

# **Biological networks: a thermodynamical approach**

A thesis submitted for the degree of  
Doctor of Philosophy  
at the University of Leicester.

by

**Yohann Grondin**

Department of Physics and Astronomy  
University of Leicester

March 2006

UMI Number: U602386

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U602386

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.  
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against  
unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

# *Declaration*

I hereby declare that this thesis has not been submitted in full or in part for any other degree at this or any other university.

The following papers have resulted from the work presented here:

*A thermodynamic view of networks*

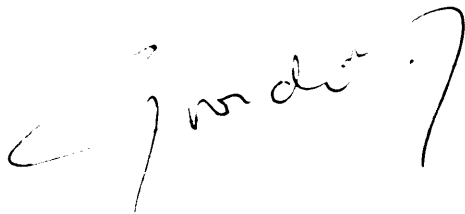
Raine D.J., Grondin Y., C R Biol. 2006 Mar;329(3):156-67.

*Networks as constrained thermodynamic systems*

Raine D.J., Grondin Y, Thellier M, Norris V., C R Biol. 2003 Jan;326(1):65-74.

Yohann Grondin

March 2006

A handwritten signature in black ink, appearing to read 'Yohann Grondin', with a stylized flourish at the end.

# *Abstract*

## **Biological networks: a thermodynamical approach**

**Yohann Grondin**

Many real systems can be represented by networks, that is a set of nodes connected to each other. The study of these systems as such has proven extremely useful as it gives access to a series of parameters that characterise their non-trivial architecture. This architecture is the product of many factors from the evolutionary mechanisms that shape the system during its growth to the functional dynamics on a shorter time scale. Gaining knowledge on the architecture is then of importance but faces many challenges in particular in the study of biological networks.

The first challenge is in terms of the method used to generate networks as we need to adopt an approach that, we expect, would allow us to understand those constraints and forces that shape the network.

The second challenge is that of understanding the relationship between the architecture of the system and its dynamics and functionality.

The third challenge is to get access using suitable techniques to the network architecture from expression data, such as mRNA abundances, for example.

We first show in this thesis that it is possible to generate networks from a thermodynamical viewpoint. This approach allows us to relate the architecture of network to some constraints. Furthermore, we show that some information on the structure resides in the non-randomness of the links between nodes. If we were to draw an analogy with traditional thermodynamics, networks could be modelled in a first approximation as perfect gases.

On a dynamical network of our design, we show a dependence of the architecture on the distribution of the level of expression of the nodes. Surprisingly, the distribution of the periods of those networks is a power-law and independent of the underlying architecture of the system. By comparing the data obtained from our model to experimental mRNA data we found a correlation between the degree of connectivity of genes and their level of abundance.

Finally, we show how we can apply a method used traditionally in image reconstruction to inference of networks.



# *Acknowledgements*

This work has been done under the supervision of Dr. D. Raine to whom I am very grateful for his help and guidance. This work has also benefited from numerous discussions with Prof. V. Norris, Dr. F. Képès and Dr. S. Gurman.

The work in section 4.4.2 was performed using the University of Leicester Mathematical Modelling Centre's supercomputer which was purchased through the EPSRC strategic equipment initiative.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Definition and representation of networks. . . . .	2
1.2	Parameters that characterise networks . . . . .	4
1.2.1	The distribution of the degree of connectivity . . . . .	4
1.2.2	Degree of clustering . . . . .	4
1.2.3	Diameter and average path length of networks . . . . .	5
1.2.4	Assortativity . . . . .	7
1.3	Various network models . . . . .	7
1.3.1	The Erdős and Rényi model . . . . .	9
1.3.2	The Barabási and Albert model . . . . .	12
1.3.3	The Watts and Strogatz model . . . . .	15
1.3.4	Summary . . . . .	18
1.4	Network models for biological systems . . . . .	18
1.4.1	A variety of biological networks . . . . .	18
1.4.2	Microarray technique . . . . .	21
1.4.3	Summary . . . . .	23
1.5	Properties associated to network . . . . .	24
1.5.1	Robustness and tolerance . . . . .	24
1.5.2	The spread of information, diseases, ... . . . .	25

1.5.3	Structural motifs . . . . .	25
1.6	Various approaches in modelling the dynamics of (biological) networks	26
1.7	Entropy and theory of information . . . . .	27
1.8	Outline of the thesis . . . . .	29
<b>2</b>	<b>Networks as Constrained Thermodynamic Systems</b>	<b>31</b>
2.1	Introduction . . . . .	31
2.2	Maximisation of the entropy subject to constraints . . . . .	32
2.2.1	The entropy maximisation principle . . . . .	32
2.2.2	Application to network construction . . . . .	33
2.3	Application to the construction of exponential networks . . . . .	35
2.3.1	General implementation of the method . . . . .	35
2.3.2	Mechanism for exponential network . . . . .	36
2.3.3	Numerical simulation . . . . .	38
2.4	Generalisation to more complex networks . . . . .	39
2.4.1	Alteration of the mechanism . . . . .	39
2.4.2	Application to scale-free networks . . . . .	42
2.4.3	Application to Mandelbrot networks . . . . .	44
2.5	Summary . . . . .	46
<b>3</b>	<b>Towards a thermodynamical description of networks</b>	<b>47</b>
3.1	Introduction . . . . .	47
3.2	The non-randomness of the connections . . . . .	48
3.2.1	First approach . . . . .	48
3.2.2	Second approach . . . . .	49
3.2.3	Summary . . . . .	52
3.3	Intensive and Extensive variables . . . . .	52
3.4	Network Complexity . . . . .	53
3.4.1	Exponential network . . . . .	55

3.4.2	Mandelbrot network . . . . .	56
3.4.3	Effect of the ‘renormalisation’ transformation . . . . .	56
3.5	The equation of state . . . . .	58
3.6	Summary . . . . .	59
<b>4</b>	<b>Dynamical network model</b>	<b>61</b>
4.1	Introduction . . . . .	61
4.2	The $NK$ -Boolean networks . . . . .	62
4.3	Setting up the model . . . . .	63
4.3.1	Architecture and representation of the network . . . . .	64
4.3.2	Dynamics . . . . .	65
4.3.3	External input . . . . .	65
4.3.4	The production function . . . . .	67
4.3.5	Is this model a Boolean network? . . . . .	68
4.3.6	The simulation . . . . .	69
4.4	Results . . . . .	69
4.4.1	Variation of the different categories of nodes . . . . .	69
4.4.2	Period distribution . . . . .	70
4.4.3	The distribution of abundance reflects the architecture of the network . . . . .	74
4.5	Summary . . . . .	76
<b>5</b>	<b>Modelling functions of genetic regulatory networks</b>	<b>78</b>
5.1	Introduction . . . . .	78
5.2	Messenger RNA abundance . . . . .	79
5.2.1	<i>Saccharomyces cerevisiae</i> . . . . .	79
5.2.2	<i>Caulobacter crescentus</i> . . . . .	84
5.2.3	Summary . . . . .	87
5.3	Can a simple model reproduce biological data? . . . . .	87

5.3.1	Reduction of complex network into a simple model . . . . .	87
5.3.2	Desynchronisation of the population . . . . .	89
5.3.3	Delays and degradation . . . . .	90
5.4	Architectural misfit . . . . .	93
5.4.1	Stochasticity . . . . .	95
5.5	Identification of correlation between node expression and number of controls . . . . .	96
5.5.1	The <i>E. coli</i> data . . . . .	97
5.5.2	Simulation . . . . .	101
5.6	Summary . . . . .	105
<b>6</b>	<b>Maximum entropy reconstruction of networks</b>	<b>107</b>
6.1	Introduction . . . . .	107
6.2	Methods to infer networks from data . . . . .	108
6.2.1	Mutual information . . . . .	109
6.2.2	Bayesian networks . . . . .	110
6.2.3	Inference from perturbation strategy experiments . . . . .	112
6.2.4	Summary . . . . .	113
6.3	The maximum entropy method . . . . .	113
6.3.1	Traditional maximum entropy method . . . . .	114
6.3.2	Application to network reconstruction . . . . .	116
6.3.3	The simulation . . . . .	120
6.4	Results . . . . .	121
6.4.1	Determination of $\alpha$ and $\lambda$ . . . . .	121
6.4.2	Reconstruction of the degree of connectivity . . . . .	123
6.4.3	Size of the sample . . . . .	127
6.5	Conclusion . . . . .	130
<b>7</b>	<b>Conclusion</b>	<b>132</b>

<b>A</b>	<b>Supplementary figures</b>	<b>134</b>
----------	------------------------------	------------

# Chapter 1

## *Introduction*

We will start with a very simple model that any system can be described by a set of elements and a set of interactions, regardless of their detailed nature. Connecting these elements together on the basis that they are interacting yields a network in which the elements are called either nodes or vertices and the interactions are called either links or edges. Since the nature of the nodes and of the links does not need to be prescribed, a network is a versatile means by which to represent a system. Examples include social networks (Watts & Strogatz, 1998, Amaral et al. 2000, Newman 2001, Börner, Maru & Goldstone, 2004), communication networks such as the Internet (Barabási & Albert, 1999, Adamic 1999), biological networks (Jeong et al. 2000, Jeong et al. 2001, Guelzim et al. 2002), etc, which will be detailed throughout this chapter.

In term of size, we can think of networks as made up of an infinite number of nodes even though this will not be the case here. For the networks to be considered in this thesis the number of nodes will be of the order of a thousand.

By its definition, a network is a model representation of a system, whether it is explicitly called a ‘model’ network or a ‘real’ network. However, we think that the term ‘model’ network is more appropriate to systems that describe properties of real systems but are constructed without knowledge of where the interaction are exactly, while ‘real’ networks would correspond to networks constructed from known sets of interactions. An alternative way to construct networks is to use data measured from the system related to the dynamical properties of the nodes, but not directly the interactions. This involves a process of reverse engineering which yields an ‘inferred’ network. There are many approaches to infer networks, using Bayesian networks (Friedman 2004) or mu-

tual information (Liang, Fuhrman & Somogyi, 1998), including our own presented in chapter 6.

The interest in describing systems in term of networks is to access information related to the architecture and topology of the system. By extracting the parameters that characterise the architecture, it is possible, for example, to envisage the mechanisms of evolution that have led to it (Albert & Barabási, 2002, Watts & Strogatz, 1998, van Noort, Snel & Huynen, 2004), and also to explain those mechanisms in term of constraints as shown in chapter 2. A network model also permits us to study the dynamics of the system and identify how the network correlates with its behaviour, for example (Kauffman 1993, Yuan, Chen & Wang, 2004, Li et al. 2004a). To go further, and this is an old problem, understanding the network is a way forward into bridging the gap between the architecture of a system and its function. We shall be interested in the use of networks to describe and study biological systems and, in particular, the genetic regulatory networks (Guelzim et al. 2002, van Noort et al. 2004).

Though the architecture of genetic regulatory networks is being thoroughly investigated (Guelzim et al. 2002), many questions, such as how they emerged and evolved remains to be answered (van Noort et al. 2004). In this thesis, we will give possible tracks to answer some of them.

In this chapter, we will introduce some of the parameters used to describe networks. We will present the major network models and describe their properties. We will then focus on biological networks and discuss the importance of the data in generating them. We will introduce the information entropy which will be used later. Finally, we discuss the plan of the rest of this thesis.

## 1.1 Definition and representation of networks.

A network is formally described by a graph,  $G$ , that is an ordered pair of disjoint sets  $(V, E)$  where  $V$  is the set of vertices  $v_i$  and  $E$  is the set of edges  $e_i$  (Bollobás 1979). Several matrices can be associated to a graph amongst which the *adjacency* matrix and the *incidence* matrix are the most important. The adjacency matrix,  $A$ , is an  $N \times N$  matrix whose elements  $a_{ij}$  are given by

$$a_{ij} = \begin{cases} 1 & \text{if } v_i v_j \in E(G); \\ 0 & \text{otherwise.} \end{cases}$$



where  $N$  corresponds to the number of nodes. A network is said to be undirected if  $a_{ij} = a_{ji}$  that is  $i$  and  $j$  cannot be distinguished from the link pairing them. On the contrary, in a directed network  $a_{ij} = a_{ji}$  is not necessarily true. In this case, node  $i$  is either the initial or terminal vertex connected to  $j$ , and  $j$  either the terminal or initial vertex, respectively.

The incidence matrix,  $B$ , of a graph is a  $N \times M$  matrix whose elements  $b_{ij}$  are given by

$$b_{ij} = \begin{cases} -1, & \text{if } v_i \text{ is the initial vertex of the edge } e_j; \\ 1, & \text{if } v_i \text{ is the terminal vertex of the edge } e_j; \\ 0, & \text{otherwise,} \end{cases}$$

where  $M$  is the number of links in the network.

The degree of connectivity  $k_i$  of node  $i$  is defined as the number of edges connected to it. In an undirected network where the adjacency matrix is symmetrical, the degree of connectivity is given by  $k_i = \sum_j a_{ij} = \sum_j a_{ji}$ . Conversely, in a directed network, for which the adjacency matrix does not have to be symmetrical, it is possible to define a out-going degree, or out-degree, of connectivity,  $k_{out,i}$ , by considering only the initial vertices of node  $i$ , and an in-coming degree or in-degree, of connectivity,  $k_{in,i}$ , by considering only the terminal vertices of node  $i$ . The degree of node  $i$  will then be  $k_i = k_{out,i} + k_{in,i}$ .

The two matrices are related to each other by the following expression (Bollobás 1979)

$$BB^t = D - A,$$

where  $B^t$  is the transpose of  $B$  and  $D$  is the diagonal matrix whose elements  $d_{ii}$  are the degree of connectivity  $k_i$  of node  $i$ .

The mean degree of connectivity,  $\bar{k}$ , is an important parameter of networks. It can be calculated either from the adjacency matrix such that,

$$\bar{k} = N^{-1} \sum_{i,j} a_{ij},$$

or from the probability distribution such that,

$$\bar{k} = \sum_k p_k k,$$

where  $p_k$  is the probability of finding a node of degree  $k$ . For a directed network, it is possible to define the mean connectivity of the in-degrees and out-degrees, as  $\bar{k} =$

$$\bar{k}_{out} + \bar{k}_{in} \text{ and } \bar{k}_{out} = \bar{k}_{in}.$$

A sequence of edges  $n_0n_1, n_1n_2, \dots, n_{r-1}n_r$  is a walk of length  $r$ . If the edges and the nodes are all distinct then this walk is called a *path* (Beineke & Wilson, 1997). A graph is *connected* if there is a path joining each pair of vertices of  $G$ . Conversely, the graph is *disconnected*. Every disconnected graph can be split into connected subgraphs called *components*. This notion of connectedness is important since we will be dealing with sparse graphs that are then more likely to be disconnected. However, in most cases, we will want to work with connected graphs. Finally, a graph is called *k-regular* if all the nodes have the same degree of connectivity  $k$  (Beineke & Wilson, 1997).

## 1.2 Parameters that characterise networks

Many parameters are used to characterise and define networks. Some of those parameters will be introduced here as they will be used throughout this thesis while some others, such as  $\beta$ -complexity, will be introduced later on as part of the studies that will be presented.

### 1.2.1 The distribution of the degree of connectivity

The probability distribution of the degree of connectivity of the nodes, also called the degree distribution, describes and characterises, though not completely, the global architecture of networks. In an undirected network, it gives the probability  $p_r$  that a node, chosen at random, has exactly a degree of connectivity  $r$ . In a directed network it is possible to extract the distribution of the out-degree and the in-degree as well of the total degree of the nodes. Various degree distribution give rise to various networks of extremely different architectures as we will see later.

### 1.2.2 Degree of clustering

The degree of clustering of a network, also referred to as the ‘clique’, is a statistical measure that provides information on the clustering of the neighbourhood of nodes. It is given by the clustering coefficient,  $C$ , which is the average over the network of the clustering coefficient of each of the nodes (Watts & Strogatz, 1998).

The clustering coefficient,  $C_i$ , of node  $i$  is calculated as the ratio of the number of links between nodes connected to  $i$ , to the number of possible links between all those

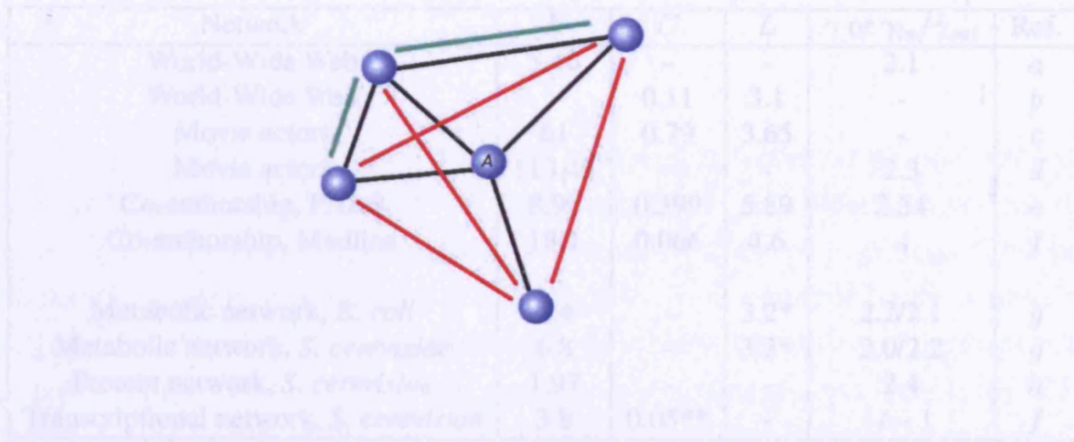


FIGURE 1.1. The clustering coefficient,  $C$ , at a node  $A$ . There are 2 triangles formed from connected neighbours of  $A$  (green) while 4 other triangles could be formed (red). The clustering coefficient is then  $C = 1/3$ .

nodes connected to node  $i$ . As illustrated in figure 1.1,  $C_A$  is as well given by the fraction of the number of existing triangles that have for summit the node  $A$  (green) to the number of possible ones from the nodes connected to  $A$  (green and red).

The number of triangles at node  $i$  is obtained from the diagonal element (counted twice) of the cubed adjacency matrix of the network. The number of possible triangles is given by  $k_i(k_i - 1)/2$ . The clustering coefficient of all the network is then

$$C = N^{-1} \sum_i \frac{a_{ii}^{[3]}}{k_i(k_i - 1)}. \quad (1.1)$$

The clustering coefficient,  $C$ , of various networks is shown in table 1.1.

Note that this formula based on the adjacency matrix will not give identical results for directed and undirected networks. In the case of directed networks, it will be preferable to symmetrise the adjacency matrix.

### 1.2.3 Diameter and average path length of networks

The diameter,  $D$ , of a network is a global parameter defined as the longest of the shortest path, with the shortest path being the minimum path between two nodes. As an example, the diameter of the network shown in figure 1.2, is given by the green path. A measure related to the diameter is the average path length,  $\langle D \rangle$ , which is the average over all the shortest paths.



Network	$k$	$C$	$L$	$\gamma$ or $\gamma_{in}/\gamma_{out}$	Ref.
World-Wide Web	5.46	-	-	2.1	a
World-Wide Web	-	0.11	3.1	-	b
Movie actors	61	0.79	3.65	-	c
Movie actors	113.43	-	-	2.3	d
Co-authorship, PNAS	8.97	0.399	5.89	2.54	e
Co-authorship, Medline	18.1	0.066	4.6	-	f
Metabolic network, <i>E. coli</i>	7.4	-	3.2*	2.2/2.1	g
Metabolic network, <i>S. cerevisiae</i>	6.8	-	3.3*	2.0/2.2	g
Protein network, <i>S. cerevisiae</i>	1.97	-	-	2.4	h
Transcriptional network, <i>S. cerevisiae</i>	3.8	0.05**	-	-/ $\sim 1$	f

Table 1.1. Comparison between parameters of various real networks. The references are (a) (Barabási & Albert, 1999), (b) (Adamic 1999), (c) (Watts & Strogatz, 1998), (d) (Amaral et al. 2000), (e) (Börner et al. 2004), (f) (Newman 2001), (g) (Jeong et al. 2000), (h) (Jeong et al. 2001), (i) (Guelzim et al. 2002), \*diameter, \*\*semi-clustering coefficient.

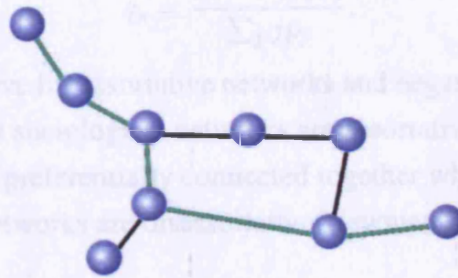


FIGURE 1.2. The diameter,  $D$ , of a network. It is the longest of the shortest paths which is represented in this picture by the green path. In this example  $D = 5$ .

Those two parameters are however very expensive to determine in term of computing time. A simple brute force algorithm on a sparse network where the shortest path between two nodes is determined by crawling can be estimated to be about  $\bar{k}^{<D>} N^2$  complex. Another parameter, called the characteristic path length,  $L$ , has instead been introduced. This is the average of the shortest paths of randomly chosen pairs of nodes, selected a number of times so that this average converges. Even though this measure is not the diameter, it is characteristic of the network (Watts 1999). The characteristic path

length of several networks is shown in table 1.1.

### 1.2.4 Assortativity

The assortative behaviour, or assortativity, of a network is defined as the preference for nodes of large degrees of connectivity to be connected to each other. Conversely, the disassortative behaviour is defined as the preference for nodes of large degrees of connectivity to be connected to nodes of small degrees. It is measured by the assortative coefficient,  $r$ . To define  $r$ , let  $e_{ij}$  be the joint probability distribution of the degrees of the nodes at the ends of a randomly chosen link, not counting this link itself in the nodal degrees (Callaway et al. 2001). Then  $r$ , ( $-1 \leq r \leq 1$ ), is given by

$$r \propto \frac{\sum_{ij} ij(e_{ij} - q_i q_j)}{\left(\sum_k k^2 q_k - (\sum_k k q_k)^2\right)},$$

where the normalised ‘remaining degree’ distribution (Callaway et al. 2000, Barabási & Albert, 1999)  $q_k$  is

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j j p_j}.$$

The coefficient  $r$  is positive for assortative networks and negative for disassortative ones. It has been measured that sociological networks are assortative that is nodes of large degrees of connectivity are preferentially connected together whereas the Internet network and various biological networks are disassortative (Newman 2002).

## 1.3 Various network models

With only the set of parameters presented above, it is already possible to describe networks with a variety of architectures and properties. The networks that will be described here, named after their authors, are very well known models and have been generalised in several ways. However, the naming of these generalisations, for example, may be misleading in relation to the properties of those networks. We will therefore try to define properly these networks and use adequate terms to name them, without entering a debate as to what is appropriate.

Along with the properties of model networks, we give relevant mechanisms, mostly intuitive, to generate them. According to these mechanisms, two contexts emerge: growing systems, in which nodes are added, and non-growing ones. This difference will be

### 1.3.1 The Erdős and Rényi model



FIGURE 1.3. Comparison between two networks with different degree distributions. The network on the left has a degree distribution that follows a Poisson law and the network on the right has a degree distribution that follows a power-law. The degree distribution has a clear impact on the architecture of the networks. The graphs are drawn using Pajek software (Batagelj & Mrvar, 2003).

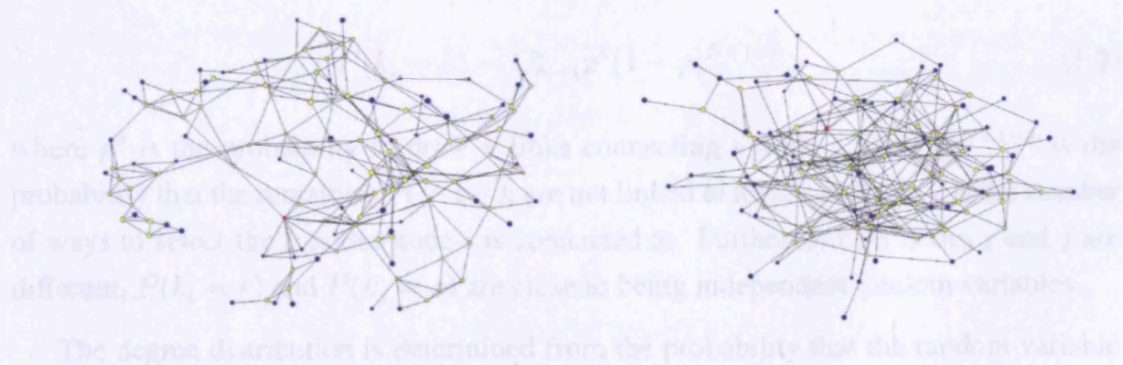


FIGURE 1.4. Comparison between two networks with different clustering coefficients. The network on the left has a higher clustering coefficient than the network on the right yielding differences in the architecture. The graphs are drawn using the Pajek software (Batagelj & Mrvar, 2003).

relevant later on.

To give a glimpse on how variations of the numerous parameters that describe a network affect their architecture, we compare in figure 1.3 two networks with a different distribution of the degree of connectivity and in figure 1.4 two networks with a different clustering coefficient. The different organisation of the network architecture leads to clearly visible differences.

Thus the number of nodes with degree  $k$  follows a Poisson distribution with mean value  $\lambda_k$  for every fixed  $k$ . The Poisson distribution decays rapidly for large values of  $k$ , the standard deviation of the distribution being  $\sigma_k = \lambda_k^{1/2}$ . With a bit of simplification, 1.3 implies that  $\lambda_k$  does not diverge much from the approximate result



### 1.3.1 The Erdős and Rényi model

The Erdős and Rényi model gives rise to a category of network called *random* networks. Those networks are the cornerstone of network studies and are used as a reference to which other networks models can be compared. We will focus on the properties of graphs that will be of interest later.

#### Characteristics

One way to introduce this model is to start with the methods to generate it. For a network constituted of  $N$  nodes, the first method consists in connecting  $n$  pairs of nodes, chosen randomly from the  $N(N-1)/2$  possible ones. Alternatively, each possible pairs of nodes are chosen independently and connected with a certain probability  $p$ , ( $0 < p < 1$ ). In the later case, the degree of connectivity  $k_i$  of a node  $i$  has binomial distribution with parameters  $N - 1$  and  $p$  such that

$$P(k_i = k) = C_{N-1}^k p^k (1 - p)^{N-1-k}, \quad (1.2)$$

where  $p^k$  is the probability to draw  $k$  links connecting a node  $i$ ,  $(1 - p)^{N-1-k}$  is the probability that the remaining  $N - 1 - k$  are not linked to node  $i$  and  $C_{N-1}^k$  is the number of ways to select the  $k$  nodes node  $i$  is connected to. Furthermore, if nodes  $i$  and  $j$  are different,  $P(k_i = r)$  and  $P(k_j = s)$  are close to being independent random variables.

The degree distribution is determined from the probability that the random variable  $X_k$ , that is the number of nodes of degree  $k$ , takes on a given value,  $P(X_k = k)$ . According to equation 1.2, the expectation value of the number of nodes with degree  $k$  is

$$E(X_k) = NP(k_i = k) = \lambda_k,$$

where

$$\lambda_k = NC_{N-1}^k p^k (1 - p)^{N-1-k}.$$

The distribution of the  $X_k$  values,  $P(X_k = k)$ , given by Bollobás (1985), approaches a Poisson distribution,

$$P(X_k = k) = e^{-\lambda_k} \frac{\lambda_k^k}{k!}, \quad (1.3)$$

Thus the number of nodes with degree  $k$  follows a Poisson distribution with mean value  $\lambda_k$ , for every fixed  $k$ . The Poisson distribution decays rapidly for large values of  $k$ , the standard deviation of the distribution being  $\sigma_k = \lambda_k^{1/2}$ . With a bit of simplification, Eq. 1.3 implies that  $X_k$  does not diverge much from the approximate result

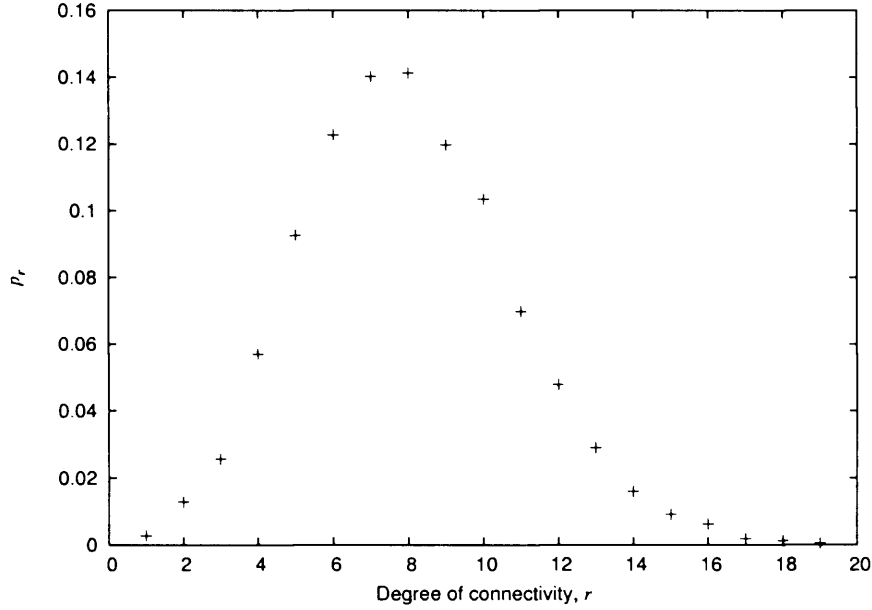


FIGURE 1.5. Distribution of the degree of connectivity of a random network with  $N = 8000$  nodes and a mean connectivity,  $\bar{k} = 8.00$ . The distribution follows a Poisson distribution centred on  $\bar{k}$ .

$X_k = NP(k_i = k)$ , valid only if the nodes are independent (Albert & Barabási, 2002). Thus to a good approximation the degree distribution of a random graph is the binomial distribution,

$$P(k) = C_{N-1}^k p^k (1-p)^{N-1-k},$$

which for large  $N$  can be replaced by a Poisson distribution (figure 1.5),

$$P(k) \approx e^{-pN} \frac{(pN)^k}{k!} = e^{-\bar{k}} \frac{(\bar{k})^k}{k!},$$

with unique characteristic parameter  $\bar{k} = pN$ , that is the mean degree of connectivity defined earlier.

In random networks, the mean connectivity determines whether the network is connected or not such as

- i. If  $\bar{k} < 1$ , a typical graph is composed of isolated trees.
- ii. If  $\bar{k} > 1$ , a giant cluster appears.
- iii. If  $\bar{k} \geq \ln(N)$ , almost every graph is totally connected.



The generation of sparse random network may then yield disconnected networks.

The maximum degree of almost all random graphs has the same order of magnitude as the average degree. Thus, despite the fact that the position of the edges is random, a typical random graph is rather homogeneous, the majority of the nodes having roughly the same number of links (Albert & Barabási, 2002).

### Clustering coefficient

In a random network, the probability that two of the nearest neighbours of a node are connected is approximately equal to the probability that two randomly selected nodes are connected. As the number of links in a network is  $N\bar{k}/2$ , the clustering coefficient is then, approximately,

$$C_{\text{random}} = \frac{N\bar{k}/2}{N(N-1)/2} \approx \frac{\bar{k}}{N}.$$

The clustering coefficient of sparse random networks is then very small.

### Diameter

Considering a sparsely connected random network is spreading, that is, it has a small clustering coefficient, it is possible to estimate the diameter with a large probability. Let  $N_l$  be the number of nodes within a distance  $l$  from a node. We can write

$$N_l = \sum_{i=0}^l \Gamma_i,$$

where  $\Gamma_i$  is the number of node at distance  $i$ . It can be approximated by

$$N_l = \sum_{i=0}^l \bar{k}^i = \frac{\bar{k}^{l+1} - 1}{\bar{k} - 1}.$$

Equating  $N_l$  to the number of nodes  $N$  in the network and taking the logarithm, we find that the diameter varies as

$$D \propto \frac{\ln N}{\ln \bar{k}}.$$

It has been shown, for  $\bar{k}/\log N > 2$ , that the range for which the diameters of these networks can vary is very small and concentrated around the value  $\ln N / \ln \bar{k}$  (Chung & Lu, 2001).

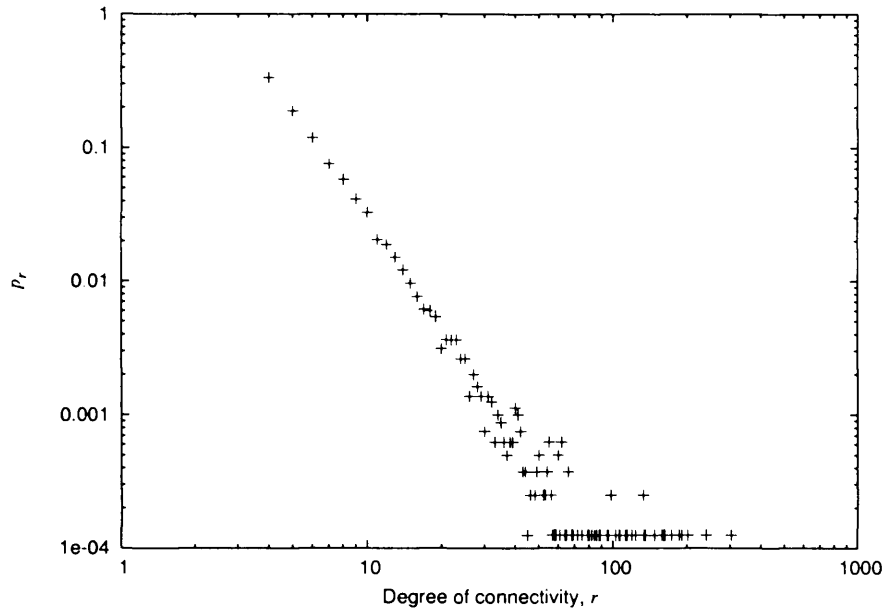


FIGURE 1.6. Distribution of the degree of connectivity of a scale-free network with  $N = 8000$  nodes and  $\bar{k} = 8.00$ . The distribution follows a power-law of slope  $\gamma = 3.00$ .

### 1.3.2 The Barabási and Albert model

The Barabási and Albert model yields a different category of networks called *scale-free* networks. Their importance lies in the fact that many real networks tend to display the same properties, as shown during the last decade.

#### Characteristics

A network is said to be scale-free when the probability  $p_r$  of having a node of degree  $r$  follows a power-law, such as

$$p_r = \frac{r^{-\gamma}}{\sum_i r_i^{-\gamma}},$$

where  $r_i$  is the degree of connectivity of node  $i$ . Here,  $\gamma$  is the characteristic parameter of the distribution as opposed to  $\bar{k}$  in random networks. The power-law distribution implies that there is a majority of nodes of low connectivity and a non-negligible proportion of nodes with a high degree of connectivity, as compared to a random network with the same  $\bar{k}$ . This change of distribution has a clear effect on the architecture of the network as illustrated in figure 1.3.

Numerous real networks in various fields present degree distributions with power-law tails. There is, for example, the World Wide Web whose nodes are the pages and the links the hyperlinks between pages (Barabási & Albert, 1999). There is also the actor-movie collaboration networks where actors, *i.e.* the nodes of the network, are connected to each other if they have played a part in a same movie (Amaral et al. 2000). There is the citation network where nodes are the papers and the links the citations therein (Redner 1998). Biological networks will be discussed later. Several parameters of these networks are compared in table 1.1.

The original, simple method, based on intuitive mechanisms, to generate such networks was introduced by Barabási & Albert, (1999). It is derived from observations of real networks. A characteristic of real networks is that new elements are being continuously added, for example new websites, actors, articles, etc. The other mechanism comes from the observation that more popular sites, actors, etc, are more likely to be acknowledged by receiving more links than others. Thus, the simple ingredients are those of the growth of networks and preferential attachment to the most popular nodes. The generative mechanism is then:

- (i) A new node  $i$  with  $m$  links is added to the network, starting from a core network.
- (ii) The new node is preferentially attached to other well connected nodes such that the probability  $p$  that the new added node is attached to node  $i$  depends on the connectivity  $k_i$  of that node. That is

$$p = \frac{k_i}{\sum_i k_i}.$$

This model yields in the stationary state a degree distribution  $p_k \propto k^{-\gamma}$ , with  $\gamma = 2.9 \pm 0.1$  (figure 1.6).

Various analytical approaches (Barabási & Albert, 1999, Barabási, Albert & Jeong, 1999, Dorogovtsev, Mendes & Samukhin, 2000, Krapivsky, Redner & Leyvraz, 2000) to the above mechanism, agreed on predicting that

- (i) The degree  $k_i$  of node  $i$ , introduced at  $t_i$  increases with time,  $t$ , as

$$k_i(t) \propto m \left( \frac{t}{t_i} \right)^\beta$$

with  $\beta = 1/2$ . The age of a node in the network is then correlated to its degree of connectivity.

(ii) The degree distribution varies asymptotically such as

$$P(k) \propto m^2 k^{-3}.$$

The distribution reaches then a steady state and is independent of the degree of the introduced nodes, in agreement with the numerical results.

More refined models have been proposed to account for features not reproduced by this simple model. These include rewiring processes (Albert & Barabási, 2000) and notions of attractiveness of nodes (Dorogovtsev, Mendes & Samukhin, 2000) which can account for the variation of  $\gamma$  ( $1 < \gamma < 3$ ) observed in real networks as shown in table 1.1. We mention also the aging phenomenon (Dorogovtsev & Mendes, 2000) or the memory effect (Klemm & Eguíluz, 2002*b*) which both attempt to decorrelate the age of the node and its probability of attracting more links, in agreement with the Internet network (Adamic & Huberman, 2000) or social networks, for example. But overall, it seems the essential condition for such networks to appear remains the asymptotic linear preferential attachment, whether it be directly implemented (Albert & Barabási, 2002, Krapivsky et al. 2000) or it indirectly derives from other mechanisms (Klemm & Eguíluz, 2002*b*).

### Clustering coefficient

It is shown numerically that the clustering coefficient of the Barabási and Albert model decreases with the number of nodes as a power-law  $C \sim N^{-0.75}$ . This is slower than the  $C = \bar{k}N^{-1}$  decay observed for random graphs. For a network with  $\bar{k} = 4$ , there is a factor difference that increases with the size of the network (Albert & Barabási, 2002) between the Barabási and Albert model and the random one. The decay of  $C$  has also been demonstrated analytically, but for large networks rather densely connected, as well as a weak dependence between the degree of connectivity of the nodes and  $C$  (Fronczak, Fronczak & Hołyst, 2003).

### Average path length

The average path length of Barabási and Albert networks is smaller than that of random networks for any number of nodes. This means that the different structural organisation of the Barabási and Albert model allows for a minimisation of the average path length. Numerically, it is shown that the average path length increases as  $\ln N$ . While

analytically proven to be the case for  $m = 1$ , it differs slightly for  $m \geq 2$  (Albert & Barabási, 2002, Bollobás & Riordan, 2004).

### **Note on the scale-free network term**

Note that ‘scale-free’ networks were named based on the fact that they present a power-law distribution, and this is all that we will understand here by ‘scale-free’. The Barabási and Albert network presents many characteristics apart from being scale-free as shown above. Note also that in the comparison between scale-free and random networks, we only compare properties related to the degree distribution. These restrictions seem appropriate in relation to the ongoing debate about scale-freeness, the interpretation of this property or even its definition (Keller 2005, Arita 2005).

### **1.3.3 The Watts and Strogatz model**

The Watts and Strogatz model introduces the class of *small-world* networks. Distinct from the two previous models, this class of network is not related to its degree distribution but rather to the clustering coefficient and the characteristic path length. Small-world networks are ubiquitous in the real world such as engineered networks, social networks or biological networks (Watts 1999) which, in the last case will be discussed in more detail below.

#### **Characteristics**

The term ‘small-world’ originates from the fact that on average two individuals in the world are not any further than 6 acquaintances from each other. This is also referred to as the phenomenon of six-degrees of separation and was introduced by the Milgram (1967) experiment.

Going back to networks, Watts & Strogatz, (1998) constructed networks with the following properties:

- (i) The clustering coefficient is large, in comparison to a random network, for example; that is, the nodes are locally highly connected to each other.
- (ii) The characteristic path length is small in comparison to an ordered network.

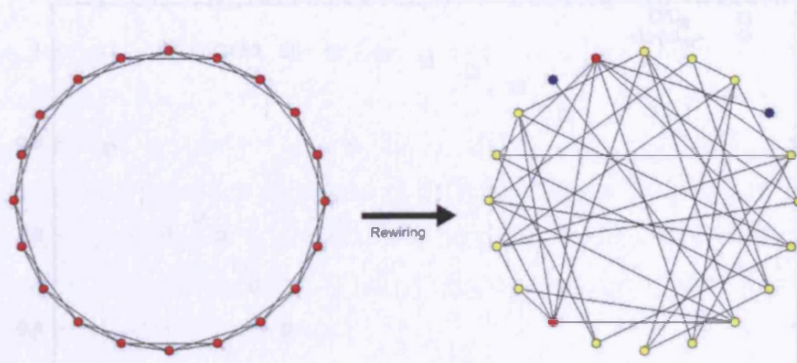


FIGURE 1.7. Construction of a small-world network. As the links of a ring lattice (left) are rewired, the architecture of the changes to that of a random network (right). The small-world network occurs at an intermediate stage.

These networks were termed ‘small-world’ and this is the definition we are going to use here even though slightly different versions may exist (Kaufmann, Lehmann & Post, 2005). As shown in table 1.1, many real networks present the small-world characteristics given above.

### Construction methods

The typical mechanism introduced by Watts & Strogatz, (1998) to obtain such a network consists of the simple random rewiring of a fraction  $p$  of the links of a given network. The model starts, that is at fraction  $p = 0$ , with a  $k$ -regular circular network of  $N$  nodes as depicted on figure 1.7. This starting network is highly clustered, with

$$C = \frac{\bar{k}^2 - (\bar{k}/2 + 1)(\bar{k}/2 + 2) + 2}{\bar{k}(\bar{k} - 1)},$$

that is  $C \rightarrow 3/4$  when  $p \rightarrow 0$  and  $\bar{k} \rightarrow \infty$  in the limit  $N \gg \bar{k} \gg \ln N$ , and  $\bar{k}$  even. On the other hand, at  $p = 1$ , all the links have been rewired and the resulting network should be random. The clustering coefficient is then much smaller with  $C \approx C_{\text{random}} \sim \bar{k}/N$ .

As shown in figure 1.8, there is a range of  $p$  for which the value of  $C$  remains high, whereas the value of  $L$  is close to that of random networks. The sharp decrease of  $L$  comes from the introduction of long-range edges, creating short-cuts and starts after a small fraction of the nodes have been rewired. To be slightly more precise, we can see

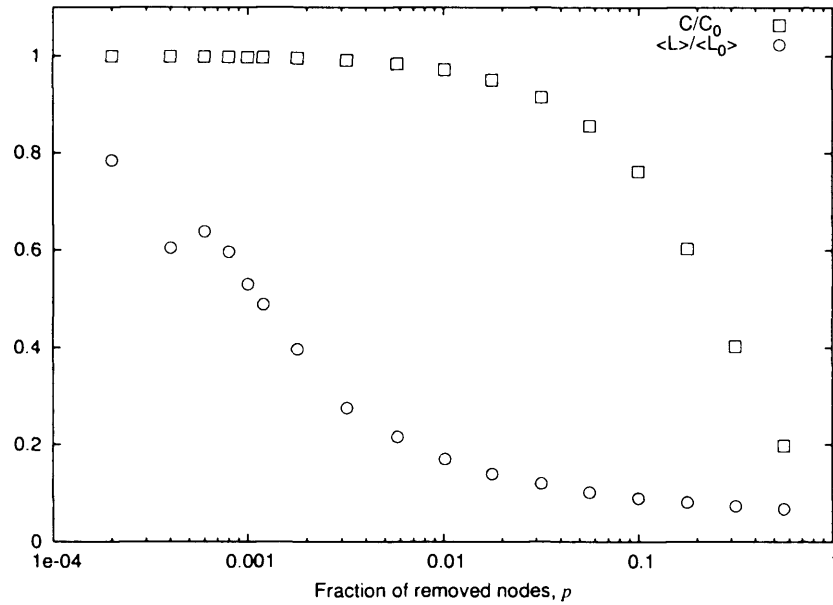


FIGURE 1.8. Variation of the clustering coefficient,  $C$ , and characteristic path length,  $L$ , as the fraction  $p$  of rewired links. The values  $C$  and  $L$  are normalised to their initial values. There is a range of  $p$  where  $C$  remains high while  $L$  is at about its lowest which characterises a small-world network.

about an 80% drop in the characteristic path length whereas the clustering coefficient drops only by about 5%.

Alternative models to generate small world networks exist. There are methods where links are added to randomly chosen pairs of nodes so as to reduce the characteristic path length without altering much the clustering coefficient (Newman & Watts, 1999a, Newman & Watts, 1999b). Another variant, (Kasturirangan 1999), consists in adding new nodes and connecting them randomly to the existing ones. Aside from probabilistic methods, there are deterministic networks, which allows an analytical description of their properties (Comellas, Ozon & Peters, 2000, Zhang, Rong & Guo, 2005) as well as the construction of small-world networks of any node distribution (Comellas & Sampels, 2002). Other methods implying growth mechanisms can also yield networks that are scale-free and small-world (Klemm & Eguíluz, 2002a).

All this shows that a class of small-world networks exists apart from the class of networks defined by their degree distribution, for example. This also indicates that small-world networks are not intermediates between regular and random networks as the first construction method described would suggest.

### 1.3.4 Summary

As we are starting to see, networks present various architectural characteristics and properties that combine to form complex systems. If different methods are capable of producing networks with various properties, those methods are also constraining a lot the final outcome so that generating networks of different degree distributions, for example, requires a complete change of approach. Furthermore, as intuitive as the methods are, they do not deliver an understanding of the natural constraints applied to the network so as to induce the proposed mechanisms.

## 1.4 Network models for biological systems

We have discussed so far the properties of model networks and given examples of real networks presenting identical properties. In this section we will focus on a more detailed description of biological networks. We will see how the description of biological systems in term of networks is closely related to the type of data available and give some examples. We will then detail the type of data that will be used later in chapters 5 and 6.

### 1.4.1 A variety of biological networks

There are many levels at which biological systems can be described but we will only discuss the molecular, and intra-cellular, level. The intra-cellular medium of a cell is composed of thousands of interacting molecules of various natures. This diversity allows for the formation and the description of various different networks. The possible elements considered here for the network are the DNA, the mRNA, the proteins and other intermediary metabolites. For the sake of modelling, we consider that those elements interact in the simplified manner as illustrated in figure 1.9. As shown, it is possible to define horizontal and vertical interactions between elements, which can be turned into several different networks. Note that apart from exemplifying the topology of these various networks mentioned above, each model addresses a different issue.

#### The protein-protein network

The first of the possible networks is the protein-protein interactions network. Such a network is computed from experimental data obtained from many organisms such as the yeast *Saccharomyces cerevisiae* (Schwikowski, Uetz & Fields, 2000, Ito et al. 2001), the



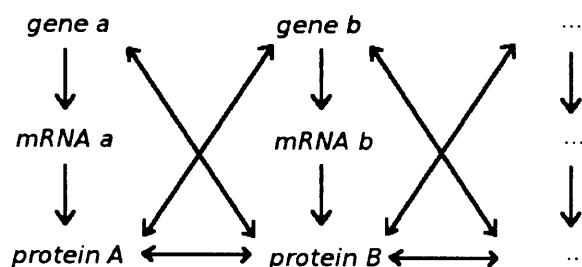


FIGURE 1.9. A simplified representation of interactions of a genetic network. The double arrows represent possible interactions among the various elements (genes, mRNAs, proteins) yielding many type of networks.

worm *Caenorhabditis elegans* (Li et al. 2004b) or the human (Rual et al. 2005). In this network, the nodes represent proteins whereas the links are the interactions between each proteins as measured by double-hybrid interaction experiments. It results in a static and undirected network which, like a map, states the possible interactions.

The main concern with networks drawn from double-hybrid data is the non-negligible proportion of false-positives as well as of false negatives (Schwikowski et al. 2000, Ito et al. 2001, Rual et al. 2005, Fields 2005). Relying only on those data may then be problematic in the absence of other parallel experiments to corroborate the results.

The main characteristics of these networks are a power-law distribution (Jeong et al. 2001, Li et al. 2004b, Han et al. 2004, Rual et al. 2005), and a disassortative behaviour, that is a tendency for highly connected nodes to interact with less highly connected nodes (Maslov & Sneppen, 2002, Rual et al. 2005). The protein-protein interaction network of *S. cerevisiae* and *C. elegans* displays a high degree of clustering and a small characteristic path length (Wagner 2001) characteristic of small-world networks. On the other hand, the protein-protein interaction network in humans does not have a such a high degree of clustering (Rual et al. 2005).

Aside of its architectural interest, those networks are used to characterise proteins of unknown function by using the fact that proteins of analogous function should be connected.

### The metabolic network

The metabolic networks are composed mainly of the intermediary metabolites found in the energy metabolism and the molecule synthesis routes. There is no clear cut way that metabolic networks should be constructed. Most commonly though, the nodes represent the substrates or the products of metabolic reactions, which are connected if they are part of a biochemical reaction (Fell & Wagner, 2000, Wagner & Fell, 2001, Jeong et al. 2000, Ma & Zeng, 2003). Another possibility is to represent the metabolic reactions as the nodes and to connect any two of them if they share a substrate or a product (Wagner & Fell, 2001). Variations in the exact method to generate these networks yield both directed and undirected networks.

There are several interests in metabolic networks, the main one of which is to study evolution through the finding of modules, that are related to metabolic functions. The idea is to find a mechanism that could explain the observed structure and, in particular this modular organisation of networks (Ravasz et al. 2002), which is also found in protein-protein interaction networks (Han et al. 2004).

It was reported that the metabolic network of *E. coli* has the property of a small-world network as well as a power-law distribution of the degree of connectivity of the nodes (Fell & Wagner, 2000, Wagner 2001, Raine & Norris, 2001). This result was extended to the other studied organisms (Jeong et al. 2000).

We mention that an alternative study in *E. coli*, in which the metabolites were connected only if they were exchanging carbon atoms, yields a different network (Arita 2004). It is found that the corresponding network has a higher average path length, questioning the small-world property of this representation of metabolic networks.

### The genetic networks

The genetic networks are built from levels of expression of the genes which are given by the measure of mRNA abundance. The resulting network, depending of the type of data, will either be that of a static network or a dynamic one. We will refer to a dynamic network as a regulatory network.

The construction of a genetic network is not as straightforward as that of the protein-protein interaction networks in the sense that the mRNA abundance bears no direct information on the interactions in the network. Various genetic networks can be generated amongst which are the co-expression networks and the transcriptional networks. The many ways to generate such networks from mRNA abundances are discussed later in

chapter 6. In these networks, the nodes represent the genes while the links might have various meaning. In the co-expression network, for example, nodes are connected to each other if the corresponding genes are co-expressed (van Noort et al. 2004). In transcriptional networks, the nodes are the genes which are connected if one is acting on the other via a transcription factor (Lee et al. 2002, Guelzim et al. 2002).

Genetic networks present the same features as the metabolic network (except for one study (Arita 2004)), and the protein-protein interaction networks. They have a power-law distribution of their degrees of connectivity and show the small-world property (van Noort et al. 2004). A more complex picture emerges when looking at the arriving and departing connectivity of the genes in *S. cerevisiae* and *E. coli* (Guelzim et al. 2002). The authors show that the arriving connectivity, defined by the number of transcription factors regulating a gene, is better fitted by an exponential distribution whereas the departing distribution, given by the number of genes a transcription factor can regulate, follows a power-law.

Alternative mechanisms to the one presented earlier have been suggested and modelled to account for the properties of the biological networks described here (Wagner 2003, Chung et al. 2003, Teichmann & Babu, 2004, van Noort et al. 2004). Those mechanisms are based on the similar concept of gene duplication.

### 1.4.2 Microarray technique

The nature of the data is an essential issue in the reconstruction and studies of biological networks. These data range from the measure of simple interactions which would allow the reconstruction of a static architecture, to measures of level of abundance of various cell constituents. The latter allow a reverse engineering procedure to reconstruct the architecture in a more or less static manner, depending on the data.

Amongst many data, mRNA levels of abundance play an important role. The mRNA abundance is a link between gene networks and protein networks in the sense that it reflects the genes expression and the regulation of the system. In that respect, mRNA level of abundance is a good mirror of cell dynamics.

A technique used to measure mRNA abundance is that of microarrays. It is very popular and measures large scale gene expression. The principle of this technique, shown in figure 1.10, is to hybridise cDNA to an array of DNA probes. The cDNAs are synthesised from mRNAs which are extracted from an experimental sample and from a reference sample. Those cDNAs are then labelled separately with two different mark-

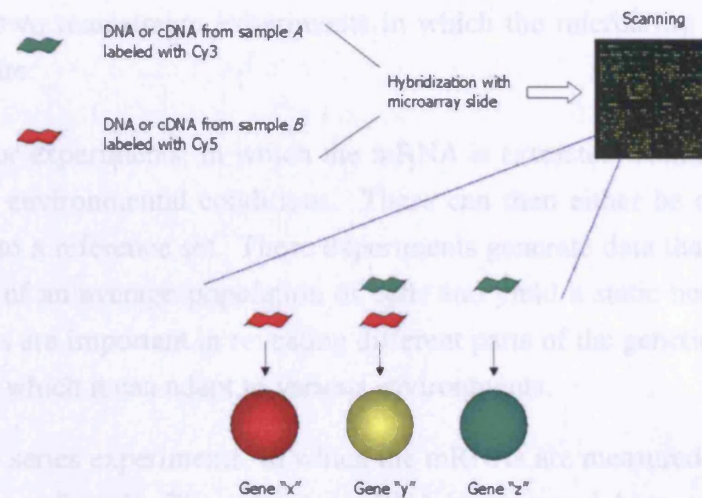


FIGURE 1.10. Principle of two-colour microarray technique.

ers whether they belong to one sample or the other sample, which are typically Cy5, a red marker, and Cy3, a green marker. Both labelled cDNAs are put to hybridise on an array where DNA probes have been spotted. This hybridisation will eventually occur if there is a labelled cDNA fragment complementary to a DNA probe. The intensity of each marker is then measured on the spots and the relative expression of the corresponding gene calculated, which is related to the colour of a spot. Then, a red spot means that the corresponding gene is over-expressed in the experimental sample or, conversely, under-expressed in the reference sample. On the contrary, a green spot means that the corresponding gene is under-expressed in the experimental sample or, conversely, over-expressed in the reference sample. A yellow spot means that the corresponding gene is equally expressed in both samples.

The microarray technique does not give a quantitative value of mRNA in a cell but only the quantitative relative expression of genes. This technique allows work on a large scale which is a clear advantage when studying a whole system. In practice, there are many considerations relative to technical issues which require the data to be treated before exploiting them. Those issues are mainly related to the measure of intensity of the spots, whether the spots are distinguishable from the background, whether the intensity is homogeneous between spots, etc, (Eisen & Brown, 1999, Nadon & Shoemaker, 2002). Furthermore, a comparative study showed, for example, that results may diverge between commercial microarrays from different manufacturers (Kothapalli et al. 2002, Tan et al. 2003) as well as between custom-made microarrays (Järvinen et al. 2004), or even

between laboratories (Irizarry et al. 2005).

There are two mainstream experiments in which the microarray technique can be used for, that are:

- (i) The factor experiments, in which the mRNA is extracted from cells grown under different environmental conditions. These can then either be compared to each other or to a reference set. These experiments generate data that are the gene expression of an average population of cells and yield a static network. These experiments are important in revealing different parts of the genetic network and the extent to which it can adapt to various environments.
- (ii) The time series experiments, in which the mRNAs are measured at different times during the cell cycle. The extracted mRNA is compared, here, to a reference population. Those experiments involve a synchronisation step of the cell which is a non-trivial operation. From these time-series data, it is possible to extract dynamical information and hence to generate a regulatory network.

The synchronisation of a cell population is a crucial operation in obtaining exploitable time series data. Synchronisation methods include chemical, physical or thermal means, which block the cells of a population at a given stage of their cell cycle. Once all the cells are blocked, the blocking agent is removed so that the cells can carry on with their cycle but they are now synchronised. This synchronisation can last over a few cycles (Spellman et al. 1998). The fact that such synchronisation may not be achievable, as well as its relevance has been the subject of a vigorous debate (Cooper & Shedden, 2003), (Cooper 2004a), (Spellman & Sherlock, 2004b), (Cooper 2004b) and (Spellman & Sherlock, 2004a). However, without entering this debate, we believe that time series data obtained from various synchronisation methods are of great interest in studying the dynamics of genetic networks.

As for the fact that results from various microarray platforms may differ, we acknowledge that the synchronisation may not be perfect and can perturb the cell cycle. Nevertheless, such data obtained from *S. cerevisiae* and *E. coli* are used later in chapter 5, when looking at general characteristics of real networks.

### 1.4.3 Summary

The trend in terms of architecture and topology in the biological networks mentioned above is to power-law distributions of the degree of connectivity as well as a small-world

network property.

On a more general note, even though these results are widely accepted, some caution is to be observed because different ways of building network can lead to various results (Arita 2004), and because of the enthusiasm in generalising conclusions (Arita 2004, Keller 2005).

## 1.5 Properties associated to network

We have so far described properties of parameters associated to the networks. We will now describe some properties related to the functionality of the network which shows how the architecture and a range of parameters, are important in determining those properties. We detail here some of them for their relevance to the architectural or dynamical part of the network we study.

### 1.5.1 Robustness and tolerance

The robustness is defined as the capacity of a network to resist ‘attacks’ while the tolerance is the capacity to be tolerant to ‘errors’. Those attacks and errors take the form of the removal of nodes either by targeting certain categories or by a random process, respectively. This notion of robustness and tolerance is important for the functionality of a network since it is expected that a network that can maintain its structural integrity is more likely to maintain its functions.

Characterising the structural robustness of a network consists in determining for what fraction of removed nodes the system falls apart. This breakdown process is measured by the size of the giant component of the network as well as the size of the components that are not part of the giant component.

It has been shown that different degree distribution (Albert, Jeong & Barabási, 2000), variation of assortativity (Newman 2002, Xulvi-Brunet & Sokolov, 2005) or degree of clustering (Grondin & Raine, 2005) affect the robustness of networks.

The behaviour of networks under attacks or errors is very different. Under attack, which in the present cases are directed to highly connected nodes, there is a critical value of the fraction of nodes removed at which the giant component breaks down into smaller components. This critical value is smaller for the Barabási and Albert networks than for the random ones, showing then a greater resilience of the latter. Similarly, for a network

with a given degree distribution, the resilience is greater for those networks with a smaller clustering coefficient. Newman (2002) argues that the robustness to node removal is related to the assortative behaviour of the network. That is assortative networks present more resilience than disassortative ones.

The networks under error show a greater resilience than that under attack as the breakdown occurs at a larger fraction of removed nodes. The effect of different clustering coefficient is also visible, with networks with higher clustering coefficient being less robust (Grondin & Raine, 2005). Note that the effect of removal of nodes at random is similar to an ‘erosion’ phenomenon as the size of the giant component decreases almost linearly.

### 1.5.2 The spread of information, diseases, . . .

Another aspect of the functionality related to the architecture is the spread of information, perturbations, diseases, etc, on networks. It has been shown that the spreading time in favourable conditions was similar to the characteristic path length of the network (Watts & Strogatz, 1998). Networks with a small characteristic path length, including small worlds are then expected to spread perturbations faster. This has been suggested as one of the possible advantages for metabolic networks to be small-world as it would help to minimise the transition time of metabolic states (Wagner & Fell, 2001).

### 1.5.3 Structural motifs

At a more detailed level, specific architectural pattern of connections, or motifs, that are statistically significant compared to random networks can be found, in a directed graph. For instance, in the transcriptional regulation network of the yeast *S. cerevisiae* and of the bacterium *E. coli* the three node motif ‘feed-forward’ loop, shown on figure 1.11, appears more than 10 times the standard deviation than it would in a random network (Milo et al. 2002). A four-node motif called ‘bi-fan’ is also present in significant numbers in those networks.

Even though we mentioned earlier the fact that we were interested beyond this level of detail and despite the fact this will not be treated in this thesis, those structural motifs may somehow be related to our studies for the two following reasons. First, we have seen that the clustering coefficient is a statistical parameter that gives information on the local structure. Similarly, it would be interesting to see whether those motifs could be described in term of a statistical parameter, and if so, how it could be integrated in our



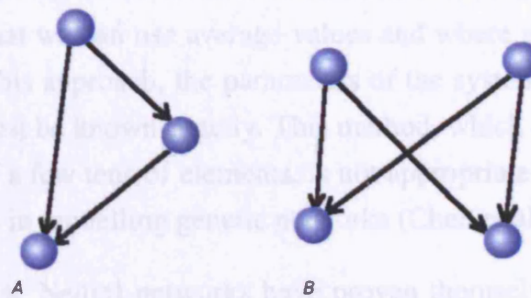


FIGURE 1.11. Local structures. (A) the feed-forward loop motif and (B) the bi-fan motif.

models. The second aspect is related to the dynamics as it is shown that specific motifs play a role in kinetic response to changes in the network (Shen-Orr et al. 2002). This kinetic response may or not affect the overall behaviour of the system.

## 1.6 Various approaches in modelling the dynamics of (biological) networks

We have so far described the architecture and properties of real or model networks. If we have mentioned that some networks could be dynamical, we have not mentioned yet how. There are actually several possibilities to model the dynamics of networks. However, the methods used will be more specifically dependent on the the system we are trying to model and the goal set with this model.

In chapters 4, 5 and 6, we will work with dynamical networks of a few thousand nodes at the most. We therefore need to choose which dynamical framework will be the most appropriate so that it gives a good insight on the behaviour of the system and it is practical to handle while realistic. We present here some of the possible ways to model dynamical biological networks.

- **Differential equation systems:** For macroscopic systems we can model the populations of interacting elements (protein, genes, etc.) as continuous variables changing continuously in time, hence as a system of coupled differential equations. The main problem is to have a set of equations that is easy to handle and that gives



results that can be interpreted biologically. The approach is appropriate for systems with a relatively small number of constituents, each of which has sufficient abundance so that we can use average values and where stochastic effects can be ignored. With this approach, the parameters of the system, such as reaction rates for example, must be known exactly. This method, which can be used typically to treat systems of a few tens of elements, is not appropriate for our purpose though it has been used in modelling genetic networks (Chen et al. 2004).

- **Neural networks:** Neural networks have proven themselves very efficient in allowing predictions from complex data, for example. However, when applied for example to genetic networks it is not clear how to extract physical or biological information from the results. For that reason this does not seem a good choice.
- **Boolean networks:** The characteristic of boolean networks is that a node is defined by its two states that are labelled, arbitrarily, as ON or OFF. In these networks, the nodes are connected to each others according to a specified architecture and the dynamics is provided by boolean functions implemented by the nodes. A benchmark boolean network in system biology is the NK-Boolean network introduced by Kauffman (1993). This model, which will be detailed in chapter 4, bears similarities to the networks we have designed to study genetic regulatory networks.

## 1.7 Entropy and theory of information

We introduce now briefly the information theory that to be used later in chapter 6.

Information theory has been developed around the idea of the use of noisy channels and encoding of messages in such a way that it can be received and decoded with a reliability approaching 100 per cent. This theory, introduced by Shannon, (Shannon & Weaver, 1949) as a measure of uncertainty of the outcome of an event. Consider an event  $X$  that can take a finite number of states  $N$  whose probabilities are  $p_1, \dots, p_N$  with  $\sum_i p_i = 1$ . The Shannon entropy of such an event,  $H(X)$ , is defined as

$$H(X) = - \sum_i^N p_i \log p_i,$$

A property of this function is that the uncertainty reaches a maximum for  $p_i = 1/N$ . This correlates to the Laplace's principle of 'insufficient reason' that states that in the absence of any information, the occurrences of an event must be equiprobable.

Similarly, the joint uncertainty of two events  $X$  and  $Y$  is defined as

$$H(X, Y) = \sum_{ij} p(i, j) \log p(i, j), \quad (1.4)$$

while

$$H(X) = \sum_{ij} p(i, j) \log p(i)$$

and

$$H(Y) = \sum_{ij} p(i, j) \log p(j),$$

where  $p(i, j)$  is the joint probability. In the case of the existence of correlations between events  $X$  and  $Y$ , the joint probability is expressed as

$$p(i, j) = p(i|j)p(j) = p(j|i)p(i),$$

where  $p(i|j)$  is the conditional probability of having  $i$  knowing  $j$  and  $p(j|i)$  the conditional probability of having  $j$  knowing  $i$ . In this case, the joint entropy is expressed as

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X),$$

where  $H(X|Y) = \sum_{i,j} p(i, j) \log p(i|j)$  and  $H(Y|X) = \sum_{i,j} p(i, j) \log p(j|i)$  are the conditional entropies corresponding to the uncertainty measure of  $X$  knowing  $Y$  and of  $Y$  knowing  $X$ , respectively. Conversely, if there are no correlations between the two events that is they are independent, we have  $p(i, j) = p(i)p(j)$ , hence  $H(X|Y) = H(X)$  and  $H(Y|X) = H(Y)$ . In this specific case, Eq. 1.4 reaches a maximum so that,

$$H(X, Y) \leq H(X) + H(Y).$$

This is an important results as it shows that if there are correlations between events  $X$  and  $Y$ , the joint entropy will be less than the sum of the entropy of the individual events.

The mutual entropy, also called information (Adami 1999) or correlation entropy, between 2 sets of random variables  $X$  and  $Y$  is defined as

$$M(X, Y) = M(Y, X) = H(X) - H(X|Y),$$

with  $M(X, Y) > 0$ , the amount of information that  $X$  and  $Y$  contain about each other.  $M = 0$  implies  $X$  and  $Y$  are independent. It corresponds to the difference between the entropy if there were no correlations and its actual entropy. The mutual entropy expresses

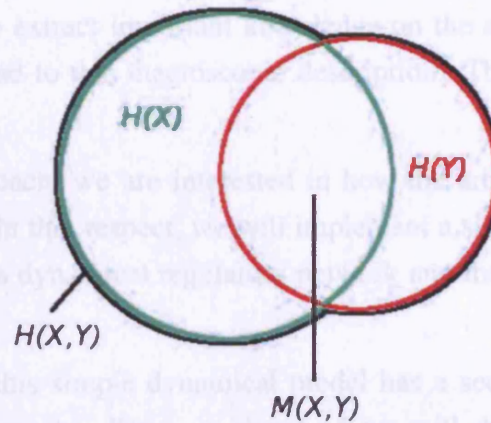


FIGURE 1.12. Venn diagram representing the mutual information,  $M(X, Y)$ , shared by both entropies  $H(X)$  and  $H(Y)$ .

then the information generated by the system. Finally, the joint entropy can be written as

$$H(X, Y) = H(X) + H(Y) - M(X, Y). \quad (1.5)$$

Thus, if correlation exists between  $X$  and  $Y$ , then by the knowledge of one of the events it is possible to extract knowledge about the other.

The concept is illustrated figure 1.12 by a Venn diagram where the intersection of the circles represents the mutual entropy and the envelope of both circles the joint entropy.

## 1.8 Outline of the thesis

In this thesis, we will study networks under three complementary approaches:

The first approach concerns the generation of networks. As mentioned earlier, the intuitive methods to construct networks fail to generate an understanding of the natural constraints yielding them. In chapter 2 we will develop a method to generate networks of any given degree distribution from a thermodynamical approach which allows a statistical treatment of the system. We will show, explicitly, that with this single conceptual method it is possible to generate networks with various degree distribution.

We expect from the thermodynamical approach applied to networks to extract the macroscopical variable that describe them. However doing this requires a deeper under-

macroscopical variable that describe them. However doing this requires a deeper understanding of the relation of the nodes and the links to the architecture of the system. On this basis, we will try to extract important knowledge on the architecture of networks which could possibly lead to this macroscopic description. This work is presented in chapter 3.

For the second approach, we are interested in how the architecture can shape the behaviour of networks. In that respect, we will implement a simple dynamics. We will describe in chapter 4 this dynamical regulatory network and the properties that emerge from it.

The formulation of this simple dynamical model has a second application related to genetic regulatory networks. Then, in chapter 5 we will show, in a trial and error approach, how the behaviour of the network is influenced and perturbed by the addition of complementary rules and how it relates to real data. In parallel, we will look at the correlation between architecture and the behaviour of real networks.

In a third approach, presented in chapter 6, we will start from generated expression data and try to infer the corresponding networks. The point will be here to see whether the data contain information on the structure of the network and whether it is enough to reconstruct it. The method used is that of maximum entropy, which will be adapted to this purpose. This method will involve knowledge gained from results of previous chapters. Compared to other methods, the maximum entropy approach will allow us to reconstruct bigger networks more rapidly. This method is illustrated by the reconstruction of networks from simulated data.

Finally, we will summarise and conclude in chapter 7 the knowledge gained from the work presented here.

# Chapter 2

## *Networks as Constrained Thermodynamic Systems*

### 2.1 Introduction

In this chapter we shall show that a thermodynamic approach allows us to construct networks as a maximum entropy configuration of nodes subject to constraints. This is not only of technical interest since the constraint can be interpreted as a cost function, which in principle represents the cost in free energy of establishing the network. This is important as the form of the cost function implicitly contains information on the evolution of the network. We therefore expect this approach to illuminate the structure of networks with possible applications in metabolic, genetic and protein networks.

To construct a network with a given nodal distribution we need to derive the relevant constraints that will lead to the given distribution. Starting from a given arbitrary network we then numerically evolve the connections to maximise the entropy of the network, subject to constraints. The trick here is to find an evolutionary dynamics that respects the constraints. We shall show that such an approach can yield a network of any nodal distribution directly.

We first briefly introduce the maximum entropy principle as used in statistical mechanics and then show how it applies to network reconstruction. We then illustrate the approach with an application to the simple case of exponential networks and then apply it to scale-free and a related class of networks following what Mandelbrot has called the ‘simplified canonical’ distribution (Mandelbrot 1954, Vohradský & Ramsden, 2001).

## 2.2 Maximisation of the entropy subject to constraints

### 2.2.1 The entropy maximisation principle

The maximum entropy principle, introduced some 50 years ago by Jaynes (1957), makes the link between information theory and statistical mechanics. Though it is very well known, we briefly recall it here. This will provide clarity when it comes to a justification of our approach to constructing networks, while giving the clues to develop further a thermodynamics of networks.

In this approach, the entropy function is the main concept used to infer probability distributions, which arise as maxima of the entropy subject to certain constraints.

Consider discrete values  $x_i$  with corresponding probabilities  $p_i$  which are not known, but for which the expectation value of the function,  $f(x)$  is known so that

$$\langle f(x) \rangle = \sum_{i=1}^n p_i f(x_i). \quad (2.1)$$

The probabilities are subject to the condition

$$\sum_{i=1}^n p_i = 1. \quad (2.2)$$

Solving exactly the probability distribution would require a set of  $n$  independent equations. However, this cannot be done here since the two expressions above provide the only known information on the system. As guessing the missing information would be inappropriate, the problem is reduced to that of finding probability assignments which avoid bias while agreeing with whatever information is given.

The solution is provided by information theory which provides the unique and unambiguous criterion for the ‘amount of uncertainty’ represented by a discrete probability distribution. This positive quantity, which increases with increasing uncertainty, is

$$H(p_1 \dots p_n) = -K \sum_i p_i \ln p_i, \quad (2.3)$$

where  $K$  is a positive constant (Shannon & Weaver, 1949). This expression corresponds to the expression of the entropy in statistical mechanics.

Then, the problem of inferences on the basis of partial information is solved by the probability distribution whose entropy is maximum subject to whatever is known. This

is the only unbiased assignment possible and using another would be an arbitrary assumption. The maximisation of equation 2.3 subject to constraints 2.1 and 2.2 requires the introduction of Lagrangian multipliers  $\lambda$  and  $\mu$ , such that

$$\frac{\partial}{\partial p_i} (H + \lambda p_i + \mu p_i f(x_i)) = 0,$$

which results in

$$p_i = e^{-\lambda - \mu f(x_i)}.$$

The Lagrangian multipliers are then identified using Eqs. 2.1 and 2.2, so that

$$\langle f(x) \rangle = -\frac{\partial}{\partial \mu} \ln Z(\mu)$$

and

$$\lambda = \ln Z(\mu),$$

where  $Z(\mu)$  is the partition function, so that

$$Z(\mu) = \sum_i e^{-\mu f(x_i)}.$$

In statistical mechanics, this approach allows the identification of macroscopic parameters such as temperature and free energy (Schrödinger 1952). In particular, all thermodynamical quantities can be calculated from the partition function and the entire macroscopic behaviour of matter at equilibrium can then be predicted by its evaluation.

A property of this method is that no possibility is ignored as positive weight is assigned to every situation that is not absolutely excluded by the given information.

### 2.2.2 Application to network construction

We apply the maximum entropy principle to deduce networks of a given degree distribution. Note that if our aim were simply to obtain a network exhibiting a specified degree distribution of nodes  $q_r$  of each degree  $r$  we could proceed as follows. Take any network as starting point (say a random one) with a distribution  $p_r = p_r^0$  of node degrees. Furthermore take any cost function of a form that approaches a maximum as  $p_r \rightarrow q_r$ , for example  $-\sum_r (p_r - q_r)^2$  or the relative entropy  $-\sum_r p_r \log(p_r/q_r)$ , and rewire the network, changing  $p_r$ , in such a way that the cost function is maximised. This is, of course, a straight forward approach if we are seeking only a specific nodal distribution.

However, this is not the most relevant for our purpose since our aim is to express the network construction in thermodynamic terms.

The problem is here the reverse of the one presented above in section 2.2.1. In the present case, we know the probability distribution of the network we are aiming at but not the constraints associated to obtain it. The problem is then to find the constraints that correspond to a given degree distribution.

Consider a system of  $N$  nodes with  $p_r$  the probability to find a node of degree  $r$ . The point entropy of the network,  $\Omega$ , associated to the nodal distribution is given by Eq. 2.3, that is (up to a constant)

$$\Omega = - \sum_r p_r \log p_r, \quad (2.4)$$

with

$$p_r = \frac{n_r}{N},$$

and where  $n_r$  is the number of nodes of degree  $r$ . Suppose  $C(p_r)$  is the constraint, yet unknown, associated with the system. The maximisation of the entropy such that  $C(p_r)$  remains constant is solved by introducing the Lagrangian multiplier,  $\beta$ . The expression to maximise is then

$$\Omega - \beta C(p_r). \quad (2.5)$$

One can verify that this is achieved if

$$C(p_r) = -\beta^{-1} \sum_r p_r \log q_r, \quad (2.6)$$

with  $\sum \delta p_r = 0$  where  $p_r$  are the probabilities to be ‘found’, and  $p_r = q_r$  is the required solution. Thus, from expression 2.6 and knowing the probability distribution we can directly deduce the appropriate constraint. In other language, the cost function mentioned earlier is  $-\Omega + \beta C$ . (To avoid confusion note that the point we are making here is not the form of the result of Eq. 2.6, which is in any case essentially just the relative entropy again, but that this form allows us to use the standard thermodynamical methods).

In this context, the distribution of probability, dependent on the constraint is of the form

$$q_r \propto e^{-\beta \frac{\partial C}{\partial p_r}}. \quad (2.7)$$

Finally, the form of the partition function will then be, up to a constant,

$$Z(\beta) = \sum_r e^{-\beta C(p_r)},$$



or, generalised to many constraint

$$Z(\beta) = \sum_{i,r} e^{-\beta_i C_i(p_r)}.$$

Thus, for any degree distribution of a network, the approach shows there exist an associated constraint and a corresponding partition function. This method needs now to be turned into a suitable algorithm.

## 2.3 Application to the construction of exponential networks

### 2.3.1 General implementation of the method

In order to justify later a thermodynamic interpretation of network properties, we need to respect the following required conditions when applying the method described above:

- (i). The size of the network is constant.
- (ii). The entropy subject to the constraint needs to be maximised to find the most probable nodal distribution associated to the given constraint.
- (iii). The constraint needs to be maintained constant so as to satisfy the use of the Lagrangian multiplier  $\beta$ .

The crux here is to show that the last condition can be satisfied, that is that one can find a mechanism that preserves any given constraint. In particular, we show that it can be satisfied exactly when generating an exponential network.

The networks are generated in a non-growing context, that is the number of nodes,  $N$ , remains constant over the evolution process. The starting network is arbitrarily chosen to be a random network (described in section 1.3.1), although any other class could be envisaged. The evolution from one network to another implies alteration of the degree of connectivity of the nodes, which can be achieved by mechanisms including rewiring, addition or removal of links. The way the modifications are carried out is an important step, since the variation of the degree of connectivity influences directly the constraint (Eq. 2.6). The mechanism is described in the next section.

Once a modification mechanism is defined we apply it to the network and monitor how the entropy varies as a consequence. If it is increased, we accept the new network and continue. If it is not, then we accept the new configuration with probability

$$e^{-\frac{\Delta(\Omega - \beta C)}{\alpha}}, \quad (2.8)$$

where  $\alpha$  is a constant set to improve the convergence of the process. The mechanism is continued until the constrained entropy is maximised.

Note that the similarity here to the Metropolis algorithm (Metropolis et al. 1953) is superficial only. We are not suggesting that the steady state properties of the network are determined by an average over this ensemble. Rather the approach is intended to prevent the network settling into a configuration with a shallow local maximum of entropy. Even so, the solution is not expected to be unique in those cases where the distribution of nodal degrees does not completely characterise the network. (For example, various different networks with the same scale-free node distribution are known (Krapivsky & Redner, 2001).)

### 2.3.2 Mechanism for exponential network

We call exponential networks those networks for which the degrees of connectivity are distributed as

$$q_r \propto e^{-\beta r}.$$

Comparing with Eq. 2.7 gives the required constraint,

$$C(p_r) = \sum r p_r. \quad (2.9)$$

This simplifies considerably the modification mechanism since  $\sum r p_r = \bar{k}$ . The condition to maintain the constraint constant reduces then into the keeping degree of connectivity constant, that is to say, to conserve the number of links. Maintaining the number of links is relatively straightforward using just a rewiring process (figure 2.1), that is:

- (i). Randomly pick a node  $i$ .
- (ii). Randomly choose a node, say  $j$ , of degree  $r$ , that is connected to  $i$  and disconnect it from  $j$ . On removing one connection from node  $j$ , the constraint  $NC$  varies by

$$\Delta(Np_r) = -1,$$

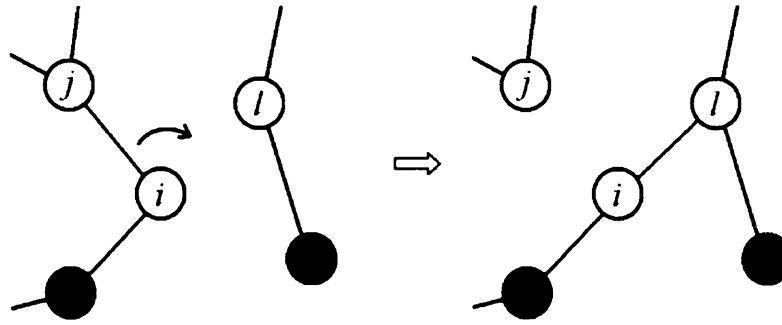


FIGURE 2.1. Simple rewiring mechanism from node  $j$  to node  $l$  where node  $i$  is the central node.

for the loss of a node of degree  $r$ , and by

$$\Delta(Np_{r-1}) = +1.$$

for the gain a node of degree  $r - 1$ . The overall variation of the constraint for the lost of a connection at a node of degree  $r$ ,  $\Delta_-$ , is then

$$\Delta_- = r\Delta(Np_r) + (r - 1)\Delta(Np_{r-1}) = -1$$

- (iii). Randomly pick a node, say  $l$  of degree  $s$ , that is not already connected to node  $i$ .
- (iv). Connect node  $i$  to  $l$ . On connecting node  $l$ , the system loses a node of degree  $s$  and gains a node of degree  $s + 1$ . As above, the overall variation of the constraint for the gain of a connection at a node of degree  $s$ ,  $\Delta_+$ , is

$$\Delta_+ = r\Delta(Np_s) + (s + 1)\Delta(Np_{s+1}) = +1$$

Hence the total variation for the rewiring process is

$$N\Delta C = \Delta_- + \Delta_+ = 0$$

and the constraint is maintain constant exactly. Note that here, the constraint is directly controlled by the rewiring mechanism and that only the entropy needs to be maximised.

### 2.3.3 Numerical simulation

We make the following general remarks and requirement on the choice of the parameters for the numerical simulation.

- i. The only adjustable parameter of importance is  $\alpha$ , in Eq. 2.8, which determines how a new configuration is accepted although it decreases the entropy. Its appropriate value for the construction of the expected network is unknown beforehand. It needs then to be determined by varying it over a wide range of values.
- ii. The number of nodes is chosen to give networks large enough to represent an infinite network over a wide enough range of node classes, yet to still give a reasonable computing time.
- iii. The mean degree of connectivity of the original network is not here an important parameter. The only requirement is for the starting network not to be disconnected. This condition will be maintained throughout the simulation.

We proceed to the simulation starting from a random network of 5000 nodes that is generated according to the method given in section 1.3.1. The mean connectivity is fixed at  $\bar{k} = 10$  which ensures a connected network (section 1.3.1). In the simulations, the rewirings are carried out until the value of the entropy, calculated according Eq. 2.4, reaches a plateau.

We give in figure 2.2 the variation of the entropy for various values of  $\alpha$  as the random network evolves towards the exponential one. The constraint is not shown since in this case it is kept constant by construction. The results in figure 2.2 show that the entropy increases as the rewiring proceeds for low values of  $\alpha$  until it reaches a plateau. At that stage the network is expected to have converged and the structure should not be affected by further rewirings. The figure shows also that the values at which the entropy plateaus depends upon  $\alpha$ , with the highest value reached for  $\alpha \lesssim 10^{-6}$ . Note that  $\alpha \rightarrow 0$  is also effective but implies that only the modifications that increase the entropy are accepted.

We show in figure 2.3 the degree distribution of the nodes obtained at maximum entropy for values of  $\alpha$  used in figure 2.2. It shows that the degree distribution of the network for  $\alpha = 10^{-6}$  is exponential. This result validates the method and the approach to generate exponential networks. The figure demonstrates also the importance of a correct  $\alpha$ . For values of  $\alpha$  greater than  $10^{-6}$  the degree distribution of the networks departs from an exponential towards the distribution of a random network. The convergence of

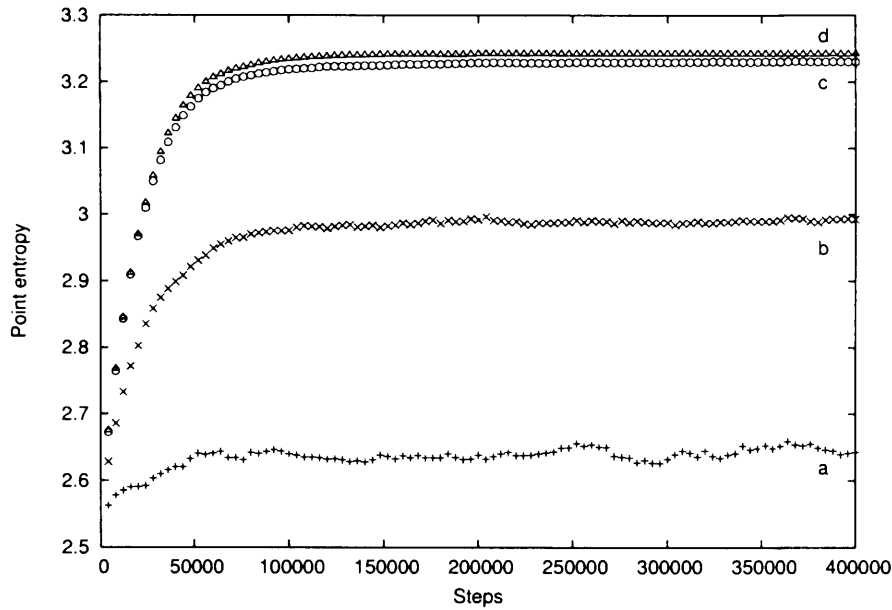


FIGURE 2.2. Variation of the entropy of the network during the maximisation process. Each curve corresponds to the entropy at different values of  $\alpha$ , where  $\alpha = 10^{-3}$  for (a),  $\alpha = 10^{-4}$  for (b),  $\alpha = 10^{-5}$  for (c) and  $\alpha = 10^{-6}$  for (d). The smaller  $\alpha$  the higher the value that the entropy converges to.

the entropy towards a plateau is not enough on its own as the plateau has to reach the highest possible value.

The distribution of the connection probabilities between nodes of degree  $r$  and  $s$  is shown in figure 2.4. We did not provide any constraint on the way the nodes should connect to each other depending on their degrees. Note that there is apparently a large probability for nodes of high degree to be connected together. This arises because of the way the figure has been constructed and results simply because the high degree nodes have more connections. Most of the connections obviously involve low degree nodes because those of high degree are exponentially small in number.

## 2.4 Generalisation to more complex networks

### 2.4.1 Alteration of the mechanism

For exponential networks the procedure to maintain the constraint constant is straightforward, but for other networks the constraint function is somewhat more complicated.

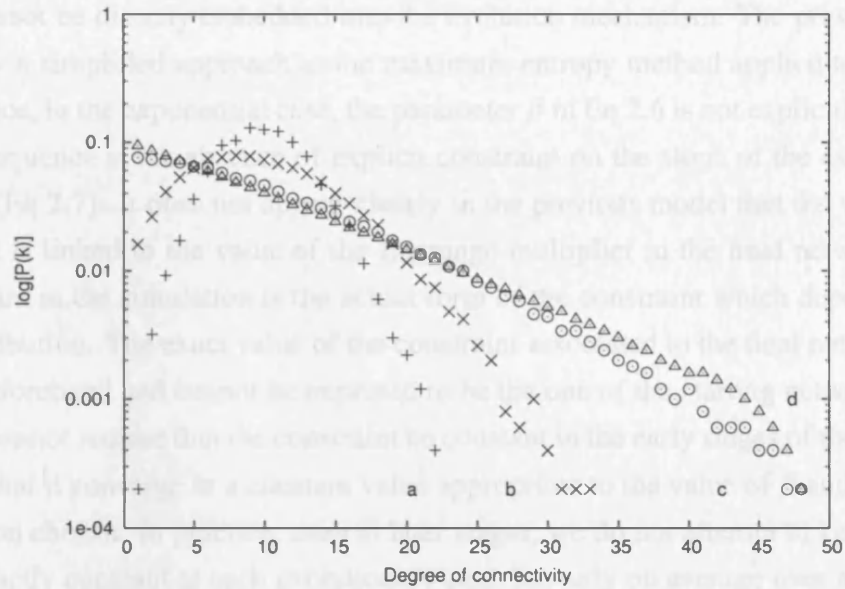


FIGURE 2.3. The degree distribution of the nodes for various values of  $\alpha$ . The distribution (a) corresponds to the degree distribution of the starting random network. Distributions (b), (c) and (d) give the degree distribution of the networks at the end of the maximisation process for  $\alpha = 10^{-4}$ ,  $10^{-5}$  and  $10^{-6}$ , respectively. The distribution (d) decays exponentially.

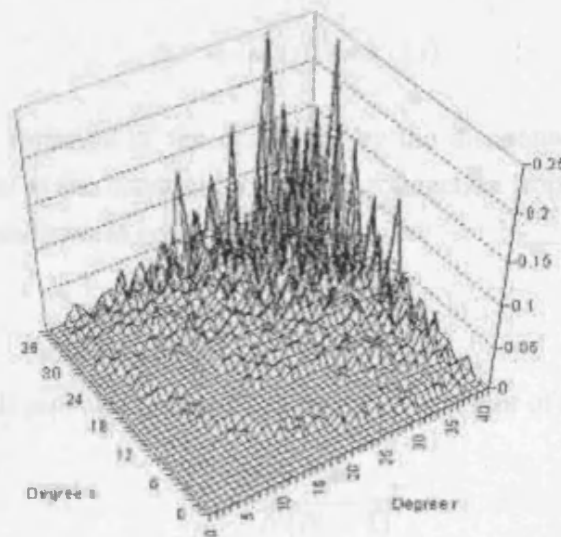


FIGURE 2.4. Distribution of the connection probability in an exponentially connected network between nodes of degree  $r$  and  $s$ .

It involves the modification of the previous algorithm since the maintenance of the constraint cannot be directly embedded into the evolution mechanism. The previous model is actually a simplified approach to the maximum entropy method applied to networks. For instance, in the exponential case, the parameter  $\beta$  in Eq 2.6 is not explicitly included. The consequence is the absence of explicit constraint on the slope of the expected distribution (Eq 2.7). It does not appear clearly in the previous model that the value of the constraint is linked to the value of the Lagrange multiplier in the final network. What is important in the simulation is the actual form of the constraint which depends on the final distribution. The exact value of the constraint associated to the final network is not known beforehand and cannot be expected to be the one of the starting network. Therefore, we cannot require that the constraint be constant in the early stages of the iterations, but only that it converge to a constant value appropriate to the value of  $\beta$  and the degree distribution chosen. In practice, even at later stages, we do not attempt to keep the constraint exactly constant at each evolutionary step, but only on average over a number of steps.

The simple rewiring mechanism turns out to be inefficient and not flexible enough to maintain the constraint in general. Instead, we implement a new evolution mechanism that includes addition and removal of links as follow: A pair of nodes, say  $i, j$ , is selected randomly. If this pair is connected then the link is removed with probability  $(1 - P_{rs})$ , where  $r$  and  $s$  are the degrees of nodes  $i$  and  $j$ , respectively. As before, the removal of one connection will decrease the constraint by

$$\Delta_- = \Delta_-(r) + \Delta_-(s)$$

where  $\Delta_-(r)$  is the variation of the constraint by the disconnection from node  $i$  and  $\Delta_-(s)$  is the variation of the constraint by the disconnection from node  $j$ . The variation of the constraint for such event occurs with probability

$$Q(1 - P_{rs}),$$

where  $Q$  is the overall probability that a randomly chosen pair of nodes in the network is linked. It is given by

$$Q = \frac{2n}{N(N-1)},$$

where  $n$  is the number of links. Similarly, if the nodes  $i$  and  $j$  are not connected then we connect them with probability  $(1 - P_{rs})$ . Such event varies the constraint by  $\Delta_+$  and

occurs with probability

$$(1 - Q)P_{rs}.$$

These probabilities are introduced to exert control over the variation of the constraint. The probability  $P_{rs}$  is calculated at each stage in order to ensure that the constraint evolves toward a constant value according to the following condition:

$$Q(1 - P_{rs})\Delta_- + (1 - Q)P_{rs}\Delta_+ = 0. \quad (2.10)$$

That is

$$P_{rs} = \frac{Q\Delta_-}{Q\Delta_- - (1 - Q)\Delta_+}.$$

With this condition the constraint evolves toward a stable value. As before, modifications leading to completely disconnected nodes are not accepted and any node has to be connected to a unique network. In practice, the choice of the probabilities  $P_{rs}$  is not unique and an appropriate fixed probability for each kind of modification (removal, addition) can also lead to a constant value of the constraint.

The decision on keeping the new configuration of the network after its modification is the one indicated in section 2.3.1. In the rest of this chapter, we will implement those modifications to the construction of scale-free and Mandelbrot networks.

## 2.4.2 Application to scale-free networks

By definition the degree distribution of a scale free network follows a power law

$$q_r \propto r^{-\beta}. \quad (2.11)$$

According to expression 2.7 this implies that the constraint be set as

$$C(p_r) = \sum_r p_r \log r. \quad (2.12)$$

The expression to be maximised is consequently

$$\Omega - \beta C(p_r) = - \sum_r p_r \log p_r - \beta \sum_r p_r \log r \quad (2.13)$$

The variable  $\beta$  is this time a parameter of the simulation and corresponds to the slope of the degree distribution of the network once Eq 2.13 is maximised.

The same considerations as before apply for the values of  $N$  and  $\bar{k}$ . For the purpose of



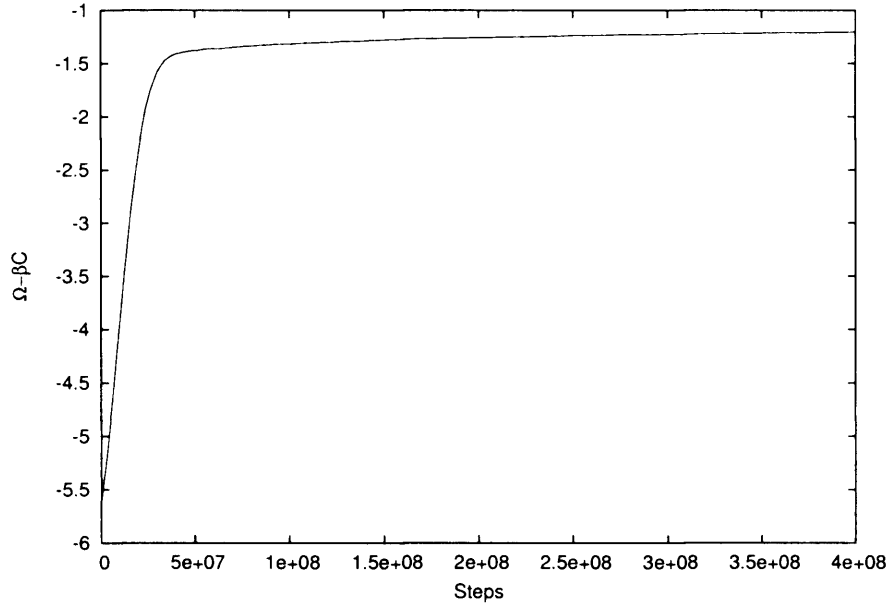


FIGURE 2.5. Variation of the constraint,  $C(p_r)$ , as defined by Eq. 2.12, reaching a plateau during the maximisation process.

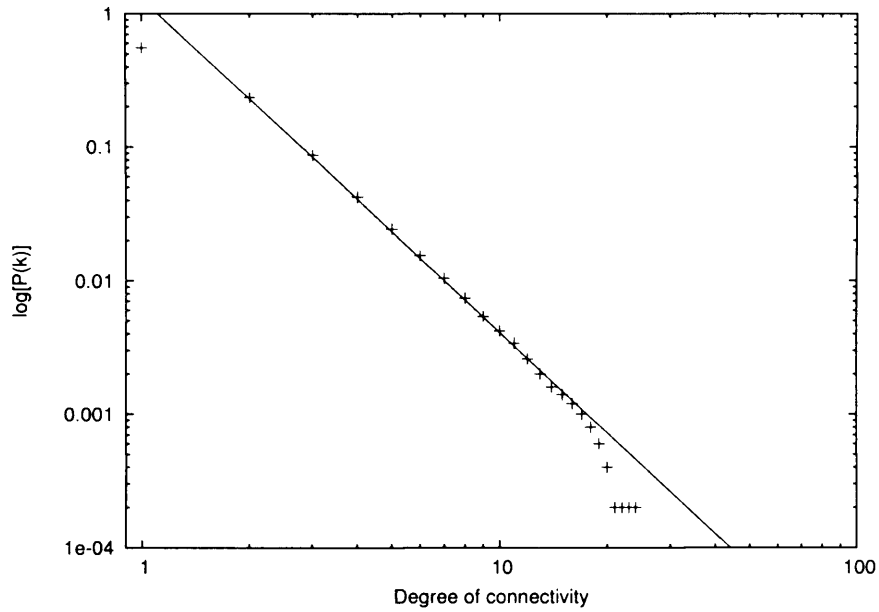


FIGURE 2.6. Degree distribution of the nodes obtained numerically (crosses) compared to the expected theoretical distribution (line) for a scale free network. The theoretical distribution is set to have  $\beta = 2.5$ .

consistency, we chose the same values in section 2.3.3 that are  $N = 5000$  and  $\bar{k} = 10.0$  for starting random networks. In the following simulation we fix  $\beta = 2.5$ , which is

chosen arbitrarily to illustrate the method.

Since the evolution mechanism has been modified the appropriate value for  $\alpha$  has to be determined as previously. It turns out, as before, that  $\alpha = 10^{-6}$  gives expected results and that  $\alpha \rightarrow 0$  is as well effective. We set  $\alpha = 10^{-6}$  for the following simulations.

Figure 2.5 shows that the value of the constraint reaches a plateau and stabilises after a certain time. Even though we do not expect the value to remain exactly constant we do not expect it to vary significantly had we pursued the maximisation further.

The degree distribution of the network at the end of the simulation is given in figure 2.6. As shown on the graph, the degree distribution follows a power law with a slope close to the expected  $\beta = 2.5$ . There is a cut-off at nodes of high connectivity (low probability), which is a reflection of the finite size of the network.

Note that until the finite size effects set in at high connectivity the method gives results that are highly reproducible with little scatter about the power law distribution. The simulation has also been tried for various values of  $\beta$  giving every time a degree distribution with a slope closed to the expected value.

### 2.4.3 Application to Mandelbrot networks

The Mandelbrot network is a generalised version of the scale free network. It is so named because its degree distribution follows the Mandelbrot law, also called the simplified canonical law (Mandelbrot 1954). This law is defined as

$$P_r \propto \frac{1}{(r + \rho)^\beta}.$$

The parameters are referred to as the diversity,  $\rho$ , and the inverse information temperature,  $\beta$ .

The new constraint  $C(r)$  is a generalisation of expression 2.11 (and consequently so is the corresponding expression for entropy) and is given by

$$C(r) = \sum_r p_r \log(r + \rho). \quad (2.14)$$

The variable  $\rho$  is then included as a parameter of the model.

The starting networks used in the following simulation are the ones previously defined in sections 2.3.3 and 2.4.2. As previously, an appropriate value of  $\alpha$  needs to be set so as to reach a maximised value of the entropy subject to the constraint. Apart from the

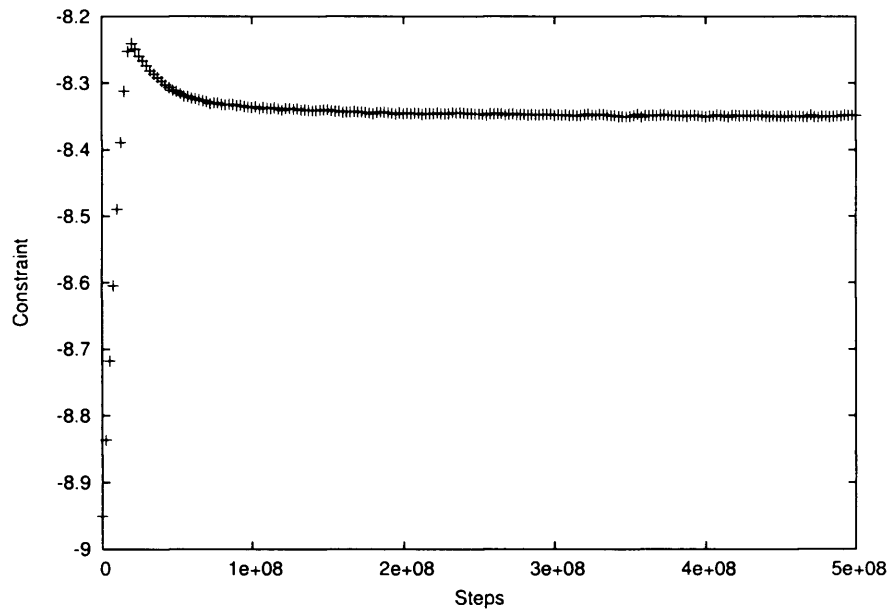


FIGURE 2.7. Variation of the constraint,  $C(P_r)$ , as defined by Eq. 2.14, reaching a plateau during the maximisation process.

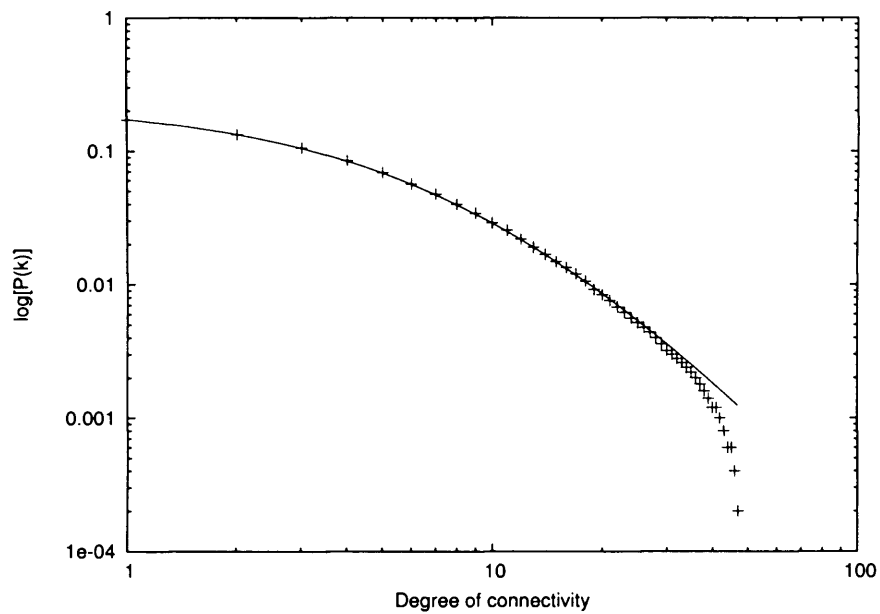


FIGURE 2.8. Degree distribution of the nodes computed numerically for a Mandelbrot network (crosses) compared to the expected theoretical one (line). The theoretical distribution is set to have  $\beta = 3.0$  and  $\rho = 10.0$ .

change of the expression of the constraint, the model is not modified. We then assume there is no need to change parameter  $\alpha$  which is then set as in the previous section to  $\alpha = 10^{-6}$ .

We present in figure 2.7 the variation of constraint obtained with the parameters  $\beta = 3$  and  $\rho = 10$ . The constraint at this stage varies around a constant value of the order of unity with an amplitude of  $\pm 2.5 \cdot 10^{-4}$  and can therefore be considered as constant.

The degree distribution of the network at the end of the simulation is given in figure 2.8. The distribution fits the simplified canonical law very well except for nodes of high degree, where again the effects of finite network size come into play.

Finally, we repeated the simulations for various values of parameters  $\beta$  and  $\rho$  which give networks in very good agreement with what was expected.

## 2.5 Summary

We have shown in this chapter that a thermodynamical approach allows us to construct networks of, virtually, any given degree distribution. This shows that the construction of networks may not rely only on intuitive mechanisms but may also involve a standard thermodynamical formalism. The obvious advantage of this approach is that it opens new ways of understanding network evolution.

This approach allows us to treat networks as thermodynamical systems. In the work presented above, the use of point entropy to infer the networks assumes we neglect the correlation between degrees of connectivity. The issue is now to see whether we can make a better thermodynamical description of the network and what knowledge to gain from this. This is presented in the next chapter.

# Chapter 3

## *Towards a thermodynamical description of networks*

### 3.1 Introduction

We have seen in the previous chapter that a thermodynamical approach enables us to generate networks of given degree distribution (at equilibrium). The method in itself is not enough to allow a complete thermodynamical description of the networks. A deeper investigation of the parameters and properties of those networks is needed if we are to extract the relevant information that will give rise to such a complete description.

Our approach allows us to think of networks in term of thermodynamics and hence in terms of the information provided by the entropy function. In such terms, the networks generated so far are not completely analogous to perfect gases, because the probability of a connection between nodes of given degrees is not independent of the degree.

In this chapter, we show first the existence of correlations in the connections of the nodes and estimate the effect of this on the network entropy which differs from random by a relatively small amount (*i.e.* that in this sense the connections are close to random). We then discuss how this thermodynamical approach might be presented in terms of intensive and extensive variables of networks. We suggest that a useful description might involve an intensive measure of ‘complexity’ of a network through what we call the  $\beta$ -complexity parameter and we give two examples. We expect this type of analysis of the networks to provide us with the macroscopic variables appropriate to life, even though this goal seems still remote.

## 3.2 The non-randomness of the connections

We have so far assumed that the entropy (disorder) of the networks under consideration is contained solely in the distribution of the connectivity of the nodes (the ‘point’ entropy). This is equivalent to assuming that the nodes are randomly connected in the sense that the probability of a link between nodes of degrees  $r$  and  $s$  is independent of  $r$  and  $s$ . This is not the case: in both scale-free and (hence) Mandelbrot networks correlations exist between nodes (Albert & Barabási, 2002). We need to show that this can be neglected in a first approximation and an analogy with a gas might be illuminating. In a perfect gas the molecules are independent so the energy (and entropy) is proportional to the number density of particles. This is analogous to a network with independent links. In an imperfect gas the particles are correlated and the energy is associated not just with the particles, but also with their interactions. This is like a network with any non-random distribution of links. The correlations contribute to the order and hence should be included in the entropy.

We show, in the following, the existence of those correlations by calculating two different and complementary entropy ratios. The first one compares the point entropy of the network to the entropy when taking into account the links, whereas the second one compares the link entropy of a network to an entropy ignoring correlations.

### 3.2.1 First approach

In this approach, we compare the value of the model entropy  $-\sum n_r \log n_r$  associated with the degree distribution of the nodes (the ‘point’ entropy) with the true statistical entropy of the network, taking into account correlations between the connections of the nodes.

Let  $n_{rs}$  be the number of nodes of degree  $r$  that are connected to nodes of degree  $s$ . The probability of finding a connection between a node of degree  $r$  and a randomly selected node of degree  $s$  is  $n_{rs}/n_s$ . If the connections were made randomly this probability is independent of the degree of the node to which the connection is made, so we should have

$$\frac{n_{rs}}{n_s} = \frac{n_{rp}}{n_p}$$

for any  $s$  and  $p$ . This is, equivalently,

$$\frac{n_{rs}}{n_{rp}} = \frac{n_s}{n_p},$$

from which we deduce that  $n_{rs} \propto n_s$ , hence, by symmetry  $n_{rs} \propto n_r n_s$  and finally, for uncorrelated networks,

$$n_{rs} = Q n_r n_s, \quad (3.1)$$

where  $Q = \bar{k}/N$  is approximately the average probability that a randomly chosen pair of nodes is connected (taking  $N(N-1) \approx N^2$  for large  $N$ ).

Let now  $S_L$  be the entropy related to the links  $n_{rs}$  such as

$$S_L = \sum_{r,s} n_{rs} \log n_{rs},$$

which, according to Eq. 3.1, becomes for uncorrelated networks

$$S_L = \sum_{r,s} Q n_r n_s \log Q n_r n_s.$$

This can then be expressed in term of the point entropy such as

$$S_L = 2NQ \sum_r n_r \log n_r + 2N^2 Q \log Q.$$

Hence the ratio

$$R_L = \frac{\frac{1}{2NQ} \sum_{r,s} n_{rs} \log n_{rs} - \frac{N}{2} \log Q}{\sum_r n_r \log n_r},$$

corresponds to the entropy contributed by the links compared to that of the nodes. This ratio, which depends on the size and mean connectivity of the network, is equal to unity for random, that is uncorrelated networks and smaller than unity for any other networks.

We show in figure 3.1 the variation of the ratio  $R_L$  for exponential networks of various mean degree of connectivity. The figure shows that for exponential networks, the ratio is not unity but slightly less, which means that, in exponential networks, the links are connected close to randomly.

### 3.2.2 Second approach

In this second approach, very similar to the first one, we compare the entropy related to the links in a network to the entropy expected in an uncorrelated network. The difference is that this time we do not consider the point entropy.

The probability  $P(s)$  of finding a connection to a node of degree  $s$ , is given by

$$P(s) = \frac{n_s k_s}{N \bar{k}}, \quad (3.2)$$

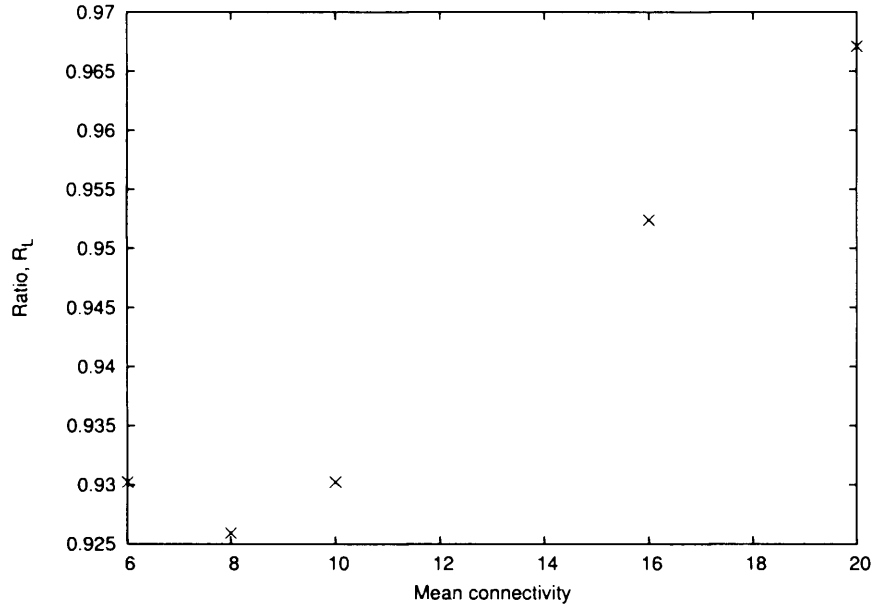


FIGURE 3.1. Ratio,  $R_L$ , of the entropy contributed by the nodes to that of the links plotted against the mean connectivity for exponential networks. This ratio is 1 for random networks.

where  $n_s$  is the number of nodes of degree  $s$  and  $k_s$  the degree of connectivity of nodes of degree  $r$ . If there are no correlations between the degrees of connectivity, then

$$P(r, s) = P(r)P(s) \quad (3.3)$$

where  $P(r, s)$  is the probability that a link joins nodes of degree  $r$  and  $s$  which is given by

$$P(r, s) = \frac{2n_{rs}}{N\bar{k}}, \quad (3.4)$$

with  $n_{rs}$  the number of links between nodes of degree  $r$  and  $s$ . Let  $Q_{rs}$  denote the probability that a pair of nodes of degree  $r$  and  $s$  are linked, so

$$Q_{rs} = \frac{2n_{rs}}{N(N-1)}.$$

Using Eqs. 3.2, 3.3 and 3.4, if the links between nodes are uncorrelated

$$Q_{rs} = \frac{2rn_rsn_s}{N^2(N-1)\bar{k}} = Q_{random},$$

We now look at the correlation in a network by computing the entropy per link rel-



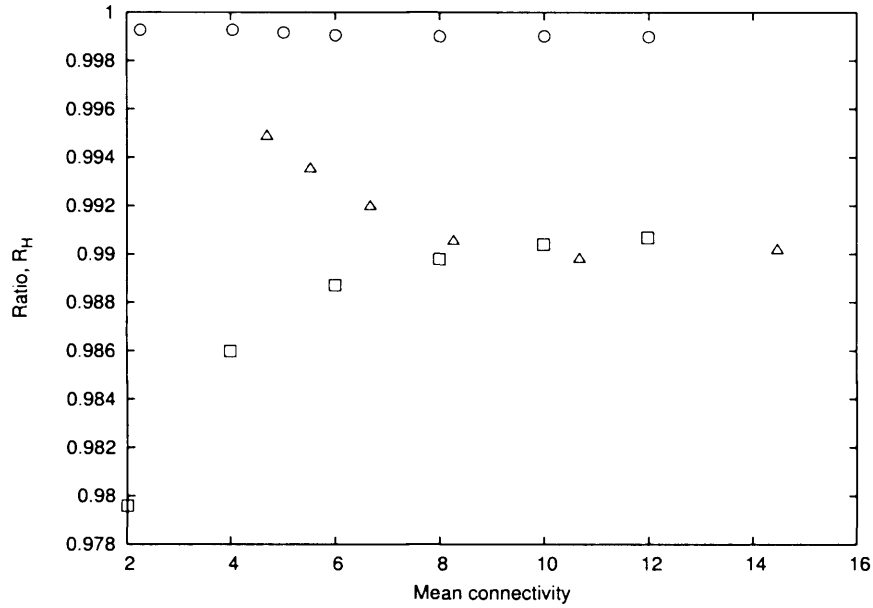


FIGURE 3.2. Ratio,  $R_H$ , of the entropy for links independent of the nodal degrees to that of the dependent ones. The circles are for random networks, the triangles are for scale-free networks obtained using the Barabási and Albert approach and the squares for the Mandelbrot networks. Each value is the average over 10 repeats.

ative to a random network. In any network a pair of nodes can either be linked, with probability  $Q_{rs}$ , or not linked with probability  $1 - Q_{rs}$ . The mean entropy per pair of nodes is therefore

$$H_L = -\frac{2}{N(N-1)} \sum_{pair} [Q_{rs} \log Q_{rs} + (1 - Q_{rs}) \log (1 - Q_{rs})].$$

We expect the entropy of a network where nodes are randomly connected to each other to be higher than that of any other network. Thus the ratio

$$R_H = \frac{-\sum_{pair} [Q_{rs} \log Q_{rs} + (1 - Q_{rs}) \log (1 - Q_{rs})]}{-\sum_{pair} [Q_{random} \log Q_{random} + (1 - Q_{random}) \log (1 - Q_{random})]} \leq 1$$

will be unity for independent linkage and less than unity for dependant links.

For a random network the numerically computed quantity  $R_H$  is close to that expected (so the ratio is close to unity) thus verifying our procedures. Figure 3.2 shows the calculation of the ratio for the Mandelbrot networks that we have constructed above. It is clear that part of the entropy resides in the non-randomness of the links.

### 3.2.3 Summary

Our conclusion is that to a first approximation the macroscopic properties of these networks can be obtained from an entropy function related to the node distribution. The fact that the entropy is not entirely contained in the node distribution means that there will be more than one local maximum of the full entropy function for a given node distribution, corresponding to different correlations (Krapivsky & Redner, 2001).

Note that our model is the opposite of Ising-like models of interacting nodes where the energy (and entropy) is associated entirely with the interactions of nodes. Our model leads to the desired connectivity distributions in chapter 2 (compare (Berg & Lassig, 2002)).

## 3.3 Intensive and Extensive variables

The cost function plays the role of an internal energy, which can be changed in two ways: by rewiring (adding heat) which will change the node distribution, or by growth of the network without changing the node distribution (doing work):

$$\delta C = -\beta^{-1} \sum \log n_r \delta n_r - \beta^{-1} \sum n_r \delta \log n_r. \quad (3.5)$$

The first term on the right is just the change in entropy of the nodes (multiplied by  $\beta^{-1}$ ). The second term on the right represents, in some sense, the cost per node associated with varying the external parameters. The external parameter might be, for example, the total number of nodes, or the total number of connections; for each choice,  $X$ , there will be an associated intensive variable,  $x$ , defined by 3.5

$$x = \left( \frac{\partial C}{\partial X} \right)_s.$$

The thermodynamic approach therefore provides a natural characterisation of the macroscopic variables associated with any network. In a sense this is the end of the story, but we would clearly like to acquire an intuitive feeling for the nature of the macroscopic variables associated with a network, just as we derive an intuitive understanding of the pressure of a gas or the magnetisation of a medium.

### 3.4 Network Complexity

In this section we consider the case for a quantity  $X$  as a measure of the complexity of a network, related to what we have called  $\beta$ -complexity (Raine & Norris, 2002).

A number of approaches to defining the complexity of networks have been proposed. In graph theory the complexity is derived from the properties of the spanning trees. Structural complexity (Crutchfield 1994) can be defined in terms of the number of parameters required to define the network. Edge complexity (Watts 1999) has been defined in terms of the variability of the second shortest path between nodes. What all of these definitions have in common, and in contrast to the notion of algorithmic complexity in computer science, is that both ordered systems and random ones have zero complexity.

We have proposed a rather different approach to network complexity, which arises from the following consideration. Another feature that random and ordered networks have in common is that the entropy per node is independent of the number of nodes. This can be paraphrased by saying that local information is sufficient to determine the large scale structure of the graphs. We want the notion of complexity to capture the extent to which this fails to be the case. Thus this aspect of the complexity of a graph is encoded in the relation between the distributions of nodes linked by one connection, two connections and so on. Random graphs and ordered graphs remain random and ordered under these ‘renormalisation’ transformations. Similarly, exponential graphs and scale-free graphs remain approximately exponential or scale-free respectively under these transformations. This suggests that the complexity of such graphs can be captured by any pair of links in this chain of connections. In other words, the complexity of networks of this form can be expressed by a single parameter. For example, we can concentrate on the first and last links in the chain, namely the clustering coefficient  $C$  and the characteristic path length  $L$  (Watts 1999).

To investigate this we consider networks from an algebraic point of view in term of their adjacency matrix,  $A$ , as presented in section 1.1. The square of such a matrix will give the number of different paths of length 2 between nodes of index  $i$  and  $j$ , the cubic matrix, the number of different paths of length 3, etc. The resulting matrix, rewritten by replacing all non-zero values with 1, and diagonal values with zero, gives the adjacency matrix of the second neighbours, third neighbours, etc.

We are interested in the diagonal terms of the product matrices. The trace of the successive products of degree  $n$  gives a hierarchy of coefficients  $\Gamma_n$ . For example, the trace of the squared adjacency matrix is related to the mean connectivity of the network

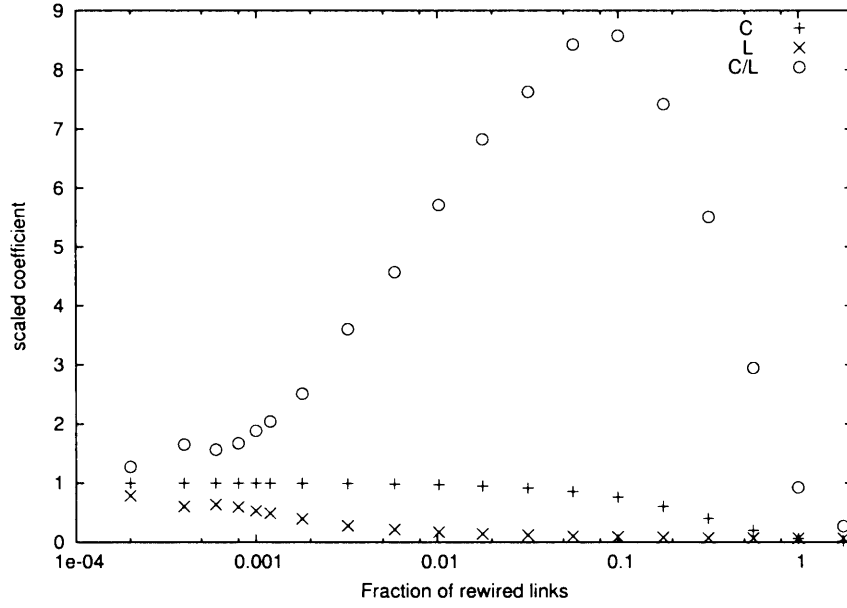


FIGURE 3.3. The ratio of clustering coefficient to average path length of the small-world network of Watts and Strogatz (Watts 1999) behaves as a complexity parameter. The coefficients are scaled to 1 for  $p = 0$ .

such as

$$\Gamma_2 = \frac{1}{2N} \sum_i a_{ii}^{[2]} = \bar{k}, \quad (3.6)$$

and the trace of the cubed adjacency matrix is related to the clustering coefficient and we have  $\Gamma_3 = C$ , according to Eq. 1.1. The number of these coefficients that provide new information on the network structure is of order of the network diameter  $L$ , defined in section 1.2.3, after which every node is connected to almost every other node. In some sense therefore this transformation illustrates the extent to which the local properties of a network are reflected in its global properties.

The degree of clustering in relation to the network diameter,  $C/L$ , is small for both random and ordered networks and larger otherwise as, for example, for the small world networks of Watts and Strogatz (Watts & Strogatz, 1998) as shown in figure 3.3. It shows the parameter  $C/L$  does have the property of a complexity parameter, which has been called the  $\beta$ -complexity of a network (Raine et al. 2003).

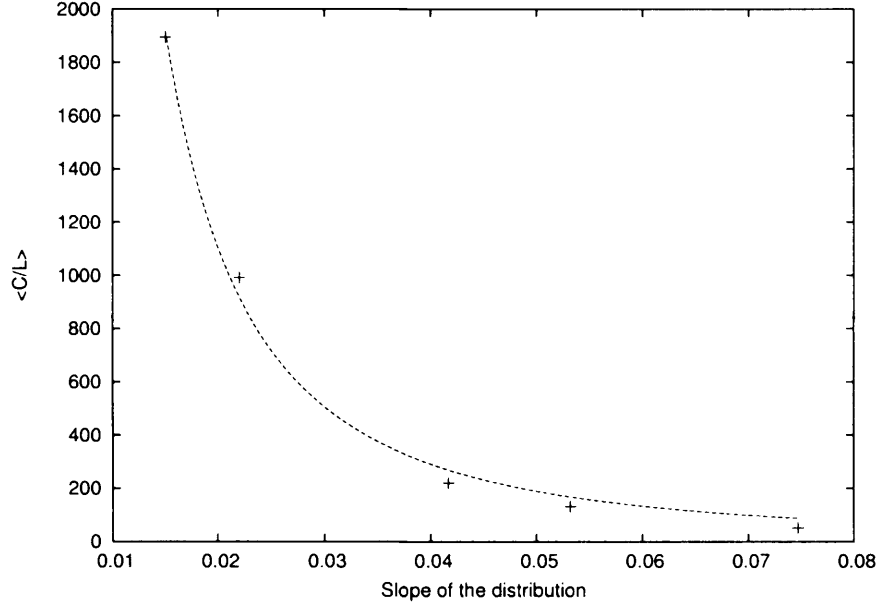


FIGURE 3.4. Plot of  $C/L$  against the slope of the degree distribution for various exponential networks. The best fit shows that  $C/L \propto \beta^{-3}$ .

### 3.4.1 Exponential network

For the exponential network we show the average value of  $C/L$  as a function of the slope of the degree distribution in figure 3.4. In terms of the parameter  $\beta$  the curve is fitted fairly closely by  $C/L \propto \beta^{-3}$ . If we interpret  $\beta$  as an inverse temperature then this says that the complexity decreases with decreasing temperature. This is to be expected since for small  $\beta$  the distribution of the degrees of the nodes spreads more with the majority of the nodes being of degree  $< 1/\beta$ . Equivalently, the ‘cost’ per node, as measured by the value of the constraint, decreases with  $\beta$ .

We now want to compare this with the thermodynamic definition of network complexity, which, according to our hypothesis above, will be of the form

$$\text{complexity} \propto \left( \frac{\partial C}{\partial X} \right)_s$$

where  $X$  is an appropriate external parameter. In view of the renormalisation steps discussed above, and the fact that the characteristic length  $L$  is of the order of the number of steps required to make a complete network, a natural choice for  $X$  is  $X = L$ , *i.e.* the complexity of a network is the change in cost per unit change in ‘volume’ (where ‘volume’ is here one-dimensional). For the exponential network this gives a complexity

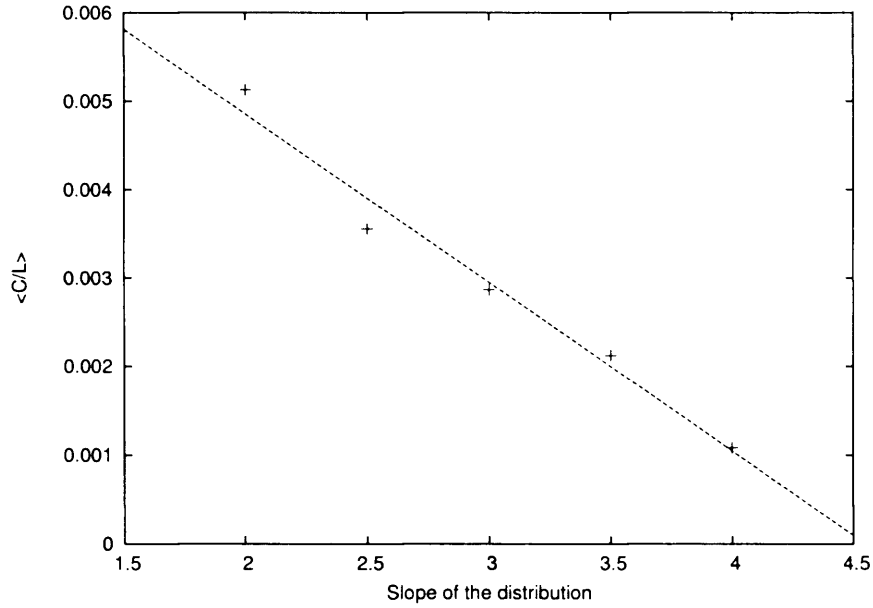


FIGURE 3.5. The value  $C/L$  plotted against the slope of the distribution for the Mandelbrot networks constructed as described in section 2.4.3.

$\propto \beta^1$ . Thus we should have

$$\beta - \text{complexity} = \beta^{-4} \left( \frac{\partial C}{\partial X} \right)_s$$

in order to get a  $\beta$ -complexity of  $\beta^{-3}$ .

### 3.4.2 Mandelbrot network

We show in figure 3.5 the run of  $\beta$ -complexity over the slope  $\beta$  of the degree distribution for the Mandelbrot networks constructed by our thermodynamic method in chapter 2. Note that steeper slopes correspond to fewer highly connected nodes and a lower value of  $C/L$ ; this is consistent with the interpretation of  $\beta$  as a measure of complexity.

### 3.4.3 Effect of the ‘renormalisation’ transformation

To investigate this further we have looked at the nodal distribution for the iterated adjacency matrix, that is of the renormalisation transformation cited above. For this, we have constructed the sequence of adjacency matrices for several classes of networks. We give

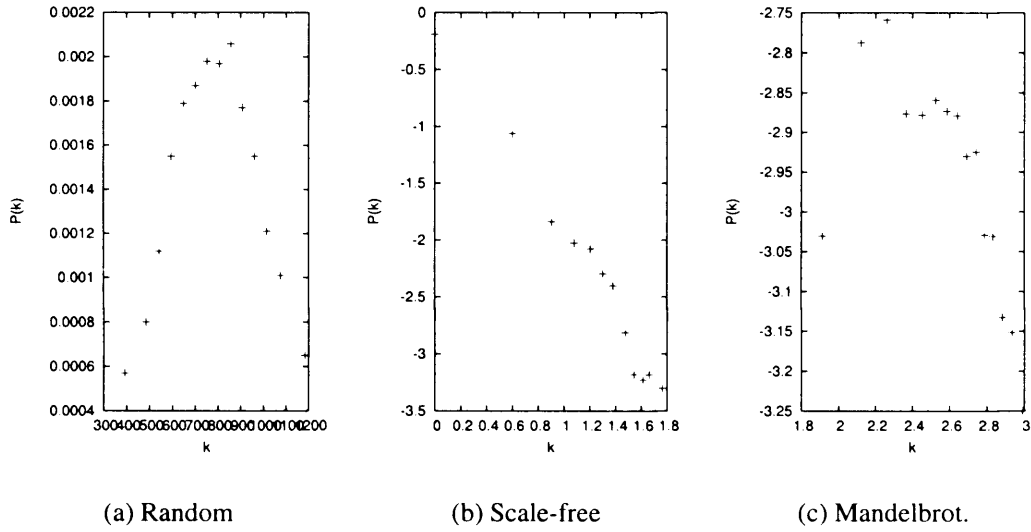


FIGURE 3.6. Degree distribution of nodes for several network classes calculated from their respective cubic adjacency matrix. (a) Random network, (b) scale free network, (c) Mandelbrot network. The noise has been reduced by binning the data.

in figure 3.6 the degree distribution function corresponding to these. For networks of a few thousand nodes the successive distributions soon become noisy, so we have taken running averages over adjacent degrees on each side of a given one.

For each class, the degree distribution functions are of approximately the same shape as the original, aside from the fact that the transformations affect the mean degree of connectivity. This denotes, from a statistical point of view, an approximate invariance of the structure of the network connectivity. Eventually this sequence must come to an end because after sufficient iterations (approximately the diameter of the network) every node is connected to every other. The  $\beta$ -complexity must therefore change along this sequence. The difficulty of seeing this change above the noise in the initial iterations is clearly related to the difficulty of constructing networks with different complexity.

The degree distribution however, does not fix the  $\beta$ -complexity of a network. This is known experimentally from the fact that the Barabási and Albert construction yields a scale-free network that is not a small world, while small world scale free networks, such as yeast co-expression networks and hierarchical networks are known and can be modelled (see van Noort et al. (2004), Ravasz & Barabasi, (2003) and references therein). This can be demonstrated in the network we constructed by varying, for example, the clustering parameter,  $C$ . This can be carried out by a simple mechanism of cross-rewiring two pairs of nodes. That is, choose node  $i$  connected to node  $j$  and node  $k$  connected to

node  $l$  with the condition there is no link between pair  $i, l$  or pair  $k, j$ . Then rewire node  $i$  to node  $l$  and node  $k$  to node  $j$ . This process ensures an alteration of the clustering coefficient without modifying the degree distribution of the network.

Random cross-rewiring is carried out over a number of steps greater than  $N\bar{k}$  on Mandelbrot networks generated according the method described in section 2.4.3. During the process, the maximum and the minimum  $C$  are recorded and the ratio  $R = C_{max}/C_{min}$  taken. Over several occurrences of that process of generation and cross-rewiring we obtain a ratio  $\langle R \rangle = 1.65 \pm 0.16$ . That is, at identical degree distribution and assuming the characteristic path length did not vary significantly during the cross-rewiring, the  $\beta$ -complexity can vary significantly. We have also witnessed large variations of clustering coefficient when constructing highly clustered networks (Grondin & Raine, 2005).

### 3.5 The equation of state

In the context of the thermodynamic approach to the construction of networks we see how the slope  $\beta$  is not only analogous to a ‘temperature’ but plays exactly the role of a Lagrange parameter multiplying the ‘energy’ constraint. It is then a matter of computation to determine the other thermodynamic functions of the state of the network from the entropy. It is in fact a general result that any constraint of the form  $\sum p_r \epsilon(r) = \text{constant}$  leads to the perfect gas law whatever the form of  $\epsilon(r)$ . We can show this as follows.

Let  $p_r$  be the probability that a node has  $r$  links and let the ‘energy’ at a node be  $\epsilon(r)$ . Then the nodal (point) free energy of a network of  $N$  nodes is given by

$$\frac{F}{N} = - \sum p_r \log p_r - \beta \sum p_r \epsilon(r) - \alpha \sum p_r.$$

Maximising the free energy as usual leads to

$$n_r = \frac{N}{Z} e^{(-\beta \epsilon(r))}, \quad (3.7)$$

where the partition function  $Z$  is defined by  $Z = \sum e^{-\beta \epsilon(r)}$ . For a perfect gas in a volume  $V$  this has the form  $zV$ . We now derive the equation of state. Define the internal energy as

$$U = \sum \epsilon(r) n_r.$$



Using 3.7 and the definition of  $Z$ , the entropy  $S = \sum p_r \log p_r$  becomes

$$S = -\frac{\partial}{\partial \beta} (\beta \log Z).$$

Thus

$$U = \frac{N}{Z} \sum \epsilon(r) e^{-\beta \epsilon(r)} = -\frac{S}{\beta} + \frac{N}{\beta} \log Z.$$

From this, using  $Z = zV$ , we can find

$$P = \left( \frac{\partial U}{\partial V} \right)_s = \frac{N}{\beta} \frac{1}{V}.$$

Thus, for the network, the crucial point is to define the volume in such a way that  $Z = zV$ . To do this, let  $M$  be the total number of links. Then

$$M = \sum r n_r = \frac{N}{Z} \sum r e^{-\beta \epsilon(r)}$$

and hence

$$Z = \frac{N}{M} \sum r e^{-\beta \epsilon(r)} = Vz,$$

if we take  $V = N/M = \bar{k}^{-1}$ . The result is independent of the form of  $\epsilon(r)$ . However, with this identification  $V \neq L$  in contrast of what we suggested above.

It is clear that this gives us the ‘perfect gas’ approximation to the equation of state of the network because, by using the node entropy, we are failing to take into account the interaction between the nodes. To obtain a deeper relationship between network entropy and thermodynamics we should have to use an entropy that includes the correlations between nodes, for example leading to a class of networks of a fixed complexity.

## 3.6 Summary

The thermodynamic approach allows us to characterise the macroscopic variables of a network, and, hence, by extension, of any system that can be described in network terms. In particular, therefore, the approach holds out the possibility of a macroscopic characterisation of living systems.

In the development of the thermodynamics of material systems in thermal equilibrium the experimental facts were discovered first and the underlying microscopic description only later. In the theory of networks we start from the microscopic viewpoint

and face the challenge of trying to derive macroscopic variables to characterise network properties in a useful way.

The fact that we now have a thermodynamic view of networks should enable us to develop such a macroscopic characterisation based on the entropy function. In networks where most of the entropy is contained in the distribution of the degrees of the nodes we expect that an approximate macroscopic description will involve a small number of global variables. We have speculated that in general the macroscopic variables will include a parameter describing in some way the complexity of the networks.

# Chapter 4

## *Dynamical network model*

### 4.1 Introduction

In the previous chapter, our approach to networks was merely structural. However, this is not sufficient when applied to real systems and in particular to the genetic regulatory network we are interested in. Focussing only on the architecture of systems ignores the relationship that may exist between architecture and function. If we are looking at constraints that shape the architecture of the system as mentioned earlier, then the function and the behaviour of the system are certainly part of them and ought to be considered.

Prior to the identification of those constraints, we need to study and understand the relationship of the architecture of networks to their function and behaviour. Indeed, the behaviour of the system is expected to influence and to be influenced by the architecture of the system. Some of the aspects of this relationship have been invoked in chapter 1 but what we want is actually to relate it to real networks.

Modelling adequately a network requires therefore a dynamics as well as an architecture, which we are going to implement. As seen previously, networks emerge as complex systems solely from the architectural point of view. Therefore, in order to draw a clear understanding from the relationship between architecture and behaviour, we want the dynamics to be simple. Furthermore, in order to be comparable to biological systems, and genetic regulatory networks in particular, we want the network to express its dynamics through some production function.

We will present in this chapter a model related to the dynamics we want to implement. We will then introduce our model and detail its properties under various changes of the parameters that define it. We shall see that our model presents a variety of behaviours of fixed point, periodic and chaotic phases. A striking result is the unexpected distribution of period, independent of the architecture. We show also the direct effect the architecture has on the distribution of abundance related to the production function associated to the system.

## 4.2 The $NK$ -Boolean networks

$NK$ -Boolean networks (Kauffman 1993) are directed networks whose essential parameters are the number of nodes,  $N$ , the number of links a node connects to,  $K$ , and a bias parameter,  $P$ , which will not be discussed here. A Boolean network is a dynamical system in which the nodes are defined by a two-state variable that takes the values 0, (off) or 1, (on). Each node is regulated by  $K$  other nodes that correspond to its inputs. The state of a node at a time  $t$  is given by a Boolean function, such as *and*, *or*, etc, and whose variables are the state of its the input nodes at time  $t - 1$ . All the nodes are updated together at each time step.

The ensemble of states at a given time defines the configuration of the network with a maximum of  $2^N$  different configurations, where  $N$  is the number of nodes. The sequence of configurations from one time to another represents the trajectory of the network in the configuration space. Even if the number of configurations becomes rapidly very large as  $N$  increases, its number remains finite for a finite network. As time goes on, a configuration will eventually appear twice and, since the network is deterministic, the trajectory will loop on itself. The network has reached an attractor with all trajectories leading to it forming its basin of attraction. A network with non-external inputs from the system is considered autonomous and the attractor in which it will fall will be dependent on the initial state of the nodes in the system.

There can be many attractors for a network and it is suggested that those many attractors could be the many cell types if the networks were those of genetic regulators but also represent the immune states in the case of immune networks, etc (see Kauffman (1993) and references therein).

The study of Boolean networks where the nodes are connected at random and the functions are connected at random shows that as  $K$  decreases from  $N$  to 1, the system goes from a chaotic behaviour to a crystallised one with the phase transition occurring

at  $K = 2$ . Relevant features of Boolean networks for our work, with respect to  $K$  are presented here (Kauffman 1993).

**For  $N \geq K \geq 5$**

It is shown that the median number of configurations in an attractor, that is its length, scales approximately as  $2^N$ . Furthermore, the behaviour of the network shows a sensitivity to initial conditions and since the length of the attractor increases exponentially with  $N$ , the system is defined as chaotic. It is also estimated that the number of attractors is of the order of  $N$ . Finally, another interesting feature is the non-uniformity of the size of the basin of attraction, with few enormous basins with large attractors and many small basins with small attractors.

**For  $K = 2$**

The behaviour of the network undergoes a change of regime as  $K \rightarrow 2$ . For  $K = 2$ , numerical simulations show that the median cycle length as well as the number of cycles scale as  $N^{1/2}$ . The attractors represent then a much smaller fraction of the configuration space compared to networks with  $K \geq 5$ . It is noticed also that the distribution cycle is skewed compared to a Gaussian distribution, when plotted as the logarithm of the cycle length, and that few attractors have very long cycles. Furthermore, the system is stable to minimal transient perturbations and when a change of attractor occurs, the new attractor is in the neighbourhood of the departing one. Finally, not all the nodes have a variable state and about 70% of the nodes remain in a fixed on or off state.

**For  $K = 1$**

For  $K = 1$ , the system falls into disconnected components. But, despite the disconnections, the network shows a dynamics where loops exist, and the overall behaviour of the network is then the product of the isolated components.

## 4.3 Setting up the model

We introduce in the next section our non-autonomous dynamical network model. In terms of dynamics alone, this model presents similarities to that of the Boolean networks

that have just been presented above. A production function has been associated with the dynamics of the network so as to model levels of expression at the nodes.

### 4.3.1 Architecture and representation of the network

We consider a directed network where the nodes, representing a production machinery of control elements, are associated with a binary state. The production at a node is state dependent such that it is producing a control element if it is ON and not producing anything if it is OFF. The state of a node is dependent on the other nodes that it is controlled by and may vary with time. A directed link represents the control interaction from a node to another and can have either a positive effect (contributing to turning a node ON) or negative effect (contributing to turning a node OFF) on the production at the controlled node. The positive or negative interactions are fixed once and for all in a simulation. In practical terms, the networks are first built with only positive links. A proportion of those links is then reassigned at random as negative and we label this proportion  $\mu$ .

The configuration at time  $t$  of the network is given by the states of all the nodes at that time. It is represented by vector  $S(t)$  whose components,  $s_i(t)$  are

$$s_i(t) = \begin{cases} 0, & \text{if node } i \text{ is OFF;} \\ 1, & \text{if node } i \text{ is ON.} \end{cases}$$

The network itself, is represented by the matrix  $A$  whose elements  $a_{ij}$  are

$$a_{ij} = \begin{cases} 0, & \text{if there is no link from the node } j \text{ to the node } i; \\ 1, & \text{if node } j \text{ is connected and directed to node } i \text{ and acts as an inducer on } i; \\ -1, & \text{if node } j \text{ is connected and directed to node } i \text{ and acts as a repressor on } i. \end{cases}$$

In terms of architecture, we focus only on the distribution of the degree of connectivity and we ignore other architectural features. We will use networks with degree distribution that follow either Poisson distributions or power-laws. The interesting point here is that, as the networks are directed, it is possible to make the distinction between the incoming and outgoing degree distribution. This will be of interest later in chapter 5

### 4.3.2 Dynamics

The state of a node is determined according to a simple dynamical rule, which could be viewed as a majority rule, that we call also the balance, of its incoming links. In simple terms, if there are more positive interactions incoming than negative ones, the node will be ON. Otherwise, it will be OFF. The extra condition is here that only nodes that are ON can exert a control over other nodes. Finally, a node  $i$  will be ON at  $t$  if

$$\sum_j a_{ij} s_j(t-1) > b_a, \quad (4.1)$$

where  $b_a$  is a preset level acting as a bias of activation and  $s_j$  is the state of the node. The node will be OFF otherwise.

### 4.3.3 External input

We have defined so far the network and its dynamics. We need now to give initial conditions; there are many ways to do this. What we want though is a network whose behaviour would be dependent on an environment with the possibility that changes in this environment affect the network. A simple approach implies that the network is permanently affected by the environment as long as this is not modified. A proportion of nodes is then set to receive permanent external inputs, which can be seen as many additional links integrating the influence of that environment in which the network is located. Those external inputs are given in the vector  $e$  as

$$e_i = \begin{cases} 0, & \text{if node } i \text{ does not receive external inputs;} \\ \epsilon, & \text{if node } i \text{ receives external inputs,} \end{cases}$$

where  $\epsilon$  is a value high enough to ensure that the corresponding node receives a sufficient induction to remain ON whatever its upstream controls might be. Its actual value is not important.

In a biological experiment, our network, with some of its nodes receiving extra inputs, would correspond to a continuous culture where the medium defines the growth conditions and where the supplied nutrients, hormones, etc, are not a limiting factor. The simplification in our model is that the state of these input nodes is never dependent on any of their upstream controls, *i.e.* the set of incoming links connecting them.

The networks are first generated with undirected links, that is  $a_{ij} = a_{ji}$ . A direction

is then set to the link by setting element  $a_{ij}$  to zero with a probability  $p$  and  $a_{ji}$  to zero with probability  $(1 - p)$ . As a consequence, three sorts of nodes emerge, which are:

- The nodes without incoming links:

A node without incoming links does not receive any control signal and is then set OFF by default. In the present case, the average number of nodes without incoming links,  $\langle I \rangle$ , is given by

$$\langle I \rangle = \sum_i n_i p^{(k_i-1)},$$

where  $n_i$  is the number of nodes of degree of connectivity  $k_i$ . Thus, for a given network architecture, increasing the mean connectivity decreases the probability of finding a node without incoming links. Likewise, diminishing  $\bar{k}$  sees the number of those nodes increasing, which is intuitive. Furthermore, changing the degree distribution of the nodes will have also consequences on  $\langle I \rangle$ . Finally, these nodes can only be ON if they are chosen as input nodes.

- The nodes without outgoing links:

These nodes do not influence the behaviour of other parts of the network as they have no nodes to control. Setting them as input nodes has a null effect on the dynamics of the network. Similarly, the average number of nodes without outgoing nodes,  $\langle O \rangle$ , is

$$\langle O \rangle = \sum_i n_i (1 - p)^{(k_i-1)}.$$

- The nodes with incoming and outgoing links:

These nodes have no need to be set as an input to play an active control part in the network, but there is no restriction in assigning them.

There are several possibilities for the nodes on which to apply the external inputs. It could be on all the nodes without incoming links since otherwise they would not play an active role in the network. It could as well be on a random subset of those nodes. This is a more sensible way if one expects to study how the system behaves under variation of external inputs, etc. Certainly, the number of nodes without incoming links varies a lot between networks of different given degree distribution and may, in some cases, be very small. Focussing only on those nodes is then not sufficient if one expects to study how the system behaves under variation of this set of external input nodes. We choose to set external input on a random subset of all the nodes whatever the above category they belong to. The proportion of input nodes is designated by  $\eta$ .



### 4.3.4 The production function

The rate of production at a ON node is some positive function of the balance at that node, and the level of product at  $t$  is related to the state of the nodes at  $t - 1$ . As the time stepping in the model is discrete, the value of the abundance of node  $i$  at one time step is equal to the value of the rate of production of that node at that time. This assumes that the product disappears, for any reason, after one time step.

The abundance at all the nodes at time  $t$  is given by the vector  $P(t)$  where the component  $p_i(t)$  is the abundance at  $t$  of the element produced at node  $i$ . The abundance is given by the relation

$$P(t + 1) = A.S(t), \quad (4.2)$$

with  $S(t)$  the vector, introduced earlier, giving the configuration of the network at  $t$ , with components

$$s_i(t) = \begin{cases} 0, & \text{if } P_i(t) + e_i + b_p \leq 0; \\ 1, & \text{if } P_i(t) + e_i + b_p > 0. \end{cases}$$

where  $b_p$  is the bias of production or base level and the same for all nodes. A biological interpretation of the production function is given later in chapter 5.3.

A node that is OFF not only does not exert controls on any other nodes but also has a zero production. Then, if all the nodes controlling a given node  $i$  have a null production, none of those nodes will exert control on node  $i$ . This node  $i$  will be OFF by default.

Many categories of nodes can be described with respect to the state of their production during the simulation.

- (i). There are the nodes that receive no controls throughout the simulation and are OFF by default. They are not considered as an active part of the system though they may become active under some changes of external inputs. By defining  $N_a$  the number of active nodes, the number of inactive nodes is  $N - N_a$ .
- (ii). With respect to the state of the nodes, there are the nodes that can be constantly ON or constantly OFF and the nodes that display a variable pattern of ON and OFF states. The number of nodes of variable state are expressed by  $N_s$ .
- (iii). With respect to the level of abundance of the nodes, there are the nodes with a constant production and those with a variable production. Note that nodes of constant state ON can nonetheless have a variable production. The nodes of variable production are given by  $N_v$ .

## 4.3.5 The simulation

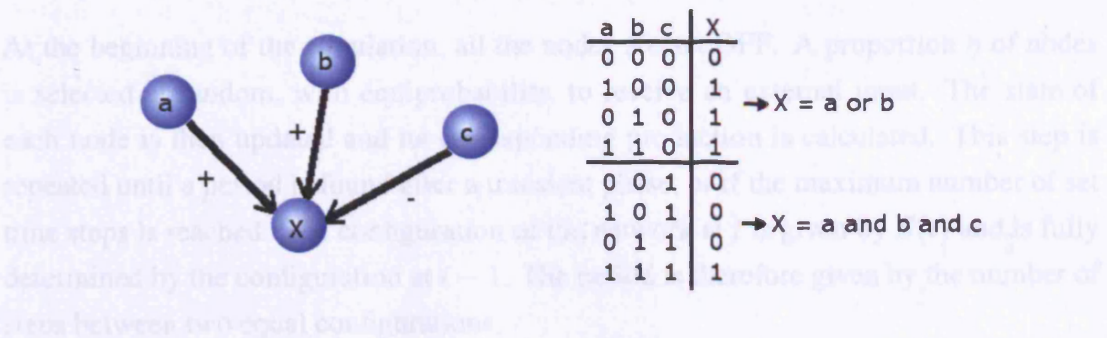


FIGURE 4.1. Representation of an example of control at a node. The truth table gives the update of the state of node  $X$  according to the dynamics of our model as well as the corresponding Boolean function.

Note that since there are active and inactive nodes, it is meaningful to recalculate values such as the mean degree of connectivity or the proportion of negative links of the network made of the active nodes. This will be used below.

### 4.3.5 Is this model a Boolean network?

Our model is that of a Boolean network with respect to the state of the nodes but not in term of its dynamics, and this despite the fact that Boolean functions could eventually be used instead. It is indeed possible to express the dynamics given by Eqs. 4.1 and 4.2 with Boolean functions, but this would be somewhat masochistic. Actually, Boolean functions can be used in computing many problems, including this one, but the fact of not using them is a matter of convenience with regards to the problem treated.

The peculiarities that make the use of Boolean functions inadequate are that the nodes that are OFF have no effect on the network, that the state of the node is the sum of the nodes ON that control it and that of the negative links. As a consequence, the writing of a Boolean function at a node becomes more and more complex as the number of regulator adds up to the regulated node, as illustrated in figure 4.1. The point we are making is that, if the dynamics of our networks can be modelled by a boolean function, it is of a very special case, one just readily described in Booleans terms.

### 4.3.6 The simulation

At the beginning of the simulation, all the nodes are set OFF. A proportion  $\eta$  of nodes is selected at random, with equiprobability, to receive an external input. The state of each node is then updated and its corresponding production is calculated. This step is repeated until a period is found after a transient phase, or if the maximum number of set time steps is reached. The configuration of the network at  $t$  is given by  $S(t)$  and is fully determined by the configuration at  $t - 1$ . The period is therefore given by the number of steps between two equal configurations.

## 4.4 Results

In this section we are going to study the effects of the parameters of the networks on the behaviour as well as the production of the network. We shall first look at the effect of  $\mu$  and  $\eta$  on the different categories of nodes and the periodicity of the network whenever possible. We will also see the effect of two different degree distributions. Note that in the following, the parameters of the model  $b_a$  and  $b_p$  are set to 0. Unless stated otherwise, the maximum time step is set to 30,000, with networks of 2000 nodes and a small mean connectivity of  $\bar{k} = 4.0$  as we are interested in sparse networks.

### 4.4.1 Variation of the different categories of nodes

The simulations show that the behaviour of the network can be classified into three distinct phases with respect to the periodicity and depending upon the parameter  $\mu$  (see figure 4.2): these are the static phase, the periodic phase and the chaotic phase.

**Static phase:** This phase corresponds to a case where the network reaches a fixed point configuration after the transient period. This occurs for small and large fractions of negative links, typically at a values of  $\mu$  such that  $\mu \lesssim 0.20$  or  $\mu \gtrsim 0.85$ . This phase can also obviously be referred to as periodic with a period of 1.

**Periodic phase:** The configurations of the network show that the network reaches a cyclic attractor. More precisely, only a subset of the nodes displays a periodic pattern as shown in figure 4.2. The size of this subset as well as the length of the period varies with respect to  $\mu$ .

**Chaotic phase:** In a large finite network, the number of possible configuration is enormous ( $2^N$ ) yet finite. Even for small  $N$  finding the periodicity of the network could

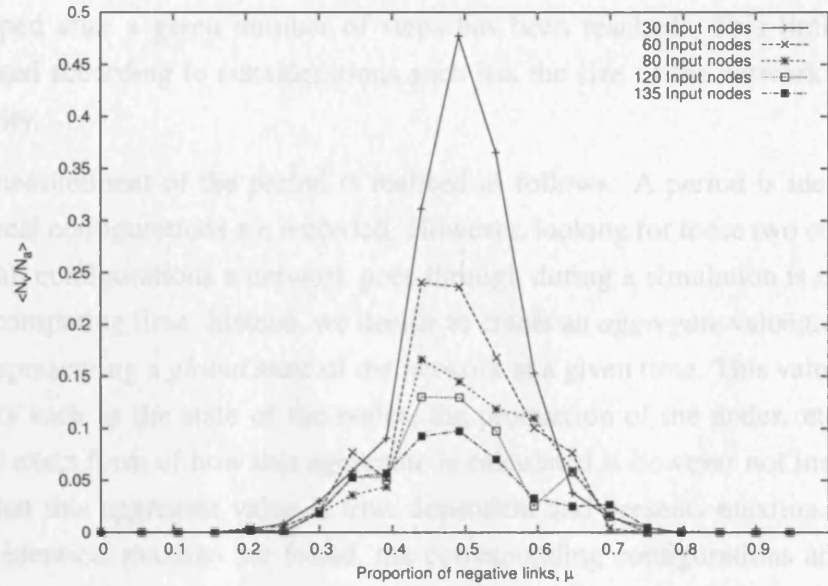


FIGURE 4.2. Distribution of the variable nodes according to the proportion of negative links,  $\mu$ .  $\langle N_v/N_a \rangle$  is the average ratio of the number of nodes of variable production  $N_v$  to the number of nodes connected to the network  $N_a$ . There are 500 realisations of the network for each value of  $\mu$ . There are no variable nodes for the values of  $\mu$  such that  $\mu \lesssim 0.20$  and  $\mu \gtrsim 0.85$  and a variable number of these nodes for  $0.20 \lesssim \mu \lesssim 0.85$ .

be too time consuming. In these cases we assume that, if the network were infinite, such periodicity would not be present. This is the chaotic phase.

In term of the proportion of the different categories of nodes, no differences can be observed whether the degree distribution of the nodes follow a Poisson distribution or a power-law.

#### 4.4.2 Period distribution

We are interested in the distribution of period of the networks for various values of parameters and various architecture. This implicitly involves that we select  $\mu$  so that the network is likely to be periodic. The period is measured for a given network with a given set of external inputs. Since the number of possible sets of external inputs as well as the number of configurations a network can go through are huge, we can only proceed to a sampling of periods. This is done by changing the set of external inputs, leaving the rest of the network unchanged.

Note that the period cannot always be found in a reasonable time. The search is then stopped after a given number of steps has been reached. This limit is variable and adjusted according to considerations such as the size of the network or the mean connectivity.

The measurement of the period is realised as follows. A period is identified when two identical configurations are recorded. However, looking for those two configurations amongst all configurations a network goes through during a simulation is exhausting in terms of computing time. Instead, we decide to create an *aggregate* value that is a single number representing a *global* state of the network at a given time. This value aggregates parameters such as the state of the nodes, the production of the nodes, etc, at a given time. The exact form of how this aggregate is calculated is however not important. The point is that this aggregate value is time dependent and presents maxima. Then, each time two identical maxima are found, the corresponding configurations are compared, which tell us whether a period has been reached.

A problem that arises is to discriminate between attractors of identical periods. This cannot be done exactly even though there are circumstances that allows to remove duplicate attractors from the sampling. For example, the number of nodes of variable production during a cycle is known. Therefore two attractors with an identical period but with a different number of nodes of variable production are considered different. Other parameters such as the mean connectivity or the proportion of negative links are also used. However, we are only sure that two attractors of identical period are different in the limit of the parameters we use.

We look first at the effect of the architecture on the period distribution. For this, we compare a randomly generated network having a Poisson distribution to a scale-free network. The figure 4.3 shows that the period distributions for those networks are two very similar power-laws. The distribution of periods is therefore independent of the architecture of the network, and since no difference can be noticed, we work, in the following, with only random networks.

We look then at the effect of the size of the network on the period distribution. Figure 4.4 shows that again the distributions are power-laws and that no distinction can be made between networks of various sizes.

We are also interested in the effect of the variation of the proportion of negative links, shown in figure 4.5. This time we can see a difference in the slope of the distribution as  $\mu$  varies: the probability of finding a network with a higher period increases with  $\mu$ . This confirms that when  $\mu \rightarrow 0.5$  the network tends to be chaotic.



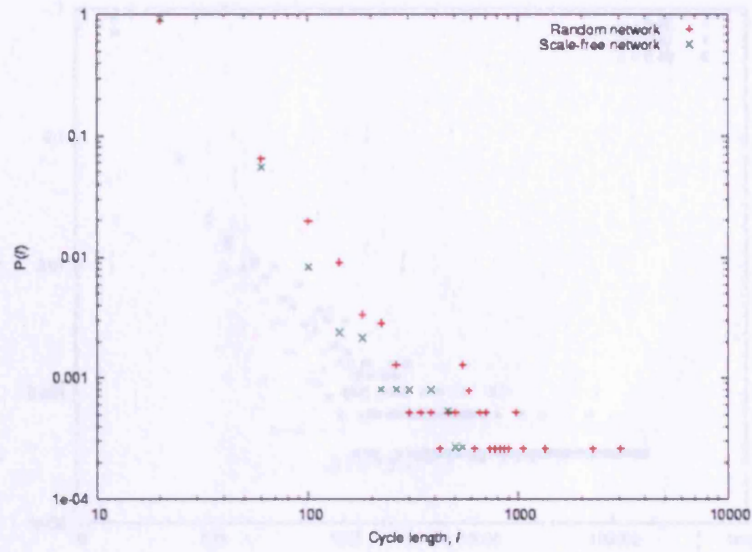


FIGURE 4.3. Distribution of period for random networks and scale-free networks. The period  $\tau$  is the length of the cyclic attractor. The networks are generated with 500 realisations using a maximum of 30,000 time steps to identify a period. The distributions are power-laws.

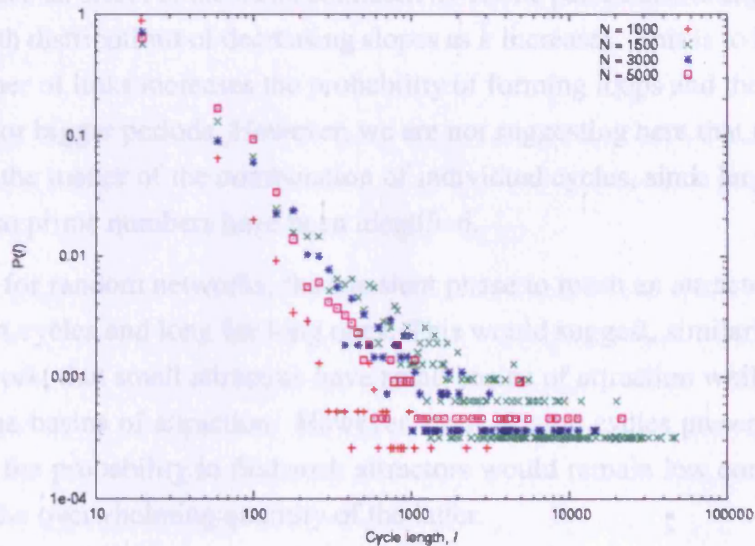


FIGURE 4.4. Distribution of periods for random networks for different sizes of the network, with  $\mu = 0.35$  and  $\bar{k} = 8.0$ . The distributions decay as power-laws.

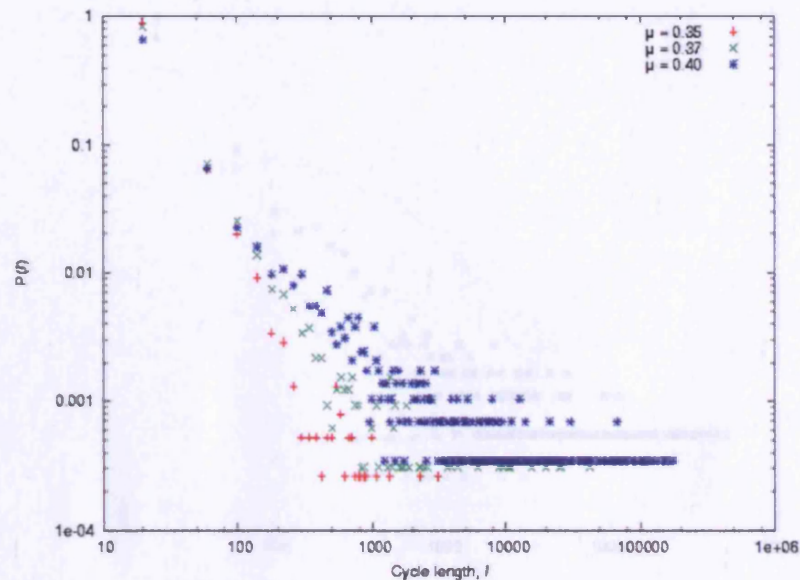


FIGURE 4.5. Distribution of periods for random networks with different proportion of negative links, with  $N = 1000$  and  $\bar{k} = 8.0$ . The distributions decay as power-laws.

There is also an effect of the mean connectivity on the period distribution as shown in figure 4.6, with distributions of decreasing slopes as  $\bar{k}$  increases. This is to be expected as a larger number of links increases the probability of forming loops and therefore greater possibilities for bigger periods. However, we are not suggesting here that the length of a cycle is only the matter of the combination of individual cycles, since large cycles with length equal to prime numbers have been identified.

Note that for random networks, the transient phase to reach an attractor is relatively short for short cycles and long for long ones. This would suggest, similarly to the  $NK$ -boolean network, that small attractors have small basins of attraction while large attractors have large basins of attraction. However, even if large cycles present large basins of attraction, the probability to find such attractors would remain low compare to short ones, due to the overwhelming quantity of the latter.



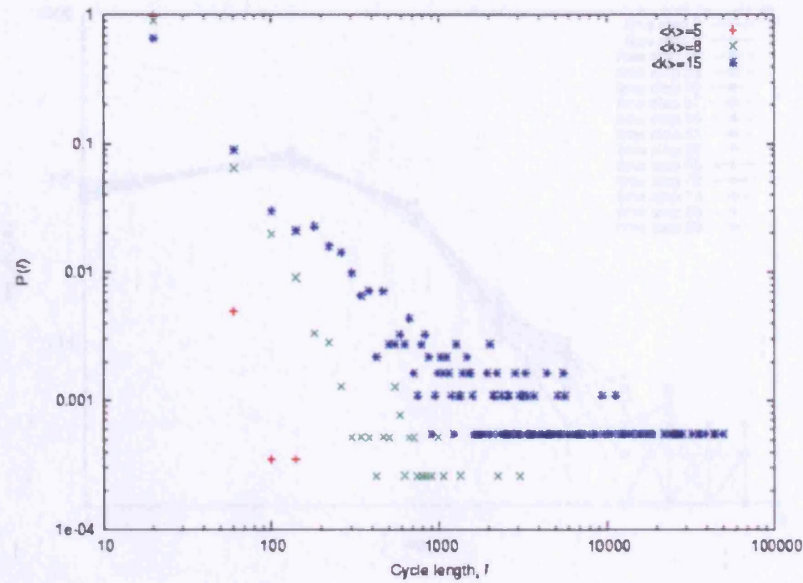


FIGURE 4.6. Distribution of periods for random networks for different mean connectivity, with  $N = 1000$  and  $\mu = 0.35$ . The distributions decay as power-laws of various slope.

#### 4.4.3 The distribution of abundance reflects the architecture of the network

When it comes to plotting the distribution of abundance, which is equivalent to the instantaneous production, there are several choices possible, with the distribution of nodes  $N_a$  and nodes  $N_v$  being the most relevant for our purpose.

We are primarily interested in the distribution of abundance of periodic networks as the static phase, in which each node is constantly expressed, is not particularly relevant to the many cells that undergo major variations in gene expression. We show in figures 4.7 and 4.8 the distribution of abundance for a scale-free and a random network. The distribution of abundance depends in the first instance on the dynamics of the system. However, our results show that, even in the time-dependent case, it reflects also the architecture of the underlying network: the distribution is power-law for a scale-free network and presents an exponential tail for random networks.

By scaling these data so that the mean abundance is shifted to zero, and the standard deviation is rescaled to one, we display the number of nodes at any one time that have their product at a particular level of abundance relative to their means. This rescaling,



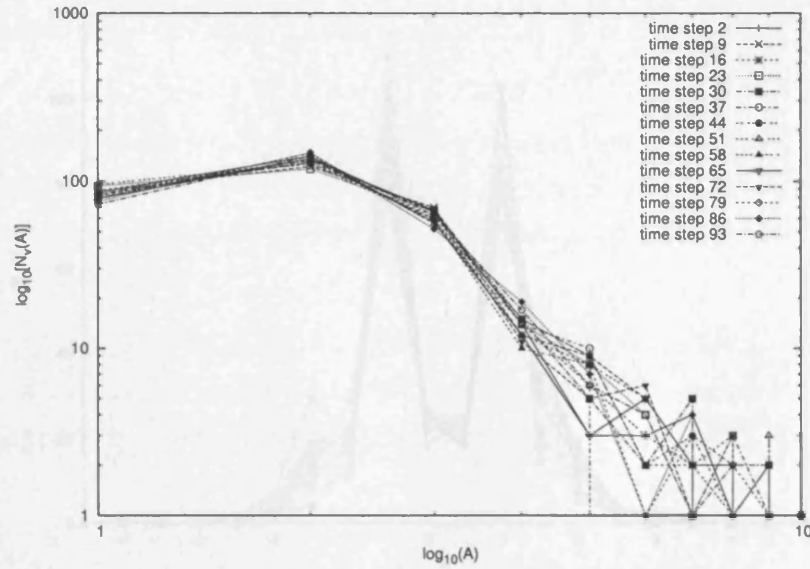


FIGURE 4.7. Distribution of abundance for a scale-free network over a period  $\tau = 94$  time steps. The number of nodes with abundance,  $A$  (x-axis), at a given time is  $N_v$ . The distribution of abundance is represented every 7 time steps beginning from the time of identification of a periodic phase. The tail of the distribution follows a power-law.

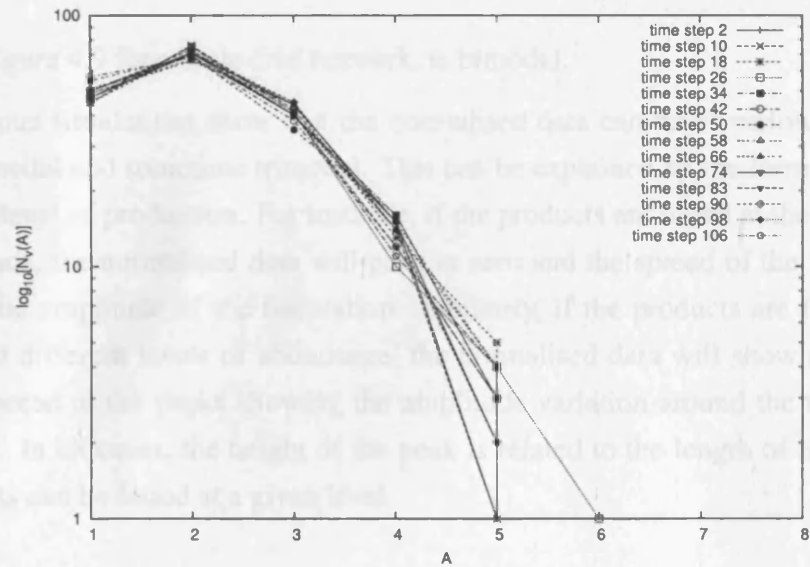


FIGURE 4.8. Distribution of abundance for a random network over a period  $\tau = 104$  cycles. The number of nodes with abundance,  $A$  (x-axis), at a given time is  $N_v$ . The distribution of abundance is represented every 8 time steps beginning from the time of identification of a periodic phase. The distribution is not a power-law.

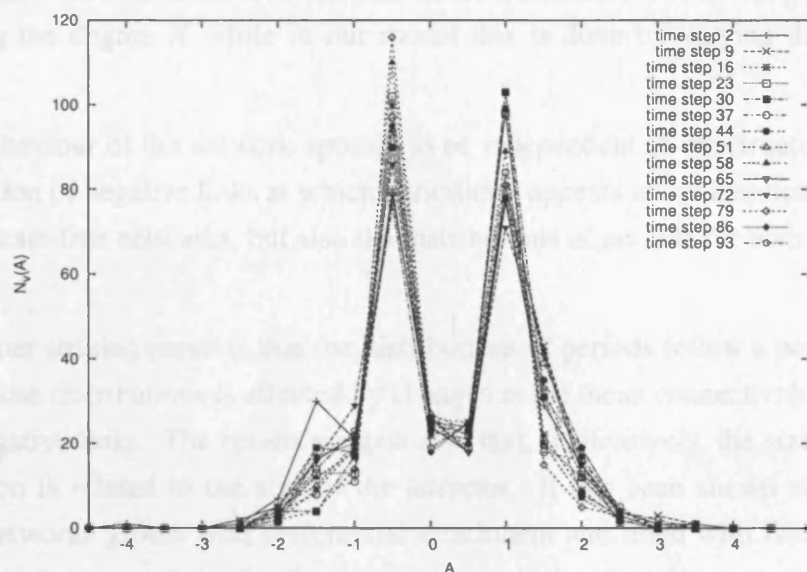


FIGURE 4.9. Re-scaled distribution of abundance for a scale-free network over a period. This network has a period  $\tau = 94$  cycles. The number of nodes with re-scaled abundance,  $A$  (x-axis), at a given time is  $N_v$ . The distribution of abundance is represented every 7 time steps beginning from the time of identification of a periodic phase.

shown in figure 4.9 for a scale-free network, is bimodal.

Numerous simulations show that the normalised data can be of various types: unimodal, bimodal and sometime trimodal. This can be explained by the form of the variation of the level of production. For instance, if the products are found at their mean level of abundance, the normalised data will peak at zero and the spread of the function will represent the amplitude of the fluctuation. Similarly, if the products are found mainly around two different levels of abundance, the normalised data will show as a bimodal with the spread at the peaks showing the amplitude variation around the two levels of abundance. In all cases, the height of the peak is related to the length of time at which the products can be found at a given level.

## 4.5 Summary

The very simple model we have constructed presents an interestingly rich behaviour. Our model presents similarities to the  $NK$ -Boolean network in that there are three regimes at which the network can exist; static or crystallised, periodic and chaotic. The difference

lies in the fact that in the  $NK$ -Boolean model, the transition between regimes is achieved by varying the degree  $K$  while in our model this is done by varying the fraction of inhibitors.

The behaviour of the network appears to be independent of the structure. Not only is the fraction of negative links at which periodicity appears about identical in both random and scale-free networks, but also the distributions of periods for both networks are similar.

The other striking result is that the distributions of periods follow a power-law. The slope of those distributions is affected by changes in the mean connectivity and the fraction of negative links. The results suggest also that, qualitatively, the size of the basin of attraction is related to the size of the attractor. It has been shown elsewhere that directed networks grown with preferential attachment and fitted with Boolean dynamics, presented a power-law distribution of the cycle lengths (Yuan et al. 2004). The authors show that such behaviour is insensitive to the parameter  $K$  and is similar to the  $NK$ -Boolean network of Kauffman (1993) for  $K = 2$ . Furthermore, results using a non-growing network model, similar to ours, show that the sizes of the basins of attraction, for its fixed-points attractors, are distributed as a power-law (Li et al. 2004a). This was found for a small ( $N = 20$ ) autonomous network. A general picture seems to emerge from all these results that the power-law distribution of the length of attractors, and seemingly of the sizes of basins of attraction, are universal. It is not however possible to argue whether this distribution of periods is relevant to a critical phase as it has been suggested (Kauffman 1993, Yuan et al. 2004). It is, nevertheless, certain that this characteristic is central and should be considered if we are to compare the attractors of the networks to cell types or immune states, etc.

Finally, we have shown that the distribution of abundance is related, in our model, to the architecture of the network. Thus, random networks yield random distributions of abundance while scale-free architectures yield power-law distributions of abundance.

# Chapter 5

## *Modelling functions of genetic regulatory networks*

### **5.1 Introduction**

In this chapter, the idea is to use relevant biological data sets to determine general features of genetic regulatory networks. In that respect, gene expression data are a good mirror of the dynamics of regulation in cells. It is expected that integrating those data into a network of interactions would give insight into the correlation between the regulatory functions of the system and its architecture. Such a network can be generated, provided the data exist, for any cell type. Therefore, using networks for different cell types should allow us to reveal some of their common features.

We use time series mRNA data over the cell cycle from microarray experiments (see section 1.4.2 for details). The interest of cell cycle data lies also in its universality as opposed to other non-dynamical sets which relate more to some environmental conditions or adaptative processes. The main approach we will have here is to look at the distribution of the mRNA abundances over the time series. Our aim is to obtain a network model that is able to reproduce these results. We are not interested in the fine details of their dynamics but rather in what simple but global elements are needed in reproducing the general form of the results.

We take in this chapter two examples of organisms for which time series mRNA data are available. We first look at the main characteristics of the distribution of mRNA abundances in comparison to our model. We then try to improve our model so as to

reproduce better the data and look at its limitations. In parallel, we identify the correlations between parts of the structure of the transcriptional network and the mRNA level of abundance. We show such a correlation between the out degree of connectivity and the level of abundance which directs us to modify our model.

## 5.2 Messenger RNA abundance

We present in the following the distribution of mRNA abundance measured during the cell-cycle of two organisms, namely *Saccharomyces cerevisiae* and *Caulobacter crescentus*. Since the organisms used and the physiological conditions from which the mRNA measurement have been carried out are different, we are more interested in the general features that characterise those distributions rather than in the specific details.

### 5.2.1 *Saccharomyces cerevisiae*

The yeast *Saccharomyces cerevisiae* is the model organism for the study of eukaryotic cells. It has a number of genes estimated at about 6200 with a typical cell-cycle, in standard growth conditions, that lasts for about 90 minutes. Several mRNA time series microarray experiment sets are available (Spellman et al. 1998, Cho et al. 1998). They use various chemical and physical methods to synchronise the population of cells from which the data are extracted. In the following we will be using data taken from the Spellman et al. (1998) experiments. For instance we will use the datasets from populations synchronised by *elutriation* and by the  $\alpha$ -factor.

In those experiments, growth conditions like the temperature and the source of carbon or the time point of the cell cycle at which the populations are arrested vary. Furthermore, two different yeast strain backgrounds were used. As a consequence, the growth rate and the synchronicities between the two sets were different. For instance, cells were synchronised during one cell cycle by elutriation, and during two cell cycles by  $\alpha$ -factor. From those datasets, the authors have identified about 800 genes being cell cycle dependent.

#### The elutriation experiment

The elutriation method is a physical means of synchronisation. It is based on cell size discrimination by isolating the smallest ones which are typically at the beginning of the

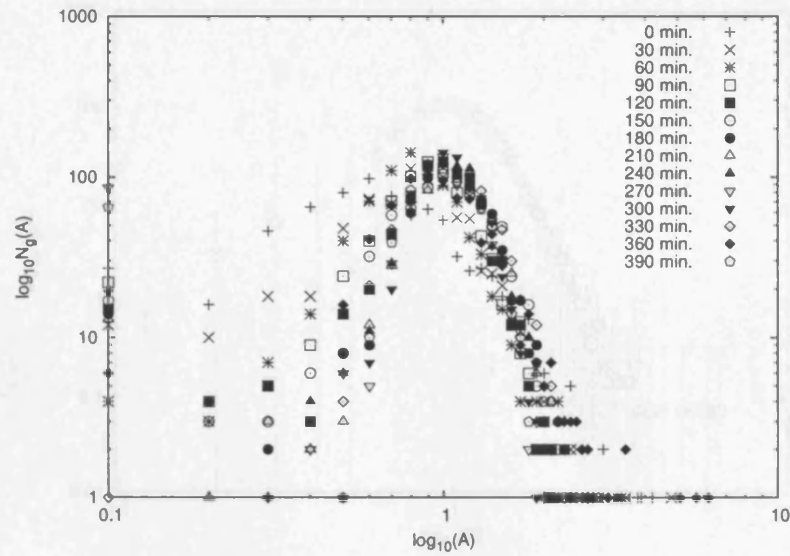


FIGURE 5.1. Distribution of mRNA abundance of *S. cerevisiae* over the first cell cycle following the synchronisation by elutriation. The number of genes of mRNA abundance,  $A$  (x-axis), is given by  $N_g$ . The mRNA abundance is relative to an unsynchronised reference population. Each distribution consists of 763 genes out of the 800 genes identified as cell cycle regulated. The right hand side of these distributions follow a power-law.

cell cycle (G1 phase) (Spellman & Sherlock, 2004b). In the present case, the cells maintained the synchronisation during one cell cycle that lasted 390 minutes. It resulted in 14 time points measurements yielding as many mRNA abundance distributions. Those distributions, for only the genes identified as cell cycle dependent, are presented in figure 5.1. The large variability between the distributions at each time is only an effect of the tail mainly below the abundance point of  $\sim 0.9$ . The large standard deviation of the decaying tail is only an effect of the logarithm scale.

This appears more clearly on figure 5.2 where we show the distributions averaged over time during the synchronised period and the standard deviation at each level of abundance. The averaged distribution with the normalised standard deviation is shown in figure A.3 of the appendix. The decaying part of the averaged distribution is exponential with a power-law tail. This is confirmed when looking independently at the distributions recorded at each time, with the exception of four of them at time  $t = 90$  min,  $t = 150$  min,  $t = 180$  min and  $t = 240$  min. Those graphs are shown in figures A.1 and A.2 in the appendix.

By scaling these data so that the mean abundance of mRNA is shifted to zero, and the



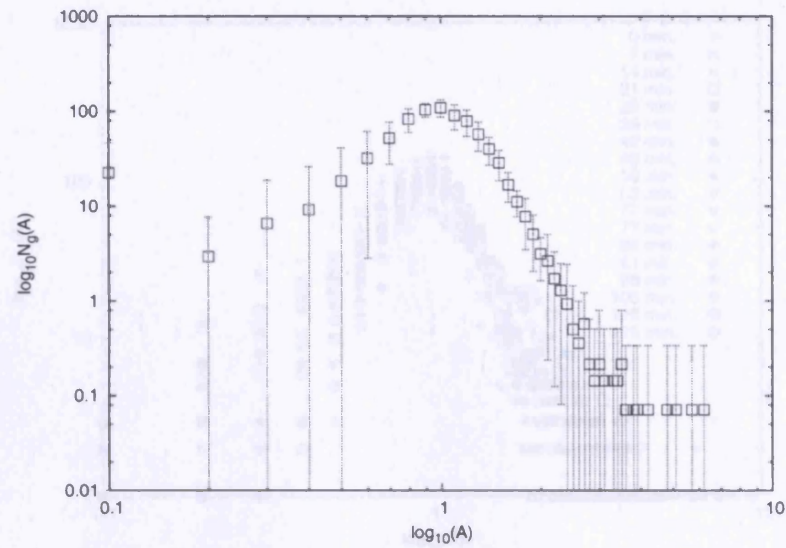


FIGURE 5.2. Distribution of mRNA abundance of *S. cerevisiae* averaged over the cell cycle following the synchronisation by elutriation. The number of genes of mRNA abundance,  $A$  (x-axis), is given by  $N_g$ . The mRNA abundance is relative to an unsynchronised reference population. The distribution consists of 763 genes identified as cell cycle regulated. The right hand side of these distributions follow a power-law

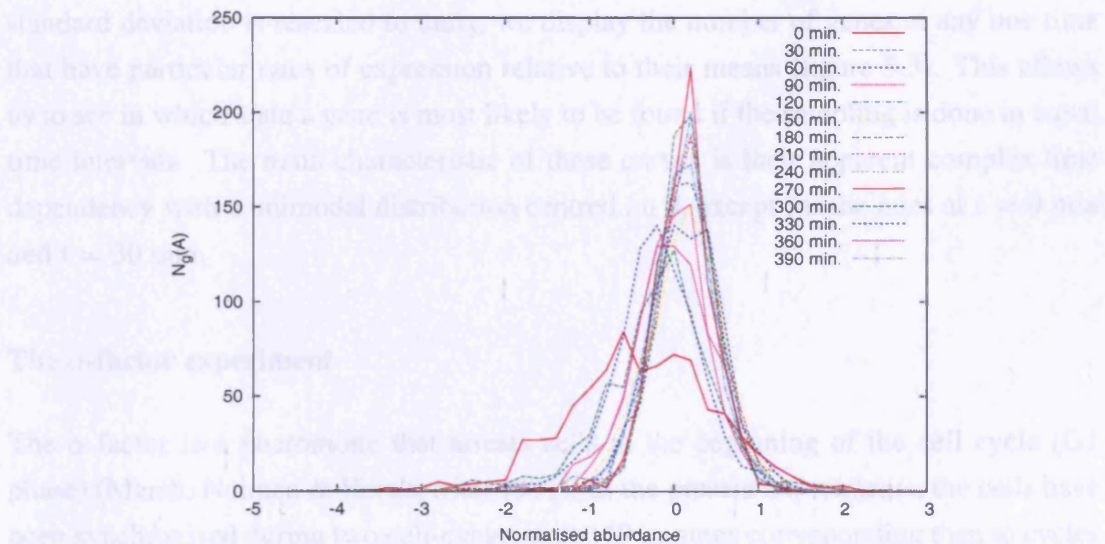


FIGURE 5.3. Re-scaled distribution of relative mRNA abundance of *S. cerevisiae* over the first cell cycle following the synchronisation by elutriation. The number of genes of normalised mRNA abundance (x-axis) is given by  $N_g$ . The data are re-scaled so that the abundance of a gene over the time series is centred to 0 and its standard deviation is 1.

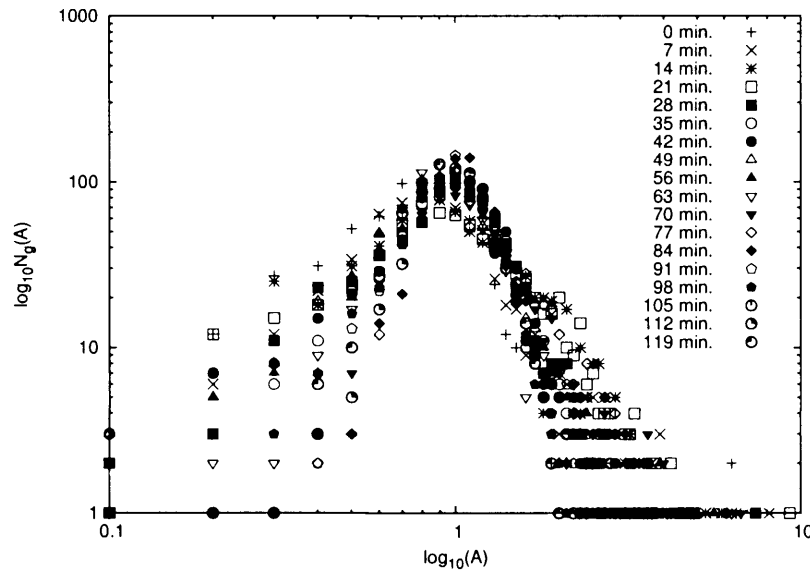


FIGURE 5.4. Distribution of mRNA abundance, of *S. cerevisiae* over the first cell cycle following the synchronisation by  $\alpha$ -factor. The number of genes of mRNA abundance,  $A$  (x-axis), is given by  $N_g$ . The mRNA abundance is relative to an unsynchronised reference population. Each distribution consists of 763 genes out of the 800 genes identified as cell cycle regulated. The right hand side of these distributions follow a power-law.

standard deviation is rescaled to unity, we display the number of genes at any one time that have particular rates of expression relative to their means (figure 5.3). This allows us to see in which state a gene is most likely to be found if the sampling is done in equal time intervals. The main characteristic of these curves is their apparent complex time dependency with a unimodal distribution centred on 0, except for the ones at  $t = 0$  min and  $t = 30$  min.

### The $\alpha$ -factor experiment

The  $\alpha$ -factor is a pheromone that arrests cells at the beginning of the cell cycle (G1 phase) (Marsh, Neiman & Herskowitz, 1991). In the present experiments, the cells have been synchronised during two cell-cycles over 119 minutes corresponding then to cycles of shorter period, compared to the elutriation experiments. This resulted in 18 time series data from which the cell cycle dependent genes were extracted. The corresponding mRNA abundance distributions are presented in figure 5.4. The figure has the same characteristics as in the elutriation case, with the large variability of the tails being again an effect of the logarithm scale. The abundance distribution averaged over time and the standard deviation at each level of abundance is shown in figure 5.5 and the averaged



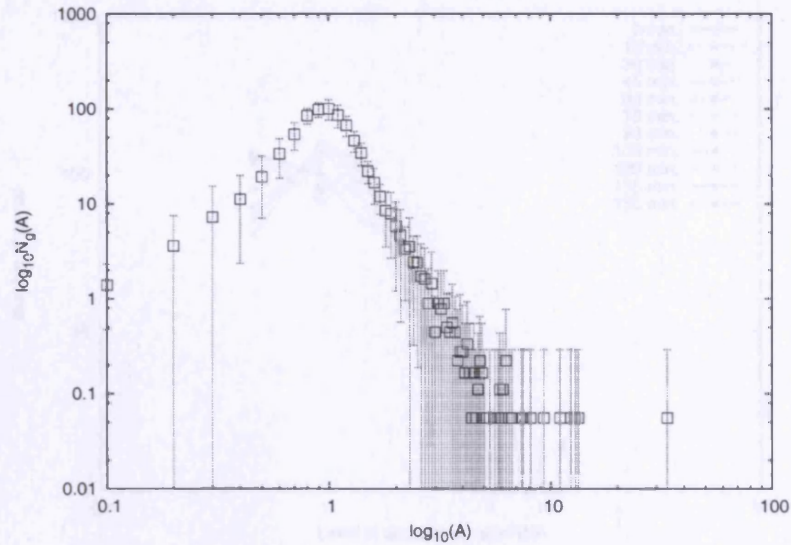


FIGURE 5.5. Distribution of mRNA abundance, of *S. cerevisiae* over the first cell cycle following the synchronisation by  $\alpha$ -factor. The number of genes of mRNA abundance,  $A$  (x-axis), is given by  $N_g$ . The mRNA abundance is relative to an unsynchronised reference population. Each distribution consists of 763 genes out of the 800 genes identified as cell cycle regulated. The right hand side of these distributions follow a power-law.

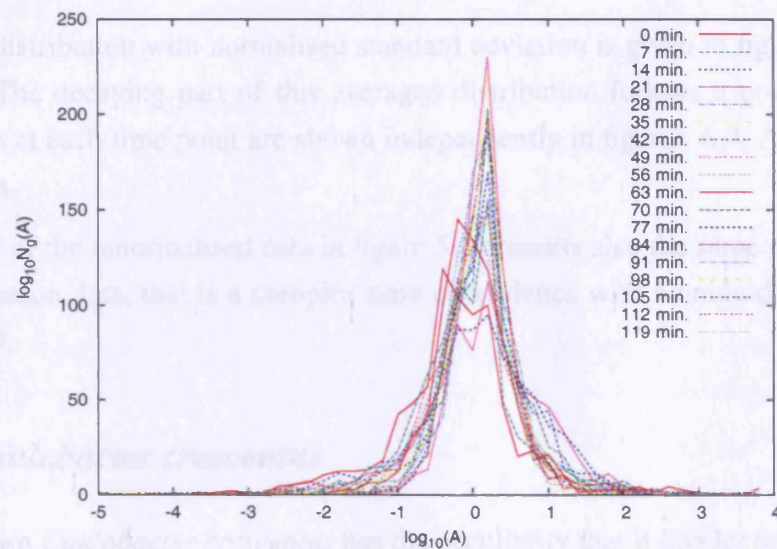


FIGURE 5.6. Re-scaled distribution of mRNA abundance, of *S. cerevisiae* over the first cell cycle following the synchronisation by  $\alpha$ -factor. The number of genes of mRNA abundance,  $A$  (x-axis), is given by  $N_g$ . The mRNA abundance is relative to an unsynchronised reference population. Each distribution consists of 763 genes out of the 800 genes identified as cell cycle regulated.

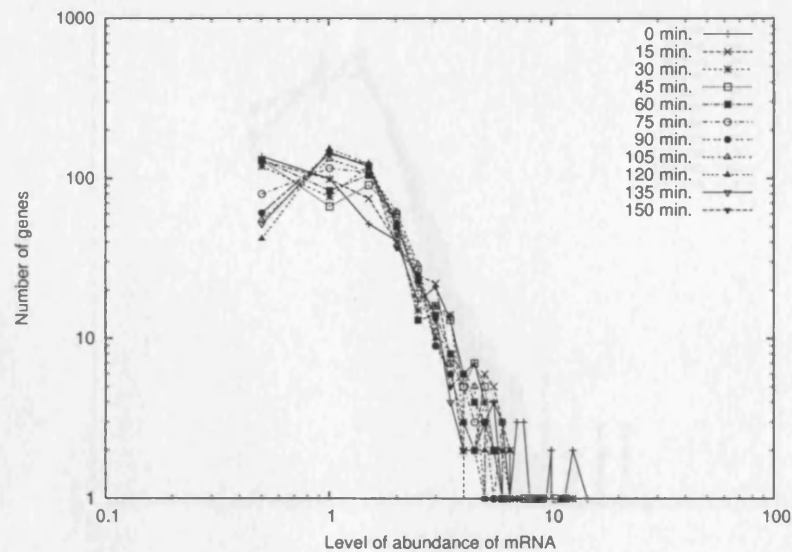


FIGURE 5.7. Distribution of mRNA abundance of *C. crescentus* over a cell cycle of the genes identified as cell-cycle dependent. The mRNA abundance is relative to a mixed and unsynchronised reference population. All the genes with missing data in the time series have been removed. Only 422 genes remain after this selection. The distribution follows a power-law.

abundance distribution with normalised standard deviation is given in figure A.7 in the appendix. The decaying part of this averaged distribution follows a power-law. The distributions at each time point are shown independently in figures A.4, A.5 and A.6 in the appendix.

The plot of the renormalised data in figure 5.6 presents also the same characteristics as the elutriation data, that is a complex time dependence with a unimodal distribution centred on 0.

### 5.2.2 *Caulobacter crescentus*

The bacterium *Caulobacter crescentus* has the peculiarity that it divides asymmetrically during the cell cycle. A stalk cell gives two daughter cells: a stalk and a swarmer cell whose functions and structures are different from each other. The genome of *C. crescentus* is smaller than that of *S. cerevisiae* with only about 3800 genes (Nierman et al. 2001). A typical division cycle in standard growth conditions lasts for about 130 min.

We use here the mRNA time-series data from Laub et al. (2000) in which the pop-

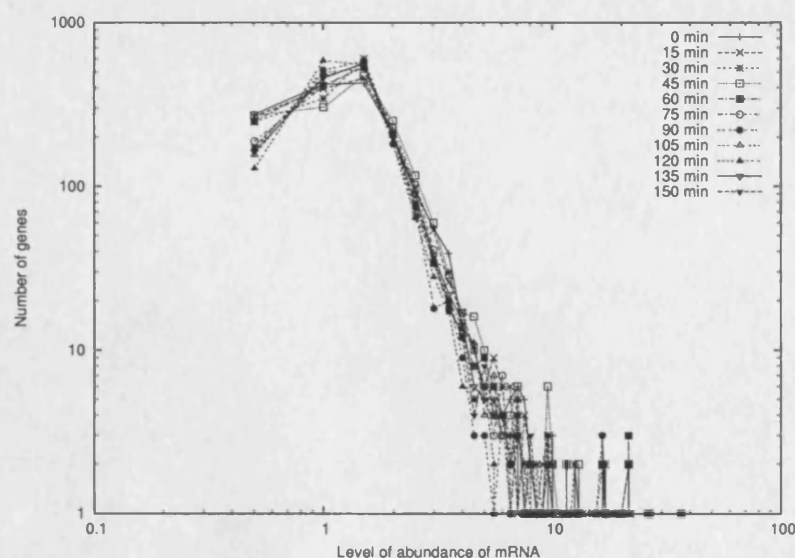


FIGURE 5.8. Distribution of mRNA abundance of *C. crescentus* over a cell cycle. The mRNA abundance is relative to a mixed and unsynchronised reference population. All the genes with missing data in the time series have been removed. Only 1593 genes remain after this selection. The distribution follows a power-law.

ulation of cells is synchronised using physical means (Quon, Marczyński & Shapiro, 1996, Evinger & Agabian, 1977). From those experiments, the authors have identified 553 genes out of the 2,966 tested to be cell cycle dependent. Of these 553 genes, only 422 present a complete time series, and only those will be considered in the following as cell-cycle related genes.

As previously, we plot the distribution of mRNA abundance of the cell cycle dependent genes at various recorded times (figure 5.7). The form of the distributions is different from that observed for *S. cerevisiae* as it does not present an increasing part at low abundances values. The tails of the distributions are power-laws of invariant slopes. However, these characteristics are not specific to the cell-cycle dependent genes and are found when all the genes are considered. Similarly, we plot in figure 5.8 the distribution of mRNA abundance of all the genes whose mRNA abundance values are known over the measured period, that is 1593 genes. Those distributions have a power-law tail of invariant slope.

We show in figures 5.9 and 5.10 the normalised distribution of mRNA abundance for the cell cycle related genes and the 1593 genes, respectively. Both the distributions present the same main characteristics of a complex time dependency with unimodal dis-

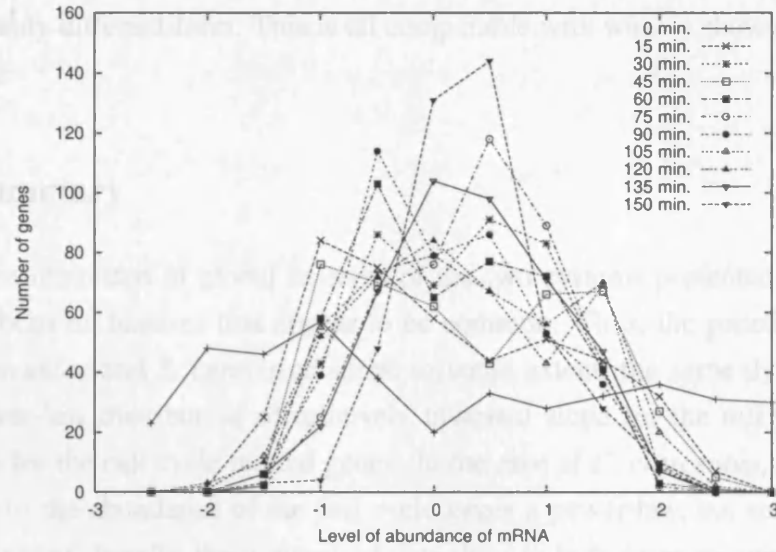


FIGURE 5.9. Re-scaled distribution of mRNA abundance of *C. crescentus* over a cell cycle for the genes identified as cell-cycle dependent. The number of genes of re-scaled mRNA abundance (x-axis) is given by  $N_g$ . The data are rescaled in the same way as in figure 5.3.

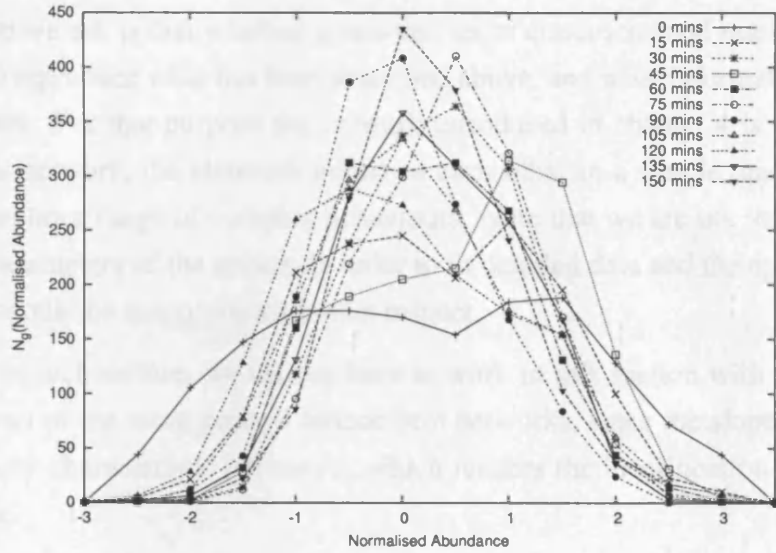


FIGURE 5.10. Re-scaled distribution of mRNA abundance of *C. crescentus* over a cell cycle for 1593 genes. The number of genes of re-scaled mRNA abundance (x-axis) is given by  $N_g$ . The data are rescaled in the same way as in figure 5.3.

tributions centred on 0. Furthermore, the distributions at  $t = 0$  minute, in both cases, are of a remarkably different form. This is all comparable with what is shown in the case of *S. cerevisiae*.

### 5.2.3 Summary

Since we are interested in global features of the two systems presented here, we will retain and focus on features that appear to be common. Thus, the genetic networks of both *C. crescentus* and *S. cerevisiae* show, to some extent, the same dynamics. They share a power-law distribution of relatively invariant slope for the mRNA abundance over a cycle for the cell cycle related genes. In the case of *C. crescentus*, not only is the distribution of the abundance of the cell cycle genes a power-law, but so also is that of most of the genes. Finally, the normalised data show in both cases an apparent complex time dependence for the normalised distributions of their mRNA abundances, as well as unimodal distributions centred on 0.

## 5.3 Can a simple model reproduce biological data?

The question we ask is that whether a minimal set of conditions and rules can constrain the model to reproduce what has been described above, and what information it provides on the system. For that purpose the network introduced in chapter 4 is an appropriate start. In this network, the elements influence each other in a simple manner and yet it is able to display a range of complex behaviours. Note that we are not interested in fine tuning the parameters of the system in order to fit detailed data and the approach we are taking here would be inappropriate in that respect.

In term of architecture, we choose here to work in this section with scale-free networks, instead of the more general Mandelbrot networks, since the slope of the power-law is the only characteristic parameter, which renders the identification of such distribution easier.

### 5.3.1 Reduction of complex network into a simple model

Going from a biological system to a model requires a number of simplifications given that we are looking at a minimal system able to reproduce observed characteristics. We make then the following assumptions and simplifications:



- i. The triplet gene-mRNA-protein is reduced to the couple 'gene'-product which is equivalent to assuming a linear dependence between mRNA and protein. There are many ways to represent such a network where 'genes' and products are two distinct elements. But, since there is a unique link between a 'gene' and its product, the product element can be abstracted from the network, leaving only the 'gene'. A node of the network represents then a control element at a certain level of abundance in which the sequence of event from gene to protein is abstracted. This element corresponds to a node of the network.
- ii. A node exerting a control over another node is referred to as a regulator. Conversely, a node on which control is exerted is referred to as the regulated. This control is represented by the links. In our model, regulators can either have a positive or a negative effect, that is as activator or inhibitor, respectively.
- iii. We assume the regulation function and the production function introduced in section 4.3. Furthermore, at a more general level, there is often no correlation between the measured abundance of an mRNA and that of the protein it encodes (Gygi et al. 1999) (Greenbaum, Jansen & Gerstein, 2002). This may be due to the different half-lives of mRNA - as a result of post-transcriptional processes, which we therefore include in our model (see 5.3.3). Note that our networks are constructed to test hypotheses about network behaviour rather than being trained like neural networks to output particular data.
- iv. The abundance of the product tells whether a node has been activated and more importantly, at what level. This abundance is given in absolute quantities as opposed to the relative abundance of mRNA data from microarray measurements. However, we are not interested by a quantitative one to one comparison of the level of expression between real and simulated data as it would require fine tuning of the parameters of the model which is not the focus of this study.
- v. The system is defined with the same parameters as in chapter 4.

Using an etymology inherited from biological systems to define the elements of our network does not turn it into a biological system. We are not trying here to model the biological system in its exact form. However, we are looking for mechanisms, rules, etc, as simple as they can be to relate, somehow, to what is known about the architecture of the network and to what is observed of the dynamics of the network.

In contrast to some similar cases of linear systems, the rate of synthesis at a node does not depend on the quantity of the regulators or of time dependent functions (Chen

et al. 2004), but upon the coupling of the regulators. We make this simplification because the alternative model is more complicated but not necessarily biologically more realistic.

The use of time-series data constrains the network to a range of  $\mu$  corresponding to periodic behaviour, that is  $0.3 \lesssim \mu \lesssim 0.7$ . We have seen in chapter 4 that the distribution of production was directly related to the architecture. According to this, the observed distribution of decaying power-law tails is dependent on a power-law architecture. This draws a direct relationship between architecture and dynamics. For the next section we will then consider the networks to be scale-free.

However, this model cannot account for the complex time dependence. Nor it can account for the unimodality shown by the normalised distribution of real data as shown above in section 5.2. We go through several hypotheses in order to try to improve the model by considering desynchronisation phenomenon, taking into account action delays and degradations of products. We consider independently those phenomena.

The proportion of input nodes is set to 0.05 of the total nodes and  $\mu$  is set to 0.4, unless stated otherwise. Networks of  $N = 2000$  nodes are used, that is about 33% of the *S. cerevisiae* network and about 53% of the *C. crescentus* network (67% of the tested genes). Furthermore, a feature of genetic networks is believed to be the low number of links involved in direct transcriptional regulation (Guelzim et al. 2002). The mean in-coming and out-going connectivities of the networks are set to 3.

### 5.3.2 Desynchronisation of the population

The hypothesis is here that the complex time dependence is introduced by dephasing the population.

A perfectly synchronised population is equivalent to simulating only one network. The desynchrony of population is introduced in the model by averaging the production at any one time over a number of time steps, or what we call a desynchronisation window. The window is defined from 1 to the period of the network,  $\tau$ , that is from a perfectly synchronised population to a population where the abundance of a product is its average abundance. In practice, the sum of the production over the window is chosen instead of the average, which differs only by a multiplicative constant. This has no consequence on the interpretation of the result but it increases the resolution of the distribution. This is due to the fact that the distribution spreads linearly on a wider range of abundance.

We present in figure 5.11 the distribution of abundance for a window of 2 that represents a desynchronisation of only about 2%, for a scale-free network. Thus, a slight

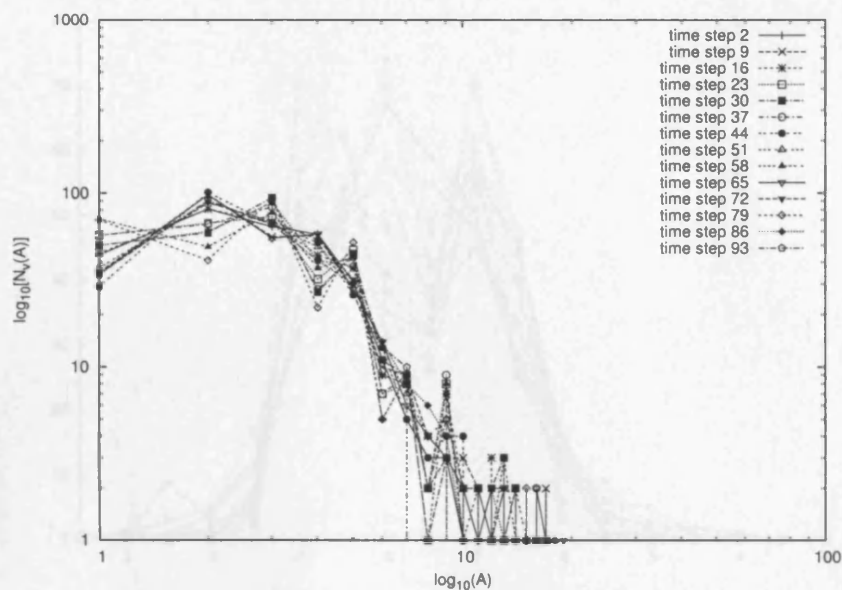


FIGURE 5.11. Distribution of abundance for a desynchronised scale-free network. This is for the same network as in figure 4.7. The number of nodes with abundance,  $A$  (x-axis), is given by  $N_v$ . The desynchronisation is for 2 time steps.

desynchronisation is enough to render the illusion of complex time dependence. This can also be seen in figure 5.12 which corresponds to normalised data. As the desynchronisation increases, the tail of the distribution of abundance becomes an exponential (figure not shown). This is consistent, as the more desynchronised the population, the more random the distribution (the mRNA abundance is then characterised by an averaged value).

### 5.3.3 Delays and degradation

We investigate here the effect of synthesis and degradation kinetics of the regulator on the behaviour of the network and the distribution of regulator. The mRNA and protein half-lives in a cell are not uniform and their quantities present in a cell depends on the rate of both synthesis and degradation. Those rates may vary according the variation of the events, and their kinetics, such as splicing, transcription initiation, mRNA transport, translation initiation, post-translational modifications, etc. The consequence may be, for example, that gene  $a$  exerts indirect control over gene  $b$  although (i) gene  $a$  is no longer activated or (ii) its corresponding mRNA has been degraded, since the protein it encodes may continue to be active over several time steps. We consider both cases independently.



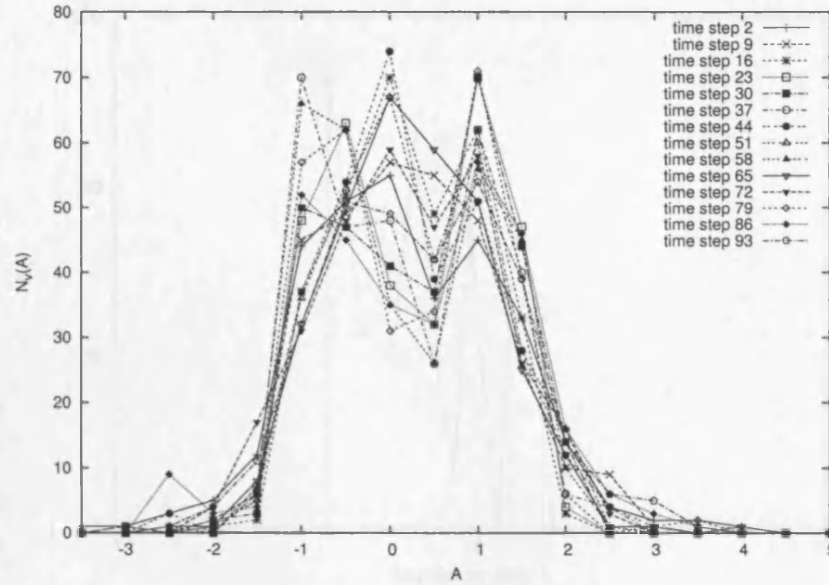


FIGURE 5.12. Re-scaled distribution of abundance for a desynchronised scale-free network. This is for the same network as in figure 4.7. The data are re-scaled in the same way as in figures section 5.2. The desynchronisation is for 2 time steps.

### Regulator lifetime

The degradation rate is indirectly taken into account by including a lifetime parameter,  $l$ . The quantity of a regulator produced at time  $t$  is summed to its quantity produced at  $t, t-1, \dots, t-l$  and disappears thereafter. Thus at  $t$  the abundance of regulator  $i$  is  $\sum_{j=0}^l P_i(t-j)$ .

The distribution of abundance is expected to vary in two ways following the modification of the model. The first one is the stretch of the distribution over the abundance. The second one is that the abundance of the nodes of constant level of expression will lie at values  $l, 2l, \dots$ .

As an example, we plot in figure 5.13 the distribution of abundance for  $l = 3$  chosen at an arbitrary time. Despite focussing only on the nodes of variable level of expression, the distribution shows outliers at level of abundance 3, 6,  $\dots$ . This shows that the level of expression of the nodes fluctuates but most of the time is spent at a given level. All in all, this does not account for what is observed.

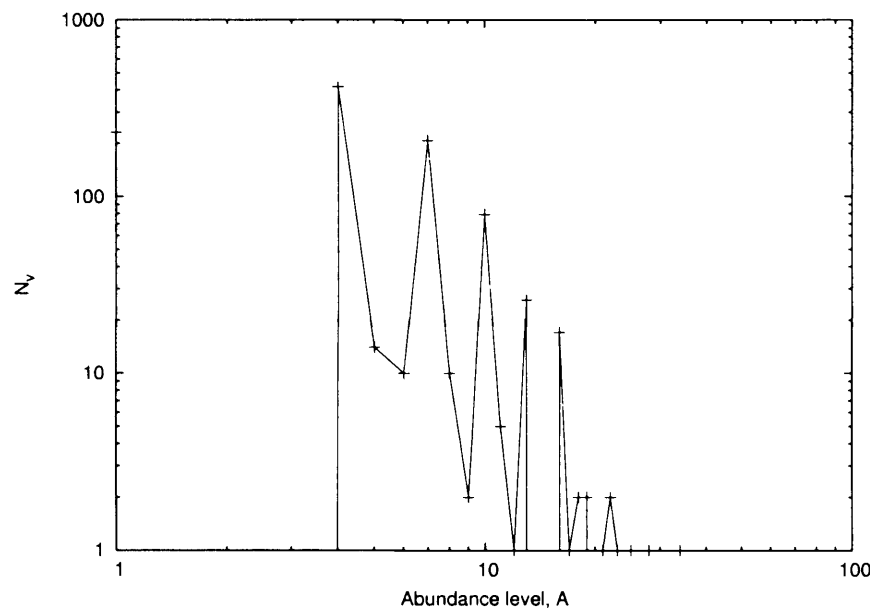


FIGURE 5.13. Distribution of abundance of the node of variable level of expression for a delay  $l = 3$  for a scale-free network. The points of abundance 3, 6, ... outlie the distribution.

### Delay of action

Delay of action is a general term to regroup any mechanism that delays the action of a regulator on its targets. The regulation in biological system is complex and many long steps may be necessary before a regulatory signal may be effective. To that effect, we introduce delays between the production and effect of a regulator that are implemented by inserting a series of delay nodes in between two originally connected nodes (figure 5.14). The number of delay nodes corresponds to the delay time of activation,  $t_d$ . Each delay node will be activated one at a time if the regulator is ON. When all the delay nodes in between two given nodes are activated, the effect of the regulatory node is exerted on the regulated node. This occurs then at a time  $t_d$  after the regulator is activated.

In the simulations, we randomly introduce delay nodes so that the average number is  $\langle t_d \rangle$ . We find that the delay does not affect the distribution if  $\langle t_d \rangle$  is small, but completely destabilises the network, which loses its periodicity, if it is large. The later can be thought of as showing that the introduction of random delays naturally destabilises the period; there is no more cohesion in the behaviour of the network and different parts of it can not be synchronised any longer.

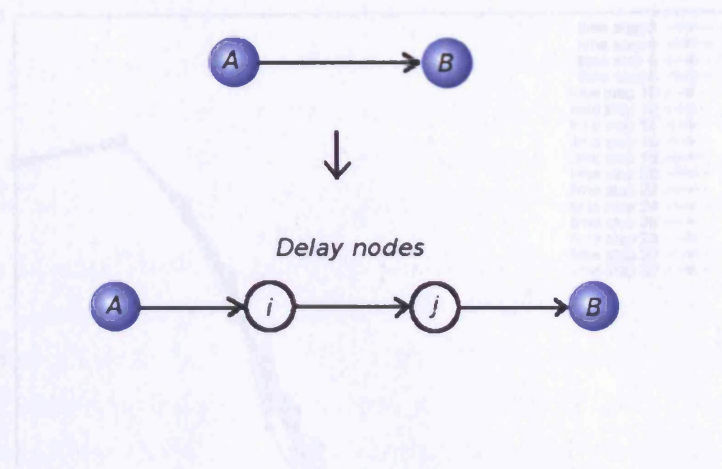


FIGURE 5.14. The delay between nodes  $A$  and  $B$  is modelled by introducing delay nodes  $i$  and  $j$  between them. Node  $B$  is not controlled anymore by the state of  $A$  at  $t - 1$  but its state at  $t - 3$ .

## 5.4 Architectural misfit

It is now clear the architecture of the network is important in determining the distribution of abundance. We here test to what extent this is true by generating networks with different in and out distributions. This is done by first generating the matrix  $A$  for a scale-free network and subsequently randomising either its rows or its columns. This yields networks where one degree distribution is scale-free and the other follows a Poisson distribution.

The figures 5.15 and 5.16 show that in the context we have defined, only the networks with incoming power-law distribution are able to yield a distribution of abundance that follows a power-law. In the case where the distribution of the incoming links is random, the distribution of abundance is exponential. This is in total contradiction with observations of transcriptional networks (Guelzim et al. 2002). For *S. cerevisiae* and *E. coli*, the distribution of the nodes of the incoming links appears to be exponential, whereas the distribution of the out-going links appears to follow a power-law.

There is a clear inconsistency of the in and out distribution at the nodes between the biological system and the model. In the context of the model, the abundance of the network is related to the distribution of the incoming links. The problem is that of turning an exponential distribution into a power-law distribution. There are two possible ways that might achieve this: by modifying either the regulatory mechanism or the production functions at the node that were given in section 4.3.

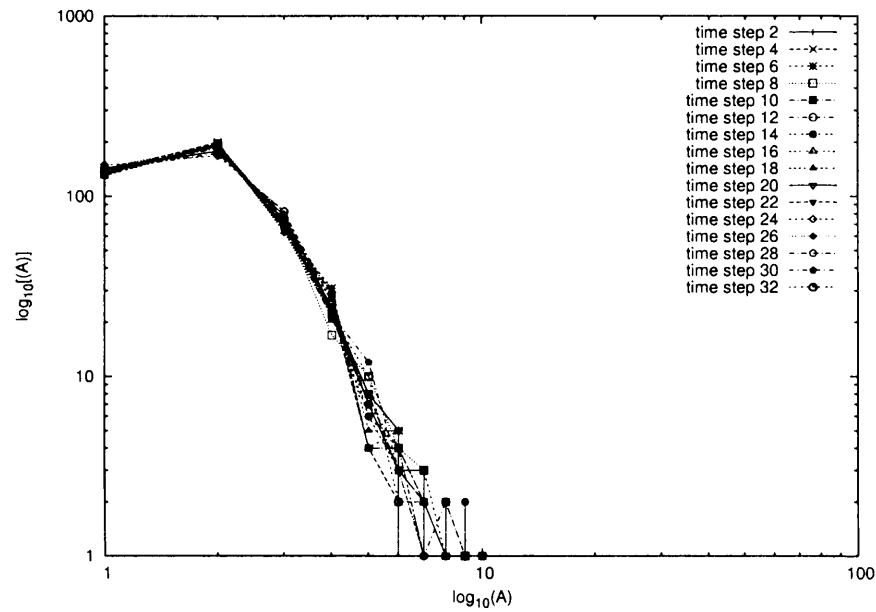


FIGURE 5.15. Distribution of abundance,  $A$ , for a scale-free network with a random distribution of its outgoing links and a power-law distribution of its incoming links. The distributions decay as a power-law.

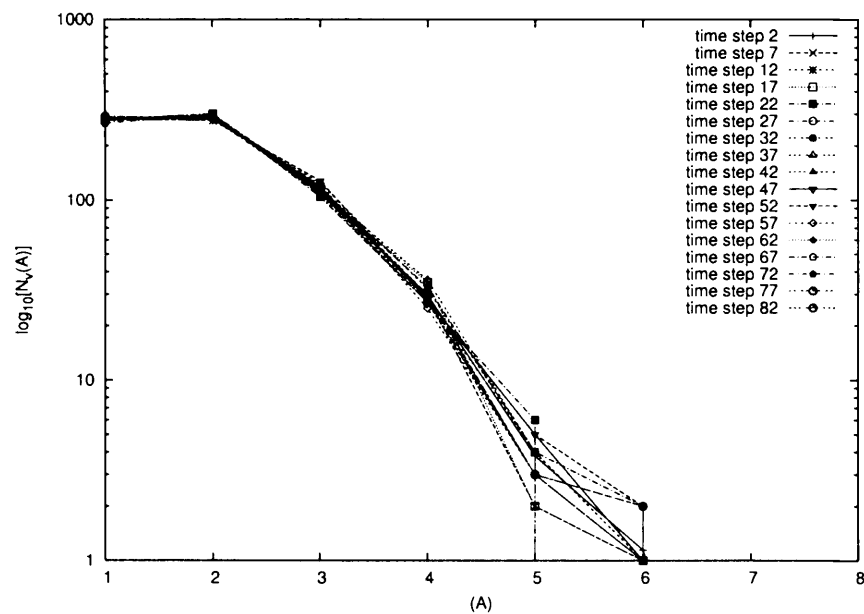


FIGURE 5.16. Distribution of abundance,  $A$ , for a scale-free network with a random distribution of its incoming links and a power-law distribution of its outgoing links.

### 5.4.1 Stochasticity

There are two options to try to reconcile the structure of the network and the distribution of abundance. The first one consists in modifying the function of production given by Eq. 4.2 so that the exponential distribution of abundance is turned into a power-law. The second route, which is the one chosen here, is to modify the regulation mechanism.

The model is modified so as to replace the deterministic mechanism of regulation by a stochastic one. Some biological arguments could be in favour of such modification. For example, the small quantities of mRNAs may alter the process of regulation while another argument may arise from spatial considerations. Indeed, nodes that are spatially close to each other (Képès 2004) may be regulated, while they may not be otherwise.

The stochasticity is implemented in the original model, that is without the consideration of delays and lifetime, as such:

- (i). The number of nodes that a given node can regulate is given by its out-going connectivity. A random number of these links will then be selected to be regulators of the nodes they are connected to and the others will be ignored, that is they will not have a regulatory role.
- (ii). The links to be regulators are reselected at each time.

All the rest of the model is as described in section 4.3. To counter-balance the fact that the introduction of stochastic dynamics lowers the levels of abundance, we record the abundance of a node over a window of time.

The networks used for these simulations have a Poisson distribution of the in-degrees of connectivity and a power-law distribution for the out-degrees. The fraction of negative links is fixed at  $\mu = 0.0$ , which would yield a static network with the deterministic dynamics. We illustrate the result of the simulations with one network.

The main consequence on the dynamic of the network is that the fixed-point attractor disappears and no other attractor is reached. As a result, the nodes cannot be distinguished according the categories mentioned earlier and therefore, all the nodes are considered for the distribution of abundance, which is shown in figure 5.17. The decaying part of the distribution of abundance follows an exponential, with a fat tail that stretches across high levels of abundance. The fact that the network display nodes with high degree of abundance is positive, even though the distribution remains exponential.

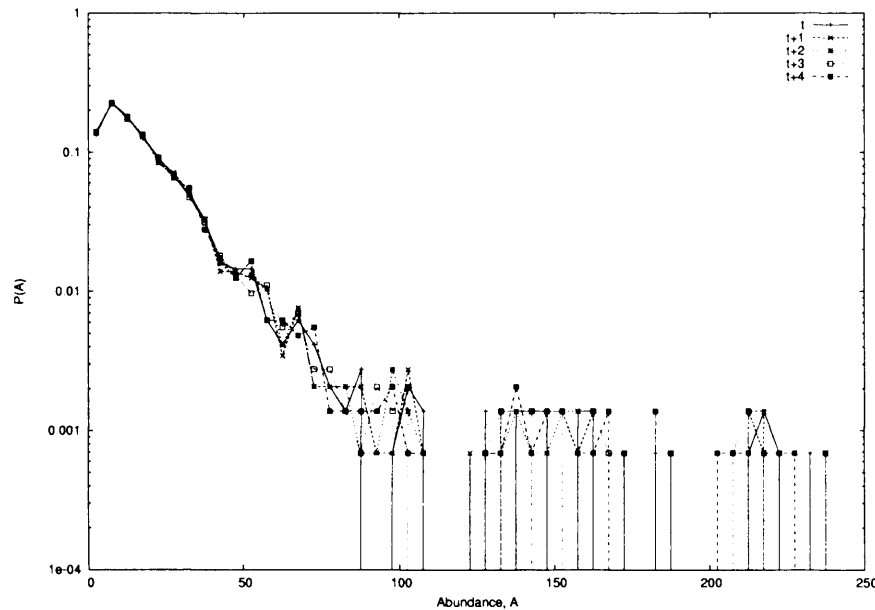


FIGURE 5.17. Distribution of abundance for a stochastic dynamics.

## 5.5 Identification of correlation between node expression and number of controls

We investigate in the following the effect of changing the function of production. In the original model, the production is related to the distribution of incoming links. Since the expected distribution of abundance follows a power-law as does the distribution of the out-degree, a simple and straightforward way would be to relate the distribution of abundance to the distribution of out-degrees. However, doing this would be a major conceptual shift. In the former, the abundance is related to the number of regulators while, in the later, the abundance of a node is related to the number of nodes it can regulate.

In the biological network of a cell, this reduces to showing that the correlations between the mRNA expression data and the architecture of the genetic regulatory network can be used to construct a model regulatory network that could simulate the qualitative features of the observed one. This, including autonomous periodicity and power law abundance distributions, we now proceed to do. For that we look at correlations between nodes expression and the number of controls in *Escherichia coli*.

### 5.5.1 The *E. coli* data

We have on one hand the in and out degree of connectivity of the genes of *E. coli* obtained from the RegulonDb database (Salgado et al. 2004). The data there connects transcription factors, (TFs), to genes and according to it, genes fall into two categories: whether or not they have a non-zero out-going degree of connectivity. The first one corresponds to the transcription factors where both degree of connectivities are greater or equal to zero. The second one corresponds to the regulated genes which do not have out-going links, that is out degrees of connectivity.

On the other hand, the mRNA data are obtained from micro-arrays experiments (Allen et al. 2003) for *E. coli* grown in normal growth conditions. Those data are available from the ASAP database (Glasner et al. 2003). The set of experiments has been repeated up to 10 times and the mRNA abundances used here correspond to the averaged abundances over all experiments. Knowing the degree of the nodes and the relative level of expression of the genes, we are able to extract potential correlations between in-degree of connectivity and mRNA abundance and between out-degree of connectivity and mRNA abundance.

#### In-degree correlation

We look first for a correlation between the in-degree distribution of the genes and the averaged mRNA abundances for those genes. We show in figure 5.18 the data for the expected correlation between in-degree and the averaged mRNA abundance corresponding to the genes of that degree. The error bars are the standard deviation of the mRNA abundances found for each in-degree. No correlation is present. However, the large scatter of mRNA production for each degree suggests that the average is probably not a correct parameter to look at and that the degree abundance is not randomly distributed around a mean. For that, we look at the mRNA abundance for each in-degree separately. We find a very interesting picture: as seen in figure 5.19, where there are sufficient data to exhibit a trend, each of the distributions shows a power law tail.

Furthermore, the power law is the same for each in-degree (figure A.8 in appendix A). Thus, for a given in-degree of connectivity, there are many mRNAs produced in low quantity and few at higher. This behaviour is independent of the degree and confirms the lack of correlation.

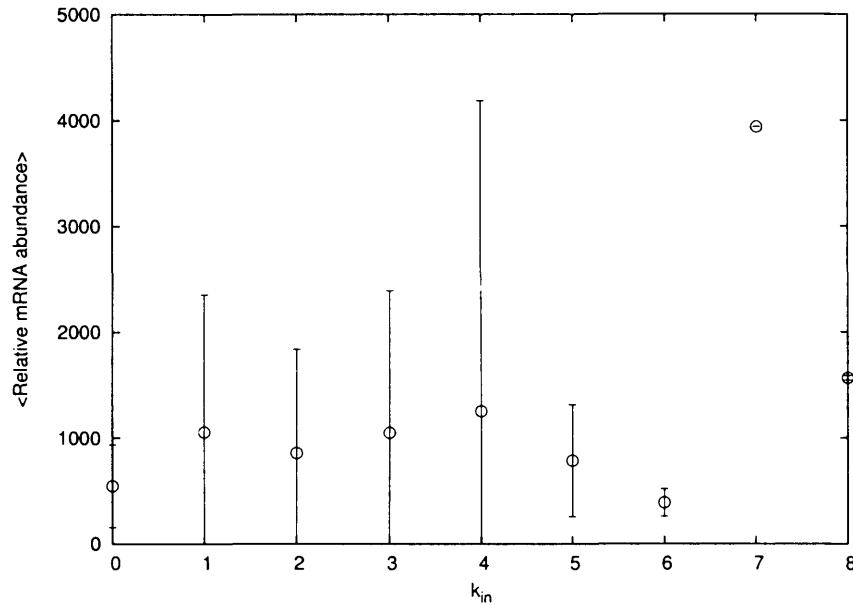


FIGURE 5.18. Averaged relative mRNA abundance measured in *E. coli* versus the in-coming degree of connectivity of the corresponding genes.

### Out-degree correlation

We proceed as above with this time only the genes of the first category, that is  $k_{out} > 0$ . Since for *E. coli* the out-degree follows a power-law (Guelzim et al. 2002) and we are dealing with fewer genes, we expect the distribution of mRNA abundance per out-degree to be more noisy. We show in figure 5.20 a correlation plot between the out-degree and the averaged mRNA abundance for that degree. Compared to figure 5.18, the standard deviations are much smaller here. For out-degrees less than  $\sim 10$ , no correlation is noticeable in the log-linear plot. However, for the other genes, it looks like a trend emerges: the higher the degree, the higher the abundance. Overall, the data are best fitted by a linear curve, showing a linear dependence between abundance and out-going degree of connectivity.

Note that there are insufficient data for each nodal degree to look effectively at the distribution of mRNA for each out-degree.

Finally we look whether there are some correlations between the out-degree of the transcription factors and the abundance of the level of expression of the genes those transcription factors regulate. This is plotted in figure 5.21, which shows no evident correlation.



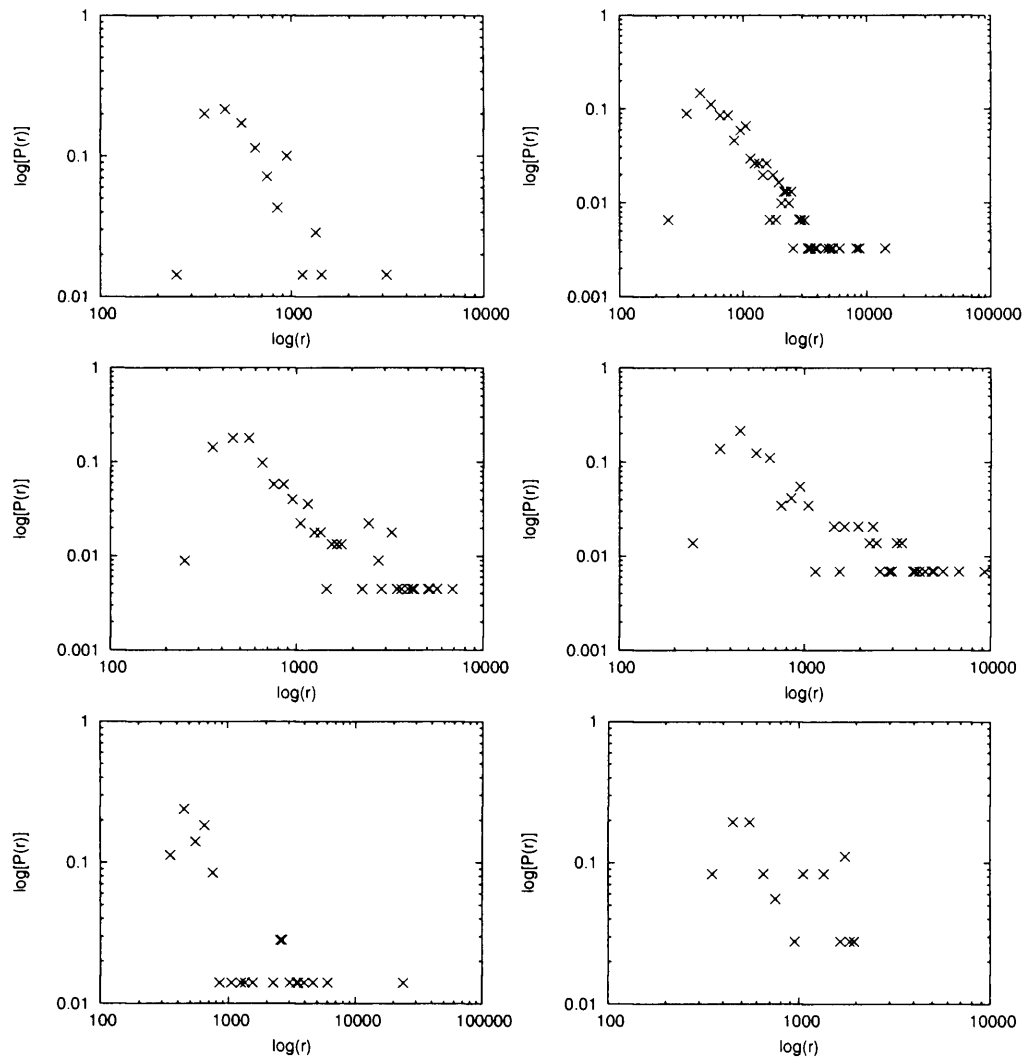


FIGURE 5.19. Distribution of mRNA abundance in *E. coli* for various in-coming degrees of connectivity such as (top left) in-coming degree of 1, (top right) in-coming degree of 2, (middle left) in-coming degree of 3, (middle right) in-coming degree of 4, (bottom left) in-coming degree of 5 and (bottom right) in-coming degree of 6. The distributions decay as power-laws.

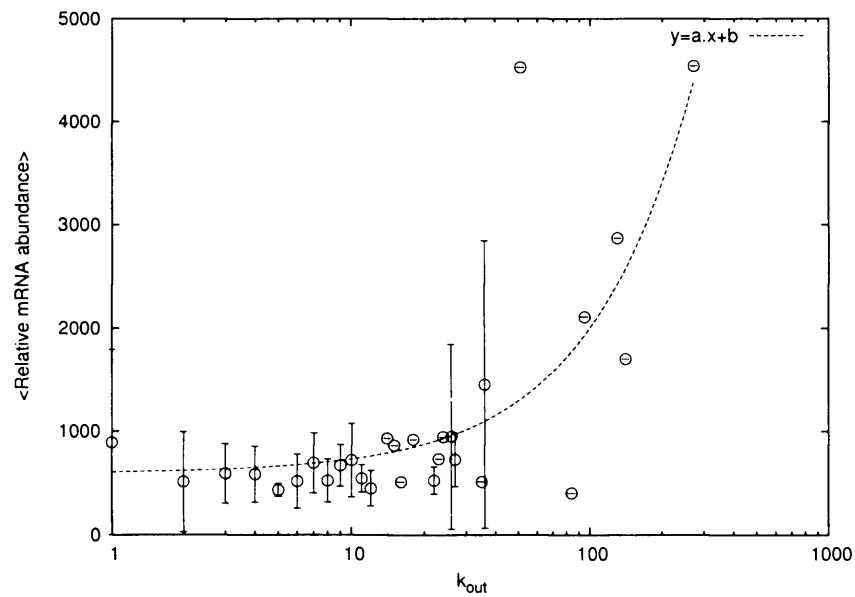


FIGURE 5.20. Averaged relative mRNA abundance measured in *E. coli* versus the out-going degree of connectivity of the corresponding genes. The distribution is best fitted by a linear function  $y = ax + b$  with  $a \approx 14$  and  $b \approx 591$ .

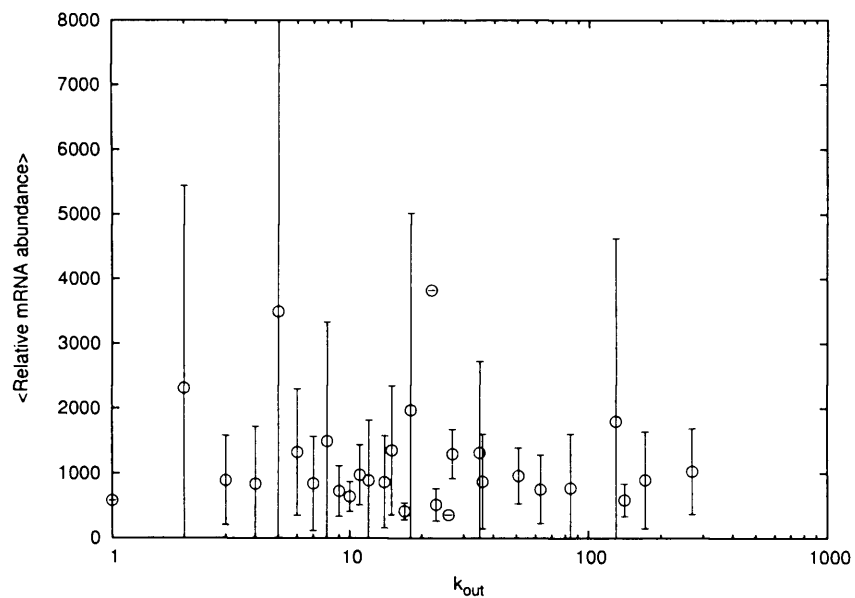


FIGURE 5.21. Averaged relative mRNA abundance measured in *E. coli* versus the out-going degree of connectivity of the transcription factor regulating the corresponding genes.

### 5.5.2 Simulation

#### Network model

From the original model, we modify the function of production at a node, given by Eq. 4.2, so as to take into account the correlation between the out-degree and regulator. The new production function at node  $i$  becomes

$$p_i(t+1) = s_i(t) \sum_j |a_{ji}|,$$

with  $a_{ij}$  an element of the adjacency matrix  $A$ , and the elements  $p_i$  and  $s_i$  as given in section 4.3.4. The rest of the model is defined as follows.

- (1). In accordance with the data we assume a Poisson distribution of the in-coming degrees of connectivity and a power-law distribution of the out-going degrees of connectivity.
- (2). In the data some nodes are connected to links of dual control that is they can be positive and negative. For simplicity we ignore these in the model: such links are arbitrarily assigned to be either positive or negative with equal probability.
- (3). The proportion of negative links is taken as 0.4.
- (4). Mean degree of connectivity  $\bar{k}$  is taken as 8, with  $\bar{k}_{in} = \bar{k}_{out} = 4$ .
- (5). The model has 1500 nodes in line with the data.

Note that the mean degree of connectivity and the proportion of negative links are those specified in the construction of the network. The actual parameters of the connected network may vary from simulation to simulation, because a variable subset of nodes remains unconnected in the construction process.

Note also that in our model, the abundance of the regulator is given by its exact value, that is the real abundance. On the contrary for *E. coli*, the mRNA abundance corresponds to a value relative to a standard population.

#### Results

We are interested in having a periodic network whose period is big enough so that the distributions are not approximated by the fixed-point attractor. However, the probability

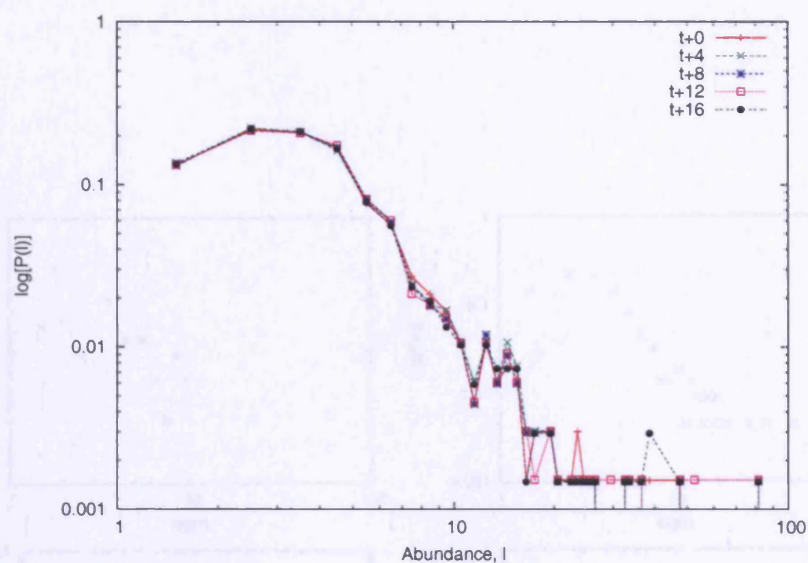


FIGURE 5.22. Distribution of level of abundance of the nodes of the simulated network. The distribution decays as a power-law.

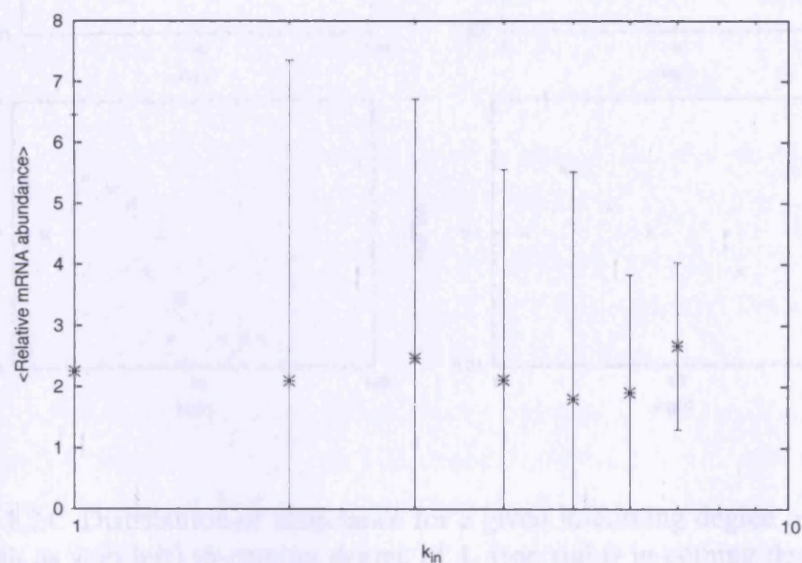


FIGURE 5.23. Averaged distribution of the abundance versus the in-coming degree of connectivity of the corresponding nodes in the simulated network.

of finding a network with a large period is low since the probability of finding a network of a given period decays as a power-law, as shown in section 4.4.2. We present the

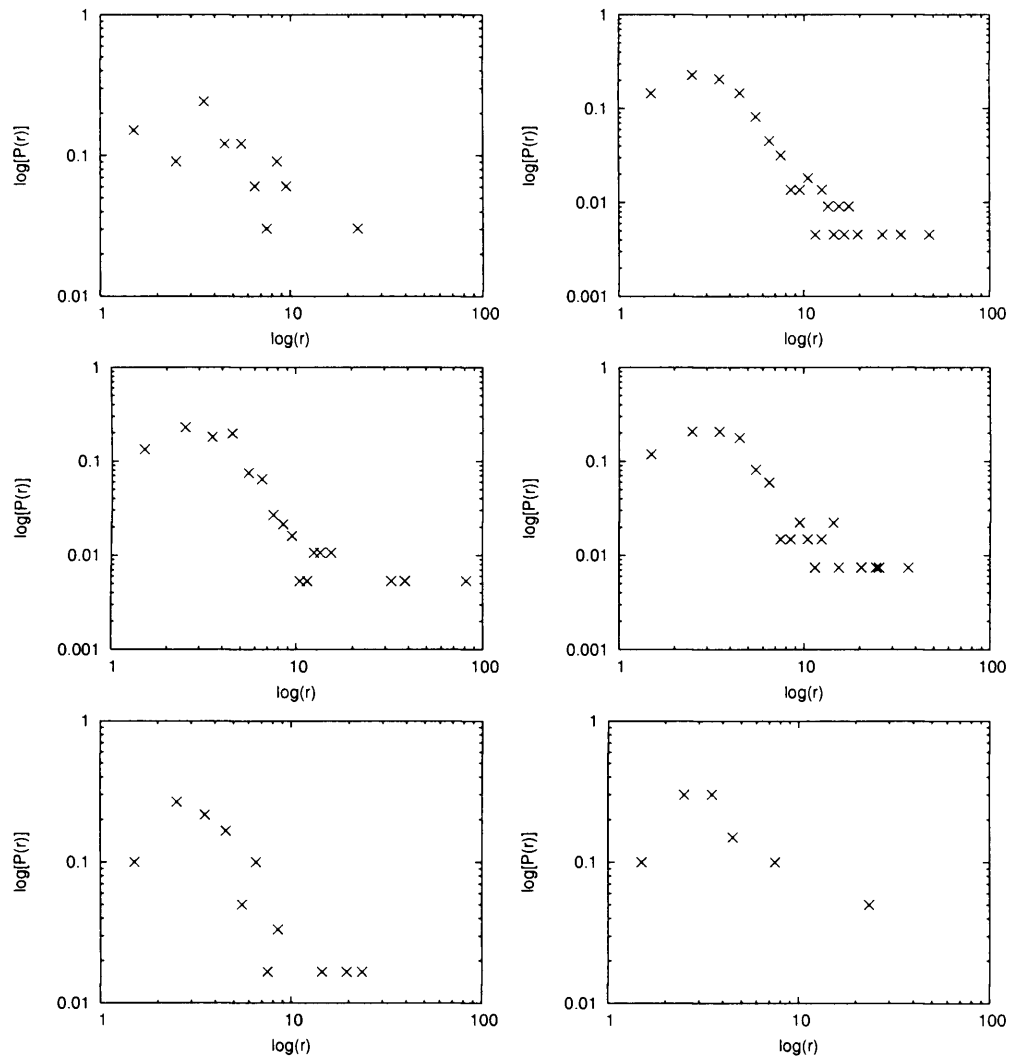


FIGURE 5.24. Distribution of abundance for a given in-coming degree of connectivity such as (top left) in-coming degree of 1, (top right) in-coming degree of 2, (middle left) in-coming degree of 3, (middle right) in-coming degree of 4, (bottom left) in-coming degree of 5 and (bottom right) in-coming degree of 6. The distributions decay as power-laws..

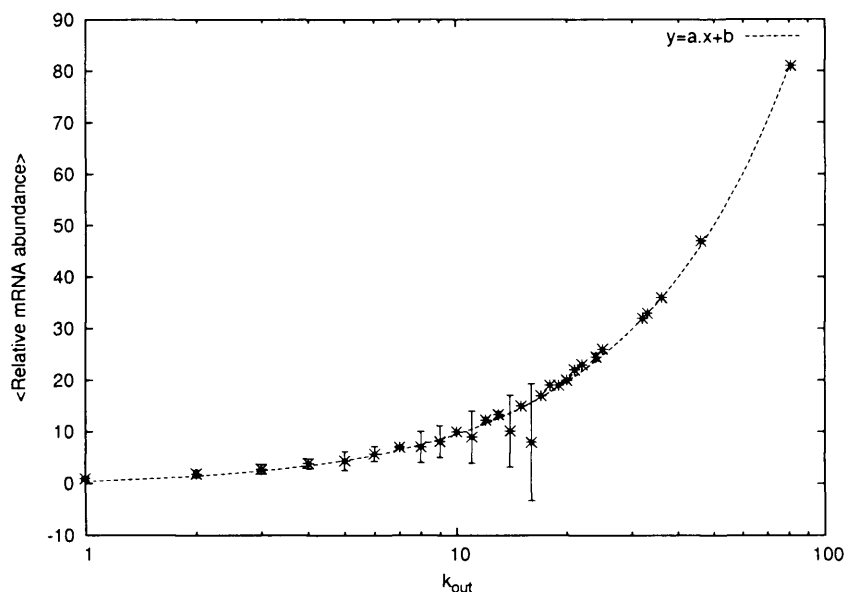


FIGURE 5.25. Averaged distribution of the abundance versus the out-going degree of connectivity of the corresponding nodes in the simulated network. The distribution is best fitted by a linear function  $y = ax + b$  with  $a \approx 0.1$  and  $b \approx -0.1$ .

resulting mRNA abundances for a network with period 32. About 87 per cent of the nodes are active, of which about 25 per cent have periodic production rates.

We show in figure 5.22 the distribution of abundance of the regulator. Each graph gives the distribution at a different time during one period.

All the active genes are taken into account, that is those of variable production rate and those of constant production rate. The tail of the distribution follows a power-law. Very slight variations from one time to another are noticeable.

As for *E. coli*, we check how in and out-degrees correlate to the abundance of regulators in the model. Figure 5.23 shows correlation data between in-degree and the averaged abundance of regulator at that degree. Similar to the *E. coli* experimental data, there is no noticeable correlation between degree and abundance. In this case again, the standard deviation for each value is high, and the distributions of the regulators show power-law tails.

We show in figure 5.24 the distribution of abundance for each of the in-degree of connectivity. The figures are similar to the one observed earlier from the *E. coli* dataset.

Finally, figure 5.25 shows correlation data between out-degree and the averaged abundance of regulator at that degree. The correlation appears clearly on this figure

and is linear. Thus, the higher the out-degree of connectivity, the higher the abundance. This figure looks similar to the corresponding graph in *E. coli*, where for small out-degree of connectivity the increase in abundance is relatively small and for higher degrees, the increase is much larger.

## 5.6 Summary

In this chapter we have approached the problem of linking what is known of the topology of genetic regulatory network and functional observation made of it. From a simple network but yet with complex behaviour we have seen that it was possible to reproduce features of the measured data.

We have seen that desynchronisation of the network may be a cause of the observed complex time dependence between time-series. This reinforces the idea that the synchronisation of a cell population cannot be accurate.

Other effects like those of lifetime dependence or of elapse time activity on the behaviour are important phenomena to take into account. However, they do not seem to play a major role at this scale.

Note that a possible explanation for the unimodality of the normalised distribution of abundance of mRNA for *S. cerevisiae* and *C. crescentus* could be sought in the way the cell cycle dependent patterns of expression are extracted. In the *C. crescentus* dataset for example, the identification is done using a Discrete Transform Analysis.

But the most important point was the inconsistency of the in and out distribution at the nodes between the biological system and the model. False assumptions about the architecture of a regulatory network, particularly about the probability distribution of the numbers of links per node and the form of the node function, can nevertheless yield networks that give accurate simulations of the mRNA abundance data. From this, a regulatory network cannot be inferred from a time series of mRNA abundance data alone.

We have shown that the experimental data for the mRNA abundance and genetic regulation in *E. coli* show no correlation between the number of regulators of a gene and the corresponding mRNA abundance. There is, on the contrary, a distinct correlation between the number of genes regulated and the mRNA abundance. It is perhaps unexpected that we should find such correlation, since this appears to require genes to 'fulfil their duty' by producing more transcription factor the more controls require it. While

strange at first sight (how does a gene know how many others it regulates?) this might be expected on evolutionary grounds, namely that mutations can be beneficial only if the structures already exist to take advantage of them. Others say, mutations can only be exploited if the controls exist, and so we shall let the data speak for themselves.

Our simulation of a simple regulatory model taking into account the correct correlation assumption allows us to reproduce, at least qualitatively, the observed data on the distribution of mRNA abundances. The model uses realistic parameters derived from the *E. coli* network, namely the same proportion of negative links, the same mean connectivity and the same dependence of mRNA production on out-going links. (We can show that connecting the abundance with the incoming distribution does not work.) These properties alone, and not the intimate details of the regulatory network, are sufficient to reproduce the qualitative features of the measured mRNA abundances in *E. coli*.



# Chapter 6

## *Maximum entropy reconstruction of networks*

### 6.1 Introduction

High throughput experiments have generated a large amount of data on genetic expression. The fact is that those data do not contain explicitly the interactions between elements of the systems they originate from. However, it is assumed that those data should contain information on the underlying structure and topology of the system. Inferring the network related to the data is achievable by a process of reverse engineering which differs from all the methods presented so far.

Reverse engineering a network from genetic expression is not however a simple task. Ideally it would be if the technology was not limiting in generating enough data and in providing powerful machines to treat them. This is not the case however and numerous methods have been proposed to infer networks on the basis that there is not enough data to reconstruct them directly and exactly. In the field of astronomy, amongst many others, the method of the maximum entropy has been proven efficient for reconstructing images from noisy and partial data. We apply here this method to the reconstruction of genetic regulatory networks since it occurs in the same context of partial and noisy data.

In this chapter, we will first present the relevant methods of genetic network inference based on mutual information and Bayesian network theories. We will then introduce the maximum entropy method in its most traditional way and see how it can be applied to the reconstruction of genetic regulatory networks. Finally, we will benchmark this

method by reconstructing the structure of known networks from simulated datasets. Results show that reconstructed networks present a similar structure to the expected one. This also shows the extent to which the data contain information on the structure of the underlying network. Compared to other methods, the maximum entropy method is capable of reconstructing relatively large scale networks in a relatively short time.

## 6.2 Methods to infer networks from data

Several methods have been developed to infer the structure of biological networks from mRNA data. These methods differ in that some of them only yield the architecture of the network, providing the correlation between nodes or the causal relationships, whereas some others may in addition provide the regulatory functions associated to the nodes. Of course, for the latter case, the function relates to the dynamical context of the network, that is whether it is a discretised or a continuous model, that is Boolean or based on differential equations. Here, we will only focus on the former.

Varied as the methods are, and in particular the more refined ones, they present the common characteristic of using prior knowledge of the system. This is a very important step, since it reduces the complexity of the problem by constraining the space of possible solutions. The other important notion that emerges from those various models is the trade-off, in order not to bias the results, between making appropriate assumptions on the system and the use of algorithms making a neutral use of the data. In that respect, the example of the reconstruction of transcriptional co-regulatory networks is a good illustration. In this case, the general assumption on the system is that of the similarity of the pattern of expression of genes that are co-regulated by the same transcription factor. The measure of expression of the elements of the network is the relative mRNA abundances obtained from micro-array experiments, and the identification of similar patterns is carried out via clustering algorithms. It turns out that the result of that clustering is highly dependent on the algorithms used (D'haeseleer, Liang & Somogyi, 2000) and therefore is not 'neutral'. Nothing can then distinguish them beforehand except, to some extent, the prior knowledge on the system. Furthermore, the assumption bases the co-regulation on one sort of mechanism while many others, such as combined co-regulation for example, could be considered.

The inference of other genetic networks is far more complex than that of co-regulated genes. We will present here the main methods used to infer networks on which the latest up-to-date methods are based. The first of those methods is related to the use of mutual

information to reconstruct genetic networks as Boolean networks. The second one is that of Bayesian networks. This method assumes nothing about the regulatory functions associated to the nodes and considers only probabilities of interaction between elements of the network. Finally, we present an alternative method that infers genetic networks, as Boolean networks, from perturbed genetic expression.

### 6.2.1 Mutual information

In this method, the genetic network is represented by a Boolean network where the nodes are the genes and the links represent the functional interactions between genes. Of course in this case, the regulatory functions of the genes are implicitly assumed to be Boolean.

The method relies on state transition and temporal response of gene expression patterns to perturbations and changes. It is based on the use of mutual information,  $M$ , defined in section 1.7, which is used to extract relationships between input and output states of the nodes in the Boolean network (Liang et al. 1998). In such networks, the input state of node  $i$ ,  $X_i$ , is completely determined by the output state of one node  $j$ ,  $Y_j$ , if

$$M(X_i, Y_j) = H(X_i),$$

or equivalently, according to Eq. 1.5, if

$$H(X_i, Y_j) = H(Y_j).$$

If it is not completely determined by the output state of node  $j$  then correlations involving 2 output nodes are sought. This is repeated, up to  $N - 1$  output nodes, until the input state of node  $i$  is completely determined by the output states of  $k$  nodes, that is

$$H(X_i, Y_j, \dots, Y_{j+k}) = H(Y_j, \dots, Y_{j+k}).$$

This gives the minimal number of regulators and characterises the existence of  $k$  links, from the corresponding nodes, to node  $i$ . The representation of those entropies in a Venn diagram would give the overlapping of each of the ensembles  $X_i$ , and  $Y_j, \dots, Y_{j+k}$ .

The directionality of the links, in the present case, is determined from the causal relationship between the state of the nodes which is given by the temporal response and not from the expression for mutual information which is symmetrical. In any case, correlation alone is not enough to determine the causal relationship between variables as  $X$  acting on  $Y$  and  $Y$  acting on  $X$  are equivalent in that respect. Another way to

determine causal links between nodes, aside of using time series, is to proceed through perturbation experiments as explained later.

The quantity of data required for such methods is very high for large networks as for a Boolean network, all possible  $2^N$  input-output pairs need to be known. The number of data needed decreases when constraints are introduced, for example on the mean connectivity. This is an important consideration since the quality of the reconstruction depends on it (D'haeseleer et al. 2000).

Other algorithms using mutual information have been proposed (Akutsu, Miyano & Kuhara, 2000*b*), including one inferring Boolean networks from noisy data (Akutsu, Miyano & Kuhara, 2000*a*).

### 6.2.2 Bayesian networks

Another way to infer the structure of gene regulatory networks is to consider the probabilistic model of Bayesian networks. This approach is based on the idea that if the expression level of a gene  $a$  is regulated by genes  $b$  and  $c$ , the expression level of  $a$  is then the joint activity levels of  $b$  and  $c$  (Friedman 2004). As before, the directed links represent the direct dependencies between the nodes. This approach does not require the formulation of assumptions on the underlying mechanisms of interaction or regulation between the nodes.

In a Bayesian network, each of the expression levels is considered as a random variable  $X_i$  of level  $x_i$ . The network itself is defined as the representation of a joint probability distribution of all the random variables of interest, that is the measured variables and the other adequate variables (Friedman et al. 2000). The joint distribution of the network is represented as the product of conditional probabilities. For a random variable  $X_i$ , the associated conditional probability is given by  $P(X_i|U_i)$ , where  $U_i$  is the set of variables called the parents of  $X_i$ . This corresponds to the nodes directly controlling  $X_i$ . For the complete network, the joint probability is then

$$P(X_1, \dots, X_N) = \prod_i P(X_i|U_i),$$

which requires an acyclic network and assumes variables  $X_i$  are independent of their non-descendants, given their parents. It is assumed that only the value of the parents influences its descendant. Indirect relationships between variables are therefore considered also as direct which may, in return, have an effect on the architecture of the network. The

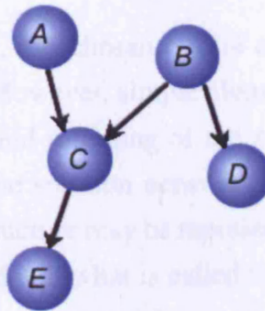


FIGURE 6.1. Directed network.

causal relationship parent-descendant arises again from temporal data or from perturbation experiments, and the resulting network is in this case directed. Take for example the network in figure 6.1. The relation between the joint probability of the complete network and the conditional probabilities associated to the nodes is given by

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B)P(E|C).$$

The inference of the architecture of the network is done via a learning process using a training data set.

In the present case, the variables are the mRNA expression levels which can be used as is or discretised in order to account for non-linear dependencies. Reconstructing networks of thousands of genes requires a huge number of expression profiles. For instance, reconstructing a network of 35 genes with binary states, where each of the genes is controlled by a maximum number of four other genes, requires a number of the order of 20,000 expression profiles (Le, Bahl & Ungar, 2004). Reconstructing networks of thousands of genes is then a challenge relative to the quantity of data available.

In general and in practice, many different instances of the reconstructed network may satisfy the data. This variety of solutions is also reinforced by the use of a dataset smaller than required. It is then necessary to use specific methods, based on scoring functions, to discriminate between those different solutions. In the latter case, the reconstruction may present a certain sensitivity to the dataset used to infer the network (Friedman & Koller, 2003). This adds to the problem of evaluating the correct solution. On the other hand, prior knowledge of the network structure, from a literature search for example, significantly reduces the amount of data required and then narrows the space of possible solutions (Le et al. 2004). But in general, the number of solutions grows exponentially

with the size of the network.

Amongst the many solutions, the chosen one is considered to be the most likely one, given the scoring function. However, simple algorithms yield equally high scoring solutions of different structures and sampling of the many solutions is preferred. The idea behind sampling is that if the solution networks appear to be different in details, strong correlations within their structure may be represented over all the networks. Those special features are then extracted and what is called the posterior probability,  $P(f|D)$ , is measured. This is the Bayesian score which corresponds to how likely the feature  $f$  appears to be considering the data  $D$ .

There are other methods to infer the structure of networks from expression data and other probabilistic models to represent them (Friedman 2004). Other methods mix different approaches (Li et al. 2005) including, for example, the use of Bayesian and mutual information methods which has been recently proposed (Li & Chan, 2004).

### 6.2.3 Inference from perturbation strategy experiments

As seen above, the data used in inferring networks is a major issue. A complementary approach to improve the reconstructed network is to use perturbation strategy experiments. Those perturbations, by changing the level of expression of genes by either direct interventions onto the genes or by less direct routes, should help to distinguish between equivalent classes of networks, for example. This implies the reconstruction of static networks as the effect of perturbations that are measured at steady state expression level of the genes.

Finally, there is an inference technique based solely on perturbation experiments (Ideker, Thorsson & Karp, 2000). In this method, the states of the genes are represented in a matrix,  $E$ , where element  $e_{ij}$  is the Boolean state of gene  $j$  in the perturbation experiment  $i$ . The interactions between genes are determined as follows. For a given gene  $j$ , all pairs of permutation experiments  $(k, l)$  where  $e_{kj} \neq e_{lj}$  are selected. For each of those pairs, the set  $S_{kl}$  of the genes that are in different states are extracted and, a minimal subset is identified so as to contain at least one gene present in all of the  $S_{kl}$  sets. The function at node  $j$  is then identified according to the data from  $E$  directly. All possible subsets are considered as possible regulators and further perturbation experiments are required to discriminate between them.

However, the range of possible perturbations is wide and the problem for such an approach is the selection of the perturbations so as to benefit the inference of the network.

In the present case, the next perturbation to be carried out is the one that maximises the entropy, or information, of the experiment matrix.

This perturbation strategy has been also approached in the context of Bayesian networks. One of the techniques used to find the useful perturbations is to determine which perturbed element is going to split the equivalence class into many subclasses of equivalent size (Pournara & Wernisch, 2004).

#### 6.2.4 Summary

We have seen the importance of a sufficient amount of data to infer network structure. Typically, the methods of network inference are adequate to reconstruct small network structures. But even in these cases, those methods may yield many solutions that satisfy the data. Furthermore, the solutions turn out to correspond to different networks. In addition, the method of discrimination between solutions varies from one method of inference to another which, despite their apparent pertinence, may be somewhat arbitrary.

### 6.3 The maximum entropy method

The maximum entropy method is used in the reconstruction of images from imperfect data and has been proven successful in various fields. It is used, for example, in astronomy to reconstruct images using gravitational lenses or in biology for the reconstruction of microscopy images, etc (Buck & Macaulay, 1991). This method applies to cases where there are not enough data to reconstruct exactly the image. A simple fitting process, for example, would lead to many possible solutions and leave us with the problem of deciding which one is the correct one. In the maximum entropy method, this is taken care of by the maximisation of the entropy ensuring that the reconstructed image is the most probable image given the data.

The lack of data is also the fate in genetic regulatory networks inference, as mentioned earlier, with the identical problem of discriminating between the many solutions. It would be therefore of interest to apply the maximum entropy method to the reconstruction of such networks, that would be here equivalent to the reconstructed image. In this section, we first present the classical maximum entropy method used in image reconstruction. We then show how to adapt this method in the context of inferring the structure of networks.

### 6.3.1 Traditional maximum entropy method

The image of an object seen through some equipment, with Gaussian noise, can be expressed as

$$D = R.f + \sigma.n, \quad (6.1)$$

where  $D$  are the acquired data from the image  $f$ ,  $R$  is the effect of the equipment on the image, also called the *point spread function*,  $\sigma$  is the noise standard deviation and  $n$  is a random number drawn from a standard normal distribution. The aim of the method is to reconstruct the image in the case that there are not enough data to do it exactly. The problem is then to look at the most probable image that reproduces the data, which turns out to be equivalent to maximising the entropy of the system, here the image, while diminishing the  $\chi^2$  of the fit to the given data. We present here, in more details, the principle of the historical maximum entropy method as used in image reconstruction.

The principle behind image reconstruction derives from the Bayes' theorem which states that

$$P(f|D) = \frac{P(f)P(D|f)}{P(D)}$$

with the probabilities described as follows:

- $P(f)$  corresponds to the *prior probability*, that is the probability distribution assigned to image  $f$  before acquiring data. It plays an important part in this context of reconstruction where not enough data are available to reconstruct exactly the image.
- $P(D|f)$  is called the *likelihood*, that is the conditional probability to acquire the particular data  $D$  given the image  $f$ . It is dependent on the equipment used to acquire the data as well as on some noise which is related to imprecision in the measure of data  $D$ .
- $P(D)$  corresponds to the *evidence* and is considered in this case as a scaling constant.
- $P(f|D)$  is the probability we are looking for. Indeed, since we want an image that is the best reconstruction, we are looking for the most probable one. This probability quantifies then the inference about the image.

The prior probability is related to the entropy,  $S$ , of the image as

$$P(f) \propto e^{\lambda.S}.$$



If the difference between the measured data and reconstructed one is non zero, it is due to noise. In the case the noise has a Gaussian distribution, that is  $n$  drawn from a unit normal distribution, the likelihood varies, using Eq. 6.1, as

$$P(D|f) \propto e^{-\frac{1}{2}\chi^2}.$$

The probability we are looking for,  $P(f|D)$ , varies then as

$$P(f|D) \propto e^{\lambda.S - \frac{1}{2}\chi^2},$$

and finding the most probable image is then equivalent to maximising expression

$$\lambda.S - \frac{1}{2}\chi^2.$$

From here, the approach is very similar to the one used before in chapter 2 that consists, in part, in the maximisation of the entropy associated to the system. There are several ways to introduce the maximum entropy method for image reconstruction but we will present here the one relevant to our purpose.

The overall method consists in minimising the Lagrangian function  $Q = \lambda.S - \chi^2$  where  $\lambda$  is the Lagrangian parameter.

The entropy,  $S$ , of an image is defined as

$$S(f) = - \sum p_i \log p_i,$$

where  $p_i$  is the probability to find an object (pixel) of intensity  $f_i$  such that  $p_i = \frac{f_i}{\sum_j f_j}$ . The log in the expression requires the positivity of the intensity, which arises as prior information on the system.

Let  $h$  be the default distribution of probabilities when there are no constraints associated to the entropy, so that when the entropy is maximised we have  $p_i = h_i$ . The entropy function can then be expressed as

$$S(f) = - \sum p_i \log \frac{p_i}{h_i}. \quad (6.2)$$

The statistic between the reconstructed image and the real data is given by the  $\chi^2$  function so that

$$\chi^2(f) = \sum_i \frac{(D - D')^2}{\sigma^2},$$

where  $D'$  are the simulated data compare to the observed ones,  $D$ . The function to maximised is then

$$\lambda.S(f) - \chi^2(f) = -\lambda \sum p_i \log \frac{p_i}{h_i} - \sum_i \frac{(D - D')^2}{\sigma^2},$$

which, regarding its nature and the size of the problem, is done iteratively (Skilling & Bryan, 1984).

### 6.3.2 Application to network reconstruction

The attributes of the networks we want to infer are similar to the ones discussed in chapter 4. It has nodes connected to each other by directed links. A links can be either an activator or an inhibitor. As before, the network is described by its adjacency matrix, say  $A$  of elements  $a_{ij}$  where

$$a_{ij} = \begin{cases} +1, & \text{if there is an activator directed link connecting node } j \text{ to node } i \\ -1, & \text{if there is an inhibitor directed link connecting node } j \text{ to node } i \\ 0, & \text{if there are no directed links connecting node } j \text{ to node } i \end{cases}$$

Knowing the image reconstruction method by maximum entropy, we apply it to network reconstruction. We describe in this section how the method is adapted to network reconstruction and how the model is implemented.

An analogy can be made between image and network reconstruction in such a way: the nodes of the network to be reconstructed correspond to the pixels of the image. The level of abundance of the product of the nodes, that is mRNA in the case of genetic networks, is analogous to the intensity of the pixels. In the same way, the production function at a node is similar to the point spread function mentioned earlier. And finally the noise accounts for all the possible imprecision that may arise, from the regulatory mechanisms themselves in the case of real systems, but also from the acquisition techniques, for example. (However, in the present model, imprecision noise from regulatory mechanisms is not implemented). In a general form, the relation between network and data can be expressed as

$$D = g(O) + \sigma.n,$$

where  $D$  corresponds to the data,  $g$  is the production function,  $O$  is the matrix representing the original network,  $\sigma$  is the noise standard deviation and  $n$  is a random number drawn from a standard normal distribution.

Similar to the image reconstruction, what we seek here is the reconstruction of a network that gives the best fit to the measured data by maximising the entropy,  $S$ , of the system while minimising the  $\chi^2$ , that is maximising the Lagrangian function

$$Q = \lambda.S - \chi^2, \quad (6.3)$$

with  $\lambda$  a Lagrangian multiplier.

### Maximisation process

The elements of the adjacency matrix of the network we want to reconstruct are chosen from a set of values  $\{-1, 0, 1\}$ . A straight forward way to maximise the function  $Q$ , (Eq. 6.3), is to use a Monte Carlo like process. The general algorithm to the reconstruction of the network is described as follows. An adjacency matrix  $A$  is first generated at random and the data  $D'$  are calculated using a function  $g$  such that

$$D' = g(A).$$

The value  $Q_o$  of  $Q$  is then calculated from the entropy function  $S$  and from  $\chi^2$ , which are given explicitly below. An element of the matrix  $A$  is then modified and the data  $D'$  corresponding to the new network are recalculated as well as the new value  $Q_n$ . The modification of  $A$  is then accepted if  $Q_n > Q_o$ , and accepted with a probability  $< e^{\frac{Q_n - Q_o}{\alpha}}$  if not, where  $\alpha$  is a hyper-parameter analogous to a temperature. The new configuration is rejected otherwise. The modification process of the adjacency matrix  $A$  and its acceptance or rejection is carried out until  $Q$  converges to a maximum.

### Function of production

The production function that gives the level of abundance of the element produced at each time step,  $g$ , is at the heart of the dynamics of the networks and determines the behaviour of the system. This function needs first to be defined in order to provide an explicit expression for  $\chi^2$ , which depends on it. In the following, we assume that the function  $g$  associated with the network to be reconstructed is known.

As an illustration of the reconstruction method, we choose the function to be, arbitrarily, a simplified version of the one previously introduced in section 4.3.4. In this version, the level of expression does not depend on the state of the nodes but only on the in-coming degrees of connectivity of the nodes. The level of expression of node  $i$  at

$t + 1$ ,  $d_i(t + 1)$ , is dependent on the abundance vector  $d(t)$  such that

$$d_i(t + 1) = \sum_j o_{ij} d_j(t),$$

where  $o_{ij}$  are the elements of the adjacency matrix  $O$  of the network. The production function is then

$$g(O) = O \cdot d(t).$$

from which it is possible to build a time series of data  $d(t_0), d(t_1), \dots, d(t_n)$ .

The other difference with the simulations in chapters 4 and 5 is that here, the elements in the matrix of the network representing the incoming links are normalised so that their sum is unity. The exact production function we use is then

$$d_i(t + 1) = \sum_j \frac{o_{ij}}{\sum_l o_{il}} d_j(t). \quad (6.4)$$

This normalisation step prevents the time series level of abundance from ‘exploding’ after a few time steps. This, however does not alter the fact that links are still either activating ( $a_{ij} > 0$ ) or inhibiting ( $a_{ij} < 0$ ).

### Expression for $\chi^2$

Consider the set  $\{d(t_0), d(t_1), \dots, d(t_n)\}$  of the measured data from the unknown network represented by adjacency matrix  $O$  and the set  $\{d'(t_1), d'(t_2), \dots, d'(t_n)\}$  of data obtained from an arbitrary network represented by its adjacency matrix  $A$  such that

$$d'_i(t_1) = \sum_j \frac{a_{ij}}{\sum_l a_{il}} d_j(t_0),$$

and

$$d'_i(t_2) = \sum_j \frac{a_{ij}}{\sum_l a_{il}} d'_j(t_1),$$

with  $d'(t + 1) = d(t + 1)$  for  $A = O$ , for example. The difference that arises between datasets measured from network  $O$  or generated from network  $A$  is given by  $\chi^2$  so that

$$\chi^2 = \sum_{i=1, t=1}^{N, n} \frac{(d_i(t) - d'_i(t))^2}{\sigma^2},$$

where  $\chi^2 \rightarrow 0$  when  $A$  approaches a solution. This solution is the most probable but may not be  $O$ .

### Entropy function

As shown in chapters 2 and 3, there are many ways to account for the entropy of a network. The general simple form of the entropy, which we recall here, is, up to a constant

$$S = - \sum_i p_i \log p_i. \quad (6.5)$$

The issue is then to define the probabilities  $p_i$  so as to serve adequately the reconstruction method. In the case of the reconstruction of networks, there are many possibilities, depending on what we wish to reconstruct.

In the traditional use of the maximum entropy method in image reconstruction, what is reconstructed of the image is the intensity  $f_i$  of the pixels, and the probability  $p_i$  is then defined as the probability of finding a pixel of intensity  $i$ . By analogy to a network, the adjacency matrix could be seen as an image where each of its elements is a pixel. In the simple case of our adjacency matrix  $A$ , there would be three levels of intensity corresponding to the three possible elements:  $-1$ ,  $0$  and  $+1$ . The probability  $p_i$  used in Eq. 6.5 is then of finding a negative a positive or no link at all, and the reconstructed network would have the most probable proportion of negative links and the most probable mean degree of connectivity given the data.

Conversely to images, the network architecture can be defined from a variety of probability distributions amongst which is the distribution of the degree of connectivity, defined earlier. In the present case, this is the most appropriate choice as we want to reconstruct the architecture of the network from information contained in expression data and that we know, from the production function we use, the data should bear such information. Furthermore, since the networks are directed the probability  $p_i$  will be that of finding a node of in-coming degree  $i$ . Other possible probability distributions related to the architecture could also be used, such as the one that relates to the degree of clustering of the nodes or of other coefficient  $\Gamma_i$ , as given in section 3.4, where distributions of probability could be extracted.

We assign also a default probability distribution  $q$  to Eq. 6.5 such that the entropy function becomes

$$S = - \sum_i p_i \log \frac{p_i}{q_i}. \quad (6.6)$$

If the data bear no information at all on the architecture with respect to the probability  $p_i$ , the entropy  $S$  will be maximum for  $p = q$ . Any departure of  $p$  from  $q$  will indicate there is some information on the architecture of the network contained in the data. A

reasonable assumption is to set  $q_i$  exponentially distributed so that Eq. 6.5 becomes

$$S = - \sum_i p_i \log \frac{p_i}{e^{-(i-\kappa)^2}},$$

where  $\kappa$  is the prior mean connectivity that has then to be assigned beforehand.

### 6.3.3 The simulation

We realise a series of simulations to illustrate and validate the method, in which the measured data time series is generated from the known adjacency matrix  $O$ . The simplification we made on the production function  $g$  has the consequence on the dynamics that the abundance vectors fall rapidly into a fixed point. This does not allow a great flexibility for testing the effects on the size of the sample data adequate for the reconstruction of the network.

A workaround is to generate data from various initial data vectors  $d^{(0)}(t_0), \dots, d^{(n)}(t_0)$  so that the measured data set, using matrix  $O$ , is  $d^{(0)}(t_1), \dots, d^{(n)}(t_1)$  while the data set generated from the inferred network  $A$  is  $d''^{(0)}(t_1), \dots, d''^{(n)}(t_1)$ . This allows us to further simplify the model so as to speed up the simulation.

The maximisation process requires us to modify an element  $a_{ij}$  of the network and thereafter to recalculate the effect of that modification on the data. Replacing element  $a_{ij} = x$  by element  $a_{ij} = y$  affects only data  $d_i^{(j)}(t_1)$  and only those data need to be recalculated. This simplification can be pushed a little bit further. Let  $\eta_i$  be the number of in-coming links at node  $i$  before the modification so that

$$\eta_i = \sum_j a_{ij},$$

and  $\eta'_i$  the number of in-coming links at node  $i$  after the modification, that is

$$\eta'_i = \eta_i - |x| + |y|.$$

Following the modification, only the values  $d_i''(t)$  are modified. Let us label the values after modification with a double prime. We have

$$d_i''^{(k)}(t+1) = \left( \sum_j \frac{a_{ij}}{\eta_i - |x| + |y|} + \frac{-|x| + |y|}{\eta_i - |x| + |y|} \right) d_j^{(k)}(t),$$

that is, with Eq. 6.4

$$d_i^{(k)}(t+1) = \frac{d_i^{(k)}(t+1)\eta_i + (|y| - |x|)d_j^{(k)}(t)}{\eta'_i}.$$

Thus, during the process of maximisation of  $Q$  one need only update a number of data corresponding to the number of time measurements which would be the square of the number of nodes in the network otherwise.

Finally, the function to maximise is

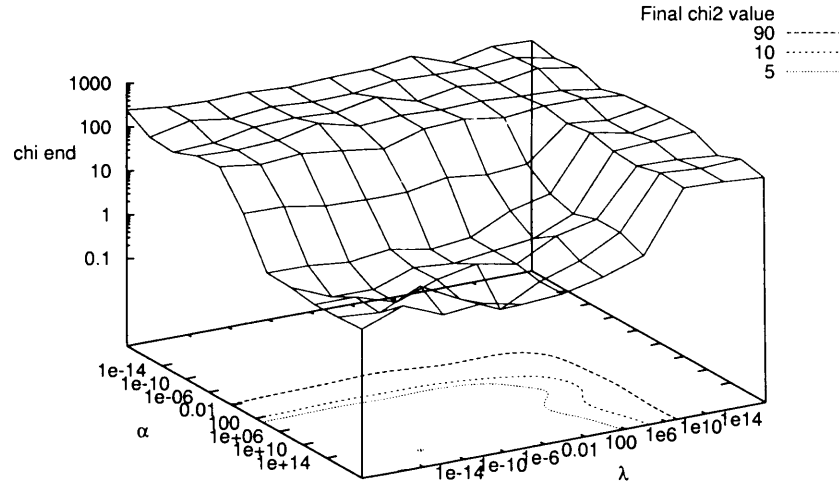
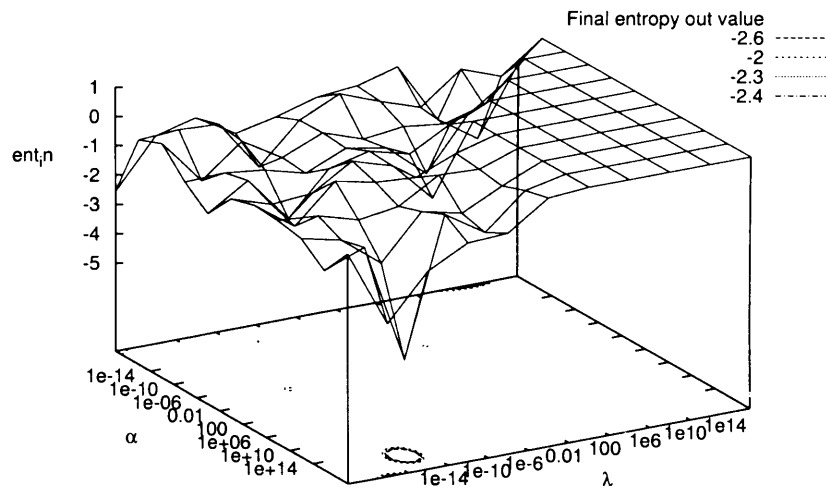
$$Q = \lambda.S - \chi^2 = \lambda \sum_i p_i \log \frac{p_i}{e^{-(i-\kappa)^2}} - \sum_{i,j} \frac{(d_i^{(j)}(t_0) - d_i^{(j)}(t_1))^2}{\sigma^2}.$$

## 6.4 Results

In this section, we first illustrate the maximum entropy method for the reconstruction of networks using a data set generated from a known network. In the following, the generated data are also referred to as the ‘measured’ data as it would be what the reconstruction is based on. The networks we use to generate the data have different degree distributions and it is this particular property of the network we are interested in reconstructing. We then look at the effect of the size of the generated dataset in order to show, qualitatively, the effect of under and over sampling the system we want to reconstruct. This effect should show up as a departure of the reconstructed degree distribution from the function  $q$  in Eq. 6.6.

### 6.4.1 Determination of $\alpha$ and $\lambda$

Before proceeding, we need to determine the range of values for which parameters  $\alpha$  and  $\lambda$  yield a result where  $\chi^2$  is minimised while  $S$  is maximised. This is done by doing a gross screening over a wide range of values for the reconstruction of a small network of 30 nodes. Figures 6.2 and 6.3 gives the values of  $\chi^2$  and  $S$ , respectively, after a large simulation time step so that they are considered to have converged. The value of  $\chi^2$  is relatively small for  $10^2 \lesssim \alpha \lesssim 10^{14}$  and  $10^{-14} < \lambda < 10^{-1}$ . However, it is not guaranteed that any set of values within those ranges will yield the solution of maximum  $Q$  and not a solution corresponding to a local maximum. However, this issue relates mainly with the algorithm we use, and in an attempt to reach the best solution, we allow ourselves to change parameters  $\alpha$  and  $\lambda$  within the proposed range during the simulation

FIGURE 6.2. Final  $\chi^2$  for various values of  $\alpha$  and  $\lambda$ .FIGURE 6.3. Final entropy for various values of  $\alpha$  and  $\lambda$ .



so as to improve the result. We assume that the range of parameters is valid for the rest of the simulation, and in the results presented below the value of the parameter  $\alpha$  is chosen in the range  $[10^1, 10^6]$  and the value of parameter  $\lambda$  is chosen in the range  $[10^{-10}, 1]$ .

### 6.4.2 Reconstruction of the degree of connectivity

In this section, we apply the maximum entropy method to the reconstruction of a random and a power-law network. Both networks have  $N = 100$  nodes,  $\bar{k} = 8.0$  and a proportion of negative links  $\mu = 0.0$ . We have generated 500 data points which are used as the ‘measured’ data set to reconstruct the network.

#### Random network

The random network used to generate the set of data is constructed as described in chapter 1.

We show in figure 6.4 the variation of  $\chi^2$  and  $S$  during the reconstruction process. The value of  $\chi^2$  which has to be minimised decreases as expected. Conversely, the value of  $S$  decreases while we are looking to maximise it. This is easily explained by the fact that the reconstruction of the network starts from a random network for which the entropy is maximal. The entropy of the final reconstructed network has therefore to be less or equal to that of the initial network. The scatter plot of the reconstructed data versus the ‘measured’ ones presented in figure 6.5 show that the reconstructed data are close to the ‘measured’ ones.

The reconstructed in-coming degree distribution has a form similar to that of the original one as shown in figure 6.6.

#### Power-law network

The power-law network used to generate the set of data is constructed using the Barabási and Albert method as described in chapter 1.

We show in figure 6.7 the variation of  $S$  and  $\chi^2$  during the reconstruction process, and in figure 6.8 the scatter plot of the reconstructed data versus the ‘measured’ data for the random network. The figures present the same characteristics as in the case of the random network: both variables  $\chi^2$  and  $S$  converge and the scatter plot shows that even if the data are not reconstructed exactly, they are close to the original ones. The

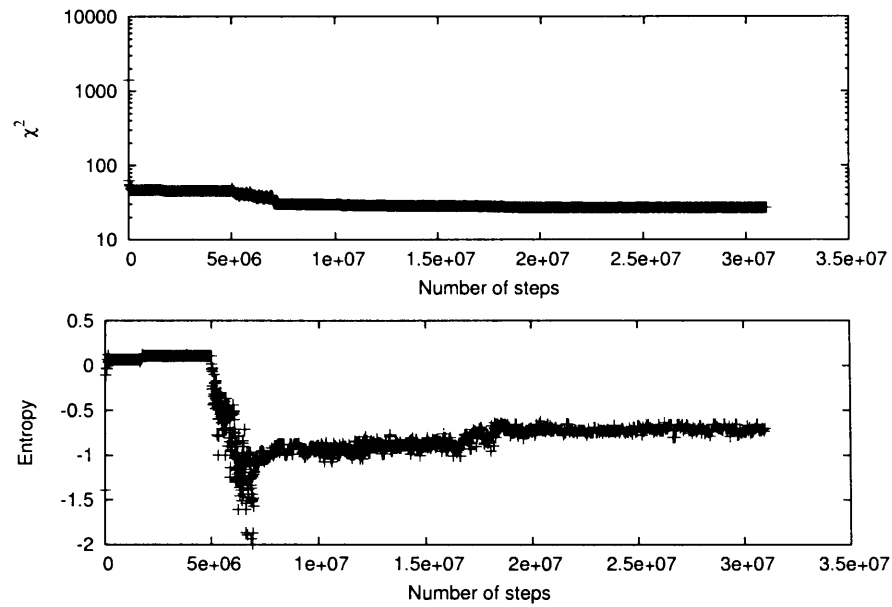


FIGURE 6.4. Variation of  $\chi^2$  (top) and of the entropy (bottom) during the maximisation process. The data used to reconstruct the network have been generated from a random network.

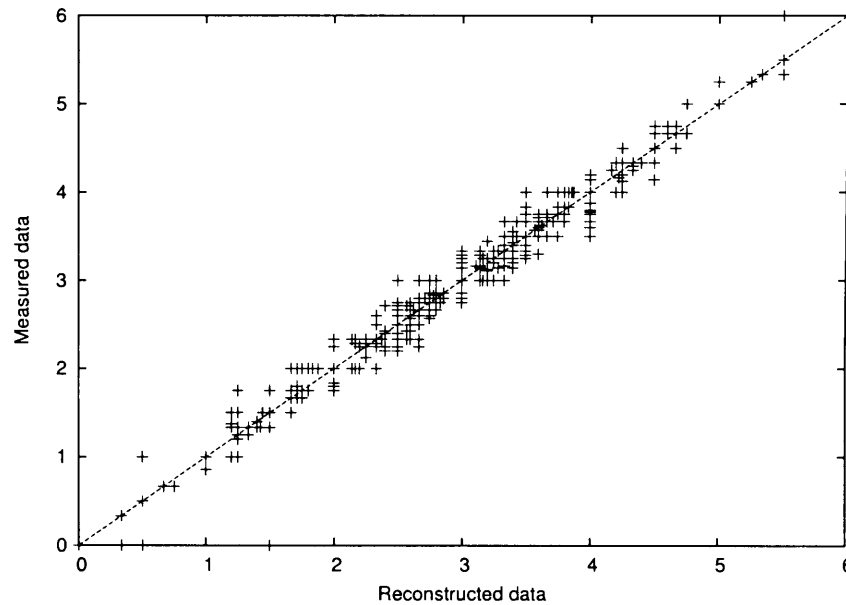


FIGURE 6.5. Scatter plot of the reconstructed data versus the original ‘measured’ data generated from a random network.

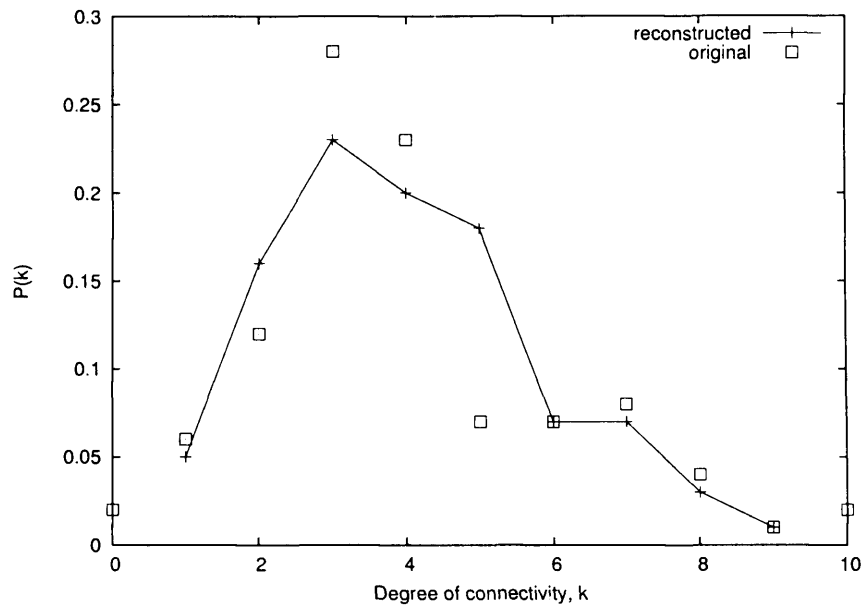


FIGURE 6.6. Comparison of the in-coming degree distribution between the original random network and the reconstructed network.

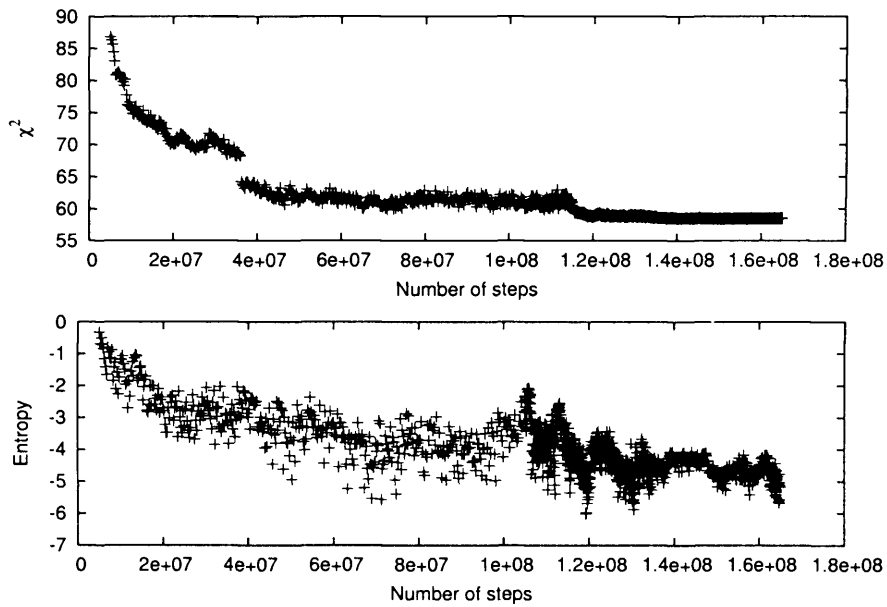


FIGURE 6.7. Variation of  $\chi^2$  and of the entropy during the maximisation process. The data used to reconstruct the network have been generated from a scale-free network.

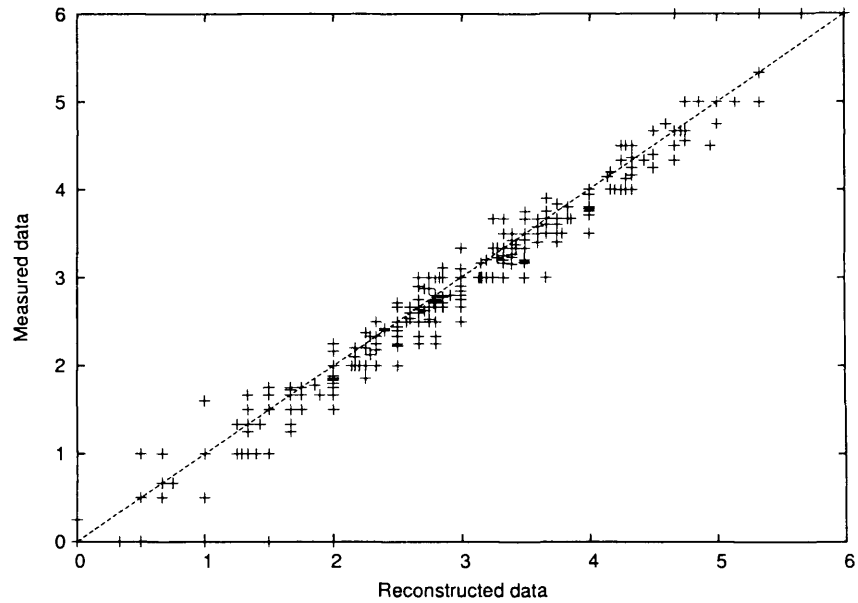


FIGURE 6.8. Scatter plot of the reconstructed data versus the original data generated from a network with a power-law degree distribution.

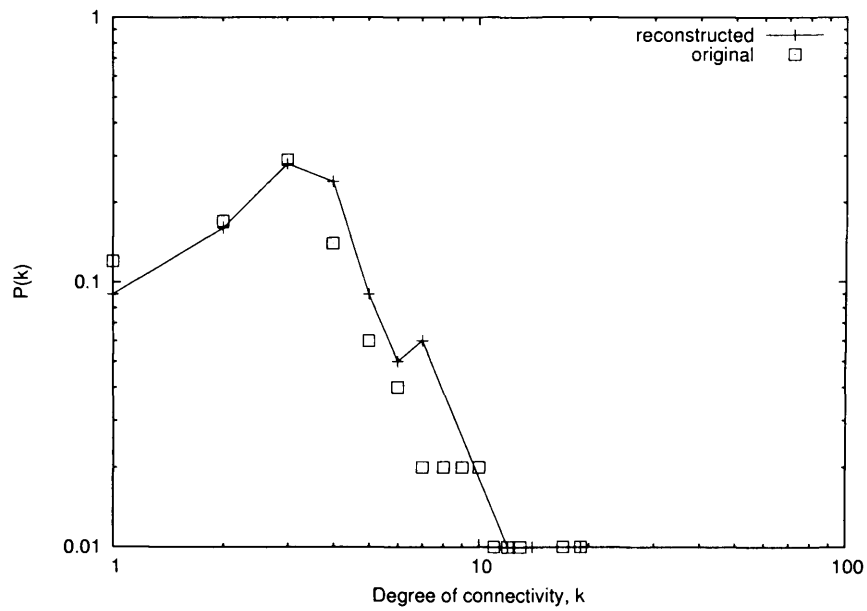


FIGURE 6.9. Comparison of the in-coming degree distribution between the original scale-free network and the reconstructed network.

reconstructed in-coming degree distribution has a form similar to that of the original one

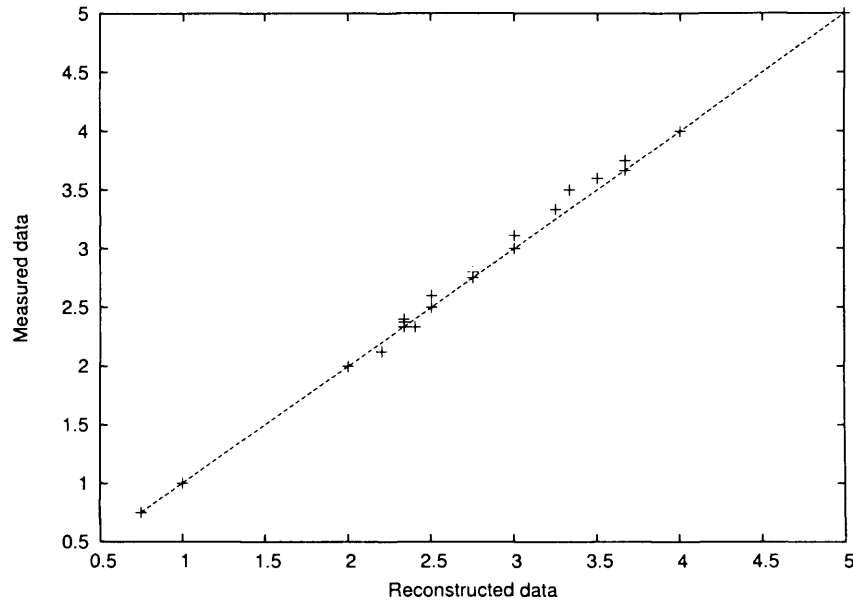


FIGURE 6.10. Scatter plot of the reconstructed data versus the original data for 30 data points.

as shown in figure 6.9.

### 6.4.3 Size of the sample

We look at the effect the size of the data sample has on the reconstruction of networks. We are particularly interested in identifying under and over sampling effects. For this we use a small size network of 30 nodes.

We show first in figures 6.10, 6.11 and 6.12 the scatter plot of the reconstructed data versus the ‘measured’ ones for 60, 150 and 450 data points to reconstruct, respectively. The  $\chi^2$  value increases as the number of data points to reconstruct increases which is shown as a bigger scattering of the points as the number of data points increases.

We show in figures 6.13, 6.14 and 6.15 the in-coming degree distribution of the reconstructed network compared to the original one using a sample size of 60, 150 and 450 data points, respectively.

For a small size sample, the reconstructed in-coming degree distribution departs from the original one. As the size of the sample increases, the degree distribution gets closer to the original distribution. About 150 data points are sufficient to reconstruct the network as more data do not give any more information.

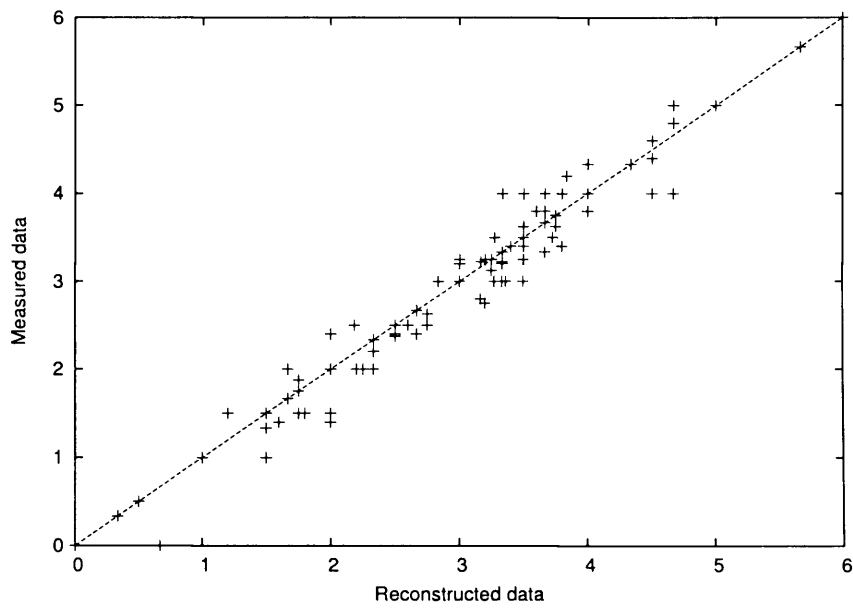


FIGURE 6.11. Scatter plot of the reconstructed data versus the original data for 150 data points.

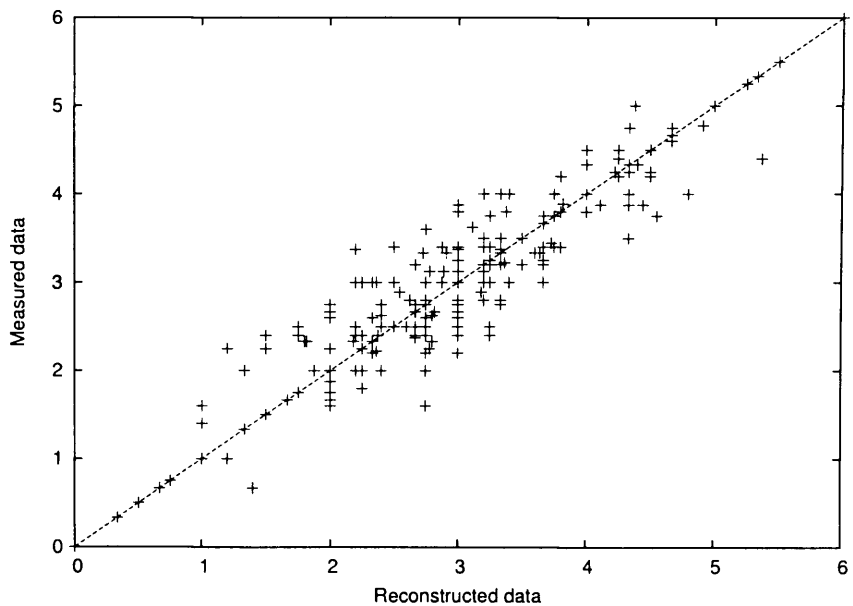


FIGURE 6.12. Scatter plot of the reconstructed data versus the original data for 450 data points.

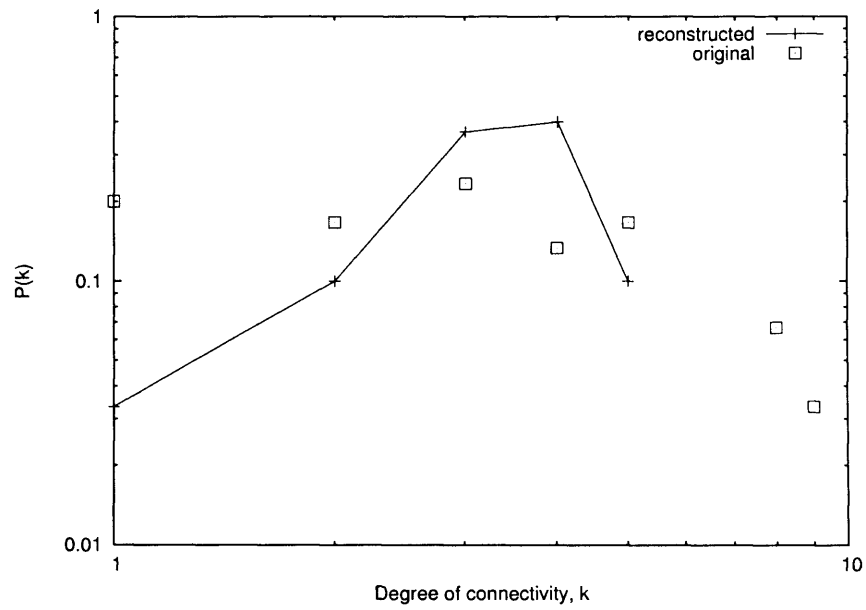


FIGURE 6.13. Comparison of the in-coming degree distribution between the original network and the reconstructed network for 30 data points.

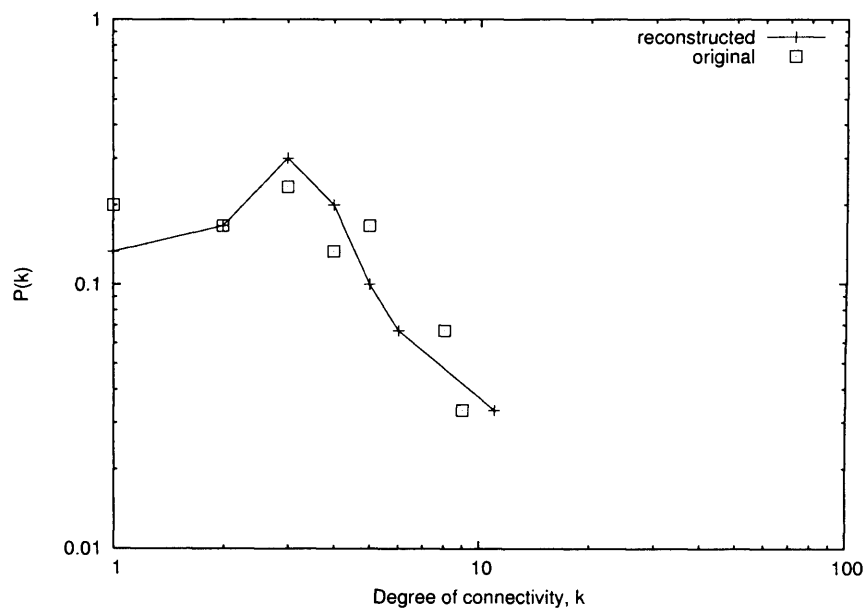


FIGURE 6.14. Comparison of the in-coming degree distribution between the original network and the reconstructed network for 150 data points.

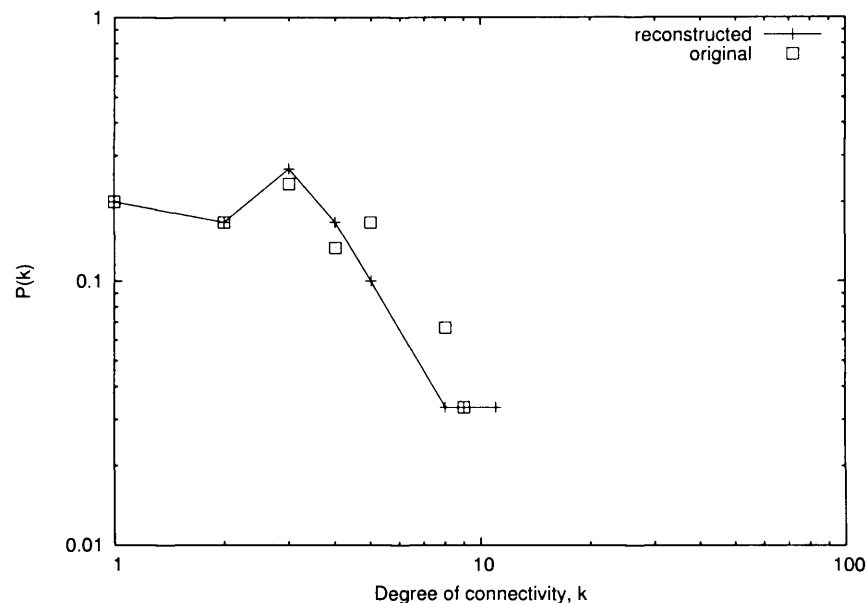


FIGURE 6.15. Comparison of the in-coming degree distribution between the original network and the reconstructed network for 450 data points.

## 6.5 Conclusion

We have shown in this chapter that the maximum entropy method applied to network reconstruction gives promising results and could be used to reconstruct networks from genetic regulatory data. It is an interesting method since it is based on the assumption that the best network to reconstruct the data is the most likely. In genetic network reconstruction, one needs also to assume the expression function at the nodes which may not be known beforehand and this is where arbitrariness comes in the model. However, we have seen in chapter 5 that a correlation could be found between elements of the architecture we are interested in and the level of abundance of the corresponding genes.

Reconstruction of networks is the first interest of the method when the function that relates the architecture to some structural aspects of the network is known. But this method could be also use to test specific functions provided the architecture of the system and its level of abundance is known.

In the method presented here, we have assumed our total ignorance of the detailed structure of the network to be reconstructed. In biology this is not always the case and some parts of genetic networks of certain systems are known. This information could be implemented as prior knowledge on the network so as to gain a more detailed recon-



struction.

# Chapter 7

## *Conclusion*

In this thesis, we have approached the study of networks from many aspects.

In chapter 2 we have developed a method to generate networks from any given degree distribution using a thermodynamical approach which allowed a statistical treatment of the system. This method has then been successfully applied to the construction of networks with various degree distributions.

In chapter 3 we have shown that information was contained in the non-randomness of the links between nodes and that, in a first approximation, the connections between nodes could be that of a random network.

In chapter 4, we have implemented a simple dynamics on networks so as to look at the effect of the architecture on the behaviour of the system. We have shown that our simple dynamics could yield networks of complex behaviour. We have found numerically that the distribution of periods of those networks is power-law regardless of their architecture. However, in the context of genetic regulation the architecture is important in determining the distribution of the level of expression of the nodes.

In chapter 5, we have compared the data obtained from our network model to mRNA data expression obtained from microarray experiments. We have shown that the dynamics implemented on the model was not sufficient to explain the time dependent complexity of the data. The mere alteration of the dynamics in the light of biological mechanisms has not been proven successful. We have shown, however, the existence of correlation between the out-degree of connectivity and the level of abundance in *E. coli*, which was then implemented in the model.

Finally in chapter 6, we have adapted the maximum entropy technique used to reconstruct imperfect images to the inference of networks. We have been able to reconstruct the degree distribution of model regulatory networks from data expression.

# Appendix A

## *Supplementary figures*

We give here the supplementary figures related to chapter 5.

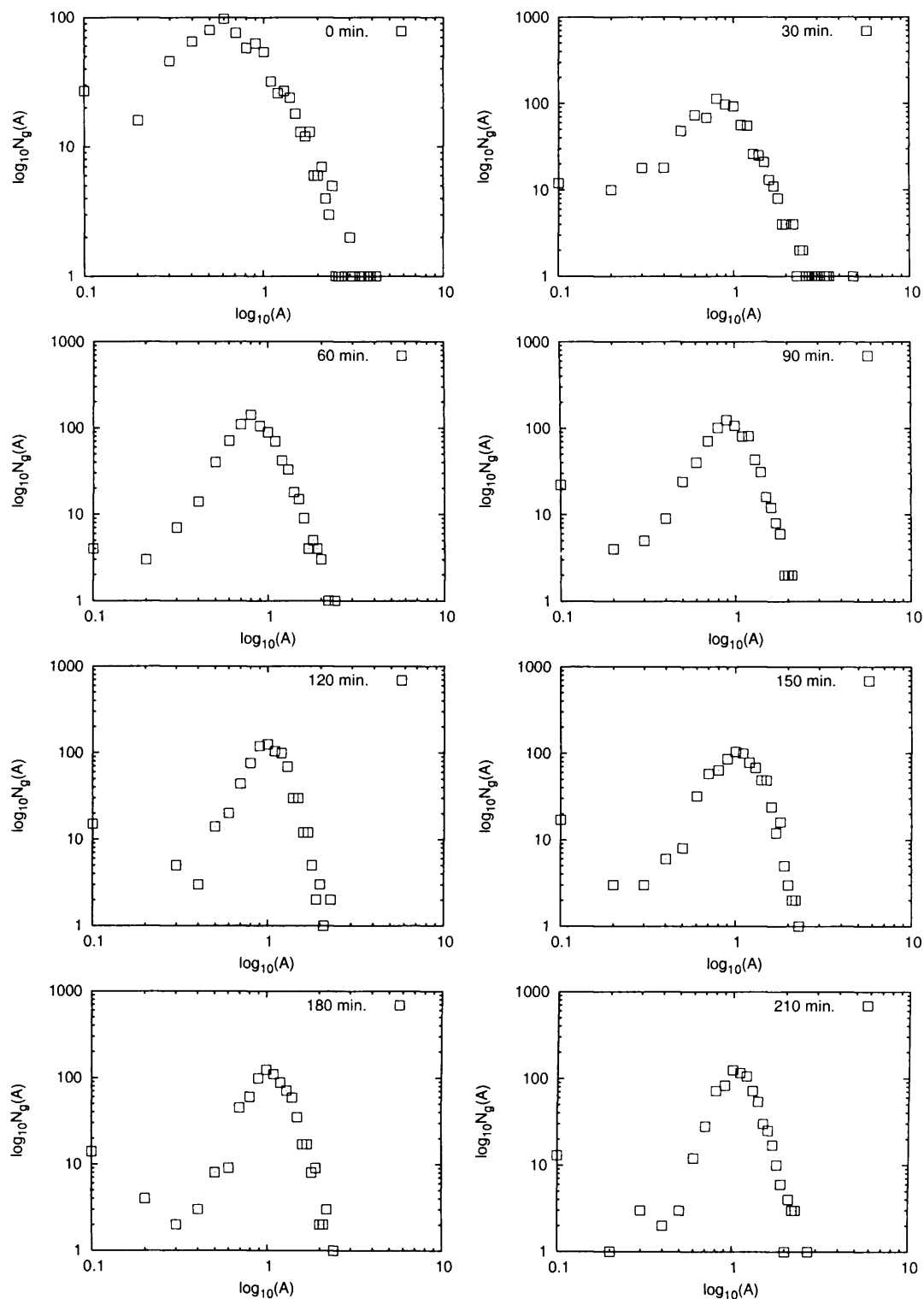


FIGURE A.1. Distribution of mRNA level of abundance in *S. cerevisiae* at different time points from  $t = 0$  to  $t = 210$  min. The cell population was synchronised by elutriation.

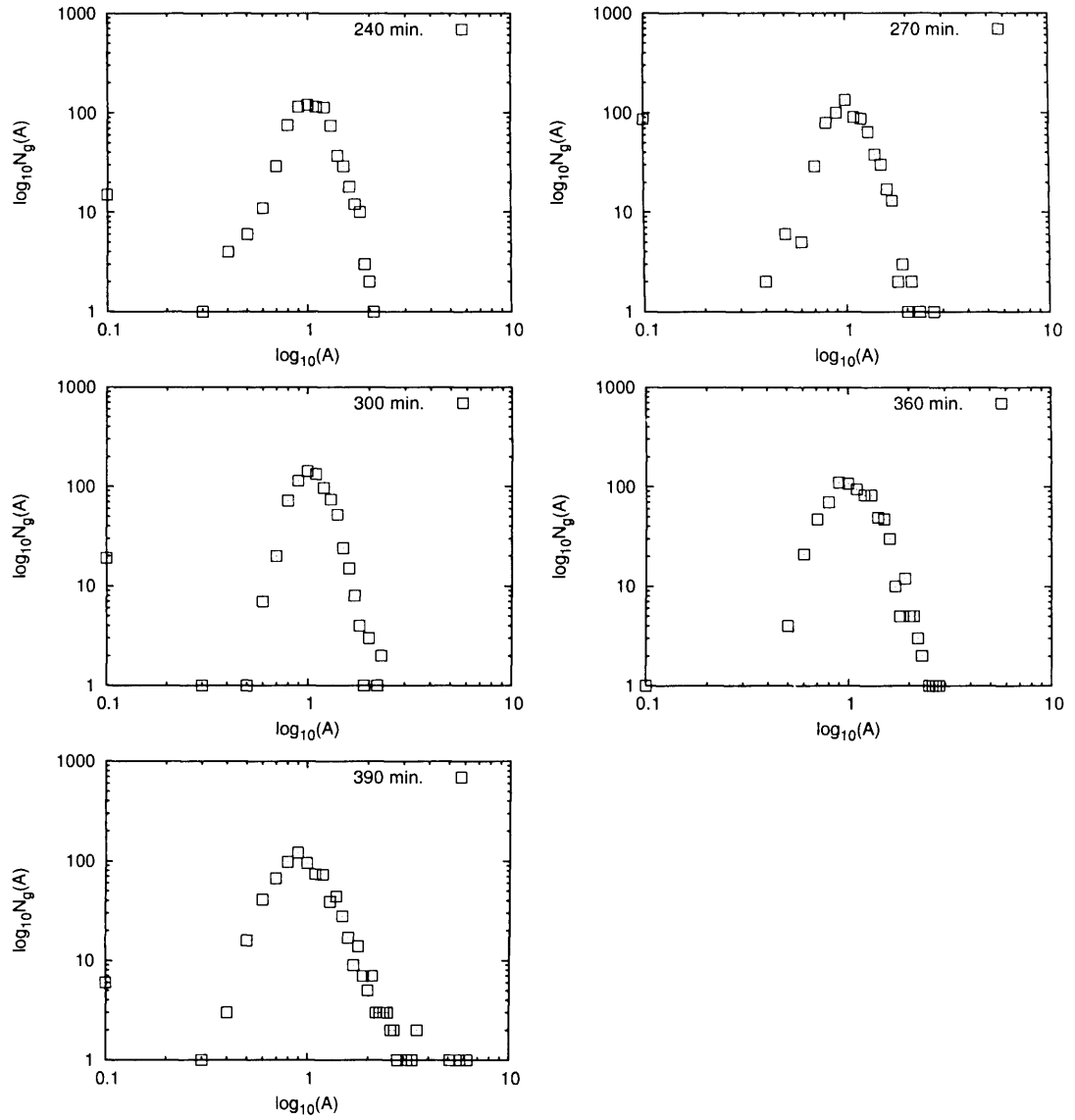


FIGURE A.2. Distribution of mRNA level of abundance in *S. cerevisiae* at different time points from  $t = 240$  to  $t = 390$ min. The cell population was synchronised by elutriation.

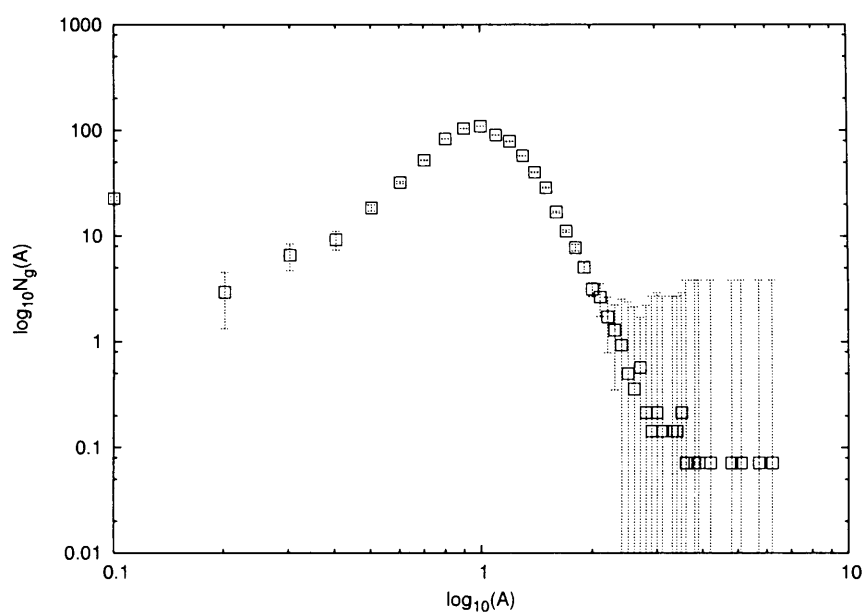


FIGURE A.3. Level of mRNA abundance distribution in *S. cerevisiae* averaged over the time series. The cell population was synchronised by elutriation. The errorbars give the standard deviation over the average.

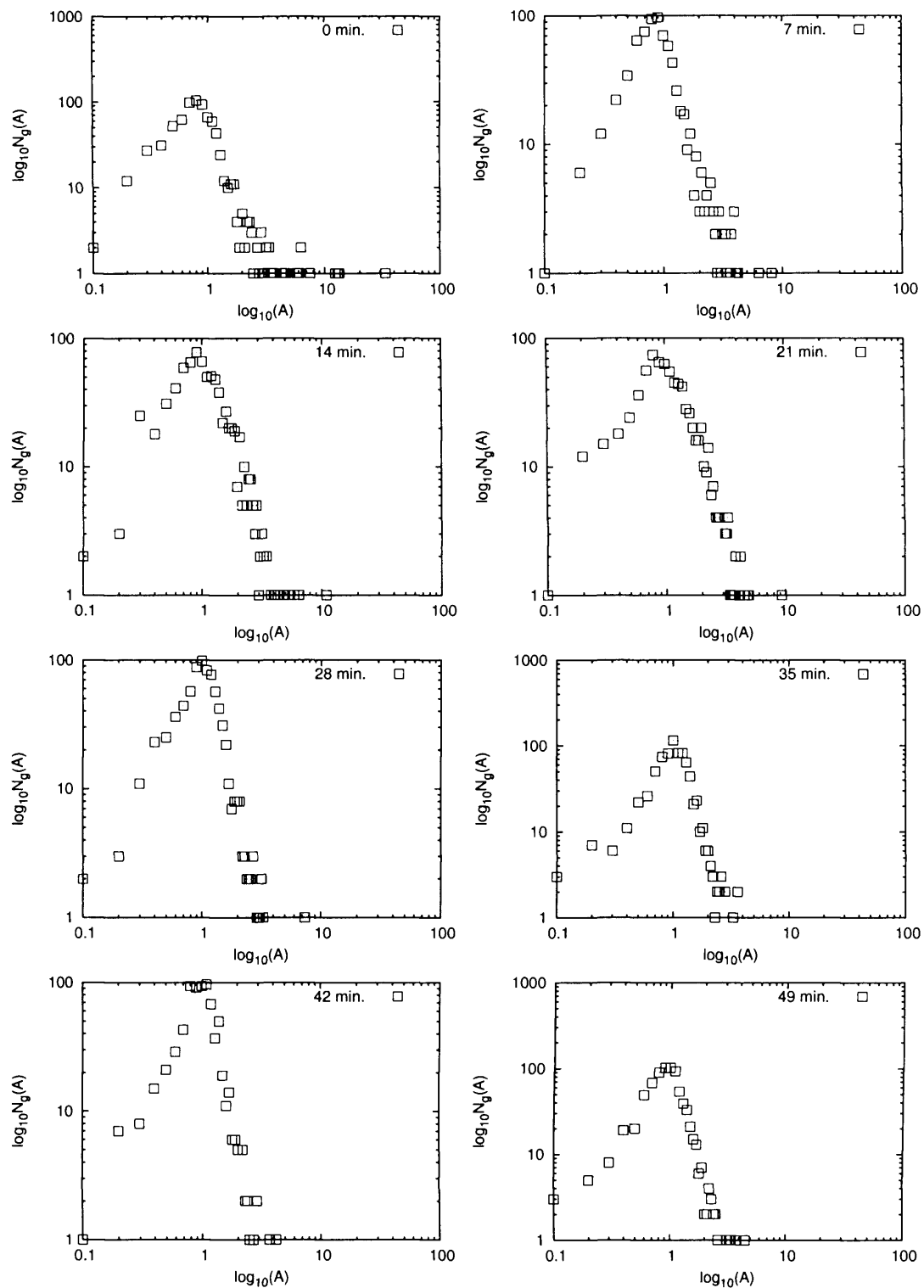


FIGURE A.4. Distribution of mRNA level of abundance in *S. cerevisiae* at different time points from  $t = 0$  to  $t = 49$  min. The cell population was synchronised using  $\alpha$ -factor.



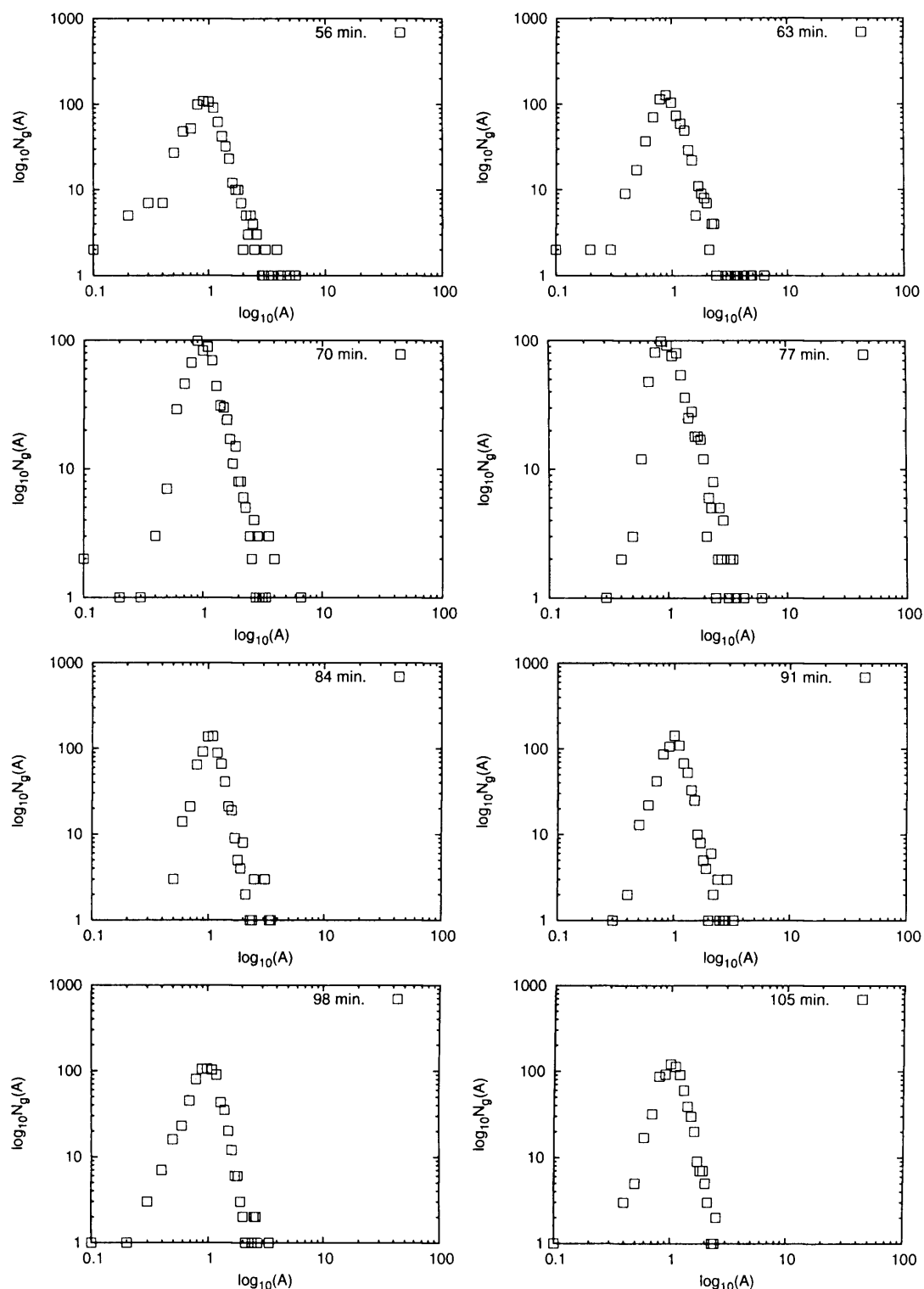


FIGURE A.5. Distribution of mRNA level of abundance in *S. cerevisiae* at different time points from  $t = 56$  to  $t = 105$  min. The cell population was synchronised using  $\alpha$ -factor.

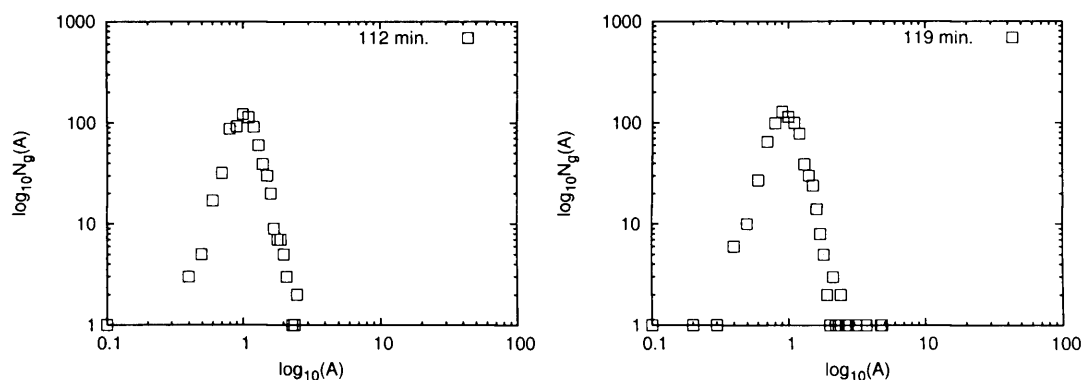


FIGURE A.6. Distribution of mRNA level of abundance in *S. cerevisiae* at different time points from  $t = 112$  to  $t = 119$ min. The cell population was synchronised using  $\alpha$ -factor.

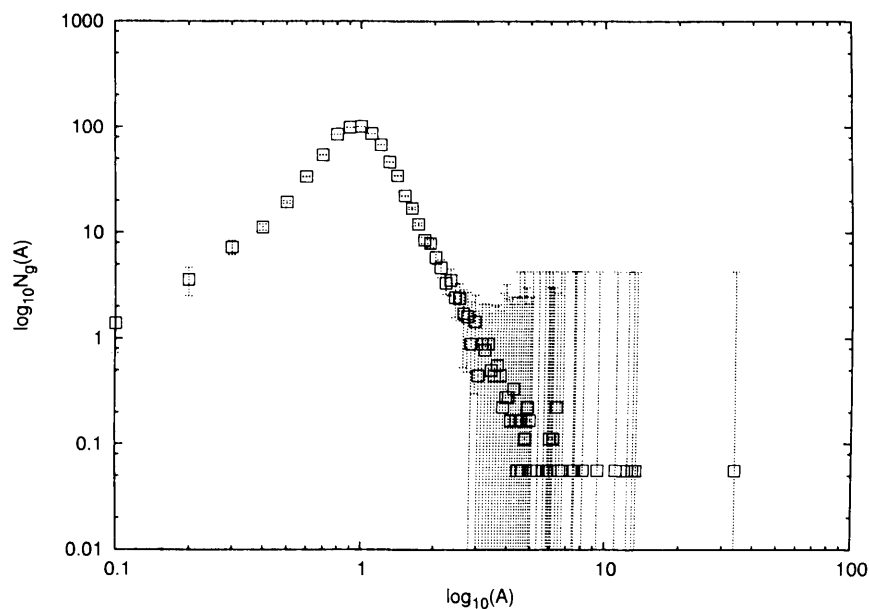


FIGURE A.7. Level of mRNA abundance distribution in *S. cerevisiae* averaged over the time series. The cell population was synchronised by  $\alpha$ -factor. The errorbars give the standard deviation over the average.

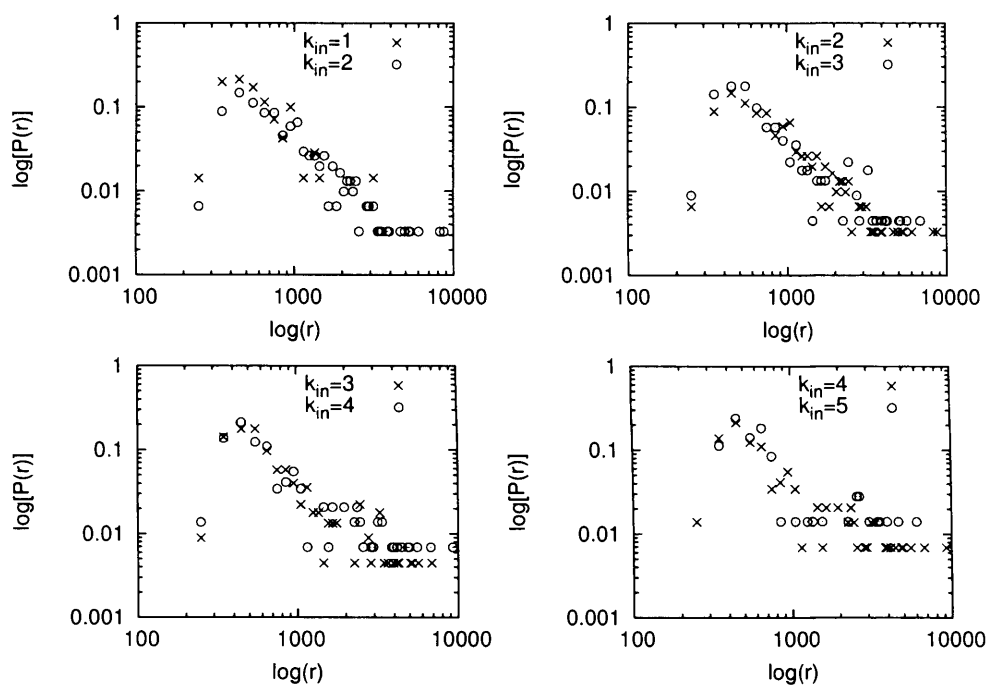


FIGURE A.8. One to one slope comparison between the distributions of mRNA abundance in *E. coli* for different in-coming degree of connectivity of the corresponding genes. The slopes are invariant

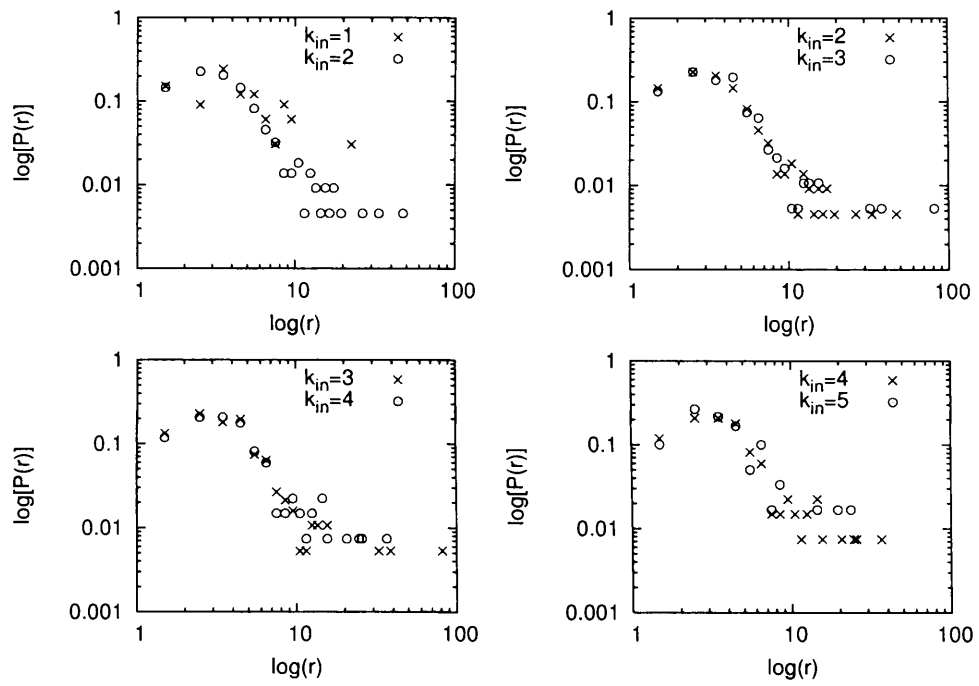


FIGURE A.9. One to one slope comparison between the distributions of product abundance in the simulated network for different in-coming degree of connectivity of the corresponding nodes. The slopes are invariant

# Bibliography

- Adami, C. (1999), *Introduction to artificial life*, Springer-Verlag.
- Adamic, L. A. (1999), The small world web, in 'ECDL '99: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries', Springer-Verlag, London, UK, pp. 443–452.
- Adamic, L. & Huberman, B. (2000), 'Power-law distribution of the world wide web', *Science* **287**, 2115–2116.
- Akutsu, T., Miyano, S. & Kuhara, S. (2000a), 'Algorithms for inferring qualitative models of biological networks.', *Pacific Symposium on Biocomputing* **5**, 293–304.
- Akutsu, T., Miyano, S. & Kuhara, S. (2000b), 'Inferring qualitative relations in genetic networks and metabolic pathways.', *Bioinformatics* **16**, 727–734.
- Albert, R. & Barabási, A. L. (2000), 'Topology of evolving networks: local events and universality', *Physical Review Letters* **85**(24), 5234–5237.
- Albert, R. & Barabási, A. L. (2002), 'Statistical mechanics of complex networks', *Review of Modern Physics* **74**, 47–97.
- Albert, R., Jeong, H. & Barabási, A. L. (2000), 'Error and attack tolerance of complex networks', *Nature* **406**(6794), 378–382.
- Allen, T., Herrgård, M., Liu, M., Qiu, Y., Glasner, J., Blattner, F. & Palsson, B. (2003), 'Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets', *Journal of Bacteriology* **185**(21), 6392–6399.
- Amaral, L. A., Scala, A., Barthélemy, M. & Stanley, H. E. (2000), 'Classes of small-world networks.', *Proceedings of the National Academy of Science USA* **97**(21), 11149–11152.
- Arita, M. (2004), 'The metabolic world of *Escherichia coli* is not small.', *Proceedings of the National Academy of Science USA* **101**(6), 1543–1547.
- Arita, M. (2005), 'Scale-freeness and biological networks', *The Journal of Biochemistry* **138**(1), 1–4.

- Barabási, A. L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.
- Barabási, A. L., Albert, R. & Jeong, H. (1999), 'Mean-field theory for scale-free random networks', *Physica A* **272**, 173–187.
- Batagelj, V. & Mrvar, A. (2003), '<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>'.
- Beineke, L. & Wilson, R. (1997), *Graph connections : relationships between graph theory and other areas of mathematics*, Oxford University Press.
- Berg, J. & Lassig, M. (2002), 'Correlated random networks', *Physical Review Letters* **89**(22), 228701.
- Bollobás, B. (1979), *Graph theory: an introductory course*, Springer-Verlag.
- Bollobás, B. (1985), *Random graphs*, Academic press, New York.
- Bollobás, B. & Riordan, O. (2004), 'The diameter of a scale-free random graph', *Combinatorica* **24**(1), 5–34.
- Börner, K., Maru, J. T. & Goldstone, R. L. (2004), 'The simultaneous evolution of author and paper networks.', *Proceedings of the National Academy of Science USA* **101**, 5266–5273.
- Buck, B. & Macaulay, V. (1991), *Maximum Entropy in Action*, Oxford: Clarendon Press.
- Callaway, D., Hopcroft, J., Kleinberg, J., Newman, M. & Strogatz, S. (2001), 'Are randomly grown graphs really random?', *Physical Review E* **64**(4), 041902.
- Callaway, D., Newman, M., Strogatz, S. & Watts, D. (2000), 'Network robustness and fragility: percolation on random graphs', *Physical Review Letters* **85**(25), 5468–5471.
- Chen, H.-C., Lee, H.-C., Lin, T.-Y., Li, W.-H. & Chen, B.-S. (2004), 'Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle', *Bioinformatics* **20**(12), 1914–1927.
- Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998), 'A genome-wide transcriptional analysis of the mitotic cell cycle.', *Molecular Cell* **2**, 65–73.
- Chung, F. & Lu, L. (2001), 'The diameter of sparse random graphs', *Advances in Applied Mathematics* **26**, 257–279.
- Chung, F., Lu, L., Dewey, T. G. & Galas, D. J. (2003), 'Duplication models for biological networks.', *Journal of Computational Biology* **10**(5), 677–687.
- Comellas, F., Ozon, J. & Peters, J. G. (2000), 'Deterministic small-world communication networks', *Information Processing Letters* **76**(1), 83–90.

- Comellas, F. & Sampels, M. (2002), 'Deterministic small-world networks', *Physica A* **309**, 231–235.
- Cooper, S. (2004a), 'Is whole-culture synchronization biology's 'perpetual-motion machine'?', *Trends in Biotechnology* **22**(6), 266–269.
- Cooper, S. (2004b), 'Rejoinder: whole-culture synchronization cannot, and does not, synchronize cells.', *Trends in Biotechnology* **22**(6), 274–276.
- Cooper, S. & Shedden, K. (2003), 'Microarray analysis of gene expression during the cell cycle.', *Cell & Chromosome* **2**, 1–12.
- Crutchfield, J. (1994), 'The calculi of emergence - computation, dynamics and induction', *Physica D* **75**, 11–54.
- D'haeseleer, P., Liang, S. & Somogyi, R. (2000), 'Genetic network inference: from co-expression clustering to reverse engineering.', *Bioinformatics* **16**, 707–726.
- Dorogovtsev, S. N. & Mendes, J. F. (2000), 'Evolution of networks with aging of sites', *Physical Review E* **62**(2), 1842–1845.
- Dorogovtsev, S. N., Mendes, J. F. & Samukhin, A. N. (2000), 'Structure of growing networks with preferential linking.', *Physical Review Letters* **85**(21), 4633–4636.
- Eisen, M. B. & Brown, P. O. (1999), 'DNA arrays for analysis of gene expression.', *Methods in Enzymology* **303**, 179–205.
- Evinger, M. & Agabian, N. (1977), 'Envelope-associated nucleoid from *Caulobacter crescentus* stalked and swarmer cells.', *Journal of Bacteriology* **132**(1), 294–301.
- Fell, D. A. & Wagner, A. (2000), 'The small world of metabolism', *Nature Biotechnology* **18**(11), 1121–1122.
- Fields, S. (2005), 'High-throughput two-hybrid analysis. the promise and the peril.', *FEBS Journal* **272**, 5391–5399.
- Friedman, N. (2004), 'Inferring cellular networks using probabilistic graphical models.', *Science* **303**, 799–805.
- Friedman, N. & Koller, D. (2003), 'Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks', *Machine learning* **14**, 95–125.
- Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), 'Using bayesian networks to analyze expression data.', *Journal of Computational Biology* **7**, 601–620.
- Fronczak, A., Fronczak, P. & Hołyst, J. A. (2003), 'Mean-field theory for clustering coefficients in Barabási-albert networks', *Physcal Review E* **68**, 046126+.
- Glasner, J., Liss, P., Plunkett, G., Darling, A., Prasad, T., Rusch, M., Byrnes, A., Gilson, M., Biehl, B., Blattner, F. & Perna, N. (2003), 'ASAP, a systematic annotation package for community analysis of genomes.', *Nucleic Acids Research* **31**, 147–51.

- Greenbaum, D., Jansen, R. & Gerstein, M. (2002), 'Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts', *Bioinformatics* **18**, 585–596.
- Grondin, Y. & Raine, D. J. (2005), Clustering and robustness in networks, in 'Proceedings of the First European Conference on Complex Systems'.
- Guelzim, N., Bottani, S., Bourguin, P. & Kepes, F. (2002), 'Topological and causal structure of the yeast transcriptional regulatory network', *Nature Genetics* **31**, 60–63.
- Gygi, S., Rochon, Y., Franza, B. & Aebersold, R. (1999), 'Correlation between protein and mRNA abundance in yeast', *Molecular and Cellular Biology* **19**, 1720–30.
- Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. & Vidal, M. (2004), 'Evidence for dynamically organized modularity in the yeast protein-protein interaction network', *Nature* **430**(6995), 88–93.
- Ideker, T., Thorsson, V. & Karp, R. (2000), 'Discovery of regulatory interactions through perturbation: inference and experimental', *Pacific Symposium on Biocomputing* pp. 305–316.
- Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q. & Yu, W. (2005), 'Multiple-laboratory comparison of microarray platforms', *Nature Methods* **2**(5), 345–350.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001), 'A comprehensive two-hybrid analysis to explore the yeast protein interactome.', *Proceedings of the National Academy of Science USA* **98**(8), 4569–4574.
- Järvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O. P. & Monni, O. (2004), 'Are data from different gene expression microarray platforms comparable?', *Genomics* **83**(6), 1164–1168.
- Jaynes, E. T. (1957), 'Information theory and statistical mechanics', *Physical Review* **106**, 620.
- Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001), 'Lethality and centrality in protein networks', *Nature* **411**, 41–42.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. & Barabasi, A. (2000), 'The large-scale organization of metabolic networks', *Nature* **407**, 651–4.
- Kasturirangan, R. (1999), Multiple scales in small-world networks, Technical report, Cambridge, MA, USA.



- Kauffman, S. (1993), *The origins of order: self-organization and selection in evolution*, Oxford University Press.
- Kaufmann, M., Lehmann, K. & Post, H. (2005), On small-world generating models, in 'Proceedings of the First European Conference on Complex Systems'.
- Keller, E. F. (2005), 'Revisiting "scale-free" networks.', *Bioessays* **27**(10), 1060–1068.
- Képès, F. (2004), 'Periodic transcriptional organization of the *E. coli* genome.', *Journal of Molecular Biology* **5**(340), 957–964.
- Klemm, K. & Eguíluz, V. M. (2002a), 'Growing scale-free networks with small-world behavior', *Physical Review E* **65**(5), 057102.
- Klemm, K. & Eguíluz, V. M. (2002b), 'Highly clustered scale-free networks', *Physical Review E* **65**(3), 036123.
- Kothapalli, R., Yoder, S. J., Mane, S. & Loughran, T. P. (2002), 'Microarray results: how accurate are they?', *BMC Bioinformatics* **3**, 1–10.
- Krapivsky, P. L. & Redner, S. (2001), 'Organization of growing random networks', *Physical Review E* **63**(6), 066123.
- Krapivsky, P. L., Redner, S. & Leyvraz, F. a. (2000), 'Connectivity of growing random networks', *Physical Review Letters* **85**(21), 4629–4632.
- Laub, M., McAdams, H., Feldblyum, T., Fraser, C. & Shapiro, L. (2000), 'Global analysis of the genetic network controlling a bacterial cell cycle', *Science* **290**, 2144–2148.
- Le, P., Bahl, A. & Ungar, L. (2004), 'Using prior knowledge to improve genetic network reconstruction from microarray data.', *In Silico Biology* **4**, 335–353.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. & Young, R. A. (2002), 'Transcriptional regulatory networks in *Saccharomyces cerevisiae*.', *Science* **298**(5594), 799–804.
- Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. (2004a), 'The yeast cell-cycle network is robustly designed', *Proceedings of the National Academy of Science USA* **101**, 4781–4786.
- Li, H., Lu, L., Manly, K., Chesler, E., Bao, L., Wang, J., Zhou, M., Williams, R. & Cui, Y. (2005), 'Inferring gene transcriptional modulatory relations: a genetical genomics approach.', *Human Molecular Genetics* **14**, 1119–1125.
- Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot,

- L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhoute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. & Vidal, M. (2004b), 'A map of the interactome network of the metazoan *C. elegans*.', *Science* **303**(5657), 540–543.
- Li, Z. & Chan, C. (2004), 'Inferring pathways and networks with a bayesian framework.', *FASEB Journal* **18**, 746–8.
- Liang, S., Fuhrman, S. & Somogyi, R. (1998), 'Reveal, a general reverse engineering algorithm for inference of genetic network', *Pacific Symposium on Biocomputing* **3**, 18–29.
- Ma, H. & Zeng, A. P. (2003), 'Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.', *Bioinformatics* **19**(2), 270–277.
- Mandelbrot, B. (1954), 'Structure formelle des textes et communication', *Word* **10**, 1–27.
- Marsh, L., Neiman, A. M. & Herskowitz, I. (1991), 'Signal transduction during pheromone response in yeast.', *Annual Review of Cell Biology* **7**, 699–728.
- Maslov, S. & Sneppen, K. (2002), 'Specificity and stability in topology of protein networks', *Science* **296**(5569), 910–913.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–1092.
- Milgram, S. (1967), 'The small-world problem', *Psychology Today* **2**, 60–67.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002), 'Network motifs: simple building blocks of complex networks.', *Science* **298**, 824–827.
- Nadon, R. & Shoemaker, J. (2002), 'Statistical issues with microarrays: processing and analysis.', *Trends in Genetics* **18**(5), 265–271.
- Newman, M. E. (2001), 'The structure of scientific collaboration networks.', *Proceedings of the National Academy of Science USA* **98**(2), 404–409.
- Newman, M. E. (2002), 'Assortative mixing in networks.', *Physical Review Letters* **89**(20), 8701–8704.
- Newman, M. E. & Watts, D. J. (1999a), 'Renormalization group analysis of the small-world network model', *Physics Letters A* **263**(4), 341–346.
- Newman, M. E. & Watts, D. J. (1999b), 'Scaling and percolation in the small-world network model.', *Physical Review E* **60**(6), 7332–7342.

- Nierman, W. C., Feldblyum, T. V., Laub, M. T., Paulsen, I. T., Nelson, K. E., Eisen, J. A., Heidelberg, J. F., Alley, M. R., Ohta, N., Maddock, J. R., Potocka, I., Nelson, W. C., Newton, A., Stephens, C., Phadke, N. D., Ely, B., DeBoy, R. T., Dodson, R. J., Durkin, A. S., Gwinn, M. L., Haft, D. H., Kolonay, J. F., Smit, J., Craven, M. B., Khouri, H., Shetty, J., Berry, K., Utterback, T., Tran, K., Wolf, A., Vamathevan, J., Ermolaeva, M., White, O., Salzberg, S. L., Venter, J. C., Shapiro, L., Fraser, C. M. & Eisen, J. (2001), 'Complete genome sequence of *Caulobacter crescentus*.' , *Proceedings of the National Academy of Science USA* **98**(7), 4136–4141.
- Pournara, I. & Wernisch, L. (2004), 'Reconstruction of gene networks using bayesian learning and manipulation experiments' , *Bioinformatics* **20**(17), 2934–2942.
- Quon, K., Marczyński, G. & Shapiro, L. (1996), 'Cell cycle control by an essential bacterial two-component signal transduction protein.' , *Cell* **84**, 83–93.
- Raine, D. J., Grondin, Y., Thellier, M. & Norris, V. (2003), 'Networks as constrained thermodynamic systems' , *Comptes Rendus Biologies* **326**, 65–74.
- Raine, D. J. & Norris, V. (2001), 'Network structure of metabolic pathways' , *Journal of Biological Physics and Chemistry* **1**, 89–94.
- Raine, D. & Norris, V. (2002), Network complexity, in 'Modeling and simulation of biological processes in the context of the genome. Conference Proceedings, Autrans', P. Amar, F. Kepes, V. Norris, P. Tracqui (Eds.), pp. 67–75.
- Ravasz, E. & Barabasi, A.-L. (2003), 'Hierarchical organization in complex networks' , *Physical Review E* **67**(2), 026112.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. (2002), 'Hierarchical organization of modularity in metabolic networks' , *Science* **297**(5586), 1551–1555.
- Redner, S. (1998), 'How popular is your paper? An empirical study of the citation distribution' , *European Physical Journal B* **4**, 131–134.
- Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P. & Vidal, M. (2005), 'Towards a proteome-scale map of the human protein-protein interaction network' , *Nature* **437**, 1173–1178.
- Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Daz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C. & Collado-Vides, J. (2004), 'RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12' , *Nucleic Acids Research* **32**, D303–6.

- Schrödinger, E. (1952), *Statistical thermodynamics : a course of seminar lectures delivered in January-March 1944 at the School of Theoretical Physics, Dublin Institute for Advanced Studies*, Cambridge U.P.
- Schwikowski, B., Uetz, P. & Fields, S. (2000), 'A network of protein-protein interactions in yeast.', *Nature Biotechnology* **18**(12), 1257–61.
- Shannon, C. & Weaver, W. (1949), *The mathematical theory of communication*, University of Illinois Press.
- Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002), 'Network motifs in the transcriptional regulation network of *Escherichia coli*.' , *Nature Genetics* **31**, 64–68.
- Skilling, J. & Bryan, R. K. (1984), 'Maximum entropy image reconstruction - general algorithm', *MNRAS* **211**, 111.
- Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. & Futcher, B. (1998), 'Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization', *Molecular Biology of the Cell* **9**, 3273–3297.
- Spellman, P. T. & Sherlock, G. (2004a), 'Final words: cell age and cell cycle are unlinked.', *Trends in Biotechnology* **22**(6), 277–278.
- Spellman, P. T. & Sherlock, G. (2004b), 'Reply: whole-culture synchronization - effective tools for cell cycle studies.', *Trends in Biotechnology* **22**(6), 270–273.
- Tan, P. K., Downey, T. J., Spitznagel, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. & Cam, M. C. (2003), 'Evaluation of gene expression measurements from commercial microarray platforms.', *Nucleic Acids Research* **31**(19), 5676–5684.
- Teichmann, S. A. & Babu, M. M. (2004), 'Gene regulatory network growth by duplication.', *Nature Genetics* **36**(5), 492–496.
- van Noort, V., Snel, B. & Huynen, M. (2004), 'The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model', *EMBO Reports* **5**(3), 280–284.
- Vohradský, J. & Ramsden, J. J. (2001), 'Genome resource utilization during prokaryotic development', *FASEB Journal* **15**, 2054–2056.
- Wagner, A. (2001), 'The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes.', *Molecular Biology and Evolution* **18**(7), 1283–1292.
- Wagner, A. (2003), 'How the global structure of protein interaction networks evolves.', *Proceedings of the Royal Society B* **270**(1514), 457–466.
- Wagner, A. & Fell, D. A. (2001), 'The small world inside large metabolic networks.', *Proceedings of the Royal Society B* **268**(1478), 1803–1810.

## Bibliography

---

- Watts, D. (1999), *Small worlds : the dynamics of networks between order and randomness*, Princeton University Press.
- Watts, D. J. & Strogatz, S. H. (1998), 'Collective dynamics of 'small-world' networks.', *Nature* **393**(6684), 440–442.
- Xulvi-Brunet, R. & Sokolov, I. (2005), 'Changing correlations in networks: Assortativity and dissortativity', *Acta Physica Polonica B* **36**(4), 1431–1455.
- Yuan, B., Chen, K. & Wang, B. (2004), 'Growing Directed Networks: Organization and Dynamics', *Arxiv preprint cond-mat/0408391* .
- Zhang, Z., Rong, L. & Guo, C. (2005), 'A deterministic small-world network created by edge iterations', *ArXiv Condensed Matter e-prints* .