

# Unravelling the Complexity of Metabolic Networks

A Thesis by Nicole Pearcy

in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy.

Department of Physics & Mathematics

Nottingham Trent University

August 2015.

©Nicole Pearcy, 2015.

## Copyright Statement

This work is the intellectual property of the author.

# ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Dr. Jonathan Crofts for the constant support and guidance he has provided throughout my PhD. I am also very grateful to Professor Nadia Chuzhanova, my second supervisor, for her help and support. I would also like to acknowledge the Mathematics department for their help and support during my time at Nottingham Trent University. The financial support via an RAE funded PhD scholarship provided by the Nottingham Trent University is also highly appreciated.

Outside of NTU, I am grateful to Professor Kazuhiro Takemoto for generously providing the reaction lists from the KEGG database, which were used to reconstruct the metabolic networks used in this work. Thanks are also owed to Kazuhiro for his advice and guidance on the reconstruction process of the metabolic networks.

I would also like to acknowledge Funds for Women Graduates (FfWG) for providing additional financial support during the last year of my PhD.

Finally, I would like to thank my family and friends for their support and patience throughout my PhD. In particular, I would like to thank my parents for their constant encouragement throughout the years.

# ABSTRACT

Network science provides an invaluable set of tools and techniques for improving our understanding of many important biological processes at the systems level. A network description provides a simplified view of such a system, focusing upon the interactions between a usually large number of similar biological units. At the cellular level, these units are usually interacting genes, proteins or small molecules, resulting in various types of biological networks. Metabolic networks, in particular, play a fundamental role, since they provide the building blocks essential for cellular function, and thus, have recently received a lot of attention. In particular, recent studies have revealed a number of universal topological characteristics, such as a small average path-length, large clustering coefficient and a hierarchical modular structure. Relations between structure, function and evolution, however, for even the simplest of organisms is far from understood.

In this thesis, we employ network analysis in order to determine important links between an organism's metabolic network structure and the environment under which it evolved. We address this task from two different perspectives: (i) a network classification approach; and (ii) a more physiologically realistic modelling approach, namely hypernetwork models. One of the major contributions of this thesis is the development of a novel graph embedding approach, based on low-order network motifs, that compares the structural properties of large numbers of biological networks simultaneously. This method was prototyped on a cohort of 383 bacterial networks,



and provides powerful evidence for the role that both environmental variability, and oxygen requirements, play in the forming of these important networked structures.

In addition to this, we consider a hypernetwork formalism of metabolism, in an attempt to extend complex network reasoning to this more complicated, yet physiologically more realistic setting. In particular, we extend the concept of network reciprocity to hypernetworks, and again evidence a significant relationship between bacterial hypernetwork structure and the environment in which these organisms evolved. Moreover, we extend the concept of network percolation to undirected hypernetworks, as a technique for quantifying robustness and fragility within metabolic hypernetworks, and in the process find yet further evidence of increased topological complexity within organisms inhabiting more uncertain environments. Importantly, many of these relationships are not apparent when considering the standard approach, thus suggesting that a hypernetwork formalism has the potential to reveal biologically relevant information that is beyond the standard network approach.

# LIST OF FIGURES

1.1	Front cover of the January 2015 issue of Molecular BioSystems featuring work from this thesis. . . . .	12
2.1	Illustration of undirected and directed network models. . . . .	14
2.2	Illustration of a weighted graph model. . . . .	15
2.3	Illustration of a bipartite graph and two possible graph projections. .	15
2.4	Illustration of undirected and directed hypergraph models. . . . .	16
2.5	Illustration of the adjacency matrices associated with the different graph models. . . . .	18
2.6	The average path-length and corresponding distance matrix for (a) an undirected graph ( $\ell = \langle D \rangle = 1.5$ or $\varepsilon = 0.75$ ); and (b) a directed graph ( $\ell = \langle D \rangle = \infty$ and $\varepsilon = 0.47$ ). . . . .	20
2.7	Local and global clustering as defined in Equations (2.7–2.8): (left) $C = \{0, 0, 0, 0\}$ and $\langle C \rangle = 0$ , (middle) $C = \{1, 1, 1, 1\}$ and $\langle C \rangle = 1$ , (right) $C = \{1, 1/3, 0, 1\}$ and $\langle C \rangle = 7/12$ . . . . .	23
2.8	The degree distribution for the metabolic network of the bacteria <i>E.coli</i> : (a) histogram of $P(k)$ versus $k$ , and (b) a log-log plot of $P(k)$ versus $k$ . . . . .	23
2.9	Classical random graph, $G_{n,p}$ , with 12 vertices and varying probabilities ( $p$ ) for connecting pairs of vertices together. . . . .	26

- 2.10 The Watts-Strogatz model. Here we start with a regular graph with 12 vertices and 24 edges, such that each vertex is connected to its nearest neighbours, leading to a maximum clustering coefficient ( $C = 1$ ). Each edge is randomly rewired by some probability  $p$ . As  $p \rightarrow 1$  the initial regular graph is transformed into an ER random graph (right). For intermediate values of  $p$ , such that  $0 < p < 1$ , a small-world network arises - highly clustered like a regular graph, yet with a small average path-length like a random graph. . . . . 29
- 2.11 Degree distributions of two graph models: (a) a classical random graph (top) together with its degree distribution (bottom); and (b) a scale-free graph (top) together with its degree distribution (bottom). 30
- 2.12 The Barabasi-Albert growth model. (a) We start with a graph with  $N_0 = 3$  vertices and  $E_0 = 2$  edges. (b) A fourth vertex preferentially attaches itself to 2 out of the 3 existing vertices, where the probability that an existing vertex receives an edge is proportional to its degree ( $\mathbf{p} = [0.5, 0.25, 0.25]$ ). (c) A fifth vertex now preferentially attaches itself to 2 out of the 4 existing vertices with probabilities  $\mathbf{p} = [0.375, 0.125, 0.250, 0.250]$ . (d) Finally, a sixth vertex attaches to 2 out of the 5 existing vertices with probabilities  $\mathbf{p} = [0.333, 0.083, 0.167, 0.250, 0.167]$ . Notice here that we have arrived at a network that is highly heterogeneous in terms of the degrees of the vertices. . . . . 31
- 2.13 Examples of network motifs: (a) three-chain motif, (b) feed-forward loop motif, (c) bi-parallel motif, (d) bi-fan motif, and (e) single-input motif. . . . . 32

2.14	Example of the rewiring step of the switching method. The two edges $E_{A,B}$ and $E_{C,D}$ are reconnected to make $A$ connected to $D$ and $C$ connected to $B$ . . . . .	33
3.1	A schematic showing the relationship between catabolism and anabolism in a cell. . . . .	37
3.2	A minimal system of glycolysis. Note that here $G_x$ , $P_x$ and $GI_x$ are external and so are not considered in the system of ODEs. The abbreviations correspond to the following: $G$ , glucose, $P$ , pyruvate and $TP$ , triosephosphate. (a) A minimal set of reactions, where one unit of glucose is converted into pyruvate. (b) A graphical illustration of the reaction system. (c) The set of ordinary differential equations (ODEs) modelling the change in concentration of each metabolite. (d) Matrix representation of the system of ODEs. . . . .	39
3.3	Illustration of constraint based modelling of metabolism. At first the flux distribution of the metabolic network may lie anywhere in the solution space (left). The mass balance constraint and the lower and upper capacity bounds are imposed, reducing the allowable solutions to a convex space (middle). An optimisation of an objective function allows FBA to identify an optimal solution, which lies on the edge of the convex space. . . . .	41
3.4	An illustration of simple graph models available for representing metabolic networks. . . . .	43
4.1	Illustration of our algorithmic approach. . . . .	57

- 4.2 A biplot representation of the data matrix for an ensemble of 235 networks consisting of 115 bacterial networks (black), 40 Erdős-Rényi networks (blue), 40 small-world networks (green), and 40 scale-free networks (red). The axis represent the three dimensions obtained via non-negative matrix factorisation transformation. . . . . 59
- 4.3 2-dimensional projections of the biplot in Figure 4.2 on to (a) columns 1 and 2; (b) columns 1 and 3; and (c) columns 2 and 3. As before, bacterial networks are shown in black, Erdős-Rényi networks in blue, small-world networks in green, and scale-free networks in red. . . . . 61
- 4.4 Relationship between environmental variability and the mean global significance score  $\langle P_{\text{global}} \rangle$  for the six bacterial habitats: **O**bligate, **S**pecialised, **A**aquatic, **F**acultative, **M**ultiple and **T**errestrial. Vertical bars represent standard errors. . . . . 65
- 4.5 Relationship between growth requirements, more specifically oxygen requirements, and the mean global significance score  $\langle P_{\text{global}} \rangle$ . Vertical bars represent standard errors. . . . . 67
- 4.6 Mean normalised frequency for the 263 metabolites obtained within motif 5 for the 115 metabolic networks. Blue bars represent the specialised class and red bars represent the varied class. Here, the metabolites are in descending order of the metabolite frequencies for the varied class. . . . . 69

- 4.7 Mean normalised frequency for the significant metabolites obtained within motif 5 with  $p < 0.01$  (Fisher's Exact test). Vertical bars are standard errors. Asterisks indicate large significant differences between metabolic networks from a specialised and varied environment, where \*, \*\*, and \*\*\* correspond to  $p < 0.001$ ,  $p < 0.0001$  and  $p < 0.00001$ . Metabolite names are provided for the most significant metabolites. . . . . 70
- 4.8 Mean normalised frequency for the 54 metabolites obtained within motif 9 for the 115 metabolic networks. Blue bars represent the specialised class and the red bars represent the varied class. Here, the metabolites are in descending order of the metabolite frequencies for the varied class. . . . . 71
- 4.9 Mean normalised frequency for the 54 metabolites obtained within motif 9 for the 115 metabolic networks. Vertical bars are standard errors. Asterisks indicate levels of significance, with \*, \*\*, and \*\*\* corresponding to  $p < 0.05$ ,  $p < 0.01$  and  $p < 0.001$ , respectively. Metabolite names are provided for the most significant metabolites. . 72
- 4.10 Mean normalised frequency for the 291 metabolites obtained within motif 5 for the 383 metabolic networks. Blue bars represent the aerobic-facultative class and red bars represent the anaerobic class. Metabolites are displayed in descending order of the metabolite frequencies for the aerobic-facultative class. . . . . 74

4.11	Mean normalised frequency for the significant metabolites obtained within motif 5 with $p < 0.01$ (Fisher's Exact test). Vertical bars are standard errors. Asterisks indicate large significant differences between metabolic networks from the aerobic-facultative and anaerobic class, where *, **, and *** correspond to $p < 0.001$ , $p < 0.0001$ and $p < 0.00001$ . Metabolite names are provided for the most significant metabolites. . . . .	75
4.12	Mean normalised frequency for the 65 metabolites obtained within motif 9 for the 383 metabolic networks. Blue bars represent the aerobic-facultative class and the red bars represent the anaerobic class. Here, the metabolites are in descending order of the metabolite frequencies for the aerobic-facultative class. . . . .	76
4.13	Mean normalised frequency for the 65 metabolites obtained within motif 9 for the 383 metabolic networks. Vertical bars are standard errors. Asterisks indicate levels of significance, with *, **, and *** corresponding to $p < 0.05$ , $p < 0.01$ and $p < 0.001$ , respectively. Metabolite names are provided for the most significant metabolites. .	77
5.1	Hypergraph model of metabolism: (a) an example of a hypothetical reaction system; (b) the accompanying hypergraph model; (c) its incidence matrix representation; and (d) its adjacency matrix representation. . . . .	81
5.2	Illustration of a directed hypergraph with its corresponding underlying hypergraph. . . . .	84
5.3	The in- and out-degree distributions (a) and the in- and out- cumulative degree distributions (b) for the directed hypergraph of <i>E. coli</i> . .	85
5.4	The hyperarc out-degree (red) and in-degree (blue) distributions of <i>E. coli</i> . . . . .	86

5.5	Toy hypergraph models to illustrate the local clustering coefficient, $HC_{\text{local}}$ , for a variety of different scenarios. . . . .	89
5.6	The scaling of the local clustering coefficient for (a) the metabolic hypergraph and (b) the metabolic standard graph of <i>E.coli</i> . Note that the blue crosses are the averaged local clustering coefficients for each unique degree, whereas the red circles are the local clustering coefficients. . . . .	90
5.7	Spy plots of (a) the incidence matrix and (b) the adjacency matrix of the directed hypergraph representation of metabolism in <i>E. coli</i> . . . .	92
5.8	Relationship between average hypernetwork reciprocity, $\langle r \rangle$ , and environmental variability. The six bacterial habitats along the x-axis are in order of environmental variability according to the NCBI classification scheme: <b>O</b> bligat, <b>S</b> pecialised, <b>A</b> quatic, <b>F</b> aculative, <b>M</b> ultiple and <b>T</b> errestrial. Here vertical bars denote the standard error of the mean. . . . .	94
5.9	Relationship between average standard reciprocity, $\langle r \rangle$ , and environmental variability. Note that we average over the six bacterial habitats along the x-axis, which are ordered according to the NCBI classification scheme of environmental variability: <b>O</b> bligat, <b>S</b> pecialised, <b>A</b> quatic, <b>F</b> aculative, <b>M</b> ultiple and <b>T</b> errestrial. Here vertical bars denote the standard error of the mean. . . . .	96
5.10	Figures (a)-(c) highlight two iterations of the site percolation process for a toy hypernetwork with $n = 15$ and $m = 9$ starting from a configuration with $p = 8/15$ . Note that here active nodes are highlighted in black, whilst inactive nodes are in grey. The nodes and hyperedges that are highlighted red and blue correspond to those that become activated in the first and second iterations, respectively. . . . .	99



- 5.11 Site percolation for the two bacterial hypernetworks: *Buchnera* (top row) and *E. coli* (bottom row). Here, we compare the percolating properties,  $S$  and  $\chi$  of the original hypernetworks (red solid lines) against an ensemble of 100 rewired hypernetworks (blue solid lines). Note that the dashed lines indicate the percolation thresholds of the corresponding hypernetwork:  $p_c = 0.53$ , for *E. coli* ( $\langle p_c^{\text{rand}} \rangle = 0.50 \pm 0.0026$ ) and  $p_c = 0.61$  for *Buchnera* ( $\langle p_c^{\text{rand}} \rangle = 0.57 \pm 0.0048$ ). . . . . 100
- 5.12 (a) Plot of the percolation threshold for each of the 115 metabolic networks versus the mean percolation threshold for matched random graphs. (b) Plot of the global clustering coefficient for each of the 115 metabolic networks versus the mean global clustering coefficient for matched random graphs. . . . . 102
- 5.13 Relationship between the average percolation threshold,  $\langle p_c \rangle$ , and the environmental variability. Note that the six bacterial habitats along the x-axis are in order of environmental variability: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. The vertical bars represent the standard error of the mean. . . . . 103
- A.1 Subplots showing the relationship between environmental variability and the mean global significance score  $\langle P_{\text{global}}(i) \rangle$  for the 6 bacterial habitats: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. Each subplot shows the result for a different value of  $k$ . . 136
- A.2 Subplots showing the relationship between growth conditions, more specifically oxygen requirements, and the mean global significance score  $\langle P_{\text{global}}(i) \rangle$  for the 3 classes: **A**erobic, **F**acultative and **A**naerobic. Each subplot shows the result for a different value of  $k$ . . . . . 137

- B.1 In (a), (c) and (e) we plot the relationship between three basic network measures and environmental variability for the 6 environmental classes. Note that here we plot the mean value over each environmental class: **O**bligate, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. In (b), (d) and (f) we show the three basic network measures plotted against the global significance score  $P_{\text{global}}$  for the 115 bacterial networks. . . . . 142
- B.2 In (a), (c) and (e) we plot the relationship between three basic network measures and growth conditions (oxygen requirements) averaged over the 3 classes: **A**erobic, **F**acultative and **A**naerobic. In (b), (d) and (f) we show the three basic network measures plotted against the global significance score  $P_{\text{global}}$  for the 383 bacterial networks. . . . 143
- B.3 In (a) we plot the relationship between the genome size (total number of genes) and environmental variability for the 6 environmental classes. Note that here we plot the mean value over each environmental class: **O**bligate, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. In (b) we show the total number of genes plotted against the global significance score  $P_{\text{global}}$  for the 115 bacterial networks. . . 144
- B.4 In (a) we plot the relationship between the genome size (total number of genes) and growth conditions (oxygen requirements) averaged over the 3 classes: **A**erobic, **F**acultative and **A**naerobic. In (b) we show the total number of genes plotted against the global significance score  $P_{\text{global}}$  for the 383 bacterial networks. . . . . 144

# LIST OF TABLES

2.1	The average path-length and clustering coefficient of several real-world networks. For each network we provide the number of vertices ( $n$ ), the average degree ( $\langle k \rangle$ ), the average path-length ( $\ell$ ) and the average clustering coefficient ( $\langle C \rangle$ ). We also provide the average path-length ( $\ell_{\text{rand}}$ ) and average clustering coefficient ( $\langle C \rangle_{\text{rand}}$ ) for the corresponding random graphs. Note that these random graphs have the same number of vertices and average degree as the real network. .	28
4.1	Network motifs obtained for the 120 network models and 115 metabolic networks using the new algorithm. . . . .	62
4.2	Network statistics for the reaction graphs of the 115 bacterial species studied in this work classified according to environmental variability. According to the NCBI, obligate bacteria have the most constant environment, followed by specialised and aquatic, and then facultative, multiple and terrestrial bacteria in that order. . . . .	63
4.3	Network statistics for the reaction graphs of the 383 bacterial species studied in this work classified according to species' oxygen requirements. The degree of oxygen required increases in the order anaerobic, facultative and aerobic. . . . .	64
4.4	Motifs significantly overrepresented in networks pertaining to a specialised and varied environment. . . . .	68

---

4.5	Motifs significant to networks with differing oxygen requirements. . .	73
5.1	Network statistics for the reaction graphs of the 115 bacterial species studied in this work, classified according to environmental variability.	92
A.1	Frequencies of the significant motifs obtained for the specialised and varied environmental classes after 1000 approximations of $H$ using non-negative matrix factorisation. . . . .	135
A.2	Frequencies of the significant motifs obtained for the aerobic and anaerobic classes after 1000 approximations of $H$ using non-negative matrix factorisation. . . . .	135

# LIST OF ABBREVIATIONS

[ER]	Erdős-Rényi .....	5
[WS]	Watts and Strogatz .....	5
[SW]	Small world .....	5
[BA]	Barabasi-Albert .....	6
[PPI]	Protein-protein interaction .....	7
[GCC]	Giant connected component .....	34
[ODE]	Ordinary differential equation .....	38
[FBA]	Flux balance analysis .....	41
[KEGG]	Kyoto Encyclopedia of Genes and Genomes .....	45
[GSCC]	Giant strong connected component .....	49
[GWCC]	Giant weak connected component .....	49
[NCBI]	National Centre of Biotechnology Information .....	50
[SF]	Scale-free .....	60
[KW]	Kruskal-Wallis .....	64

# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	A Brief Historical Background . . . . .	4
1.2	Networks in Biology . . . . .	7
1.3	Overview of Thesis . . . . .	9
1.4	Publications and Presentations . . . . .	11
<b>2</b>	<b>Preliminaries in Network Science</b>	<b>13</b>
2.1	An Introduction to Graph Theory and Complex Networks . . . . .	13
2.1.1	Types of Networks . . . . .	13
2.1.2	Adjacency Relations in Networks . . . . .	17
2.2	Network Measures . . . . .	19
2.3	Random Graph Models . . . . .	24
2.4	Further Related Topics in Network Science . . . . .	31
2.4.1	Network Motifs . . . . .	31
2.4.2	Percolation and Network Robustness . . . . .	34
2.5	Summary . . . . .	35
<b>3</b>	<b>Metabolic Networks</b>	<b>36</b>
3.1	The Metabolic Process . . . . .	36
3.2	Mathematical Models of Metabolism . . . . .	38
3.2.1	Kinetic Models . . . . .	38
3.2.2	Steady-State Models . . . . .	40
3.2.3	Topological Models . . . . .	42
3.3	Metabolic Network Construction Used in This Study . . . . .	45

3.3.1	The KEGG Database . . . . .	45
3.4	Bacterial Metabolic Data Used in This Study . . . . .	50
3.4.1	Habitat Lifestyle . . . . .	50
3.4.2	Oxygen Requirements . . . . .	51
3.5	Summary . . . . .	52
<b>4</b>	<b>A Novel Approach to Network Classification Based on Network Motifs</b>	<b>53</b>
4.1	Current State-of-the-Art . . . . .	53
4.2	A New Graph Embedding Approach . . . . .	55
4.2.1	Motif Frequency Vectors . . . . .	55
4.2.2	Global and Local Motif Significance Scores . . . . .	58
4.2.3	Detection of Network Motifs and Choice of Null Model . . . . .	59
4.3	Test on Synthetic Data . . . . .	60
4.4	Tests on Directed Metabolic Networks . . . . .	63
4.4.1	Environmental Variability Promotes Network Complexity . . . . .	64
4.4.2	The Effect of Oxygen Requirement on Network Structure . . . . .	66
4.4.3	Motifs Responsible for the Observed Differences . . . . .	67
4.5	Summary . . . . .	78
<b>5</b>	<b>Hypernetwork Models of Metabolism</b>	<b>80</b>
5.1	Complex Hypernetworks . . . . .	80
5.1.1	Hypernetwork Preliminaries . . . . .	82
5.1.2	Current Complex Hypernetwork Measures . . . . .	84
5.2	Novel Measures of Metabolic Hypernetwork Robustness . . . . .	91
5.2.1	Hypernetwork Reciprocity as a Measure of Chemical Equilibrium . . . . .	93
5.2.2	Percolation in Complex Metabolic Hypernetworks . . . . .	95
5.3	Summary . . . . .	105
<b>6</b>	<b>Summary and Outlook</b>	<b>106</b>
6.1	Summary . . . . .	106
6.1.1	Classification of Biological Networks . . . . .	107

---

6.1.2 Hypernetwork Models . . . . .	109
6.2 Outlook . . . . .	110
<b>Bibliography</b>	<b>112</b>
<b>A Non-Negative Matrix Factorisation</b>	<b>133</b>
<b>B Additional Properties of the Metabolic Networks</b>	<b>138</b>



# CHAPTER I

## INTRODUCTION

In the last decade, the area of complex networks has developed significantly, due primarily to the availability of genome scale data and to advances in computational techniques. The field is inter-disciplinary by nature and joins together such diverse areas as physics, mathematics, biology, social sciences and many more. The Internet, social and biochemical networks provide just a few examples of networks that have been widely studied to date [1, 2, 3]. Remarkably, despite the differences that exist between these networked systems, important overarching rules have emerged: scale-free [4] and small-world [5] phenomena (the definitions of which are given in Chapter 2), for example. Moreover, the topology of real-world networks has been shown to deviate significantly from random graphs, such as those studied by Erdős and Rényi, rather, real-world network organisation tends to lie somewhere between completely random and regular lattice-like structures.

### 1.1 A Brief Historical Background

Although the emerging field of complex networks witnessed its major developments in the last 15 years, its origin actually dates back to as early as the 18th century. Indeed, the study of networks traditionally belongs to a branch of discrete mathematics called *graph theory*. The story began when Leonhard Euler, a Swiss mathematician, published his solution to the famous Königsberg bridge problem in 1736. It was this

paper that motivated the analysis of real problems using graph structures, leading to substantial developments in graph theory [6, 7, 8].

Since Euler many other mathematicians have contributed to graph theory, however their main focus was in determining the properties of small, structured graphs, such as  $k$ -regular graphs, trees or lattice graphs. It was not until the 1950s that Paul Erdős and Albert Rényi, two Hungarian mathematicians, diverted the attention towards *random graphs*. In their seminal 1959 paper [9] they set about studying large (possibly infinite) irregular graphs. Importantly, this work led to the now famous Erdős-Rényi (ER) random graph model (cf. §2.3) that is still widely used to this very day as a null model for comparative purposes when studying real-world networks.

In the 1950s, 60s and 70s graph theory gained further prominence through its use in a number of different social experiments, perhaps the most famous of which were those by the social psychologist Stanley Milgram, that led to the introduction of the small-world (SW) concept [10]. Other notable work in this area includes that of Katz [11], Freeman [12] and Zachary [13]. The work of Milgram, in particular, has led to a number of references in popular culture to the small-world phenomena, such as The Oracle of Bacon (<http://oracleofbacon.org/>), six-degrees of separation, and even to the saying of ‘it’s a small world’.

In the late 1990s the study of complex networks took a major step towards its current status. The area received an explosion of renewed interest and research, and scientists from a wide variety of disciplines started using network models to describe a range of diverse systems. Notably, it was the pioneering work described in two seminal papers that triggered this renewed interest. The first was that of Watts and Strogatz (WS) [5], who, motivated by the SW property, proposed a simple, yet remarkable model that interpolated between a regular ring lattice and a classical random graph. Using the *clustering coefficient* and *characteristic path-length* (the definitions of which are given in Chapter 2) to quantify ‘small-worldness’, they were able to show that not only did this new model exhibit the SW property, but it was comparable to a number of real-world networks that have a small average path length and high clustering coefficient (e.g. neural network of *C. elegans*, the US power grid and a network of Hollywood film collaborations), in contrast to the traditionally

used random graph models.

The next big-step forward came just a year later. Physicists Albert Barabási and Réka Albert (BA) noticed what they deemed to be a universal property amongst complex systems, and that was that many of these systems exhibited a decentralised structure, which deviates significantly from the structure of both the random ER graph model and the new model of Watts and Strogatz. Instead, these networks were found to consist of a few highly connected vertices which they labelled *hubs*, along with many vertices with very few connections. Such a structure is something of a recurring theme in physics and so Barabási and Albert immediately recognised that the *degree* structure of such networks must follow a power-law distribution. Immediately, they set about constructing a model capable of reproducing this structure, and the result was the now famous *preferential attachment* model (a.k.a. the BA model) [4].

Motivated by these unexpected discoveries, a plethora of new network measures and models have been proposed [14, 15], of which, a few key concepts stand out due to the dominant role they have played in the field. *Centrality measures*, for example, that provide a means of ranking network nodes according to their ‘importance’ within the network, have revealed a number of interesting results concerning network resilience, stability and functionality [16, 17]. For example, many real-world networks have been found to be resilient against random failures, yet are vulnerable to targeted ‘attacks’ on the ‘important’ vertices [18, 19, 20], and thus these networks display a surprising level of robustness. An additional feature of many real-world systems is that they exhibit a so-called *modular structure*, that is, the network is organised into groups of densely interconnected vertices with sparse connections between them [21, 22, 23]. A number of algorithms exist for estimating the modularity of a network, the most common of which is the Newman-Girvan algorithm [24, 25]. Modularity is particularly important to biological networks, since it is considered to be one of the major underlying principles of these systems. Another important concept, especially for biological networks, is that of *network motifs*, that is, small recurring patterns of interconnected vertices across a network. Importantly, network motifs are often considered to be the “building blocks” of real complex networks, and a number of studies have evidenced significant relationships between network

motifs and functional and evolutionary properties of cellular networks [26, 27, 28].

Despite the many advances made over the last 15 years or so, the area of network science is still in its infancy. Many open questions remain and a number of controversies need readdressing, for example, previous results suggest that the universality of the scale-free phenomena may actually be a result of network misrepresentation, either due to inappropriate models or due to a lack of confidence and coverage in network data – particularly in the biological sciences, due to genome databases containing missing and erroneous annotations [29, 30, 31, 32]. Thus, an important focus of future research in network science is in the development of more ‘realistic’ network models, e.g. hyper-graphs [33], temporal networks [34], multiplex networks [35], etc. Also, despite the vast array of network measures that have been proposed to date, a strict framework for characterising a network via its structural properties remains out of reach. Clearly, network classification has considerable applications in biology and beyond, and thus presents an important future challenge for researchers working in network science. Future progress in these areas promises to provide greater understanding and insight into the behaviour of a number of diverse complex systems, which are currently far from being fully understood.

## 1.2 Networks in Biology

The focus of this piece of work is biological networks, which can range from cellular networks, the most important of which are protein-protein interaction (PPI) networks, gene regulatory networks and metabolic networks, to neuronal networks (across varying scales), disease networks and many more. Typically, a network description provides a greatly simplified view of the system at hand, focusing upon interaction patterns between a usually very large number of similar ‘units’, e.g. in a PPI network the units are proteins and we say that two proteins interact if they can physically bind to each other. Thus, in this section we briefly describe some of the most common networks arising in biology.

**Cellular networks:** Recent developments in high throughput techniques have resulted in a large amount of cellular data becoming available for a variety of organ-

isms [36, 37, 38, 39]. This cellular data describes molecular interactions between genes, proteins and small molecules that take place within the cell. There are three cellular networks that have been of particular interest in the literature: *transcriptional regulatory networks* which describe the regulations occurring between genes [40], *metabolic networks* which describe the entire set of biochemical reactions occurring within the cell in order to transform one compound into another [41], and *protein-protein interaction networks* which describe the physical interactions between proteins that bind together [42]. In such networks vertices represent the biological entities, that is the genes, metabolites, proteins, etc., and edges represent the functional, chemical or physical interactions between these molecules.

**Neuronal networks:** Graphical representations of neuronal networks provide a new framework for analysing the anatomical and functional structure of the brain [43]. At the cellular scale the network is formed via inter-neuronal connectivity, that is, the vertices denote individual neurons and edges denote synapses. The topology of the neuronal network of the roundworm *C. elegans* has been studied at the microscopic level (see for example [44]). However, when considering more complex species, the number of neurons and connections becomes unmanageable, and coupled with a lack of available data at this level, means that most neuronal networks are studied at a larger scale, e.g. inter-regional connectivity. In that case the vertices denote brain regions and edges denote axonal (either individual or bundles) projections.

**Disease networks:** Disease networks [45, 46, 47] provide a framework for analysing associations between certain (disease) genes and a range of different medical disorders, and can therefore help to unravel the genetic evolution for a range of complex diseases. More specifically, the disease network is a bipartite graph (see §2.1), where one set of vertices denotes the known genetic disorders and the other set denotes the known disease genes. A gene and disorder are connected by an edge if the disorder is implicated by a mutation of that gene.

**Food webs:** Another well studied example of a biological network is the food web [20, 48, 49, 50]. These networks describe the feeding relationships between species in a particular habitat. That is, the species in a habitat are either eaten by or eat

another specie, and thus form a large complex network of predator-prey interactions. A directed graph is often used to represent the predator-prey interactions, where the vertices denote species and the edges are directed from the prey to the predator.

## 1.3 Overview of Thesis

In this thesis, we employ novel network techniques to investigate relations between metabolic network structure and environmental variability for a large class of bacterial species. We propose a new form of graph embedding that uses so-called motif feature vectors alongside data-mining techniques to provide a low-dimensional representation of the data, thus greatly improving our ability to analyse large numbers of networks of possibly differing order. In particular, we find strong evidence for increased complexities within those metabolic networks that have evolved in more variable environments. We then move on to consider hypernetwork models of metabolism, and introduce a number of new concepts including an extension of *network reciprocity* and *site percolation* to this more complicated setting. Again we find strong correlations between environmental variability and metabolic network structure, and perhaps more importantly, we find that many of these results are not supported when a standard network analysis is performed.

We start in Chapter 2 by providing an overview of some of the fundamental definitions and theories within the area of network science. This includes different network representations, as well as a look at some of the most important network measures, and a number of different network models. These tools and techniques have been used to characterise a large variety of different real-world networks and will prove invaluable in what follows. Additionally, we provide background on two specific areas of network science that will be used in our work: *network motifs* and *percolation theory*.

In Chapter 3 we provide an overview of metabolic modelling, focussing on metabolic networks. We begin by discussing the metabolic process from a biological point-of-view before discussing metabolic modelling in general. We then focus on metabolic networks and discuss the pros and cons of the different network approaches available. We then go on to describe the process of constructing metabolic network

models from the raw biological data. We then conclude the Chapter with a discussion on the classification of different bacterial species according to environmental variability, a concept of key importance to the investigations of this thesis.

In Chapter 4 we present a new network classification technique. This new approach is based on the concept of graph embedding, and uses so-called motif feature vectors in order to embed a large number of networks into a low-dimensional space. We illustrate the new technique on a number of synthetic networks including small-world, scale-free and random graphs, as well as a large cohort of bacterial species. Importantly we are able to show that the network motif structure of the bacterial metabolic networks correlates with both environmental variability and changing growth conditions. Moreover, we are able to determine exactly which motifs are driving the results and can then attempt to explain the underlying biology behind the structural differences observed. Further evidence of the added value of the new approach follows from the observation that the aforementioned correlations disappear when other, more standard network measures are considered.

In Chapter 5 we leave behind the standard network approach and instead consider so-called hypernetwork models of metabolism. To begin we provide an overview of the current state of the art in hypernetwork modelling of real-world systems, and define a number of complex hypernetwork analogues (e.g. degree, clustering coefficient, etc.) of the standard network measures. Importantly, a number of results fail to hold in this more general setting and we shall illustrate this using a number of examples. Next, we consider an extension of the *reciprocity* measure to the case of hypernetworks and demonstrate its usefulness on a cohort of 115 metabolic hypernetworks. Finally, we introduce a percolation strategy, adapted to the hypernetwork framework, and use this in order to determine the relationship between environmental variability and metabolic hypernetwork robustness.

We conclude in Chapter 6 by summarising the work presented in this thesis and by discussing possible avenues for future work.

## 1.4 Publications and Presentations

The material presented in Chapter 4 as well as some of the material discussed in Chapter 3 has been written as an article:

*Network motif frequency vectors reveal evolving metabolic network organisation*, N Percy, J J Crofts and N Chuzhanova, *Molecular BioSystems*, 2015, **11**(1), 77–85.

This work was chosen to feature on the front cover of the January 2015 issue of the journal *Molecular BioSystems* (see Figure 1.1).

The material presented in Chapter 5 has been written as two separate articles: a conference proceedings

*Hypergraph models of metabolism*, N Percy, J J Crofts and N Chuzhanova, *International Journal of Biological, Veterinary, Agricultural and Food Engineering*, 2014, **8**(8), 732–736;

and a journal paper

*Complexity and robustness in hypernetwork models of metabolism*, N Percy, N Chuzhanova and J J Crofts, in preparation (2015).

In addition, the material in Chapter 4 as well as some of the work discussed in Chapter 3 was first given as a presentation entitled “Classification of biological networks using graph theoretical techniques” at Nottingham Trent University, School of Science and Technology, 7th Annual Research Conference in May 2013, and then was given as a poster presentation at ICSB (International Conference on Systems Biology), Copenhagen, Denmark, September 2013.

The material presented in Chapter 5 and again some of the work discussed in Chapter 3 was given as a presentation entitled “Hypergraph models of metabolism” at ICBCSB (International Conference of Bioinformatics, Computational and Systems Biology), London, UK, August 2014.



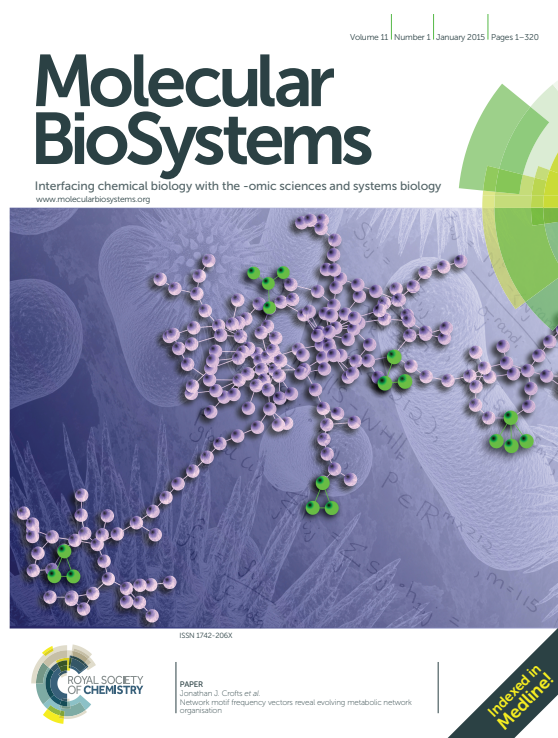


Figure 1.1: Front cover of the January 2015 issue of *Molecular BioSystems* featuring work from this thesis.

# CHAPTER II

## PRELIMINARIES IN NETWORK SCIENCE

### 2.1 An Introduction to Graph Theory and Complex Networks

This thesis is concerned with the mathematical modelling of complex biological systems through the use of graphs, or networks. In what follows we shall present both an empirical and mathematical overview of networks. We start by describing the variety of graph models that are used to represent complex networks, followed by an overview of the most common network measures that are used to characterise network topology. We then review the main network prototypes developed within the last decade or so in order to provide insight into the mechanisms underlying these intricate and complex systems. Finally, we discuss two topics in network theory that will be of great importance to our studies: network motifs and percolation theory.

#### 2.1.1 Types of Networks

In the language of graph theory, a network is represented by a pair  $G = (V, E)$ , where  $V = \{v_1, v_2, \dots, v_n\}$  is a finite set of nodes and  $E = \{e_1, e_2, \dots, e_m\}$  the set of

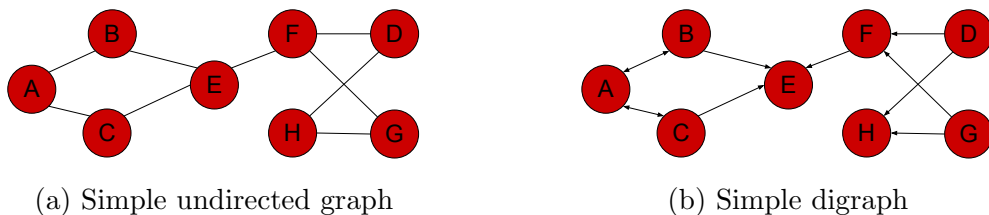


Figure 2.1: Illustration of undirected and directed network models.

edges representing connections between the nodes. Each  $e_i$  is made up of subsets of  $V$ , i.e.  $e_i \subseteq V$  for  $i = 1, 2, \dots, m$ . In the case of a *simple graph*, i.e. one free of loops and multiple edges,  $|e_i| = 2$ , where  $|x|$  denotes the cardinality of  $x$ . Such simple representations are common place in biology, where nodes are used to describe biological units of interest (e.g. genes, proteins or metabolites) and edges represent the different types of interaction that can occur amongst these units (e.g. regulation, binding and reactions). However, many biological systems are more naturally described using more complicated mathematical objects, e.g. directed and/or weighted networks, hypernetworks, multiplex networks, etc., the most important of which we describe below.

### Undirected Vs Directed Networks

Simple, undirected graphs are widely used to represent real world networks. This type of graph model, however, is only suitable when relationships between elements in a network are mutual (i.e. bi-directional,  $A \rightarrow B$  and  $A \leftarrow B$ ), which is often not the case for many real-world networks. Take for example, a reaction network [51]. Here we have the possibility of reversible and irreversible reactions, and so using an undirected graph would ignore the information regarding the directionality of the reaction. A few more examples where non-mutual relationships exist include the world wide web, food webs, and economic networks. These networks are more suitably represented by directed graphs, also referred to as *digraphs*. Vertices in directed graphs are connected by directed edges, or *arcs*. In Figure 2.1 we provide toy examples of both a simple undirected graph and a digraph.

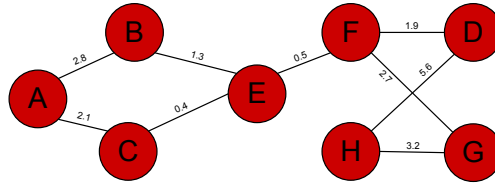


Figure 2.2: Illustration of a weighted graph model.

### Weighted Networks

Not only do some real world networks have non-mutual relationships but they can also have varying intensities, weights or strengths. To encapsulate this additional structure one can define a weighted network by considering a graph  $G = (V, E)$  along with an associated weight function  $w : E \mapsto \mathbb{R}$ , where  $\mathbb{R}$  denotes the real number line (see for example Figure 2.2). Almost all data emerging from the life sciences is weighted in some sense; however, it is worth noting that the majority of network studies tend to threshold the weights to obtain a standard binary description, as network measures for non-weighted graphs are simpler and more readily interpreted.

### Bipartite Networks

A bipartite (more generally multi-partite) graph can be either directed or undirected, and is a special instance of a simple graph in which the vertex set  $V$  admits a partition into two sets  $V_1, V_2$ , such that  $(u, v) \in E$  implies that either  $u \in V_1, v \in V_2$  or  $v \in V_1, u \in V_2$ . Applications of this type of graph have been used to represent both metabolic and disease networks, as discussed in [46] and [52].

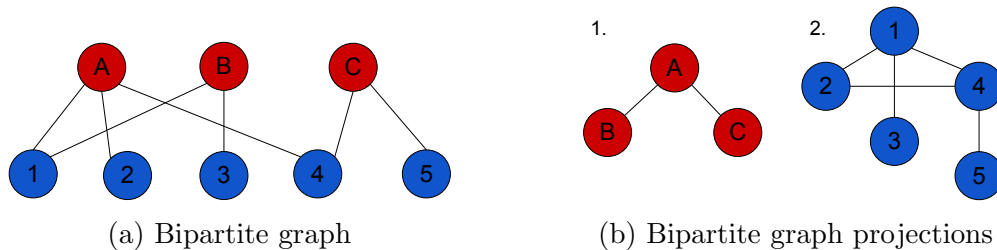


Figure 2.3: Illustration of a bipartite graph and two possible graph projections.

To apply standard network techniques to these graphs it is common to first project these bipartite models onto some simple, unipartite representation. Note

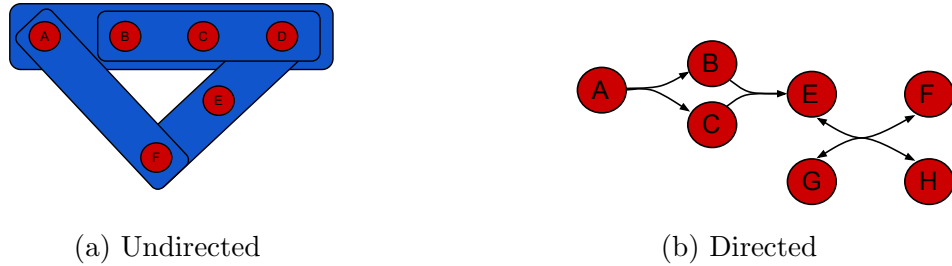


Figure 2.4: Illustration of undirected and directed hypergraph models.

however, that the projection of a bipartite graph is not unique, and that a variety of unipartite graph representations exist. Figure 2.3 shows an example of a bipartite graph together with two possible unipartite projections. This has led to a number of controversies in the literature. Universal topological features such as the small-world and scale-free phenomena, are now being questioned. In fact, a number of recent studies now argue that many of these properties are a consequence of inappropriate graph projections [53, 54, 55].

### Hypernetworks

In some real-world networks links can join more than two nodes at a time. For example, we might want to consider a metabolic network representing chemical reactions in the cell. Chemical reactions often involve more than one substrate and/or product and thus standard network models can prove inadequate in some instances, *e.g.* path analysis. Hypergraphs provide a solution to this problem by allowing edges to consist of two or more nodes, forming a hyperedge, and thus preserving the additional information (see Figure 2.4a for an illustration). More formally, a hypergraph consists of two elements: a vertex set,  $V$ , and a hyperedge set,  $E = \{E_1, E_2, \dots, E_m\}$ . Each  $E_i$  is made up of subsets of  $V$ , i.e.  $E_i \subseteq V$  for  $i = 1, 2, \dots, m$ . Note that  $2 \leq |E_i| \leq n$ . If in addition, one considers directed hypergraphs then each hyperarc is further subdivided into two sets - the tail set  $T$  and the head set  $H$ , allowing us to take into account directionality, as is the case in Figure 2.4b.

The importance of hypergraphs to biology, where quite often the functional relationships are extremely complicated and naturally multi-dimensional, has recently

been highlighted by a number of authors - see for example [33]; however, so far the network community has largely resisted, mainly due to the lack of available techniques for the analysis of these more complicated objects.

### 2.1.2 Adjacency Relations in Networks

The most common way to represent a complex network is in matrix form. Given a simple graph  $G$  then we may define the  $n \times n$  *adjacency matrix*  $A$  as follows

$$A_{ij} = \begin{cases} 1, & \text{if } i \sim j, \\ 0, & \text{otherwise.} \end{cases} \quad (2.1)$$

Here  $i \sim j$  denotes that nodes  $i$  and  $j$  are adjacent to each other. Two vertices are said to be adjacent (or neighbours) to one another if they are connected by an edge. Note that if the network is undirected then its adjacency matrix is symmetric since  $a_{ij} = a_{ji}$ . In the case that the network is weighted we may simply set  $a_{ij} = w_{ij}$ , where  $w_{ij}$  is the result of the weight function acting on the edge between nodes  $i$  and  $j$ .

Examples and matrix representations describing the aforementioned network models are given in Figure 2.5. Notice that the diagonal entries are zero in all cases since we do not allow self-loops.

From this representation it is easy to see that a number of important network measures can be obtained directly from the adjacency matrix. For example, it is easy to see that for the undirected network the vector of *degrees*  $\mathbf{k}$  is given by

$$\mathbf{k} = (\mathbf{e}^T A)^T = A^T \mathbf{e} = A \mathbf{e}, \quad (2.2)$$

where  $\mathbf{e}$  is the  $n \times 1$  vector of all ones. Recall that the  $i$ th term of the vector  $\mathbf{k}$  counts the number of edges incident to node  $i$ , that is, the number of non-zeros in the  $i$ th column, or row, of  $A$ . Of course, in the case of a directed network  $(\mathbf{e}^T A)^T \neq A \mathbf{e}$ , in general. Instead we have two different types of node degrees: the *in-degree*, which counts the number of links pointing towards a node, and the *out-degree*, which counts

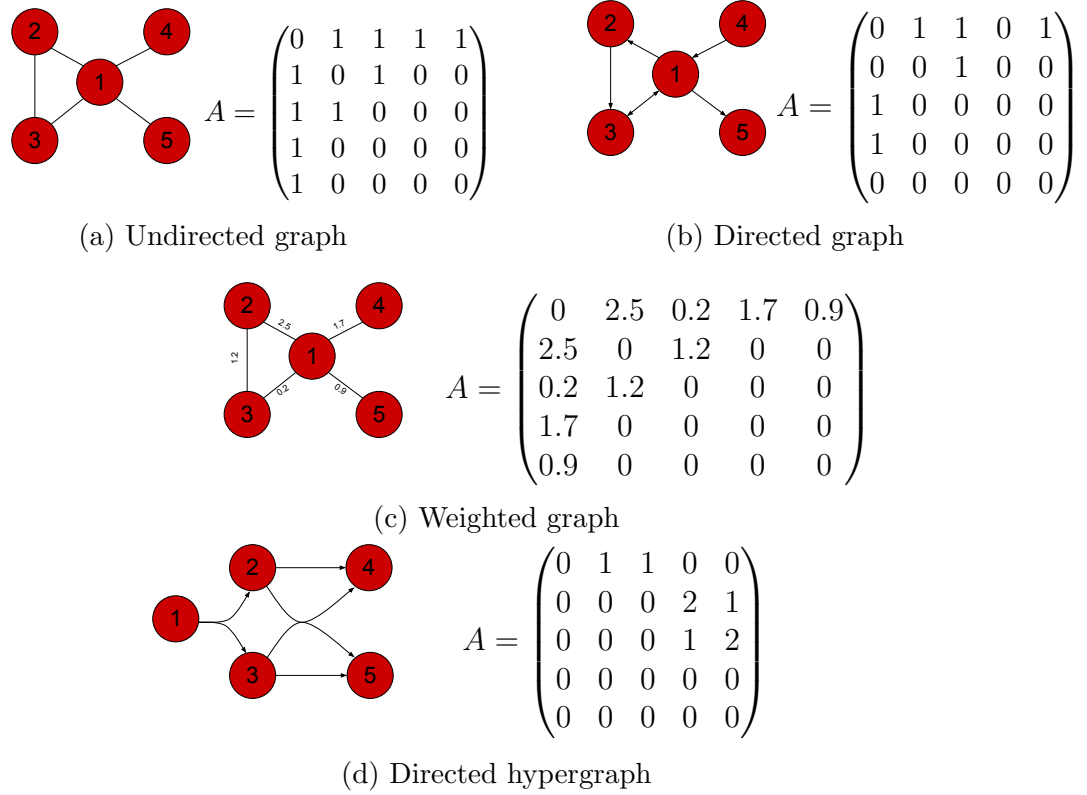


Figure 2.5: Illustration of the adjacency matrices associated with the different graph models.

the number of links leaving a node.

$$\mathbf{k}^{\text{in}} = (\mathbf{e}^T A)^T \quad \text{and} \quad \mathbf{k}^{\text{out}} = A\mathbf{e}. \quad (2.3)$$

An important result, closely related to Euler's famous *handshaking lemma* [56], is the following formula for the mean degree:

$$\langle \mathbf{k} \rangle = \frac{2|E|}{n}.$$

This formula can be easily derived from the adjacency matrix by noting the following identity

$$\mathbf{e}^T A \mathbf{e} = \sum_{i=1}^n k_i = 2|E|.$$

The above formula states that the sum of all entries of  $A$  is equal to twice the number

of edges in the network.

Perhaps the most useful property concerning adjacency matrices is their ability to count *walks* on a graph. A walk of length  $k$  starting at node  $i$  and ending at node  $j$  is any traversal through the network that follows  $k$ , not necessarily distinct edges. A useful observation is that the  $ij$ th element of the adjacency matrix raised to the  $k$ th power, i.e.

$$(A^k)_{ij} = \sum_{r_1=1}^n \sum_{r_2=1}^n \cdots \sum_{r_{k-1}=1}^n a_{i,r_1} a_{r_1,r_2} a_{r_2,r_3} \cdots a_{r_{k-2},r_{k-1}} a_{r_{k-1},j}, \quad (2.4)$$

counts the number of walks of length  $k$  that start at node  $i$  and finish at node  $j$  – this result is easily proved by induction [57]. Importantly, as well as providing an efficient means for computing graph invariants, e.g. the *clustering coefficient* (see §2.2), the idea of ranking nodes according to the number of walks passing through them has generated a whole new class of centrality measures [17, 58, 59].

## 2.2 Network Measures

The literature is awash with various network measures (see [14] for a review); however, three concepts in particular have played a leading role in the development of the field to date.

### Average Path-Length

A *path* is defined as a walk between two vertices, such that the vertices and edges within the sequence are all unique. The path between two vertices that has the shortest number of steps, or minimum weight as is the case in weighted graphs, is referred to as the *shortest path*. The total weight of the shortest path is then defined as the *distance*  $d_{ij}$  between the vertices  $i$  and  $j$ . To compute the average path-length  $\ell$  one then simply takes the average distance over all pairs of vertices in the network, such that

$$\ell = \langle d_{ij} \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}. \quad (2.5)$$

The above concept, however, is somewhat troublesome when applied to real world



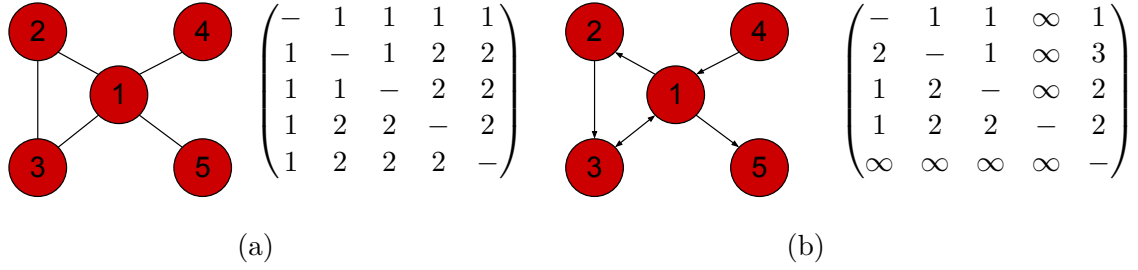


Figure 2.6: The average path-length and corresponding distance matrix for (a) an undirected graph ( $\ell = \langle D \rangle = 1.5$  or  $\varepsilon = 0.75$ ); and (b) a directed graph ( $\ell = \langle D \rangle = \infty$  and  $\varepsilon = 0.47$ ).

networks. Real world networks, for instance, often contain isolated components which result in non-existent paths between pairs of vertices in the network. A common approach to avoid this problem is to restrict the analysis using this measure to the giant connected component only. In addition to this, directed networks also result in the same problem, since it is possible that a path only exists in one direction. In this situation, one may wish to use an alternative but related distance measure referred to as the *efficiency* of a network

$$\varepsilon = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}. \quad (2.6)$$

Note that here  $d_{ij} = \infty$  if no path exists between the vertices  $i$  and  $j$ . Figure 2.6a provides an example of an undirected graph alongside its corresponding distance matrix containing the shortest path-lengths. The average path-length is obtained by taking the average of the matrix  $D$ ,  $\ell = \langle D \rangle = 1.5$ . Figure 2.6b provides an example of a directed graph alongside its corresponding distance matrix. The average path-length cannot be obtained for this directed graph, since some pairs of vertices are not connected by a path, therefore resulting in an infinite distance between them. In this case, we can only compute the efficiency of the network,  $\varepsilon = 0.47$ .

### Clustering Coefficient

Clustering describes the interconnectedness of a node's nearest neighbours. Consider, for example, a friendship network. Typically, lots of close ties tend to be formed in such networks, as, after all, you are more likely to know a friend of a friend than a randomly selected person from the network. Thus, friendship networks tend to exhibit high levels of clustering.

In their seminal work on ‘small-world’ networks [5], Watts and Strogatz introduced the *local clustering coefficient* to quantify the extent of a node's clustering. For a given node,  $i$  say, they considered the following ratio

$$C(i) = \frac{2C_3(i)}{k_i(k_i - 1)}. \quad (2.7)$$

Here,  $C_3(i)$  counts the number of triangles connected to node  $i$ , and the quantity  $k_i(k_i - 1)/2$  counts the number of connected triples centred on node  $i$  ( $k_i$  is the degree of node  $i$ ). Informally,  $C(i)$  is the ratio between existing edges connecting the nearest neighbours of node  $i$  and all possible edges between these nearest neighbours. Note that  $C_3(i)$  can be easily computed using the adjacency matrix as follows

$$C_3(i) = \frac{1}{2} A_{ii}^3.$$

The average value of the clustering coefficient is then given by

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^n C(i). \quad (2.8)$$

This quantity has been widely used, along with the characteristic (or average) path-length, in order to quantify the extent to which a network is considered as being ‘small-world’.

More generally, the clustering coefficient defined in Equation (2.7) measures the extent to which a single vertex forms cliques within a network, i.e. a fully connected subgraph on  $k_i + 1$  nodes. The value of  $C(i)$  is bounded between 0 and 1 (as is  $\langle C \rangle$ ), with a value of one implying that all the vertex's nearest neighbours are also nearest neighbours, whilst a score of zero means that none of a vertex's neighbours

are neighbours.

An alternative to the *global clustering coefficient* of Equation (2.8), which has often been applied to social networks, is the *transitivity coefficient* [60]. The transitivity coefficient measures the proportion of triangles present within the network, such that

$$T = \frac{3C_3}{P_2}, \quad (2.9)$$

where  $C_3$  counts the number of triangles and  $P_2$  the number of 2-paths (or, connected triples) present within the network. The factor of three accounts for the fact that each triangle contains three different connected triples.

In terms of the adjacency matrix, the number of triangles in a network is computed as

$$6C_3 = \text{Tr}(A^3),$$

where the trace of an  $n \times n$  matrix  $B$  is given by

$$\text{Tr}(B) = \sum_{i=1}^n b_{ii}.$$

To compute the number of connected triples we recall that the quantity  $(A^2)_{ij}$  counts all walks of length 2 starting from node  $i$  and terminating at node  $j$ , and also, that a connected triple is nothing other than a 2-path, i.e. a walk of length 2 starting and ending at distinct nodes. It follows that the number of connected triples is given by

$$2P_2 = \mathbf{e}^T A^2 \mathbf{e} - \text{Tr}(A^2).$$

The second term here discounts those walks that are not 2-paths (i.e. walks by which the start and end nodes are not distinct).

There are important differences between the two global clustering coefficients,  $\langle C \rangle$  and  $T$ . For example, the latter measure gives equal weight to each vertex, whereas the former gives equal weight to each triangle. Network transitivity is arguably the more intuitive measure since it gives equal weighting to low- and high-degree vertices. The average  $\langle C \rangle$ , on the other hand, weights low-degree vertices more highly, since these vertices will have a small denominator in Equation (2.8).

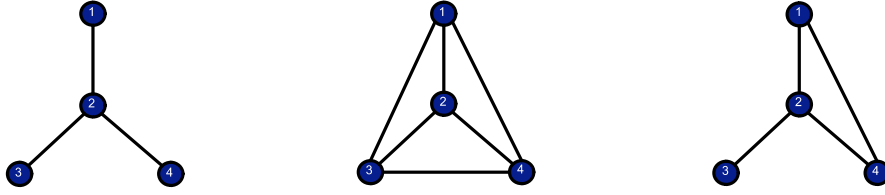


Figure 2.7: Local and global clustering as defined in Equations (2.7–2.8): (left)  $C = \{0, 0, 0, 0\}$  and  $\langle C \rangle = 0$ , (middle)  $C = \{1, 1, 1, 1\}$  and  $\langle C \rangle = 1$ , (right)  $C = \{1, 1/3, 0, 1\}$  and  $\langle C \rangle = 7/12$ .

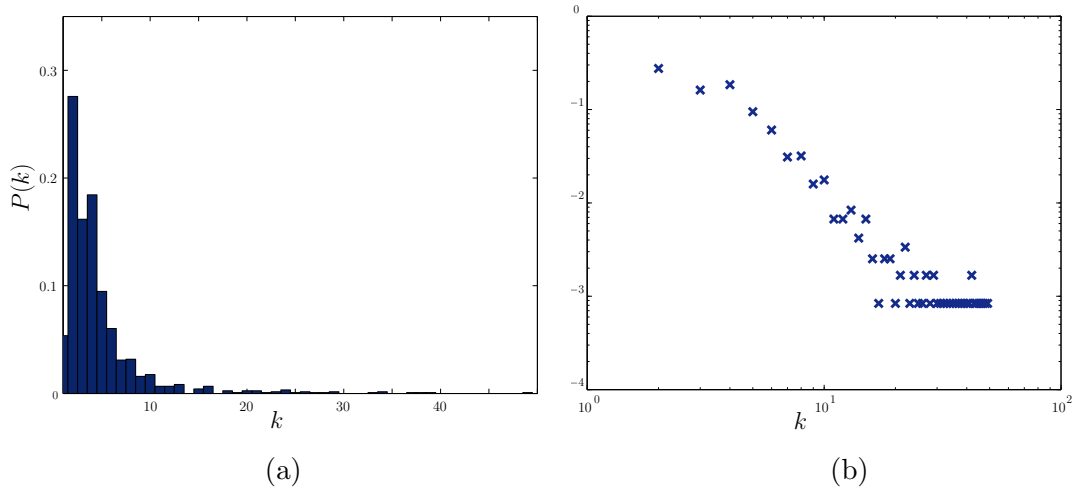


Figure 2.8: The degree distribution for the metabolic network of the bacteria *E.coli*: (a) histogram of  $P(k)$  versus  $k$ , and (b) a log-log plot of  $P(k)$  versus  $k$ .

Consider Figure 2.7, for example, here  $T = \{0, 1, 3/5\}$  whereas  $\langle C \rangle = \{0, 1, 7/12\}$ . In this example, the two measures coincide for the two extreme cases (left and middle) yet are different for the graph on the right.

We finish by noting that several authors have proposed generalisations of the above clustering coefficient to take into account directed or weighted networks for example, and the interested reader is directed to the following two examples [61, 62].

### Degree Distribution

The node degree vector,  $\mathbf{k}$ , provides a useful means for quantifying certain structural properties of networks. For example, one of the salient features of real-world networks are the broad differences seen in their *degree distributions* in comparison say, to popular random network models (cf. §2.3).

The degree distribution may naturally be considered as a probability distribution

as follows. Let  $n_k$  denote the number of nodes having degree  $k$ . Then the probability that a randomly chosen node has degree  $k$  is given by

$$P(k) = \frac{n_k}{n}, \quad (2.10)$$

where  $n$  is the size/order of the network. A plot of  $P(k)$  versus  $k$  represents the degree distribution of the network. Figure 2.8 provides an illustration of the degree distribution for a simple undirected network representation of metabolism in the bacterial species *Escherichia coli*.

Note that for directed networks we compute two such distributions: the in- and out-degree distributions as described in Equation (2.3).

## 2.3 Random Graph Models

A variety of simple network models have been proposed in the literature, the purpose of which is twofold. Firstly, they act as reference models, enabling us to investigate whether observed characteristics in real-world networks are significant, or if they could have arisen by chance in some simple random graph model. Secondly, network models can provide a mechanism to explain how certain characteristics have developed, and provide insights into a range of complex real-world systems. Here we provide a basic introduction into the common network models (for a more detailed review see, for example [15, 63, 64, 65, 66]).

### Classical Random Graphs

In their seminal work in the 1950's and 60's, Paul Erdős and Alfréd Rényi established the field of random graph theory, an area devoted to studying the properties of certain large, random graphs [9, 67]. The *Erdős-Rényi random graph* is a statistical ensemble whose members are all possible labelled graphs on  $n$  nodes and  $M$  edges, and is typically denoted  $G_{n,M}$ . Each member has an equal probability of occurring. At around about the same time, Edgar Gilbert an American mathematician was studying a related statistical ensemble of graphs [68]. Denoted by  $G_{n,p}$ , this new ensemble consisted of networks of order  $n$ , such that any two nodes are connected

with equal probability  $p$  (see Figure 2.9 for a small example). Remarkably, these two different random networks may be regarded as being equivalent for sufficiently large sparse networks [69].

Indeed, in practice, it is often the case that the two representations are considered equivalent, both being referred to as Erdős-Rényi (ER) random graphs. Typically, it is the Gilbert model  $G_{n,p}$  that is used since it is easier to analyse in general. Consider, as an example, the degree distribution of the Gilbert model. The probability that a node has degree  $k$  is given by

$$P(k) = C_k^{n-1} p^k (1-p)^{n-1-k}. \quad (2.11)$$

Here  $p^k$  is the probability that the node connects to  $k$  neighbours,  $(1-p)^{n-1-k}$  is the probability that the remaining  $n-1-k$  edges are not present, and  $C_k^n = n!/(n-k)!k!$  is the binomial coefficient. It follows immediately that the mean degree of a classical random graph is given by  $\langle k \rangle = p(n-1)$ . Furthermore, as  $n \rightarrow \infty$  the degree distributions, for both the Gilbert and ER models, tend towards that of a Poisson one:

$$P(k) = e^{-\langle k \rangle} \langle k \rangle^k \frac{1}{k!}. \quad (2.12)$$

Note, that we require  $\langle k \rangle$  to remain bounded in the above limit. As a result of the above, we see that classical random graph models are highly homogeneous, in the sense that most vertices have a ‘similar’ degree, being symmetrically distributed about  $\langle k \rangle$ .

Another important feature of classical random graphs is that they exhibit a small characteristic path-length. To see why this property arises, consider the following heuristic. For a classical random graph with mean degree  $\langle k \rangle$ , we have that each vertex within the network has approximately  $\langle k \rangle$  nearest neighbours. Now, extending to non-nearest neighbours, each vertex has  $\langle k \rangle^2$  second neighbours,  $\langle k \rangle^3$  third neighbours, and so on. Using this idea we can compute the number of edges, say  $\ell$ , required to reach every vertex in the network. That is, we want to find  $\ell$  for which  $\langle k \rangle^\ell$  is approximately equal to the total number of vertices. Thus, by equating  $\langle k \rangle^\ell$  to  $n$ , and solving for  $\ell$ , we find that the average number of edges required to reach

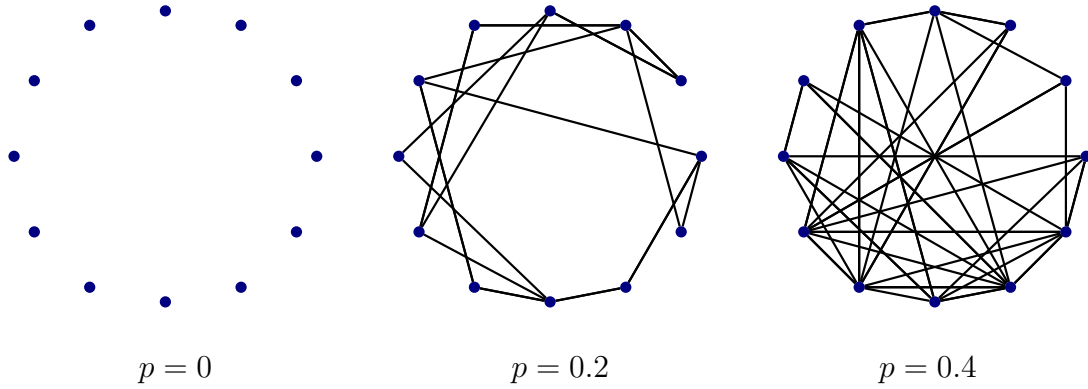


Figure 2.9: Classical random graph,  $G_{n,p}$ , with 12 vertices and varying probabilities ( $p$ ) for connecting pairs of vertices together.

any vertex in the network is approximately

$$\langle k \rangle^\ell \sim n \implies \ell \sim \frac{\log n}{\log \langle k \rangle}.$$

The clustering coefficient provides yet another example of why it is preferable to work with the  $G_{n,p}$  network model. In that case the probability that a node's neighbours are connected is always  $p$ , and thus

$$C = p \approx \frac{\langle k \rangle}{n}, \quad (2.13)$$

where the latter equation holds in the large  $n$  limit – recall that in this case we have that  $\langle k \rangle = p(n-1) \approx pn$ . In particular, the clustering coefficient for a large sparse random graph approaches zero. If we contrast this against reported measurements for many real-world networks, we see a difference of several orders of magnitude.

It is worth noting at this point that when studying real-world networks, we often need to quantify the relative importance of our observations by undertaking a comparative analysis against an appropriate null model – typically some form of random network model. As we shall see later on, classical random graphs are usually a poor choice of null model, and we should instead consider a statistical ensemble that conserves certain network properties, e.g. degree, density, etc. For example, the so-called *configuration model* [70] is a generalisation of the classical random graph model, which generates a random graph by selecting uniformly at random from an

ensemble of graphs with a prescribed in- and out-degree. We shall discuss such extensions at greater lengths in Chapter 4.

### Watts-Strogatz Model

In 1967 Stanley Milgram, a famous social psychologist, conducted an experiment that provided the first evidence of the so-called *small-world phenomenon* [10]. In this famous experiment, the motivation of which was to measure distances within acquaintance networks in the United States, Milgram posed the following question: how many intermediate links separate two randomly selected individuals? To answer this question, Milgram selected at random a large number of individuals living in Omaha, and sent them each a letter instructing them to forward it to a randomly selected individual in Nebraska, Boston, subject to the following rules:

1. if you know the target person by name then forward it immediately;
2. else mail the letter plus instructions to a person whom you know by first name and who is more likely to know the target person.

Surprisingly, it was found that of those letters that reached the target, the average number of links was  $l = 5.5$ . This result is well-known as the ‘six degrees of separation’.

The importance of this result follows from its ubiquity throughout a large number of networked systems across a range of disciplines, all of which exhibit a remarkably small characteristic path-length relative to system size,  $n$  (see for example [66, 71, 72, 73]). Table 2.1 provides a few examples whereby the characteristic path-length is small compared to the system’s size. Theoretically one considers a network to be ‘small-world’ if its path-length is proportional to the log of network size [74], that is

$$\ell \propto \log n. \quad (2.14)$$

Real-world networks, however, tend to display increased levels of clustering as well (see Table 2.1), and for this reason, it is common practice to classify a network as being small-world if it exhibits both a small characteristic path-length and a large amount of clustering, as compared to a ‘random’ network.



Network	$n$	$\langle k \rangle$	$\ell$	$\ell_{rand}$	$\langle C \rangle$	$\langle C \rangle_{rand}$
Metabolic network [41]	766	9.62	2.56	3.04	0.67	0.02
Protein interaction [19]	1879	2.40	6.8	7.85	0.1	0.05
Neural network [5, 75]	297	14.46	2.45	2.42	0.29	0.05
Coauthors (network science) [76]	1461	3.75	5.77	5.39	0.69	0.002
Power grid [5]	13188	2.67	18.99	8.22	0.8	0.00006
Karate club [13]	34	4.6	2.34	2.32	0.57	0.11
Word adjacencies [76]	112	7.67	2.51	2.47	0.17	0.06

Table 2.1: The average path-length and clustering coefficient of several real-world networks. For each network we provide the number of vertices ( $n$ ), the average degree ( $\langle k \rangle$ ), the average path-length ( $\ell$ ) and the average clustering coefficient ( $\langle C \rangle$ ). We also provide the average path-length ( $\ell_{rand}$ ) and average clustering coefficient ( $\langle C \rangle_{rand}$ ) for the corresponding random graphs. Note that these random graphs have the same number of vertices and average degree as the real network.

The above discussion leads naturally to a certain group of real-world networks that are classified by the presence of two important structural properties: (i) small average path-length and (ii) a large clustering coefficient. Classical random graphs display a very small average path-length, however, they exhibit extremely low levels of clustering, making them poor models of such systems. At the other extreme are regular, lattice like graphs, which display very high levels of clustering but relatively large average path-lengths, e.g. the average path-length of a  $n_d$  dimensional lattice goes as

$$l \sim n^{1/n_d}. \quad (2.15)$$

In their seminal work, Watts and Strogatz [5] had the ingenious idea to construct a random network model that interpolates between a regular and random structure, thus leading to a model that attains both required properties simultaneously. Starting with a  $k$ -regular lattice on  $n$  vertices (e.g. a ring in 1-d), each edge is rewired with probability  $p$ , such that each end vertex is reconnected with a different randomly chosen vertex – see Figure 2.10 for an illustration. For small values of  $p$  this process has very little effect on network clustering (recall we start off with a lattice ( $p = 0$ )), however, the addition of these random ‘short-cuts’ has a non-linear effect on  $\ell$  since it affects not only the nearest neighbour structure, but also the second neighbours, third neighbours and so on. The result is a random graph model with both the SW property and a high clustering coefficient. For large values of  $p \approx 1$

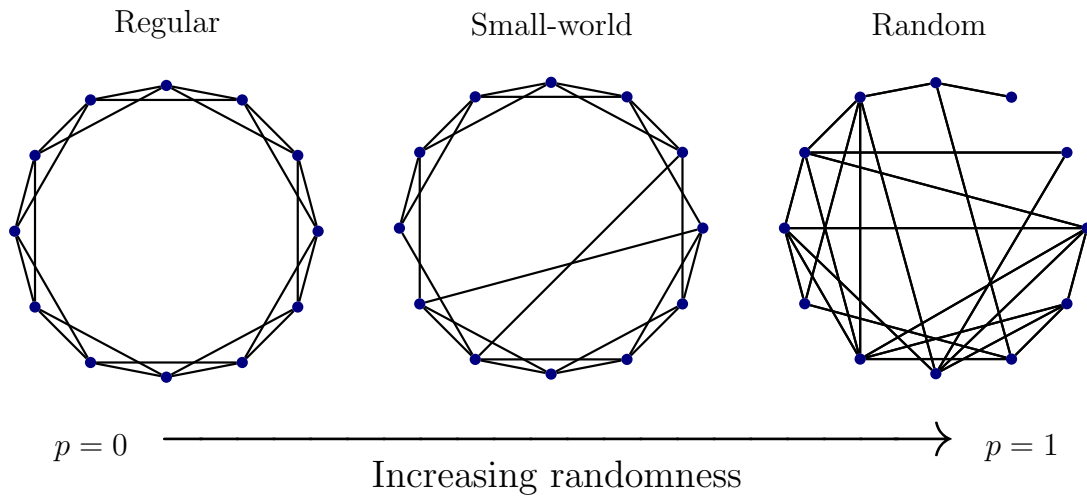


Figure 2.10: The Watts-Strogatz model. Here we start with a regular graph with 12 vertices and 24 edges, such that each vertex is connected to its nearest neighbours, leading to a maximum clustering coefficient ( $C = 1$ ). Each edge is randomly rewired by some probability  $p$ . As  $p \rightarrow 1$  the initial regular graph is transformed into an ER random graph (right). For intermediate values of  $p$ , such that  $0 < p < 1$ , a small-world network arises - highly clustered like a regular graph, yet with a small average path-length like a random graph.

we obtain a classical random graph.

It is worth noting here that Newman and Watts [77] proposed a slight variation to the WS model, in which new edges were added at random, rather than the usual rewiring of edges. This model has the advantage that it remains connected at all times, which greatly simplifies the resulting analysis – an extensive review of both models is provided in [78].

### Barabási-Albert Model

A limitation of classical random graph models and WS model is that unlike real-world networks the vertices are highly homogeneous in terms of their connectivity (or degree). Instead, many real-world networks have been found to have highly heterogeneous vertices, whereby a few vertices are highly connected, referred to as the *hubs* in the network, whilst many vertices have just a few connections [41, 63, 79, 80]. This heterogeneous set up results in a degree distribution with no scale, and thus networks found to be organised in this manner are characterised as *scale-free* networks [81]. More importantly, the degree distribution of scale free networks are

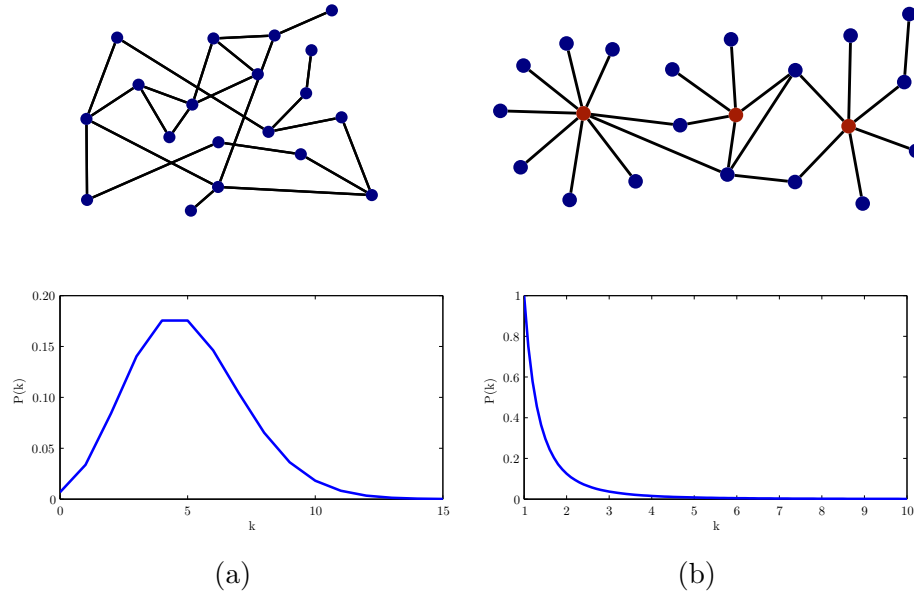


Figure 2.11: Degree distributions of two graph models: (a) a classical random graph (top) together with its degree distribution (bottom); and (b) a scale-free graph (top) together with its degree distribution (bottom).

found to follow a power law (for large  $k$ ) such that

$$P(k) \sim k^{-\gamma}, \quad (2.16)$$

where  $\gamma$  is the scaling parameter – note that such a power law is termed scale free since  $f(\alpha x) = \alpha^n x^n \sim x^n$ , i.e. all such power laws are equivalent up to constant factor<sup>1</sup>. Figure 2.11 shows an example of a scale-free network together with its degree distribution. Notice that the hubs are highlighted in red.

Barabási and Albert proposed a network model that possesses this scale-free degree distribution [4]. Their model starts at time  $t_0$  with a network consisting of  $N_0$  vertices, and  $m_0$  edges between them. At each time step  $t_{i+1}$  a new vertex is introduced to the network, and becomes connected to  $m \leq |N_0|$  vertices present in the network at time  $t_i$ . Each vertex already present in the network at time  $t_i$  has the following probability  $p(i)$  of receiving a connection to the new vertex

$$p(i) = \frac{k_i}{\sum_j k_j}.$$

<sup>1</sup>Actually power laws are examples of homogeneous functions and thus are naturally scale-free

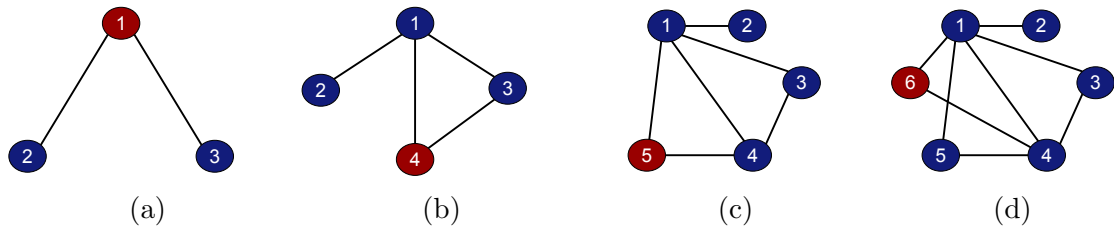


Figure 2.12: The Barabasi-Albert growth model. (a) We start with a graph with  $N_0 = 3$  vertices and  $E_0 = 2$  edges. (b) A fourth vertex preferentially attaches itself to 2 out of the 3 existing vertices, where the probability that an existing vertex receives an edge is proportional to its degree ( $\mathbf{p} = [0.5, 0.25, 0.25]$ ). (c) A fifth vertex now preferentially attaches itself to 2 out of the 4 existing vertices with probabilities  $\mathbf{p} = [0.375, 0.125, 0.250, 0.250]$ . (d) Finally, a sixth vertex attaches to 2 out of the 5 existing vertices with probabilities  $\mathbf{p} = [0.333, 0.083, 0.167, 0.250, 0.167]$ . Notice here that we have arrived at a network that is highly heterogeneous in terms of the degrees of the vertices.

In other words, the probability that a vertex (*i* say) receives a new edge is proportional to its degree. An illustration of the growth process is shown in Figure 2.12. This approach follows the *preferential attachment rule*, whereby new vertices introduced to the network prefer to attach to the vertices in the network that are already highly connected. This idea is commonly referred to as the ‘rich get richer’ phenomenon, where it is easy for the highly connected vertices (rich) to become even more connected, and very difficult for the weakly connected vertices (poor) to receive connections. Importantly, the Barabási-Albert (BA) model provides us with a possible explanation as to why scale-free degree distributions are so ubiquitous throughout science and technology.

## 2.4 Further Related Topics in Network Science

To complete our brief primer on Network Science we finish by giving an overview of two specific topics of particular interest to the work presented in this thesis.

### 2.4.1 Network Motifs

Network motifs are patterns of interconnected vertices that recur across a network a significantly greater number of times than they would in an ensemble of random

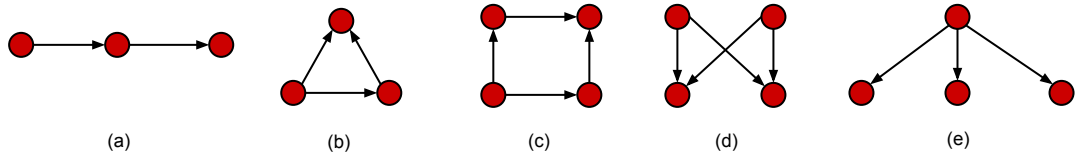


Figure 2.13: Examples of network motifs: (a) three-chain motif, (b) feed-forward loop motif, (c) bi-parallel motif, (d) bi-fan motif, and (e) single-input motif.

graphs. They are often considered to be the basic building blocks for real-world networks, and thus are likely to contain important information regarding the underlying network architecture.

Mathematically a *motif* is defined to be a subgraph that appears more frequently in real-world networks than it does across a suitably defined null model (*e.g.* Erdős-Rényi random graph). Figure 2.13 provides an example of several 3- and 4-node motifs. Null models are typically constructed so as to preserve important network features. For example, given a network  $G$  denote by  $\Omega(G)$  the family of randomised (see below) graphs of  $G$ , sharing the same number of vertices and the same degree distribution. Given such an ensemble, it is straightforward to compute a z-statistic and determine statistical significance, as we can sample repeatedly from the ensemble and compute an average motif frequency as well as computing its spread (see Chapter 4 for further details).

Biological networks, in particular, have been found to have a number of specific motifs that have been linked to functionality. The feed forward loop, for example, is thought to be important in transcriptional regulatory networks, whereby they serve as either sign-sensitive accelerators (*i.e.* speeding up the response time of the target gene expression) or sign-sensitive delays (*i.e.* delaying the response time) [27, 82]. An over-abundance of a particular motif in a real-world network is thought to be a consequence of the evolutionary process, and the resultant motif is likely to exhibit desirable structural and/or functional properties [83, 84, 85, 86, 87]. Indeed, we shall see further evidence for this hypothesis in later chapters.

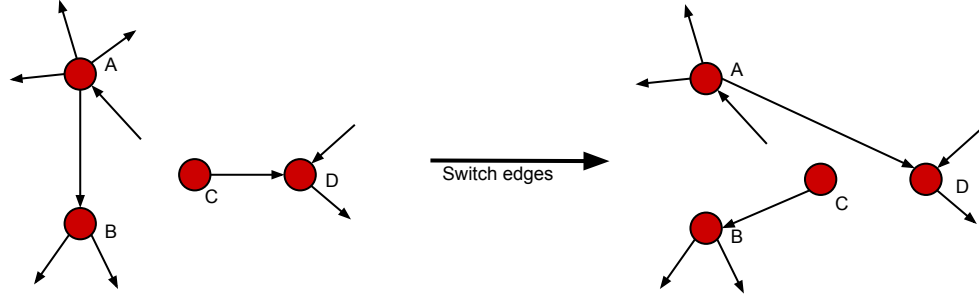


Figure 2.14: Example of the rewiring step of the switching method. The two edges  $E_{A,B}$  and  $E_{C,D}$  are reconnected to make  $A$  connected to  $D$  and  $C$  connected to  $B$ .

### Generating Randomised Null Models in Motif Analysis

As mentioned above, a variety of different random graph models can be used for statistical testing of network motifs; however, the choice of model can highly influence the result. All randomised networks should at least preserve some shared structural property of the real network, such as the number of vertices and the number of edges. ER random graph models could be used in this instance. However, it is argued that a more accurate null model should preserve not only the number of vertices and edges but also the degree distribution of the real network [88]. By doing so we guarantee that the significance of the motifs is not just a consequence of this underlying structure. In addition, it is also argued that the null model should preserve the number of  $(q - 1)$ -node motifs, where  $q$  is the size/order of the motif. This is to ensure that an over-abundance of a certain motif has not just occurred due to it being a substructure of a larger highly significant motif [83]. A common randomisation method used to preserve the aforementioned properties is the so-called *switching method* [89]. Here, two edges are randomly selected, let's say  $E_{A,B}$  and  $E_{C,D}$ . The two edges are rewired (or exchanged) to reconnect vertex  $A$  to vertex  $D$  and vertex  $C$  to vertex  $B$ . However, the exchange only takes place if the reconnection does not alter the required properties (see Figure 2.14). This process is repeated until the network is randomised. Note that rigorous methods for determining the point by which the network becomes fully randomised do not exist. In general, however, the process is repeated  $Q \cdot M$  times, where  $M$  is the number of edges and  $Q$  is chosen to ensure that the number of exchanges that actually take place is large [89].

### 2.4.2 Percolation and Network Robustness

One of the earliest and simplest processes studied in complex networks is *percolation* - the process by which a fraction of nodes together with their adjacent edges are removed (or added) from a network [90]. This process has been used to model a number of real world phenomena [16, 18, 91, 92, 93]. Consider, as an example, a social network used to analyse the spread of a disease [94]. Here, the disease is assumed to spread across a population due to connections between individuals. The vaccination of an individual would prevent further spreading of the disease through this individual, and thus has a “knock on” effect. By carefully targeting particular individuals for vaccination, this knock on effect could result in eradicating the disease from the population. In this example, we can think of the individuals receiving vaccination as the nodes that are removed from the network. In other networks, one may want to investigate a slightly different percolation process, which investigates the effect of removing edges, rather than nodes [16]. When considering the internet, for example, whereby connections (edges) are between routers, a fault in a connection would disrupt the communication regardless of whether the routers are functioning correctly.

#### Network Robustness

The percolation transition can be used to analyse the behaviour of a system [18, 95, 96]. A network that breaks down into small isolated components that can no longer communicate with each other, is unlikely to be able to carry out its function. A network that maintains a *giant connected component* (GCC) (i.e. a group of connected nodes that span a large proportion of the entire network) after the removal of nodes, on the other hand, has a higher chance of carrying out (at least partially) its function. For this reason, the size of the GCC, after a fraction of nodes have been removed, has been widely used as a measure of a network’s ability to function properly and so is closely related to the concept of network robustness [16, 18, 41, 97]. The rate by which the network function (in this case, the size of the GCC) decreases after a series of random failures (errors) or targeted attacks, is used to understand the networks resilience to various perturbations. For targeted attacks, note that it

is usual to remove nodes according to some centrality measure, such as node degrees or betweenness centrality [16]. In this case, the nodes are removed from the network in order of their centrality value from highest to lowest, and so such an approach targets the nodes with the highest chance of causing a “knock on” effect to the system.

## 2.5 Summary

In this Chapter we have given a brief overview of some of the most fundamental aspects of the theory of complex networks with a focus on those topics of most relevance to the current work. However, our review is not exhaustive. For example, we have not considered so-called community detection algorithms [24, 98, 99, 100, 101, 102, 103], the aim of which is to divide a network into nearly interdependent modular units, such that the intra-connections are dense, whereas the inter-connections are sparse. We have briefly mentioned centrality measures and their use in providing an ‘importance ranking’ of a networks nodes, however, the literature on these methods is vast (see, for example, [12, 17, 104, 105]), and they have also been used as a tool for detecting network communities [24], as well as a method for exploring network robustness [16, 93]. Finally spectral methods provide a powerful approach to studying networks [106]. The graph spectra is the set of eigenvalues of the adjacency matrix (or related matrix representations) and is related to a number of important structural properties, such as the diameter, average path-length and betweenness [106]. The graph spectra can also be used as an efficient method for computing subgraphs across the network, and has resulted in a number of proposed network measures, such as *subgraph centrality* [17] and *returnability* measures [107]. Note that the graph spectra of the *laplacian matrix* can also be used to reveal important structural information [108, 109, 110].



# CHAPTER III

## METABOLIC NETWORKS

At the cellular level, networks emerge via interacting proteins, and other macromolecules, resulting in various biochemical nets, such as gene regulatory networks, protein-protein interaction networks and protein residue networks. In this regards, the metabolic process in particular plays a fundamental role, providing the building blocks (nucleic acids and amino acids) that enable genes to interact effectively, and thus for the cell to function properly.

In this thesis we are interested in revealing a number of important features of metabolic systems using mathematical modelling techniques. In this chapter we present an overview of metabolic modelling. We start with a brief description of the metabolic process, followed by a discussion of the three main metabolic network modelling techniques. We then explain the procedure for reconstructing genome scale metabolic networks used in this work, and finish with a description of the 383 bacterial metabolic networks that have been reconstructed and used in this thesis.

### 3.1 The Metabolic Process

Metabolism is the entire set of biochemical reactions that take place within a living cell, in order for them to extract, convert and store energy from nutrients (molecules) within the environment. These biochemical reactions are crucial for the survival and proper functioning of a cell. The metabolites in each biochemical reaction are

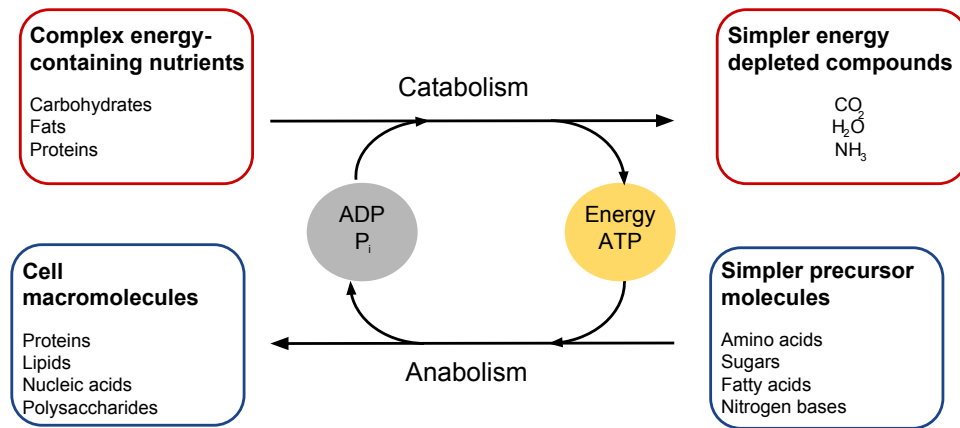


Figure 3.1: A schematic showing the relationship between catabolism and anabolism in a cell.

usually small molecules, such as glucose and amino acids, or macromolecules such as polysaccharides and glycan. Each biochemical reaction is usually catalysed by a specific enzyme (protein) that encourages the *substrate* metabolites to react together to form the *product* metabolites. Some reactions can be catalysed by more than one enzyme, whereas others do not require any. Additionally, it is possible that an enzyme can catalyse more than one reaction. The set of biochemical reactions can be divided into two categories.

- *Catabolism*: the process by which molecules, absorbed from the external environment, such as carbohydrates, fats and protein, are broken down into simpler molecules to produce and store energy.
- *Anabolism*: the process by which the energy is then used to synthesize molecules essential for producing the components within a cell, such as essential proteins, lipids, nucleic acids and polysaccharides. Figure 3.1 shows a schematic of the relationship between catabolism and anabolism.

A series of biochemical reactions form metabolic pathways, such that the products of one reaction act as the substrates to the subsequent reaction. Glycolysis is an example of one of the most ancient metabolic pathways, which converts glucose into pyruvate to release energy required for the formation of ATP and NADH

(high-energy compounds). The end product of a metabolic pathway may be used instantly, or may be consumed in another pathway. The reactions may be involved in multiple pathways, and multiple pathways may consume and produce the same metabolites. The entire set of biochemical reactions, and pathways, form a highly interconnected metabolic network, allowing one to consider the structure and function of the metabolic system as a whole.

## 3.2 Mathematical Models of Metabolism

The metabolic process is commonly modelled using a systems biology approach [111, 112, 113, 114], which provides a powerful set of tools and techniques for analysing the structural and dynamical behaviour that occurs within all living organisms. Metabolism can be studied from a static or dynamic point of view, the models of which are distinct in each case and vary in the amount of complexity and metabolic information they require. In the following we provide a brief description of the three most popular systems-like approaches to studying metabolism: kinetic models, constraint based models and topological (or network) models.

### 3.2.1 Kinetic Models

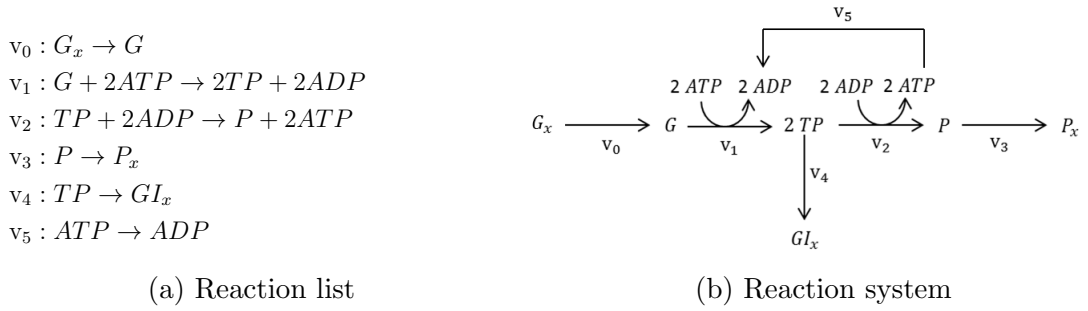
One of the most widely known methods for analysing metabolic processes replaces the chemical reaction network by a system of ordinary differential equations (ODEs), that is the metabolic network is represented as a system of mass balance equations, such that each equation describes the rate at which the concentration of a metabolite changes. Mathematically, this is written as

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^m s_{i,j} \nu_j(\mathbf{x}, \mathbf{k}),$$

or, in matrix notation

$$\frac{d\mathbf{x}}{dt} = S\boldsymbol{\nu}(\mathbf{x}, \mathbf{k}), \quad (3.1)$$

where  $S \in \mathbb{R}^{n \times m}$  denotes the stoichiometric matrix, that is, a matrix whose elements represent the number of metabolites produced ( $s_{i,j} > 0$ ) or consumed ( $s_{i,j} < 0$ ) in a single reaction step, and  $\nu(\mathbf{x}, \mathbf{k}) \in \mathbb{R}^m$  is a vector of rate equations. Note that each rate equation is a function of  $\mathbf{x} \in \mathbb{R}^n$ , a vector of metabolite concentrations, and a vector of kinetic parameters  $\mathbf{k} \in \mathbb{R}^d$ . Figure 3.2 illustrates the above in the case of glycolysis [115].



$\frac{d(G)}{dt} = \nu_0 - \nu_1$   
 $\frac{d(TP)}{dt} = 2\nu_1 - \nu_2 - \nu_4$   
 $\frac{d(P)}{dt} = \nu_2 - \nu_3$   
 $\frac{d(ATP)}{dt} = -2\nu_1 + 2\nu_2 - \nu_5$   
 $\frac{d(ADP)}{dt} = 2\nu_1 - 2\nu_2 + \nu_5$

(c) System of ODEs

$$\frac{d}{dt} \begin{bmatrix} G \\ TP \\ P \\ ATP \\ ADP \end{bmatrix} = \begin{bmatrix} +1 & -1 & 0 & 0 & 0 \\ 0 & +2 & -1 & 0 & -1 \\ 0 & 0 & +1 & -1 & 0 \\ 0 & -2 & +2 & 0 & 0 \\ 0 & +2 & -2 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \nu_0 \\ \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_5 \end{bmatrix}$$

(d) System of ODEs in matrix form

Figure 3.2: A minimal system of glycolysis. Note that here  $G_x$ ,  $P_x$  and  $GI_x$  are external and so are not considered in the system of ODEs. The abbreviations correspond to the following:  $G$ , glucose,  $P$ , pyruvate and  $TP$ , triosephosphate. (a) A minimal set of reactions, where one unit of glucose is converted into pyruvate. (b) A graphical illustration of the reaction system. (c) The set of ordinary differential equations (ODEs) modelling the change in concentration of each metabolite. (d) Matrix representation of the system of ODEs.

Kinetic models are an invaluable tool as they allow us to understand and predict a systems functional behaviour. Using such dynamic models has proved particularly useful in contributing to our understanding of the function and regulation of metabolic systems to date [111, 116, 117, 118, 119]. There are likely to be a number

of difficulties, however, when using kinetic models for metabolic modelling. One of the most challenging problems is determining kinetic parameters. It is highly likely, for instance, that the activity of many enzymes are dependent on a number of physiological conditions, such as temperature, pH, etc., and in a complicated and non-linear manner. For that reason, it is very difficult to obtain reliable estimates for kinetic parameters. The accuracy when using kinetic models to represent a large metabolic network consisting of 100's or 1000's of reactions, coupled with unreliable kinetic parameters, is likely to be scrutinized. For that reason, the application of dynamical models is currently only applicable to small systems or for studying individual pathways.

### 3.2.2 Steady-State Models

Unlike the kinetic models described above, steady-state models require only a small amount of information regarding the kinetic parameters of the system. This is achieved by assuming that the system has reached a steady-state, i.e. the rate by which the metabolites are consumed is equal to the rate by which they are produced. Let us recall that Equation (3.1) measures the concentration variation of the metabolites over time, which is proportional to the rate by which the metabolites are produced and consumed. If we assume that the system is at a steady state then Equation (3.1) reduces to the following set of linear equations,

$$S\boldsymbol{\nu} = 0,$$

where as before  $S \in \mathbb{R}^{n \times m}$  is the stoichiometric matrix and  $\boldsymbol{\nu} \in \mathbb{R}^m$  is the vector of unknown fluxes. Note that metabolic networks usually contain fewer metabolites than reactions, thus leading to an under determined system of equations, such that there are more variables than equations and thus potentially infinitely many solutions (of which not all will be feasible).

The most widely used steady-state model in the area is *flux balance analysis* (FBA), and unlike kinetic models, can be readily applied to large-scale metabolic networks. In FBA a linear set of constraints and an objective function are formulated

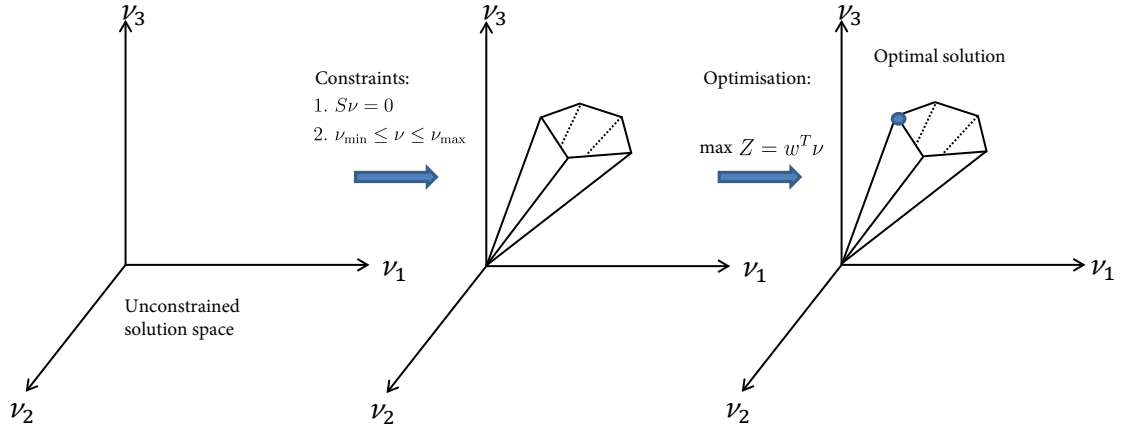


Figure 3.3: Illustration of constraint based modelling of metabolism. At first the flux distribution of the metabolic network may lie anywhere in the solution space (left). The mass balance constraint and the lower and upper capacity bounds are imposed, reducing the allowable solutions to a convex space (middle). An optimisation of an objective function allows FBA to identify an optimal solution, which lies on the edge of the convex space.

to reduce the solution space of the flux distributions. The formation of a FBA model would be as follows,

$$\begin{aligned}
 &\text{maximise } Z = \mathbf{w}^T \boldsymbol{\nu} \\
 &\text{subject to : } S\boldsymbol{\nu} = 0 \\
 &\text{and } \boldsymbol{\nu}_{\min} \leq \boldsymbol{\nu} \leq \boldsymbol{\nu}_{\max},
 \end{aligned}$$

where here  $Z$  is the objective function, usually chosen to maximise some growth condition, and the lower and upper limits of  $\boldsymbol{\nu}$  are known capacity constraints. Common objective functions include maximising the yield of biomass and maximising production of ATP for energy. In Figure 3.3 we provide an illustration to show how this constraint-based approach identifies optimal solutions. FBA has been used in many applications of bioengineering, where the technique is used to identify modifications to pathways in micro-organisms that would lead to optimal yield of industrial chemicals, such as biofuels [120, 121, 122]. It has also played a role in identifying suitable drug targets for treatments of cancer and pathogens [123, 124, 125].

### 3.2.3 Topological Models

Applying topological (network) models has become a popular choice in recent years, particularly due to the wide set of network measures and tools that can be applied with a relatively small amount of computational effort. Unlike the aforementioned models, using network models only requires information regarding the connectivity of metabolites, and thus can be applied to a large number of organisms with 1000s of nodes. From a topological perspective, metabolic networks are most commonly represented by simple graphs, whereby nodes and edges represent metabolites and biochemical reactions, respectively. There exist a number of ways of characterising metabolism as a simple graph, each of which captures different biological information. The following provides a brief description of each simple graphical representation.

- *Substrate-Substrate network*: edges connect either a pair of substrates or products of the same reaction.
- *Substrate-Product network*: edges connect substrates and products of the same reaction.
- *Substance network*: edges connect all substrates involved in the reaction.
- *Reaction network*: nodes represent reactions and an edge connects a pair of reactions if they share a common substance.

(See Figure 3.4 for an illustration of each representation.)

A plethora of network measures and tools exist (see Chapter 2 for a review of the main concepts in the field) that can be applied to graphical models with relatively little computational cost. These developments have led to new insights and understanding into the structure, organisation and behaviour of metabolic networks [29]. Perhaps the two most important findings to date, are that metabolic networks exhibit characteristics that are typical of the small world [71] and scale-free [41] phenomena. These characteristics are a generic feature of many non-biological and biological networks, and promote robustness of a network to random errors. In terms of metabolic networks, these characteristics are thought to be design principles that have evolved to enhance the survival and growth of the organism, particularly to

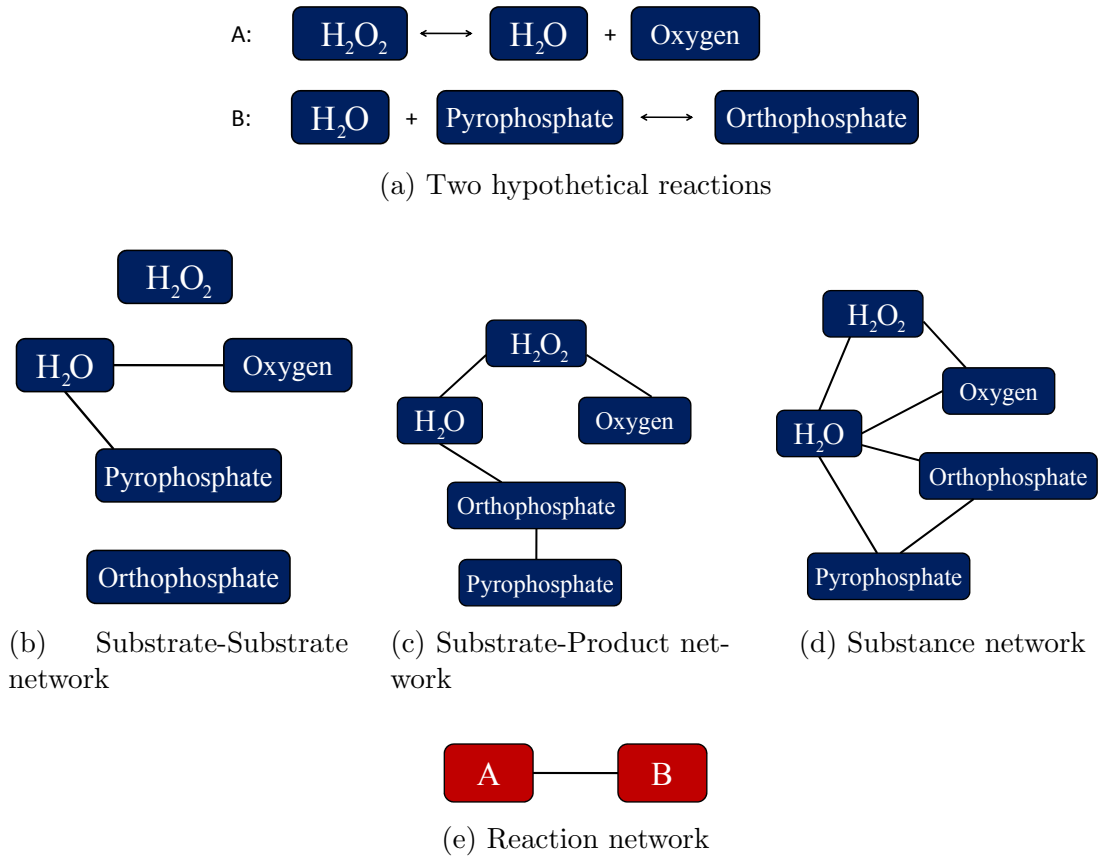


Figure 3.4: An illustration of simple graph models available for representing metabolic networks.

their environmental habitat. For example, a small average path-length is thought to allow the metabolic system to respond efficiently to either internal mutations or external fluctuations in an organisms environment.

Modularity has also been shown to be a prevalent feature within metabolic networks [21, 126], and has, for example, been related to important biological properties such as robustness [127, 128] and evolvability [128, 129, 130]. However, metabolic networks are by no means perfectly modular; their inter-module connectivity is relatively high, leading some authors to conclude that these networks are better described as being hierarchically structured [21], that is metabolic networks may be considered to possess fractal-like properties, such as self-similarity.

Another popular approach for analysing metabolic networks is provided by *network motifs*, [131], i.e. recurrent, statistically significant subgraphs. Motifs are of particular interest since they are typically associated with certain biological func-



tions, and their relative over-abundance is considered to be an evolutionary result reflecting their “importance” to the organisms involved [86]. Moreover, they constitute the basic structural units from which complex metabolic networks are formed, and thus provide a simplified framework for probing large-scale topologies.

Importantly, however, the majority of these studies carry out their metabolic analysis using simple graphical approaches such as those presented in Figure 3.4. In some circumstances, however, such an approach is likely to provide an inadequate description of metabolism, since (i) the choice of simple graph accentuates different aspects of the metabolic process; and (ii) information is inevitably lost when reducing the full system to a simple graph, and so one risks oversimplifying the system of interest in a potentially significant way. Zhou and Nakhleh [32], for example, recently provided evidence countering the reported scaling between network degree and the local clustering coefficient for the metabolic network of *E. coli*, arguing that previous sightings were due to a misrepresentation of the system as a simple graph.

An alternative approach is provided by bipartite graphs, whereby the graph consists of two types of nodes - metabolites and reactions, say. Indeed, several authors have recently considered such models of metabolism [99, 132, 133]. This approach though, can prove problematic. For example, bipartite networks do not exhibit clustering, and hence it makes no sense to talk about, say, the small-worldness of the network – in the Watts-Strogatz sense, anyway. Worse still, the majority of studies tend to project back onto some simple, unipartite representation before applying the standard tools of network science; again, the choice of projection is likely to heavily bias any results [53].

Hypernetworks, on the other hand, provide an attractive alternative since they allow for the description of more general interactions consisting of more than two nodes. Metabolic networks are particularly amenable to such an approach, with nodes representing different metabolites and hyperedges, that is sets of nodes, representing chemical reactions [33]. Note that we shall consider such models at length in Chapter 5.

### 3.3 Metabolic Network Construction Used in This Study

The availability and improved accuracy of genome scale data for a variety of different species, from bacteria to humans, has enabled the reconstruction and analysis of large-scale metabolic networks for organisms at a systems level. Metabolic data is now available from a number of online resources, some of the most widely used include: Kyoto Encyclopedia of Genes and Genomes (KEGG) [134], EcoCyc [135], BioCyc [136] and MetaCyc [137]. These databases offer a platform for researchers to easily reconstruct organism specific metabolic networks. In this section we describe the construction process for the metabolic networks used in our work.

#### 3.3.1 The KEGG Database

The KEGG database is an integrated platform, consisting of genomic, chemical and network information for multiple species. The database has four main sections including: systems information, chemical information, genome information and health information [138]. To reconstruct metabolic networks, we are interested in the chemical information, which is more commonly known as KEGG LIGAND. The two files ‘reaction\_name.lst’ and ‘reaction.lst’ can be downloaded from the KEGG ftp site (available at: <ftp://ftp.bioinformatics.jp/kegg/ligand/reaction/> to subscribed users only). The file ‘reaction\_name.lst’ contains all reactions that are within the ENZYME section and KEGG/PATHWAY database, whereby each reaction is identified by an R number. An example of a reaction within ‘reaction\_name.lst’ is:

R00259: Acetyl - CoA + L - Glutamate  $\longleftrightarrow$  CoA + N - Acetyl - L - glutamate.

In the file ‘reaction.lst’ the compounds are converted to C numbers, such that the reaction above would be converted to

R00259: C00024 + C00025  $\longleftrightarrow$  C00010 + C00624.

The reactions within these files include the full chemical equations (i.e. with stoichiometric coefficients and currency metabolites). To generate organism specific reaction lists, XML files can be downloaded from the KEGG ftp website, which contain R numbers (e.g. R00010), and indicate which metabolic reactions should be extracted from ‘reaction.lst’ to derive the organism specific reaction list.

### Exclusion of Currency Metabolites

When constructing metabolic networks, the inclusion or exclusion of currency metabolites should be considered. Currency metabolites, such as  $H_2O$ , ATP, NADH, etc., are metabolites that are mainly used for transferring electrons and specific functional groups (i.e. phosphate groups, carbon atoms, amino group, etc.). Consider, as an example, the reaction  $\text{Glucose} + \text{ATP} \rightarrow \text{G6P} + \text{ADP}$ . Here, the metabolites ATP and ADP transfer phosphate to glucose. Importantly, the connections between currency metabolites can sometimes result in biologically meaningless results. The shortest distance between a pair of nodes, for instance, is drastically reduced due to meaningless short-cuts across the network that are created by currency metabolites.

To derive reaction lists that exclude currency metabolites, the file ‘reaction\_mapformula.lst’ can be used, which like ‘reaction.lst’, is obtained from the KEGG ftp site. In this file, only the main reactant pairs in the metabolic reactions are considered, as is shown in the KEGG pathway diagrams, and thus the currency metabolites are excluded. The XML files for each organism are then used to extract R numbers to indicate which metabolic reactions should be extracted from ‘reaction\_mapformula.lst’, in order to derive the organism specific reaction list.

### Removing Mistakes and Inconsistencies from the Reaction Lists

A number of mistakes and inconsistencies exist within the reactions of the KEGG database, and therefore care must be taken when reconstructing metabolic networks from the original reaction lists. In the following we describe our efforts to correct such mistakes in the data used in this thesis.

First of all, the KEGG database derives reaction lists by curating several chemical pathway maps from the KEGG database. Thus, since a reaction may be present

within multiple metabolic maps there exist some reactions that are repeated within the list. Any repeat reactions with the same reaction ID need to be removed from the reaction list. However, due to errors within the KEGG database some of these repeats are not identical. This is due to the fact that chemical equations in different chemical pathway maps are catalysed by the same enzyme and thus have the same reaction ID, yet sometimes the reactions differ. These non-trivial cases, where we have differences between either the metabolites involved or the directionality of the reaction, were treated by taking the most comprehensive equation. In the following we provide 3 examples of the difficulties we faced, and describe the action that was taken.

**Example 1: include the most comprehensive reaction**

Reaction ID	Map ID	Chemical equation
R00212	00620	$C00058 + C00024 \longleftrightarrow C00022$
R00212	00650	$C00022 \longrightarrow C00024$

In this scenario the two equations have the same reaction ID but differ by the directionality and number of metabolites included. Thus in our work we would include the reaction corresponding to map 00620.

**Example 2: merging of substrates and products**

Reaction ID	Map ID	Chemical equation
R00114	00250	$C00026 \longrightarrow C00025$
R00114	00910	$C00064 \longrightarrow C00025$

In this example we have two equations that have the same reaction ID, however, they contain different substrates. In this situation we merge the substrates together to form:



Note that the same approach is taken for a scenario where we have two chemical equations with equivalent substrates but different products (i.e. the products would be merged together).

### Example 3: merging multiple reactions

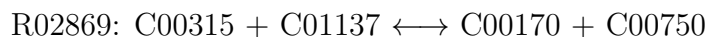
Finally, in this example we show the most complicated case, such that multiple reactions exist and the substrates and products of each reaction sometimes overlap.

Reaction ID	Map ID	Chemical equation
R02869	00270	$C01137 \longrightarrow C00170$
R02869	00330	$C00315 + C01137 \longrightarrow C00750$
R02869	00410	$C00750 \longrightarrow C00315$
R02869	00480	$C00315 \longrightarrow C00750$

Here, the equations for the maps 00330, 00410 and 00480 all include the information that either  $C00315 \longrightarrow C00750$  or  $C00750 \longrightarrow C00315$ , and so the three equations can be merged to form:



The substrate C01137 is a substrate in the maps 00330 and 00270, and therefore we can form the final reaction as



It is worth noting here that the reactions requiring manual edits as described above make up a relatively small proportion of the total number of chemical reactions considered, and so the choice by which we form the final reaction is not expected to have too significant an effect on our results.

Another problem arises due to condensation and polymerization reactions. These type of reactions often involve compounds occurring as both a substrate and product

in the reaction, which can be problematic when representing a reaction system in matrix form. The incidence matrix  $C(H)$ , for example, uses a  $-1$  to represent the substrates and  $1$  to represent the products of a reaction in order to distinguish between the different sides of a reaction. Obviously, reactions involving overlapping substrates and products cannot be represented in this manner. For that reason, we exclude such reactions. Several other studies, such as those using flux based analysis [139, 140], have also excluded these reactions due to the fact that they are imbalanced.

### Restricting to the Giant Connected Component

Despite the inter-connectivity between the biochemical reactions, metabolic networks are far from fully connected [141]. Instead, the connectivity structure of metabolic networks can be divided into several sub-networks called components. Components are clusters of nodes, such that a path exists between all pairs of nodes within the cluster. Recall from Chapter 2 that the largest component of a network is called the giant connected component (GCC). In directed graphs, the giant connected component commonly refers to the *giant strongly connected component* (GSCC), which is defined as the largest subset of nodes, such that a path exists in both directions between all possible node pairs. The *giant weakly connected component* (GWCC), on the other hand, relaxes the condition on the direction, such that a path only has to exist in one direction between each node pair.

It is common in the literature to restrict the analysis of metabolic networks, to consider the GCC (or GWCC if one is considering directed graphs) of the network only [129, 130, 142], since many global network measures are not amenable to disconnected components. The average path-length, for example, cannot be computed between pairs of nodes lying on distinct components (typically we set  $l = \infty$ ). Additionally, a number of issues can arise when constructing randomised networks due to the rewiring of disconnected components.

In our work we consider the GWCC of all of our metabolic networks (or hyper-networks) as this avoids the issues touched upon above, but also allows for a fair comparison to previous studies. In total we have data for 383 bacterial organisms. We found that the size of the GWCC relative to the entire network was on aver-

age 0.7497, and the majority of components only contained between 2 and 4 nodes. The size of the GWCC of the metabolic networks compared with all other components, therefore indicates that this is the most complex and core component of the metabolic networks that we wish to investigate, and so it is reasonable to restrict our analysis to this component.

## 3.4 Bacterial Metabolic Data Used in This Study

In total we have constructed metabolic networks for some 383 bacterial species. In addition to the reaction data, we also have information pertaining to biological features such as habitat lifestyle and growth conditions, both of which we describe in more detail below.

### 3.4.1 Habitat Lifestyle

For 115 of the metabolic networks we have detailed information regarding the variability of the environments in which they evolved. According to the National Centre of Biotechnology Information (NCBI) these bacteria can be classified according to the following six categories:

1. *Obligate bacteria*: species that obligately associate with a host, either intracellular or extracellular, and have no (or very little) contact with the environment outside the host.
2. *Specialised bacteria*: species that live in specialised conditions, such as marine thermal vents.
3. *Aquatic bacteria*: species that live in fresh water and never live within a host.
4. *Facultative bacteria*: free-living bacteria, that can associate with a host. An example, is *E.coli*, a bacteria that can live in a number of different conditions, including a host environment.
5. *Multiple bacteria*: species that can live in a wide range of environments, and can associate with a variety of different hosts.

6. *Terrestrial bacteria*: species that live in soil, where the conditions are highly heterogeneous.

In [129], these categories were ranked according to the heterogeneity in their environment based on a number of physiological conditions, such as temperature, pH, oxygen availability, etc. It was found that the obligate class were the bacteria that lived in the most isolated environments, being surrounded by very few if any other species, thus resulting in less competition for nutrients. For that reason, they were ranked as the most specialised. The ordering between the specialised and aquatic classes is somewhat ambiguous and are often considered to have equivalent levels of heterogeneity in their environments. These bacteria have adapted to specialised ecological conditions (marine thermal vents and fresh water), however, the amount of competition from other species is higher than the obligate class. Both the facultative and multiple classes can live both inside a host and in the outside environment, however, the multiple class, can span a wider range of hosts and environments. The terrestrial class is then considered the most varied environment, since soil is considered to have the most heterogeneous ecological conditions.

### 3.4.2 Oxygen Requirements

In addition, we also have data concerning oxygen requirements for all 383 species. More specifically, we can group the bacteria into three groups based upon their oxygen dependence:

1. *Aerobic bacteria*: species that can only survive in the presence of oxygen.
2. *Facultative aerobic bacteria*: species that can survive both in the presence and absence of oxygen in the environment.
3. *Anaerobic bacteria*: species that live in the absence of oxygen in the environment.

A number of recent studies [143, 144, 130] have highlighted the effect of oxygen on metabolic performance and so we would hypothesise that metabolic networks evolving under aerobic conditions will be structurally different to those evolving under anaerobic conditions.



## 3.5 Summary

The main benefit of a topological approach above the other methods discussed in this chapter is that it allows for large-scale studies of metabolism both within and across different species. Whilst such models undoubtedly represent a very abstract view of metabolism, lacking in detail and complexity, recent evidence suggests that these complex interaction patterns may reflect the evolutionary origins of important functional changes within cellular metabolism [129, 142, 144]. Thus, understanding their topology promises to unravel important features of biological organisation at the systems level. To finish, it is worth mentioning at this point, that reconstructing metabolic networks using the procedure described above, is by no means perfect. Despite the KEGG database being one of the most well established databases, the data still remains incomplete, and could therefore result in false positives (or false negatives) when carrying out the analysis on the networks. Indeed, this provides an additional motivation for restricting our analysis to the giant connected component of the global network, since this component has been shown to contain the most accurate data [145].

# CHAPTER IV

## A NOVEL APPROACH TO NETWORK CLASSIFICATION BASED ON NETWORK MOTIFS

In this Chapter we begin by giving a brief overview of the network classification problem before describing a number of recent works in the area. We then go on to introduce a new graph embedding approach that uses low-order network motifs to map a network into Euclidean space, thus allowing us to use data-mining tools to determine network clusters. We finish by illustrating the new method on a large ensemble of bacterial metabolic networks as well as a number of synthetic tests on model networks.

### 4.1 Current State-of-the-Art

Classifying networks based upon global features such as clustering coefficients, degree distributions, average path lengths, etc., has proved popular in network science. For example, networks are commonly classified as being small-world or scale-free [2, 41, 146, 147]; in fact some authors have taken this a step further and introduced a number of additional classes based upon, for example, the type of power law being

followed [148, 149]. However, such approaches have received heavy criticism of late. Indeed, in a recent study by Li *et al.* [30], it was shown to be relatively easy to construct two large networks with identical global properties, yet considerably different structure. These findings are particularly pertinent in the case of biological networks, where the incompleteness of the large, noisy data sets under consideration can be misleading. Thus, network classification is becoming an increasingly important problem in the field, the goal of which is to cluster networks according to the similarities/differences in their topological structures.

To address this problem a suitable approach for the comparison of two (or more) networks is necessary. However, this is a non-trivial task in general. Obtaining the exact structural differences between two large networks is in fact infeasible, since it requires solving the graph isomorphism problem which is NP-complete [150]. In practice, heuristic approaches are developed that solve a related, ‘simpler’ problem, resulting in a suitable approximation.

One such approach is to construct a suitable graph metric which can be used to cluster networks. For example, Sanfeliu and Fu [151] introduced the so-called graph edit distance as a measure of similarity between two graphs, whereby edit costs are associated to the relabelling, deletion and addition of vertices and edges, required to map one graph to another. An additional related approach was put forward by Kondor and Lafferty [152], who introduced the idea of graph kernels as a way of computing pairwise similarities, which, in conjunction with statistical machine learning techniques (e.g. Support Vector Machine), allowed for the determination of network clusters.

More recently, a class of approaches that represent a graph in a (possibly high-dimensional) vector space, i.e. graph embeddings, have been proposed. For example, in [153] graph and vertex attributes (e.g. efficiency, degree, etc.) have been used to map functional brain networks into a low-dimensional space in which to perform a clustering analysis. Another related approach was proposed by Ren *et al.* [154] in which feature vectors consisting of the so-called Ihara graph coefficients were constructed, before applying a PCA analysis to categorise a large number of networks.

Other notable, recent attempts include Guimera *et al.* [155] who proposed a technique for classifying networks according to their modular structure. Whilst a

related approach by Onnela *et al.* attempts to provide a taxonomy of networks based upon a scaled measure of community structure [156]. Finally, in a series of papers by Przulj *et al.* [157, 158, 159, 160, 161, 162] the idea of graphlet distributions has been introduced and used as a technique for network classification.

## 4.2 A New Graph Embedding Approach

Motivated by the prominent role that network motifs have played to date in the analysis of biological networks (see Chapter 2), we propose a new, lossy graph embedding technique based on low-order motifs. The proposed technique is lossy in the sense that the original network cannot be recovered from the corresponding vector-space representation. Importantly, such an approach takes a difficult and unwieldy problem, i.e. the analysis of many large, complex biological networks of differing order, and replaces it by one which is ‘easier’ to manipulate - a plethora of tools and techniques from statistical machine learning already exist for the analysis of the resultant embedded data.

### 4.2.1 Motif Frequency Vectors

Motif frequencies can be used to directly compare different metabolic networks as they provide a ‘unique’ network signature [74]. Alternatively, networks can be compared by calculating a feature vector of z-scores, computed in the usual way, i.e.

$$z_{i,j} = \frac{N_j^i - \langle N_j^{\text{rand}_i} \rangle}{\sigma_j^{\text{rand}_i}},$$

where here,  $N_j^i$  denotes the rate of recurrence of the  $j$ th motif within the  $i$ th network whilst  $\langle N_j^{\text{rand}_i} \rangle$ , and  $\sigma_j^{\text{rand}_i}$  denotes the mean and standard deviation of the rate of recurrence of the  $j$ th motif in an ensemble of randomised networks [83]. Using z-scores in this way assumes that motif frequencies are normally distributed across a network and that each motif is independent, and whilst this is not always guaranteed, such an approach is standard throughout the literature [85, 86, 93, 163].

In this way, we can compute, for each network of interest, a feature vector  $\mathbf{z}_i$ ,

whose elements are the z-scores of each network motif. For example, if we consider all 3- and 4-node motifs then the result is a vector  $\mathbf{z}_i \in \mathbb{R}^{212}$  representing the  $i$ th network.

Note that it is typically the case that the networks we wish to compare are of varying order and as such we need to take care that network size does not bias any results. To handle this issue one can consider instead of the z-scores defined above, a so-called *significance profile* [85] defined by

$$s_{i,j} = \frac{z_{i,j}}{\sqrt{\sum_m z_{i,m}^2}}. \quad (4.1)$$

The motif significance profile for the  $i$ th network,  $\mathbf{s}_i$ , is simply the normalised vector of z-scores. The motif significance profile allows for direct comparisons between networks of different sizes. This is important due to the fact that motifs in larger networks tend to exhibit larger z-scores than they do in smaller networks [85]. Note also, that the motif significance profile lies in the interval  $[-1, 1]$ .

In the work presented in this thesis we threshold the network significance profiles such that any entries  $s_{i,j} < 0$  are set to zero as we are only interested in those motifs that are over represented. Motifs that are under-represented are known as anti-significant motifs, or anti-motifs, and although we do not consider them in this study, the approach forwarded here can easily be extended to that case. This results in a matrix

$$S = [\mathbf{s}_1, \dots, \mathbf{s}_m]^T \geq 0,$$

i.e. a non-negative matrix, whose rows consist of the significance profiles (thresholded) for the  $m$  networks under investigation.

To analyse the matrix  $S$  we use a matrix decomposition to compute a low-rank approximation of the data [164]. Since our data is non-negative it is natural to decompose it using a non-negative matrix factorisation (for the actual algorithmic details of the method see Appendix A). Such an approach is akin to a principal component analysis, that reduces the dimension of the problem, thus allowing us to

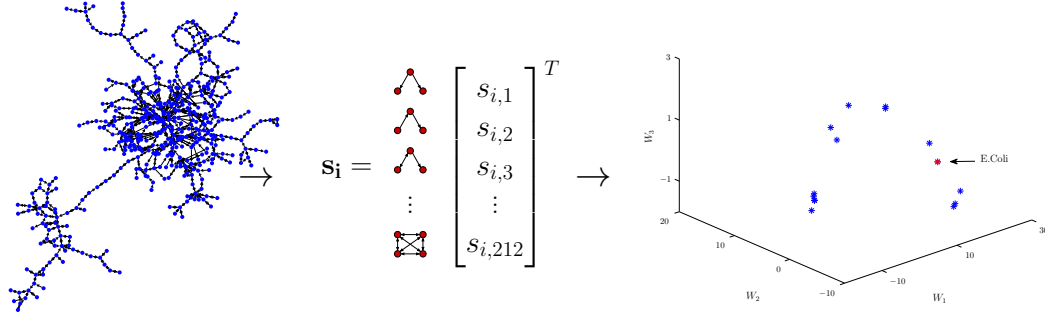


Figure 4.1: Illustration of our algorithmic approach.

detect important network features. Mathematically, we approximate  $S$  as follows

$$S \approx WH, \quad (4.2)$$

where  $W \in \mathbb{R}^{m \times k}$  and  $H \in \mathbb{R}^{k \times 212}$  are non-negative matrices. Here,  $k$  is the rank of the approximation and  $m$  the number of networks being considered. Importantly, both the columns and rows of  $W$  and  $H$  can be used to reveal important network features [165, 166].

The approach can be concisely summarised into the following three basic steps (see Figure 4.1 for a schematic description):

**Step 1:** For each metabolic network compute the significance profile,  $\mathbf{s}_i \in \mathbb{R}^{212}$ , consisting of the normalised significance scores for each of the 212 three- and four-node motifs.

**Step 2:** Compute a low-dimensional ( $k \ll 212$ ) representation of the thresholded matrix of significance scores,  $S = [\mathbf{s}_1, \dots, \mathbf{s}_m]^T$ , using a non-negative matrix factorisation.

**Step 3:** Use the columns/rows of  $W/H$  to determine network features.

### 4.2.2 Global and Local Motif Significance Scores

In order to determine the relative importance of the  $j$ th motif in the  $i$ th network we construct the following local motif significance score:

$$P(i, j) = s_{i,j} \cdot h_{1,j}. \quad (4.3)$$

Note that this results in a matrix  $P \in \mathbb{R}^{m \times 212}$  (recall that  $m$  denotes the number of networks under investigation), whose rows encapsulate the network motif structure for each metabolic network, and whose columns provide information pertaining to the relative importance of specific motifs across the network ensemble.

In the experiments in the next section, we derive a *global significance score* for each network by summing the rows of  $P$  as follows

$$\begin{aligned} P_{\text{global}}(i) &= \sum_j P(i, j) = \sum_j s_{i,j} \cdot h_{1,j} \\ &= \mathbf{s}_i \cdot \mathbf{h}_1, \end{aligned} \quad (4.4)$$

As alluded by the second row in Equation (4.4), this is equivalent to projecting the significance vector  $\mathbf{s}_i$  onto  $\mathbf{h}_1$ , the first row of  $H$ . Note that when computing the factors  $W$  and  $H$ , MATLAB normalises the rows of  $H$  to be of unit length, and orders the columns of  $W$  in decreasing order of their magnitude. In practice, the magnitude of the columns of  $W$  provide a measure of the role that each of the  $k$  factors play in describing the objects under consideration (metabolic networks in our case). For that reason, the first row of  $H$  is likely to provide the optimal single-variable projection of the data [164].

Importantly, we consider the global significance score in Equation (4.4) to be a proxy for network complexity, in the sense that a large value indicates the presence of a relatively large number of network motifs, whereas a low value indicates a simpler, more tree-like structure.

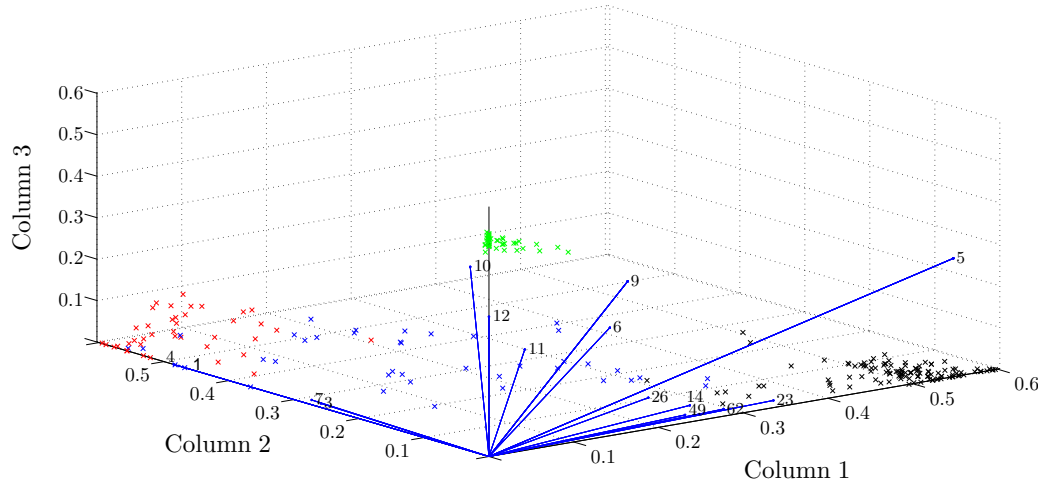


Figure 4.2: A biplot representation of the data matrix for an ensemble of 235 networks consisting of 115 bacterial networks (black), 40 Erdős-Rényi networks (blue), 40 small-world networks (green), and 40 scale-free networks (red). The axis represent the three dimensions obtained via non-negative matrix factorisation transformation.

### 4.2.3 Detection of Network Motifs and Choice of Null Model

Network motif frequencies were computed using the open-source software *mfinder*, available at <http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html>. This software uses a semi dynamic programming algorithm in order to reduce the computational running time, and thus provides an efficient way of computing low order motifs in relatively large networks. Note here that algorithm complexity increases with the size of the network under analysis. In general, the number of motifs tends to be higher in larger networks, and therefore increases the computational runtime. In addition, the number of motifs of size  $q$  grows exponentially, and thus reaches enormous quantities for relatively low-order motifs. The number of motifs of order 3, 4, 5 and 6 is equal to 13, 199, 9364 and 1530843 respectively. For these reasons we restrict the work presented here to include 3- and 4-node motifs only.

To determine significance, motif frequencies were computed against frequency distributions for some 1000 random graphs, chosen so as to preserve both the in- and out-degree, as well as  $(q - 1)$ -node motifs (see Chapter 2 for details on the algorithmic procedure). Note that the latter condition ensures that the enrichment of  $q$ -node motifs is not simply due to the presence of highly significant subgraphs.



## 4.3 Test on Synthetic Data

In this section we present a ‘proof of principle’ of our new algorithm by attempting to correctly classify a large ensemble of networks consisting of both model and real-world networks. More specifically, we consider 115 metabolic networks (see Chapter 3 and the next section for a detailed description) as well as some 120 model networks. The model networks used are the standard ones described in Chapter 2: (i) Erdős-Rényi (ER) random graphs; (ii) small-world (SW) networks and (iii) scale-free (SF) networks. For each model type we construct 40 instances, each having  $n = 200$  nodes and approximately  $m \approx 500$  edges, giving a density of about 0.01; note that such a density is typical of that found in the corresponding metabolic networks. The bacterial networks are highly variable, with the number of nodes ranging between 65 and 892, and the corresponding densities in the range  $[0.05, 0.001]$ .

We have applied the new algorithm (i.e. steps 1 - 3 above) to the above ensemble of networks thus obtaining feature vectors consisting of the 212 motif significance scores for each of the 235 networks, resulting in a  $235 \times 212$  data matrix. Using a non-negative matrix factorisation each data point is mapped to a 3-dimensional space (see Figure 4.2). Here, the results are presented using a biplot, which allows us to visualise the data from two points of view - the transformed observations (i.e. networks) are displayed as points, and the corresponding variables (i.e. network motifs) as vectors [167]. The points in the biplot can be interpreted in the same way as a standard scatterplot in that data points that are located close together (far apart) can be considered as being similar (dissimilar), with respect to their motif structure. The corresponding vectors are aligned in the direction that is most strongly related to a particular variable, or motif in our case, and the length of each vector conveys the strength of the relationship [168]. Note that each vector is labeled with its corresponding motif ID.

As can be clearly seen, the approach successfully determines distinguishable clusters for the bacterial networks (black), the SW networks (green) and the SF networks (red). The ER networks, on the other hand, appear more dispersed across the graph, which, is perhaps not too surprising since these graphs contain the most random structure, thus the likelihood of two graphs from this group having the same signif-

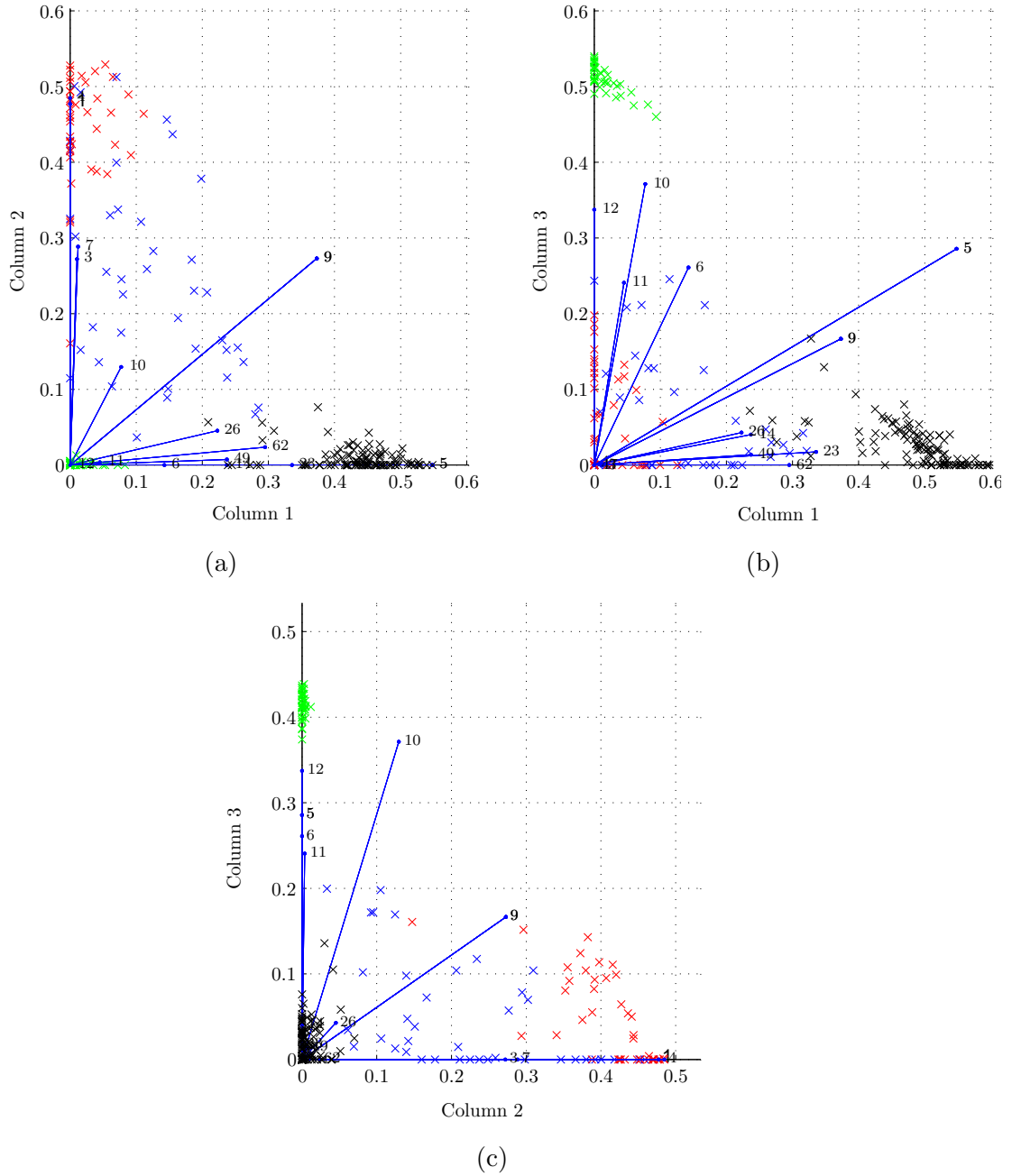


Figure 4.3: 2-dimensional projections of the biplot in Figure 4.2 on to (a) columns 1 and 2; (b) columns 1 and 3; and (c) columns 2 and 3. As before, bacterial networks are shown in black, Erdős-Rényi networks in blue, small-world networks in green, and scale-free networks in red.

icant motifs is small. Importantly, we can use the biplot to determine which motifs are driving the observed differences. Illustrations of the motifs obtained in the biplots are displayed in Table 4.1. In Figure 4.3 we present 2-dimensional projections,

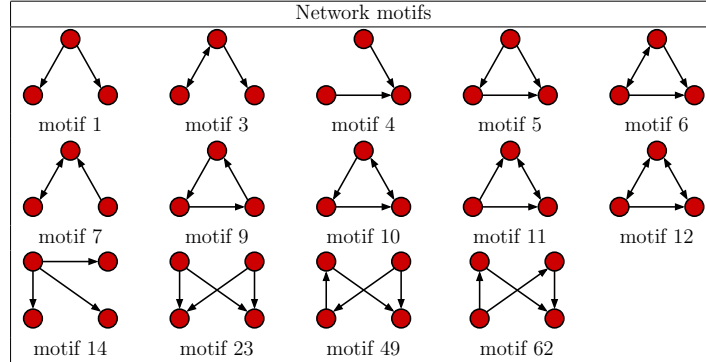


Table 4.1: Network motifs obtained for the 120 network models and 115 metabolic networks using the new algorithm.

in order to obtain a clearer visualisation of the role of each column. Firstly, we observe that motif 9 is consistent within all 3 columns (Figure 4.3a – 4.3c), thus suggesting that motif 9 is generic to all the networks considered here. In addition, motifs 5 and 6 are found to be a common feature in both the metabolic networks and small-world networks, motif 6, however, appears slightly more significant in the latter class.

In terms of each of the clusters, we find that the metabolic networks are mainly explained by column 1, since the biplot in Figure 4.3c, more or less, shows the cluster of metabolic networks lying very close to the origin. The motifs 14, 23, 49 and 62 all appear to be specific to the class of metabolic networks. Small world networks, on the other hand, are almost completely explained by column 3, since this class of networks lie extremely close to origin in Figure 4.3a. The motifs 11 and 12 appear to be specific to the class of SW networks. Motif 10 also appears to be a prominent feature of SW networks, however, the biplots suggest that this motif is also featured in ER networks and/or SF networks but to a lesser extent. Furthermore, the SF networks appear to be most explained by column 2, however, some information appears to be contained within column 1 and 3, since the SF networks do not lie near the origin in Figure 4.3b. This result suggests that SF networks have a slightly more randomised structure, in terms of network motifs, than either the metabolic or SW networks, but exhibit a considerably higher amount of ordering than ER networks. In particular, the motifs 1, 3, 4 and 7 appear to be specific to the class of SF networks.

Environment	Nodes			Edges		
	min	median	max	min	median	max
Obligate (34)	78	273	620	91	340	840
Specialised (5)	442	480	541	566	641	692
Aquatic (4)	541	580	647	700	751	868
Facultative (41)	90	652	809	101	890	1160
Multiple (28)	430	615	800	560	821	1119
Terrestrial (3)	557	689	693	779	944	966
Total (115)	78	541	809	91	730	1160

Table 4.2: Network statistics for the reaction graphs of the 115 bacterial species studied in this work classified according to environmental variability. According to the NCBI, obligate bacteria have the most constant environment, followed by specialised and aquatic, and then facultative, multiple and terrestrial bacteria in that order.

## 4.4 Tests on Directed Metabolic Networks

Next we go on to consider metabolic networks in more detail. The metabolic data in this study was derived from the KEGG database on May 20th, 2011. In total we studied 383 bacterial species (see Tables 4.2 and 4.3 for an overview of some basic network properties), each being characterised by a number of shared biological features (e.g. environmental variability, oxygen requirements and genome size), using graph theoretical techniques. For this study we represent the metabolic networks as substrate-product graphs whereby nodes and edges correspond to metabolites and reactions, respectively. Note that a potential caveat of such an approach is that it can lead to the detection of erroneous pathways. However, since we are not considering a path analysis here and for the ease of comparability with previous studies, we consider the substrate-product representation in all our experiments. Moreover, the currency metabolites, such as  $H_2O$ , ATP and NADH were removed from the analysis as they tend not to be involved in higher order functions, and if included, typically lead to physiologically meaningless pathways. Finally, to further simplify the analysis, we consider only the largest connected component for each network. For more details of the network construction please refer back to Chapter 3.

As an illustration of the new approach we carried out two experiments with the aim of testing the hypothesis that organism adaptability is manifested *via* the network motif structure of the corresponding metabolic networks.

Environment	Nodes			Edges		
	min	median	max	min	median	max
Aerobic (154)	65	605	892	74	809	1210
Facultative (180)	78	602	816	91	825	1168
Anaerobic (49)	307	488	681	381	645	969
Total (383)	65	581	892	74	789	1210

Table 4.3: Network statistics for the reaction graphs of the 383 bacterial species studied in this work classified according to species’ oxygen requirements. The degree of oxygen required increases in the order anaerobic, facultative and aerobic.

#### 4.4.1 Environmental Variability Promotes Network Complexity

The first experiment undertaken considered 115 metabolic networks, each being categorised according to their environmental habitat (see Table 4.2). The organisms can be found in a variety of conditions, ranging from highly specialised (e.g. symbiotic bacteria living within a host), to extremely heterogeneous conditions such as soil, and thus have evolved under very different selective pressures. For more details on the environmental habitats please refer to Chapter 3.

Figure 4.4 shows a plot of the mean global motif score,  $P_{\text{global}}$ , versus environmental variability for the 115 different bacterial networks. Note that the average here is taken over each of the 6 environmental classes: obligate, specialised, aquatic, facultative, multiple and terrestrial. Importantly, we found that motif frequency, and thus network complexity, increased significantly with environmental variability. The lowest motif frequency is found for the bacteria within the obligate class, followed by a relatively steep increase to the specialised and aquatic classes, then higher again for the facultative and multiple classes, and then highest for the terrestrial class. The group differences shown in Figure 4.4 are significant by the Kruskal-Wallis (KW) test (p-value:  $p < 10^{-9}$ ). Note, that for the result presented here, the non-negative matrix factorisation was carried out using  $k = 3$ . For more details of the choice of  $k$  see Appendix A.

This result relates to a number of previous studies that have found a relationship between environmental variability and network complexity [129, 142, 169, 170]. Biologically, this increase in network complexity manifests via an increase in genome

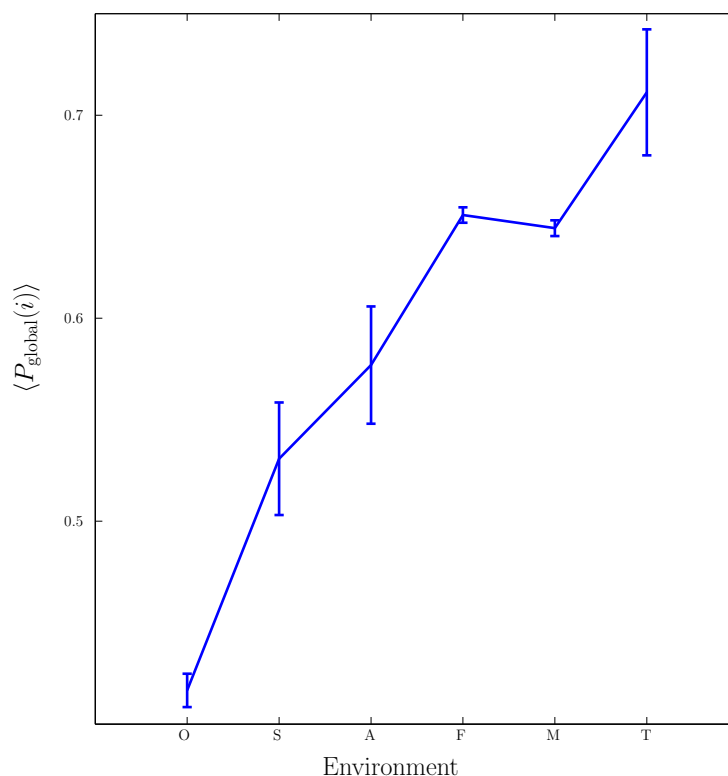


Figure 4.4: Relationship between environmental variability and the mean global significance score  $\langle P_{\text{global}} \rangle$  for the six bacterial habitats: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. Vertical bars represent standard errors.

complexity, as has been evidenced in a number of recent studies. For example, Zhou *et al.* [171] recently found that microbes that inhabit heterogeneous environments display a larger metabolome, as compared against those leading a more specialised lifestyle. The obligate symbiont *Buchnera aphidicola* provides such an example. Indeed, in a recent study by Pérez-Brocal *et al.* [172] a significant reduction in genome length was discovered in a comparative analysis of previously sequenced strains. We further investigated the relationship between environmental variability and genome size by considering the total number of genes in the 115 metabolic networks. As expected, we find a significant relationship between the six environmental classes and the total number of genes using the Kruskal-Wallis test (p-value:  $p < 10^{-8}$ , see appendix 2). As a control for the effect of genome size on  $P_{\text{global}}$ , we computed the Spearman’s partial correlation between  $P_{\text{global}}$ , environment variability and the total number of genes, and find that the results in Figure 4.4 remain significant ( $c = 0.4239$ ,  $p = 10^{-5}$ , see Appendix B).

The results presented above provide further evidence that environmental variability promotes metabolic network complexity. From a biological point-of-view this result may naturally be considered a product of the evolutionary process: due to the existence of the steady supply of metabolites that a constant environment provides, certain pathways become redundant and are removed over time. Whilst, contrary to this, the uncertainty faced by those species inhabiting varied environments, means that the additional costs, associated with more complex metabolic structures, are a necessary requirement for survival.

#### 4.4.2 The Effect of Oxygen Requirement on Network Structure

Next, we considered the effects of oxygen requirements on metabolic network structure. We studied some 383 bacterial species which were categorised into 154 aerobes, 180 facultative aerobes and 49 anaerobes.

Figure 4.5 shows a plot of the mean global motif score versus growth conditions for the 383 different bacterial species. Interestingly, we find that networks that have evolved in the presence of oxygen, that is, aerobes and facultative aerobes, have a significantly larger number of network motifs. The group differences shown in Figure 4.5 were found to be significant using the Kruskal-Wallis (KW) test (p-value:  $p < 10^{-4}$ ). Note, as before,  $k = 3$  was chosen in the non-negative matrix factorisation.

Importantly, this result suggests that bacterial networks that are exposed to oxygen are able to form additional pathways and motifs, compared to those that are oxygen deprived. Thus enabling them to evolve more complex network architectures, a result that is in agreement with a number of recent studies (see for example, the paper by Raymond and Segré [143]). Indeed, Raymond and Segré found that the number of reactions and metabolites was approximately 1.5 fold higher in aerobic bacteria compared to anaerobic bacteria, and thus supports the view that oxygen availability induced network complexity. As with the habitat data, we investigated the relationship between growth conditions and genome size. We find a significant relationship between the availability of oxygen and the total number of genes using

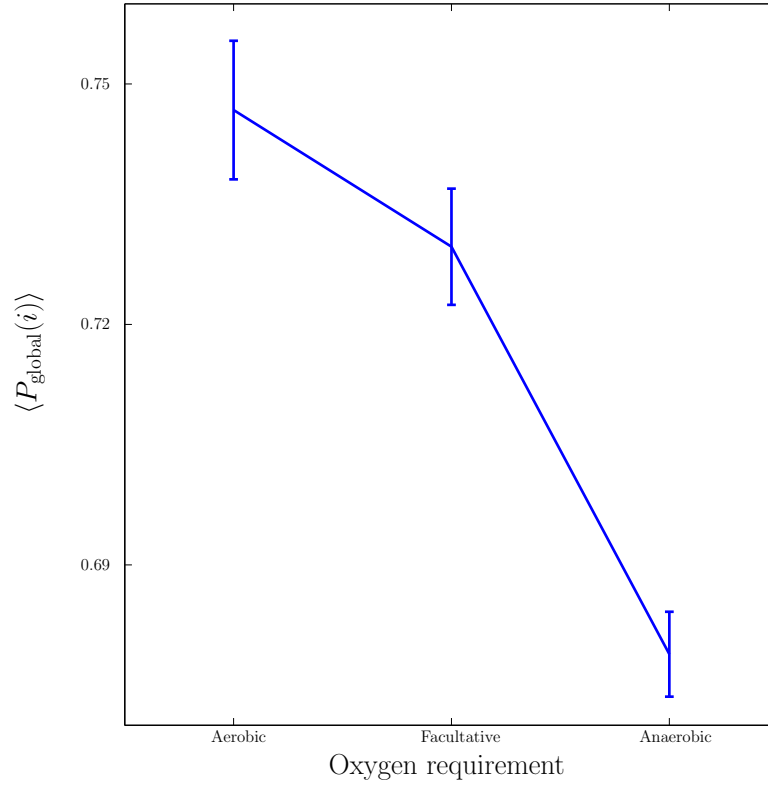


Figure 4.5: Relationship between growth requirements, more specifically oxygen requirements, and the mean global significance score  $\langle P_{\text{global}} \rangle$ . Vertical bars represent standard errors.

the Kruskal-Wallis test (p-value:  $p < 10^{-2}$ , see appendix 2). As a control for the effect of genome size on  $P_{\text{global}}$ , we computed Spearman’s partial correlation between  $P_{\text{global}}$ , oxygen requirements and the total number of genes, and find that the result in Figure 4.5 is still significant ( $c = -0.1020$ ,  $p = 0.0464$ , see Appendix B).

#### 4.4.3 Motifs Responsible for the Observed Differences

To determine the specific motifs driving the observations of the previous section, we considered the quantity  $\sum_i P(i, j)$ , that is, the column sum of the matrix  $P$  defined in Equation (4.3) – recall that the columns of  $P$  contain information specific to individual motifs. Moreover, by restricting the sum above to a particular subgroup of interest (specialised, obligate, multiple, etc.), it is possible to detail the extent to which any particular motif featured within that group. In the following we consider a motif to be significant within a particular group, if the mean local significance score



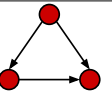
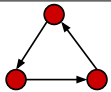
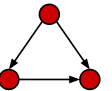
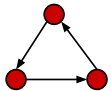
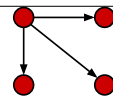
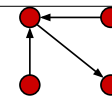
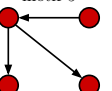
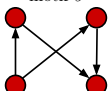
Environment	Significant motifs			
Specialised (43)	 motif 5	 motif 9		
Varied (72)	 motif 5	 motif 9	 motif 14	 motif 26
	 motif 15	 motif 62		

Table 4.4: Motifs significantly overrepresented in networks pertaining to a specialised and varied environment.

of that motif (restricted to the group of interest) is at least 2 standard deviations greater than the mean score across the entire network ensemble.

It is important to note that due to the nature of the algorithms used to compute the non-negative matrix factorisation (see Appendix A for details), the significant motifs discovered by the above approach may vary as we repeat the experiment. For this reason, we perform the matrix factorisation a total of 1000 times and use the frequency with which a particular motif occurs as a measure of significance. More specifically, motifs reported in this chapter must have been found to be significant in at least 75% of the 1000 factorisations performed (additional details can be found in Appendix A).

### Habitat Variability

In order to simplify the analysis we considered two groups: specialised (consisting of the obligate, specialised and aquatic classes) and varied (consisting of facultative, multiple and terrestrial classes). The significant motifs are displayed in Table 4.4. The first thing to note is that motifs 5 and 9, a feed forward loop and closed cycle, respectively, are prominent throughout the entire ensemble of networks, regardless of environmental factors. This is perhaps not too surprising as both of these patterns are considered to play important functional roles in many biological networks. The addition of a feed forward loop to a linear cascade of biochemical reactions, for example, has been hypothesised to accelerate the metabolic process [173]. Importantly, we found the number of significant motifs to be greatest in those metabolic

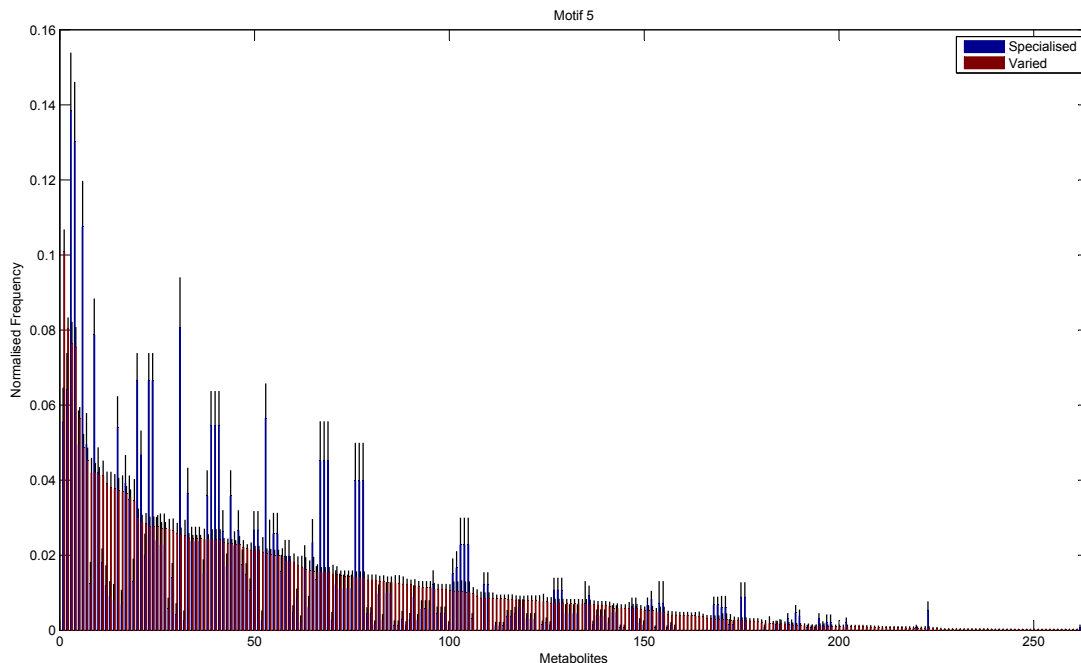


Figure 4.6: Mean normalised frequency for the 263 metabolites obtained within motif 5 for the 115 metabolic networks. Blue bars represent the specialised class and red bars represent the varied class. Here, the metabolites are in descending order of the metabolite frequencies for the varied class.

networks exposed to more variable environments: 2/212 for specialised and 6/212 for varied (see Table 4.4). Clearly, this represents only a very small percentage of available 3- and 4-node motifs ( $\approx 1\text{-}3\%$ ), and so the differences observed in Figure 4.4 can be attributed to a small set of motifs more or less specific to the different kinds of bacteria.

The increased numbers of network motifs present within the varied class indicates a potentially significant growth in network redundancy within those organisms inhabiting fluctuating environments, and can be considered as further evidence of so-called *functional redundancy mediated robustness* [174], that is, the observed perseverance of systems level redundancies prevalent in metabolic, as well as more general, cellular networks. More specifically, of the 4 additional significant motifs found in the varied class, motifs 14 and 15 may be considered variants of the single-input motif, motif 62 a bi-parallel fan, and motif 26 a multi-input motif, all of which have

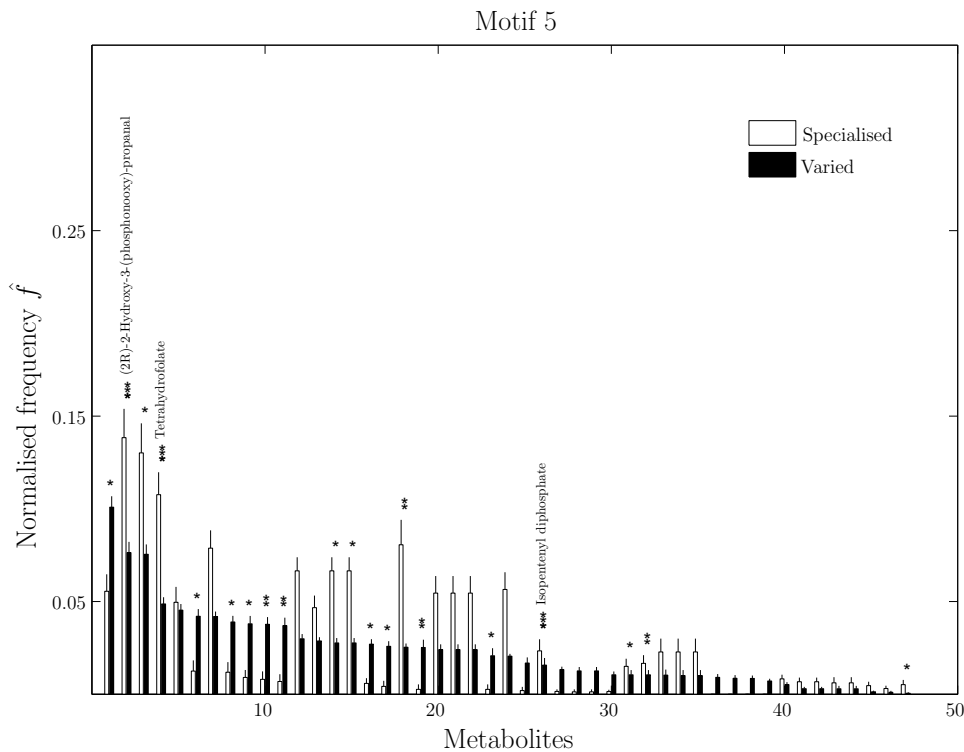


Figure 4.7: Mean normalised frequency for the significant metabolites obtained within motif 5 with  $p < 0.01$  (Fisher’s Exact test). Vertical bars are standard errors. Asterisks indicate large significant differences between metabolic networks from a specialised and varied environment, where \*, \*\*, and \*\*\* correspond to  $p < 0.001$ ,  $p < 0.0001$  and  $p < 0.00001$ . Metabolite names are provided for the most significant metabolites.

been implicated as potential indicators of network redundancy. For example, in the context of metabolism the single-input motif consists of a substrate  $X$  that is consumed in multiple reactions, the result of which are the products  $Y, Z, \dots$ ; whilst the bi-parallel fan implies the presence of multiple, or compensatory, pathways whose efficiencies may vary according to alterations in environmental conditions. Indeed, these findings are in agreement with a number of recent studies relating genetic robustness and organism adaptability [174, 175], and suggest that bacteria that live in more variable environments typically display a greater abundance of redundant metabolic reactions.

In addition to the topological differences observed between varied and specialised bacteria, we found that the distribution of those metabolites occurring within motif

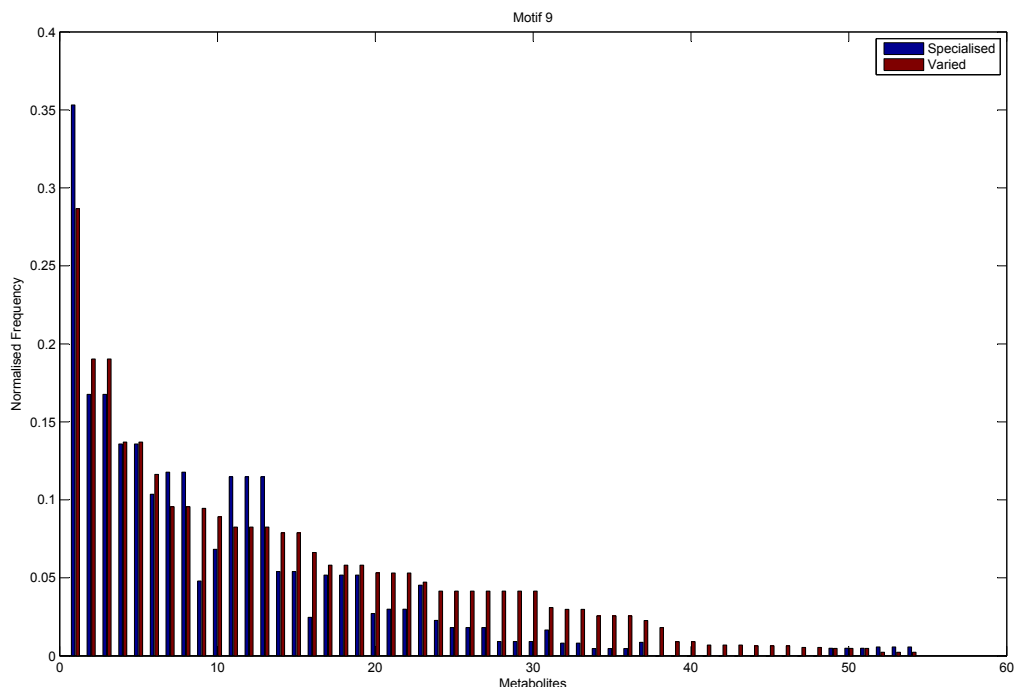


Figure 4.8: Mean normalised frequency for the 54 metabolites obtained within motif 9 for the 115 metabolic networks. Blue bars represent the specialised class and the red bars represent the varied class. Here, the metabolites are in descending order of the metabolite frequencies for the varied class.

structures present across the entire network ensemble, i.e. motifs 5 and 9, also differed significantly. Figures 4.6 and 4.7 show the mean frequency for metabolites occurring within motif 5 for the 115 metabolic networks, again grouped into the specialised (blue bars) and varied classes (red bars). Note that the frequencies plotted in Figures 4.6 and 4.7 have been normalised to remove any bias due to network size (the normalisation procedure is described in more detail towards the end of this section), and metabolites are displayed in decreasing order according to the varied class. Figure 4.6 displays the distribution for those 263 metabolites that occurred at least once within motif 5 across the two classes under consideration. Interestingly, we see that the distribution for the varied class is relatively broad, with a large number of metabolites occurring with a relatively low frequency, whereas the distribution for the specialised class is more akin to a *scale-free* or *power-law* distribution, consisting of a small set of relatively high frequency metabolites.

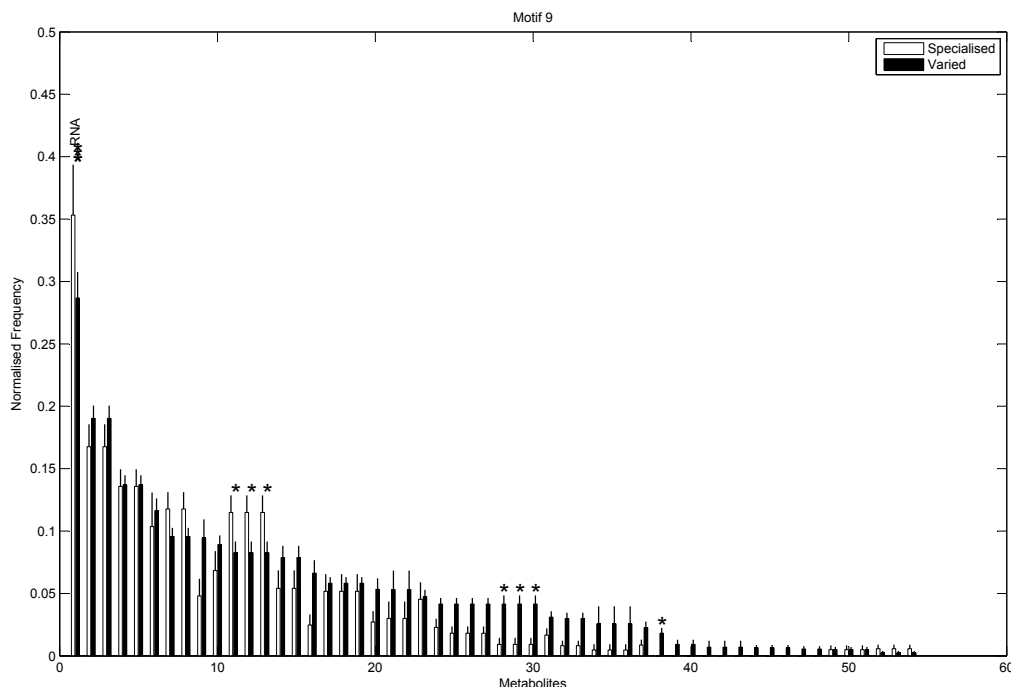


Figure 4.9: Mean normalised frequency for the 54 metabolites obtained within motif 9 for the 115 metabolic networks. Vertical bars are standard errors. Asterisks indicate levels of significance, with \*, \*\*, and \*\*\* corresponding to  $p < 0.05$ ,  $p < 0.01$  and  $p < 0.001$ , respectively. Metabolite names are provided for the most significant metabolites.

Next we used a Chi-square test (Fisher's exact test,  $p < 0.01$ ) to explore the differences in proportions of the individual metabolites between the two groups. Figure 4.7 shows the 47/263 metabolites for which a significant difference in proportions was found in motif 5, again displayed according to decreasing frequency of the varied class. Metabolites displaying the most significant differences (Fisher's Exact test,  $p < 10^{-5}$ ) included (2R)-2-Hydroxy-3-(phosphonoxy)-propanal, Tetrahydrofolate and Isopentenyl diphosphate, all of which were overrepresented in the specialised group compared to the varied group. Note that the aforementioned overrepresented metabolites are required for biosynthesis of various amino acids, folates and terpenoids and are also responsible for the regulation of carbohydrate metabolism in many bacterial species.

Note that similar results were found for motif 9. Figure 4.8 shows the mean frequency for metabolites occurring within motif 9 for the 115 metabolic networks, grouped into the specialised (blue bars) and varied (red bars) classes. Here metabo-

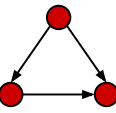
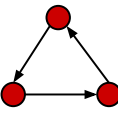
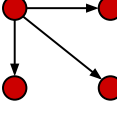
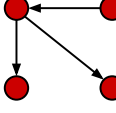
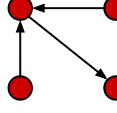
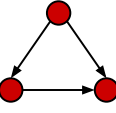
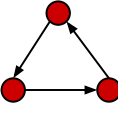
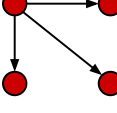
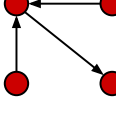
Environment	Significant motifs				
Aerobic & Facultative (334)	 motif 5	 motif 9	 motif 14	 motif 15	 motif 26
Anaerobic (49)	 motif 5	 motif 9	 motif 14	 motif 26	

Table 4.5: Motifs significant to networks with differing oxygen requirements.

lites are displayed in decreasing order according to the varied class. Figure 4.8 shows the 54 metabolites that were found at least once across the 115 metabolic networks. We find that the distribution of metabolites is slightly broader for the varied class, similar, but less prominent, to the results obtained for motif 5. Using Chi-square tests (Fisher's Exact test) we explored group differences for the individual metabolites. Figure 4.9 identifies only one metabolite, RNA, for which significant differences were found (Fisher's Exact test,  $p < 0.001$ ).

### Oxygen Requirements

Similar to the above, we then investigated which motifs were driving the observed differences between metabolic networks that evolved in the presence or absence of oxygen. Again, for simplicity we divided the bacteria into two separate groups: anaerobic and aerobic (including facultative aerobes). The significant motifs are displayed in Table 4.5. For aerobic networks 5/212 possible motifs were found to be significant, whilst for the anaerobic networks 4/212 were found to be significant. Again, motifs 5 and 9 were significant across the entire cohort, along with motifs 14 and 26 in this instance. The only motif that differed between the two groups was motif 15, which was specific to the aerobic class. Interestingly, the study by Raymond and Segré [143] found that the effects of oxygen exposure on metabolic network structure was most prolific at the periphery of the network, that is, network alterations were largely due to the addition of new reactions and pathways, rather than network rewiring. Thus, the enrichment of motif 15 is a natural consequence, as it acts as a branch point on these newly formed peripheral reactions and pathways.

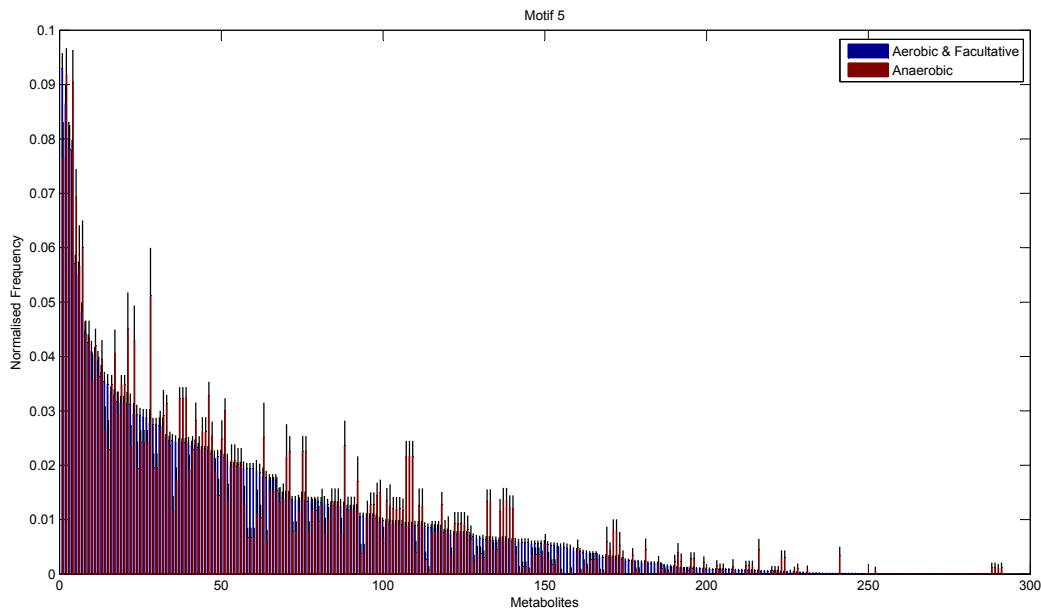


Figure 4.10: Mean normalised frequency for the 291 metabolites obtained within motif 5 for the 383 metabolic networks. Blue bars represent the aerobic-facultative class and red bars represent the anaerobic class. Metabolites are displayed in descending order of the metabolite frequencies for the aerobic-facultative class.

Figures 4.10 and 4.11 show the distribution of metabolites across motif 5 for the two groups, ordered according to decreasing metabolite frequency for the aerobic class (blue bars). Note, that the aerobic class exhibits a fairly broad distribution, whilst the anaerobic distribution tails off slightly quicker, in a similar but less pronounced manner to that displayed by the specialised bacteria in Figure 4.6. Figure 4.11 shows those metabolites that displayed a significant group difference. Interestingly, the majority of metabolites, some 37/52, were found to be overrepresented in the aerobic group compared to the anaerobic group, the most significant of which were Isopentenyl diphosphate, Fatty acid, trans-Farnesyl diphosphate, Phosphatidylethanolamine, Phosphatidylserine, L-Threonine, L-2-Amino-3-oxobutanoate, 2-Acyl-sn-glycero-3 phosphocholine, Phosphatidylcholine, 3'.5'-Cyclic GMP (Fisher's Exact test,  $p < 10^{-5}$ ). These metabolites are known to be involved in the biosynthesis of a range of amino acids and secondary metabolites.

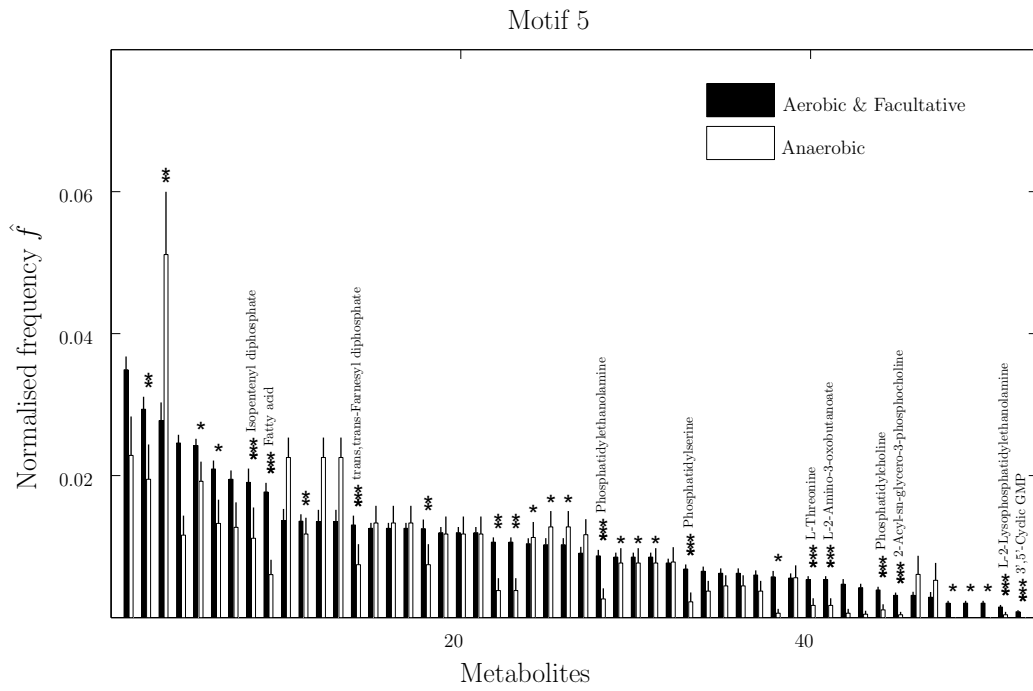


Figure 4.11: Mean normalised frequency for the significant metabolites obtained within motif 5 with  $p < 0.01$  (Fisher's Exact test). Vertical bars are standard errors. Asterisks indicate large significant differences between metabolic networks from the aerobic-facultative and anaerobic class, where \*, \*\*, and \*\*\* correspond to  $p < 0.001$ ,  $p < 0.0001$  and  $p < 0.00001$ . Metabolite names are provided for the most significant metabolites.



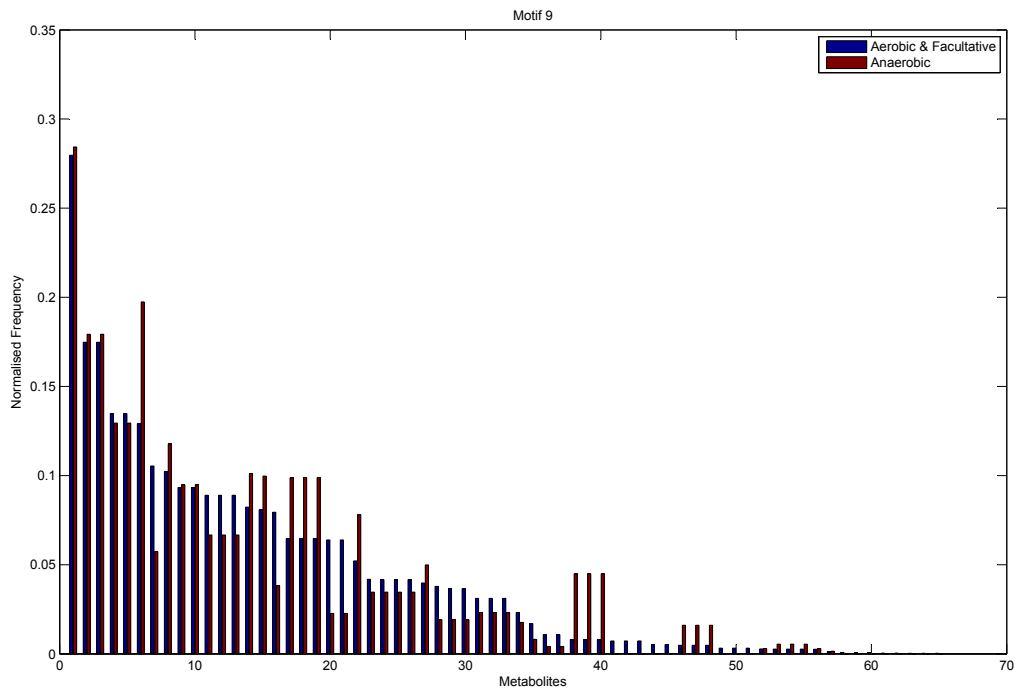


Figure 4.12: Mean normalised frequency for the 65 metabolites obtained within motif 9 for the 383 metabolic networks. Blue bars represent the aerobic-facultative class and the red bars represent the anaerobic class. Here, the metabolites are in descending order of the metabolite frequencies for the aerobic-facultative class.

Again, similar results were found for motif 9. Figure 4.12 shows the mean frequency for metabolites occurring within motif 9 for the 383 metabolic networks that evolved in either the presence or absence of oxygen. Here metabolites are displayed in decreasing order according to the aerobic-facultative class (blue bars). Figure 4.12 shows the 65 metabolites that were found at least once across the two classes. Note that the distribution for the aerobic-facultative class and anaerobic class for motif 9 are a lot closer than that obtained for motif 5. Figure 4.13 shows that the metabolites with the most significant differences (Fisher's Exact test,  $p < 0.001$ ) included Glutathione, L-Arginine, L-Citrulline, N-(L-Arginino)succinate, Succinate, Succinyl-CoA and O-Succinyl-L-homoserine.

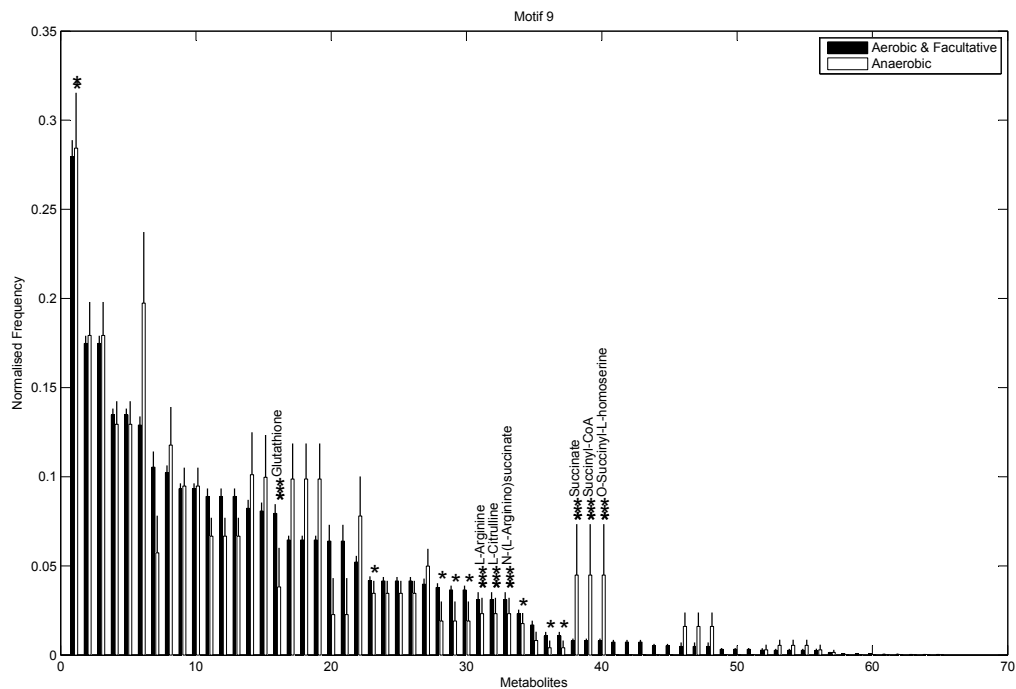


Figure 4.13: Mean normalised frequency for the 65 metabolites obtained within motif 9 for the 383 metabolic networks. Vertical bars are standard errors. Asterisks indicate levels of significance, with \*, \*\*, and \*\*\* corresponding to  $p < 0.05$ ,  $p < 0.01$  and  $p < 0.001$ , respectively. Metabolite names are provided for the most significant metabolites.

### Determining Significant Metabolites

We conclude this section by providing a few additional details regarding the statistical analysis performed above. Note, that when considering the differences between the frequency of metabolites occurring in a motif of interest (5 or 9 in our case) care must be taken to eliminate the influence of network size on the analysis. This bias is due to the increased number of motifs exhibited by larger networks which naturally leads to greater frequencies of metabolites. Thus, given a network  $i$  and a metabolite  $j$ , we denote by  $f_{i,j}$  the frequency with which metabolite  $j$  appears within the motif of interest, motif  $q$  say, for the  $i$ th network. Now, in order to remove any bias due to network size we normalise the statistic  $f_{i,j}$  by dividing it by the frequency with which motif  $q$  appears in network  $i$ , which we denote by  $f_{i,\text{mot}_q}$ . This then leads to the following normalised statistic:

$$\hat{f}_{i,j} = \frac{f_{i,j}}{f_{i,\text{mot}_q}},$$

describing the relative importance of metabolites via their participation within specific motifs. It is the normalised frequency  $\hat{f}$  that is displayed in Figures 4.6–4.13.

## 4.5 Summary

In this Chapter, we have introduced a new graph embedding approach for studying large numbers of networks, of possibly differing order, and employed it to investigate the effect of environmental variability on the metabolic network structure of a large cohort of bacterial species. As a proof of principle, we first applied the new approach to a large ensemble of networks consisting of both bacterial networks and a range of standard network models (Erdős-Rényi networks, small-world networks and scale-free networks), providing evidence that the technique can correctly differentiate between various types of networked structures. Moreover, applying the new technique to a large cohort of metabolic networks, we found evidence supporting the view that organisms that evolve in more uncertain environments exhibit more complex metabolic connectivity structures than those evolving under more stable

conditions. Note, that the motif based approach forwarded here strongly supports the view that environmental conditions play a pivotal role in shaping the resultant metabolic networks, and is robust in the sense that the patterns described in Figures 4.4 and 4.5 are reproducible in both the latest and older, less complete versions of the data [176] (data not shown). This is in contrast to recent studies in which network features that were found to correlate with environmental variability (e.g. modularity) disappeared when tested on newer versions of the data [130, 171]. Importantly, these findings suggest that alterations in the motif signature provide a robust indicator of adaptability and evolvability in bacterial metabolic networks.

# CHAPTER V

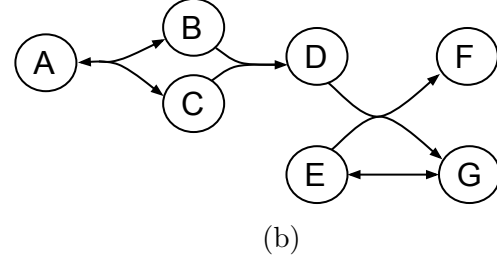
## HYPERNETWORK MODELS OF METABOLISM

In this Chapter we start by giving a brief overview of some recent works in network science that use hypernetwork models, with an emphasis on biological applications. We then provide some preliminaries detailing the extensions of a number of network definitions and theories to this more complicated setting, including our own extension of the concept of network reciprocity. Finally, we extend a number of ideas from percolation theory to hypernetworks in order to quantify metabolic hypernetwork robustness, and provide a biological interpretation of our results.

### 5.1 Complex Hypernetworks

Whilst the topological characterisation of complex networks has received considerable attention over the past decade [177, 178], the theory of complex hypernetworks is far less developed, which, coupled with the increased algorithmic complexities that accompany such an approach, perhaps explains why this more natural framework has not been more widely adopted in the study of biological networks to date. Recently, however, a number of studies have attempted to extend complex network reasoning to this more complicated setting. For example, the commonly used clustering coefficient, a measure of the probability that any two neighbours of a given

**Reaction 1:**  $A \rightarrow B + C$   
**Reaction 2:**  $B + C \rightarrow D$   
**Reaction 3:**  $D + E \rightarrow F + G$   
**Reaction 4:**  $E \rightarrow G$   
**Reaction 5:**  $B + C \rightarrow A$   
**Reaction 6:**  $G \rightarrow E$



$$C(H) = \begin{matrix} & R1 & R2 & R3 & R4 & R5 & R6 \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \end{matrix} & \begin{pmatrix} -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 \end{pmatrix} \end{matrix}$$

(c)

$$A(H) = \begin{matrix} & A & B & C & D & E & F & G \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

(d)

Figure 5.1: Hypergraph model of metabolism: (a) an example of a hypothetical reaction system; (b) the accompanying hypergraph model; (c) its incidence matrix representation; and (d) its adjacency matrix representation.

node are also neighbours, has been extended to hypernetworks [32, 179]. In [32] in particular, it was shown that the inverse scaling between network degree and clustering, typically reported in standard network analyses of metabolism and considered indicative of a hierarchical network structure [180], may actually be an artefact due to misrepresentation. Another important topological parameter that has been generalised to this more complicated setting is the subgraph centrality [179]. Centrality measures provide a measure of the relative importance of each node within a network, and the generalisation given in [179] provides such a characterisation for the nodes of a hypernetwork. Other notable works include the extension of random graph models such as Erdős-Rényi and Barabási-Albert to hypernetworks [181]; the use of random walks to infer information flow and network architecture [182, 183]; and novel community detection algorithms for determining modular hypernetwork structure [184, 185].

### 5.1.1 Hypernetwork Preliminaries

A complex hypernetwork can be described by a pair of objects  $H = (V, E)$ , where  $V = \{v_1, v_2, \dots, v_n\}$  is a set of  $n$  vertices and  $E = \{E_1, E_2, \dots, E_m\}$  the corresponding hyperedge set. Each hyperedge consists of subsets of  $V$ , such that  $\bigcup_i E_i = V$  and  $E_i \neq \emptyset$ . We say that two vertices  $v_i$  and  $v_j$  are *adjacent* if they are contained within the same hyperedge, i.e.  $v_i, v_j \in E_k$ . For a directed hypernetwork each hyperedge is further divided into two subsets (the tail set  $X$  and the head set  $Y$ ) allowing us to distinguish between bidirectional and unidirectional relationships. Directed hyperedges are more commonly known as *hyperarcs*.

A hypernetwork can be represented by a variety of different matrices [186], the most popular of which is the *incidence matrix*, an  $n \times m$  matrix  $C(H)$  representing the relationships between the  $n$  nodes and  $m$  hyperarcs. The entries of the matrix  $C(H)$  are given by

$$C_{ij} = \begin{cases} 1, & \text{if } v_i \in E_j, \\ 0, & \text{otherwise,} \end{cases}$$

that is,  $C_{ij}$  equals 1 if the node  $v_i$  belongs to the  $j$ th hyperedge. In the case of directed hypernetworks, the incidence matrix is able to distinguish between nodes that belong to the tail/head set of a hyperarc, such that the entries are given by

$$C_{ij} = \begin{cases} -1, & \text{if } v_i \in X_j, \\ 1, & \text{if } v_i \in Y_j, \\ 0, & \text{otherwise.} \end{cases}$$

Alternatively, a directed hypernetwork can be represented by two incidence matrices - the negative (or outer) incidence matrix  $C_-(H)$  and the positive (or inner) incidence matrix  $C_+(H)$ , representing the tail sets and head sets of the hyperarcs respectively. That is, the elements of  $C_-(H)$  are equal to 1 if  $v_i \in X_j$  and 0 otherwise. The elements of  $C_+(H)$ , on the other hand, are equal to 1 if  $v_i \in Y_j$  and 0 otherwise [183].

Importantly, given the incidence matrix it is straightforward to compute the

*adjacency matrix* of a hypernetwork as follows

$$A(H) = C(H)C(H)^T. \quad (5.1)$$

More formally, the elements of  $A(H)$  are defined as

$$A_{ij} = |\{E_k \in E : v_i, v_j \in E_k\}|, \text{ for } v_i, v_j \in V, i \neq j. \quad (5.2)$$

Note that we can use the outer and inner incidence matrices to generalise the above to the case of a directed hypernetwork as

$$A(H) = C_-(H)C_+(H)^T. \quad (5.3)$$

Again, this can be written more formally as

$$A_{ij} = |\{E_k \in E : \{v_i \in X_k, v_j \in Y_k\} \subset E_k\}|, \text{ for } v_i, v_j \in V, i \neq j,$$

which, in words, states that the  $ij$ th element is given by the cardinality of the set of hyperarcs, such that  $v_i$  belongs to the tail set (i.e.  $v_i \in X_k$ ) and  $v_j$  belongs to the head set (i.e.  $v_j \in Y_k$ ). It should be noted here that, unlike the incidence matrix, the adjacency matrix loses information regarding the hypergraph's structure. That is, an adjacency matrix cannot be used to determine whether a group of nodes belong to the same hyperedge. Instead, the adjacency matrix is akin to a multigraph, such that it allows for multiple links between nodes, since it contains non-negative integers as its entries.

The next definition that we require is that of the *underlying hypernetwork*, which is the hypernetwork  $U(H)$  that results after removing the directionality of all hyperarcs. The underlying hypernetwork provides an elegant way to normalise directed hypernetwork measures, and thus allows us to compare across directed hypernetworks of differing order. Figure 5.2 shows an example of a hypernetwork together with its corresponding underlying hypernetwork.



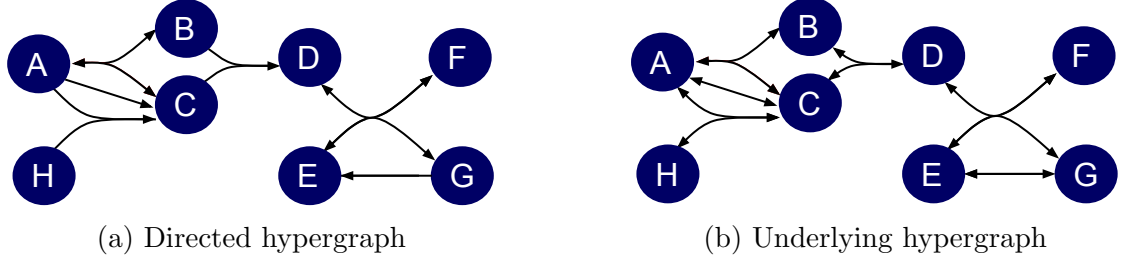


Figure 5.2: Illustration of a directed hypergraph with its corresponding underlying hypergraph.

### 5.1.2 Current Complex Hypernetwork Measures

Next we give a brief overview of the complex hypernetwork analogues of some of the measures introduced in Chapter 2 with a particular focus on the extension of some of the most familiar network measures, namely the degree distribution and the clustering coefficient.

#### Hypernetwork Degree Distributions

Perhaps the simplest measure to extend is that of degree. The degree  $k(v_i)$  of a node  $v_i$  in a hypernetwork is defined as the cardinality of the set of hyperedges, such that  $v_i \in E_k$ . Note that a vector of hypernetwork node degrees is easily computed using the incidence matrix as follows,

$$\mathbf{k}(V) = C(H)\mathbf{e}^T,$$

where  $\mathbf{e}$  is a  $1 \times m$  vector of all ones. In the case of a directed hypernetwork, we consider two types of node degrees: the in-degree,  $k^{\text{in}}(v_i)$ , which is defined as the cardinality of the set of hyperarcs containing  $v_i$  in the head set, and the out-degree,  $k^{\text{out}}(v_i)$ , which is defined as the cardinality of the set of hyperarcs containing  $v_i$  in the tail set. More formally, we have the following

$$k^{\text{out}}(v_i) = |\{E_k \in E : v_i \in X_k\}|$$

and

$$k^{\text{in}}(v_i) = |\{E_k \in E : v_i \in Y_k\}|.$$

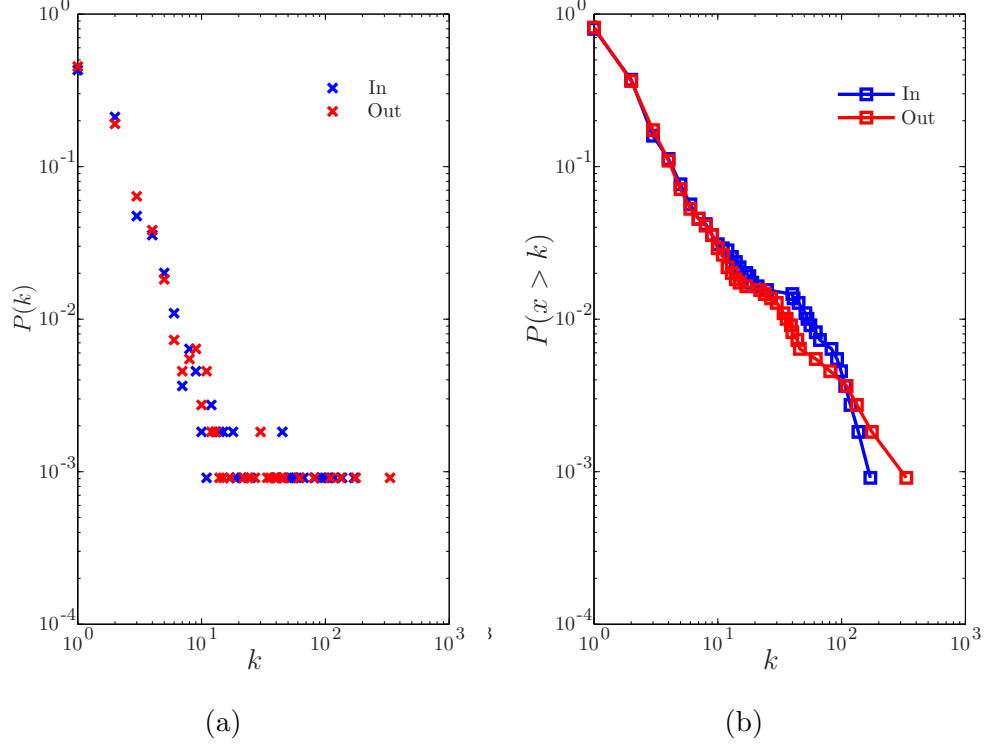


Figure 5.3: The in- and out-degree distributions (a) and the in- and out- cumulative degree distributions (b) for the directed hypergraph of *E. coli*.

Similar to the undirected case, vectors of in- and out-degrees are easily computed using the positive and negative incidence matrices, in the following way

$$\mathbf{k}^{\text{out}}(V) = C_-(H)\mathbf{e}^T \quad \text{and} \quad \mathbf{k}^{\text{in}}(V) = C_+(H)\mathbf{e}^T.$$

Figure 5.3 illustrates the above in the case of *E. coli*: 5.3a displays the degree distribution whilst 5.3b displays the cumulative degree distribution. In this example it is seen that both the in- and out-degree distributions are very similar, and appear to follow a power-law distribution.

In addition to the standard degree one can consider the so-called hyperedge cardinality, and plot its distribution. The hyperedge cardinality counts the number of nodes contained within each hyperedge, that is

$$k(E_j) = |\{v_i \in V : v_i \in E_j\}|.$$

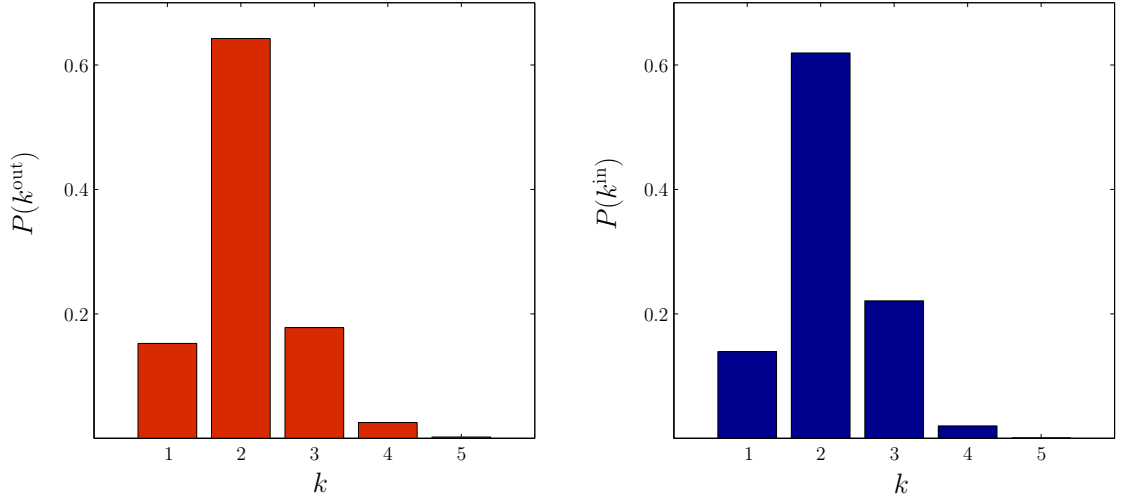


Figure 5.4: The hyperarc out-degree (red) and in-degree (blue) distributions of *E.coli*.

Similar to the node degrees, a vector of hyperedge degrees can be easily computed from the incidence matrix as,

$$\mathbf{k}(E) = \mathbf{e}C(H),$$

where here  $\mathbf{e}$  is a  $1 \times n$  vector of all ones. In the case of directed hypernetworks, we again consider two types of degrees, the out-degree  $k^{\text{out}}(E_i)$ , corresponding to the number of substrates in the hyperarc, and the in-degree  $k^{\text{in}}(E_i)$ , corresponding to the number of products in the hyperarc. Mathematically, we have the following

$$k^{\text{out}}(E_j) = |\{v_i \in V : v_i \in X_j\}|$$

and

$$k^{\text{in}}(E_j) = |\{v_i \in V : v_i \in Y_j\}|.$$

Again, vectors of out- and in-degrees can easily be computed using the negative and positive incidence as follows

$$\mathbf{k}^{\text{out}}(E) = \mathbf{e}C_-(H) \quad \text{and} \quad \mathbf{k}^{\text{in}}(E) = \mathbf{e}C_+(H).$$

In Figure 5.4 we plot the hyperarc out-degree (red) and in-degree (blue) distri-

butions for the hypernetwork of *E.coli*. Again, we find that the out- and in-degree distributions are very similar, both of which have a mean degree of 3 and a range from 1 to 5. Unlike the node-degree distributions, the hyperarc cardinality distributions appear to be more akin to a poisson or binomial distribution.

### Hypernetwork Clustering Coefficients

As mentioned previously several clustering coefficients have been proposed for hypernetworks. The first attempt was by Estrada and Rodriguez [187] who provided a generalisation of the network transitivity measure to hypernetworks. Recall that the transitivity coefficient of a network is defined as

$$T(G) = \frac{3 \times \text{the number of triangles}}{\text{number of paths of length 2}} = \frac{3C_3(G)}{P_2(G)},$$

where  $C_3$  is the number of triangles and  $P_2$  the number of paths of length 2.

To generalise the above to hypernetworks, Estrada and Rodriguez defined a hyper-triangle to be a sequence of three distinct nodes and hyperedges, such that the nodes are mutually adjacent, i.e.  $v_i E_p v_j E_q v_k E_r v_i$ . Similarly, they defined a two-path to be a sequence of the form  $v_i E_p v_j E_q v_k$ ; here all hyperedges and nodes are distinct. It is then tempting (recalling the methods described in Chapter 2) to calculate the ratio of triangles to two-paths in a hypernetwork using the adjacency matrix as before. Unfortunately, proceeding in this manner leads to incorrect results due to the inclusion of so-called *false hyper-triangles*. Such false triangles result when closed walks of length three are formed within one (or two) hyperedges. Thus due to the construction of the adjacency matrix (see Equation (5.2)) the usual methods of counting triangles severely over counts and so any sensible definition must account for this.

To this end Estrada and Rodriguez put forward the following definition of the hypernetwork transitivity coefficient

$$T(G_H) = \frac{6 \times \text{the number of hyper-triangles}}{\text{number of hyper-paths of length 2}} = \frac{6C_3(G_H) - 6t}{P_2(G_H) - 6t}.$$

Here  $t$  denotes the number of false hyper-triangles which is defined as

$$t = \sum_{j=1}^m (-1)^{j+1} a_j,$$

where  $a_k = \sum_{\alpha_{i_1, i_2, \dots, i_k}} \binom{\alpha_{i_1, i_2, \dots, i_k}}{3}$  and  $\alpha_{i_1, i_2, \dots, i_k} = \left| \bigcap_{r=1}^k E_{i_r} \right|$ . Note that this is a consequence of the inclusion-exclusion principle.

More recently, Zhou and Nakhleh [32] have proposed extensions of both the local and global clustering coefficient to undirected hypernetworks employing the notion of hyperedge extra overlap.

Given a hypernetwork  $H$  the extra overlap for any pair of overlapping hyperedges is given by

$$EO(E_i, E_j) = \frac{|N(D_{ij}) \cap D_{ji}| + |N(D_{ji}) \cap D_{ij}|}{|D_{ij}| + |D_{ji}|}, \quad (5.4)$$

where  $D_{ij} = E_i - E_j$  denotes the set difference, and  $N(v_k)$  is the neighbourhood of the node  $v_k$ , i.e.

$$N(v_k) = \{v_l | \{v_l, v_k\} \subseteq E_j \text{ for some } E_j \in E\}.$$

To extend this definition to sets, one simply takes the union of the neighbourhoods of all nodes contained within the set. Informally, the extra overlap between a pair of hyperedges is the fraction of vertices that are connected via a separate hyperedge (or hyperedges); see Figure 5.5 for an illustration of the extra overlap for a toy hypernetwork on five nodes.

Given Equation (5.4) we can then define the following local clustering coefficient

$$HC_{\text{local}}(v_k) = \begin{cases} \binom{|\mathcal{M}(v_k)|}{2}^{-1} \sum_{\substack{E_i, E_j \\ \in \mathcal{M}(v_k)}} EO(E_i, E_j), & \text{if } k(v_k) > 1, \\ 0, & \text{if } k(v_k) = 1. \end{cases} \quad (5.5)$$

Here,  $k(v_k)$  denotes the degree of node  $v_k$  and  $\mathcal{M}(v_k)$  is the set of hyperedges containing  $v_k$ . The above definition naturally extends to a global clustering coefficient

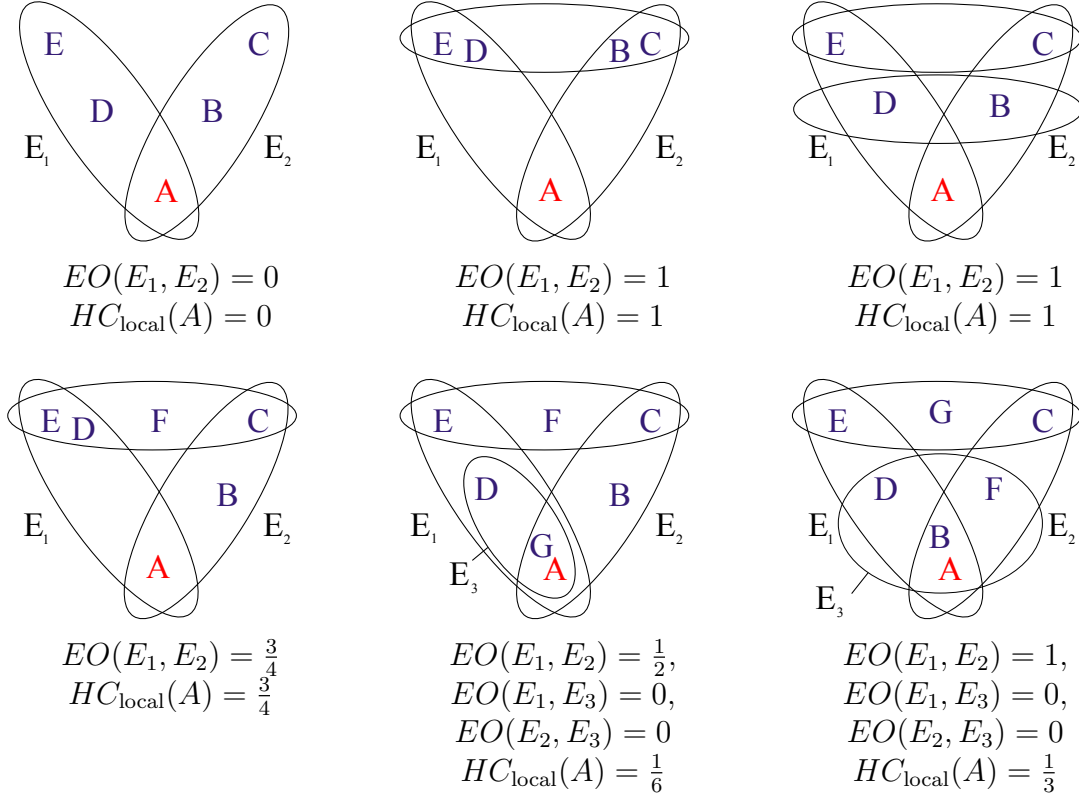


Figure 5.5: Toy hypergraph models to illustrate the local clustering coefficient,  $HC_{\text{local}}$ , for a variety of different scenarios.

as follows

$$HC_{\text{global}}(H) = \begin{cases} \frac{1}{|\mathcal{I}|} \sum_{\substack{E_i, E_j \\ \in \mathcal{I}}} EO(E_i, E_j), & \text{if } \mathcal{I} \neq \emptyset, \\ 0, & \text{if } \mathcal{I} = \emptyset. \end{cases} \quad (5.6)$$

Here,  $\mathcal{I} = \{\{E_i, E_j\} \subset E : E_i \cap E_j \neq \emptyset \text{ and } E_i \neq E_j\}$ , is the set of hyperedge pairs with non-empty intersection. Informally, the local clustering coefficient is simply the hyperedge extra overlap, averaged over those hyperedges containing the node of interest; whereas the global clustering coefficient averages the extra overlap over all hyperedge pairs containing at least one common vertex.

Importantly, Zhou and Nakhleh [32] applied their method to the metabolic hypernetwork of *E. coli* and found a number of inconsistencies between their work and previous studies. In particular, they found that the power law scaling between the clustering coefficient and network degree disappears when applying their new measure  $HC_{\text{local}}$  to metabolic networks represented as hypernetworks. In Figure 5.6

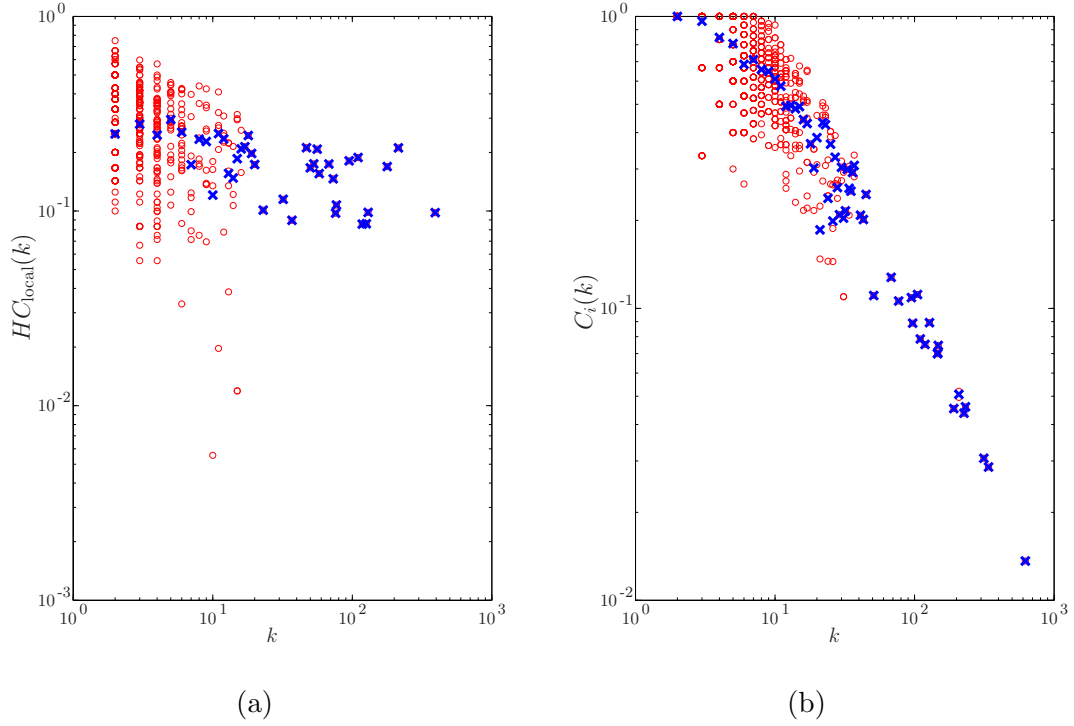


Figure 5.6: The scaling of the local clustering coefficient for (a) the metabolic hypergraph and (b) the metabolic standard graph of *E. coli*. Note that the blue crosses are the averaged local clustering coefficients for each unique degree, whereas the red circles are the local clustering coefficients.

we repeat the analysis performed by Zhou and Nakhleh on our data (see Chapter 3 and the next section for a detailed description of the data) by plotting the clustering coefficient versus degree for (a) the standard network and (b) the hypernetwork representation of *E. coli*. Importantly, our results are consistent with those reported in [32], in that the scaling is far less evident when the hypernetwork clustering measure is used. However, this result is far from conclusive since (i) only one metabolic network has been used, and (ii) the definition of the hypernetwork clustering coefficient used here is not necessarily a meaningful one. A local version of the hypernetwork transitivity coefficient described above, for instance, could provide a more accurate measure of the hypernetworks structure.

## 5.2 Novel Measures of Metabolic Hypernetwork Robustness

Having discussed some preliminaries in the theory of complex hypernetworks we wish to now consider hypernetwork models of metabolism. As before, we concentrate on bacterial species. A hypernetwork model of metabolism consists of  $n$  nodes representing the metabolites and  $m$  hyperedges, one for each chemical equation. Note that in the directed case, each hyperedge is further subdivided into tail and head sets in order to account for the substrates and products of the reactions, respectively.

To reconstruct metabolic hypernetworks, we used the procedure described in Chapter 3. The metabolic data for 115 different organisms was downloaded from the KEGG database on the 12th October 2013. First, the reaction lists for each organism were imported into MATLAB, and the data was searched for any mistakes and inconsistencies (e.g. repeated reactions). Using string comparisons, repeated reactions could be identified using their reaction ID's, and then the substrates and products could be checked using their compound ID's. Modification/removal of the repeated reactions was carried out using the rules described in Chapter 3.

The data for each organism was then represented by an incidence matrix, which allows for the easy construction of the adjacency matrix using Equation (5.1). All self-loops are removed from the adjacency matrix by setting the diagonal elements to zero. Note that in the case of directed hypernetworks, a pair of incidence matrices are employed, the negative incidence matrix ( $C_-(H)$ ) and the positive incidence matrix ( $C_+(H)$ ), which, as before, can then be used to form the corresponding adjacency matrix. It is worth noting that in the directed case, reversible reactions need to be considered separately, that is, a reaction of the form  $A + B \leftrightarrow C + D$  is treated as  $A + B \rightarrow C + D$  and  $C + D \rightarrow A + B$ , each of which will have its own corresponding column in the incidence matrices. The negative incidence matrix contains the information regarding the substrates in each reaction (i.e.  $C_-(i, j) = 1$  if metabolite  $i$  is a substrate in the  $j$ th reaction), whereas the positive incidence matrix contains the information regarding the products in each reaction (i.e.  $C_+(i, j) = 1$  if metabolite  $i$  is a product in the  $j$ th reaction).



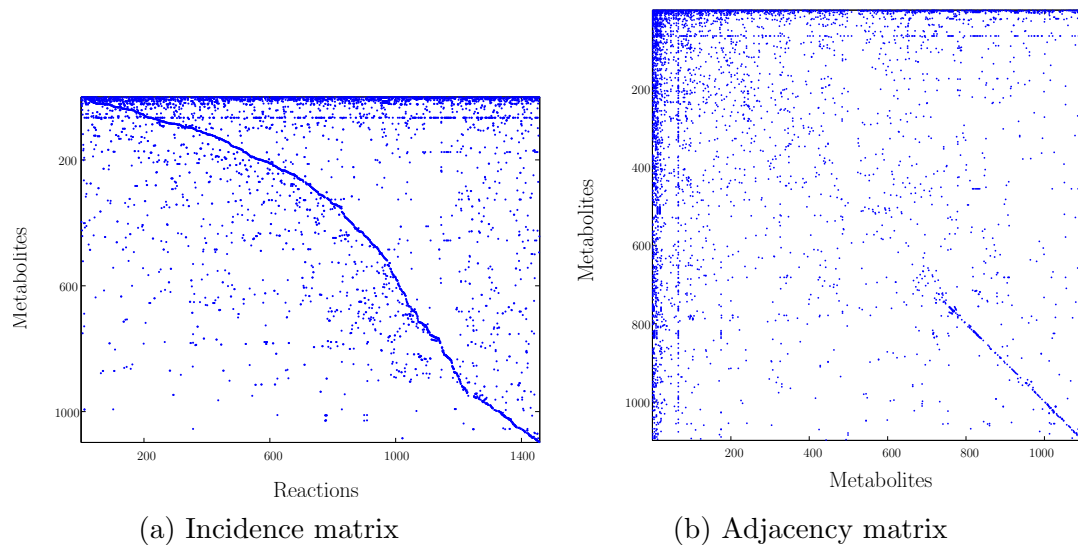


Figure 5.7: Spy plots of (a) the incidence matrix and (b) the adjacency matrix of the directed hypergraph representation of metabolism in *E. coli*.

Environment	Nodes			Hyperedges			Hyperarcs		
	min	median	max	min	median	max	min	median	max
Obligate (34)	224	441	979	143	337	883	197	443	1156
Specialised (5)	643	695	743	554	627	651	707	805	841
Aquatic (4)	754	851	1014	645	747	896	819	944	1146
Facultative (41)	244	947	1308	153	883	1199	204	1155	1510
Multiple (28)	631	900	1226	545	821	1143	712	1078	1468
Terrestrial (3)	890	942	955	832	912	936	1086	1205	1219
Total (115)	224	748	1308	143	695	1199	197	895	1510

Table 5.1: Network statistics for the reaction graphs of the 115 bacterial species studied in this work, classified according to environmental variability.

Figure 5.7 shows spy plots of the two matrix representations described above for a directed hypernetwork model of metabolism for *E. coli*. The hypernetwork of *E. coli* has 1097 nodes and 1457 hyperarcs. In Table 5.1 we provide an overview of the basic network statistics (i.e. number of nodes and edges) for the hypernetworks of the 115 bacteria that have been reconstructed. Note that the bacteria are the same as studied in Chapter 4, and so again are grouped according to environmental variability.

### 5.2.1 Hypernetwork Reciprocity as a Measure of Chemical Equilibrium

In this section we extend the reciprocity measure [188] to the more general case of hypernetworks and use it to quantify the effect of environmental variability on the set of available chemical reactions for a large cohort of bacterial species, and, in particular, the extent to which reversible reactions are present.

#### Reciprocity in Complex Hypernetworks

Reciprocity in standard networks measures the proportion of mutual relationships within the network, or, in other words, the probability that an edge from B to A exists given that an edge from A to B exists. More formally, reciprocity is defined as

$$r = \frac{L^{\leftrightarrow}}{L}, \quad (5.7)$$

where  $L^{\leftrightarrow}$  is the number of bidirectional edges and  $L$  is the total number of edges in the graph. Now, in order to extend the above measure of reciprocity to directed hypernetworks it is useful to recount the following theorem, which can be found, for example, in [189].

**Theorem 1.** *The number of walks ( $i \neq j$ ) or closed walks ( $i = j$ ) of length  $k$  in a directed hypernetwork is equal to the  $(i, j)^{th}$  element of the matrix  $A^k$ .*

Despite the usefulness of the above theorem, using the adjacency matrix  $A(H)$  does not uniquely define a walk in a hypergraph, for the reasons discussed previously in §5.1.1. However, since reciprocity is not concerned with the individual paths taken in each walk but rather their count, the above theorem can be used to rewrite Equation (5.7) in terms of the adjacency matrix, and thus to generalise reciprocity to the case of directed hypernetworks as follows:

$$r(H) = \frac{\text{trace}(A^2)}{\text{trace}(U^2)}. \quad (5.8)$$

Here  $A$  and  $U$  are the adjacency matrices of the directed hypernetwork and underlying directed hypernetwork, respectively, and the trace of a matrix is the sum of the

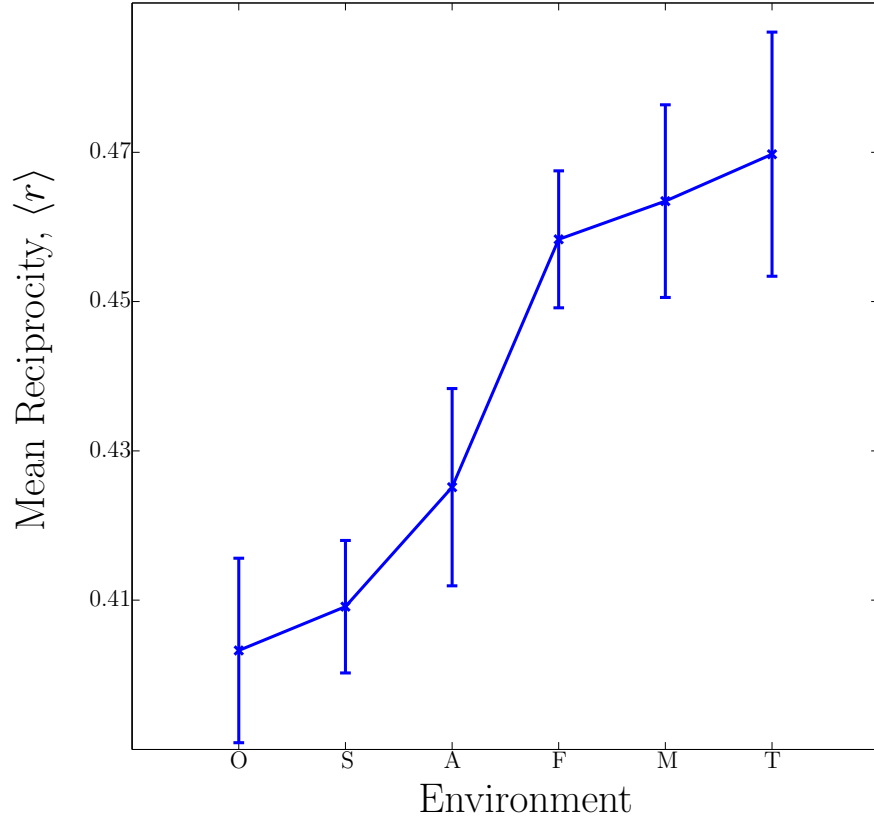


Figure 5.8: Relationship between average hypernetwork reciprocity,  $\langle r \rangle$ , and environmental variability. The six bacterial habitats along the x-axis are in order of environmental variability according to the NCBI classification scheme: **O**bligat, **S**pecialised, **A**quatic, **F**aculative, **M**ultiple and **T**errestrial. Here vertical bars denote the standard error of the mean.

diagonal elements. Note that in the case of a standard graph, Equations (5.7) and (5.8) are equivalent. Importantly, unlike the standard reciprocity measure, Equation (5.8) includes information regarding the number of mutual hyper-connections that exist between a pair of vertices.

Figure 5.8 shows a plot of the average hypernetwork reciprocity,  $r(H)$ , versus environmental variability for the different bacterial networks. Note that the average here is taken over each of the 6 environmental classes: obligate, specialised, aquatic, facultative, multiple and terrestrial. Importantly, we found that the hypernetwork reciprocity increased significantly with environmental variability. The lowest value of reciprocity is found for the bacteria within the obligate class, followed by a slight increase for the specialised class, and then again slightly higher for the aquatic class, there is then a relatively steep increase to the facultative, multiple and terrestrial

classes with a relatively small increase between each class. The group differences shown in Figure 5.8 are significant by the Kruskal-Wallis (KW) test ( $p\text{-value} < 10^{-4}$ ).

The results obtained using the new reciprocity measure (Figure 5.8) support the idea that habitat lifestyle plays an important role in an organism's hypernetwork topology. This result is consistent with a number of studies, including the work carried out in Chapter 4, that have found a relationship between network structure and variations in the environment, using standard network models [129, 142, 169, 170, 190]. Importantly, the reciprocity measure applied to the standard network representation does not support this result (see Figure 5.9) suggesting that the hypernetwork formalism, and our new measure of reciprocity, in particular, adds significant value beyond the standard approach, as well as providing further evidence that habitat variability has a significant impact on metabolic network structure. In the current context, the relationship found in Figure 5.8 can be viewed as an evolutionary adaptation caused by a larger amount of uncertainty present within a more varied environment, and thus the ease by which metabolites are reciprocated is higher to enable greater adaptability to fluctuations within the environment.

### 5.2.2 Percolation in Complex Metabolic Hypernetworks

In this section we adapt widely studied percolation-based approaches [90, 177] in order to probe complex metabolic hypernetwork topology and quantify the robustness and fragility of these systems (refer to Chapter 2 for an introduction to percolation and network robustness). Historically, a number of studies have investigated the resilience of metabolic networks to random mutations (and targeted attacks) [41, 146, 191], and an important open question is to what extent these results are a consequence of the oversimplification of the network models employed. This is particularly the case for metabolic networks since they admit a variety of standard network representations [54, 55], with each one likely to accentuate a different aspect of the metabolic process.

One approach to address the effect of random mutations on network function is to use percolation theory. Here, we consider a *site percolation process*, in which nodes are referred to as *sites* which can be found in one of two states: *active* or *inactive*.

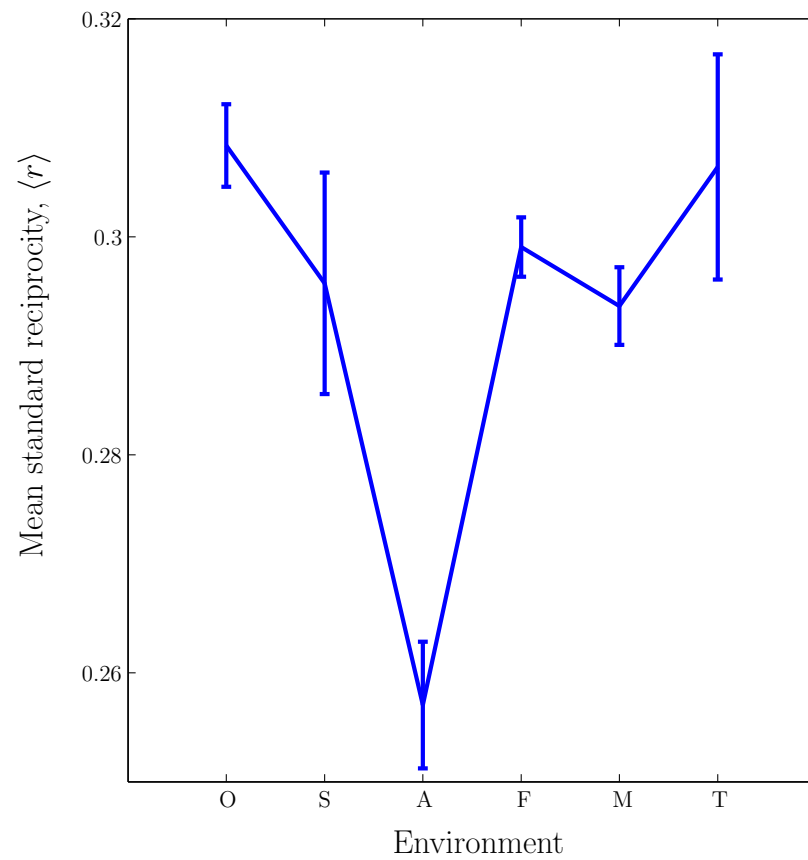


Figure 5.9: Relationship between average standard reciprocity,  $\langle r \rangle$ , and environmental variability. Note that we average over the six bacterial habitats along the x-axis, which are ordered according to the NCBI classification scheme of environmental variability: **O**bligat, **S**pecialised, **A**quatic, **F**aculative, **M**ultiple and **T**errestrial. Here vertical bars denote the standard error of the mean.

Such a process starts from an initial formation in which all states are inactive, sites are then turned on at random, and the networks edges added whenever two adjacent nodes become active, until the system achieves full activation. In this way we can observe the formation of so-called percolation clusters which form as the proportion,  $p$ , of active network nodes increases. In practice a *percolation threshold*,  $p = p_c$ , exists at which point a phase transition occurs and the system goes from a *non-percolating phase*, containing lots of small microscopic clusters, to a *percolating phase* in which a single dominant cluster, comparable to system size  $n$ , forms. Importantly, the critical point,  $p = p_c$ , separating the two different phases can be considered a proxy for network robustness [95, 192] – at this point the network is seen to disintegrate into isolated components and ceases to function correctly.

The key difference between our approach and the standard one (described above) lies in the criteria by which network nodes/edges are added. In the standard approach nodes are activated at random and edges placed between activated, adjacent node pairs. In the case of hypernetworks, we impose the more stringent requirement that all nodes within a hyperedge must be activated before any links are added. Or in terms of metabolism, all substrates and products of a reaction must be present before a reaction can occur. Below we provide algorithmic details for site percolation in an undirected hypernetwork; see Figure 5.10 for a graphical illustration of our approach.

1. Starting from an empty hypernetwork in which all nodes are inactive (i.e.  $p = 0$ ), set  $S$ , the relative size of the GCC, equal to 0.
2. Activate a randomly chosen node,  $i$  say.
3. Loop through all hyperedges containing node  $i$  adding those hyperedges for which all its nodes are active.
4. Compute the relative size of the giant connected component,  $S$ .
5. Repeat steps 2 to 4 until all sites are activated (i.e.  $p = 1$ ,  $S = 1$ ).

The above steps constitutes one realisation of our algorithm. To generate a statistically reliable estimate of  $S(p)$  we repeat the process a large number,  $M$  say, of

times. Importantly, employing a hypernetwork formalism alongside an appropriate percolation model allows us to quantify metabolic hypernetwork resilience to random mutations/errors, and to probe important evolutionary processes behind their hypernetwork structure.

Note that we use an adaptation of the Newman-Ziff algorithm [193] in all the computations we perform which is significantly faster than the usual breadth-first search, and that the GCC can be computed using standard network algorithms applied to the adjacency matrix of the hypernetwork. To determine the percolation threshold we employ the network *susceptibility* function as defined in [194], which is given by

$$\chi = \frac{\langle S^2 \rangle - \langle S \rangle^2}{\langle S \rangle}. \quad (5.9)$$

Note that the above is the variance-to-mean ratio, and is used to measure the dispersion of  $S$  in the  $M$  realisations, for the corresponding  $p$  value. The peak  $\chi_{\max}$  is an indication that the network is undergoing a phase transition, such that the giant connected component has emerged, and so importantly the corresponding value of  $p$  can be used as an estimate of the percolation threshold  $p_c$  [192, 194]. In our work we consider these two values to be equivalent in the sense that we refer to the point at which the susceptibility is maximised as the percolation threshold.

### Metabolic Hypernetwork Robustness

In the following we perform a site percolation analysis in order to investigate hypernetwork topology for the 115 metabolic hypernetworks for which we have habitat variability data (see Table 5.1), as well as considering in more detail two well-studied model organisms in *E. coli* and *Buchnera*, for which we provide some additional detail below.

- (i) *E. coli*: a facultative bacteria that is free-living within a variety of different hosts, and thus its environment is relatively varied, requiring the system to be highly adaptable. The hypernetwork of *E. coli* consists of 1097 vertices (metabolites) and 1117 hyperedges (reactions).
- (ii) *Buchnera*: a symbiotic bacteria that associates with one host, and thus lives within a very controlled environment. The hypernetwork of *Buchnera* consists

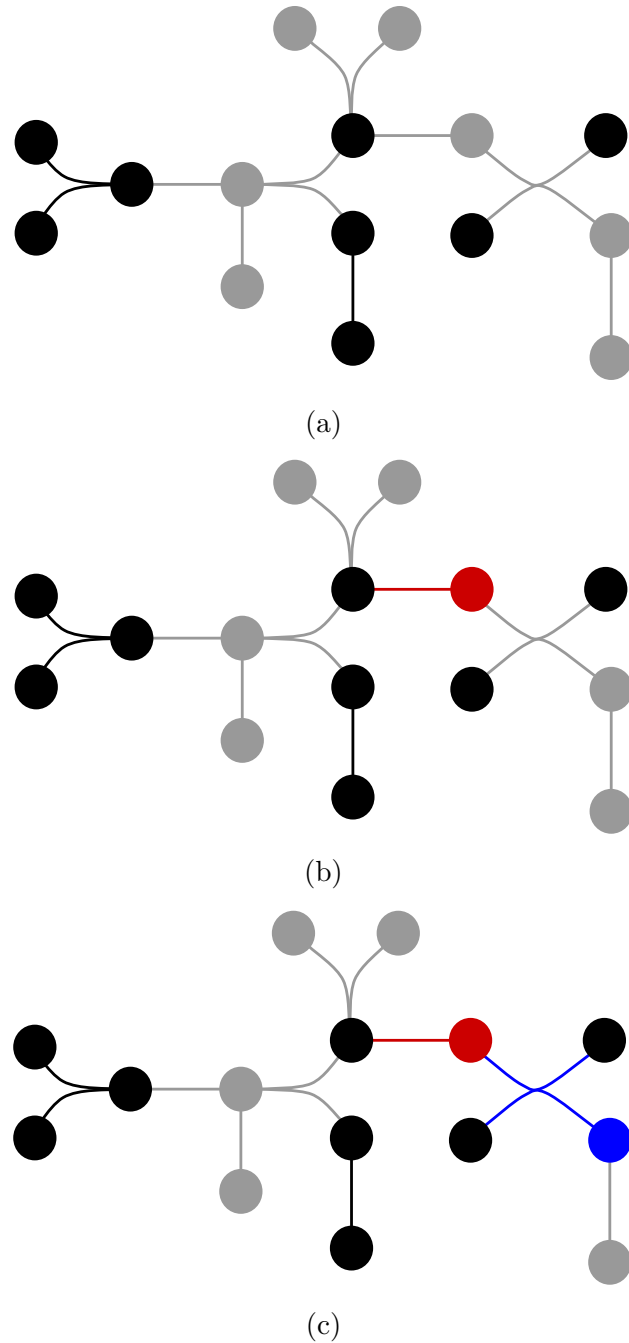


Figure 5.10: Figures (a)-(c) highlight two iterations of the site percolation process for a toy hypernetwork with  $n = 15$  and  $m = 9$  starting from a configuration with  $p = 8/15$ . Note that here active nodes are highlighted in black, whilst inactive nodes are in grey. The nodes and hyperedges that are highlighted red and blue correspond to those that become activated in the first and second iterations, respectively.



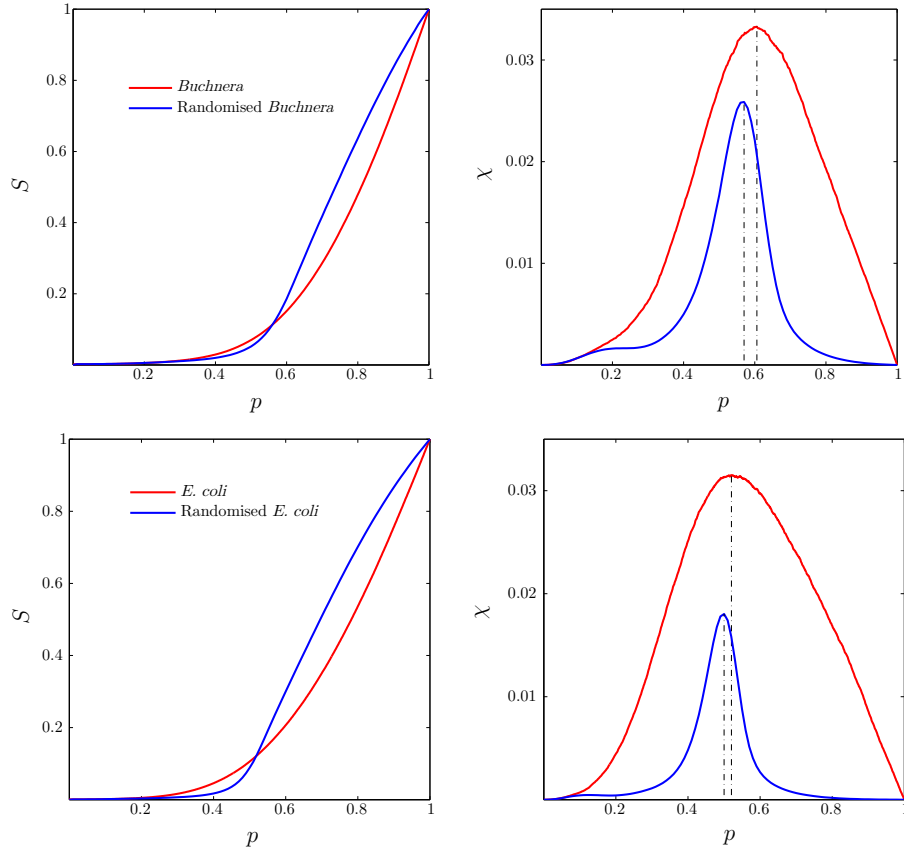


Figure 5.11: Site percolation for the two bacterial hypernetworks: *Buchnera* (top row) and *E. coli* (bottom row). Here, we compare the percolating properties,  $S$  and  $\chi$  of the original hypernetworks (red solid lines) against an ensemble of 100 rewired hypernetworks (blue solid lines). Note that the dashed lines indicate the percolation thresholds of the corresponding hypernetwork:  $p_c = 0.53$ , for *E. coli* ( $\langle p_c^{\text{rand}} \rangle = 0.50 \pm 0.0026$ ) and  $p_c = 0.61$  for *Buchnera* ( $\langle p_c^{\text{rand}} \rangle = 0.57 \pm 0.0048$ ).

of 444 nodes (metabolites) and 332 hyperedges (reactions).

Note that unlike the analysis performed in the previous section we restrict to undirected hypergraphs here, as our algorithm doesn't differentiate between the substrates and products of a reaction. Of course it is possible to extend the algorithm to consider directed hypernetworks but this lies outside the scope of the current work and rather provides an avenue for possible future research.

The results shown in Figures 5.11 and 5.13 were produced by running the site percolation process, as described in the previous section, 10,000 times for each of the 115 metabolic hypernetworks. To analyse the percolation properties we plot the size of the GCC averaged over all realisations,  $\langle S \rangle$ , as a function of  $p$  (i.e. the

proportion of active nodes). In addition we determine the percolation threshold for each organism by using the susceptibility function defined in Equation (5.9) to measure fluctuations in  $\langle S \rangle$ . Note that by repeating the percolation process 10,000 times we ensure that the percolation threshold for each network is statistically significant.

Figure 5.11 compares the results of our percolation analysis for the two model organisms of *E. coli* and *Buchnera*. Additionally, the plots show a comparative analysis against ‘equivalent’ random graphs for both organisms. Random graphs were produced using the rewiring algorithm forwarded by Zhou and Nakhleh [32]. In this algorithm the hyperedges of a metabolic hypernetwork are randomly rewired, whilst preserving both the size of the hypernetwork (i.e. number of nodes and hyperedges) and the hyperedge degree distribution. In our work, we construct random hypernetworks by applying 20,000 edge swaps, and results shown are for ensemble averages over some 100 realisations. Note that the randomisation procedure described above can produce networks that are not completely connected, and so in our experiments we restricted each realisation to the GCC; however, in practice this led to a maximum of 2-3% difference in network sizes overall, and so is unlikely to have any undue effects on our analysis.

The first point of note in Figure 5.11 is the difference between the percolation thresholds of the two organisms:  $p_c = 0.53$  for *E. coli* and  $p_c = 0.61$  for *Buchnera*. Note that the earlier appearance of the GCC in *E. coli* as opposed to *Buchnera*, is suggestive of a more robust network structure in the sense that *E. coli* would seem to be less susceptible to random attacks or errors. This may be considered a consequence of the greater evolutionary pressures/competition associated with the more variable habitat that *E. coli* has evolved in, as compared to *Buchnera*.

Another interesting outcome of Figure 5.11, is that both organisms would appear to be less robust than their random counterparts, with the difference being slightly greater for *Buchnera*. To check the validity of this result, we compared percolation thresholds for each of the 115 metabolic networks against average percolation thresholds,  $\langle p_c^{\text{rand}} \rangle$ , for matched random graphs. Importantly, we found the percolation threshold to be significantly larger in 114 of the 115 metabolic hypernetworks as compared to their random counterparts (see Figure 5.12a). Note that whilst

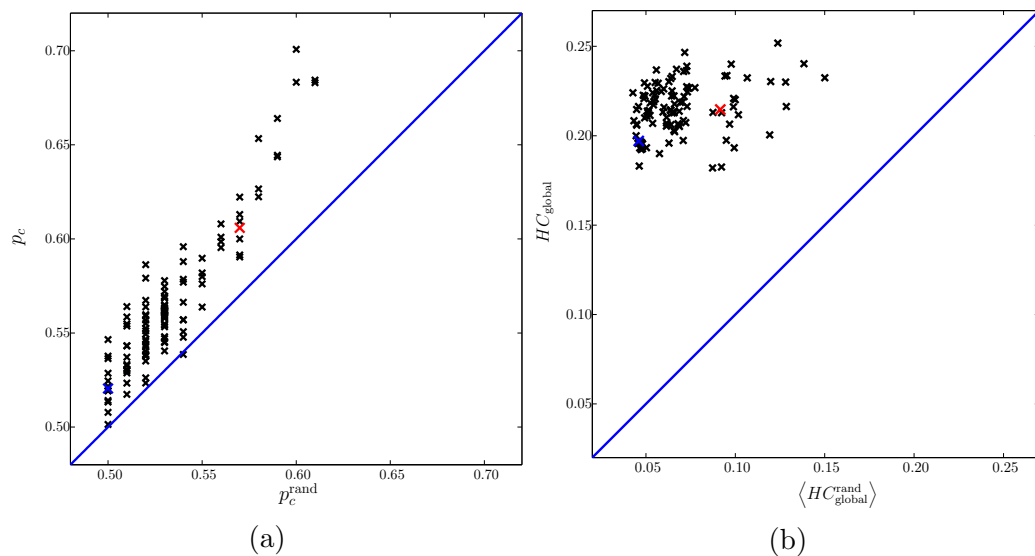


Figure 5.12: (a) Plot of the percolation threshold for each of the 115 metabolic networks versus the mean percolation threshold for matched random graphs. (b) Plot of the global clustering coefficient for each of the 115 metabolic networks versus the mean global clustering coefficient for matched random graphs.

this result might sound counter-intuitive, suggesting as it does, that these organisms are less resilient to node failure than matching random graphs, similar results have been reported for standard networks, where it has been shown that increased levels of clustering (as compared to random surrogates), coupled with high levels of heterogeneity, can lead to a so-called *core-periphery structure* [194]. In such a structure, the network is organised into a highly connected core, whilst the remainder of the nodes form the periphery [195]. A high amount of clustering within the network leads to a large highly entangled core that is very difficult to break down, and therefore decreases the percolation threshold (see, for example, [196, 197] and references therein). High clustering within the periphery of the network, however, leads to small sparsely interconnected cliques, which are very fragile to random mutations, and thus increases the percolation threshold [194]. Importantly, a number of recent studies have suggested that metabolic networks are organised into a highly modular core-periphery type structure, such that the core module connects the central metabolites and carries out basic metabolic functions, whilst periphery modules perform highly specific functions with minimum interactions with other modules [169, 198, 199].

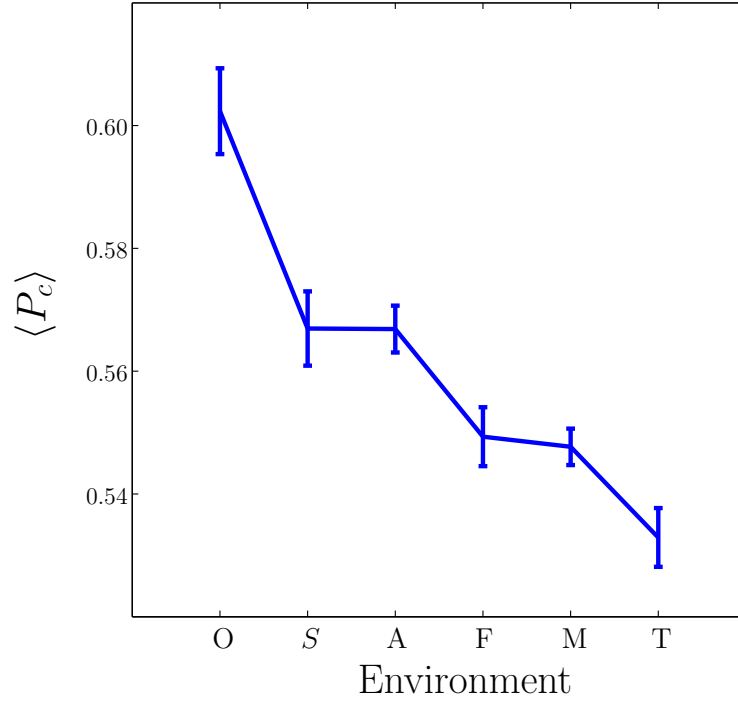


Figure 5.13: Relationship between the average percolation threshold,  $\langle p_c \rangle$ , and the environmental variability. Note that the six bacterial habitats along the x-axis are in order of environmental variability: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. The vertical bars represent the standard error of the mean.

A necessary (but not sufficient) condition for such a core-periphery structure, therefore, is the presence of an increased level of clustering. To measure the extent to which these networks are clustered, we compute the global clustering coefficient, as defined in Equation 5.6, for both the 115 metabolic hypernetworks ( $HC_{\text{global}}$ ), and their matched random graphs ( $\langle HC_{\text{global}}^{\text{rand}} \rangle$ ). The results are shown in Figure 5.12b which displays  $HC_{\text{global}}$  versus  $\langle HC_{\text{global}}^{\text{rand}} \rangle$  as well as the line  $HC_{\text{global}} = \langle HC_{\text{global}}^{\text{rand}} \rangle$ . Clearly, data points positioned above/below this line represent networks displaying higher/lower levels of clustering than their random surrogates. Importantly, we find that the data points for all 115 metabolic networks lie high above the line  $HC_{\text{global}} = \langle HC_{\text{global}}^{\text{rand}} \rangle$ , indicating increased levels of clustering relative to matched random graphs. Note that whilst this result, coupled with the heterogeneity evidenced in §5.1.2, is indicative of a core-periphery structure, additional investigations would be necessary (e.g. a k-core analysis) before any definitive conclusions can be made – again, providing additional avenues of possible future research.

Further evidence supporting the idea that network resilience, as measured using percolation thresholds, is correlated with the variability within an organisms environment is provided by Figure 5.13. Here, we have again split the bacterial species into the 6 different environmental classes (see §3): obligate, specialised, aquatic, facultative, multiple and terrestrial. We then computed the mean percolation threshold,  $\langle p_c \rangle$ , for each class and plotted it against increasing environmental variability. As can be readily seen, we find that the average percolation threshold decreases with increased variability, backing up our previous investigations of *E. coli* and *Buchnera*. More specifically, we find that the obligate class has a significantly larger percolation threshold than the other five classes, providing further evidence to suggest that host-associated bacteria are more vulnerable to random failures. This is perhaps a consequence of the symbiotic lifestyle of the obligate bacteria, where a metabolite-rich environment is provided by the host. Bacteria, which live in such an environment, are believed to have experienced a genome reduction throughout evolution, such that only the essential genes necessary for survival within that host were retained [200, 201]. Any random error within a bacteria that has undergone such a process is likely to be detrimental to the networks function. The next interesting observation is that the percolation thresholds for the specialised and aquatic classes are almost identical. This is not too surprising however, since these two classes are often considered to be equivalent regarding their environment variability [129]. Therefore, the bacteria from these two classes are likely to have a similar tolerance towards random errors, despite these bacteria being exposed to very different conditions. We then observe a relatively large decrease in the percolation threshold for the facultative and multiple classes, suggesting a higher resilience to random failures. Again, this comes as no surprise, since the bacteria within these two classes live in various different environments, and therefore are required to maintain their function in conditions where metabolite availability is uncertain. Finally, the smallest percolation threshold is observed for the terrestrial class, as expected, due to the highly heterogeneous conditions that bacteria living in soil are exposed to. The group differences shown in Figure 5.13 are significant by the Kruskal-Wallis (KW) test (p-value  $< 10^{-10}$ ).

## 5.3 Summary

In this Chapter, we have introduced several new network concepts based on a hypernetwork formalism of metabolism, and applied them to investigate whether this more complicated, yet more physiologically realistic setting can reveal further structural differences between metabolic networks. In particular, we have extended the reciprocity measure to hypernetworks, which allows us to characterise a hypernetwork in terms of its bidirectional relationships, which is a proxy for the extent to which reactions are reversible within these organisms. Moreover, we have adapted percolation strategies to undirected hypernetworks as a technique for investigating the robustness and vulnerability of metabolic networks. Importantly, using our new percolation-based approach, we found further evidence of the increased network complexities that greater environmental pressures and challenges induce. Importantly, many of the above findings fail to hold when a standard network approach is considered, and so such investigations provide great promise in forwarding our understanding of a range of complex biological processes.

# CHAPTER VI

## SUMMARY AND OUTLOOK

### 6.1 Summary

In this thesis we have applied complex network theory to analyse biological networks, with the aim of revealing important structure, function and evolutionary relationships. Typically, a network description provides a greatly simplified view of such a system, focusing on the interaction patterns between a very large number of similar ‘units’. Our investigations have focussed on metabolic networks, where the units are metabolites (i.e. small molecules or macromolecules), and we consider two metabolites to interact if they are involved in the same biochemical reaction. In this regard, the metabolites are modelled as nodes and the reactions are typically modelled by edges.

When analysing metabolic networks, the first thing to consider is a suitable reconstruction procedure. Recently, due to the availability of genome scale data and advances in computational techniques, a number of online databases have been built that contain metabolic data for a variety of different organisms. In Chapter 3 we describe the reconstruction process for the 383 metabolic networks studied in this work. It is important to note that despite the KEGG database being one of the most well established online resources, available data still remains incomplete. For that reason, the KEGG database is constantly being updated, such that missing (or spurious) reactions and metabolites are added (or removed). Therefore, the

reaction lists downloaded and used in this work are likely to be out-of-date and/or incomplete, and so it is possible for false positive (or false negative) errors to arise when analysing the metabolic data. The metabolic networks used in this work, however, are restricted to the giant connected component of the global network, which has been shown to contain the most accurate data and therefore any updates to the data are unlikely to have a huge effect on our results.

Importantly, the 383 bacterial networks that we consider can be classified according to different environmental pressures, such as growth conditions and habitat lifestyle. Using this data, the aim of the thesis was to determine important links between metabolic network structure and the environmental conditions to which the 383 organisms have evolved. The approach taken has been twofold:

- improve the current state-of-the-art in network classification; and
- develop new, more physiologically realistic network models and accompanying measures.

In the following we describe the contribution of this thesis, as well as suggestions for future directions of each approach.

### 6.1.1 Classification of Biological Networks

Network classification has recently become an increasingly important problem in network science, where the aim is to cluster networks according to their structural similarities/differences. Current approaches have mainly involved comparing network structure based upon global properties, such as the average path length, clustering coefficients and degree distributions, and have led to many diverse networks being classified as small-world and/or scale-free. Recently, however, such approaches have received criticism, since it is possible to find two networks with identical global properties, yet a considerably different structure [30]. Moreover, computing global properties for the metabolic data, which, as mentioned previously is incomplete and noisy, may result in misleading findings.

Motivated by this, in Chapter 4 we introduced a novel graph embedding classification approach, which involves computing so-called feature vectors that are based



on low-order network motifs (i.e. small interaction patterns of nodes). A non-negative matrix factorisation was then used to reduce the size of the problem, and to improve our ability to find important patterns and features within the data. One of the main advantages of this approach as compared to other methods, is that it can be employed to compare the local structure of networks of varying order with a relatively small amount of computational effort. The approach can therefore be applied to any cohort of networks, providing they can be represented as simple graphs. A slight caveat of the approach results from the non-uniqueness of the non-negative matrix factorisation, due to (a) the ambiguity surrounding the choice of the parameter  $k$ , and (b) the iterative nature of the non-negative matrix algorithm, which means that the solution depends on both our choice of  $k$  and the initial guess. To overcome these drawbacks the experiments were repeated 1000 times and for values of  $k$  ranging from 1 to 20, and importantly, we found our results remained consistent.

As a proof of principle, we applied the new approach to a large ensemble of networks consisting of both bacterial networks and a variety of different network models (Erdős-Rényi networks, small-world networks and scale-free networks), and were able to provide evidence that the technique can correctly differentiate between the various networked structures.

Furthermore, we applied our new approach to an ensemble of 383 bacterial networks as described in the previous section. By introducing a new global significance score, which can be used as a proxy for network complexity, we were able to demonstrate a number of significant correlations between environmental factors, such as growth conditions and habitat variability, and network motif structure, providing evidence that organism adaptability leads to increased complexities in the resultant metabolic networks. Note, that our new motif based approach is robust in the sense that results obtained are reproducible in both the latest data set and an older (data not shown), less complete version of the data.

A direct extension of this work would be to carry out the analysis using larger order network motifs. In this work we only considered network motifs of order 3 and 4, so it would be interesting to investigate whether higher order network motifs reveal further differences between the different species. However, since the number of motifs of size  $q$  grows exponentially, and therefore reaches enormous quantities, the

computational cost will be much greater than the algorithm forwarded in this thesis. In terms of metabolism, an extension of this approach to directed hypernetworks is desirable. However, as it stands, a well defined definition for network motifs within a hypernetwork does not exist, and any consideration of such an extension is likely to exhibit a much greater amount of complexity than the standard approach.

Other possible network classification techniques using feature vectors could also be explored. One possible idea is to construct feature vectors based upon algebraic graph theory [154]. It is well known for instance, that the spectrum of the adjacency matrix (or the closely related Laplacian matrix) can be used to extract important structural properties of a network. The coefficients of the characteristic polynomial of the adjacency matrix, for instance, are associated with cycles (i.e. subgraphs) of different sizes occurring across a network [57]. Computing the coefficients of the characteristic equation for large networks, however, still remains a challenging task.

### 6.1.2 Hypernetwork Models

From a modelling perspective an important open question concerns the determination of a suitable network model for metabolism. Depending upon the questions being asked simple graph models may not be appropriate, as information is lost when moving from the full chemical reaction system to a simple graph. Recently, several authors have argued that hypernetworks provide the most physiologically realistic representation of metabolism, since they allow for hyperedges to connect multiple nodes, thus preserving the information regarding the dependency between metabolites.

In Chapter 5 we investigated the extent to which hypernetwork models capture more biologically relevant information compared to the standard models. In particular, we proposed a generalisation of the reciprocity measure (i.e. the proportion of bidirectional relationships in the network) to hypernetworks, and evidenced relations between bacterial metabolic hypernetwork structure and environmental variability. Moreover, we have adapted a widely used percolation method (i.e. the process by which nodes are randomly activated within a network) to a hypergraph formalism, in order to quantify the robustness and vulnerability of metabolic hypernetworks to

random failures on the nodes. The key difference in our new approach compared to the standard approach, lies in the way by which connections between nodes are activated in the network. In our approach, motivated by biology, a hyperedge only becomes activated in the network if all nodes (metabolites) involved in the hyperedge (reaction), are currently active (available). Using our new percolation-based approach we were able to find further evidence to suggest that environmental pressures play a pivotal role in metabolic network complexity.

Importantly, many of the relationships found between metabolic network structure and environmental variability, were not apparent when applied to a simple graph representation, thus suggesting that a hypernetwork formalism adds significant value beyond the standard network approach. To date only a handful of complex network measures have been extended to hypernetworks and so the extension of other network measures could potentially lead to further insights into important structure, function and evolutionary relationships within biological networks. For example, based on the results from our new reciprocity measure, it would be interesting to extend the idea of network returnability [107] (i.e. the proportion of loops within a network) to directed hypernetworks, in order to determine whether the aforementioned relationship holds for closed walks of length greater than two, as is the case for the reciprocity measure. Moreover, an extension of our new hypernetwork percolation model incorporating the effects of cascading failures on network structure promises to provide further insights into the mechanisms underlying metabolism.

## 6.2 Outlook

For the last century biological research has been dominated by the reductionist approach, such that individual cellular components have been analysed in great detail. More recently, however, it has been recognised that cellular functions are unlikely to be attributed to one component, and instead arise due to the intricate web of interactions between the cellular components, such as proteins, and other macro-molecules, which result in various biochemical nets. The recent development of network science has allowed for biological networks to be analysed at the systems level, which has allowed for significant advances in the understanding of cellular organisation.

Many studies, including the work carried out in this thesis, focus their analysis on a subset of the cellular network, such as gene regulatory networks, protein-protein interaction networks, and protein residue networks. We have focused our analysis on metabolic networks, and found strong evidence to suggest that structure and robustness are interlinked with the functioning of a system and its ability to adapt to the environment.

These networks, however, do not function in isolation, and are in fact highly interconnected. The end products of metabolic reactions, for example, provide the protein molecules that mediate interactions between genes. We then also have that metabolic reactions are catalysed by enzymes that are end products of gene interaction networks. One of the main directions of network science is therefore to integrate all biological interactions within the cell, with the aim of providing further insight into its behaviour. Multiplex, or more generally multilayer, networks provide a promising network representation for analysing the collection of cellular networks as a whole, since they allow nodes to be connected via different types of connections, each of which constitutes to a layer in the multiplex network [202, 203, 204]. In the current context the different layers would consist of the different types of interaction that exist in the cell, and connections between layers would account for interactions between these different structures (e.g. a layer describing metabolic reactions would connect to a layer describing protein-protein interactions via the synthesis of proteins and amino acids).

# BIBLIOGRAPHY

- [1] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1):69–77, 2000. [1](#)
- [2] Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. [1](#), [4.1](#)
- [3] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007. [1](#)
- [4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. [1](#), [1.1](#), [2.3](#)
- [5] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998. [1](#), [1.1](#), [2.2](#), [2.3](#)
- [6] Norman L Bigg, Edward Keith Lloyd, and Robin James Wilson. *Graph Theory: 1736-1936*. Oxford University Press, 1976. [1.1](#)
- [7] Béla Bollobás. *Graph Theory*. Elsevier, 1982. [1.1](#)
- [8] Jonathan L Gross and Jay Yellen. *Handbook of Graph Theory*. CRC Press, 2004. [1.1](#)
- [9] Paul Erdős and Alfréd Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959. [1.1](#), [2.3](#)

- [10] Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967. [1.1](#), [2.3](#)
- [11] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. [1.1](#)
- [12] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. [1.1](#), [2.5](#)
- [13] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, pages 452–473, 1977. [1.1](#)
- [14] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007. [1.1](#), [2.2](#)
- [15] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, 2006. [1.1](#), [2.3](#)
- [16] Petter Holme, Beom Jun Kim, Chang No Yoon, and Seung Kee Han. Attack vulnerability of complex networks. *Physical Review E*, 65(5):056109, 2002. [1.1](#), [2.4.2](#), [2.4.2](#), [2.5](#)
- [17] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005. [1.1](#), [2.1.2](#), [2.5](#)
- [18] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000. [1.1](#), [2.4.2](#), [2.4.2](#)
- [19] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001. [1.1](#)
- [20] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Science of the United States of America*, 99(20):12917–12922, 2002. [1.1](#), [1.2](#)

- [21] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002. [1.1](#), [3.2.3](#)
- [22] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999. [1.1](#)
- [23] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003. [1.1](#)
- [24] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. [1.1](#), [2.5](#)
- [25] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. [1.1](#)
- [26] Stephan Wuchty, Zoltán N Oltvai, and Albert-László Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179, 2003. [1.1](#)
- [27] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):11980–11985, 2003. [1.1](#), [2.4.1](#)
- [28] Tae-Hwan Kim, Junil Kim, Pat Heslop-Harrison, and Kwang-Hyun Cho. Evolutionary design principles and functional characteristics based on kingdom-specific network motifs. *Bioinformatics*, 27(2):245–251, 2011. [1.1](#)
- [29] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and M-F Sagot. An introduction to metabolic networks and their structural analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 5(4):594–617, 2008. [1.1](#), [3.2.3](#)
- [30] Lun Li, David Alderson, John C Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005. [1.1](#), [4.1](#), [6.1.1](#)

- [31] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024, 2007. [1.1](#)
- [32] Wanding Zhou and Luay Nakhleh. Properties of metabolic graphs: biological organization or representation artifacts? *BMC Bioinformatics*, 12(1):132, 2011. [1.1](#), [3.2.3](#), [5.1](#), [5.1.2](#), [5.1.2](#), [5.2.2](#)
- [33] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009. [1.1](#), [2.1.1](#), [3.2.3](#)
- [34] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012. [1.1](#)
- [35] Mikko Kivelä, Alex Arenas, Marc Barthélemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 3(3):203–271, 2014. [1.1](#)
- [36] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280, 2004. [1.2](#)
- [37] Gary D Bader, Doron Betel, and Christopher WV Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003. [1.2](#)
- [38] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002. [1.2](#)
- [39] Veá Matys, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E. Kel, Olga V. Kel-Margoulis, et al. Transfac®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003. [1.2](#)



- 
- [40] M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, 2004. [1.2](#)
- [41] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000. [1.2](#), [2.3](#), [2.4.2](#), [3.2.3](#), [4.1](#), [5.2.2](#)
- [42] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004. [1.2](#)
- [43] Edward Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. [1.2](#)
- [44] Olaf Sporns, Dante R Chialvo, Marcus Kaiser, and Claus C Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, 2004. [1.2](#)
- [45] Joseph Loscalzo, Isaac Kohane, and Albert-Laszlo Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular Systems Biology*, 3(1), 2007. [1.2](#)
- [46] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, 2007. [1.2](#), [2.1.1](#)
- [47] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011. [1.2](#)
- [48] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, 5(4):558–567, 2002. [1.2](#)

- [49] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Network structure and robustness of marine food webs. *Marine Ecology Progress Series*, 273:291–302, 2004. [1.2](#)
- [50] Jennifer A Dunne. The network structure of food webs. *Ecological Networks: linking structure to dynamics in food webs*, pages 27–86, 2006. [1.2](#)
- [51] Andrew V Zeigarnik and Danail Bončev. *Chemical Reaction Networks: A Graph-Theoretical Approach*. CRC Press, 1996. [2.1.1](#)
- [52] Jacques van Helden, Lorenz Wernisch, David Gilbert, and SJ Wodak. Graph-based analysis of metabolic networks. In *Bioinformatics and Genome Analysis*, volume 38, pages 245–274. Springer, 2002. [2.1.1](#)
- [53] Raul Montañez, Miguel Angel Medina, Ricard V Sole, and Carlos Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays*, 32(3):246–256, 2010. [2.1.1](#), [3.2.3](#)
- [54] Petter Holme. Model validation of simple-graph representations of metabolism. *Journal of The Royal Society Interface*, 6(40):1027–1034, 2009. [2.1.1](#), [5.2.2](#)
- [55] Petter Holme and Mikael Huss. Substance graphs are optimal simple-graph representations of metabolism. *Chinese Science Bulletin*, 55(27-28):3161–3168, 2010. [2.1.1](#), [5.2.2](#)
- [56] Oystein Ore. *Graphs and Their Uses*, volume 34. Cambridge University Press, 1990. [2.1.2](#)
- [57] Norman Biggs. *Algebraic Graph Theory*. Cambridge University Press, 1993. [2.1.2](#), [6.1.1](#)
- [58] Ernesto Estrada, Desmond J Higham, and Naomichi Hatano. Communicability betweenness in complex networks. *Physica A: Statistical Mechanics and its Applications*, 388(5):764–774, 2009. [2.1.2](#)
- [59] Ernesto Estrada and Desmond J Higham. Network properties revealed through matrix functions. *SIAM Review*, 52(4):696–714, 2010. [2.1.2](#)

- 
- [60] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, 1994. [2.2](#)
- [61] Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004. [2.2](#)
- [62] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007. [2.2](#)
- [63] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002. [2.3](#), [2.3](#)
- [64] Mark EJ Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. [2.3](#)
- [65] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001. [2.3](#)
- [66] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20, 2003. [2.3](#), [2.3](#)
- [67] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960. [2.3](#)
- [68] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959. [2.3](#)
- [69] Béla Bollobás. *Random Graphs*, volume 73. Cambridge University Press, 2001. [2.3](#)
- [70] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978. [2.3](#)

- 
- [71] David A Fell and Andreas Wagner. The small world of metabolism. *Nature Biotechnology*, 18(11):1121–1122, 2000. [2.3](#), [3.2.3](#)
- [72] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999. [2.3](#)
- [73] Jose M Montoya and Ricard V Solé. Small world patterns in food webs. *Journal of Theoretical Biology*, 214(3):405–412, 2002. [2.3](#)
- [74] Björn H Junker and Falk Schreiber. *Analysis of Biological Networks*, volume 2. John Wiley & Sons, 2008. [2.3](#), [4.2.1](#)
- [75] John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode *caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, 1986.
- [76] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.
- [77] Mark EJ Newman and Duncan J Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341–346, 1999. [2.3](#)
- [78] Mark EJ Newman. Models of the small world. *Journal of Statistical Physics*, 101(3-4):819–841, 2000. [2.3](#)
- [79] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115, 2000. [2.3](#)
- [80] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3):035103, 2002. [2.3](#)
- [81] Albert-László Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412–413, 2009. [2.3](#)
- [82] Shmoolik Mangan, Alon Zaslaver, and Uri Alon. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, 334(2):197–204, 2003. [2.4.1](#)

- [83] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. [2.4.1](#), [2.4.1](#), [4.2.1](#)
- [84] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, 2002. [2.4.1](#)
- [85] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004. [2.4.1](#), [4.2.1](#), [4.2.1](#)
- [86] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005. [2.4.1](#), [3.2.3](#), [4.2.1](#)
- [87] Robert J Prill, Pablo A Iglesias, and Andre Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology*, 3(11):e343, 2005. [2.4.1](#)
- [88] Tom Britton, Maria Deijfen, and Anders Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006. [2.4.1](#)
- [89] Ron Milo, Nadav Kashtan, Shalev Itzkovitz, Mark EJ Newman, and Uri Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv preprint cond-mat/0312028*, 2003. [2.4.1](#)
- [90] Amnon Aharony and Dietrich Stauffer. *Introduction to Percolation Theory*. Taylor & Francis, 2003. [2.4.2](#), [5.2.2](#)
- [91] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1):309–320, 2000. [2.4.2](#)

- [92] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626, 2000. [2.4.2](#)
- [93] Baharan Mirzasoleiman and Mahdi Jalili. Failure tolerance of motif structure in biological networks. *PloS One*, 6(5):e20512, 2011. [2.4.2](#), [2.5](#), [4.2.1](#)
- [94] Mark EJ Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002. [2.4.2](#)
- [95] Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85(25):5468, 2000. [2.4.2](#), [5.2.2](#)
- [96] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011. [2.4.2](#)
- [97] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Science of the United States of America*, 98(2):404–409, 2001. [2.4.2](#)
- [98] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. [2.5](#)
- [99] Julien Gagneur, David B Jackson, and Georg Casari. Hierarchical analysis of dependency in metabolic networks. *Bioinformatics*, 19(8):1027–1034, 2003. [2.5](#), [3.2.3](#)
- [100] Alex Arenas, Leon Danon, Albert Diaz-Guilera, Pablo M Gleiser, and Roger Guimera. Community analysis in social networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):373–380, 2004. [2.5](#)
- [101] Dennis M Wilkinson and Bernardo A Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5241–5248, 2004. [2.5](#)
- [102] Petter Holme, Mikael Huss, and Hawoong Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532–538, 2003. [2.5](#)

- 
- [103] Petter Holme and Mikael Huss. Discovery and analysis of biochemical subnetwork hierarchies. *arXiv preprint q-bio/0309011*, 2003. [2.5](#)
- [104] Linton C Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979. [2.5](#)
- [105] Stephen P Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005. [2.5](#)
- [106] Fan RK Chung. *Spectral Graph Theory*, volume 92. American Mathematical Soc., 1997. [2.5](#)
- [107] Ernesto Estrada and Naomichi Hatano. Returnability in complex directed networks (digraphs). *Linear Algebra and Its Applications*, 430(8):1886–1896, 2009. [2.5](#), [6.1.2](#)
- [108] Anirban Banerjee and Jürgen Jost. Spectral plots and the representation and interpretation of biological data. *Theory in Biosciences*, 126(1):15–21, 2007. [2.5](#)
- [109] Anirban Banerjee and Jürgen Jost. Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, 157(10):2425–2431, 2009. [2.5](#)
- [110] Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann. The Laplacian spectrum of graphs. *Graph Theory, Combinatorics, and Applications*, 2:871–898, 1991. [2.5](#)
- [111] Reinhard Heinrich, SM Rapoport, and TA Rapoport. Metabolic regulation and mathematical models. *Progress in Biophysics and Molecular Biology*, 32:1–82, 1978. [3.2](#), [3.2.1](#)
- [112] Andreas Karoly Gombert and Jens Nielsen. Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11(2):180–186, 2000. [3.2](#)
- [113] Ralf Steuer. Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry*, 68(16):2139–2151, 2007. [3.2](#)

- 
- [114] Jörg Stelling. Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, 7(5):513–518, 2004. [3.2](#)
- [115] Ralf Steuer and Bjorn H Junker. Computational models of metabolism: stability and regulation in metabolic networks. *Advances in Chemical Physics*, 142:105, 2009. [3.2.1](#)
- [116] David Fell and Athel Cornish-Bowden. *Understanding the Control of Metabolism*, volume 2. Portland Press London, 1997. [3.2.1](#)
- [117] Reinhart Heinrich and Stefan Schuster. The modelling of metabolic systems. structure, control and optimality. *Biosystems*, 47(1):61–77, 1998. [3.2.1](#)
- [118] Johann M Rohwer. Kinetic modelling of plant metabolic pathways. *Journal of Experimental Botany*, 63(6, SI):2275–2292, 2012. [3.2.1](#)
- [119] Kirill Peskov, Ekaterina Mogilevskaya, and Oleg Demin. Kinetic modelling of central carbon metabolism in Escherichia coli. *FEBS Journal*, 279(18):3374–3385, 2012. [3.2.1](#)
- [120] Patrik R Jones. Improving fermentative biomass-derived h<sub>2</sub>-production by engineering microbial metabolism. *International Journal of Hydrogen Energy*, 33(19):5122–5130, 2008. [3.2.2](#)
- [121] Sung Kuk Lee, Howard Chou, Timothy S Ham, Taek Soon Lee, and Jay D Keasling. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*, 19(6):556–563, 2008. [3.2.2](#)
- [122] Akihiko Kondo, Jun Ishii, Kiyotaka Y Hara, Tomohisa Hasunuma, and Fumio Matsuda. Development of microbial cell factories for bio-refinery through synthetic bioengineering. *Journal of Biotechnology*, 163(2):204–216, 2013. [3.2.2](#)
- [123] Nathan E Lewis and Alyaa M Abdel-Haleem. The evolution of genome-scale models of cancer metabolism. *Frontiers in Physiology*, 4(4), 2013. [3.2.2](#)



- [124] Karthik Raman, Preethi Rajagopalan, and Nagasuma Chandra. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS Computational Biology*, 1(5):e46, 2005. [3.2.2](#)
- [125] Karthik Raman, Kalidas Yeturu, and Nagasuma Chandra. targettb: a target identification pipeline for mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology*, 2(1):109, 2008. [3.2.2](#)
- [126] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005. [3.2.3](#)
- [127] Petter Holme. Metabolic robustness and network modularity: a model study. *PloS One*, 6(2):e16605, 2011. [3.2.3](#)
- [128] Arend Hintze and Christoph Adami. Evolution of complex modular biological networks. *PLoS Computational Biology*, 4(2):e23, 2008. [3.2.3](#)
- [129] Merav Parter, Nadav Kashtan, and Uri Alon. Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, 7(1):169, 2007. [3.2.3](#), [3.3.1](#), [3.4.1](#), [3.5](#), [4.4.1](#), [5.2.1](#), [5.2.2](#)
- [130] Kazuhiro Takemoto. Does habitat variability really promote metabolic network modularity? *PloS One*, 8(4):e61348, 2013. [3.2.3](#), [3.3.1](#), [3.4.2](#), [4.5](#)
- [131] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. [3.2.3](#)
- [132] Didier Croes, Fabian Couche, Shoshana J Wodak, and Jacques van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1):222–236, 2006. [3.2.3](#)
- [133] Romain Bourqui, Ludovic Cottret, Vincent Lacroix, David Auber, Patrick Mary, Marie-France Sagot, and Fabien Jourdan. Metabolic network visualization eliminating node redundancy and preserving metabolic pathways. *BMC Systems Biology*, 1(1):29, 2007. [3.2.3](#)

- [134] Minoru Kanehisa. The KEGG database. In *Novartis Found Symp*, volume 247, pages 91–101, 2002. [3.3](#)
- [135] Peter D Karp, Monica Riley, Milton Saier, Ian T Paulsen, Julio Collado-Vides, Suzanne M Paley, Alida Pellegrini-Toole, César Bonavides, and Socorro Gama-Castro. The ecocyc database. *Nucleic Acids Research*, 30(1):56–58, 2002. [3.3](#)
- [136] Peter D Karp, Christos A Ouzounis, Caroline Moore-Kochlacs, Leon Goldovsky, Pallavi Kaipa, Dag Ahrén, Sophia Tsoka, Nikos Darzentas, Victor Kunin, and Núria López-Bigas. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 33(19):6083–6089, 2005. [3.3](#)
- [137] Cynthia J Krieger, Peifen Zhang, Lukas A Mueller, Alfred Wang, Suzanne Paley, Martha Arnaud, John Pick, Seung Y Rhee, and Peter D Karp. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, 32(suppl 1):D438–D442, 2004. [3.3](#)
- [138] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids research*, 40(D1):D109–D114, 2012. [3.3.1](#)
- [139] Areejit Samal and Olivier C Martin. Randomizing genome-scale metabolic networks. *PloS One*, 6(7):e22295, 2011. [3.3.1](#)
- [140] Areejit Samal, Andreas Wagner, and Olivier C Martin. Environmental versatility promotes modularity in genome-scale metabolic networks. *BMC Systems Biology*, 5(1):135, 2011. [3.3.1](#)
- [141] Hong-Wu Ma and An-Ping Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11):1423–1430, 2003. [3.3.1](#)
- [142] Jonathan J Crofts and Ernesto Estrada. A statistical mechanics description of environmental variability in metabolic networks. *Journal of Mathematical Chemistry*, 52(2):675–688, 2013. [3.3.1](#), [3.5](#), [4.4.1](#), [5.2.1](#)

- [143] Jason Raymond and Daniel Segrè. The effect of oxygen on biochemical networks and the evolution of complex life. *Science*, 311(5768):1764–1767, 2006. [3.4.2](#), [4.4.2](#), [4.4.3](#)
- [144] Kazuhiro Takemoto and Suritalatu Borjigin. Metabolic network modularity in archaea depends on growth conditions. *PloS One*, 6(10):e25874, 2011. [3.4.2](#), [3.5](#)
- [145] AR Pah, R Guimera, AM Mustoe, and LAN Amaral. Use of a global metabolic network to curate organismal metabolic networks. *Scientific Reports*, 3, 2013. [3.5](#)
- [146] Andreas Wagner and David A Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, 2001. [4.1](#), [5.2.2](#)
- [147] Reka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, 2005. [4.1](#)
- [148] Luis A Nunes Amaral, Antonio Scala, Marc Barthélemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000. [4.1](#)
- [149] Kwang-Il Goh, Eulsik Oh, Hawoong Jeong, Byungnam Kahng, and Doochul Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, 99(20):12583–12588, 2002. [4.1](#)
- [150] Stephen A Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151–158. ACM, 1971. [4.1](#)
- [151] Alberto Sanfeliu and King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 13(3):353–362, 1983. [4.1](#)
- [152] Risi Imre Kondor and John D Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Confer-*

- ence on Machine Learning*, pages 315–322. Morgan Kaufmann Publishers Inc., 2002. [4.1](#)
- [153] Jonas Richiardi, Sophie Achard, Edward Bullmore, and Dimitri Van De Ville. Classifying connectivity graphs using graph and vertex attributes. In *Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on*, pages 45–48. IEEE, 2011. [4.1](#)
- [154] Peng Ren, Richard C Wilson, and Edwin R Hancock. Graph characterization via Ihara coefficients. *Neural Networks, IEEE Transactions on*, 22(2):233–245, 2011. [4.1](#), [6.1.1](#)
- [155] Roger Guimera, Marta Sales-Pardo, and Luís AN Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, 3(1):63–69, 2006. [4.1](#)
- [156] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009. [4.1](#)
- [157] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007. [4.1](#)
- [158] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257, 2008. [4.1](#)
- [159] Tijana Milenković, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121, 2010. [4.1](#)
- [160] Wayne Hayes, Kai Sun, and Nataša Pržulj. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491, 2013. [4.1](#)
- [161] Kai Sun, Joana P Gonçalves, Chris Larminie, et al. Predicting disease associations via biological network analysis. *BMC Bioinformatics*, 15(1):304, 2014. [4.1](#)

- [162] Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. Revealing the hidden language of complex networks. *Scientific Reports*, 4(4547), 2014. [4.1](#)
- [163] Conner I Sandefur, Maya Mincheva, and Santiago Schnell. Network representations and methods for the analysis of chemical and biochemical pathways. *Molecular bioSystems*, 9(9):2189–2200, 2013. [4.2.1](#)
- [164] David Skillicorn. *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. CRC Press, 2007. [4.2.1](#), [4.2.2](#)
- [165] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. [4.2.1](#), [A](#)
- [166] Clare Lee, Desmond Higham, Daniel Crowther, and J Keith Vass. Non-negative matrix factorisation for network reordering. *Monografias de la Real Academia de Ciencias de Zaragoza*, 33:39–53, 2010. [4.2.1](#)
- [167] Jan Graffelman. *calibrate: A Guide to Scatterplot and Biplot Calibration*, 2012. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source). [4.3](#)
- [168] Wendy L Martinez, Angel Martinez, and Jeffrey Solka. *Exploratory data analysis with MATLAB*. CRC Press, 2004. [4.3](#)
- [169] Anat Kreimer, Elhanan Borenstein, Uri Gophna, and Eytan Ruppin. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, 105(19):6976–6981, 2008. [4.4.1](#), [5.2.1](#), [5.2.2](#)
- [170] Sarath Chandra Janga and M Madan Babu. Network-based approaches for linking metabolism with environment. *Genome Biology*, 9(11):239, 2008. [4.4.1](#), [5.2.1](#)
- [171] Wanding Zhou and Luay Nakhleh. Convergent evolution of modularity in metabolic networks through different community structures. *BMC Evolutionary Biology*, 12(1):181, 2012. [4.4.1](#), [4.5](#)

- [172] Vicente Pérez-Brocal, Rosario Gil, Silvia Ramos, Araceli Lamelas, Marina Postigo, José Manuel Michelena, Francisco J Silva, Andrés Moya, and Amparo Latorre. A small microbial genome: the end of a long symbiotic relationship? *Science*, 314(5797):312–313, 2006. [4.4.1](#)
- [173] Advait A Apte, John W Cain, Danail G Bonchev, and Stephen S Fong. Cellular automata simulation of topological effects on the dynamics of feed-forward motifs. *Journal of Biological Engineering*, 2(1):2, 2008. [4.4.3](#)
- [174] Zhi Wang and Jianzhi Zhang. Abundant indispensable redundancies in cellular metabolic networks. *Genome Biology and Evolution*, 1:23–33, 2009. [4.4.3](#)
- [175] Richard Harrison, Balázs Papp, Csaba Pál, Stephen G Oliver, and Daniela Delneri. Plasticity of genetic interactions in metabolic networks of yeast. *Proceedings of the National Academy of Sciences*, 104(7):2307–2312, 2007. [4.4.3](#)
- [176] H W Ma and A P Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003. [4.5](#)
- [177] Mark EJ Newman. *Networks: An Introduction*. Oxford University Press, 2010. [5.1](#), [5.2.2](#)
- [178] Ernesto Estrada. *The structure of Complex Networks: Theory and Applications*. Oxford University Press, 2011. [5.1](#)
- [179] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006. [5.1](#)
- [180] Mark Buchanan. *Networks in Cell Biology*. Cambridge University Press, 2010. [5.1](#)
- [181] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004. [5.1](#)

- [182] Abdelghani Bellaachia and Mohammed Al-Dhelaan. Random walks in hypergraph. In *Proceedings of the 2013 International Conference on Applied Mathematics and Computational Methods*, pages 187–194, 2013. [5.1](#)
- [183] Aurélien Ducournau and Alain Bretto. Random walks in directed hypergraphs and application to semi-supervised image segmentation. *Computer Vision and Image Understanding*, 120:91–102, 2014. [5.1](#), [5.1.1](#)
- [184] Alexei Vazquez. Finding hypergraph communities: a Bayesian approach and variational solution. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):P07006, 2009. [5.1](#)
- [185] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012. [5.1](#)
- [186] Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2):177–201, 1993. [5.1.1](#)
- [187] Ernesto Estrada and Juan A Rodriguez-Velazquez. Complex networks as hypergraphs. *arXiv preprint physics/0505137*, 2005. [5.1.2](#)
- [188] Diego Garlaschelli and Maria I Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701, 2004. [5.2.1](#)
- [189] JA Rodríguez. On the Laplacian spectrum and walk-regular hypergraphs. *Linear and Multilinear Algebra*, 51(3):285–297, 2003. [5.2.1](#)
- [190] Nicole Pearcy, Jonathan J Crofts, and Nadia Chuzhanova. Network motif frequency vectors reveal evolving metabolic network organisation. *Molecular BioSystems*, 11(1):77–85, 2015. [5.2.1](#)
- [191] Abdelhalim Larhlimi, Sylvain Blachon, Joachim Selbig, and Zoran Nikoloski. Robustness of metabolic networks: a review of existing definitions. *Biosystems*, 106(1):1–8, 2011. [5.2.2](#)

- 
- [192] Filippo Radicchi. Predicting percolation thresholds in networks. *Physical Review E*, 91(1):010801, 2015. [5.2.2](#), [5.2.2](#)
- [193] MEJ Newman and RM Ziff. Efficient Monte Carlo algorithm and high-precision results for percolation. *Physical Review Letters*, 85(19):4104, 2000. [5.2.2](#)
- [194] Pol Colomer-de Simón and Marián Boguñá. Double percolation phase transition in clustered complex networks. *Physical Review X*, 4(4):041020, 2014. [5.2.2](#), [5.2.2](#), [5.2.2](#)
- [195] Petter Holme. Core-periphery organization of complex networks. *Physical Review E*, 72(4):046111, 2005. [5.2.2](#)
- [196] Mark EJ Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, 2003. [5.2.2](#)
- [197] M Ángeles Serrano and Marián Boguñá. Clustering in complex networks. ii. percolation properties. *Physical Review E*, 74(5):056115, 2006. [5.2.2](#)
- [198] Jing Zhao, Guo-Hui Ding, Lin Tao, Hong Yu, Zhong-Hao Yu, Jian-Hua Luo, Zhi-Wei Cao, and Yi-Xue Li. Modular co-evolution of metabolic networks. *BMC Bioinformatics*, 8(1):311, 2007. [5.2.2](#)
- [199] Márcio Rosa da Silva, Hongwu Ma, and An-Ping Zeng. Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks. *Proceedings of the IEEE*, 96(8):1411–1420, 2008. [5.2.2](#)
- [200] John P McCutcheon and Nancy A Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26, 2012. [5.2.2](#)
- [201] Nancy A Moran. Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586, 2002. [5.2.2](#)
- [202] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010. [6.2](#)



- 
- [203] Sergio Gomez, Albert Diaz-Guilera, Jesus Gomez-Gardeñes, Conrad J Perez-Vicente, Yamir Moreno, and Alex Arenas. Diffusion dynamics on multiplex networks. *Physical Review Letters*, 110(2):028701, 2013. [6.2](#)
- [204] Kyu-Min Lee, Byungjoon Min, and Kwang-Il Goh. Towards real-world complexity: an introduction to multiplex networks. *The European Physical Journal B*, 88(2):1–20, 2015. [6.2](#)
- [205] Zhirong Yang and Jorma Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1):363–373, 2007. [A](#)
- [206] H Sebastian Seung and Daniel D Lee. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001. [A](#)
- [207] Duncan Cramer. A cautionary tale of two statistics: partial correlation and standardized partial regression. *The Journal of Psychology*, 137(5):507–511, 2003. [B](#)

# APPENDIX A

## NON-NEGATIVE MATRIX FACTORISATION

Non-negative matrix factorisation (NMF) is a relatively new dimension reduction technique, whereby a non-negative matrix  $A$  is factorised into two lower rank non-negative matrices  $W$  and  $H$ . A more formal definition of NMF of a matrix  $A$  is given by

$$A \approx WH, \text{ or}$$

$$A_{ij} \approx (WH)_{ij} = \sum_{a=1}^k W_{ia} H_{aj},$$

where here  $A \in \mathbb{R}^{m \times n}$ ,  $W \in \mathbb{R}^{m \times k}$  and  $H \in \mathbb{R}^{k \times n}$  for some rank  $k \ll \min(m, n)$ . From this approximation we have that each data vector of  $A$  is approximated by a linear combination of the columns of  $W$ , weighted by the components of  $H$ .

In order to find approximations to the matrices  $W$  and  $H$  one first needs to define a cost function that measures the accuracy of the approximation. The cost function regularly used in the literature is the root mean squared residual such that

$$D = \|A - WH\|_F^2, \tag{A.1}$$

subject to the constraints  $W \in \mathbb{R}_+^{m \times k}$  and  $H \in \mathbb{R}_+^{k \times n}$ . There exist several algorithmic approaches that can be used to find the approximations of  $W$  and  $H$ , whilst

minimising the cost  $D$ .

In this study we use the *Multiplicative Update rule* [205]. This algorithm uses an iterative method that starts with initial random matrices for  $H$  and  $W$  and then the following equations are used to update  $H$  and  $W$  alternatively,

$$H^{i+1} = H^i \cdot * \frac{W^{i^T} A}{W^{i^T} W^i H^i + 10^{-19}}$$

$$W^{i+1} = W^i \cdot * \frac{A H^{i^T}}{W^i H^i H^{i^T} + 10^{-19}}$$

In words, the new approximations of  $W$  and  $H$  are computed at each iteration by multiplying the current approximations by some factor that depends on the gradient of the cost function [205]. Lee and Seung prove in [206] that the accuracy of the approximation increases monotonically with further iterations. In practice, however, the solution space is nonlinear and so there is no guarantee that the method will converge to an optimal matrix factorisation. For that reason, the experiments of Chapter 4 were repeated 1000 times to ensure the approximation avoided local minima, thus improving the reliability of our results.

## Determination of a suitable value of $k$

As a rule of thumb, the rank of  $k$  is chosen so that  $(n + m)k < nm$ , thus resulting in a compressed version of the original data [165]. In Figures A.1 and A.2 we plot the results after carrying out the non-negative matrix factorisation using  $k = 1$  to  $k = 20$  on the habitat data and oxygen data respectively.

Importantly, we found that the choice of  $k$  did not alter the results considerably, with almost all values, up to and including  $k = 20$ , displaying a trend for increased complexity with habitat/environmental variability. In all the experiments in our work we chose  $k = 3$ , which not only enables much easier visual inspection of any results, but is also the standard choice in the literature.

## Significant motif frequency tables

Table A.1 and A.2 shows all those motifs that were returned as being significant in at least one run of our matrix factorisation for the two different data sets studied in Chapter 4.









Significant motif								
Specialised motif frequencies (43)	1000	900	27	0	683	0	99	591
Varied motif frequencies (72)	1000	899	903	887	221	918	4	854

Table A.1: Frequencies of the significant motifs obtained for the specialised and varied environmental classes after 1000 approximations of  $H$  using non-negative matrix factorisation.









Significant motif								
Aerobic motif frequencies (334)	1000	996	998	994	24	999	1	20
Anaerobic motif frequencies (49)	1000	994	987	0	35	987	1	316

Table A.2: Frequencies of the significant motifs obtained for the aerobic and anaerobic classes after 1000 approximations of  $H$  using non-negative matrix factorisation.

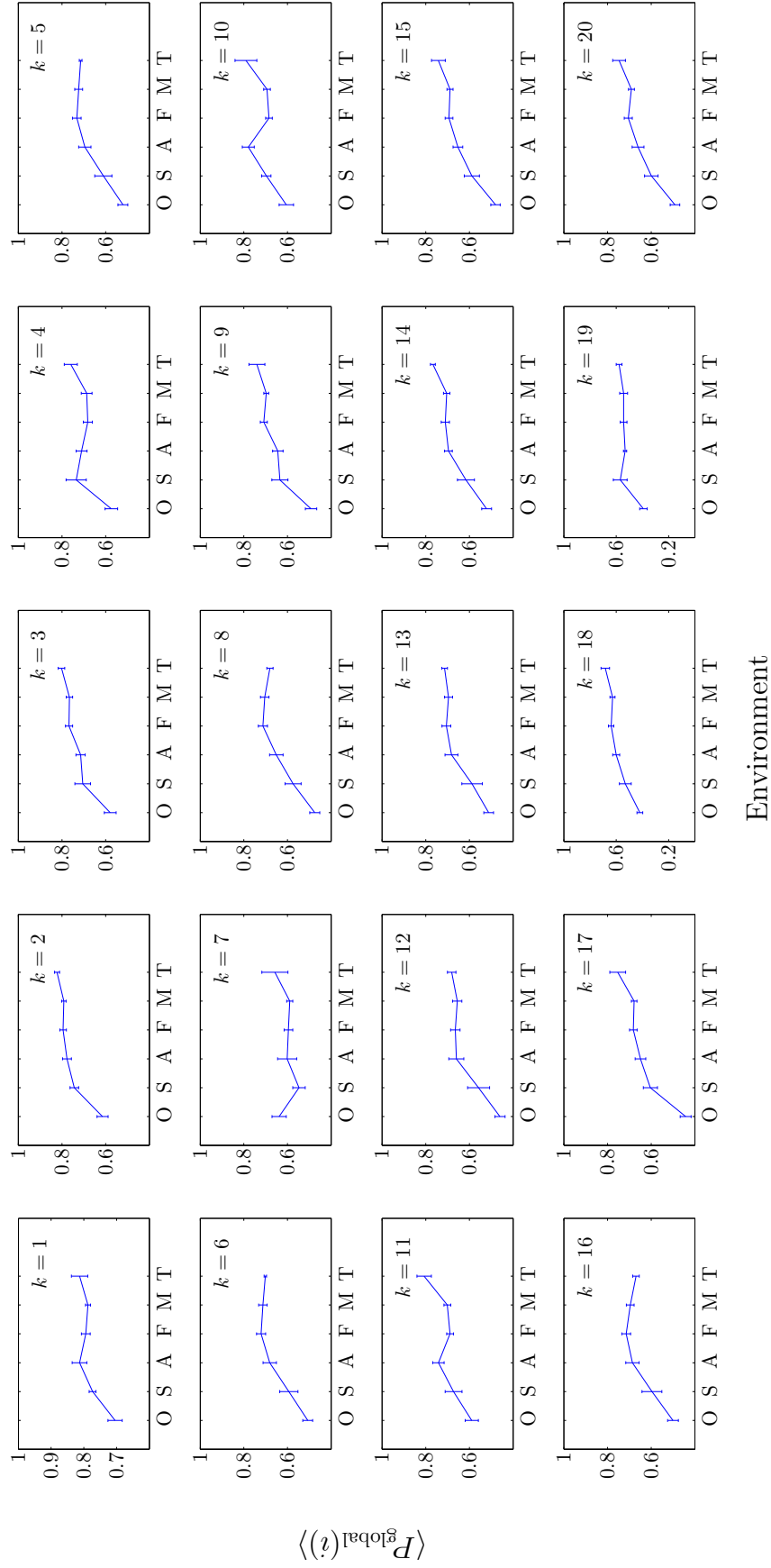


Figure A.1: Subplots showing the relationship between environmental variability and the mean global significance score  $\langle P_{\text{global}}(i) \rangle$  for the 6 bacterial habitats: **O**bligate, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. Each subplot shows the result for a different value of  $k$ .

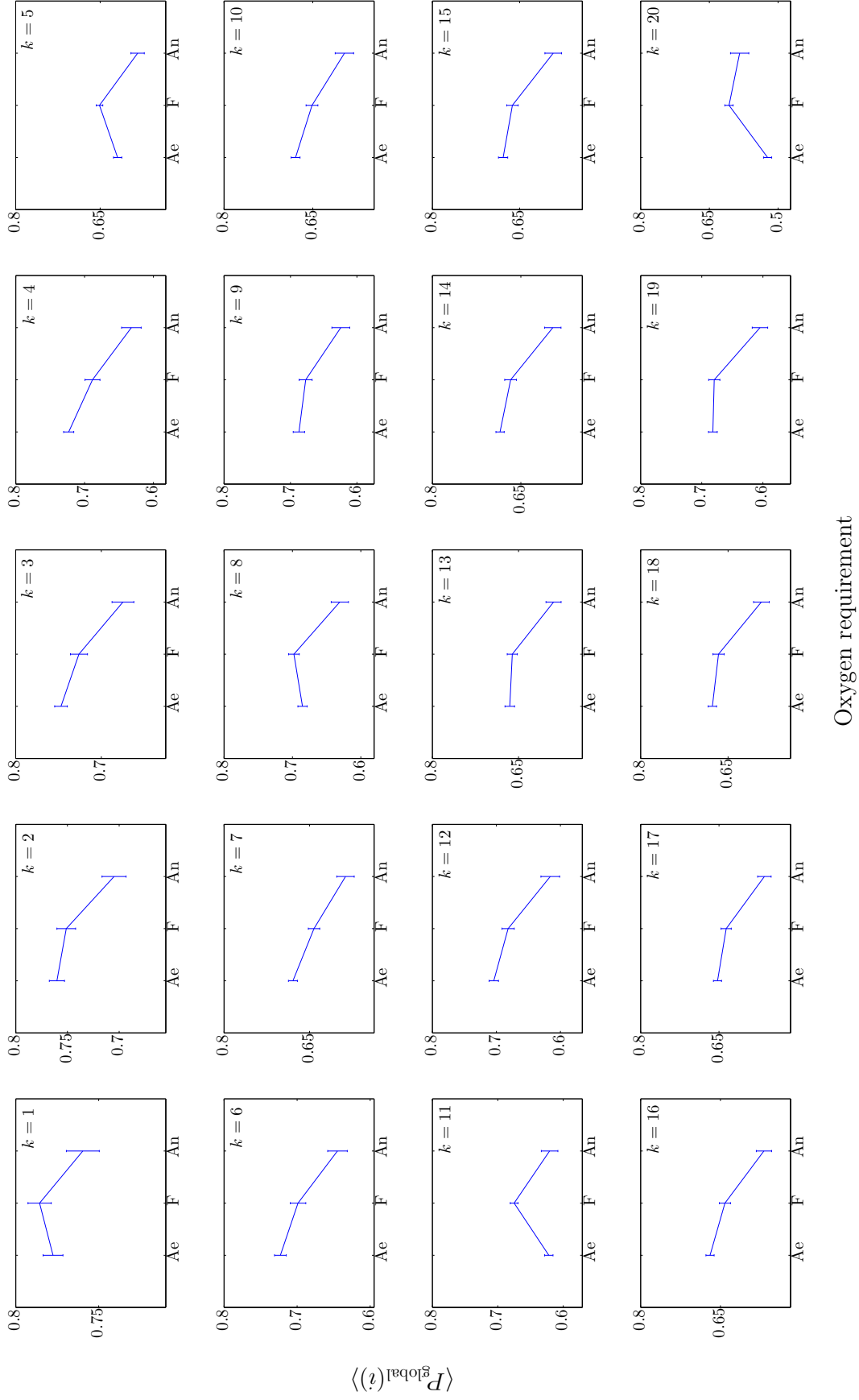


Figure A.2: Subplots showing the relationship between growth conditions, more specifically oxygen requirements, and the mean global significance score  $\langle P_{\text{global}}(i) \rangle$  for the 3 classes: **Aerobic**, **Facultative** and **Anaerobic**. Each subplot shows the result for a different value of  $k$ .

# APPENDIX B

## ADDITIONAL PROPERTIES OF THE METABOLIC NETWORKS

In the following experiments, we have investigated the relationship between growth conditions and some additional properties of the metabolic networks.

### **Relationship between growth conditions and basic network measures**

We have considered the relationship between environmental conditions and three basic network measures (total-degree, clustering coefficient and average path-length), and then investigated the extent to which they correlate with  $P_{\text{global}}$ .

### **Habitat variability**

Figure [B.1a](#) shows that the trend for the average total degree is to increase with variability in the environment, although not monotonically. In Figure [B.1b](#) we plot the average total degree against our global significance score  $P_{\text{global}}$  and find a significant correlation ( $r = 0.8983$ ,  $p < 10^{-5}$ ).

Figure B.1c shows that the average path length is the smallest for the networks within the obligate class (i.e. the most specialised) and the largest for the terrestrial class (i.e. the most varied). The specialised, aquatic, facultative and multiple class however, all have very similar values. In Figure B.1d we have plotted the average path length against the global significance score  $P_{\text{global}}$  and find a significant correlation ( $r = 0.8723$ ,  $p < 10^{-5}$ ).

Figure B.1e shows that the clustering coefficient does not follow any particular trend as regards environmental habitat. Figure B.1f shows that the average clustering coefficient is only weakly correlated with the global significance score  $P_{\text{global}}$  ( $r = 0.2531$ ,  $p < 0.01$ ).

## Oxygen requirements

Figure B.2a shows that the average total degree does not appear to follow any particular trend regarding oxygen requirements, despite being correlated to the global significance score  $P_{\text{global}}$  ( $r = 0.7932$ ,  $p < 10^{-5}$ ) (Figure B.2b).

Figure B.2c shows that the average path length is larger for the metabolic networks that evolved in the presence of oxygen, that is, the aerobic and facultative class. In Figure B.2d we have plotted the average path length against the global significance score  $P_{\text{global}}$  for the 383 metabolic networks and find a significant correlation ( $r = 0.8617$ ,  $p < 10^{-5}$ ).

Figure B.2e shows a relationship similar to the average total degree (Figure B.2a), that is, the facultative class has a significantly larger amount of clustering present than the aerobic and anaerobic classes. Figure B.2f (b) shows that the average clustering coefficient and the global significance score  $P_{\text{global}}$  are not correlated ( $r = 0.0737$ ,  $p = 0.1489$ ).



### **Spearman’s partial correlation between $P_{\text{global}}$ and environment conditioned on basic network measures**

Spearman’s partial correlation between  $X$  and  $Y$  conditioned on  $Z$  allows one to compute the correlation between  $X$  and  $Y$ , discounting the correlations between  $X$  and  $Z$  and between  $Y$  and  $Z$  [207]. We computed the correlation between  $P_{\text{global}}$  and variability conditioned on the simpler network metrics considered in the previous section (degree, path-length and clustering) and found that our results remained significant ( $c = 1$ ,  $p < 10^{-5}$ ). Note that we use Spearman’s correlation since the data consists of a mixture of both ordinal and continuous values; correlations were computed using the `partialcorr` function which is available in the MATLAB Statistics Toolbox.

### **Relationship between growth conditions and genome size**

We considered an organism’s genome size as a quantitative measure of environmental variability, and then investigated the extent to which it correlates with  $P_{\text{global}}$ .

#### **Habitat variability**

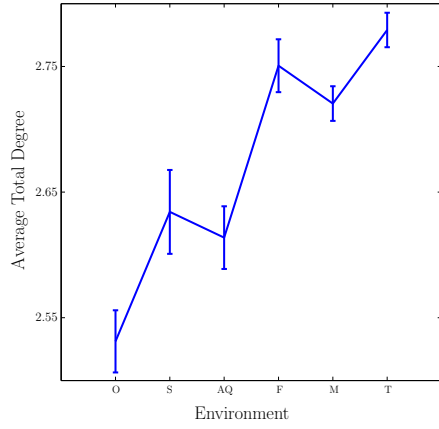
Figure B.3a shows that the trend for the total number of genes is to increase with variability in the environment. We find this relationship to be significant using the Kruskal-Wallis test (KW) (p-value:  $p < 10^{-8}$ ). In Figure B.3b we plot the total number of genes against our global significance score  $P_{\text{global}}$  and find a significant correlation ( $r = 0.7757$ ,  $p < 10^{-23}$ ).

As a control for the effect of genome size, we computed the correlation between  $P_{\text{global}}$  and environmental variability, conditioned on the total number of genes using the Spearman’s partial correlation (as described in the previous section). We find that our results remain significant ( $c = 0.4239$ ,  $p = 10^{-5}$ ).

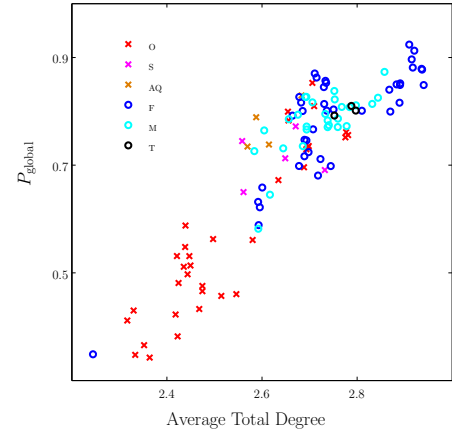
## Oxygen Requirements

Figure B.4a shows that the trend for the total number of genes decreases monotonically with oxygen availability. We find this trend to be significant using the Kruskal-Wallis test (KW) (p-value:  $p < 10^{-2}$ ). In Figure B.4b we plot the total gene number against our global significance score  $P_{\text{global}}$  and find a significant correlation ( $r = 0.6415$ ,  $p < 10^{-45}$ ).

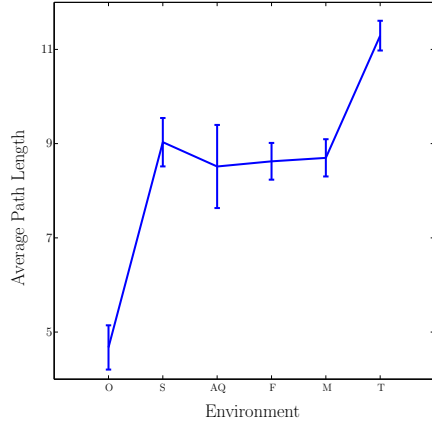
As a control for the effect of genome size, we computed the correlation between  $P_{\text{global}}$  and oxygen availability, conditioned on the total number of genes using the Spearman's partial correlation. We find that our results remain significant ( $c = -0.1020$ ,  $p = 0.0464$ ).



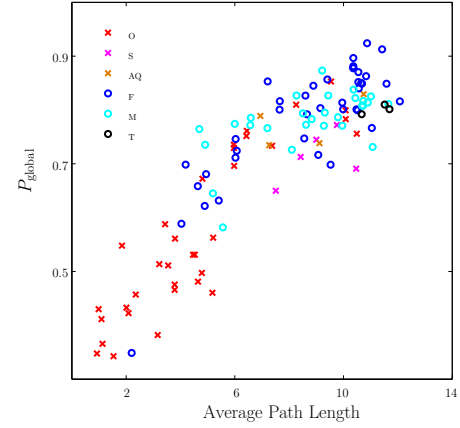
(a) Average total degree and environmental variability.



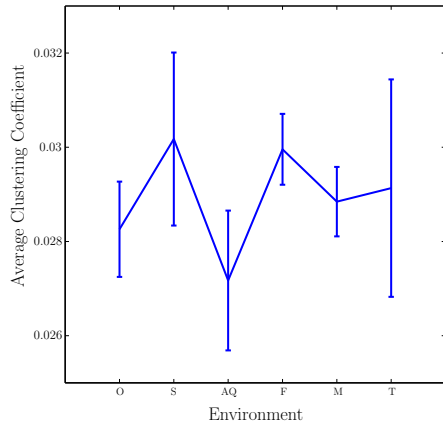
(b) Average total degree versus the global significance score  $P_{\text{global}}$ .



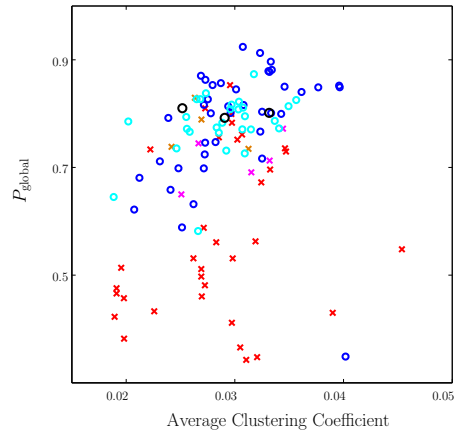
(c) Average path length and environmental variability.



(d) Average path length versus the global significance score  $P_{\text{global}}$ .

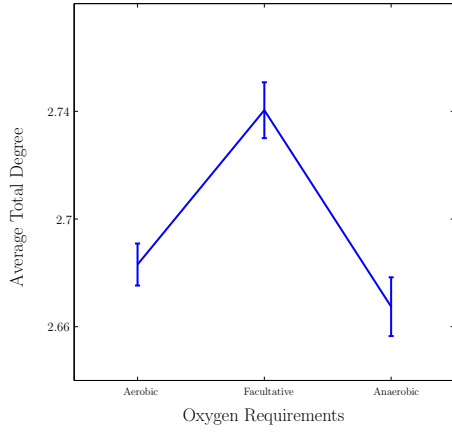


(e) Clustering coefficient and environmental variability.

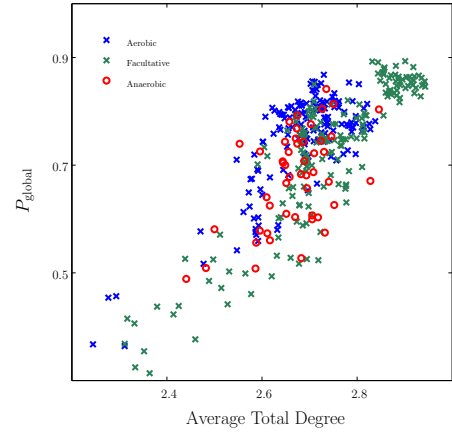


(f) Clustering coefficient versus the global significance score  $P_{\text{global}}$ .

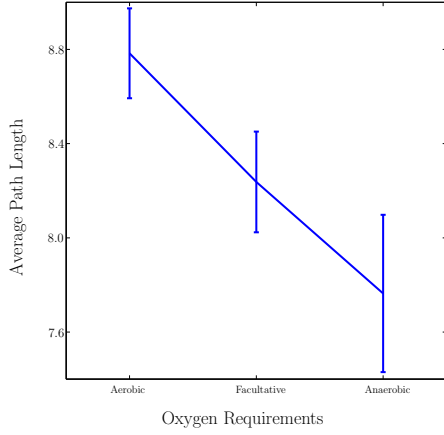
Figure B.1: In (a), (c) and (e) we plot the relationship between three basic network measures and environmental variability for the 6 environmental classes. Note that here we plot the mean value over each environmental class: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. In (b), (d) and (f) we show the three basic network measures plotted against the global significance score  $P_{\text{global}}$  for the 115 bacterial networks.



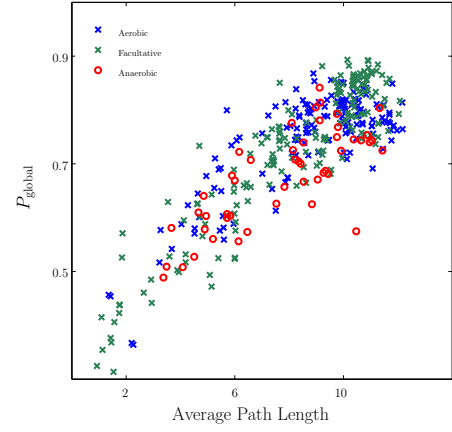
(a) Average total degree and oxygen requirements.



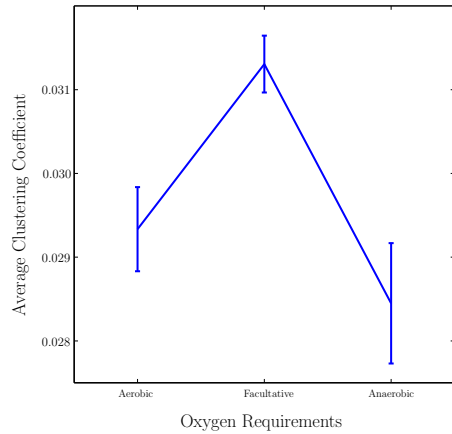
(b) Average total degree versus the global significance score  $P_{\text{global}}$ .



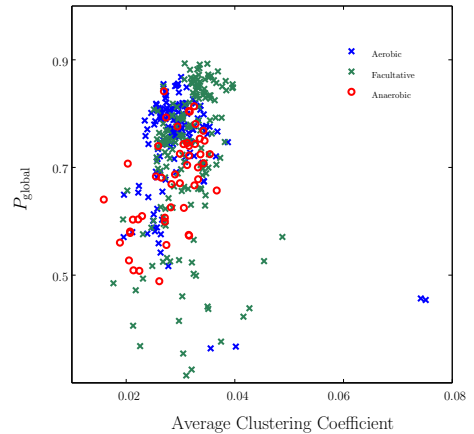
(c) Average path length and oxygen requirements.



(d) Average path length versus the global significance score  $P_{\text{global}}$ .

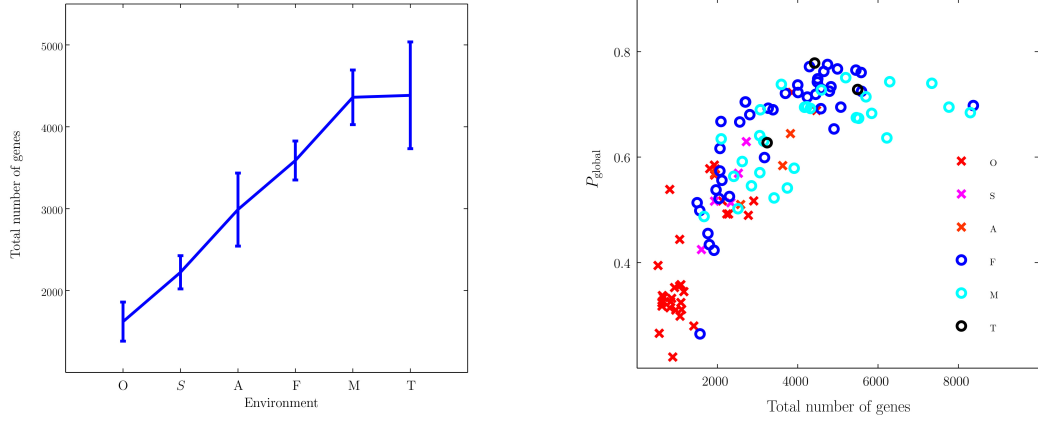


(e) Clustering coefficient and oxygen requirements.



(f) Clustering coefficient versus the global significance score  $P_{\text{global}}$ .

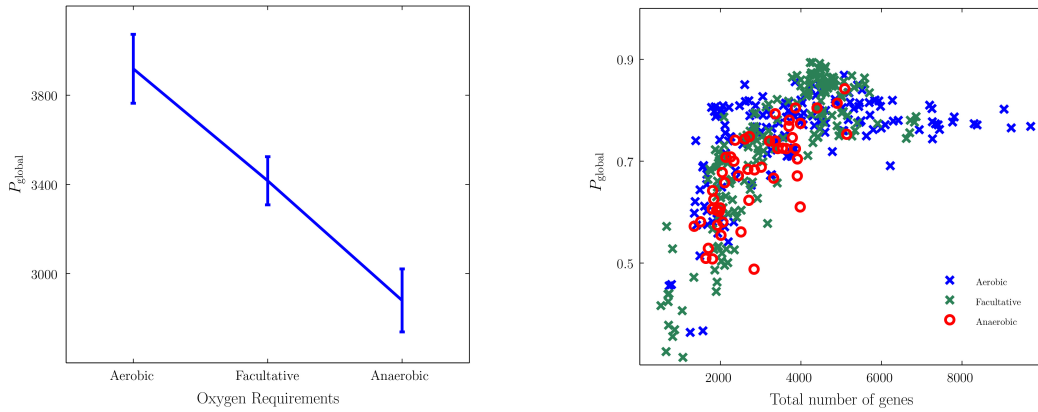
Figure B.2: In (a), (c) and (e) we plot the relationship between three basic network measures and growth conditions (oxygen requirements) averaged over the 3 classes: **A**erobic, **F**acultative and **A**naerobic. In (b), (d) and (f) we show the three basic network measures plotted against the global significance score  $P_{\text{global}}$  for the 383 bacterial networks.



(a) Total number of genes and environmental variability.

(b) Total number of genes versus the global significance score  $P_{\text{global}}$ .

Figure B.3: In (a) we plot the relationship between the genome size (total number of genes) and environmental variability for the 6 environmental classes: **O**bligat, **S**pecialised, **A**quatic, **F**acultative, **M**ultiple and **T**errestrial. In (b) we show the total number of genes plotted against the global significance score  $P_{\text{global}}$  for the 115 bacterial networks.



(a) Total number of genes and oxygen requirements.

(b) Total number of genes versus the global significance score  $P_{\text{global}}$ .

Figure B.4: In (a) we plot the relationship between the genome size (total number of genes) and growth conditions (oxygen requirements) averaged over the 3 classes: **A**erobic, **F**acultative and **A**naerobic. In (b) we show the total number of genes plotted against the global significance score  $P_{\text{global}}$  for the 383 bacterial networks.