FROM THE TOPOLOGY TO THE DYNAMICS OF COMPLEX NETWORKS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Soon-Hyung Yook, B. S., M. S.

A.-L. Barabási, Director

Graduate Program in Physics
Notre Dame, Indiana
August 2004
FROM THE TOPOLOGY TO THE DYNAMICS OF COMPLEX NETWORKS

Abstract

by

Soon-Hyung Yook

Understanding the mechanisms governing the behavior of complex networks is a prerequisite for characterizing complex systems. Frequently, networks are modelled as unweighted graphs in which each link has the same strength. However, for many real networks appearing in biological, technological and economic systems, each link has a specific weight, as nodes interact with each other with different strengths. In order to extend our understanding of network architecture to such systems, we introduce several weighted network models and investigate their scaling properties. In some real systems each node has a fixed geographical location, forcing some nodes to be connected by physical links of considerable length, such as routers connected by wires on the Internet. In such systems, we find that the physical layout of the underlying network strongly impacts the large-scale properties of the network. By combining data from several empirical databases and results from numerical simulations, we uncover the existence of three universal mechanisms which significantly affect the network’s global properties. As an application of complex network theory, we also study the large-scale properties of four yeast-protein interaction databases, finding quantitative evidence of strong correlations between the underlying network’s structure and pro-
tein classes and function. To move beyond the topological properties of networks, we investigate the dynamics of a diffusive system driven by multiplicative noise, uncovering the relationship between the fluctuations of the incoming fluxes and the underlying network topology.
## CONTENTS

**FIGURES** .......................................................... iv

**ACKNOWLEDGMENTS** .............................................. xi

**CHAPTER 1: INTRODUCTION** .................................... 1
  1.1 Networks in the Real World ................................. 2
      1.1.1 Social Networks ....................................... 2
      1.1.2 Information Networks ................................. 3
      1.1.3 Technological Networks ............................... 3
      1.1.4 Biological Networks .................................. 4
  1.2 Erdös-Rényi (ER) Model .................................... 4
      1.2.1 Degree Distribution .................................. 4
  1.3 Small-World Networks ...................................... 7
  1.4 Scale-Free (SF) Networks .................................. 7
      1.4.1 Barabási-Albert Model ............................... 8
      1.4.2 Clustering Coefficient and Modularity ............... 13
  1.5 Fluctuations of Random Walkers on Scale-Free Networks ... 15

**CHAPTER 2: WEIGHTED EVOLVING NETWORKS** .............. 17
  2.1 Introduction ................................................ 17
  2.2 Defining Weighted Network Models ......................... 19
      2.2.1 Weighted Scale-Free (WSF) Model .................. 19
      2.2.2 Weighted Exponential (WE) Model .................. 19
  2.3 Results of Numerical Simulations .......................... 20
  2.4 Analytic Derivation of $w_i(t)$ ........................... 22
      2.4.1 WE Model ............................................ 24
      2.4.2 WSF Model .......................................... 27
  2.5 Weight Driven Connectivity ................................ 27
  2.6 Discussion .................................................. 28

**CHAPTER 3: SPATIALLY CORRELATED NETWORK MODEL** ...... 30
  3.1 Introduction ................................................ 30
  3.2 Physical Layout ............................................. 32
  3.3 Distribution of Links ...................................... 34
### CHAPTER 4: FUNCTIONAL AND TOPOLOGICAL CHARACTERIZATION OF PROTEIN INTERACTION NETWORKS

4.1 Introduction .................................................. 45
4.2 Methods ....................................................... 47
   4.2.1 Protein Interactions Data Sources ...................... 47
4.3 Results ....................................................... 48
   4.3.1 Protein Interaction Databases ......................... 48
   4.3.2 Large-Scale Organization of the Protein Interaction Network 49
   4.3.3 Correlations Between Topology and Functional Organization 55
   4.3.4 Correlation Between Topology and Cellular Localization .... 63
   4.3.5 Relationship Between Functional and Localization Classes ... 66
4.4 Conclusion and Discussion .................................... 72
   4.4.1 Large-Scale Organization ............................. 77
   4.4.2 Function and Cellular Localization ..................... 79

### CHAPTER 5: ROLE OF THE CYTOSKELETON IN SIGNALLING NETWORKS

5.1 Introduction .................................................. 81
5.2 Method ....................................................... 83
   5.2.1 Quantitative Analysis .................................. 83
5.3 Results ....................................................... 84
   5.3.1 Definitions of Signalling and Cytoskeletal Proteins ....... 84
   5.3.2 Global Clustering ........................................ 85
   5.3.3 Local Clustering ......................................... 89
   5.3.4 The Special Role of the Cytoskeleton in Signalling Networks 92
   5.3.5 Discussion ............................................... 100

### CHAPTER 6: NETWORK DYNAMICS: FLUX FLUCTUATIONS WITH MULTIPLICATIVE NOISE

6.1 Introduction .................................................. 101
6.2 Model and Numerical Analysis ................................ 102
   6.2.1 Definition of Model .................................... 102
   6.2.2 Numerical Analysis ..................................... 104
6.3 Master Equation Approach ................................... 106
6.4 Summary and Discussion ..................................... 113

REFERENCES ...................................................... 115
FIGURES

1.1 Schematic diagram for Erdős-Rényi random graph with $N = 7$. Each pair of nodes (e.g. red) are connected with probability $p$.  

1.2 Example of the $P(k)$ distribution for the ER model. The number of nodes is $N = 10000$, and we used the connection probability $p = 0.001$. The solid line represents the Poisson distribution with $\langle k \rangle \simeq 10$, in good agreement with Eq. (1.2).  

1.3 Example of $P(k)$ distribution of real networks. (a) router level of the Internet: the slope of the data is $\gamma \approx 2.14$, (b) synonym network in linguistics with exponent $\gamma \approx 2.8$.  

1.4 Schematic diagram of the BA model with $m = 1$. The red circle denotes a new node. According to the Eq. (1.4), the most probable configuration of connection of the new link (red line) is the connection to the node which has $k = 4$.  

1.5 Example of $P(k)$ distribution of the BA model (a) with preferential attachment and (b) without preferential attachment. The slopes of the solid lines represent (a) $P(k) \sim k^{-3}$ and (b) $P(k) \sim exp(-k/3.5)$, respectively.  

2.1 The distribution $P(w_{ij})$ of the individual link weights, $w_{ij}$ for the (a) WE and the (b) WSF models, defined in the text ($m = 2$). The symbols correspond to different system sizes (or time), i.e. $N = 10^3(\bigcirc), 10^4(\square), 10^5(\diamond)$ and $10^6(\triangle)$. The insets show the same data on a log-linear plot, indicating that the tail decays faster than exponential.  

2.2 (a) Distribution $P(w)$ of the total connectivity $w$ assigned to individual nodes for the WE model. The symbols correspond to different values of $m$, i.e. $m = 2$ ($\bigcirc$), $3$ ($\square$), $4$ ($\diamond$) and $5$ ($\triangle$). The inset shows the connectivity distribution, $P(k)$, for the same parameters as in the main panel. (b) Time dependence of $k_i(t)$ ($\bigcirc$) and $w_i(t)$ ($\square$) for a randomly selected node $i$ for the WE model ($i = 5000$). (c) $P(k)$ ($\bigcirc$) and $P(w)$ ($\square$) distributions for the WSF model for $m = 5$. The inset shows the same data for $m = 2$. (d) $k_i(t)$ ($\bigcirc$), $w_i(t)$ ($\square$) vs. $t$ for the WSF model ($i = 10000$).
2.3 The difference \( w(t) - k(t) \) for the (a) WE and the (b) WSF models. The continuous lines in each case represent the analytic solution (11) and (13), respectively. We limited the simulations to nodes appearing at large \( t_0 \) to capture the asymptotic limit, that is predicted by our predictions (11) and (13). We find that for smaller \( t_0 \) the crossover time for the convergence to the analytic solution is numerically prohibited.

2.4 Plot of \( P(k) \) (square) and \( P(w) \) (circle) distributions obtained from numerical simulation of weight driven connectivity model. The slope of the solid line represents \( \gamma \approx 2.94 \).

3.1 Distribution of the Internet around the world. (a) World-wide router density map, obtained by using the NETGEO tool to identify the geographical location of 228,265 routers mapped out by the extensive router level mapping effort of Govindan and Tangmunarunkit. (b) Population density map based on the Columbia University’s Center for International Earth Science Information Network’s (CIESIN) population data. Both maps are shown with a box resolution \( 1^\circ \times 1^\circ \). The bar next to each map gives the range of values encoded by color code, indicating that the highest population density within this resolution is of the order \( 10^7 \) people/box, while the highest router is of the order of \( 10^4 \) routers/box. Note that while in economically developed nations there are visibly strong correlations between population and router density, in the rest of the world Internet access is sparse, limited to urban areas characterized by population density peak.

3.2 Characterizing the Internet’s physical layout and topology by using direct measurements. (a) The physical layout of the Internet was studied by using a box counting method [59, 144], applied to the map shown in Fig. 3.1. The log-log plot shows the number of boxes of size \( \ell \times \ell \) km\(^2\) with nonzero routers/AS/inhabitants in function of \( \ell \) for North America. The slope of the straight line indicates that \( D_f \approx 1.5 \pm 0.1 \) for each dataset. (b) The dependence of the router/AS density on the population density in North America, showing the average number of router/AS nodes in a \( 1^\circ \times 1^\circ \) box in function of the number of people living in the same area. Similar plots were obtained for each continent, the steepness of the curves strongly correlated to economic factors. (c) The cumulative length distribution of the links connecting routers. \( d \) is the Euclidean distance between routers and \( R = 6,378 \) km is the radius of the Earth. (d) The cumulative change \( \Delta k \) in the connectivity of AS nodes with \( k \) links. The dotted line has slope 2, indicate the linear preferential attachment of links.
3.3 Phase diagram summarizing various internet models and their expected large-scale topology. The axes represent the scaling exponents \(1/(\sigma + 1), 2 - D_f\) and \(\alpha\), governing link placement, node location, and preferential attachment, respectively. Our measurement indicates (Fig. 3.2) that within this phase space, the Internet can be found at \(1/(\sigma + 1) = 1/2, 2 - D_f = 1/2, \) and \(\alpha = 1\), identified as a red circle. The yellow boxes indicate the location of all current Internet topology generators.

WAXMAN \[152\]: Nodes are placed randomly in space \(\left( D_f = 2 \right)\) with exponential distance dependence \(\left( 1/(\sigma + 1) = 0 \right)\) and without preferential attachment \((\alpha = 0)\). TIERS \[81\]: Based on a three-level hierarchy the model has no space dependence \(\left( D_f = 1 \right)\) and no preferential attachment \((\alpha = 0)\). GT-ITM \[27\]: While based on several different models, the most used pure random transit-stub version occupies the same position in the phase space as TIERS \(\left( D_f = 2, 1/(\sigma + 1) = 1, \alpha = 0 \right)\). INET2.0 \[73\] connects randomly placed nodes by using an externally imposed power-law connectivity information. While it does not include preferential attachment explicitly, as \(P(k)\) is forced to follow a power law, we put this generator at \(D_f = 2, 1/(\sigma + 1) = 1, \alpha = 1\). Note, however, that there are significant known differences \[77\] between a static graph, such as generated by INET, and scale-free topologies generated by evolving networks, such as the Internet. BRITE \[94\]: The most advanced of all, BRITE incorporates preferential attachment \((\alpha = 1)\) combined with the Waxman rule \((1/(\sigma + 1) = 1)\) for placing the links. As BRITE has the option to produce topologies with different parameters, we denote by BRITE-1 the version with only preferential attachment \(\left( D_f = 2, 1/(\sigma + 1) = 1, \alpha = 1 \right)\), and BRITE-2 the version including the Waxman rule as well \(\left( D_f = 2, 1/(\sigma + 1) = 0, \alpha = 1 \right)\). Note that BRITE has as option to include inhomogeneous node placement, creating regions with high-node density mimicking highly populated areas. The algorithm, however, does not create a fractal, thus we choose \(D_f = 2\) for both BRITE-1 and BRITE-2. The scale-free model \[13\], which ignores the physical location of the nodes \((\sigma = 0\) thus \(D_f\) can be arbitrary) is shown as a separate blue line on the \(1/(\sigma + 1) = 1\) axis and \(\alpha = 1\). The green areas correspond to an exponential \(P(k)\) distribution, while yellow areas are characterized by gelation, indicating that the Internet strikes a delicate balance at the boundary of these two topologically distinct phases.

3.4 The dependence of the degree distribution (Left) and link length distribution (Right) on the scaling exponent \(\sigma, \alpha\) and \(D_f\). (a) and (b): The effect of changing \(\sigma\) with \(\alpha = 1\) and \(D_f = 1.5\) unaltered, on \(P(k)\)(a) and \(P(d)\)(b). Note that the exponential Waxman's rule corresponds to \(\sigma = \infty\), in which case \(P(k)\) develops an exponential tail. (c) and (d): The effect of changing \(\alpha\) with fixed \(\sigma = 1\) and \(D_f = 1.5\). \(\alpha > 1\) correspond to the gelation phase. (e) and (f): The effect of changing \(D_f\) while fixing \(\alpha = \sigma = 1\).
4.1 Relationship between the four studied databases (a) at the protein and (b) at the interaction level. The sum of the numbers within each color boundary denotes (a) the total number of proteins or (b) the total number of interactions found in the corresponding database.

4.2 Large scale characteristics of the protein interaction databases. (a) Degree distribution of the four databases, shown on a log-log plot. Note that all datasets have a power-law tail, indicating that the underlying network has a scale-free topology. The solid line is obtained from the fitting to the function $P(k) \sim k^{-\gamma}$ to the DIP data, the best fit indicating $\gamma \approx 2.5$ for DIP dataset. (b) Plot of the clustering coefficient against degree on a log-log plot. The straight line has slope -2. (c) Cluster size distribution for the four databases shown on a log-log plot. Apart from the points corresponding to the giant component (for right) the $P(n)$ curves follow a power-law. The solid line is obtained from the least square fitting to $P(n) \sim n^{-\xi}$ for the MIPS dataset, providing $\xi = 3 \sim 4$.

4.3 Functional segregation. (a), (b) Schematic illustration of the topological interpretation of the segregation parameter $m(1)$. For example, for protein 1, focusing only on interaction to proteins belonging to the same functional class (shown in light color), we have $k = 4$, $C_1 = 1/3$, $m(1) = 3/4$. Focusing on proteins that are $d = 2$ distance from protein 1, we have $m(2) = 0$. In (b) we reorganized randomly the same number of proteins and kinks. In this random configuration we find $C_1 = 0$, $m(1) = 2/4 = 1/2$, and $m(2) = 1/2$. (c)-(f) The relative segregation functions for the four databases. Each curve corresponds to a different functional class, the numbers and the corresponding functional classes being listed in Table 4.2. The four panels describe the different dataset, i.e., (c) Uetz, (d) Ito complete, (e) MIPS, and (f) DIP.

4.4 Characterization of the segregation properties of proteins classified based on their functional class (a)-(c) or subcellular localization (d)-(f). In (a)-(c) on the horizontal axis we show the number corresponding to their various functional classes described in Table 2. (a) $m(1)/m_{\text{rand}}(1)$ ratio for each functional class, shown separately for each of the five studied databases. (b) Relative clustering coefficient $C/C_{\text{rand}}$ for each functional class, shown separately for all five databases. (c) Relative number of links between proteins belonging to the same functional class, $L/L_{\text{rand}}$, for each of 14 functional classes, shown separately for the five databases. (d)-(f) The same quantities as in (a)-(c) but characterizing proteins sharing the same subcellular localization. The horizontal axis, therefore, denotes the cellular localization classes listed in Table 4.3. The color code is the same as in (a), the plots showing separately the data for each localization class (horizontal axis) and each database.
4.5 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on Uetz. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes. . . . . . . . . . . . 68

4.6 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on Ito. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes. . . . . . . . . . . . 69

4.7 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on DIP. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes. . . . . . . . . . . . 70

4.8 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on MIPS. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes. . . . . . . . . . . . 71

4.9 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on Uetz. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes. . . . . . . . . . . . 73

4.10 \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) matrix, representing the relationship between the different functional classes based on Ito. The higher value of \((L^{\lambda \phi} / L^\lambda)/(I_{\text{rand}}^{\lambda \phi} / I_{\text{rand}}^\lambda)\) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes. . . . . . . . . . . . 74
4.11 \( \left( L^\lambda / L^\phi \right) / \left( L^\phi_{\text{rand}} / L^\lambda_{\text{rand}} \right) \) matrix, representing the relationship between the different functional classes based on DIP. The higher value of \( \left( L^\lambda / L^\phi / L^\phi_{\text{rand}} / L^\lambda_{\text{rand}} \right) \) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.

4.12 \( \left( L^\lambda / L^\phi \right) / \left( L^\phi_{\text{rand}} / L^\lambda_{\text{rand}} \right) \) matrix, representing the relationship between the different functional classes based on MIPS. The higher value of \( \left( L^\lambda / L^\phi / L^\phi_{\text{rand}} / L^\lambda_{\text{rand}} \right) \) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.

5.1 The largest interconnected cluster of 1,458 interacting proteins in the Uetz et al. data. Yellow and green dots denote cytoskeletal and signaling proteins respectively. The definition of the cytoskeletal and signaling follows the criteria in the text. Proteins in red are shared by two subclasses. The analogous cluster in the D network contains 4,198 proteins. It is not shown here because the density of proteins are too high to examine visually.

5.2 Distance distribution \( P(d) \) for proteins in the various classes. The notation in the figure are as follows: \( c - c \): cytoskeletal protein and cytoskeletal protein, \( s - s \): signalling protein and signaling protein, \( c - s \): signalling protein and cytoskeletal protein. \( P(d) \) for all proteins includes distances between any two proteins in the largest cluster. Numbers in the inset are the average distances \( \langle d \rangle \) between indicated proteins. (a) Results based on the Uetz et al. protein-protein interaction data. (b) Results based on the DIP protein-protein interaction database. The largest number of proteins in the DIP dataset is manifested in a narrower distance distribution and higher peak values.

5.3 Plot of rescaled average local segregation function \( \langle m_d \rangle / \langle m_{\text{rand}} \rangle \) using Uetz et al. interaction data (left) and the DIP data (right).

5.4 The nearest neighbor segregation function \( \langle m_1(\phi / \lambda) \rangle \). Arrows point from \( \phi \) to \( \lambda \). The left panels in (a) (Uetz data) and (b) (DIP data) summarize results from Fig. 5.3, whereas the right panels results for pseudo protein classes, constructed by selecting randomly 74 (U) or 92 (D) (for pseudo \( c \) class) and 141 (U) or 207 (D) (for pseudo \( s \) class) proteins from the largest cluster.
5.5 $\langle m_d(s/\lambda) \rangle / m_{\text{rand}} (\lambda = 0, 1, \cdots, 14)$ for fifteen functional classes in addition to the $c$ class. Upper and lower panels are based on Uetz et al. data and the DIP database respectively. The 15 functional protein classes are the following: 0-metabolism, 1-energy, 2-cell growth, cell division and DNA synthesis, 3-transcription, 4-protein synthesis, 5-proteins destination, 6-transport facilitation, 7-cellular transport mechanisms, 8-cellular biogenesis, 9-cell rescue, defense, cell death and ageing, 10-ionic homeostasis, 11-cellular organization, 12-transposable elements, viral and plasmid proteins, 13-classification not yet clear cut, 14-unclassified proteins.

5.6 $\langle m_d(c/\lambda) \rangle / m_{\text{rand}} (\lambda = 0, 1, \cdots, 14)$ for fifteen functional classes in addition to the $c$ class. Upper and lower panels are based on Uetz et al. data and the DIP database, respectively. The fifteen functional classes are the same as in Fig. 5.5.

5.7 The combined $c - s$ subnetwork derived from the largest connected cluster in Fig. 5.1. Yellow and green dots denote signalling and cytoskeletal proteins, respectively, proteins in red are shared by the two subclasses. Only proteins with at least one connection are shown.

5.8 Collapse of the signalling subnetwork shown in Fig. 5.7, on knock-out of $s$ proteins connecting to the cytoskeletal proteins.

6.1 The plot of $\sigma$ against $\langle H \rangle$ on SF networks. Squares ($\square$) represent $(p + q)/(p + q + 1) = 0.0$ (maximum diffusion rate) and circles ($\bigcirc$) represent $(p + q)/(p + q + 1) = 0.8$ (low diffusion rate). The slope of the solid line is $\alpha = 1/2$, and the slopes of the dashed lines are $\alpha = 1.0$. Both low and high diffusion rates cause the same type of crossover.

6.2 Plots for the diffusion-dominant region on SF networks((a)-(c)) and on ER networks((d)-(f)). (a) and (d) show the measurement of fluctuations of incoming fluxes on a SF network and a random network respectively. Solid lines denote $\sigma \sim \langle H \rangle$ and dotted lines represent $\sigma \sim \langle H \rangle^{1/2}$. (b) and (e): average connectivity $\langle k \rangle$ as a function of average incoming flux $\langle H \rangle$. Note that in the ER graph, the largest value of $\langle k \rangle$ is around 10, corresponding to the $\alpha = 1/2$ region in SF networks. (c) and (f): average incoming flux distribution. Solid squares ($\blacksquare$) represent low connectivity region and open circles ($\bigcirc$) denote high connectivity region.

6.3 Plot of incoming flux distribution $P(H)$. The slope of the solid line represents $-1 - \mu \simeq -3.3$.

6.4 Plot of $\langle H_{nn} \rangle$ against $H$. The inset shows $\langle H_{nn} \rangle$ for $k \leq 10^8$. The solid line represents the hyperbolic decaying function $\langle H_{nn} \rangle = a + b/(c + H)$. 110
ACKNOWLEDGMENTS

Dr. Albert-László Barabási has been a great advisor. I have deeply appreciated his consistent demand for quality, his respect for my judgement, and the real concern he has shown for my development as a physicist. I also greatly appreciate all of my collaborators, especially, Dr. Hawoong Jeong, Dr. Zoltan N. Oltvai, and Dr. Gabor Forgacs and Dr. Marcio de Menezes. Of my friends I would like to thank them all, but most especially Dr. Taek-Hyun Kim, Dr. Seong-Kyun Cheong, and all of my office mates, Dr. Choong-seop Lee, Erzsebet Ravasz, and Zoltan Dezso. Of those who aided specifically in the writing of this thesis, I thank Dr. Stefan Wuchty and Dr. Eivind Almaas for their critical readings of my thesis. Lastly, I am very thankful to my parents for their constant encouragement.
Traditionally, complex web-like structures have been studied in graph theory. Traditional graph theory initially focused on regular structures, such as the binary tree. In 1960 Erdős and Rényi introduced the random graph model, generated by random connections between nodes [46]. This model had guided our thinking about complex networks for decades. However, the growing interest in complex systems has prompted scientists to reconsider this modeling paradigm and ask a simple question: are real networks indeed random and homogeneous?

In the past few years, owing to the explosive growth in computational power, network data could be digitized and stored in huge databases, making them publicly accessible. This enabled us to investigate the structural properties of various complex systems more accurately. Intensive studies of these databases have indicated that many systems do not share the structural properties of Erdős and Rényi’s random network model. These discoveries have fuelled the development of new modeling concepts, aimed to understand the basic mechanisms and structural and dynamical properties which deviate from Erdős and Rényi’s random graph model. In this chapter, we will briefly discuss the historically important random complex network models and concepts aiming to classify the topological properties of complex structures.
1.1 Networks in the Real World

As mentioned above, recent network studies have been initiated by observations focusing on the properties of real networks, fueling attempts to understand the mechanisms responsible for their formation and to capture their topological properties. Hence, looking over some examples of real networks could help us understand the characteristics of real networks. In this section, we will classify real networks into four loose categories, following Newman [106].

1.1.1 Social Networks

In general a social network is defined by the interactions taking place between people or groups of people [130, 148]. Friendships between individuals [100, 121], business relationships between companies [52, 53, 89, 98], intermarriages between families [112], and sexual partnerships [82] are examples of social networks which have been studied so far. In 1967, social psychologist Milgram performed a notable experiment on acquaintance networks between participants [97], by asking a few hundred people in Omaha to forward a letter to a “target” stranger in Boston through personal contacts. The experiment ended up with 60 completed chains of letters through on average path length of six intermediates. This experiment demonstrated the ‘small-world’ phenomena for the first time, indicating that networks are characterized by small degrees of separation (see Section 1.3). More recent studies in this category have been extended to collaboration networks such as movie actors, which is available in the on-line database (www.imdb.com) [1, 8, 108, 150], communication networks such as telephone calls [2], and email networks [107].
1.1.2 Information Networks

A well studied example in this category is the citation network between academic publications [45, 106, 125]. Another famous example of an information network is the World Wide Web (WWW), in which each web page contains a piece of information about links to other pages. The WWW has been very intensively studied especially with “.com surge” in late 1990s [4, 15, 75, 25]. The WWW appears to have scale-free behavior which will be discussed in Section 1.4. Another example in this category are peer-to-peer networks [1], which are virtual networks of computers that share files between users such as Napster and Gnutella. Semantic word networks also can belong to in this category [40].

1.1.3 Technological Networks

Technological networks are designed typically for distribution of some commodity or resource such as electricity. Compared to the previous two classes, this class of network is hardware-oriented. The electrical power grid is one example in this class. Several authors have studied the statistical properties of power grids [8, 150, 151]. Other examples in this category are the networks of airline routes [8], and railways [80]. River networks [36, 90, 126] can be classified as technological networks. The telephone network and delivery networks, such as those used by the post office or delivery companies also fall into this category. Another very widely studied technological network is the Internet, i.e. the network of physical connections between computers/routers/autonomous systems [47]. We will discuss about the Internet more extensively in Chapter 3.
1.1.4 Biological Networks

Networks appear a broad range of biological systems. One of the well known examples is the network of metabolic pathways, which captures the metabolic reactions between the biomolecules inside a cell. Typical studies of this network can be found in Refs. [71, 72, 147]. The network defined by the physical interactions between proteins also belongs to this category. We will discuss the protein-protein interaction network in Chapters 4 and 5. Another much studied example of a biological network is the food web, which defines the food chain in an ecosystem [119]. Other examples are found in the context of neural networks [101] and blood vessels [12, 153, 154].

1.2 Erdös-Rényi (ER) Model

As mentioned above, the first classic random graph model was introduced by Erdös and Rényi in 1960 [46]. The model is defined as follows: starting with \( N \) vertices, connect each pair of nodes with probability \( p \) (see Fig. 1.2). Erdös and Rényi showed that the model has a critical probability \( p_c(N) \), which is a function of the number of nodes \( N \). Thus, depending on \( p \), the model shows different characteristics, described by some property \( Q \) (for example, each node is connected to one another). Below \( p_c \), the probability that the graph has property \( Q \), \( P(Q) \), is 0 in the limit \( N \to \infty \); otherwise \( P(Q) \) is 1. This feature resembles percolation theory [135] which has been studied in the context of critical phenomena [84].

1.2.1 Degree Distribution

Several important quantities are used to characterize the topological properties of networks. Degree distribution, clustering coefficient, and average distance are ex-
amples of such quantities. Here the degree (or connectivity) of a node $i$ represents the number of links (or edges) that the node has. Erdős and Rényi studied the distribution of the maximum and minimum degree in a random graph [46], and the degree distribution was derived by Bollobás [22]. For an Erdős-Rényi random graph with connection probability $p$, the probability that a node $i$ has a degree $k$, follows a binomial distribution

$$P(k_i = k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{1.1}$$

In Eq. (1.1) $p^k$ represents the probability that node $i$ is connected to $k$ nodes among $N-1$ nodes, $(1-p)^{N-1-k}$ represents the probability that the other $(N-1-k)$ nodes are disconnected, and $\binom{N-1}{k}$ is the number of possible combinations of connecting $k$ nodes to node $i$. In the large $N$ limit, Eq. (1.1) can be approximated with the Poisson

![Figure 1.1. Schematic diagram for Erdős-Rényi random graph with $N = 7$. Each pair of nodes (e.g. red) are connected with probability $p$.](image)
distribution

\[ P(k) \approx e^{-pN} \frac{pN^k}{k!} = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \]  

(1.2)

Fig. 1.2 shows a typical example of degree distribution of an ER model with \( N = 10000 \) nodes and \( p = 0.001 \). The solid line represents the Poisson distribution with \( \langle k \rangle \approx 9 \sim 10 \) which is consistent with the expectation of Eq. (1.2).
1.3 Small-World Networks

Loosely speaking, the term “small-world”, introduced in Section 1.1.1, refers to a network characterized by a small average distance (or diameter) and large clustering coefficient (see Section 1.4.2). The distance between nodes \( i \) and \( j \) is defined by the number of links along the shortest path between \( i \) and \( j \). In 1998 Watts and Strogatz [151] proposed a model that interpolates between a regular structure in a finite dimension and a random graph. The definition of the model is as follows:

(i) Start from an one dimensional ring lattice with \( N \) nodes satisfying the periodic boundary condition and connect each node to its \( k \) neighbors.

(ii) Rewire each edge of the lattice with probability \( p \). During the rewiring the self-connection and duplication of edges are not allowed.

By changing \( p \), the system undergoes a crossover between ordered (\( p = 0 \)) and disordered phases (\( p = 1 \)). By measuring the clustering coefficient (see Section 1.4.2 for exact definition) and the average path length as a function of \( p \), Watts and Strogatz showed that for small \( p \), there is a region in which the average path length drops rapidly but the clustering coefficient stays almost constant at its value for the regular lattice, resulting in the large clustering and small path length.

1.4 Scale-Free (SF) Networks

Recent empirical studies on many real networks indicate that the structural properties of real networks are different from that of the well known classical random network. One of the most apparent differences between real networks and random networks can be found in the degree distribution. For example, the degree distribution of the WWW can be characterized by the power-law

\[
P(k) \sim k^{-\gamma},
\]  

(1.3)
with $\gamma \simeq 2.1$ for incoming degrees and $\gamma \simeq 2.45$ for outgoing degrees [4, 13]. The degree distribution of the Internet also satisfies the relation (1.3) as shown in Fig. 1.3 (a) [47]. Many other examples of this type of networks can be found in various fields including collaboration of movie actors [1, 8, 108, 150], scientific collaboration [103, 104, 105], human sexual contacts [82], cellular networks [71], ecological networks [28, 29], linguistic networks based on synonyms obtained from *Merriam-Webster's Collegiate Dictionary—Tenth Edition* (Fig. 1.3 (b)) and so on. Compared to the Poisson distribution of the ER random graph (see Eq. (1.2)), in which most nodes have a degree around $\langle k \rangle$, the discovery of the power-law degree distribution (Eq. (1.3)) in many real networks tells us that the structure of real networks is scale-free and highly inhomogeneous. Most of the nodes have only a few links. However, there exist a few nodes, called ‘hubs’, which have most of the links of the system. This inhomogeneous structure has interesting consequences for network robustness and tolerance to attacks [5, 71].

1.4.1 Barabási-Albert Model

In order to describe the fundamental mechanisms which cause power-law degree distributions, Barabási and Albert (BA) introduced a simple stochastic growth model [13]. The basic idea of the model is that the probability to connect two nodes is not uniform. By definition, the model has two basic ingredients: (i) growth and (ii) preferential attachment. To incorporate growth, a new node with $m$ edges is connected to $m$ different old nodes at each time step. The preferential attachment provides the connection probability of the $m$ new links. In general, the connection probability of a
Figure 1.3. Example of $P(k)$ distribution of real networks. (a) router level of the Internet: the slope of the data is $\gamma \approx 2.14$, (b) synonym network in linguistics with exponent $\gamma \approx 2.8$. 


new node to already present node $i$ is given by (Fig. 1.4)

$$
\Pi_i = \frac{k_i}{\sum_j k_j}.
$$

Thus if a node $i$ has high degree, then it has a high chance to be connected to, therefore, it has more chances to become a 'hub'. Fig. 1.5 (a) shows for example a numerical simulation result of the BA model with $N = 10^5$ nodes and $m = 2$. The slope of the solid line is $\gamma \approx 3.0$.

The continuum approach introduced by Barabási and Albert [13] and Barabási, Albert, and Jeong [14] calculates the time dependence of the degree $k_i$ of a certain node $i$. This degree will increase every time a new node is introduced into the system and links to the node $i$, with a probability $\Pi(k_i)$ given by Eq. (1.4). Assuming that $k_i$ is a continuous variable, it satisfies the growth equation

$$
\frac{\partial k_i}{\partial t} = m\Pi(k)
$$
Figure 1.5. Example of $P(k)$ distribution of the BA model (a) with preferential attachment and (b) without preferential attachment. The slopes of the solid lines represent (a) $P(k) \sim k^{-3}$ and (b) $P(k) \sim \exp(-k/3.5)$, respectively.
\begin{align*}
\frac{m k_i}{\sum_{j=1}^{N-1} k_j} &= \frac{k_i}{2t}.
\end{align*}

(1.5)

In the denominator of the second line, the summation is over all nodes in the system except the new node, thus in the large $t$ limit $\sum_{j=1}^{N-1} k_j = 2mt - m \simeq 2mt$. The solution of Eq. (1.5) with the initial condition $k_i(t_i) = m$ is

\begin{equation}
k_i(t) = m \left( \frac{t}{t_i} \right)^\beta,
\end{equation}

(1.6)

where $t_i$ is the time at which the node $i$ is introduced and $\beta$ for Eq. (1.5) is $1/2$. From Eq. (1.6) one can write the probability that a node has a degree $k_i(t)$ smaller than $k$ at time $t$ as

\begin{equation}
P(k_i(t) < k) = P \left( t_i > \frac{m^{1/\beta}t}{k^{1/\beta}} \right).
\end{equation}

(1.7)

Assuming that we add the nodes at equal time intervals to the network, the $t_i$ values have a constant probability density

\begin{equation}
P(t_i) = \frac{1}{m_0 + t},
\end{equation}

(1.8)

where $m_0$ is small numbers of nodes given by initial condition. By substituting Eq. (1.8) into Eq. (1.7) one obtains the relation

\begin{equation}
P \left( t_i > \frac{m^{1/\beta}t}{k^{1/\beta}} \right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(t + m_0)}.
\end{equation}

(1.9)

The degree distribution is obtained using

\begin{equation}
P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{(1/\beta)m^{1/\beta}t}{(m_0 + t)k^{1/\beta+1}}.
\end{equation}

(1.10)
In the asymptotic limit \((t \to \infty)\), Eq. (1.10) becomes

\[
P(k) \approx 2m^{1/2}k^{-\gamma},
\]

where

\[
\gamma = \frac{1}{\beta} + 1 = 3.
\]

Without preferential attachment, the rate of change of the connectivity of node \(i\) can be written as

\[
\frac{\partial k}{\partial t} = \frac{m}{m_0 + t - 1}.
\]

Solving Eq. (1.13) with initial condition \(k_i(t_i) = m\) one obtains the relation

\[
k_i = m(\ln(m_0 + t - 1) - \ln(m_0 + t_i - 1) + 1).
\]

and in the asymptotic limit the degree distribution decays exponentially [13],

\[
P(k) = \frac{e}{m} \exp \left( -\frac{k}{m} \right).
\]

Fig. 1.5 (b) shows the \(P(k)\) distribution of the model without preferential attachment which follows the exponentially decaying function Eq. (1.15). Both with and without preferential attachment, the continuum theory provides predictions in agreement with the numerical simulations.

1.4.2 Clustering Coefficient and Modularity

An important property of social networks is the existence of a strong tendency for clustering. Clusters may represent circles of friends, acquaintances, or coworkers. As in a cluster most members know each other, a cluster or group is highly inter-
connected. In order to quantify this clustering tendency, Watts and Strogatz [151] defined the clustering coefficient. Consider a node $i$ in the network with $k_i$ edges. If the nearest neighbors of the node $i$ were part of a clique, there would be $k_i(k_i - 1)/2$ edges between them. The ratio between the number of edges $n_i$ that actually exist between these $k_i$ nodes and the total number of possible connections gives the value of the clustering coefficient of node $i$,

$$C_i = \frac{2n_i}{k_i(k_i - 1)}. \quad (1.16)$$

The scale-free property and clustering phenomena are observed at the same time in many real networks including metabolic networks [72, 147], the protein interaction networks [71, 146], the WWW [4], and some social networks [103]. However, most models, that have been studied so far, have difficulty capturing the coexistence of the scale-free (SF) property and the high clustering. The numerical simulations have shown that for the BA model [13, 14], the average clustering coefficient depends on the system size $C(N) \sim N^{-0.75}$ [3, 41], which is significantly larger for large $N$ than the results for the random network, which follow $C(N) \sim N^{-1}$. Yet, this prediction still disagrees with the finding from empirical measurements, which indicate that $C$ does not depend on $N$ [3]. In order to incorporate the independency of $C$ on $N$, Ravasz and Barabási introduced the concept of hierarchical networks [16, 124, 123]. The main idea of the model is that the network has basic modules and by copying the structure of these basic modules, a hierarchical structure emerges. They showed that the hierarchical structure can provide both scale-free degree distributions and high clustering coefficients which do not depend on the size of network $N$. An important discovery is that for a hierarchical structure the clustering coefficient depends on the
degree $k$, following a power-law:

\[ C(k) \sim k^{-\delta}, \]  

(1.17)

which is not observed in random networks or simple scale-free networks. Thus, the degree dependent clustering coefficient can be used as evidence for the existence of hierarchical structure or modularity.

1.5 Fluctuations of Random Walkers on Scale-Free Networks

Dynamical processes on networks can describe several categories such as the traditional random walker problems [6, 110, 111], synchronization [109], and disease spreading [21]. Recently, in order to study the dynamical organizing principles common for a wide range of complex systems, de Menezes and Barabási studied the relationship between the average flux $\langle f_i \rangle$ and dispersion $\sigma_i$ at node $i$ [35], finding a power-law relationship between them

\[ \sigma \sim \langle f \rangle^{\alpha}, \]  

(1.18)

where $\alpha$ is a dynamical scaling exponent. Based on empirical data, they found that the value of $\alpha$ can be $1/2$ or $1$. The Internet and the logic gate network in microprocessors belong to the $\alpha \simeq 1/2$ class. On the other hand, the WWW and the highway traffic in Colorado and Vermont belong to the $\alpha \simeq 1$ class. To understand the origin of different scaling exponents, they introduced a daily variation of the number of random walkers, $\Delta W$, as an “external” factor, and they showed that if $\Delta W$ exceeds certain threshold then one can observe a crossover between $\alpha = 1/2$ to $\alpha = 1$. In Chapter 6, we will discuss the relationship between fluctuations in diffusive systems with multiplicative...
noise and the underlying network topology.
CHAPTER 2

WEIGHTED EVOLVING NETWORKS

2.1 Introduction

The recognition that real networks are fundamentally different from the random models that dominated the mathematical literature in the past forty years [23, 46] lead to a surge of activity in addressing the statistical properties of these systems [8, 13, 32, 38, 39, 102, 78, 151, 150]. In one aspect most recently developed models, aimed to describe the large-scale topology of complex networks, are incomplete when compared with real systems: they assume that all links are equivalent. But in many fields it is well known that the interaction strengths can vary widely, such variations being essential to the network’s ability to carry on its basic functions. Sociologists have repeatedly argued about the importance of assigning strengths to social links, finding that the weak links people have outside their close circle of friends play a key role in keeping the social system together [56]. Recently, Newman has showed that assigning weights to the links between scientists allows for a better characterization of the scientific collaboration web [103]. Similarly, there is an ongoing discussion about the importance of weak links between species in guaranteeing the stability of an ecosystem [18]. Finally, many transportation networks in nature, ranging from cardiovascular to respiratory networks, have well defined weights or flow rates assigned
to the links, whose magnitude is intimately determined by the network’s topology [12]. The issue of link strength has been extensively addressed in the neural network literature. The question posed in that context so far had an unique focus: given a network topology, how can one alter the link weights in a dynamical fashion to allow the network to perform certain desired functions, ranging from memory to pattern recognition [101]? Similarly, research on allometric scaling has also been concerned with assigning weights to links on a network with fixed, often tree-like topology [12]. On the other hand, the recent advances in statistical modeling of complex networks have brought the community’s attention towards large networks whose topology evolves in time. Despite the known importance of interaction strengths in the various systems these models aim to describe, in this context there have been no attempts to model networks other than binary nets, whose links have weights 0 or 1.

In this chapter we will present a systematic study of evolving networks with non-binary connectivities. We introduce and investigate two models that assign weights to new links as they are dynamically created, providing a prototype of a weighted evolving network. While we choose the simplest possible models, in which the weights are driven by the network connectivity only, numerical simulations indicate that the distribution of the total weight scales differently from the total connectivity. However, an analytical solution reveals that the different scaling behavior can be explained by strong logarithmic correction, and asymptotically the investigated weighted networks belong to the same universality class as their unweighted counterparts.
2.2 Defining Weighted Network Models

2.2.1 Weighted Scale-Free (WSF) Model

Starting from a small number \(m_0\) of vertices, at each time step we add a new node which links to \(m\) existing nodes in the system. The probability that a new node \(j\) will connect to an existing node \(i\) is

\[
\Pi_i = \frac{k_i}{\sum_j k_j},
\]

(2.1)

where \(k_i\) is the total number of links that the node \(i\) has. In assigning a weight to the newly established link \(j \leftrightarrow i\), we assume that the weight \(w_{ji}\) is proportional to \(k_i\), i.e., more connected (and therefore more “powerful”) nodes gain more weight and is symmetric \((w_{ji} = w_{ij})\). Also, one can assume that all new nodes have fairly uniform total ‘resources’ for linking to other nodes in the system, we therefore require that each new node has a fixed total weight, i.e. we normalize \(w_{ij}\) such that the sum of the weights for the \(m\) new links is \(\sum_{\{i'\}} w_{j'i'} = 1\), where \(\{i'\}\) represents a sum over the \(m\) existing nodes to which the new node \(j\) is connected. As a result of the two assumptions, each link \(i \leftrightarrow j\) of the newly added node \(j\) is assigned a weight as

\[
w_{ji} = \frac{k_i}{\sum_{\{i'\}} k_{i'}}.
\]

(2.2)

2.2.2 Weighted Exponential (WE) Model

The model is inspired by model A discussed in Refs. [4, 14] or Section 1.4.1, and is defined as follows: at every time step we add a new node with \(m(\leq m_0)\) links, connected with \textit{equal probability} to the nodes already present in the system. The weights of the links are assigned again by using Eq. (2.2).

The difference between the WSF and WE models comes in preferential attach-
ment, which is known to fundamentally alter the topology [4, 8, 14, 38, 39, 78, 79]:

The WSF model generates a scale-free network whose connectivity distribution follows \( P(k) \sim k^{-3} \), while the network generated by the WE model is exponential with the connectivity distribution following \( P(k) = \frac{c}{m} e^{-k/m} \). Since the weights of the links are driven by the connectivity, this difference is expected to lead to significant changes in the distribution of the link strengths as well.

2.3 Results of Numerical Simulations

We start by investigating the weight distribution of the two models. As Fig. 2.1 (a) and (b) show, both the WE and the WSF models lead to a peaked and skewed weight distribution, whose tails decay exponentially (or faster) for large \( w_{ij} \). The boundedness of \( P(w_{ij}) \) is due to the normalization condition, which does not allow individual weights to be larger than 1. Most important, however, we find that the distribution is stationary, i.e. \( P(w_{ij}) \) is independent of time (and system size).

While the individual weights assigned to links, \( w_{ij} \), are bounded, we get a very different picture when we study the total weight associated with a selected node. In binary networks the node’s importance is characterized by the total number of links it has, \( k_i \). Similarly, in a weighted network the importance of a node \( i \) can be measured by its total weight, obtained by summing the weights of the links that connect to it, \( w_i = \sum_{\{j\}} w_{ij} \).

Due to the normalization condition (2.2) a new node has \( w_i = 1 \), but \( w_i \) increases in time when a link from a new node is added links to \( i \). Since in both models the weights are determined by the network connectivity, we expect that \( P(w) \) closely follows \( P(k) \). In contrast, the numerical results summarized in Fig. 2.2 indicate
Figure 2.1. The distribution $P(w_{ij})$ of the individual link weights, $w_{ij}$ for the (a) WE and the (b) WSF models, defined in the text ($m = 2$). The symbols correspond to different system sizes (or time), i.e. $N = 10^3(\bigcirc), 10^4(\square), 10^5(\check{\triangle})$ and $10^6(\triangle)$. The insets shows the same data on a log-linear plot, indicating that the tail decays faster than exponential.
striking differences between $P(k)$ and $P(w)$. As Fig. 2.2 (a) shows, while for the
WE model $P(k)$ decays exponentially, $P(w)$ systematically deviates from a simple
exponential behavior. This difference is even more evident in the network dynamics:
while both $k_i(t)$ and $w_i(t)$ appear to increase logarithmically in time, they can be
fitted with a different slope on a log-linear plot (Fig. 2.2 (b)). Similar systematic
discrepancies are observed for the WSF model as well: as Fig. 2.2 (c) indicates,
while $P(w)$ and $P(k)$ can be fitted with power-laws, $P(w) \sim w^{-\sigma}$ and $P(k) \sim k^{-\gamma}$,
it appears that $\gamma = 3$ and the exponent $\sigma$ is different from $\gamma$. Furthermore, we find
that $\sigma$ depends strongly on $m$ (Fig. 2.2 (c)). Again, this difference is reflected in
the dynamical behavior of $k_i(t)$ and $w_i(t)$: as Fig. 2.2 (d) indicates, $w_i(t) \sim t^\beta$ with
$\beta > 1/2$, in contrast with $k_i(t) \sim t^{1/2}$ [13, 14] predicted by the binary scale-free model.

2.4 Analytic Derivation of $w_i(t)$

To understand the different behaviors of $w_i$ and $k_i$ uncovered by the numerical sim-
ulations, we resort to the use of an analytical method in determining the averaged
behavior of $w_i(t)$ for the discussed model. To simplify the discussion in the following
we assume $m = 2$, however, the calculations can be generalized for arbitrary $m$. The
total weight of node $i$ at time $t$ can be written as

$$
\begin{align*}
w_i(t) &= 1 + \sum_{\{j\}} w_{ij} = 1 + \int_{t_i^0}^{t} \tilde{P}_i(t') \langle w_{ij}(t') \rangle dt',
\end{align*}
$$

where $\tilde{P}_i(t)$ is the probability that the node $i$ is selected to be connected to a new
node $j$ at time $t$ and $t_i^0$ is the time at which the node $i$ has been added to the system.
$\langle w_{ij} \rangle$ is the average weight of link between $i \leftrightarrow j$ once the link is established. When
a new node $j$ and the list of $m$ nodes $\{i'\}$ to which it connects are selected, the
Figure 2.2. (a) Distribution $P(w)$ of the total connectivity $w$ assigned to individual nodes for the WE model. The symbols correspond to different values of $m$, i.e. $m = 2$ (○), 3 (□), 4 (△) and 5 (△). The inset shows the connectivity distribution, $P(k)$, for the same parameters as in the main panel. (b) Time dependence of $k_i(t)$ (○) and $w_i(t)$ (□) for a randomly selected node $i$ for the WE model ($i = 5000$). (c) $P(k)$ (○) and $P(w)$ (□) distributions for the WSF model for $m = 5$. The inset shows the same data for $m = 2$. (d) $k_i(t)$ (○), $w_i(t)$ (□) vs. $t$ for the WSF model ($i = 10000$).
weights of the links, $w_{ji'}$ are assigned according to Eq. (2.2). These weights depend on the number of links the selected nodes have, i.e. $\{k_i\}$. If we assume that node $j$ is connected to nodes $i$ and $l$ $(m = 2)$, we have

$$\langle w_{ij}(t) \rangle = \int_{m}^{\infty} w_{ji}(l) \mathcal{P}(k_l) \, dk_l$$

(2.4)

where $w_{ji}(l)$ is the weight between the $j$ and $i$ nodes, $\mathcal{P}(k_l)$ is the probability distribution of $k_l$, the total link number of node $l$. Substituting Eq. (2.4) into Eq. (2.3), we have obtained

$$w_i(t) = 1 + \int_{t_0}^{t} \int_{m}^{\infty} \tilde{P}_i(t') w_{ji}(l) \mathcal{P}(k_l) \, dk_l \, dt'.$$

(2.5)

According to (2.2) for $m = 2$, the weight $w_{ji}(l)$ is given by

$$w_{ji}(l) = \frac{k_i}{k_i + k_l},$$

(2.6)

thus Eq. (2.5) becomes

$$w_i(t) = 1 + \int_{t_0}^{t} \int_{m}^{\infty} \tilde{P}_i(t') \frac{k_i}{k_i + k_l} \mathcal{P}(k_l) \, dk_l \, dt'.$$

(2.7)

Eq. (2.7) represents a general expression for calculating $w_i(t)$ for $m = 2$. To apply it to the WE and WSF models, we need to calculate explicitly $\tilde{P}(t)$ and $\mathcal{P}(k_l)$.

2.4.1 WE Model

In the WE model the nodes to which a new node connects to are selected uniformly among all existing nodes, thus the probability that a node $i$ will be picked is independent of this node’s connectivity and is given by

$$\tilde{P}_i(t) = \frac{m}{t + m_0}.$$

(2.8)
Similarly, the connectivity distribution and the dynamical behavior of a single node are given by [14]

\[ \mathcal{P}(k) = A e^{-k/m} = \frac{e}{m} e^{-k/m}, \]  

\[ k_i(t) = m \left[ \ln(m_0 + t - 1) - \ln(m_0 + t_0 - 1) + 1 \right] \]

\[ = m \left[ \ln(at + b) + 1 \right], \]

where \( a = \frac{1}{m_0 + t_0 - 1}, \ b = \frac{m_0 - 1}{m_0 + t_0 - 1} \) and the normalization condition is 1 = \( \int_m^\infty \mathcal{P}(k) dk \).

Substituting Eq. (2.9) into Eq. (2.7), we obtain

\[ w_i(t) = 1 + e \int_t^{t_0} \int_m^\infty \frac{1}{t' + m_0 k_i(t')} \frac{k_i(t')} {k_i(t')} e^{-k_i/m} \, dk_i \, dt'. \]  

(2.10)

After performing the integration and inserting \( k_i(t) \) from (2.9), for large \( t \) we obtain

\[ w_i(t) \simeq m \ln(at + b) - m \ln(\ln(at + b) + 2) + C, \]  

(2.11)

where \( C \) is an integration constant independent of \( t \). Therefore the relation between \( w_i(t) \) and \( k_i(t) \) for large \( t \) follows

\[ w_i(t) \simeq k_i(t) - m \ln \ln t + C. \]  

(2.12)

The prediction (2.12) is fully supported by numerical simulations: in Fig. 2.3 (a) we plot the difference \( w_i(t) - k_i(t) \) as function of \( \ln \ln(t) \), showing that the difference indeed follows a double logarithmic law. This result is very interesting since it indicates that the different slopes observed in Fig. 2.2 (b) for \( k_i(t) \) and \( w_i(t) \) do not represent distinct power-law scaling behaviors, but are the result of logarithmic corrections.
Figure 2.3. The difference \((w_i(t) - k_i(t))\) for the (a) WE and the (b) WSF models. The continuous lines in each case represent the analytic solution (11) and (13), respectively. We limited the simulations to nodes appearing at large \(t_i^0\) (\(t_i^0 = 10^4\)) to capture the asymptotic limit, that is predicted by our predictions (11) and (13). We find that for smaller \(t_i^0\) the crossover time for the convergence to the analytic solution is numerically prohibited.
2.4.2 WSF Model

In the scale-free model the probability distributions and \( k_i(t) \) are given by [14]

\[
\begin{align*}
\tilde{P}_i(t) & = \frac{m k_i(t)}{\sum_j k_j} = \frac{m}{2mt} k_i(t) = \frac{k_i(t)}{2t}, \\
P(k) & = mk^{-2} \propto k \cdot P(k), \quad (2.13) \\
k_i(t) & = \frac{m}{\sqrt{t_0}} \sqrt{t}.
\end{align*}
\]

Substituting Eq. (2.13) into Eq. (2.7), and performing the integrals we have obtained

\[
\frac{w_i(t)}{k_i(t)} \approx k_i(t) - \frac{m}{8} \left( \ln \left( \frac{m^2 t}{t_0} \right) \right)^2 + \frac{m}{2} \ln (m) \ln \left( \frac{t}{t_0} \right) + C', \quad (2.14)
\]

indicating that despite a different scaling behavior suggested by the numerical simulations (Fig. 2.2 (d)), we are dealing with strong logarithmic corrections and asymptotically two scaling laws are the same. Again, the analytical prediction Eq. (2.14) is confirmed by more detailed numerical simulations shown in Fig. 2.3 (b).

Our ability to calculate analytically \( w_{ij} \) for the discussed models is based on the fact that the weights are driven by the connectivity distribution. To address the generality of our results we investigated another extensions of these two models, as discussed in the following section.

2.5 Weight Driven Connectivity

In some systems the topology could be driven by the total weights, and not by the connectivity. Thus we assume that the probability Eq. (2.1) that a new node is connected to a node \( j \) is

\[
\Pi_i = \frac{w_i}{\sum_j w_j}, \quad (2.15)
\]
Figure 2.4. Plot of $P(k)$ (□) and $P(w)$ (○) distributions obtained from numerical simulation of weight driven connectivity model. The slope of the solid line represents $\gamma \simeq 2.94$.

where $w_i$ is the weight of node $i$. The weights are then assigned following Eq. (2.2). We have found that the scaling of this network is identical to that of the scale-free model, and the evolution of the weights also follows the paradigm established for the SF model (Fig. 2.4).

2.6 Discussion

Extensive simulations of networks whose size is comparable to the real networks that are currently available indicate the emergence of new scaling exponents for the behavior of the total weights. However, the analytical solutions reveal that the results are affected by strong logarithmic corrections, and asymptotically the scaling behaviors
of the weighted and unweighted models are identical. This result raises important questions regarding our ability to uncover the correct scaling behavior of real weighted networks, should such data become available in the near future: the real exponents could be easily shadowed by corrections to scaling similar to that encountered in the investigated models here.

The results presented in this chapter represent only the starting point towards understanding weighted networks. In some real systems, diverse dynamical rules can govern the assignment of weights to links, which could result in statistical properties of the network that are different from that discussed here. In particular, we assumed that once a weight has been assigned to a link, it stays unchanged, which is often not the case in more realistic networks: weights can evolve dynamically just as the network topology does. For example, acquaintance can turn into friendship by strengthening a previously weak link. Determining the generic behavior of such complex evolving systems is a real challenge for future research. Despite these limitations, the investigated models give a glimpse into the complex behavior we are facing as we attempt to make network modeling more realistic by incorporating weights.
CHAPTER 3

SPATIALLY CORRELATED NETWORK MODEL

3.1 Introduction

In general, a network can be regarded as an infinite dimensional object. However, in some systems, for example, in the Internet, due to social and economical reasons, there could be a restriction on the distribution of resources, hence, this restrictions can affect the network’s large-scale topology. In light of extensive evidence that Internet protocol performance is greatly influenced by the network topology [83], network generators are a crucial prerequisite for understanding and modeling the Internet. Indeed, security and communication protocols perform poorly on topologies provided by generators different from which they are optimized for, and are often ineffective when released. Protocols that work seamlessly on prototypes fail to scale up, being inefficient on the larger real network. Thus to efficiently control and route traffic on an exponentially expanding Internet [81], it is important that topology generators not only capture the structure of the current Internet, but allow for efficient planning and long-term network design as well.

Our ability to design good topology generators is limited by our poor understanding of the basic mechanisms that shape the Internet’s large-scale topology. Indeed, until recently all Internet topology generators [27, 152], which are software designed
to generate realistic network topologies with several input parameters for research and
development purposes, provided various random graphs [23, 46]. The 1999 discovery
of Faloutsos et al. [47] that the Internet is a scale-free network with a power-law de-
gree distribution [13] invalidated all previous modeling efforts. Subsequent research
confirmed that the difference between scale-free and random networks are too signi-
cant to be ignored: protocols designed for random networks fare poorly on a scale-free
topology; a scale-free Internet displays high tolerance to random-node failures but is
fragile against attacks [5, 32, 33]; computer viruses spread threshold free on scale-free
networks [116] with obvious consequences on network security. These insights moti-
vated the development of a new brand of Internet topology generators [73, 94] that
provide scale-free topologies in better agreement with empirical data. Despite these
rapid advances, it is unclear that we are aware of all driving forces that govern the
Internet’s topological evolution. It is therefore of crucial importance to perform mea-
surements that directly probe and uncover the mechanisms that shape the Internet’s
large-scale topology.

Here we offer direct empirical evidence for a series of fundamental mechanisms
that drive the Internet’s evolution and large-scale structure. In contrast with the
random placement of nodes, we find that the Internet develops on a fractal support,
driven by the fractal nature of population patterns around the world. In contrast
with current modeling paradigms, which assume that the likelihood of placing a link
decays exponentially with the link’s length, we find that this dependence is only linear.
Finally, we provide quantitative evidence that preferential attachment, responsible
for the scale-free topology, follows a linear functional form on the Internet. These
results allow us to identify a class of models that could serve as a starting point for
topologically correct network generators. Surprisingly, the obtained phase diagram indicates that all current Internet network generators are in a different region of the phase space than the Internet.

3.2 Physical Layout

At the highest resolution the Internet is a network of routers connected by links. As each router belongs to some administrative authority, or autonomous systems (AS), the Internet is often considered as a network of interconnected ASs. For completeness, here we study simultaneously the router and AS level topology, using the term node to represent both routers and ASs unless specified otherwise.

Current Internet topology generators assume that routers and domains are distributed randomly in a 2D plane [27, 73, 94, 152]. In contrast, we find that routers and AS’s form a fractal set [59], strongly correlating with the population density around the world. In Fig. 3.1 (a) we show a map of the worldwide router density, obtained by using the NETGEO (www.caida.org/tools/utilities/netgeo) tool to identify the geographical coordinates of 228,265 routers provided by the currently most extensive router-level Internet mapping effort. Compared with the population density map obtained from CIESIN (http://sedac.ciesin.org/plue/gpw) (Fig. 3.1 (b)), the results indicate strong, visually evident correlations between the router and the population density in economically developed areas of the world. We used a box counting method [59, 144] to analyze the spatial distribution of router, AS, and population density. The results shown in Fig. 3.2 (a), indicate that each of the three sets form a fractal with dimension $D_f = 1.5 \pm 0.1$. The coincidence between the fractal dimension of the population and the Internet (router and AS) nodes is not unexpected: high population
Figure 3.1. Distribution of the Internet around the world. (a) Worldwide router density map, obtained by using the NETGEO tool to identify the geographical location of 228,265 routers mapped out by the extensive router level mapping effort of Govindan and Tangmunarunkit. (b) Population density map based on the Columbia University’s Center for International Earth Science Information Network’s (CIESIN) population data. Both maps are shown with a box resolution $1^\circ \times 1^\circ$. The bar next to each map gives the range of values encoded by color code, indicating that the highest population density within this resolution is of the order $10^7$ people/box, while the highest router is of the order of $10^4$ routers/box. Note that while in economically developed nations there are visibly strong correlations between population and router density, in the rest of the world Internet access is sparse, limited to urban areas characterized by population density peak.
density implies higher demand for Internet services, resulting in higher router and domain density. Fig. 3.2 (b) supports the existence of such correlations, indicating that the router and AS density increase monotonically with the population density.

3.3 Distribution of Links

Connecting two nodes on the Internet requires extensive resource and time investment. Thus network designers prefer to connect to the closest node with sufficient bandwidth, a process that clearly favors shorter links. To discourage long links, all topology generators are based on the Waxman model [152], which assumes that the likelihood of placing a link between two nodes separated by the Euclidean distance $d$ decays as $P(d) \sim \exp(-d/d_0)$, where $d_0$ is a free parameter taken to be proportional to the system size. Despite its wide use in Internet topology generators [27, 94, 152], there is no empirical evidence for such exponential form, which forbids links between faraway nodes. Intuition suggests otherwise: one would expect that the likelihood of connecting two nodes is inversely proportional with the distance between the nodes, i.e.,

$$P(d) \sim \frac{1}{d}.$$  \hspace{1cm} (3.1)

Indeed, the cost of placing a physical link between two existing routers has two components to it:

(i) a fixed technical and administrative connection cost at the two ends of the link,

(ii) a cost of the physical line and its maintenance.

The second factor is proportional to the line’s length. For large distances the distance-dependent cost dominates, potentially suggesting an $1/d$ asymptotic dependence for the probability to connect two routers. The correct functional form of $P(d)$
Figure 3.2. Characterizing the Internet's physical layout and topology by using direct measurements. (a) The physical layout of the Internet was studied by using a box counting method [59, 144], applied to the map shown in Fig. 3.1. The log-log plot shows the number of boxes of size $\ell \times \ell \text{km}^2$ with nonzero routers/AS/inhabitants in function of $\ell$ for North America. The slope of the straight line indicates that $D_f \approx 1.5 \pm 0.1$ for each dataset. (b) The dependence of the router/AS density on the population density in North America, showing the average number of router/AS nodes in a $1^\circ \times 1^\circ$ box in function of the number of people living in the same area. Similar plots were obtained for each continent, the steepness of the curves strongly correlated to economic factors. (c) The cumulative length distribution of the links connecting routers. $d$ is the Euclidean distance between routers and $R = 6,378 \text{km}$ is the radius of the Earth. (d) The cumulative change $\Delta k$ in the connectivity of AS nodes with $k$ links. The dotted line has slope 2, indicate the linear preferential attachment of links.
is crucial for Internet modeling: our simulations indicate that a network developing under the Waxman rule asymptotically converges to a network with an exponentially decaying degree distribution, in contrast with the power-law documented for the Internet. Therefore, to uncover the proper form of $P(d)$ we measured the length distribution of the documented Internet links. The results, shown in Fig. 3.2 (c), indicate that both router and AS level $P(d)$ decays linearly with $d$, excluding Waxman’s rule.

3.4 Preferential Attachment

Preferential attachment is believed to be responsible for the emergence of the scale-free topology in complex networks [5]. It assumes that the probability that a new node will link to an existing node with $k$ links depends linearly on $k$, i.e.,

$$
\Pi(k) = \frac{k}{\sum_i k_i}.
$$

(3.2)

On the other hand, in real systems preferential attachment could have an arbitrary nonlinear form. Calculations indicate, however, that for $\Pi(k) \sim k^\alpha$, with $\alpha \neq 1$ the degree distribution deviates from a power-law [78]. In the light of these results, to properly model the Internet, we need to determine the precise functional form of $\Pi(k)$. To achieve this, we use Internet AS maps recorded at 6-month intervals, allowing us to calculate the change $\Delta k$ in the degree of a AS node with $k$ links during the investigated time frame. The results indicate that the rate at which a node increases its degree is linearly proportional with the number of links the node has, offering quantitative support for the presence of linear preferential attachment (Fig. 3.2 (d)), supported by independent measurements as well [78].
Taken together, our measurements indicate that four mechanisms, acting independently, contribute to the Internet’s large-scale topology. First, in contrast with classical network models the Internet grows incrementally, being described by an evolving network [13, 30, 43, 78] rather than a static graph [23, 46]. Second, nodes are not distributed randomly, but both routers and domains form a scale-invariant fractal set with fractal dimension $D_f = 1.5$. Finally, link placement is determined by two competing mechanisms. First, the likelihood of connecting two nodes decreases linearly with the distance between them, and second, the likelihood of connecting to a node with $k$ links increases linearly with $k$. Building on these mechanisms, each supported independently by our measurements, we propose a general model that provides an integrated framework to investigate the effect of the different mechanisms on the Internet’s large-scale topology.

3.5 Modelling the Internet

Consider a map, mimicking a continent, which is a 2-dimensional surface of linear size $L$. The map is divided into squares of size $\ell \times \ell$ ($\ell \ll L$), each square being assigned a population density $\rho(x, y)$ with fractal dimension $D_f$. At each time step we place a new node $i$ on the map, its position being determined probabilistically, such that the likelihood of placing a node at $(x, y)$ is linearly proportional with $\rho(x, y)$. We assume that the new node connects with $m$ links to nodes that are already present in the system. The probability that the new node links to a node $j$ with $k_j$ links at distance $d_{ij}$ from node $i$ is

$$\Pi(k_j, d_{ij}) \sim \frac{k_j^\alpha}{d_{ij}^\gamma},$$

(3.3)
where $\alpha$ and $\sigma$ are preassigned exponents, governing preferential attachment and the cost of the node-node distance. Increasing $\alpha$ will favor linking to nodes with higher degree, whereas a higher $\sigma$ will discourage long links.

The parameters of the model can be assigned into two qualitatively different classes. First, $L$, $\ell$, and $m$ are nonuniversal parameters, as their value can be changed without affecting the network's large-scale topology. On the other hand, $\alpha$, $\sigma$, and $D_f$ are universal exponents, as their values uniquely parameterize a family of Internet models, generating potentially different large-scale topologies. Therefore, we use a 3-dimensional phase space whose axis are the scaling exponents, $2 - D_f$, $\alpha$ and $1/(\sigma + 1)$ (Fig. 3.3) to identify the possible scaling behavior predicted by the model. For easy reference, we show the location within this phase space of all currently used Internet topology generators. Our measurements (Fig. 3.2) allow us to unambiguously identify the position of the Internet within this phase space at $\sigma = 1$, $\alpha = 1$, and $D_f = 1.5$, clearly separated from all network generators. Such separation should not be a problem if some of the models and the Internet belong to a region of the phase space that share the same universal topological features. We will show next, however, that this is not the case, as deviations from the point denoting the Internet can significantly alter the network’s large-scale topology.

In order to systematically investigate the changes in the network topology as we deviate from the point denoting the Internet next we consider the effect of changing $\sigma$, $\alpha$, and $D_f$, moving separately along the three main axis.

Varying $\sigma$ while leaving $\alpha = 1$ and $D_f = 1.5$ unchanged changes the contribution of the Euclidean distance to the network topology, interpolating between the $\sigma = 0$ phase corresponding to the scale-free model and the $\sigma = \infty$ limit, corresponding to
Figure 3.3. Phase diagram summarizing various internet models and their expected large-scale topology. The axes represent the scaling exponents $1/(\sigma + 1)$, $2 - D_f$ and $\alpha$, governing link placement, node location, and preferential attachment, respectively. Our measurement indicates (Fig. 3.2) that within this phase space, the Internet can be found at $1/(\sigma + 1) = 1/2$, $2 - D_f = 1/2$, and $\alpha = 1$, identified as a red circle. The yellow boxes indicate the location of all current Internet topology generators.

WAXMAN [152]: Nodes are placed randomly in space ($D_f = 2$) with exponential distance dependence [$1/(\sigma + 1) = 0$] and without preferential attachment ($\alpha = 0$).

TIERS [81]: Based on a three-level hierarchy the model has no space dependence [$D_f = 2$, $1/(\sigma + 1) = 1$] and no preferential attachment ($\alpha = 0$). GT-ITM [27]: While based on several different models, the most used pure random transit-stub version occupies the same position in the phase space as TIERS [$D_f = 2$, $1/(\sigma + 1) = 1$, $\alpha = 0$].

INET2.0 [73] connects randomly placed nodes by using an externally imposed power-law connectivity information. While it does not include preferential attachment explicitly, as $P(k)$ is forced to follow a power law, we put this generator at $D_f = 2$, $1/(\sigma + 1) = 1$, $\alpha = 1$. Note, however, that there are significant known differences [77] between a static graph, such as generated by INET, and scale-free topologies generated by evolving networks, such as the Internet. BRITE [94]: The most advanced of all, BRITE incorporates preferential attachment ($\alpha = 1$) combined with the Waxman rule ($1/(\sigma + 1) = 1$) for placing the links. As BRITE has the option to produce topologies with different parameters, we denote by BRITE-1 the version with only preferential attachment [$D_f = 2$, $1/(\sigma + 1) = 1$, $\alpha = 1$], and BRITE-2 the version including the Waxman rule as well [$D_f = 2$, $1/(\sigma + 1) = 0$, $\alpha = 1$]. Note that BRITE has an option to include inhomogeneous node placement, creating regions with high-node density mimicking highly populated areas. The algorithm, however, does not create a fractal, thus we choose $D_f = 2$ for both BRITE-1 and BRITE-2. The scale-free model [13], which ignores the physical location of the nodes ($\sigma = 0$ thus $D_f$ can be arbitrary) is shown as a separate blue line on the $1/(\sigma + 1) = 1$ axis and $\alpha = 1$. The green areas correspond to an exponential $P(k)$ distribution, while yellow areas are characterized by gelation, indicating that the Internet strikes a delicate balance at the boundary of these two topologically distinct phases.
the Waxman rule. As Fig. 3.4 (a) shows, for an exponential distance dependence (Waxman’s rule) the degree distribution $P(k)$ develops an exponential tail, disagreeing with the power law $P(k)$ of the Internet [47]. Moving toward the $\sigma = 0$ axis, as the physical distance gradually loses relevance in Eq. (3.3), we recover the scale-free model, for which the physical layout does not influence the network topology. Changing $\sigma$ affects the link length distribution $P(d)$ as well. As shown in Fig. 3.4 (b), for $\sigma = 1$ the $P(d)$ distribution quantitatively agrees with the $1/d$ dependence uncovered by the direct measurements (Fig. 3.2 (c)), but in the exponential limit of the Waxman rule we find that $P(d)$ develops an exponential tail, a functional form that characterizes the full $1/(\sigma + 1) = 1$ plane (i.e., valid for arbitrary $\alpha$ and $D_f$ as long as $\sigma = \infty$). Finally, decreasing $\sigma$ does not seem to affect $P(k)$, but it does change $P(d)$, which for $\sigma = 0$ develops an extended plateau for $d/R < 10^{-2}$, followed by rapid decay, in disagreement with the measurements (Fig. 3.2 (c)). Consequently, moving with $\sigma$ away from the empirically determined $\sigma = 1$ value affects both the degree and the distance distribution, creating significant deviations from the known Internet topology.

Varying $\alpha$ while leaving $D_f$ and $\sigma$ unchanged has drastic immediate effects on the degree distribution, forcing on it an exponential form. Indeed, for any $\alpha < 1$, the distribution $P(k)$ develops an exponential tail, turning into a pure exponential for $\alpha = 0$, when preferential attachment is absent (Fig. 3.4 (c)). This finding agrees with the analytical predictions of Kaprivsky et al. [78], who have shown that $\alpha < 1$ destroys the power-law nature of the degree distribution in the scale-free model. The calculations predict that for $\alpha > 1$ gelation takes place, leading to a network architecture in which all nodes are connected to a central node. Our simulations fully
confirm these prediction in the vicinity of the $D_f = 1.5$ and $\sigma = 1$ point, as for $\alpha > 1$ the $P(k)$ distribution develops an elongated nonpower law tail, corresponding to a few highly connected nodes, a characteristic feature of gelation. Furthermore, our simulations indicate that the gelation phase is present in the full $\alpha > 1$ region of the phase space, colored yellow in Fig. 3.3. Consequently, we find that any deviation from $\alpha = 1$ results in a significant alternation of the network’s $P(k)$ distribution, while having little effect on $P(d)$ (Fig. 3.4 (d)).

Varying $D_f$ while leaving $\sigma$ and $\alpha$ unchanged has little visible effect on the degree distribution (Fig. 3.4 (e)). The only changes appear in $P(d)$. Indeed, we find that placing the nodes proportional to the population density, thereby creating a fractal with $D_f = 1.5$, results in a $P(d)$ distribution in agreement with the data points provided by the direct measurements (Fig. 3.4 (f)). On the other hand, a random node distribution, corresponding to $D_f = 2$, generates a $P(d)$ that is clearly different from the real data. A uniform distribution appears to lead to a long plateau in $P(d)$, followed by a faster decay than seen in the real Internet.

3.6 Summary and Discussion

In summary, our measurement and simulations indicate that the Internet takes up a very special point in the $(\alpha, D_f, \sigma)$ phase space, such that deviations from its current position identified by direct empirical measurements can significantly alter the network topology. This Internet’s position within the phase space can be understood if one inspects the network’s evolution. Indeed, the population density-driven router placement determines the $D_f$ point and the cost of the cables determines the $\sigma = 1$ point. The only potential explanation for $\alpha = 1$ is that a router’s attractiveness is
Figure 3.4. The dependence of the degree distribution (Left) and link length distribution (Right) on the scaling exponent \( \sigma \), \( \alpha \) and \( D_f \). (a) and (b): The effect of changing \( \sigma \), with \( \alpha = 1 \) and \( D_f = 1.5 \) unaltered, on \( P(k) \)(a) and \( P(d) \)(b). Note that the exponential Waxman’s rule corresponds to \( \sigma = \infty \), in which case \( P(k) \) develops an exponential tail. (c) and (d): The effect of changing \( \alpha \) with fixed \( \sigma = 1 \) and \( D_f = 1.5 \). \( \alpha > 1 \) correspond to the gelation phase. (e) and (f): The effect of changing \( D_f \) while fixing \( \alpha = \sigma = 1 \).
determined mostly by the bandwidth it offers, which appears to scale linearly with the router’s degree. Interestingly, we find that all current generators lay in a different region of the phase space than the Internet (Fig. 3), indicating that they generate networks that belong to a different topological class. Although the changes induced by not considering the fractal nature of the router distribution are less striking, we find that the use of $\sigma \neq 1$, a feature of all available network generators, has drastic topological consequences.

Identifying the location of the Internet within this phase space does not automatically provide an Internet model valid to ultimate details, as there are several nonuniversal characteristics that contribute to the network topology. For example, several studies have established that the value of the degree exponent $\gamma$ can be tuned by changing the model parameters [13, 30, 43, 78], as the relative frequency of node and link addition and removal jointly determine $\gamma$. In order to predict the value of the degree exponent one needs to carefully measure all frequencies and include internal links, which requires time-resolved Internet maps that are currently available only at the low-resolution AS-level only. Consequently, the degree exponent can take up any value, while leaving the Internet’s position in the phase space unchanged. Similarly, the Internet displays nontrivial higher-order correlations, that could be explained by incorporating node fitness [19, 115]. Therefore, to design topology generators that reproduce simultaneously the precise numerical values and the correct functional form of the Internet’s path length and degree distribution, one needs to incorporate numerous Internet-specific details. However, our results indicate that several universal constraints influence the network’s large-scale topology. That is, no matter how detailed an Internet model is, if its universal parameters ($\alpha, \sigma, D_f$) deviate from those
uncovered by measurements, the large-scale topology will inevitably differ from the current Internet.

The advantage of the model proposed here is its flexibility: it offers an universally acceptable skeleton for potential Internet models, on which one can build features that could lead to further improvements. Using an evolving network to model the Internet has the potential to predict the future of the network, as the model incorporates only time-invariant mechanisms that should continue driving the network’s development in the future. Such predictive power, combined with elements pertaining to link bandwidths and traffic predictions [34] could offer a crucial tool to uncovering potential bottlenecks and network congestions resulting from the Internet’s rapid, decentralized development. These advances are crucial for both scientific and design purposes, being a key prerequisite for developing the next-generation communication technologies. In addition, the model introduced here offers a realistic starting point for a general class of network topologies that combine the scale-free structure with a precise spatial layout [87].
4.1 Introduction

As protein-protein interactions are central to most biological processes, the systematic identification of all protein-interactions is considered a key strategy for uncovering the inner workings of a cell. Consequently, a number of experimental and computational techniques have been developed to systematically determine both the potential and actual protein-interactions in selected model organisms, primarily in *Saccharomyces cerevisiae* [7, 10, 20, 44, 49, 54, 55, 61, 67, 70, 92, 93, 96, 140, 141, 156]. This proliferation of interest and tools resulted in extensive databases of protein-interactions, covering organisms from bacteria to eukaryotes, and fuelling research aimed at understanding the large-scale organizing principles of cellular function [76, 145].

As the interactions in which a given protein participates are likely to correlate with the protein’s functional properties, protein interaction maps are frequently utilized to uncover in a systematic fashion the potential biological role of proteins of unknown functional classification [129, 140]. Also, the topology of the uncovered protein interaction networks may reflect the cell’s higher-level functional organization. Yet, despite their clear utility, there is very little understanding to what degree
the collected protein network topologies encode such functional information [145]. For example, four different protein interaction maps are currently used, often interchangeably, to approximate the protein interaction network of yeast (*S. cerevisiae*), but the limitations and quality of the four databases remains poorly studied. Of these, two independently performed systematic two-hybrid assays provide us with maps of potential pair-wise interactions [67, 141]. In addition, two hand-curated databases, MIPS [96] and DIP [156], collate experimentally-determined protein-protein interactions from the literature, but use the results of the two-hybrid experiments as well, thus incorporating both demonstrated and potential pair-wise protein interactions.

While many interactions appear in all four databases, the disparities between the four databases are notable. For example, the overlap in the interactions identified by the two independent two-hybrid datasets [67, 141] is only between 16 to 20% [60, 67]. This limited overlap could indicate that the two-hybrid techniques cover only a small percentage of the potential interactions, or could signal a high rate of false negative and positive interactions [145].

In order to determine how well the four available databases characterize the protein interaction network of *S. cerevisiae*, in this chapter we systematically analyze the relationship between the topology of the obtained protein-interaction maps and the known functional properties of the proteins. We start by demonstrating that the four protein-interaction networks are characterized by comparable degree and cluster size distributions, indicating that they are described by the same large-scale topology that is best approximated by a hierarchically modular [124] protein-interaction network structure. We also find that despite the potentially small coverage and the ambiguity of the uncovered interactions, there are clear correlations between the known
functional classification of the proteins and the underlying large-scale topology of protein-interaction network in all four datasets. Indeed, most proteins sharing similar functional roles appear segregated on well-defined regions of the protein-interaction maps and display a high degree of network-based clustering. Similar signatures of network-based segregation are obtained when we study the impact of cellular localization on the network topology, finding that proteins sharing the same subcellular localization form relatively compact subclusters in the protein-interaction network. There are noticeable differences, however, in the degree of correlations between function, subcellular localization, and topology characterizing the different databases. The developed methods and subsequent results allow us to uncover the functional relationship between the functional classes and to provide a guide for the utility of the four databases for various bioinformatics studies of the yeast proteome.

4.2 Methods

4.2.1 Protein Interactions Data Sources

The primary focus of the chapter is the potentially and experimentally determined direct physical interactions occurring between *S. cerevisiae* proteins. The information on protein interactions are deposited in four separate databases, detailed below:

(i) The Database of Interacting Proteins or DIP [156], combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both manually and automatically, using computational approaches that utilize the knowledge about the protein-protein interaction networks. The database downloaded (on August 2001) from http://dip.doembi.ucla.edu/ contains information for 5,798 yeast proteins with at least one interaction to other proteins, connected to each other by over 20,000 interactions.

(ii) The Munich Information Center for Protein Sequences [96] or MIPS (mips.gsf.de), another hand curated database for *S. cerevisiae*, on February 2002 contained information on 6552 proteins, connected via 3,797 interactions.
(iii) The Uetz dataset [141] summarizes the results of the systematic two-hybrid assay, collecting information on 2115 proteins connected via 4480 interactions. (depts.washington.edu/sfields/)

(iv) The Ito database [67] contains the results of an independent two-hybrid assay. Currently it contains information on 3280 proteins connected via 8868 interactions. For completeness we analyzed separately the full dataset, including all detected protein interactions, as well as the higher confidence core data, a subset of the full dataset containing only interactions with more than three interaction sequence tag hits. (genome.c.kanazawa-u.ac.jp/Y2H/)

(v) Finally, http://us.expasy.org/, http://www.ncbi.nlm.nih.gov/, http://dip.doebi.ucla.edu/ and http://mips.gsf.de/ databases were used to obtain the protein name conventions necessary to compare the different datasets.

4.3 Results

4.3.1 Protein Interaction Databases

For the present study we focused on four large *S. cerevisiae* protein interaction databases: DIP [156], MIPS [96], which are hand curated databases, and the two-hybrid dataset collected by Uetz et al. [141] and Ito et al. [67]. The relationship between the four datasets is summarized in Fig. 4.1 and Table 4.1, and their particulars are detailed in the Section 4.2. As the two largest datasets, DIP and MIPS, contain interaction data for a significant fraction of the yeast proteins, there is a rather large overlap between them. In contrast, the Uetz and the Ito data sets are subsets of both hand-curated databases. Yet, as noted before [60, 67], the overlap between the Uetz and the Ito data is rather small: less than 30% of the yeast proteins in the Ito data set are found in the Uetz data, and vice versa, only about 30% of the Uetz proteins appear in the Ito database as well. At the level of identified protein interactions the differences between the four databases are even more significant. In Fig. 4.1 (b); for example, only 7% of the interactions identified by the Ito dataset
overlap with those present in the Uetz data.

4.3.2 Large-Scale Organization of the Protein Interaction Network

Understanding the large-scale organizing principles of protein-interaction networks is one of the prominent goals of post-genomic biology. Rapid advances in complex network theory in general have considerably aided this quest [3, 95]. Many studies, recently performed, have shown the relevance of scale-free degree distribution which is characterized by Eq. (1.3) to cellular network as well [48, 72, 113, 146, 155]. In particular, the protein interaction network generated by the two-hybrid approaches has also been found to have a scale-free topology [71, 146]. As scale-free networks are dominated by a few highly connected nodes, or hubs, the inhomogeneous nature of the underlying network topology has important consequences on the robustness and error tolerance of the underlying cellular networks as well [5, 71]. Yet, it is unclear if the scale-free topology is a generic feature of all four protein-network maps. To investigate the generality of the scale-free concept in Fig. 4.2 (a) we show the degree distribution for all four protein interaction maps on a log-log plot. As the fit indicates, each of the four derived networks have a power-law degree distribution, indicating that they are all described by scale-free networks with comparable degree exponent $\gamma$ (Table 4.1).

An important question underlying biological organization relates to the potential existence of modules in biological networks. Indeed, following the recent proposal of modular biology [58] a series of studies have focused on identifying the biological modules in various cellular networks, ranging from the metabolism [62, 123, 124, 128] to genetic networks [63, 67]. Modularity assumes the existence of groups of proteins
Figure 4.1. Relationship between the four studied databases (a) at the protein and (b) at the interaction level. The sum of the numbers within each color boundary denotes (a) the total number of proteins or (b) the total number of interactions found in the corresponding database.
that work together to achieve some well-defined biological function. For example, it is experimentally well established that protein complexes that act as functional modules carry out many biological functions. From the network perspective, these modules should appear as distinct group of nodes that are highly interconnected with each other, but have only a few links to nodes outside of the module. Yet, the scale-free topology apparently forbids the existence of independent modules in the network, as the hub proteins ability to interact with a high fraction of each module’s components makes a module’s relative isolation all but impossible. Recently, we proposed that the network’s scale-free topology can be reconciled with its potential modularity within the framework of a hierarchical modularity [16, 123, 124]. The most important test of such hierarchical modularity is the scaling of the clustering coefficient, $C$, defined by Eq. (1.16). As mentioned in Section 1.4.2, for the random and the scale-free model such as shown in Section 1.4 the clustering coefficient of a node with $k$ links is independent of $k$ that is, on average hubs have the same clustering coefficient as small nodes do. In contrast, for a hierarchical network the clustering coefficient $C(k)$ depends on the node’s degree as [16, 37, 124, 137]

$$C(k) \sim k^{-\delta},$$

(4.1)

where $\delta$ the modularity exponent characterizing the network’s hierarchical modularity. Therefore, the $C(k)$ function, which can be measured for arbitrary networks [123], can provide direct evidence if the network has a hierarchical modularity. To test the organization of modularity in protein interaction networks we measured the $C(k)$ function for each of the four studied protein-network databases. As Fig. 4.2 (b) shows, we find that $C(k)$ is not independent of $k$, but can be well approximated by a
power law with exponent $\delta \simeq 2$, giving direct evidence of hierarchical modularity in protein-interaction networks.

Another important property of the currently available protein-interaction networks is that they are fragmented into many distinct clusters [71, 129, 146]. We find that each of the four databases are dominated by a giant cluster that contains a significant fraction of all connected proteins, such that one can find a path of protein interactions between any two proteins belonging to this giant component. A small fraction of proteins, however, are either completely isolated (i.e., do not have any known interactions to other proteins) or form small islands of isolated groups of interconnected proteins. To characterize the fragmented nature of the protein interaction network we determined the size $n$ of each isolated cluster, and prepared a normalized histogram of the results, obtaining the cluster-size distribution. As Fig. 4.2 (c) shows, each of the datasets have a giant component of approximately $10^3$ proteins. However, the giant component coexists with many isolated proteins, somewhat fewer two-protein clusters, and even fewer three-protein clusters. If we disregard the giant component, the cluster-size distribution follows a power-law, $P(n) \sim n^{-\xi}$ where $\xi$ is the cluster size exponent, with values ranging between $\xi \simeq 3 \sim 4$. This fragmentation could indicate that the existing databases contain only a small fraction of all protein-protein interactions present in $S. cerevisiae$. Indeed, if more protein interactions are uncovered, the giant component is expected to absorb a larger fraction of all proteins, and a fully connected protein network could emerge with a single giant component. Such increase of the giant cluster is a well-known result of random graph theory [23] predicting that as the number of interactions increase in a network with a fixed number of nodes, the isolated clusters will be gradually absorbed by the giant
Figure 4.2. Large scale characteristics of the protein interaction databases. (a) Degree distribution of the four databases, shown on a log-log plot. Note that all datasets have a power-law tail, indicating that the underlying network has a scale-free topology. The solid line is obtained from the fitting to the function $P(k) \sim k^{-\gamma}$ to the DIP data, the best fit indicating $\gamma \approx 2.5$ for DIP dataset. (b) Plot of the clustering coefficient against degree on a log-log plot. The straight line has slope -2. (c) Cluster size distribution for the four databases shown on a log-log plot. Apart from the points corresponding to the giant component (for right) the $P(n)$ curves follow a power-law. The solid line is obtained from the least square fitting to $P(n) \sim n^{-\xi}$ for the MIPS dataset, providing $\xi = 3 \sim 4$.

Finally, we have found that the giant component is typically highly interconnected, resulting in a small node-to-node distance (or diameter). Indeed, the average node-to-node distance for each of the four datasets varies between 4 and 8 (Table 4.1), indicating that protein interaction networks have small-world properties.

In summary, regarding the large scale topology of protein interaction networks all four databases display the same generic properties: they are all scale-free networks forming a giant cluster accompanied by many small disconnected clusters of proteins;
### TABLE 4.1

**SUMMARY OF THE GLOBAL CHARACTERISTICS OF THE FOUR STUDIED PROTEIN-PROTEIN INTERACTION DATABASES.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( N )</th>
<th>( N_{int} )</th>
<th>( L )</th>
<th>( N_{LC} )</th>
<th>( L_{LC} )</th>
<th>( D_{LC} )</th>
<th>( C/C_{rand} )</th>
<th>( \gamma )</th>
<th>( \frac{m(1)}{m_{rand}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIPS</td>
<td>6745</td>
<td>2043 (188)</td>
<td>5434</td>
<td>1441</td>
<td>4538</td>
<td>7.71</td>
<td>34.91</td>
<td>2.34</td>
<td>3.28</td>
</tr>
<tr>
<td>DIP</td>
<td>5798</td>
<td>5798 (352)</td>
<td>20098</td>
<td>4198</td>
<td>15892</td>
<td>4.9</td>
<td>117.09</td>
<td>2.50</td>
<td>3.48</td>
</tr>
<tr>
<td>Uetz</td>
<td>2115</td>
<td>1870 (74)</td>
<td>4480</td>
<td>1458</td>
<td>3941</td>
<td>6.8</td>
<td>54.64</td>
<td>2.32</td>
<td>2.28</td>
</tr>
<tr>
<td>Ito</td>
<td>3280</td>
<td>3280 (82)</td>
<td>8868</td>
<td>2840</td>
<td>8371</td>
<td>4.9</td>
<td>36.40</td>
<td>2.44</td>
<td>1.49</td>
</tr>
<tr>
<td>Ito core</td>
<td>797</td>
<td>797 (52)</td>
<td>1560</td>
<td>417</td>
<td>1055</td>
<td>6.2</td>
<td>4.94</td>
<td>2.1</td>
<td>7.06</td>
</tr>
</tbody>
</table>

The second column (\( N \)) denotes the total number of proteins in the entire dataset while the third column (\( N_{int} \)) represents the number of proteins which appear in the protein interaction network. In the parenthesis we show the number of proteins which have self-interactions. The fourth column (\( L \)) shows the number of interactions between the proteins. For the largest cluster (LC) data \( N_{LC} \) denotes the number of proteins in the largest cluster, and \( L_{LC} \) represents the number of links in the largest cluster. The diameter (\( D_{LC} \)) denotes the average node-to-node distance for the proteins in the largest cluster being shown in the seventh column. The clustering coefficient is normalized with the clustering coefficient of the random network (\( C/C_{rand} \): for details see text). The degree exponents (\( \gamma \)) are obtained from the relation \( P(k) \sim k^{\gamma} \) and the average segregation parameter reflects the tendency of the proteins to be primarily connected to proteins that belong to the same functional class.
they display a high degree of modularity with a hierarchical organization; and the giant cluster has a small diameter, an indication of its small-world property. As these properties are derived from all four databases, they appear to be generic features of the yeast protein-interaction network.

4.3.3 Correlations Between Topology and Functional Organization

To correlate the topological and functional properties of the derived protein-interaction networks we utilize the functional classification established by the MIPS database, in which each protein is assigned to one or several of 14 functional classes, based on functional information reported in the literature (Table 4.2). While a classification scheme into 44 functional classes is also available (www.proteome.com), our choice for the 14-class classification system was motivated by statistical purposes: many functional classes in the 44 class breakdown contain too few proteins to allow us to systematically analyze their segregation and clustering properties.

**TABLE 4.2**

**FUNCTIONAL CLASSES BASED ON THE MIPS DATABASE**

<table>
<thead>
<tr>
<th>ID</th>
<th>Function Name</th>
<th>( N_i )</th>
<th>Uetz</th>
<th>MIPS</th>
<th>DIP</th>
<th>Ito</th>
<th>Ito core</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Metabolism</td>
<td></td>
<td>324</td>
<td>1065</td>
<td>605</td>
<td>541</td>
<td>110</td>
</tr>
<tr>
<td>1</td>
<td>Energy</td>
<td></td>
<td>72</td>
<td>252</td>
<td>140</td>
<td>131</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>Cell growth, cell division, and DNA synthesis</td>
<td></td>
<td>485</td>
<td>836</td>
<td>586</td>
<td>435</td>
<td>125</td>
</tr>
<tr>
<td>3</td>
<td>Transcription</td>
<td></td>
<td>404</td>
<td>793</td>
<td>534</td>
<td>416</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>Protein synthesis</td>
<td></td>
<td>88</td>
<td>359</td>
<td>152</td>
<td>160</td>
<td>23</td>
</tr>
</tbody>
</table>
TABLE 4.2 (CONTINUED)

<table>
<thead>
<tr>
<th>ID</th>
<th>Function Name</th>
<th>$N_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Uetz</td>
</tr>
<tr>
<td>5</td>
<td>Protein destination</td>
<td>278</td>
</tr>
<tr>
<td>6</td>
<td>Transport facilitation</td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>Cellular transport and transport mechanisms</td>
<td>237</td>
</tr>
<tr>
<td>8</td>
<td>Cellular biogenesis</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>Cellular communication/signal transduction</td>
<td>83</td>
</tr>
<tr>
<td>10</td>
<td>Cell rescue, defense, cell death, and ageing</td>
<td>156</td>
</tr>
<tr>
<td>11</td>
<td>Ionic homeostasis</td>
<td>30</td>
</tr>
<tr>
<td>12</td>
<td>Cellular organization</td>
<td>1006</td>
</tr>
<tr>
<td>13</td>
<td>Classification not yet clear-cut</td>
<td>39</td>
</tr>
<tr>
<td>14</td>
<td>Unclassified proteins</td>
<td>489</td>
</tr>
<tr>
<td>*</td>
<td>Transposable elements, viral and plasmid proteins</td>
<td>2</td>
</tr>
</tbody>
</table>

We start from the hypothesis that proteins belonging to the same functional class have a strong tendency of working together, and thus potentially have a high number of connections between each other. If this were true, we expect the topology of the protein-interaction network to be segregated into different functional classes, such
that a given protein interacts predominantly with proteins belonging to the same functional class, and only to a lesser degree with proteins belonging to other functional classes. To investigate the validity of this hypothesis for each protein $i$ that belongs to functional class $\lambda$ we define the segregation function, $m_\lambda^i(d)$, as

$$m_\lambda^i(d) = \frac{M_\lambda^i(d)}{M_i(d)}, \quad (4.2)$$

where $M_\lambda^i(d)$ denotes the number of proteins at distance $d$ from protein $i$ that belong to the functional class $\lambda$ and $M_i(d)$ denotes the total number of proteins at distance $d$ from protein $i$. Here the distance between two proteins is defined by the number of links along the shortest path between given two proteins.

As illustrated in Fig. 4.3 (a), a protein with links only to nodes in the same functional class has $m_i(1) = 1$ Fig. 4.3 (b), while one that does not have links to any protein of the same functional class has $m_i(1) = 0$. As the topology of the network around a single protein is statistically not representative, it is useful to define $m^{\lambda}(d)$ as the average of $m_\lambda^i(d)$ over all proteins $i$ that belong to the same functional class $\lambda$. Therefore, $m^{\lambda}(d)$ offers a measure of the degree of segregation for the functional class $\lambda$. If proteins belonging to a given functional class $\lambda$ were randomly distributed in the network, then $m^{\lambda}(d)$ should be independent of the distance $d$, equal to $m^{\lambda}_{rand}$, where $m^{\lambda}_{rand}$ the average density of proteins belong to the functional class $\lambda$, given by $m^{\lambda}_{rand} = N^\lambda / N$ where $N^\lambda$ denotes the number of proteins that belong to the functional class $\lambda$ and $N$ is the total number of proteins in the protein network. In contrast, if proteins belonging to the functional class $\lambda$ have a tendency to cluster together, we expect the associated $m^{\lambda}(d)$ function to monotonically decrease, converging for large $d$ to $m^{\lambda}_{rand}$.
Figure 4.3. Functional segregation. (a), (b) Schematic illustration of the topological interpretation of the segregation parameter $m(1)$. For example, for protein 1, focusing only on interaction to proteins belonging to the same functional class (shown in light color), we have $k = 4$, $C_1 = 1/3$, $m(1) = 3/4$. Focusing on proteins that are $d = 2$ distance from protein 1, we have $m(2) = 0$. In (b) we reorganized randomly the same number of proteins and kinks. In this random configuration we find $C_1 = 0$, $m(1) = 2/4 = 1/2$, and $m(2) = 1/2$. (c)-(f) The relative segregation functions for the four databases. Each curve corresponds to a different functional class, the numbers and the corresponding functional classes being listed in Table 4.2. The four panels describe the different dataset, i.e., (c) Uetz, (d) Ito complete, (e) MIPS, and (f) DIP.
The $m^\lambda(d)$ curves obtained for each of the 14 functional classes are shown in Fig. 4.3 (c)-(f) for the four protein-interaction networks. As the number of proteins differ between the functional classes, one expects large, functionally irrelevant variations in $m^\lambda(d)$. To offset these variations, in Fig. 4.3 (c)-(f) we plot the relative segregation function $m^\lambda(d)/m^\lambda_{rand}$ all four datasets. The ratio $m^\lambda(d)/m^\lambda_{rand} > 1$ if the proteins belonging $\lambda$ display measurable topological segregation. For most functional classes we observe that $m^\lambda(d)/m^\lambda_{rand} \gg 1$ small $d$, and decreases rapidly with $d$, reaching the asymptotic limit $m^\lambda(d)/m^\lambda_{rand} \sim 1$ for $d \geq 3 \sim 4$. This indicates that most functional classes display some degree of topological localization within the protein interaction network, i.e., the immediate neighbors of a given protein belong with high probability to the same functional class. For some functional classes the segregation function for small $d$ is over 10, implying that the proteins belonging to this class are 10 times more likely to have neighbors that belong to the same functional class than proteins randomly placed in the network. For example, this high degree of segregation is seen for proteins contributing to transport facilitation (#6 in Table 4.2) in the Uetz data, or cellular communication and signal transduction (#9) in the MIPS database.

To compare directly the four protein interaction networks in Fig. 4.4 (a) we plot $m^\lambda(d)/m^\lambda_{rand}$ for each of the four datasets and the 14 functional classes. The results obtained for the four datasets correlate with each other: a high degree of functional segregation of one dataset is typically reflected as some degree of segregation in the other datasets as well. We observe a high degree of segregation for the Uetz, MIPS, DIP and the core Ito data. In contrast, with a few exceptions, the extended Ito dataset displays a smaller degree of functional segregation, while the DIP and the core Ito datasets have the highest $m^\lambda(d)/m^\lambda_{rand}$ coefficient for most functional classes.
Figure 4.4. Characterization of the segregation properties of proteins classified based on their functional class (a)-(c) or subcellular localization (d)-(f). In (a)-(c) on the horizontal axis we show the number corresponding to their various functional classes described in Table 2. (a) $m(1)/m_{rand}(1)$ ratio for each functional class, shown separately for each of the five studied databases. (b) Relative clustering coefficient $C/C_{rand}$ for each functional class, shown separately for all five databases. (c) Relative number of links between proteins belonging to the same functional class, $L/L_{rand}$, for each of 14 functional classes, shown separately for the five databases. (d)-(f) The same quantities as in (a)-(c) but characterizing proteins sharing the same subcellular localization. The horizontal axis, therefore, denotes the cellular localization classes listed in Table 4.3. The color code is the same as in (a), the plots showing separately the data for each localization class (horizontal axis) and each database.
If nodes belonging to a given functional class form cohesive groups within the protein interaction network, they should display a high degree of clustering. The degree of clustering of a complex network is often characterized by the clustering coefficient, $C$, discussed above. To determine the degree of clustering for each functional class we restricted the network to the nodes belonging to a given functional class and direct links between them, and measured the average clustering coefficient for the obtained functional sub-graph. As often this sub-graph is rather fragmented (particularly for functional classes with smaller number of nodes), the value of the clustering coefficient by itself is not particularly revealing. To obtain a meaningful measure, we calculate the relative clustering coefficient, $c^\lambda = C^\lambda / C^\lambda_{\text{rand}}$ for each functional class Fig. 4.4 (b) where, to determine $C^\lambda_{\text{rand}}$ we randomly distribute on the network $N^\lambda$ proteins (i.e., assign randomly chosen proteins to the functional class $\lambda$, without altering the network topology), and measure $C^\lambda_{\text{rand}}$ for the obtained random subnetwork $\lambda$. One can notice the high degree of correlation between the results obtained for all datasets. The pattern seen in Fig. 4.4 (a) is evident here, as well: the degree of clustering observed for the MIPS, DIP and Uetz datasets is very high. Overall the MIPS database has the highest relative clustering coefficient for most functional classes.

Finally, another measure of a network’s functional segregation can be obtained by determining the number of direct links between proteins that belong to the same functional class. Let us consider an arbitrary functional class $\lambda$, and denote by $L^\lambda$ the number of direct links between proteins that belong to $\lambda$. To obtain a meaningful measure of the topological cohesiveness of functional class $\lambda$, we calculate the ratio $\ell^\lambda \equiv L^\lambda / L^\lambda_{\text{rand}}$ where $L^\lambda_{\text{rand}}$ is the number of direct links between proteins of functional class $\lambda$ if the proteins of $\lambda$ are placed randomly on the network, without altering...
the network’s topology. A ratio \( \ell = 1 \) implies that the proteins belonging to \( \lambda \) are randomly distributed in the network. A ratio of 10, however, indicates that there are 10 times more internal links within the functional class \( \lambda \) than expected for a random protein distribution. The results again indicate large deviations from a random distribution for the DIP, MIPS, Uetz and core Ito datasets, and weak segregation for the complete Ito data.

The combination of the results of Fig. 4.3 offer a rather detailed characterization of each of the four protein-interaction networks and allow us to uncover systematic differences between the different functional classes. For example, we have found that proteins responsible for cellular communication and signal transduction (#9, Table 4.2) show a very high segregation parameter in the MIPS database, indicating that the neighbors of a protein contributing to cellular communication are 26 times more likely to belong to the same functional class, and they interact only with such proteins. This finding is corroborated by Fig. 4.4 (c) as well. The clustering coefficient of this class is not remarkable, however, see Fig. 4.4 (b). Therefore, the proteins belonging to this functional class mostly interact with each other, but they form a loose collection of nodes, with a small degree of clustering. In contrast, the functional class responsible for cellular organization (#12, Table 4.2) has a very high clustering coefficient in all databases, the corresponding proteins forming strongly interconnected clusters. It does not have an unusually high segregation parameter, however, indicating that in addition to the direct links within the same functional class, cellular organization proteins also interact with a large number of proteins from other functional classes.
4.3.4 Correlation Between Topology and Cellular Localization

Depending on the functional role they play, proteins are often localized in spatially distinct areas of the cell. This spatial compartmentalization is particularly prominent for eukaryotes, and is expected to leave its mark on the topology of the protein-interaction network as well: a protein localized in the nucleus is more likely to interact with another nuclear protein than with those localized at the cell wall. To investigate the correlation between cellular localization and the protein-network topology, we assigned each protein its cellular location based on a 28 sub-cellular localization classes (Table 4.3) obtained from the Proteome database (www.proteome.com). In order to characterize the correlations between the topology of protein interaction networks and the known sub-cellular localization properties of the yeast proteome, we used the quantities developed earlier, measuring for each localization class \( \lambda \) the function \( m^\lambda(1)/m^\lambda_{\text{rand}} \), \( C^\lambda(1)/C^\lambda_{\text{rand}} \), and \( L^\lambda(1)/L^\lambda_{\text{rand}} \). As Fig. 4.4 demonstrates, our measurements indicate that most localization classes appear segregated in the protein interaction network. In particular, proteins belonging to a few classes, such as those localized in the mitochondrial matrix or outer membrane, or nuclear pore, show an over 100 times increase in their localization coefficient compared to the randomly distributed reference set. In addition, we observe correlations between the degree of localization observed in the four studied datasets. In this case, however, the MIPS database stands out, as it shows a higher degree of segregation than any of the other databases. Most importantly, the fact that the segregation and clustering parameters are significantly higher than one for most functional classes indicates that the topology of the protein-interaction network reflects, to a considerable degree, the cell’s physical compartmentalization.
Some interesting cases are observed in this case as well. Proteins localized in the mitochondrial outer membrane (#14, Table 4.3) display a very high degree of segregation ($m(1)/m_{rand} \approx 250, 150, 40$ for MIPS, Ito respectively), yet the $L/L_{rand}$ parameter is not particularly high and the clustering coefficient of this class is not remarkable either. Therefore, while the proteins of this class interact predominantly with each other, they do not form a highly interconnected cluster.

### TABLE 4.2

CLASSIFICATION OF THE S. CEREVISIAE PROTEINS IN CELLULAR LOCALIZATION CLASSES, BASED ON THE PROTEOME (WWW.PROTEOME.COM) DATABASE:

<table>
<thead>
<tr>
<th>ID</th>
<th>Function Name</th>
<th>$N_i$</th>
<th></th>
<th></th>
<th>Ito</th>
<th>Ito core</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Uetz</td>
<td>MIPS</td>
<td>DIP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Bud neck</td>
<td>37</td>
<td>28</td>
<td>53</td>
<td>39</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>Cell Wall</td>
<td>13</td>
<td>39</td>
<td>68</td>
<td>34</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Centrosome/spindle pole body</td>
<td>54</td>
<td>44</td>
<td>70</td>
<td>49</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Cytoplasmic</td>
<td>298</td>
<td>385</td>
<td>747</td>
<td>425</td>
<td>102</td>
</tr>
<tr>
<td>4</td>
<td>Cytoskeletal</td>
<td>83</td>
<td>50</td>
<td>100</td>
<td>79</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Endoplasmic reticulum</td>
<td>101</td>
<td>119</td>
<td>225</td>
<td>136</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>Endoplasmic/endosomal vesicles</td>
<td>20</td>
<td>17</td>
<td>36</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Extracellular (excluding cell wall)</td>
<td>5</td>
<td>13</td>
<td>24</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Golgi</td>
<td>50</td>
<td>43</td>
<td>93</td>
<td>48</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>Lysosome/vacuole</td>
<td>30</td>
<td>49</td>
<td>90</td>
<td>54</td>
<td>15</td>
</tr>
</tbody>
</table>
# TABLE 4.3 (CONTINUED)

<table>
<thead>
<tr>
<th>ID</th>
<th>Function Name</th>
<th>$N_i$</th>
<th>Uetz</th>
<th>MIPS</th>
<th>DIP</th>
<th>Ito</th>
<th>Ito core</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Microsomal fraction</td>
<td>7</td>
<td>13</td>
<td>19</td>
<td>12</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Mitochondrial</td>
<td>95</td>
<td>225</td>
<td>442</td>
<td>238</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Mitochondrial inner membrane</td>
<td>30</td>
<td>72</td>
<td>146</td>
<td>76</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Mitochondrial matrix</td>
<td>14</td>
<td>37</td>
<td>68</td>
<td>39</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Mitochondrial outer membrane</td>
<td>13</td>
<td>16</td>
<td>30</td>
<td>17</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Nuclear</td>
<td>590</td>
<td>598</td>
<td>1123</td>
<td>781</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Nuclear nucleolus</td>
<td>36</td>
<td>67</td>
<td>132</td>
<td>60</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Nuclear pore</td>
<td>36</td>
<td>31</td>
<td>54</td>
<td>46</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Other vesicles of the secretory/endocytic pathways</td>
<td>34</td>
<td>37</td>
<td>65</td>
<td>41</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Peroxisome</td>
<td>16</td>
<td>30</td>
<td>49</td>
<td>31</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Plasmid membrane</td>
<td>70</td>
<td>115</td>
<td>234</td>
<td>134</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Unspecified membrane</td>
<td>62</td>
<td>123</td>
<td>275</td>
<td>103</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell ends</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contractile ring</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lipid particles</td>
<td>0</td>
<td>10</td>
<td>13</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mitochondrial intermembrane space</td>
<td>3</td>
<td>7</td>
<td>13</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nuclear matrix</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nuclear transport factor</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 4.3 (CONTINUED)

<table>
<thead>
<tr>
<th>ID</th>
<th>Function Name</th>
<th>$N_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Uetz</td>
</tr>
<tr>
<td>*</td>
<td>Secretory vesicles</td>
<td>9</td>
</tr>
</tbody>
</table>

#### 4.3.5 Relationship Between Functional and Localization Classes

The segregation of the various functional classes in separate regions of the protein-interaction network inspires a new question: how do these functional classes relate to each other? That is, knowing the overall topology of the protein-interaction network, can we establish the relationship between the different cellular functions? If the proteins were to interact *only* with proteins belonging to the same functional class, the protein-interaction network should be broken into islands corresponding to the different functional classes. This is not the case however, as there are a considerable number of interactions between proteins belonging to different functional classes [129].

The number of links between proteins of two functional classes offers a measure to what degree proteins from two functional classes may act together within functional modules, thus our goal is to use this measure to derive a global map of potential functional relationships within the yeast proteome.

To determine the degree to which proteins of functional class $\lambda$ are related to
proteins of class $\phi$, we measure the $\ell(\lambda, \phi)$ coefficient, defined as

$$\ell(\lambda, \phi) = \frac{L^{\lambda,\phi} + L^{\phi,\lambda}}{L^\lambda + L^\phi},$$

where $L^{\lambda,\phi}$ is the total number of links that proteins of class $\lambda$ have to protein members belonging to functional class $\phi$ and where $L^\lambda$ ($L^\phi$) is the total number of links between the proteins of functional class $\lambda$ ($\phi$). The $(L^{\lambda,\phi}/L^\lambda)/(L^{\lambda,\phi}/L^\lambda_{\text{rand}})$ matrices obtained for the four protein interaction networks are shown in Fig. 4.5-Fig.4.8. They indicate that the relationships between the different classes vary widely: we observe strong ties between some functional classes, while others appear only weakly related.

To uncover the relationship between these functional classes we applied a minimum linkage clustering algorithm [68] using the quantity $1/\ell(\lambda, \phi)$ as the distance metric between classes $\lambda$ and $\phi$. The algorithm places close to each other the functional classes that are topologically closely related. A hierarchical tree, generated by the clustering process, summarizes the relationship between the different functional classes. At a first glance it is evident that the hierarchical trees obtained for the four databases agree on some generic features of the cell’s internal organization. Indeed, all databases indicate that protein function for the classes #12 (cellular organization) and #3 (transcription) belong to the more connected core of the network, closely traced by proteins belonging to the classes #2 (cell growth, cell division and DNA synthesis) #7 (cellular transport and transport mechanisms), and #5 (protein destination). The rest of the functional classes surround this core in an onion-like fashion. All four databases agree regarding the two classes that show the smallest degree of interaction with any other class (and thus are delegated to the outer branches of the hierarchical tree): these are the proteins found in the classification not yet clear-cut.
Figure 4.5. $\frac{L^{\lambda,\phi}}{L^{\lambda}} / \frac{L^{\lambda,\phi}}{L^{\lambda}}$ matrix, representing the relationship between the different functional classes based on Uetz. The higher value of $\frac{L^{\lambda,\phi}}{L^{\lambda}} / \frac{L^{\lambda,\phi}}{L^{\lambda}}$ coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes.
Figure 4.6. \((L_{\phi}^{\lambda}/L^{\lambda})/(L_{\text{rand}}^{\lambda}/L_{\text{rand}}^{\lambda})\) matrix, representing the relationship between the different functional classes based on Ito. The higher value of \((L_{\phi}^{\lambda}/L^{\lambda})/(L_{\text{rand}}^{\lambda}/L_{\text{rand}}^{\lambda})\) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes.
Figure 4.7. \( \frac{L^{λφ}}{L^λ} / \frac{L_{\text{rand}}^{λφ}}{L_{\text{rand}}^λ} \) matrix, representing the relationship between the different functional classes based on DIP. The higher value of \( \frac{L^{λφ}}{L^λ} / \frac{L_{\text{rand}}^{λφ}}{L_{\text{rand}}^λ} \) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes.
Figure 4.8. \( \frac{L^{\lambda,\phi}}{L^{\lambda}} \)/\( \frac{L^{\lambda,\phi}_{\text{rand}}}{L^{\lambda}_{\text{rand}}} \) matrix, representing the relationship between the different functional classes based on MIPS. The higher value of \( \frac{L^{\lambda,\phi}}{L^{\lambda}} \)/\( \frac{L^{\lambda,\phi}_{\text{rand}}}{L^{\lambda}_{\text{rand}}} \) coefficient, the more interactions are detected between the two functional classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes.
and protein synthesis (#4).

We can perform a similar clustering based on the overlap between the proteins belonging to different cellular localization. For this we measure again the $\ell(\lambda, \phi)$ parameters defined above, but here $\lambda$ and $\phi$ denote different localization classes (see Table 4.3). The results, summarized in Fig.4.9-Fig.4.12, indicate again a relative agreement between the relationships predicted by the four databases. First, the protein-interaction network appears to be organized around nuclear proteins (#15), which interact closely with mitochondrial outer-membrane proteins (#14) in the Uetz, Ito and MIPS databases, and rather closely in the DIP data as well. The two other localization classes that are always found in the vicinity of these two core classes include cytoplasmic (#3) and nuclear pore proteins (#17). The hierarchical trees are also consistent regarding the protein groups that are far from the core: extra cellular or microsomal fraction proteins (#7 and 10) are clustered together and are far from the rest of the functional classes in most datasets.

4.4 Conclusion and Discussion

Uncovering the large-scale properties of protein interaction networks potentially offers an increased understanding of the system level properties of living organisms. Two questions are of primary importance from this perspective (a) understanding the network’s large-scale organization and (b) understanding how do these large-scale properties reflect the functional properties of the cellular compartments. The increasingly extensive protein-interaction databases, together with the functional annotation of the different proteins, allow us to address these questions in a systematic manner. In the following we briefly summarize our findings and discuss their implications on our
Figure 4.9. \(\frac{L^{\lambda \phi}/L^{\lambda}}{L_{\text{rand}}^{\lambda \phi}/L_{\text{rand}}^{\lambda}}\) matrix, representing the relationship between the different functional classes based on Uetz. The higher value of \(\frac{L^{\lambda \phi}/L^{\lambda}}{L_{\text{rand}}^{\lambda \phi}/L_{\text{rand}}^{\lambda}}\) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.
Figure 4.10. \((L^\lambda /L^\lambda)/(L^\lambda_{\text{rand}}/L^\lambda_{\text{rand}})\) matrix, representing the relationship between the different functional classes based on Ito. The higher value of \((L^\lambda /L^\lambda)/(L^\lambda_{\text{rand}}/L^\lambda_{\text{rand}})\) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.
Figure 4.11. \( \frac{L^\lambda\phi}{L^\lambda} / \frac{L^\lambda_{\text{rand}}}{L^\lambda_{\text{rand}}} \) matrix, representing the relationship between the different functional classes based on DIP. The higher value of \( \frac{L^\lambda\phi}{L^\lambda} / \frac{L^\lambda_{\text{rand}}}{L^\lambda_{\text{rand}}} \) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.
Figure 4.12. \( (L^{\lambda,\phi}/L^{\lambda})/(L_{\rm rand}^{\lambda,\phi}/L_{\rm rand}^{\lambda}) \) matrix, representing the relationship between the different functional classes based on MIPS. The higher value of \( (L^{\lambda,\phi}/L^{\lambda})/(L_{\rm rand}^{\lambda,\phi}/L_{\rm rand}^{\lambda}) \) coefficient, the more interactions are detected between the two localization classes. The vertical and horizontal axes represent each protein class. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.
ability to use these databases for various bioinformatics purposes.

4.4.1 Large-Scale Organization

Our results offer convincing evidence that networks deduced from the four protein interaction databases have the same large-scale topology. Indeed, each database generates a scale-free network, with embedded hierarchical modularity. We find that the scaling exponents characterizing both the degree distribution $P(k)$ and the modularity distribution $C(k)$ are comparable. As each of the four databases is incomplete, we need to ask if a more complete dataset would change these conclusions. The extensive studies on scale-free networks indicate that this is unlikely [3, 95]: if the underlying network is scale-free, a restricted network, obtained by randomly sampling the links of the scale-free network, will also stay scale-free. In contrast, it is impossible to obtain a scale-free network from the incomplete but random sampling of a network that does not have a power-law degree distribution.

Several investigators have proposed that the observed scale-free nature of the protein-interaction map is the result of gene-duplication, a process frequently occurring during evolution [31, 120, 132, 143]. Each gene duplication event leads to a new protein that interacts with the same proteins as the protein product of the original duplicated gene. Proteins that have a large number of links to other proteins are more likely to be connected to a duplicating gene, therefore they will be more likely to gain new interactions to the newly created protein. This subtle effect leads to both growth (after each gene duplication the network has an additional node, thus the network expands) and preferential attachment (highly connected proteins increase their number of interactions faster than their less connected counterparts, as they are more
likely connected to a randomly duplicating protein), the two necessary ingredients for
the appearance of a scale-free network [13, 14]. While the microscopic parameters
required to predict the precise value of the scaling exponents are still unknown, gene
duplication does offer the conceptual framework to understand the origin of the scale-
free behavior observed in protein interaction networks. The presence of the power-law
degree distribution in all four databases (Fig. 4.2 (a)) supports the expectation that
the scale-free topology is a generic feature of the protein-interaction network.

The fragmentation of the network into separate, isolated clusters, however, are
much more sensitive to potential data incompleteness. Recent models addressing
the potential origin of the scale-free topology in protein-interaction networks indicate
that the observed fragmentation could be an intrinsic property of the evolutionary
processes leading to the protein-interaction networks [74]. Indeed, the divergence of
the duplicated protein sequences by mutations could lead to the loss of interactions
between a protein and its interaction partners. If an isolated protein is duplicated,
several subsequent duplication events could lead to the emergence of an isolated clus-
ter of proteins. The analytical results indicate that the network emerging as a result
of gene duplication and loss of interactions due to mutations can develop a power-law
cluster-size distribution, whose exponent depends on the rate at which proteins add
links to other proteins during evolution [74]. Thus the power-law cluster size distri-
bution seen in all four databases could be another consequence of gene duplication
and divergence [74]. In the absence of a precise knowledge of gene duplication and
divergence rates it is impossible to predict whether the final network should be fully
connected or fragmented.
4.4.2 Function and Cellular Localization

The three quantities we have introduced to quantify the relationship between topology and function/localization allowed us to compare the segregation properties of the four protein interaction networks. The results indicate that the four databases show different strengths in different functional or localization classes. Despite these differences, the two hand-curated databases, DIP and MIPS, display a higher degree of correlation between the network structure and functional/localization based classification than the two two-hybrid datasets. Overall, the MIPS dataset shows the highest degree of functional localization in most functional classes, but the DIP dataset often offers a higher degree of cellular localization-based segregation. The weakest correlation between topology and functional and localization based classification are observed in the complete Ito dataset, but the core Ito data displays correlation comparable to that observed in the Uetz dataset. These results indicate that hand-curated databases not only offer a higher number of interactions, but the structure of the protein-interaction network reflects better the functional and localization features of the proteins. Therefore, these databases offer a better starting point for bioinformatics studies.

Naturally, the studied databases represent our current knowledge about protein interactions. While *S. cerevisiae* represents one of the most studied organism, the differences between the functional and localization based characteristics of the four studied databases offer a glimpse how incomplete these databases are. This incompleteness comes from two sources: the absence of many potential interactions, and the presence of false positives. While new research continues to add new interactions to these databases, thus gradually addressing the issue of data incompleteness, the
presence of false positives will be much harder to eliminate. In addition, the biological literature and the curating efforts tend to focus on scientifically and commercially more interesting subsets of proteins, such as e.g. signaling pathways. Therefore, certain functional subclasses are better mapped than others and even within a given class some proteins are better characterized than others. It is also a challenge to integrate into these databases the results of the interactions generated by mass spectroscopic studies on protein complexes [54, 61]. Indeed, two-hybrid measurements offer information only on pair-wise interactions. If two proteins cannot bind together without the help of a third protein, two hybrid datasets will likely not indicate a potential interaction between them. Recently two groups have provided information on the composition of hundreds of protein complexes under a given growth condition [54, 61]. Each of the components of a given complex is therefore a potential interaction partner. Including the complex information as pair-wise interactions in the databases, however, is misleading, as being part of the same complex is not sufficient to establish a pair-wise interaction. Yet, the emergence of these new datasets could help to strengthen the validity of known interactions, and offer the hope that with time we will acquire a quite complete map of protein interactions in such simple organism as yeast, serving as a starting point for a better understanding of its functional architecture.
CHAPTER 5

ROLE OF THE CYTOSKELETON IN SIGNALLING NETWORKS

5.1 Introduction

In Chapter 4 we have shown that the large-scale properties of protein-interaction networks for different databases share common properties. In this chapter we will discuss the clustering tendency of the interactions in a more specific case. The normal functioning of a cell requires constant interaction with its extracellular environment and with other cells, and these interactions lead to changes in cell physiology, cell shape, and gene expression. Signals from neighboring cells and the extracellular matrix are perceived by membrane-bound receptors, resulting in changes in their biochemical or physical states which typically initiate a cascade of signalling events within the cell [118, 127]. Intracellular signal transduction may involve physical processes (e.g., diffusion), chemical changes (e.g., phosphorylation) of signalling intermediates, or both. For most characterized signal transduction pathways, the initial signalling event and the end point are known, but intermediate events that transmit the signal are either partially or completely unknown. In order to fully understand intracellular signal transduction, it is essential to know intermediate signalling molecules and to understand how information flows from one to the next. These issues have been difficult to address experimentally because signalling molecules typically bind each other
transiently and with relatively low affinities.

The cytoskeleton, an interconnected assembly of actin, intermediate filament and microtubule networks that extend throughout the entire cell, has been implicated in intracellular signal transduction [26, 50, 51, 57, 64, 65, 69, 122, 131]. Experimental evidence indicates that individual filaments of the cytoskeleton can transmit mechanical perturbations, be used as tracks to move organelles within the cell, and to provide transient docking sites for proteins and lipids [66, 69, 99]. However, most of the evidence regarding the role of the cytoskeleton in signal transduction has been inferred from experiments that employed destructive perturbations to the cytoskeleton, such as those caused by drugs that depolymerize filaments. These manipulations cause a complete loss of one or more cytoskeletal elements, leading to global changes that complicate the interpretation of experiments.

Recent progress in proteomics offers the possibility to quantitatively address the role of the cytoskeleton in intracellular signalling. Analysis of protein interactions on the scale of entire proteomes by yeast two hybrid interaction screening and protein purification has generated a huge amount of information regarding protein networks within the cell. So far, these large scale experimental approaches have been applied most extensively to the budding yeast, *S. cerevisiae* [7, 10, 20, 44, 49, 54, 55, 61, 67, 70, 92, 93, 96, 140, 141, 156].

In this chapter, we have developed several independent, quantitative methods to probe for correlations between functionally defined protein classes based on the concepts of complex network. Specifically, we have tested the hypothesis that the network of interacting cytoskeletal proteins and the network of signalling proteins are segregated into subnetwork to a higher degree than are other functionally defined
classes of proteins. We have found that the correlation of signalling proteins with cytoskeletal proteins is much stronger than with 15 other protein classes examined. These results strongly suggest that without the cytoskeleton the cell’s intracellular signalling apparatus cannot properly function.

5.2 Method
5.2.1 Quantitative Analysis

In order to quantitatively study the clustering tendency of proteins in the various subclasses we employed several approaches in this chapter. For global characterization of clustering we analyzed the distance distribution $P(d_{ij})$ for all possible combinations of proteins. The distance between two proteins can be defined by the number of links along the shortest path between given two proteins. And to characterize the local structure of interaction networks we have extended the definition of segregation function in Eq. (4.2), and define the local segregation function for different protein classes, $m_d(\phi/\lambda)$ which counts all those proteins (denoted by $\lambda$) that are at distance $d$ from a given protein (denoted by $\phi$). Here $\phi$ and $\lambda$ stand for the various protein classes: for example $c = \text{[cytoskeletal protein]}$, $s = \text{[signalling protein]}$, $r = \text{[a protein that is not in the c or s classes]}$. By its definition, $m_d(\phi/\lambda)$ contains information on the number of neighbors of proteins which are belong to class $\lambda$ at distance $d$ from a given protein which belongs to class $\phi$. The $d = 1$ or nearest neighbors of a given protein are those proteins that directly interact with that protein. The nearest neighbor clustering index, $m_1(c/s)$, for a selected cytoskeletal protein $c$ is then calculated as

$$m_1(c/s) = \frac{\text{number of those nearest neighbors of } c \text{ that are } s \text{ proteins}}{\text{total number of nearest neighbors of } c}.$$ (5.1)
For a given protein, this metric measures the proportion of interactions to other proteins in a given class. Thus, we expect that if two given protein classes have strong tendency to interact each other, then the local segregation function $m_d(\phi/\lambda)$ would have large value for small $d$ and decay fast as increasing $d$ and finally goes to some asymptotic value which can be defined by average density of proteins in a given class.

5.3 Results

5.3.1 Definitions of Signalling and Cytoskeletal Proteins

In order to construct the signalling ($s$) and cytoskeletal ($c$) protein sets, we categorized each of the gene products of *S. cerevisiae* as a component of a signalling pathway, a cytoskeletal component, or neither (the random, or $r$ set). The rules used to define these sets were based on experimentally determined biochemical or genetic features of each protein, without reference to the databases that constitute the available interaction maps. As *S. cerevisiae* does not have intermediate filaments, the composition of the cytoskeleton was defined as actin, tubulin, proteins that bind actin or tubulin, proteins that bind a protein that binds actin or tubulin, and the septins, leading to the identification of 125 cytoskeletal proteins, 2.2% of the yeast proteome. This definition includes the filamentous septin, actin and tubulin networks (including known cross-linkers, capping, severing, etc. proteins), as well as most proteins that localize to actin patches, which underlie the plasma membrane and are prominent components of the yeast cytoskeleton. The set of signalling proteins included all protein and lipid kinases, phosphatases, GTPases and their auxiliary factors, heterotrimeric G-protein-linked membrane receptors, nucleotide cyclases/phosphodiesterases, and
biochemically or genetically characterized scaffolding proteins. This analysis identified 342 signalling proteins, 5.9% of the proteome. Twenty proteins were common to both sets. Importantly, the criteria used to define cytoskeletal and signalling proteins are conservative and independent of each other. Several metabolic kinases known to bind directly to the cytoskeleton (e.g., phosphofructokinase) were not included in the cytoskeleton protein set because their inclusion might obscure the more subtle interplay between the cytoskeleton and other signalling pathways. In addition, uncharacterized open reading frames with homology to known signal transduction proteins were excluded. These definitions, therefore, focused the analysis on proteins for which functional information is currently available.

5.3.2 Global Clustering

In the currently available protein interaction databases, information was available for subsets of the proteins in our defined classes. In the Uetz et al. (U) [141] and DIP (D) [156] databases, we identified 74 (U) and 92 (D) cytoskeletal proteins and 141 (U) and 207 (D) signalling proteins in the largest interconnected clusters. There were 15 (U) and 18 (D) proteins shared by the two classes in each database. Surprisingly, tubulin and tubulin-associated proteins were not present in the largest connected clusters for either the Uetz et al. or DIP databases; they formed separate connected clusters with a small number of proteins. As mentioned in some publications [92], one possibility of such unexpected results represent the possible presence of false negative or false positive.

The largest connected cluster within the U and D databases was drawn with c proteins color coded yellow, s proteins color coded green, and proteins found in
both classes color coded red (Fig. 5.1). Inspection of Fig. 5.1 qualitatively suggests
correlations between cytoskeletal and signalling proteins because the majority of these
two protein groups form relatively localized clusters within the network.

In order to quantify the clustering tendency of proteins in each class, we calculated
the distance distribution \( P(d) \) (see Section 5.2 for details) for all protein pairs in the
largest interconnected clusters Fig. 5.2. Since the distance between two proteins was
defined as the number of links required to travel from one protein to another (see
Section 1.3 for details), \( P(d) \) for all proteins in a cluster reflects the degree to which
the proteins within the cluster interact with each other. When calculated for the set
of all proteins in the largest connected cluster in the Uetz database, the peak of \( P(d) \)
was approximately at \( d = 6.8 \). As expected, the peak of the distance distributions
for the \( c \) and \( s \) proteins was shifted to smaller values, 5.4 and 6.0, respectively,
indicating that proteins within these groups preferentially interact with each other.
The corresponding values for all proteins, cytoskeletal proteins and signaling proteins
derived from the DIP dataset are 5.4, 4.0, and 4.3, respectively. It is of note that due
to our definition of the cytoskeletal protein class, the maximum value of \( d_{cc} \) derived
from an ideal interaction map should be 4, since for each protein in this class (except
for septins) the maximal distance from actin is two. (Although the distance of septins
from actin is not constrained, only 3 septins appear in the largest interconnected U
and D clusters so their effect on the maximum value of \( d_{cc} \) is negligible.). This is
clearly not reflected by the two datasets that were used, which is not surprising,
since our procedure to classify the yeast proteins is independent of these interaction
maps. It is, however, consistent with the built-in enhanced clustering of cytoskeletal
proteins in that \( \langle d_{cc} \rangle \) is the smallest among the values listed in Fig. 5.2. Here
Figure 5.1. The largest interconnected cluster of 1,458 interacting proteins in the Uetz et al. data. Yellow and green dots denote cytoskeletal and signaling proteins respectively. The definition of the cytoskeletal and signaling follows the criteria in the text. Proteins in red are shared by two subclasses. The analogous cluster in the D network contains 4,198 proteins. It is not shown here because the density of proteins are too high to examine visually.
Figure 5.2. Distance distribution $P(d)$ for proteins in the various classes. The notation in the figure are as follows: $c - c$: cytoskeletal protein and cytoskeletal protein, $s - s$: signalling protein and signaling protein, $c - s$: signalling protein and cytoskeletal protein. $P(d)$ for all proteins includes distances between any two proteins in the largest cluster. Numbers in the inset are the average distances $\langle d \rangle$ between indicated proteins. (a) Results based on the Uetz et al. protein-protein interaction data. (b) Results based on the DIP protein-protein interaction database. The largest number of proteins in the DIP dataset is manifested in a narrower distance distribution and higher peak values.
\( \langle d \rangle \) denotes the average of \( d \) over the distribution \( P(d) \). For the case of the DIP interaction map of cytoskeletal proteins, where \( \langle d_{cc} \rangle = 4 \) (Fig. 5.2), the majority of \( c - c \) connections do indeed have \( d \approx 4 \). This measurement suggests that \( P(d) \) accurately describes interactions within the networks, and as more information is obtained regarding interactions of cellular proteins, the methods we have devised should be of general use.

Using distance-distribution analysis, we also determined how closely signalling proteins are linked to cytoskeletal proteins. As can be seen from Fig. 5.2, the peak value of \( P(d_{cs}) \), the distance distribution for all pairs of \( c \) and \( s \) proteins is also moved to smaller \( d \) values, indicating that the two groups are more strongly connected to each other within the network than expected for two random sets. Interestingly, the degree to which \( s \) proteins were linked to \( c \) (as measured by \( \langle d_{cs} \rangle \)), was approximately the same as for \( s \) proteins alone (Fig. 5.2). This result suggests that signalling proteins are intimately linked to the cytoskeleton.

5.3.3 Local Clustering

In order to obtain information about the local composition of the interaction networks we calculated the local segregation function, \( m_d(\phi/\lambda) \) (see Section 5.2). This metric characterizes the proportion of proteins at distance \( d \) from a given protein in the \( \phi \) class that are members of the protein class \( \lambda \). In Fig. 5.3 we plot the average segregation function

\[
\langle m_d(\phi/\lambda) \rangle = \frac{1}{N_\phi} \sum_{i=1}^{N_\phi} m_{i,d}(\phi/\lambda), \tag{5.2}
\]
Figure 5.3. Plot of rescaled average local segregation function $\langle m_d \rangle / \langle m_{rand} \rangle$ using Uetz et al. interaction data (left) and the DIP data (right).
here $N_\phi$ is the total number of proteins in the class $\phi$ in the network. This analysis indicates that at short distances, signalling proteins and cytoskeletal proteins interact primarily with proteins of the same class. Note that $\langle m_d(c/c) \rangle$ decays fast as a function of distance and at $d \approx 4$ practically reaches its asymptotic value indicating again that the networks derived from the U and D databases are consistent with our independent definition of the set of cytoskeletal proteins.

In the absence of any clustering tendency between proteins in two different classes, $\phi$ and $\lambda$, the segregation function, $\langle m_d(\phi/\lambda) \rangle$, should be independent of distance and it should be equal to the average density of the $\lambda$ proteins in the network, $m_{\text{rand}}(\phi/\lambda) = N_\lambda/N$. Here, $N_\lambda$ denotes the total number of proteins that belong to the class $\lambda$. In contrast, if proteins belonging to the $\phi$ and $\lambda$ classes have a tendency to cluster, then $\langle m_d(\phi/\lambda) \rangle$ should be higher than $N/N_\lambda$ for small values of $d$ and it should decrease monotonically and converge to a value smaller (possibly zero) than $N/N_\lambda$ for large $d$. These expectations are indeed supported by the plots in Fig. 5.3. For example, using the DIP dataset, the proportion of $s$ proteins connected by a single link to a $c$ protein (red curve at $d = 1$) is almost three times greater than the same quantity evaluated by replacing the $c$ protein by a randomly selected protein (magenta curve at $d = 1$). Furthermore, this proportion is about 6 times higher than the proportion of $s$ proteins linked to the cytoskeleton by 6 or more bonds (red curve at $d = 6$). Similar relationships are seen for the proportion of $c$ proteins linked to $s$ proteins by few bonds compared to many bonds (green curve), whereas analysis of random protein sets shows the predicted flat distribution.

Note that because the $c$ and $s$ classes contain different numbers of members, and the local segregation function is affected by the proportion of proteins in each
class in the entire network, it was necessary to rescale the segregation function \( \langle m_d(\phi/\lambda) \rangle / m_{\text{rand}} \). The rescaled segregation functions are smaller than one already for \( d = 8 \) (the largest distance shown in Fig. 5.3), indicating that at large distances, there is no preferential interaction between proteins within the \( c \) and \( s \) classes.

To further address linkage between signalling and cytoskeletal proteins using the local segregation function, we compared the segregation function, \( \langle m_1(\phi/\lambda) \rangle \) at \( d = 1 \), calculated for all \( s \) and \( c \) proteins. In order to determine if \( s \) proteins are more closely linked to \( c \) proteins by this analysis, it was necessary to compare \( m_1 \) between these groups to \( m_1 \) obtained for proteins chosen at random. The classes of randomly chosen proteins were termed the pseudo-\( c \) and pseudo-\( s \) classes and they contained as many randomly selected proteins as there are \( c \) and \( s \) proteins in the largest interconnected clusters of the employed protein interaction maps.

In Fig. 5.4 we summarize the results of this comparison. For the \( c \) proteins \( \langle m_1(c/c) \rangle \) is about an order of magnitude larger for the true cytoskeletal class than for its pseudo analogue, which may reflect our definition of the \( c \) class. However, the difference between the true and pseudo classes remains consistently large (around a factor of three) for all the other combinations of the \( \phi \) and \( \lambda \) proteins, independently of the dataset used. These results indicate that, at least within the datasets used, the clustering tendency of the \( c \) and \( s \) proteins and the correlation between the two classes are inherent properties of these proteins.

5.3.4 The Special Role of the Cytoskeleton in Signalling Networks

The results in Fig. 5.2-Fig. 5.4 suggest that the cytoskeleton and signalling networks are linked, but this might result fortuitously from the limited nature of the interactions
Figure 5.4. The nearest neighbor segregation function $\langle m_1(\phi/\lambda) \rangle$. Arrows point from $\phi$ to $\lambda$. The left panels in (a) (Uetz data) and (b) (DIP data) summarize results from Fig. 5.3, whereas the right panels results for pseudo protein classes, constructed by selecting randomly 74 (U) or 92 (D) (for pseudo c class) and 141 (U) or 207 (D) (for pseudo s class) proteins from the largest cluster.
detected by the datasets used. To address this possibility, we studied the correlation between the class of signalling proteins and 15 other functional protein classes as defined by the MIP database [96]. We calculated local segregation functions for signalling proteins to each of the other 15 classes of proteins: \( \langle m_d(s/\lambda) \rangle / m_{rand} (\lambda = 0, 1, \cdots, 14) \), where \( i \) denotes the \( \lambda \)-th functional protein class (specified in the legend to Fig. 5). As shown in Fig. 5.5, the nearest neighbor clustering index \( m_1 \) for \( s \) proteins to \( c \) proteins (equal to 2.83 (U), 6.68 (D)), is nearly 2 fold greater than to the next most closely linked class of proteins (class 2 in Fig. 5.5), involved in cell growth, cell division and DNA synthesis (equal to 1.54 (U), 3.9 (D)). These results confirm that the cytoskeleton plays a distinguished role in the organization of cell signalling network.

The cytoskeleton represents a global structure, spanning the entire cell. Thus, its association with the various functional protein classes (in particular with the signalling network) could be expected. To see whether our analysis is consistent with this expectation, we repeated the above calculation for \( \langle m_d(c/\lambda) \rangle / m_{rand} \), the local segregation function of the cytoskeletal proteins and plotted the results in Fig. 5.6. Indeed, as comparison of Fig. 5.5 and Fig. 5.6 reveals, the association of the \( c \) proteins with the 15 functional protein classes defined in the MIPS database is quite uniform, suggesting that signalling proteins have no special role in the organization of the cytoskeleton. This is particularly well reflected by the values of \( m_1 \). The nearest neighbor segregation function for the \( c \) proteins to the \( s \) proteins and to the proteins in class 2, are relatively much closer to each other than the corresponding quantities with \( c \) replaced by \( s \): \( \langle m_1(c/s) \rangle / \langle m_1(c/2) \rangle \) is 44% (U) or 61% (D) smaller than \( \langle m_1(s/c) \rangle / \langle m_1(s/2) \rangle \).
Figure 5.5. $\langle m_d(s/\lambda)\rangle/m_{rand} (\lambda = 0, 1, \cdots, 14)$ for fifteen functional classes in addition to the $c$ class. Upper and lower panels are based on Uetz et al. data and the DIP database respectively. The 15 functional protein classes are the following: 0-metabolism, 1-energy, 2-cell growth, cell division and DNA synthesis, 3-transcription, 4-protein synthesis, 5-proteins destination, 6-transport facilitation, 7-cellular transport mechanisms, 8-cellular biogenesis, 9-cell rescue, defense, cell death and ageing, 10-ionic homeostasis, 11-cellular organization, 12-transposable elements, viral and plasmid proteins, 13-classification not yet clear cut, 14-unclassified proteins.
Figure 5.6. $\langle m_d(c/\lambda) \rangle / m_{rand}$ ($\lambda = 0, 1, \cdots, 14$) for fifteen functional classes in addition to the $c$ class. Upper and lower panels are based on Uetz et al. data and the DIP database, respectively. The fifteen functional classes are the same as in Fig. 5.5.
The quantitative analysis presented here suggests that the topological properties of intracellular signalling pathways within the protein interaction network of *S. cerevisiae* are strongly dependent on the cytoskeleton. This linkage was even more evident when only those cytoskeletal and signalling proteins that are connected to each other exclusively through *c* or *s* proteins were analyzed. The corresponding subnetwork derived from the U database is shown in Fig. 5.7. All of the proteins that directly interconnect the two classes are unusual in that they have the highest number of links (at least 4 here). That is, they are hubs and they are distributed throughout the network, indicating that the cytoskeleton and the set of signalling molecules are linked in a global manner.

The protein interaction networks analyzed here are examples of scale-free networks [13, 71, 72], with the property of being simultaneously tolerant to random errors and fragile against the removal of the most connected nodes or hubs [5]. In order to investigate the significance of the hubs in the present context we removed all signalling proteins that link the signalling subnetwork to the cytoskeleton (23 of the 28 hubs). The resulting interaction map (with only those proteins shown that have at least one connection) is plotted in Fig. 5.8. The total collapse or fragmentation of the signalling network, as seen in Fig. 5.8, strongly suggests that without communication with the cytoskeleton the cell's signalling apparatus cannot properly function.

It is perhaps not surprising that many of the most connected hubs in the subnetwork were identified as members of both the cytoskeleton and signalling subsets. Some of these proteins, for example Las17p, the yeast WASP homolog, and Cla4p, the yeast PAK1 kinase homolog, are well-characterized regulators of the cytoskeleton, and they coordinate cytoskeletal dynamics with changes in cell growth, division,
Figure 5.7. The combined $c-s$ subnetwork derived from the largest connected cluster in Fig. 5.1. Yellow and green dots denote signalling and cytoskeletal proteins, respectively, proteins in red are shared by the two subclasses. Only proteins with at least one connection are shown.
Figure 5.8. Collapse of the signalling subnetwork shown in Fig. 5.7, on knock-out of s proteins connecting to the cytoskeletal proteins.
and mating. Other hubs provide critical (possibly the only) connections between two parts of the signalling network. For example, Akr1p, an ankyrin repeat-containing cytoskeletal protein, provides a pathway in this network to transmit a signal from Gcs1p and Ste3p to other components of the mating pathway (Ste4p, Ste5p, and Ste18p).

5.3.5 Discussion

The analysis presented in this chapter provides quantitative evidence for the long-standing hypothesis that the cytoskeleton participates in an important way in intracellular signal transduction. How might the cytoskeleton be used in signal transduction pathways? The results of the network analysis suggest that the cytoskeleton may potentially serve in at least two ways. First, individual proteins of the cytoskeleton may participate directly in signal transduction by linking two or more signalling proteins. One implication of this role is that the cytoskeleton may provide alternative signal transduction routes so that there are multiple pathways to transduce a signal. Secondly, the cytoskeleton may provide a macromolecular scaffold, which spatially organizes components of a signal transduction cascade [114]. This function is analogous to the role of molecular scaffolds, such as the yeast Ste5 protein, which their multiple components of a pathway to promote signal transduction between them. The analysis presented here suggests that during eukaryotic evolution, signalling pathways have incorporated components and features of the cytoskeleton as their integral parts and this may be a general feature of eukaryotic intracellular signal transduction networks.
CHAPTER 6

NETWORK DYNAMICS: FLUX FLUCTUATIONS WITH MULTIPLICATIVE NOISE

6.1 Introduction

Nonequilibrium statistical mechanics have been an intensively studied subject [91, 136] in various physical systems [3, 9, 17, 41, 42, 88, 117, 149]. One example of nonequilibrium systems which has received a lot of attention in the last decade is self-organized criticality with potential relevance to the scale-invariant features observed in many phenomena such as the sand-pile model [11], river networks [36, 90, 126] and growth dynamics [17]. An early version of self-organized critical model was proposed by Takayasu [138, 139]. The model is believed to be relevant to economic systems because its diffusion process with aggregation, deposition and evaporation of particles resembles the “efficiency” dynamics of competing agents in economic systems [85]. In order to make direct comparisons between a model system and empirical observations, however, one should understand how the underlying interaction topology between agents affects their dynamical behavior. The effects of the underlying topology to dynamical properties such as the relaxation time, the autocorrelation function and the return probability of random walkers have been recently studied [6, 110]. Especially, it is known that diffusive particles constrained to move along the nodes and links of a
scale-free network tend to agglomerate on the high connectivity nodes or hubs [111]. Moreover, as mentioned in Section 1.5, it has been found that the average number of walkers, $\langle H_i \rangle$, reaching a node $i$ on a reasonably large time window $\tau \ll T_{\text{max}}$, scales with the fluctuations around the average $\sigma$ as $\langle H_i \rangle \sim \sigma^\alpha$, where $\alpha = 1/2$ if the total number of walkers on the system does not fluctuate too much between different time windows, and $\alpha = 1$ otherwise [35] ($\langle \cdots \rangle$ represents average over a time window $\tau$ and $T_{\text{max}}$ is the observation time). Thus, the amplitude $\Delta W$ of the variations of the average number of walkers in the system determines the value of the exponent $\alpha$. Here we show that, if the diffusive process is multiplicative and if the underlying topology is intrinsically inhomogeneous, there is a crossover from $\alpha = 1/2$ to $\alpha = 1$ on the $\langle H \rangle$ vs. $\sigma$ curve even for $\Delta W = 0$. While this crossover is evident for multiplicative diffusive processes on a scale-free network, it does not appear on random, Erdős-Rényi networks, for which the connectivity distribution is narrowly peaked at $\langle k \rangle$.

In this chapter, we will discuss the influence of scale-free (SF) topology on the fluctuation of flux both numerically and analytically. We generate SF networks using the simple BA model [13], since our main interest is finding the relationship between the connectivity and the fluctuation of flux. For comparison with other random topologies, we will also use the Erdős-Rényi (ER) network [46].

6.2 Model and Numerical Analysis

6.2.1 Definition of Model

In the Takayasu model, a non-negative mass variable $h_i$ is assigned to each site $i$ of a lattice. At each time step the total mass of a randomly selected site $l$ moves to one of its nearest neighbors $j$ and aggregates with rate 1 resulting $h'_j = h_j + h_l$. In
addition, a unit mass is deposited at a randomly chosen site with rate \( q \) [138, 139]. Majundar et al. [85, 86] generalized the model by including desorption of a unit mass from a randomly chosen site with rate \( p \) and showed that in the steady state the mass distribution \( P(h) \) follows a power-law for \( q < q_c(p) \) and decays exponentially for \( q > q_c(p) \), with [85, 86]

\[
q_c(p) = p + 2 - 2\sqrt{p + 1}.
\]  

(6.1)

In our study we assume that the amount of mass moving from a node \( i \) to one of its nearest neighbors is a random fraction of the total mass at node \( i \). Thus, we adopt the following set of rules for the diffusive process: starting from a random initial distribution of mass \( h_i \) at each node \( i \), (i) a unit of mass is deposited at a randomly chosen site \( i \) with probability \( p/(p+q+1) \), (ii) a unit mass is removed from a randomly chosen site \( i \) with probability \( q/(p + q + 1) \), and (iii) a random fraction of mass from a randomly chosen site \( i \) moves to one of its (randomly chosen) nearest neighbors \( j \) with probability \( 1/(p + q + 1) \).

We also define the flux \( H \) at node \( i \) during a given time interval \( \Delta t \) as

\[
H_i(t) = \sum_{\tau} \Delta h_{i}^{in}(t + \tau),
\]  

(6.2)

where \( h_{i}^{in}(t) \) is the incoming mass at time \( t \) and \( \tau \) is the size of the time window representing the level of coarse-graining of the measurement time. Thus, we obtain for each node \( i \) a time series \( \{H_i(0), H_i(\tau), \ldots, H_i(T_{max})\} \), where \( T_{max} \) is the total observation time and \( T_{max} \gg \tau \).
6.2.2 Numerical Analysis

In numerical simulations we first constructed the BA or ER network with \( N = 10^5 \) nodes. Then we set the parameters \( p \) and \( q \) to satisfy the conditions: (i) \( \sum_i^N h_i \rightarrow const. \) (steady state condition), and (ii) \( q \approx q_c \) in Eq. (6.1). Fig. 6.1 shows the behavior of \( \sigma(\langle H \rangle) \) for different choices of \( (p, q) \) pairs satisfying conditions (i) and (ii). As indicated in Fig. 6.1, as long as there is diffusion on SF networks, \( \sigma(\langle H \rangle) \)'s show the same type of crossover \( (\alpha = 0.5 \rightarrow 1) \) independent of the diffusion rate. This caused by the fact that diffusion governs the dynamics of the system in the thermodynamic limit. The only difference between high and low diffusion rate is the value of \( \langle H_c \rangle \) at which the crossover occurs. Therefore, without loss of generality, we can choose any finite value of \( p \) and \( q \) as long as they satisfy the conditions (i) and (ii), and we will use the value \( (p+q)/(p+q+1) = 0.1 \) throughout this chapter. Averaging \( H_i(t) \) over time and calculating \( \sigma_i \) for each node \( i \), we find that \( \langle H \rangle \sim \sigma^\alpha \), where \( \alpha = 1/2 \) for small \( \langle H \rangle \) values, and \( \alpha = 1 \) for large \( \langle H \rangle \) (Fig. 6.2.2(a)). As shown in Ref. [111], high-connectivity nodes are subject to more traffic than low-connectivity nodes, the traffic fluctuations being amplified by the multiplicative nature of the process, that is, the bigger the mass, the larger is the amount of mass transferred to the nodes, on the average. We check this by measuring the average connectivity \( \langle k \rangle \) of a node receiving an amount of mass \( \langle H \rangle \), shown in Fig. 6.2.2(b), confirming that hubs receive more traffic. From Fig. 6.2.2(a) and Fig. 6.2.2(b) one can find \( \langle k \rangle_c \), the value of \( k \) associated with the value of \( \langle H \rangle \) where the crossover occurs, and calculate the restricted distributions of mass \( P(\langle H \rangle)_{k<k_c} \) and \( P(\langle H \rangle)_{k>k_c} \). As depicted in Fig. 6.2.2(c), these distributions differ fundamentally, the low-connectivity \( (k < k_c) \) distribution being much narrower than the high-connectivity one \( (k > k_c) \). It is this
Figure 6.1. The plot of $\sigma$ against $\langle H \rangle$ on SF networks. Squares (□) represent $(p + q)/(p + q + 1) = 0.0$ (maximum diffusion rate) and circles (○) represent $(p + q)/(p + q + 1) = 0.8$ (low diffusion rate). The slope of the solid line is $\alpha = 1/2$, and the slopes of the dashed lines are $\alpha = 1.0$. Both low and high diffusion rates cause the same type of crossover.
broad range in $P((H)_{k>k_c})$ that makes the average mass $\langle H \rangle$ scales linearly with the mass fluctuations for $\langle H \rangle \gg \langle H \rangle_c$ (Ref. [35]). We verified this by repeating the same experiment on Erdős-Rényi random networks. In this case the crossover disappears (Fig. 6.2.2(d)), the average mass scales weakly with connectivity (Fig. 6.2.2(e)), and the restricted mass distributions for low and high connectivity are both narrowly distributed (Fig. 6.2.2(f)), which is reflected in the $\alpha = 1/2$ exponent measured over the whole range of $\langle H \rangle$ (Fig. 6.2.2(f)).

From these numerical simulations we conclude that the topological differences between scale-free and random networks lead to different dynamical behaviors. This can be accredited to the multiplicative nature of the diffusive process, since simple diffusion on scale-free networks leads to $\alpha = 1/2$, as reported in Ref. [35].

6.3 Master Equation Approach

To gain a deeper understanding of the crossover observed on the SF networks, we solved the master equation by a mean-field type approximation. Since, by definition, the incoming flux is a random fraction of the nearest neighbors’ mass, the incoming flux is on the average proportional to the mass at the nearest neighbors. The master equation for the change of mass at node $i$ during a unit time interval is

$$H_i(t+1) = H_i(t) + \sum_j \frac{1}{k_j} \eta_j(t) H_j(t) - \eta_i(t) H_i(t) + \xi_i(t). \quad (6.3)$$

Here the second term on the right hand side stands for the incoming mass from nearest neighbors and the third term represents the outgoing mass from node $i$ during the unit time interval. The last term represents fluctuations caused by deposition and desorption. Moreover, $\eta_i(t)$ and $\xi_i(t)$ are assumed to be Gaussian white noises.
Figure 6.2. Plots for the diffusion-dominant region on SF networks((a)-(c)) and on ER networks((d)-(f)). (a) and (d) show the measurement of fluctuations of incoming fluxes on a SF network and a random network respectively. Solid lines denote $\sigma \sim \langle H \rangle$ and dotted lines represent $\sigma \sim \langle H \rangle^{1/2}$. (b) and (e): average connectivity $\langle k \rangle$ as a function of average incoming flux $\langle H \rangle$. Note that in the ER graph, the largest value of $\langle k \rangle$ is around 10, corresponding to the $\alpha = 1/2$ region in SF networks. (c) and (f): average incoming flux distribution. Solid squares (■) represent low connectivity region and open circles (○) denote high connectivity region.
uncorrelated both in space and time. This type of diffusion equation with multiplicative noise has been studied in the context of stochastic Lotka-Volterra models, which can be characterized by the (truncated) Pareto or Lévy distribution $P(H) \sim H^{-1-\mu}$ [24, 133, 134]. Since we are focusing only on the incoming mass in the diffusion dominant region, Eq. (6.3) can be reduced to

$$H_i(t + 1) = H_i(t) + \sum_j k_j \eta_j(t) H_j(t).$$ (6.4)

In the continuum limit we can approximate the change of incoming mass as

$$\frac{dH}{dt} \approx \sum_j k_j \eta_j(t) H_j(t).$$ (6.5)

The data in Fig. 6.3 shows that the incoming flux distribution $P(H)$ also has a long power-law decaying tail even though we do not consider the outgoing flux. The value of the exponent $\mu$ is known to depend on the details of model’s parameters [24, 134, 133]. By using the least square fitting, we find the corresponding exponent to be $-1 - \mu \approx -3.3$ for the incoming flux in our model.

In order to obtain an approximate expression for $H(t)$ we assume that $\sum_j k_j \eta_j(t) H_j(t) \approx \langle k_i \rangle \langle \eta_j(t) H_j(t) \rangle$, where $\langle k_{nn} \rangle$ denotes the average degree of a node’s nearest neighbors. Since $\eta_j(t)$ and $H_j(t)$ are independent variables, Eq. (6.5) becomes

$$\frac{dH}{dt} \approx \frac{\langle k \rangle}{\langle k_{nn} \rangle} \langle \eta \rangle \langle H_{nn} \rangle \equiv A(\langle k \rangle) \langle H_{nn} \rangle,$$ (6.6)

where $\langle H_{nn} \rangle$ represents the average incoming mass on the nearest neighbors of the node. In general, the average connectivity of the nearest neighbors of a given node can be expressed as a function of connectivity of the chosen node, i.e. $\langle k_{nn} \rangle \sim \langle k \rangle^{-\nu}$. The BA model corresponds to $\nu = 0$ [142]. To express $\langle H_{nn} \rangle$ as a function of incoming
Figure 6.3. Plot of incoming flux distribution $P(H)$. The slope of the solid line represents $1 - \mu \simeq -3.3$.
Figure 6.4. Plot of $\langle H_{nn} \rangle$ against $H$. The inset shows $\langle H_{nn} \rangle$ for $k \leq 10^8$. The solid line represents the hyperbolic decaying function $\langle H_{nn} \rangle = a + b/(c + H)$.

As shown in Fig. 6.4, $\langle H_{nn} \rangle$ decreases rapidly as $\langle H \rangle$ increases and converges slowly to a constant value. The crossover between the rapidly decaying region and the almost flat region occurs around $H_c \simeq 10^8$, and this value is consistent with the value obtained from Fig. 6.2.2(a). Based on the data in Fig. 6.4, we approximated $\langle H_{nn} \rangle = \text{const.}$ for $H \geq 10^8$ and

$$\langle H_{nn} \rangle \approx \left( a + \frac{b}{c + H} \right), \quad (6.7)$$

for $H \leq 10^8$. Here $a$, $b$ and $c$ are the fitting parameters. Using the least square fitting
we obtained the value of parameters $a$, $b$, and $c$:

$$
\begin{align*}
\quad a &= 4.52581 \times 10^8, \\
\quad b &= 5.7908 \times 10^{15}, \\
\quad c &= 2.58341 \times 10^6.
\end{align*}
$$

(6.8)

The calculation for the $H \geq 10^8$ with the approximation $\langle H_{nn} \rangle = \text{const.}$ gives

$$
\langle H \rangle \sim \langle k \rangle, 
$$

and

$$
\sigma \sim \langle H \rangle.
$$

(6.9)

(6.10)

Eq. (6.9) indicates that the linear dependence of $\langle k \rangle$ on $\langle H \rangle$ for large $\langle H \rangle$ values in Fig. 6.2.2 (b) comes from the nearly constant value of $\langle H_{nn} \rangle$ leading to the linear dependence of $\sigma$ on $\langle H \rangle$.

On the other hand, the calculation of Eq. (6.4) for $H \leq 10^8$ requires somewhat lengthy integrations and expansions. From the direct integration of Eq. (6.6) (using the approximation Eq. (6.7)) we obtain

$$
-\frac{b}{a^2}\ln\left(1 + \frac{aH}{ac+b}\right) + \frac{H}{a} = A\langle k \rangle t,
$$

(6.11)

and since the coefficients $a/b \ll 1$, the Taylor expansion of $\exp(-aH/b)$ around $a/b = 0$ yields

$$
B_0H^2 - B_1H - B_2 = 0,
$$

(6.12)
where

\[ A(\langle k \rangle) \equiv \langle k \rangle^{1-\nu} \langle \eta \rangle, \]

\[ B_0 \equiv \frac{a^2}{b(ac + b)}, \]

\[ B_1 \equiv \frac{a}{ac + b} - \frac{a}{b}, \]

\[ B_2 \equiv 1 - \exp \left(-\frac{a^2 A}{b} t \right). \]

(6.13)

Now, if we consider the condition that \( H \geq 0 \), the only possible root must have positive value, i.e.,

\[ H(\langle k \rangle, t) = \frac{B_1 + \sqrt{B_1^2 + 4B_0B_2}}{2B_0} = C_0 + C_1 \sqrt{C_2} - e^{-qt}, \]

(6.14)

where \( q = a^2 A/b \), \( C_0 = B_1/2B_0 \), \( C_1 = 1/\sqrt{B_0} \), and \( C_2 = B_1^2/4B_0 \). In the large \( t \) limit \( e^{-qt} \ll 1 \), expanding \( \sqrt{C_2} - e^{-qt} \) around \( e^{-qt} = 0 \), Eq. (6.14) can be rewritten as

\[ H(\langle k \rangle, t) = D_0 + D_1 e^{-qt}. \]

(6.15)

Here \( D_0 = C_0 + C_1 C_2 \) and \( D_1 = C_1/2\sqrt{C_2} \).

Thus,

\[ \langle H \rangle \approx \frac{1}{T} \sum_t H(\langle k \rangle, t) \]

\[ \approx D_0 + \frac{D_1}{T} \int_0^T e^{-qt} dt \]

\[ \approx D_0 - \frac{D_1}{Tq} (e^{-qT} - 1), \]

(6.16)
and

\[ \langle H^2 \rangle = D_0^2 - \frac{2D_0D_1}{Tq} (e^{-qT} - 1) - \frac{D_1^2}{2qT} (e^{-2qT} - 1) . \quad (6.17) \]

Therefore, the fluctuations in the incoming mass is given by

\[
\sigma = \sqrt{\langle H^2 \rangle - \langle H \rangle^2} \\
\simeq \sqrt{\frac{D_1^2}{2qT} (1 - e^{-2qT}) + \frac{D_1^2}{T^2q^2} (1 - e^{-qT})}. \quad (6.18)
\]

In the limit \( T \to \infty \) the dominant term in Eq. (6.18) scales with \((1/T)^{1/2}\). From Eq. (6.16), we know that the average incoming flux \( \langle H \rangle \) is proportional to \((1/T)\), hence we conclude that \( \sigma \sim \langle H \rangle^{1/2} \).

### 6.4 Summary and Discussion

In summary, we have studied a diffusion process reminiscent of Takayasu’s model on scale-free and Erdös-Rényi networks. In this process, the number of particles on a given node dictates how many particles will diffuse simultaneously from it. This type of dynamics can be described by a diffusion equation with multiplicative noise, generally characterized by Pareto or Lévy distributions and is regarded as a model for systems such as population dynamics of each species in a given region or stock exchange markets in the financial systems. Through numerical simulations of the model, we found a nontrivial crossover between two different scalings of the average number of incoming particles on a node \( \langle H \rangle \) and fluctuations about the average \( \sigma \), which occurs due to the multiplicative nature of the process coupled to intrinsic inhomogeneity of the underlying scale-free topology. This result is also verified by solving the master equation of the process. Comparisons with a simpler, random
geometries offer us an insight on how topology affects dynamics, since this crossover is not present on the latter.
BIBLIOGRAPHY


