The number $\Gamma_{H(t)}$ of those pairs, $\Sigma(t)$, $\widetilde{\Sigma}(t)$, of configurations, such that $\Sigma(t)$ differs from $\widetilde{\Sigma}(t)$ at its $H(t)$ variables at time $t$, is given by

$$\Gamma_{H(t)} = \Omega \times \binom{N}{H(t)} \ , \ \ 1 \leq H(t) \leq N \ . \tag{5.3.8}$$

Thus, one has to consider all such pairs of configurations to calculate the average overlap $x(t+1)$ at the next time $t+1$ for a given value of $x(t) = 1 - N^{-1}H(t)$ at time $t$,

$$x(t+1) = 1 - \frac{1}{N\Gamma_{H(t)}} \sum_{(\Sigma(t),\, \widetilde{\Sigma}(t)) \in\, \Gamma_{H(t)}} \ \sum_{i=1}^{N} \left(\sigma_i(t+1) - \widetilde{\sigma}_i(t+1)\right)^2 \ . \tag{5.3.9}$$

In Fig. 5.10 we display $x(t+1) = \mathcal{F}(x(t))$ for different system sizes $5 \leq N \leq 13$. We find that for the Gaussian length distribution the curves stay above the diagonal for all the system sizes considered here, whereas the curve obtained for the ensemble with the exponential length distribution starts crossing the diagonal as $N \geq 10$. So in this latter case, the fixed point at $x^* = 1$ becomes unstable and another stable fixed point starts to appear at $x^* < 1$. According to the definition of the stability in the Kauffman networks this would correspond to a chaotic phase.

To see to what extent this claim is true we have calculated the average overlap $x(t+T)$ in successive time steps,

$$x(t+T) = 1 - \frac{1}{N\Gamma_{H(t)}} \sum_{(\Sigma(t),\, \widetilde{\Sigma}(t)) \in\, \Gamma_{H(t)}} \ \sum_{i=1}^{N} \left(\sigma_i(t+T) - \widetilde{\sigma}_i(t+T)\right)^2 \ , \tag{5.3.10}$$

where again, $\Sigma(t)$ differs from $\widetilde{\Sigma}(t)$ at its $H(t)$ variables at time $t$. The average is performed over $10^4$ realizations of the network, and all such pairs of configurations. Following the trajectory of this quantity under successive steps of the dynamics, we find, for each value of $N$ considered, that it converges to a set of

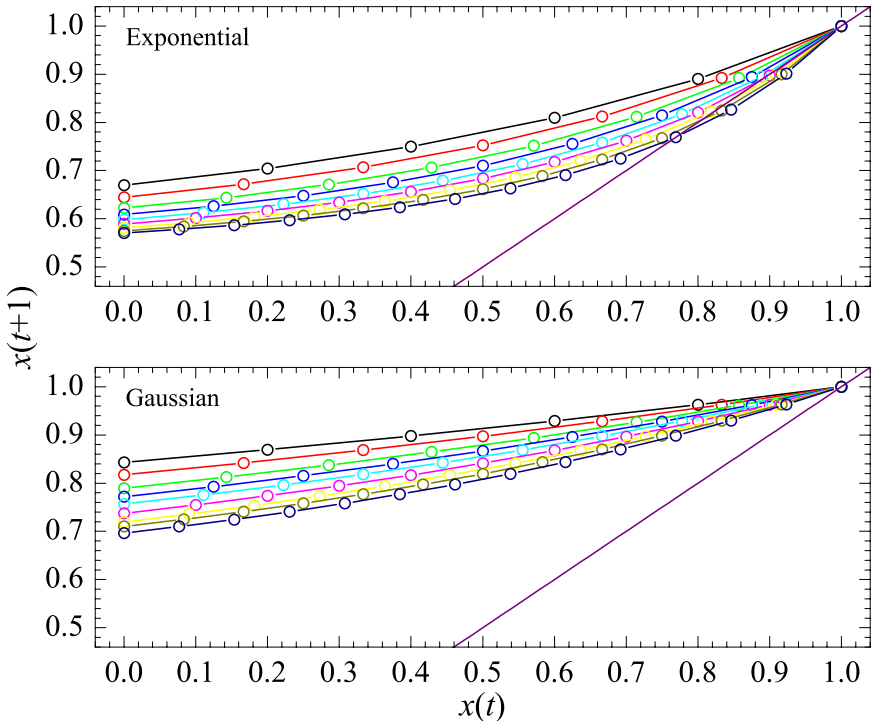


Figure 5.10: The evolution of the overlap function in one time step for different network sizes $5 \leq N \leq 13$ (from back to navy, respectively), for the exponential (above) and Gaussian (below) length distribution. Averages have been taken over $10^4$ network realizations and all pairs of configurations having initial overlap $x(t)$. The solid lines in both panels indicates the diagonal. The overlap function in the above panel stats to cross the diagonal as $N \geq 10$, that the fixed point at $x = 1$ becomes unstable and the system is called as in the "chaotic" regime.

points in a rather small but finite interval lying below unity, as shown in Fig. 5.11, which becomes shifted to smaller values for larger $N$ in the case of exponential string length distribution, as both $\langle n_{\rm a} \rangle$ and $\langle l_{\rm a} \rangle$ grow. Even for small $N$, $n_{\rm a} > 1$ with small but finite probability, and the phase points to which trajectories originating in different basins of attraction converge are separated by finite distances which are at least 1. In the presence of periodic orbits of lengths $(l_{\rm a})_i > 1$, one obtains a set of $\binom{L}{2}$ such pairs of phase points, where $L = \sum_{i=1}^{n_{\rm a}} (l_{\rm a})_i$. Thus, the persistence of distances between randomly chosen points in phase space does not automatically signal "chaotic" behavior, but the existence of multiple and/or periodic attractors.

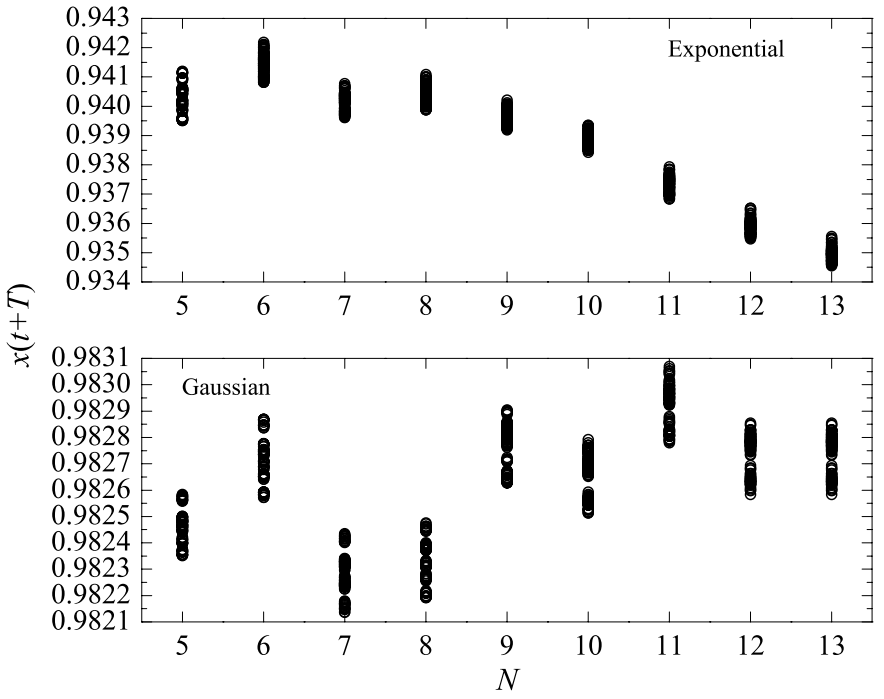


Figure 5.11: The long-time trajectories (superposition of points with $T \gg \tau$, where the trajectories have already got fixed) of the overlap function for different network sizes $5 \leq N \leq 13$, for the exponential (above) and Gaussian (below) length distribution. Averages have been taken over $10^4$ network realizations and all pairs of configurations having initial overlap $x(t) = 1 - 1/N$.

# 6 CONCLUSION

We have shown that the content-based models even with generic distributions of the lengths of random sequences associated with the nodes of the content-based networks result in network architectures which exhibit topological properties similar to real-world networks. Among these features we may list the degree distributions, the out-degrees being distributed in a very broad interval almost covering the size of the network, whereas the in-degrees being confined within a much narrower range, as observed for genetic regulatory networks. They also exhibit the small-word effect, even being of the smallest-world type, which might be important for these systems, particularly in reacting to stimuli and for the dynamical processes taking place. Another important property is the robustness of the network structure against random removal of nodes, needed and exhibited by most of the complex systems. We have also shown that the most important inputs to model networks are the length distributions of sequences which determine the overall topology, but this also provides us with a direct control on the degree of complexity of the system.

We have modelled the TRN of yeast within our content-based approach, with the distribution of the amount of shared information coded in the binding sites, recognized by the TFs of the organism under consideration, being the most important biological input to our model. We have made a very detailed comparison of the topological properties of the TRNs of yeast with those of the content-based model networks. The content-based model is able to reproduce all the topological features of the yeast TRN. The close structural similarity between the model and real networks may guide us to claim that they are members of the same statistical ensemble of networks. We may also claim that since the content-based networks, whose nodes are in association with random codes having appropriate length distributions, capture the properties of these regulatory networks, they could have arisen spontaneously and did not have to be engineered for the specific regulatory

functions they perform.

The model provides us with an understanding of the origin of the topological properties, such as the disassortative nature, i.e., the nodes with high degrees are preferentially connected with the nodes with low degrees. By adapting null-null models, we have also determined essential ingredients needed to model such networks. For example, we have observed that the existence of two node-types (those coding TFs, and others which do not) is important, and even with a modified Erdös-Rényi model one may capture some properties of topological coefficients. We have also shown that when the lengths of the promoter regions are fixed at the same relatively large length rather than being distributed with a broad tail, then the resulting model networks start to differ from the yeast TRN and are not able to capture all of its properties, if they do some of them. Via several randomization procedures, we have tried to identify the topological constraints of both the model and the real network. In both cases, we have seen that the total degree distribution is not a determinant of the overall network structure, whereas we have shown that when the out- and in-degree distributions are conserved, the topological features stay essentially invariant.

We have also introduced, as a null-null model, the hidden-variable version of our content-based networks where only pairwise connectivities are considered and further correlations are neglected. Comparison of the topological properties of the hidden-variable networks with those of the content-based model and yeast networks has revealed that these networks share very close structures. This result motivates that the analytical calculations based on this model can be meaningful and may provide us with further predictions on the features of TRNs. We have calculated the degree distributions and topological coefficients analytically and have shown that our theoretical results are in very good agreement with the simulations of the hidden-variable model.

Because of the highly distinct and seemingly advantageous architectures that the content-based networks have, we have adapted random Boolean dynamics to our content-based approach. Our results on small model networks with generic length distributions are promising but need much further investigation. Studies

on large networks, such as the content-based model of the TRN of yeast, may help us establish the connection between the underlying network structure of genetic regulation and its effects on the functioning of the system.

## REFERENCES


[1] **Fontana, W. and Ballati, S.**, 1999. Complexity, *Complexity*, **4**, 14-16.

[2] **Mitchell, M.**, 2006. Complex systems: Network thinking, *Artificial Intelligence*, **170**, 1194-1212.

[3] **Dawkins, R.**, 1986. The Blind Watchmaker, W.W. Norton and Company, New York.

[4] **Kauffman, S.A.**, 1993. The origins of order: Self-organization and selection in evolution, Oxford University Press, Oxford.

[5] **Balcan, D. and Erzan, A.**, 2004. Random model for RNA interference yields scale free network, *Eur. Phys. J. B*, **38**, 253-260.

[6] **Mungan, M., Kabakçıoğlu, A., Balcan, D. and Erzan, A.**, 2005. Analytical solution of a stochastic content-based network model, *J. Phys. A*, **38**, 9599-9620.

[7] **Calcott, B., Balcan, D. and Hohenlohe, P.**, 2005. Modeling the evolution of development, in *Complex Systems Summer School Final Project Papers*, Santa Fe Institute, Santa Fe, NM.

[8] **Bilge, A.H., Erzan, A. and Balcan, D.**, 2004. The shift-match number and string matching probabilities for binary sequences, q-bio.GN/0409023.

[9] **Balcan, D., Kabakçıoğlu, A., Mungan, M. and Erzan, A.**, 2006. A content-based approach to modeling the topological properties of the transcriptional regulation network of yeast, Submitted; q-bio.MN/0605045.

[10] **Balcan, D. and Erzan, A.**, 2007. Content-based networks: a pedagogical overview, Submitted.

[11] **Balcan, D. and Erzan, A.**, 2006. Dynamics of content-based networks, in *International Conference on Computational Science*, LNCS, **3993**, pp. 1083-1090, Eds. Alexandrov, V.N. et al., Springer-Verlag, Berlin.

[12] **Erzan, A. and Balcan, D.**, 2006. Content-based networks - From topology to dynamics, *Nonlinear Phenomena in Complex Systems*, To appear.

[13] **Amaral, L.A.N. and Ottino, J.M.**, 2004. Complex networks. Augmenting in the framework for the study of complex systes, *Eur. Phys. J. B*, **38**, 147-162.

[14] **Albert, R. and Barabasi, A-L.**, 2001. Statistical mehanics of complex networks, *Rev. Mod. Phys.*, **74**, 47-97; cond-mat/0106096.

[15] **Dorogovstsev, S.N. and Mendes, J.F.F.**, 2002. Evolution of networks, *Adv. Phys.*, **51**, 1079-1187; cond-mat/0106144.

[16] **Newman, M.E.J.**, 2003. The structure and funtion of complex networks, *SIAM Review*, **45**, 167-256; cond-mat/0303516.

[17] **Pastor-Satorras, R. and Vespignani, A.**, 2004. Evolution and Structure of the Internet: A Statistical Physics Approach, Cambridge University Press, Cambridge.

[18] **Erdös, P. and Rényi, A.**, 1960. On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 17-60.

[19] **Şengün, Y. and Erzan, A.**, 2006. Content-based network model with duplication and divergence, *Physica A*, **365**, 446-462.

[20] **Maslov, S. and Sneppen, K.**, 2002. Specificity and stability in topology of protein networks, *Science*, **296**, 910-913.

[21] **Pastor-Satorras, R., Vázquez, A., and Vespignani, A.**, 2001. Dynamical and correlation properties of the Internet, *Phys. Rev. Lett.*, **87**, 258701(4).

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

[23] **Dorogovtsev, S.N., Goltsev, A.V. and Mendes, J.F.F.**, 2002. Pseudofractal scale-free web, *Phys. Rev. E*, **65**, 066122(4).

[24] **Szabó, G., Alava, M. and Kertész, J.**, 2003. Structural transitions in scale-free networks, *Phys. Rev. E*, **67**, 056102(5); cond-mat/0208551.

[25] **Catanzaro, M., Boguñá, M. and Pastor-Satorras, R.**, 2005. Generation of uncorrelated random scale-free networks, *Phys. Rev. E*, **71**, 027103(4).

[26] **Zhou, S. and Mondragon, R.J.**, 2004. The rich-club phenomenon in the Internet topology, *IEEE Comm. Lett.*, **8**, 180-182.

[27] **Colizza, V., Flammini, A., Serrano, M.A. and Vespignani, A.**, 2006. Detecting rich-club ordering in complex networks, *Nature Physics*, **2**, 110-115.

[28] **Batagelj, V. and Zaveršnik, M.**, 2002. Generalized cores, cs.DS/0202039.

[29] **Alvarez-Hamelin, I., Dall'Asta, L., Barrat, L., Vespignani, A.**, 2005. $k$-core decomposition: a tool for the visualization of large scale networks, cs.NI/0504107.

[30] **Alvarez-Hamelin, I., Dall'Asta, L., Barrat, L., Vespignani, A.**, 2005. $k$-core decomposition: a tool for the analysis of large scale Internet graphs, cs.NI/0511007.

[31] **Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. and Shir, E.**, 2006. MEDUSA - New model for Internet topology using $k$-shell decomposition, cond-mat/0601240.

[32] **Alberts, B., Johnson, A., Lewis, J., Raff, M., et al.**, 2002. Molecular biology of the cell, Garland Science, New York.

[33] **Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., et al.**, 2003. The evolution of transcriptional regulation in eukaryotes, *Mol. Biol. Evol.*, **20**, 1377-1419.

[34] **Lockhart, D.J. and Winzeler, E.A.**, 2000. Genomics, gene expression and DNA arrays, *Nature*, **405**, 827-836.

[35] **Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., et al.**, 2002. Transcriptional regulatory networks in saccharomyces cerevisiae, *Science*, **298**, 799-804.

[36] **Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., et al.**, 2004. Genomic analyis of regulatory network dynamics reveals large topological changes, *Nature*, **431**, 308-312.

[37] **Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., et al.**, 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae, *Nucl. Acids. Res.*, **34**, D446-451.

[38] **Kınıkoğlu, B. and Kırdar, B.**, 2006. Submitted to Yeast.

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

[40] **Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., et al.**, 2004. Global mapping of the yeast genetic interaction network, *Science*, **303**, 808-813.

[41] **Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., et al.**, 2004. Transcriptional regulatory code of a eukaryotic genome, *Nature*, **431**, 99-104.

[42] **Kauffman, S.A.**, 1969. Metabolic stability and epigenesis in randomly constructed genetic nets, *J. Theor. Biol.*, **22**, 437-467.

[43] **Barabasi, A.L. and Albert, R.**, 1999. Emergence of scaling in random networks, *Science*, **286**, 509-512.

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

[45] **Strogatz, S.H.**, 2001. Exploring complex networks, *Nature*, **410**, 268-276.

[46] **Ihmels, J., Levy, R., and Barkai, N.**, 2004. Principles of transcription control in the metabolic network of *S. cerevisiae*, *Nature Biotechnology*, **22**, 86-92.

[47] **Huang, S.**, 2004. Back to the biology in systems biology: What can we learn from biomolecular networks?, *Brief. Funct. Gen. Prot.*, **2**, 279-297.

[48] **Vazques, A., Dobrin, R., Sergi, D., Eckmann, J.P., et al.**, 2004. The topological relationship between the large-scale attributes and local interaction patterns of complex networks, *Proc. Natl. Acad. Sci. USA*, **101**, 17940-17945.

[49] **Bergmann, S., Ihmels, J. and Barkai, N.**, 2004. Similarities and differences in genome-wide expression data of six organisms, *PLoS Biol.*, **2**, 85-93.

[50] **Perelson, A.S. and Weisbuch, G.**, 1997. Immunology for physicists, *Rev. Mod. Phys.*, **69**, 1219-1267.

[51] **Reil, T.**, 1999. Dynamics of gene expression in an artificial genome, in *Proceedings of the 5th European Conference on Advances in Artificial Life*, LNCS, **1674**, pp. 457-466, Springer-Verlag, London.

[52] **Geard, N. and Wiles, J.**, 2003. Structure and dynamics of a gene network model incorporating small RNAs, in *Proceedings of the 2003 Congress on Evolutionary Computation (CEC).*

[53] **Watson, J., Geard, N. and Wiles, J.**, 2004. Towards more biological mutation operators in gene regulation studies, *BioSystems*, **76**, 239-248.

[54] **van Noort, V., Snel, B. and Huynen, M.A.**, 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model, *EMBO Rep.*, **5**, 280-284.

[55] **Banzhaf, W. and Kuo, P.D.**, 2004. Network motifs in natural and artificial transcriptional regulatory networks, *J. Biol. Phys. Chem.*, **4**, 85-92.

[56] **Wagner, A.**, 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization, *Proc. Natl. Acad. Sci. USA*, **91**, 4387-4391.

[57] **Wagner, A.**, 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.*, **18**, 1283-1292.

[58] **Sole, R.V. and Pastor-Satorras, R.**, 2002. Complex networks in genomics and proteomics, in *Handbook of Graphs and Networks*, Eds. Bornholdt, S. and Schuster, H.G., Wiley-VCH Verlag, Berlin.

[59] **Hannon, G.J.**, 2002. RNA interference, *Nature*, **418**, 244-251.

[60] **Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C.**, 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*, *Nature*, **391**, 806-811.

[61] **Montgomery, M.K. and Fire, A.**, 1998. Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression, *Trends Genet.*, **14**, 255-258.

[62] **Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G. and Mello, C.C.**, 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing, *Cell*, **106**, 23-34.

[63] **Guibas, L.J. and Odlyzko, A.M.**, 1981. Periods in strings, *J. Comb. Theory A*, **30**, 19-42.

[64] **Guibas, L.J. and Odlyzko, A.M.**, 1981. String overlaps, pattern matching, and nontransitive games, *J. Comb. Theory A*, **30**, 183-208.

[65] **Shannon, C.E.**, 1949. Communication in the presence of noise, *Proc. IRE*, **37**, 10-21.

[66] **Almirantis, Y. and Provata, A.**, 1999. Scaling properties of coding and non-coding DNA sequences, *J. Stat. Phys.*, **97**, 233-262.

[67] **Derényi, I., Palla, G. and Vicsek, T.**, 2005. Clique percolation in random networks, *Phys. Rev. Lett.*, **94**, 160202.

[68] **Molloy, M. and Reed, B.**, 1995. A critical point for random graphs with a given degree sequence, *Random Structures and Algorithms*, **6**, 161-180.

[69] **Caldarelli, G., Capocci, A., De Los Rios, P. and Munoz, M.A.**, 2002. Scale-free networks from varying vertex intrinsic fitness, *Phys. Rev. Lett.*, **89**, 258702.

[70] **Kauffman, S.**, 2004. A proposal for using the ensemble approach to understand genetic regulatory networks, *J. Theor. Biol.*, **230**, 581-590.

[71] **Huang, S. and Ingber, D.E.**, 2000. Shape-dependent control of cell growth, diferentiation and apoptosis: Switching between attractors in cell regulatory networks, *Exp. Cell Res.*, **261**, 91-103.

[72] **Albert, R. and Othmer, H.G.**, 2003. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *drosophila melanogaster*, *J. Theor. Biol.*, **223**, 1-18.

[73] **Mendoza, L. and Alvarez-Buylla, E.R.**, 2000. Genetic regulation of root hair development in arabidopsis thaliana: a network model, *J. Theor. Biol.*, **204**, 311-326.

[74] **Espinosa-Soto, C., Padilla-Longoria, P. and Alvarez-Buylla, E.R.**, 2004. A gene regulatory network model for cell-fate determination during *arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles, *The Plant Cell*, **16**, 2923-2939.

[75] **Derrida, B. and Flyvbjerg, H.**, 1986. Multivalley structure in Kauffman model - Analogy with spin-glasses, *J. Phys. A*, **19**, 1003-1008.

[76] **Derrida, B. and Pomeau, Y.**, 1986. Random networks of automata - a simple annealed approximation, *Europhysics Letters*, **1**, 45-49.

[77] **Derrida, B. and Flyvbjerg, H.**, 1987. Distribution of local magnetizations in random networks of automata, *J. Phys. A*, **20**, L1107-L1112.

[78] **Kadanoff, L., Coppersmith, S. and Aldana, M.**, 2002. Boolean dynamics with random couplings, nlin.AO/0204062.

[79] **Wilke, C.O., Ronnenwinkel, C. and Martinetz, T.**, 2001. Dynamic fitness landscapes in molecular evolution, *Physics Reports*, **349**, 395-446.

[80] **Kauffman, S.A.**, 1995. At home in the universe: the search for laws of self-organization and complexity, Oxford University Press, Oxford.

[81] **Aldana, M.**, 2003. Dynamics of boolean networks with scale-free topology, *Physica D*, **185**, 45-66.

[82] **Serra, R., Villani, M. and Agostini, L.**, 2004. On the dynamics of random Boolean networks with scale-free outgoing connections, *Physica A*, **339**, 665-673.

[83] **Bilke, S. and Sjunnesson, F.**, 2001. Stability of the Kauffman model, *Phys. Rev. E*, **65**, 016129.

## BIOGRAPHY

Duygu BALCAN was born in Istanbul in December, 1979. In 1997, she started her undergraduate education in the Department of Physics after attending the English preparation class for one year in the Department of Languages and History at Istanbul Technical University. She received her BSc degree in 2001 and her MSc degree in 2003 from the same department. She has been a PhD candidate in the Department of Physics at Istanbul Technical University since October, 2003. She has been working as a research assistant in the Faculty of Sciences and Letters at Istanbul Technical University since November, 2001.