Spectral Analysis of Neutral Evolution

Dissertation presented for the degree of Master of Science in the Department of Computer Science, University of Cape Town, October 2017, Supervised by Dr. Geoff Nitschke and Prof. Dr. Agoston Eiben.

David Shorten

Abstract

It has been argued that much of evolution takes place in the absence of fitness gradients. Such periods of evolution can be analysed by examining the mutational network formed by sequences of equal fitness, that is, the neutral network. It has been demonstrated that, in large populations under a high mutation rate, the population distribution over the neutral network and average mutational robustness are given by the principal eigenvector and eigenvalue, respectively, of the network's adjacency matrix. However, little progress has been made towards understanding the manner in which the topology of the neutral network influences the resulting population distribution and robustness. In this work, we build on recent results from spectral graph theory and utilize numerical methods to enhance our understanding of how populations distribute themselves over neutral networks. We demonstrate that, in the presence of certain topological features, the population will undergo an *exploration catastrophe* and become confined to a small portion of the network. We further derive approximations, in terms of mutational biases, for the population distribution and average robustness in networks with a homogeneous structure. The applicability of these results is explored, first, by a detailed review of the literature in both evolutionary computing and biology concerning the structure of neutral networks. This is extended by studying the actual and predicted population distribution over the neutral networks of H1N1 and H3N2 influenza haemagglutinin during seasons between 2005 and 2016. It is shown that, in some instances, these populations experience an exploration catastrophe. These results provide insight into the behaviour of populations on neutral networks, demonstrating that neutrality does not necessarily lead to an exploration of genotype/phenotype space or an associated increase in population diversity. Moreover, they provide a plausible explanation for conflicting results concerning the relationship between robustness and evolvability.

CONTENTS

NOMENCLATURE

AI Artificial Intelligence, page 15

EC Evolutionary Computing, page 7

XBS Xulvi-Brunet - Sokolov, page 22

CHAPTER 1

INTRODUCTION

When an entity undergoes evolutionary change, much of this change may not be due to a response to selective pressure, but rather due to the discovery of variants with equivalent fitness. It has been argued that the majority of genetic change in natural organisms is due to such neutral mutations (Kimura, 1983). In *Evolutionary Computing* (EC) (Eiben & Smith, 2015), it has been found that many fitness functions result in a substantial proportion of mutations being neutral (Galván-López et al., 2011). It has even been proposed that neutrality plays an important role in the evolution of technological innovations (Lobo et al., 2004).

A highly productive abstraction for studying neutrality is the *neutral network* (van Nimwegen, 2006; Van Nimwegen et al., 1999). This is a network whose nodes represent genotypes of a given fitness and where an edge connects two nodes if the associated genotypes differ by a single point mutation. It has been shown that, under certain assumptions, these networks permeate sequence space and that every phenotype is reachable by traversing the network (Reidys et al., 1997).

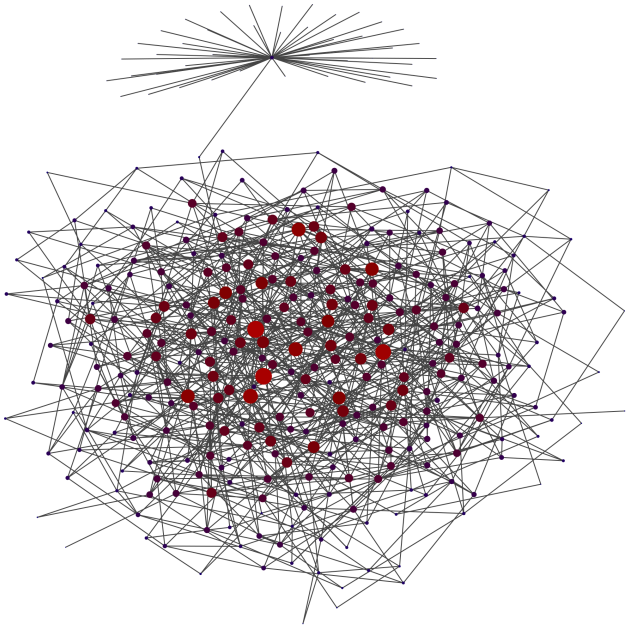
A variety of authors have demonstrated the substantial impact of neutrality on evolutionary dynamics (Koelle et al., 2006; Van Nimwegen et al., 1999; Newman & Engelhardt, 1998). Much of this analysis has focused on how, in instances where no advantageous mutations exist, neutrality prevents the population from getting stuck at a certain point in sequence space. Instead, it can explore the neutral network until it finds an advantageous phenotype lying ad-

jacent to the network (Fontana & Schuster, 1998; Gavrilets, 1997). Moreover, it has been demonstrated that larger neutral networks allow for more such "stepping off points" (Wagner, 2008), facilitating the discovery of adaptive and innovative phenotypes. It has further been shown that large neutral networks allow the population to spread out and gain standing variation. This facilitates the population's adaptive response to changes in its environment (Masel & Trotter, 2010). However, there is some ambiguity as to whether neutrality is universally beneficial to evolution (Cuevas et al., 2009; Elena & Sanjuán, 2008; Galván-López et al., 2011).

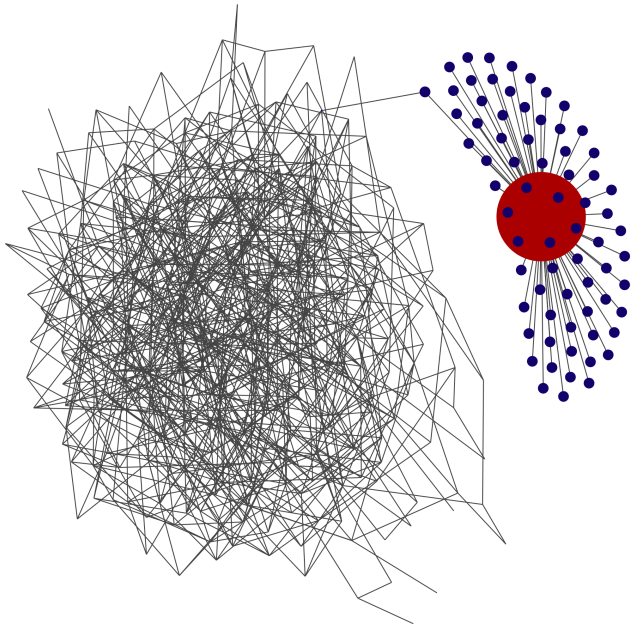
The seminal work in the modeling of evolutionary dynamics is that of Erik van Nimwegen, James P. Crutchfield and Martijn Huynen (1999). By employing a straightforward model of neutral evolution, the authors demonstrated the existence of two distinct behavioural regimes. If $M\mu \ll 1$, where M is the population size and μ is the per genome mutation rate, then the population is *monomorphic* (Bloom et al., 2007). Mutations either fix or go extinct, that is they either become present in the entire population or disappear from it completely. Conversely, if $M\mu \gg 1$, then the population is polymorphic and mutations do not fix. The population distributes itself over a number of nodes in the network. More specifically, the population's distribution is given by the network's principal eigenvector and its average robustness (average number of neutral neighbours) is given by the network's principal eigenvalue. Random walks are a very well described phenomenon (Lovász, 1993), and so this work focuses exclusively on the, more interesting, polymorphic case.

As closed-form solutions to the eigenvalues and eigenvectors of graphs do not exist, these are somewhat opaque quantities. However, various authors have been able to draw some conclusions from this result. Firstly, the population spreads out, or diffuses, over the neutral network, gaining variation (Manrubia & Cuesta, 2010; Crutchfield & Schuster, 2003; Hu et al., 2011; Masel & Trotter, 2010). Secondly, the population will become more concentrated on the "most connected" nodes and, in so doing, increase the average robustness of the population (van Nimwegen, 2006; Van Nimwegen et al., 1999; Banzhaf & Leier, 2006). These conclusions are well founded, as the average degree of a network is a lower bound on the principal eigenvalue (Cioabă et al., 2010) and the principal eigenvector is a measure of centrality in a network, the *eigenvector centrality* (Bonacich, 1972), and, as such, assigns a non-zero centrality score to each node.

In this thesis, it is demonstrated that this description of the behaviour of



(a) Hub of degree 45.



(b) Hub of degree 70.

Figure 1.1: A localisation transition. A hub (star network) is connected to an Erdős-Renyi network by adding an edge between one of the star's peripheral nodes and a random node of the Erdős-Renyi network. The original Erdős-Renyi networks contained 400 vertices and 1200 edges. Node sizes are proportional to the corresponding component of the principal eigenvector of the adjacency matrix which is equal to the proportion of the population found on the node. Moreover, nodes with a higher eigenvector component are more red and nodes with a lower eigenvector component are more blue.

polymorphic populations can be refined. Although the average population robustness will always be higher than the network's average degree, we can construct examples where the population concentrates on a region of the network which does not agree with our intuition of "most connected". Moreover, networks can be constructed where the population concentrates on a small number of vertices and does not spread out, or diffuse, over it. Take, for instance, the two networks shown in figure 1.1. Both of these networks consist of an Erdős-Renyi network (Erdős & Renyi, 1959) with 400 vertices and 1200 edges connected to a hub (star network), where the connection to the hub is made via one of its peripheral vertices. In the first network, the hub is of degree 45 and in the second it is of degree 70. Despite the similarity of these two networks, the equilibrium distribution of the population over them is vastly different. In the first network, the population behaves roughly as we would expect and distributes itself fairly evenly over the network, being more concentrated on the more central nodes of the Erdős-Renyi component. It is worth noting that only a very small proportion of the population (around 0.5%) is found on the hub or its neighbours. However, in the second network, around 99.5% of the population is concentrated on the hub and its neighbours. This behaviour is observed regardless of the size of the Erdős-Renyi component, so long as the average degree of this component is kept constant.

It is worth briefly contemplating the implications of this behaviour. At equilibrium, the population is not exploring the neutral network, but is localized on a small part of it. Moreover, the amount of genetic variation within the population is small, given that almost all members are within a Hamming distance of one from the hub sequence. Furthermore, although the average robustness of the population is high, the average degree of the hub region on which it is concentrated is low (star networks have an average degree of two). This is in contrast with the higher average degree of the Erdős-Renyi component, which has an average degree of six. The population is, therefore, not located on the "most connected" part of the network.

There is a further consideration which would be beneficial to include in any description of neutral evolution. This is the effect that mutational biases have on the evolutionary process. Various biases present in landscapes have been studied, including neighbouring genotypes containing similar mutational neighbourhoods (Greenbury et al., 2016), the correlation of the robustness of neighbouring genotypes (Payne et al., 2014) and the overrepresentation of certain phenotypes

(Schaper & Louis, 2014).

A well known result in graph theory is that the neighbours of nodes in a network have an average degree higher than the average degree of the network. This result is named the *friendship paradox* (Feld, 1991), referring to the fact that, on average, people's friends have more friends than they do [1]. This result is due to the fact that a sampling of vertices at the ends of edges is biased towards higher degree vertices. The implication of this result is that, when a sequence undergoes mutation, it is biased towards more mutationally robust sequences. This mutational bias should have a significant impact on evolutionary dynamics during neutral epochs.

Furthermore, we can question whether the distribution of robustness amongst mutants is dependent on the robustness of the parent genotype and the manner in which this influences neutral evolution. This would represent an additional bias on the effects of mutations on the robustness of offspring. Moreover, as Darwinian evolution requires *heritable* variation, we should expect that the evolution of high levels of robustness would require that the robustness of offspring be correlated with the robustness of parents. Such a correlation has a direct analogue in terms of neutral network properties: the network's *assortativity* (Newman, 2003). Assortativity refers to the correlation in the degrees at either end of an edge. We should, therefore, expect assortative neutral networks to produce populations with a higher average robustness than disassortative networks.

The principal eigenvectors and eigenvalues of graphs are of great importance to a variety of problems (Restrepo et al., 2007), principally synchronization phenomena and the spread of epidemics. Since the publication of van Nimwegen et. al.'s work, there has been substantial progress in approximating these quantities in terms of network properties (Goltsev et al., 2012). In this work, we build on these results in order to incorporate the above observations and intuitions into a more complete understanding of the evolution of polymorphic populations on neutral networks.

Ancel & Fontana (2000) demonstrated that, for evolving populations of RNA sequences with plastogentic congruence, the population could undergo an *exploration catastrophe*, whereby it would be confined to a small portion of the neutral network. It has recently been demonstrated (Martin et al., 2014) that the principal eigenvector is a poor measure of centrality in networks. This is

[1] It is interesting to note that, on average, people believe that they have more friends than their friends do (Zuckerman & Jost, 2001)

due to the fact that certain structural heterogeneities can cause the eigenvector to localize on certain portions of the network, assigning almost all of its weight to these portions and very little to the rest. We make the argument here that this localisation phenomenon has important implications for the neutral evolution of asexual populations at high mutation rates. Specifically, in neutral networks with certain topological features, the population will undergo an *exploration catastrophe*. Moreover, this phenomenon will occur without the presence of special properties of the genotypes or phenotypes, such as plastogenetic congruence, and occurs independent of mutation rate. We use computational methods to confirm that this localisation of the eigenvector occurs in biologically plausible neutral networks. We further demonstrate novel modes of eigenvector localisation not yet explored in the literature.

On the other hand, for networks with a homogeneous topology, which can be well approximated by a mean-field approach (Gleeson et al., 2012), we derive an approximation of the equilibrium distribution of the population in terms of the mutation sampling bias provided by the friendship paradox and the network's assortativity. More specifically, we show that, in the absence of a correlation between parent and offspring robustness, the average robustness of the population is equal to what would be found through a sampling of genotypes by randomly selecting and following allowed mutations on the neutral network. It rises above or below this in the presence of positive or negative assortativity, respectively.

The specific contributions of this thesis are as follows.

- The argument that the localization phenomenon of the principal eigenvector on graphs implies the possibility of an *exploration catastrophe* occurring when large populations evolve asexually on neutral networks.
- A detailed analysis of the localization phenomenon on network models capturing biologically relevant features, namely, hubs loosely connected to random networks (§ 4.5) and the multiple hubs of the Barábasi-Albert model (§ 4.6).
- The confirmation that localisation can occur on graphs imbedded in Hamming space (§ 4.7 and § 4.8).
- The demonstration of two novel modes of localisation, namely: localisation across loosely connected components (§ 4.7) and localisation on Hamming balls connected to random graphs in Hamming space (§ 4.8).

- The derivation of approximations, in terms of sampling biases, for the equilibrium population distribution on neutral networks with an homogeneous structure (§ 3.2).

- An analysis of the neutral networks of H3N2 and H1N1 influenza during the seasons between 2005 and 2016. This analysis focuses on both the predicted and the actual population distribution over the networks, and demonstrates that, in some instances, the population is localised (chapter 7).

- A detailed analysis of the literature in evolutionary computing and biology reporting on the structure of neutral networks, and the implications of the above results to evolution on those networks (chapter 5 and chapter 6).

  The relationship between genetic robustness and evolvability has emerged as an important topic of research (Masel & Trotter, 2010; Wagner, 2008). A specific focus is the attempt to reconcile (Stern et al., 2014) conflicting experimental results (Cuevas et al., 2009; McBride et al., 2008) on whether robustness promotes or hinders evolvability. This work provides an additional plausible explanation for why robustness might sometimes be correlated with evolvability, whereas, in other instances, it is anti-correlated. Robustness that leads to a population spreading out over a neutral network should increase evolvability, whereas, robustness that results in an exploration catastrophe is likely to decrease evolvability.

  Furthermore, the existence of the exploration catastrophe has important implications for the *in vitro* evolution of proteins (Matsuura & Yomo, 2006) and evolutionary computing. In both these instances, diversification of the population during neutral epochs is highly desired, and the exploration catastrophe is a phenomenon which system designers would want to avoid.

  The position of a population on its neutral network determines the phenotypes to which it can mutate. While an exploration catastrophe limits the amount of variation to which a population is exposed, this limitation on available trajectories could plausibly facilitate the prediction of the population's future evolution (Łuksza & Lässig, 2014).

  The study of mutational biases has similarly emerged as a pertinent research theme (Greenbury et al., 2016; Payne et al., 2014; Schaper & Louis,

2014). The expressions for the population distribution and average robustness derived in § 3.2 are significant in that they directly relate biases to important metrics of the consequences of evolution on a given landscape, namely, genetic robustness and population distribution. Moreover, the derived expressions shed light on a mechanism for the evolution of robustness, an interesting question in it own right (de Visser et al., 2003).

CHAPTER 2

BACKGROUND

2.1 Representation, neutrality and robustness

When organisms undergo natural evolution, mutation does not act directly on their form, but rather on the genetic code. Similarly, in *Evolutionary Computing* (EC), a *representation* of the problem, upon which mutation can occur, must be identified. The problem of choosing such a representation, the *representation problem*, has been identified as a critical issue within EC (Eiben & Smith, 2015), as well as *Artificial Intelligence* (AI) in general (Nilsson, 2009). Historically, a large research focus within AI, particularly in computer vision (Lowe, 1999), was toward designing representations. However, much of the recent progress within the field has been due to the development of algorithms capable of *learning* representations (Bengio et al., 2013), as opposed to designing them by hand.

Representations, and genetic code, require a mapping between themselves (genotype) and the organism or resulting problem solution (phenotype): the G→P map. The developmental process which translates genetic information into various biological organisms is not well understood (Pigliucci, 2010). Yet, it has become clear that this mapping is neither one-to-one nor linear (Gjuvsland et al., 2013). In many organisms and *Ribonucleic Acid* (RNA) folding (Draper, 1992), it has been found that genetic change resulting from mutation is not proportional to phenotypic change (Pigliucci, 2010; Wagner, 2008; Parter et al., 2008). Moreover, the G→P map is highly degenerate, that is, many genotypes might en-

code for an identical phenotype (Pigliucci, 2010).

There exists great variation in the mappings between representations and candidate solutions used in EC. On the one hand, in genetic algorithms, the relationship between representation and solution is often somewhat straightforward (Eiben & Smith, 2015). However, within the field of *generative and developmental systems* (Devert, 2009), many highly complicated mappings between representations and evolved forms have been proposed. Such mappings have been applied to a variety of tasks, including robot morphologies and organisms in artificial life studies (Stanley & Miikkulainen, 2003). Although the properties of individual mappings depend on their definition, some have been shown to be highly degenerate.

Degeneracy introduces the possibility that, when mutated, a genotype will still map to the same phenotype. This implies that the mutation has no effect on fitness and so can be labeled as *neutral*. Kimuraet al. (1968), along with King & Jukes (1969), brought the importance of neutral mutations to the attention of the scientific community through what has come to be known as the *neutral theory of molecular evolution*. This posits that the majority of evolutionary change is the result of the fixation of neutral mutations, as opposed to mutations which confer a selective advantage. Although the level of importance that such genetic drift has on evolution has been controversial (Nei, 2005), it is beyond doubt that certain mutations of certain organisms and structures are selectively neutral (Noirel & Simonson, 2008; Wagner, 2014; Bornberg-Bauer, 1997).

If the genetic code is a string of characters, as opposed to, say, a vector of real numbers, then one can construct networks out of genotypes coding for a given phenotype (Van Nimwegen et al., 1999). Here the vertices represent genotypes, and an edge connects two vertices if there exists a point mutation between their associated genotypes, that is their genetic codes are a Hamming distance of one apart. These networks have come to be known as *neutral networks* (Galván-López et al. (2011) credit Harvey & Thompson (1997) as being the originators of the term). Neutral networks have been studied extensively (Van Nimwegen et al., 1999; Aguirre et al., 2009; Bornberg-Bauer, 1997; Noirel & Simonson, 2008) and it has been shown that, under certain assumptions, these networks permeate sequence space and that any common phenotype can be reached by traveling along them (Reidys et al., 1997).

An important associated concept is that of *mutational robustness* (Taverna & Goldstein, 2002). This refers to the proportion of mutations which leave the

phenotype unchanged. The greater the mutational robustness of the genotypes, the larger their neutral networks will be (Wagner, 2011). This has an impact on the *evolvability* of these genotypes, as they can access a greater variety of phenotypes through neutral drift. Moreover, populations can evolve so as to occupy the most connected parts of the network (Van Nimwegen et al., 1999), thus increasing their average robustness.

2.2 Quasispecies

When populations evolve asexually at high mutation rates, they conform to what are known as *quasispecies dynamics* (Andino & Domingo, 2015; Domingo et al., 2012). As the topic of this thesis concerns asexual neutral evolution at high mutation rates, a discussion of these dynamics is necessary. Under a high mutation rate, a population is no longer concentrated on a single optimal genotype (Lauring & Andino, 2010). Instead, it spreads out in sequence space in what is labeled either a mutant 'cloud' or 'swarm'. This has the implication that it is not the fitness of a single sequence which determines the evolutionary trajectory of a population in genotype space, but, rather, the fitnesses of sequences within an area of this space. This is due to the fact that, at high mutation rates, the frequency of a given genotype in a population, is not determined solely by its fitness, but also by the frequency of its mutational neighbours. Their frequency is, in turn, determined by both their fitness and the frequency of their mutational neighbours.

A significant phenomenon within quasispecies dynamics is what is known as 'survival of the flattest' (Wilke et al., 2001). This is the phenomenon whereby a sub-population situated on a single highly fit genotype surrounded by low fitness neighbours can be out-competed by a sub-population situated on a region of genotype space consisting of multiple medium fitness genotypes, each of which is well connected to one another. Here, the term 'survival of the flattest' refers to a visualisation of this effect on a two-dimensional genotype space, where fitness is represented by height. The single genotype would be represented by a single spike, whereas the region of medium fitness genotypes would be represented by a lower hill with a flat top. This phenomenon is important in the context of this thesis for two reasons. Firstly, it emphasizes the importance of neutrality for evolution under these dynamics (Lauring & Andino, 2010). If the

topology of the neutral network allows the sub-population to achieve a high level of mutational robustness, this will allow it to out-compete sub-populations with lower robustness. The second reason is that, as will be discussed later on in this thesis (see, in particular, § 4.5 and § 4.8), the topology of the neutral network can have a substantial impact on the average level of robustness evolved by a population. The fact that the robustness of sub-populations can allow them to out-compete sub-populations situated elsewhere in seuquence space implies that the topology of the neutral network has implications beyond just the isolated evolution on the network itself, but could play an important role in determining on *which* network the population converges.

As stated above, one of the contributions of this thesis is the demonstration that, under certain topological conditions, a population evolving on a neutral network undergoes a localisation transition: the *exploration catastrophe*. Quasispecies exhibit a well-studied delocalisation [1] transition: the *error catastrophe* (Tejero et al., 2011; Summers & Litwin, 2006). This occurs when the mutation rate is raised such that the population is no longer located on its mutant cloud, but is instead spread out over genotype space. Under this regime, adaptive evolution is unable to take place. It has further been argued that, in certain fitness landscapes, the population might undergo multiple localisation-delocalisation transitions (Tannenbaum & Shakhnovich, 2004; Tejero et al., 2011). As the mutation rate is increased, it occupies broader regions of sequence space with higher robustness and lower fitness until, eventually, it spreads out over the entire space.

2.3 Modeling Neutral Evolution

The seminal work on modeling neutral evolution is that of Van Nimwegen et al. (1999). The model used is the application of Manfred Eigen's original, ubiquitous, model of quasispecies evolution (Eigen, 1971), to evolution on neutral networks. This thesis makes use of this model. As some of the reasoning presented in this thesis is dependent on the details of this model, it is necessary to go through it in some detail.

A population of constant size M resides on the neutral network of size N. Each generation, M genotypes are selected with replacement from the popula-

[1] It is worth noting that, assuming that the controlling parameter can be changed in both directions, a localisation transition can be turned into a delocalisation transition by altering the direction of change in the parameter.

tion and undergo mutation with probability μ. Those genotypes which undergo mutation will either stay on the network, or mutate off it.

Van Nimwegen et al. (1999) found that two behavioural regimes emerge. Given a population size M and a mutation rate μ, then if $M\mu \ll 1$ the population is *monomorphic* (Bloom et al., 2007). Mutations either fix or disappear, that is they either become present in the entire population or none of it. Thus, the entire population is concentrated on a single node of the neutral network. Throughout the neutral epoch the population performs a random walk over the network. On the other hand, if $M\mu \gg 1$, the population is polymorphic and spreads out over the neutral network (Wagner, 2011). Populations of self-replicating RNA, viruses and bacteria are polymorphic , whereas larger organisms are monomorphic (Wagner, 2011). Given the simple dynamics of the monomorphic case, this work focuses exclusively on polymorphic populations.

Van Nimwegen et al. (1999) further showed that the equilibrium distribution of a polymorphic population is given by the principal eigenvector of the adjacency matrix of the neutral network and that the average robustness of the population is given by the principal eigenvalue. An important aspect of this result, which has implications for the work presented in this thesis, is that the population distribution is not dependent on either the mutation rate, or the difference in fitness values between genotypes on and off the network (although genotypes adjacent to the network are required to have lower fitness).

2.4 Interpreting the Principal Eigenvector of Graphs

Closed-form solutions to the principal eigenvector and eigenvalue of graphs do not exist. This makes the result of Van Nimwegen et al. (1999) somewhat difficult to interpret. However, as mentioned in the introduction, various conclusions have been drawn. The population spreads out, or diffuses, over the neutral network, gaining variation (Manrubia & Cuesta, 2010; Crutchfield & Schuster, 2003; Hu et al., 2011; Masel & Trotter, 2010). Moreover, the population tends to concentrate on the "most connected" nodes and, in so doing, increase the average robustness of the population (van Nimwegen, 2006; Van Nimwegen et al., 1999; Banzhaf & Leier, 2006). These conclusions can be justified by the facts that the average degree of a network is a lower bound on the principal eigenvalue (Cioabă et al., 2010) and the principal eigenvector is a measure of centrality in

a network, the *eigenvector centrality* (Bonacich, 1972), which assigns a non-zero centrality score to each node.

There has been a certain amount of work towards refining this picture. Reeves et al. (2016) were able to derive an upper limit to the principal eigenvalue in terms of the size of the network, by utilising the fact that neutral networks are subgraphs of a hypercube graph (that is, they are embedded in Hamming space). This work, however, said nothing about the effect of other topological features and, moreover, has no implications for the principal eigenvector. Noirel & Simonson (2008) were able to show, in simulation, that degree assortativity and the existence of hubs increased the average robustness of populations. Bornberg-Bauer & Chan (1999) studied the population distribution on protein neutral networks, using Hamming balls as an abstraction for their structure. It was found that this structure lead to a slight concentration of the population towards the center of the ball.

2.5 Complex Networks

Network science has emerged as a powerful tool for studying complex phenomena involving the interactions of a large number of components (Barabási, 2016). Furthermore, it has already been fruitfully applied to the interrogation of biological structure and function (Wuchty et al., 2006; Barabasi & Oltvai, 2004). Techniques from network science have been successfully applied to a wide variety of domains, including gene regulatory networks (Pechenick et al., 2012), protein-protein interaction networks (Jeong et al., 2001), metabolic networks (Zhao et al., 2006), neural networks (Sporns, 2010) and ecological networks (Bascompte, 2010). They have also been applied to studying the mutational structure of genotypes coding for given phenotypes (Samal et al., 2010; Wagner, 2014) as well as neutral networks (Aguirre et al., 2011) (see chapter 6). The rest of this section will go through some concepts and techniques from network science which are used in this thesis.

Connected Components

A network can be separated into its connected components (Barabási, 2016). These are sets of nodes, where each node in the set can be reached from another node in the set by traversing edges. The result of Van Nimwegen et al. (1999)

assumes that the network is connected, that is, it has a single connected component. As this work builds on that result, we are also required to make this assumption. However, the term "neutral network" is often used to refer to all genotypes of a given fitness. This network could have multiple connected components. Before applying the results of Van Nimwegen et al. (1999) and those in this thesis, one would have to separate the network into its connected components. However, constantly referring to "a connected component of a neutral network" is awkward. Therefore, for the remainder of this thesis, we abuse terminology slightly, and use the term "neutral network" to refer to a connected component of a neutral network.

Assortativity

Degree assortativity (from hereon, referred to as *assortativity*) refers to the Pearson correlation between the degrees of the nodes at either end of the edges in a network (Newman, 2003). It is measured by the *assortativity coefficient* r:

$$r = \frac{\sum_{jk} jk\left(e_{jk} - q_j q_k\right)}{\sigma_q^2} \tag{2.1}$$

Here, j and k are the remaining degree of the vertices at either end of the edge. That is, the degree of the vertex excluding the edge which we are observing. q_k is the distribution of the remaining degree, e_{jk} the joint distribution of the remaining degree and σ_q^2 is the variance of this joint degree distribution.

Friendship Paradox

This effect is named after the phenomenon where, in social networks, the average number of friends of friends is higher than the average number of friends. Moreover, this effect is present in all networks, where the average number of neighbors of neighbors is higher than the average number of neighbors of nodes in the network. The cause of this paradox is that sampling the degrees of neighbors is equivalent to sampling the degrees of nodes at the ends of edges, which is biased towards higher degree nodes. The relationship between these two averages can be expressed as: (Feld, 1991)

$$\hat{\lambda} = \langle k \rangle + \frac{\sigma_n^2}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} \tag{2.2}$$

where $\langle k \rangle$ is the average degree (robustness) of genotypes on the neutral network, σ_n^2 is the variance of these degrees and $\hat{\lambda}$ is the average degree of single mutation neighbors. An implication of this result, as demonstrated by van Nimwegen et. al. (Van Nimwegen et al., 1999), is that random walks on neutral networks result in an average neutrality equal to $\hat{\lambda}$.

Xulvi-Brunet - Sokolov Algorithm

Later sections of this thesis make heavy use of randomly generated networks. In order to study networks with varying values of the assortativity coefficient r, the Xulvi-Brunet - Sokolov (XBS) algorithm was used (Xulvi-Brunet & Sokolov, 2004). This algorithm takes, as input, a given graph and rewires it to be more or less assortative. It operates by randomly selecting two edges in the graph. These edges are removed, however, if the goal is to create an assortative network, then, with probability α, the two nodes with the highest degrees are connected by a new edge. Similarly, the two nodes with the lowest degrees are also connected. Otherwise, the four nodes are randomly connected. if the goal is to create a disassortative network, then, with probability α, the highest degree node is connected to the lowest degree node and the remaining two nodes are also connected. Otherwise, the four nodes are randomly connected.

It is important to note that this algorithm is *degree preserving*. The degrees of the nodes remain unchanged. Moreover, the repeated application of this rewiring rule produces an ergodic Markov chain. This Markov chain reaches a stationary state in which the networks have a certain level of assortativity. Once this state has been reached, the rewiring provides a uniform sampling of networks with the given degree sequence and level of assortativity (Ray et al., 2014).

There are, however, two disadvantages to this technique. The first is that there is no simple relationship between the parameter α and the resulting network assortativity r of the stationary distribution. The value of r produced is also dependent on aspects of the network's degree sequence (Barabási, 2016). However, using a value of $\alpha = 1$ will produce maximally assortative (or disassortative) networks. Therefore, by using a range of α values between 0 and 1 one can generate networks with all realizable levels of assortativity. Secondly, there are no precise guarantees on the mixing time of the Markov chain, that is the number of rewirings required before the algorithm produces an unbiased sampling of networks with the given degree sequence and level of assortativity (Ray

et al., 2014). That being said, it has been argued (Ray et al., 2014) that $5E - 30E$ rewirings should be sufficient, where E is the number of edges in the graph. In all the following simulations, 1 million rewirings were performed, as this was found to be the largest number which could be used without making each run prohibitively time consuming. The largest number of edges in any of the networks produced was 70 000.

CHAPTER 3

HOMOGENEOUS NETWORKS

3.1 Intuition

We would not expect populations to converge on an average level of robustness substantially lower than what a random walk provides. As stated in § 2.5, this is equal to the mutation sampling bias provided by the friendship paradox, $\hat{\lambda} = \langle k^2 \rangle / \langle k \rangle$. Although robust genotypes have a selective advantage in that they produce more viable offspring, if these offspring themselves are not robust it is difficult to see how the population could converge on this lineage. Therefore, the selection of robustness is facilitated by the existence of robust vertices whose offspring are also robust. This sort of higher-order mutational bias is provided by network assortativity, that is, correlation in the degrees of the vertices at the end of edges (Newman, 2002) (§ 2.5). In the following section, we derive an expression for the average population robustness, demonstrating that it is equal to the mutational sampling bias and rises above or below this figure depending on whether the network has positive or negative degree assortativity.

3.2 Derivation of Approximation

Mean-field analysis has proven to be an effective approach for approximating dynamics occurring on networks (Gleeson et al., 2012). Moreover, it has already proven to be effective for deriving approximations to the principal eigenvalue

of network adjacency matrices (Restrepo et al., 2007). In this chapter, we use a mean-field approach to derive expressions relating the principal eigenvalue and eigenvector of the adjacency matrix of the neutral network to mutational biases present in the landscape.

We simplify the model of Van Nimwegen et al. (1999) (§ 2.3) by assuming that the genomic mutation rate is exactly one and that mutations off the network are lethal. These assumptions do not affect the resulting approximations of the eigenvectors and eigenvalues as those authors demonstrated that the population distribution over the neutral networks is independent of the mutation rate and off-network fitness (so long as the off-network fitness is lower than the on-network fitness).

The simplified model is then as follows. The population of constant size M resides on a neutral network of size N. The total number of neighbours, neutral and non-neutral, that a given genotype can have is given by U, this limit is determined by the length of the genetic code and the size of the alphabet. Each generation, M genotypes are selected with replacement from the population. These individuals then undergo mutation. With probability k_i/U the individual remains on the network, where k_i is the degree of the node representing the individual's genotype. If the individual stays on the network, it moves to one of its neighbouring nodes, chosen at random. If it mutates off the network then it is ineligible for selection in the subsequent generation.

In the delocalized regime, progress on approximating the population's distribution and robustness can be made by assuming that, at equilibrium, for every node in the network, the average population concentration on nodes at a given distance l is equal. That is, we utilise a mean-field approximation at a given distance l. This average concentration is the uniform concentration, that is the population size divided by the number of nodes. This is equivalent to assuming that the correlation length for the degrees is low. It has been found that, for most real-world networks, the correlation length is low (Mayo et al., 2015). Using this assumption we can approximate the proportion of the population which mutates onto a given node, and hence the population distribution and average robustness.

For the cases $l = 2$ and $l = 3$ we make use of the *annealed network approximation* (Dorogovtsev et al., 2008), whereby all nodes with a given degree k are approximated as having the same nearest neighbour degree distribution, which is the aggregate distribution over the neighbours of all nodes with degree k. This

has the implication that all nodes of degree k have the same average nearest neighbours degree, that is $\bar{k}_{nn}(i) = \bar{k}_{nn}(k_i)$, where i is a node's index and k_i the associated degree. We also use the approximation:

$$\bar{k}_{nn}(k) \approx \hat{\lambda} + \left(k - \hat{\lambda}\right) r \tag{3.1}$$

Where $\bar{k}_{nn}(k)$ is the average nearest-neighbour degree of nodes of degree k, $\hat{\lambda} = \langle k^2 \rangle / \langle k \rangle$ ($\langle k \rangle$ being the average degree $\langle k^2 \rangle$ being the average of the squares of the degrees) and r is the assortativity coefficient (the Pearson correlation between the degrees at either end of an edge, see § 2.5) (Newman, 2002). This approximation is derived by considering that r is the root of the coefficient of determination of the linear regression between the degrees of the nodes at either end of an edge.

We introduce the notation λ_1^l to denote the approximation of the principal eigenvalue (population average robustness) based on the assumption of equal average distribution at distance l. Similarly, we use $f_i\left(\lambda_1^l\right)$ to denote the i^{th} component of the principal eigenvector (the proportion of the population having the genotype represented by the i^{th} node), based on the assumption of equal average distribution at distance l.

Zero-hop Case

The simplest case is that we assume that the average population concentration at a distance zero from each node is equal, that is we assume that the population is uniformly distributed. The average robustness of the population is therefore, trivially, the average degree. Thus, we have:

$$f_i\left(\lambda_1^0\right) = \frac{1}{N} \tag{3.2}$$

$$\lambda_1^0 = \langle k \rangle \tag{3.3}$$

Where $\langle k \rangle$ is the average degree and N is the size of the network.

One-hop Case

The next case assumes that, at equilibrium, the average population concentration of the neighbours of each node is equal. Therefore, each generation, an average of $k_i M / NU$ individuals mutate onto node i. Normalizing, we arrive at

$$f_i\left(\lambda_1^1\right) = \frac{k_i}{N \langle k \rangle} \tag{3.4}$$

Multiplying by the robustness (k_i) and summing over all the nodes we arrive at the average robustness of:

$$\lambda_1^1 = \frac{\langle k^2 \rangle}{\langle k \rangle} = \hat{\lambda} \tag{3.5}$$

Two-hop Case

If we assume an average uniform population concentration two hops from each node, then, each generation, by the annealed network approximation, the nodes neighbouring node i will receive, on average, $\bar{k}_{nn}(k_i) M/NU$ mutants. This implies that node i will receive $k_i \bar{k}_{nn}(k_i) M/NU^2$ individuals. Substituting in equation (3.1) and normalizing we arrive at

$$f_i\left(\lambda_1^2\right) = \frac{1}{N\langle k^2 \rangle}\left(k_i \hat{\lambda} + k_i\left(k_i - \hat{\lambda}\right) r\right) \tag{3.6}$$

Multiplying through by the node's robustness (k_i) and summing over the nodes, we arrive at

$$\lambda_1^2 = \hat{\lambda} + \frac{r\sigma_e^2}{\hat{\lambda}} \tag{3.7}$$

Where $\sigma_e^2 = \langle k^3 \rangle / \langle k \rangle - \langle k^2 \rangle^2 / \langle k \rangle^2$ is the variance of the node's degrees when sampled by following edges. (3.7) is equivalent to the approximation derived by Goltsev et. al. (Goltsev et al., 2012) through the use of a power iteration.

Three-hop Case

Our final approximation is based on the assumption that, from any given node, the average population density at nodes three hops away is equal. We consider the node i', a neighbour of i. Each generation, this node will receive an average of $k_{i'} \bar{k}_{nn}(k_{i'}) M/NU^2$ mutants from its neighbours. We then average this over all neighbours i' of node i, that is we want to find

$$I = \frac{M}{NU^2}\left\langle k_{i'} \hat{\lambda} + k_{i'}^2 r - k_{i'} \hat{\lambda} r \right\rangle_{i'} \tag{3.8}$$

Using the fact that $\langle k_{i'}^2 \rangle_{i'} \approx \sigma_e^2 + \langle k_{i'} \rangle_{i'}^2$, where the equality is approximate as σ_e^2 is the global variance and not specific to the neighbours of nodes of degree $k_{i'}$, we can arrive at

$$I \approx \frac{M}{NU^2}\left(\hat{\lambda}^2 + \hat{\lambda}\left(k_i - \hat{\lambda}\right)r + \hat{\lambda}\left(k_i - \hat{\lambda}\right)r^2 + \left(k_i - \hat{\lambda}\right)^2 r^3 + \sigma_e^2 r\right) \tag{3.9}$$

The number of mutants that a node i receives is $k_i I/U$. When we come to normalise this, we find that the total population is

$$P \approx \frac{M}{U^3}\left(\langle k\rangle \hat{\lambda}^2 + \langle k\left(k - \hat{\lambda}\right)^2\rangle r^3 + \langle k\rangle \hat{\sigma}_e^2 r\right) \tag{3.10}$$

The second two terms in the parentheses are much smaller than the first and so, for mathematical expediency, we ignore them (In the numerical analysis of the following section, we use the symbol $\tilde{\lambda}_1^3$ to refer to the approximation which results from including all three terms in the parentheses). This results in

$$f_i\left(\lambda_1^3\right) \approx \frac{1}{N\langle k^2\rangle}\left(k_i\hat{\lambda} + k_i\left(k_i - \hat{\lambda}\right)r + k_i\left(k_i - \hat{\lambda}\right)r^2 + \frac{k_i\left(k_i - \hat{\lambda}\right)^2 r^3}{\hat{\lambda}} + \frac{k_i\sigma_e^2 r}{\hat{\lambda}}\right) \tag{3.11}$$

As previously, we multiply by each node's robustness (k_i) and sum over all nodes to arrive at the approximation for the eigenvalue (population average robustness).

$$\lambda_1^3 \approx \hat{\lambda} + \frac{2r\sigma_e^2}{\hat{\lambda}} + \frac{r^2\sigma_e^2}{\hat{\lambda}} + \frac{r^3\left(\langle k^4\rangle - 2\hat{\lambda}\langle k^3\rangle + \hat{\lambda}^2\langle k^2\rangle\right)}{\langle k^2\rangle\hat{\lambda}} \tag{3.12}$$

3.3 Erdős-Renyi Networks

In order to ascertain the accuracy of the approximations for the equilibrium distribution of populations in homogeneous networks, derived in the previous section (§ 3.2), Erdős-Renyi (Erdős & Renyi, 1959) networks were generated. The generation of networks conforming to this model is performed by instantiating N vertices and E edges. Each end of each edge is connected to a node, chosen randomly. As such, the Erdős-Renyi model is a model of maximally random networks. Figure 3.1 shows a diagram of an Erdős-Renyi network along with the population distribution over it.

In order to study the bulk properties of this model, 1100 large networks were generated. All generated networks contained $N = 5000$ vertices and either $E = 35000$ or $E = 70000$ edges, providing average degrees of $\langle k\rangle = 14$ and $\langle k\rangle = 28$, respectively. The lower value was chosen as it was found that this was the lowest value which could be used for which all generated networks were connected. As specified in § 2.5, our analysis assumes connected networks.

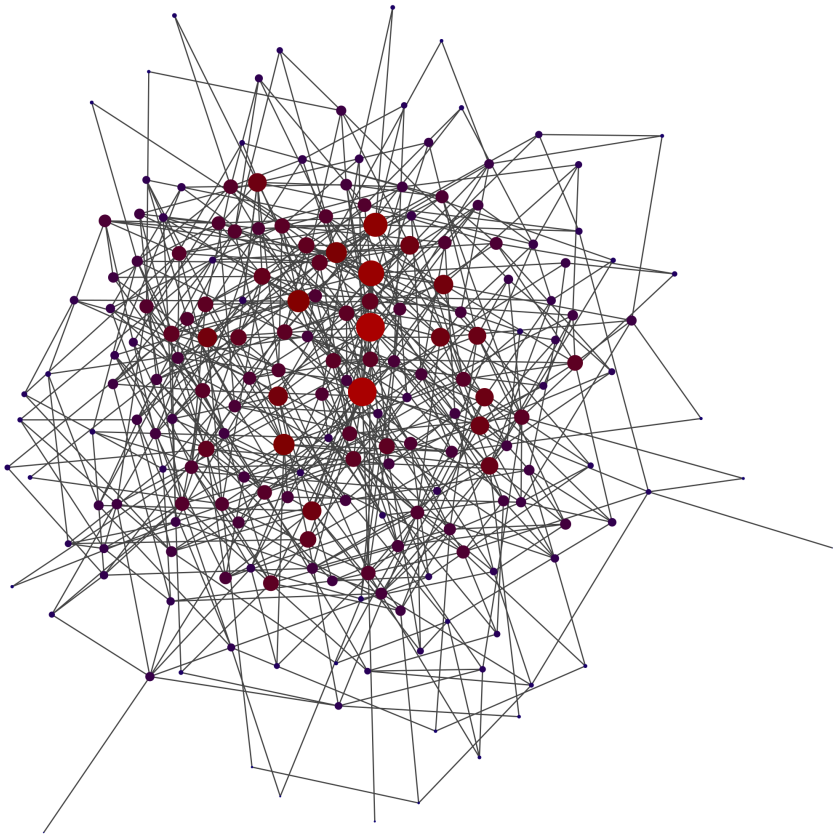
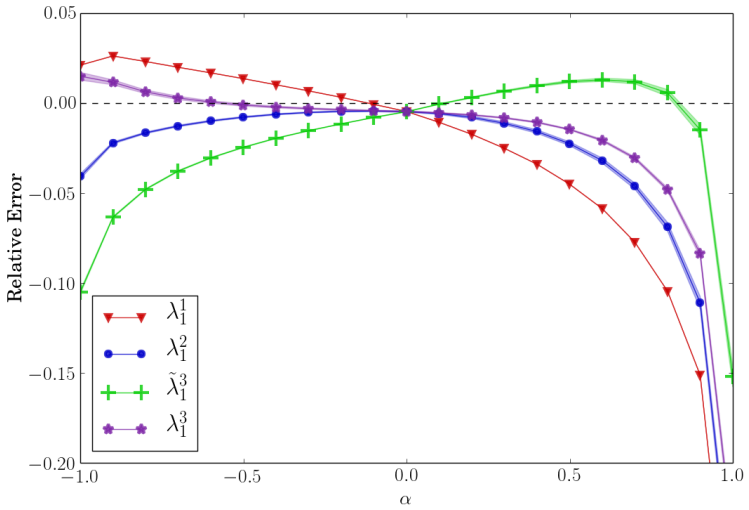


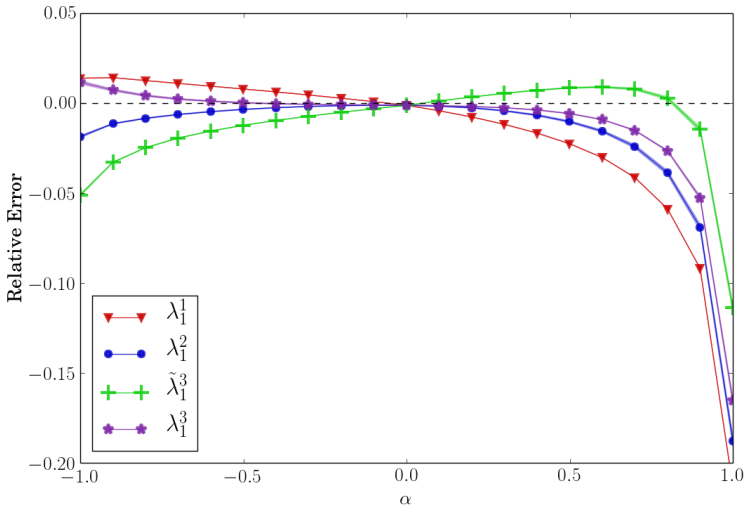
Figure 3.1: An Erdős-Renyi network with $N = 200$ nodes and $E = 600$ edges. The node size is proportional to the proportion of the population that is located on it. Moreover, nodes with a higher population concentration are more red and nodes with a lower concentration are more blue. The layout was determined the Fruchterman-Reingold force directed layout (Fruchterman & Reingold, 1991)

In order to study the accuracy of the derived approximations at various levels of network assortativity, the XBS rewiring algorithm (§ 2.5) was employed. The algorithm was run in both the assortative and disassortative variants for the 11 values of the parameter α between 0 and 1, inclusive, at spacings of 0.1 . for each of these 21 values (the zero value is equivalent in both modes of the algorithm), 100 Erdős-Renyi networks were generated and subsequently rewired.

The relative errors of the various average population robustness (eigenvalue) predictions are shown in figure 3.2.



(a) $\langle k \rangle = 7$



(b) $\langle k \rangle = 14$

Figure 3.2: The relative error of the population average robustness (principal eigenvalue) approximations derived in § 3.2, tested on Erdős-Renyi networks rewired using the XBS algorithm to exhibit positive or negative degree assortativity. Although the parameter α of this algorithm is always positive, the negative values in this figure signify that the algorithm was being used in its disassortative mode. The value $\alpha = -1$ produced an average assortativity coefficient of $r = -0.97$. Similarly, $\alpha = 1$ produced an average assortativity coefficient of $r = 0.97$

3.4 Discussion

fig. 3.2 demonstrates that the derived expressions for average population robustness are fairly accurate. As would be expected, the approximations based on a mean-field assumption at a greater distance from a given node turned out to be more accurate. It is interesting to note that, for highly assortative networks, the approximation $\tilde{\lambda}_1^3$ was more accurate than λ_1^3. This can be explained by the fact that λ_1^3 is smaller than λ_1, and that the approximation of the denominator used for $\tilde{\lambda}_1^3$ reduced its size (for positive values of r).

The approximations λ_1^2 and λ_1^3 demonstrate that our intuition (§ 3.1) was broadly correct. That is, absent assortativity, the population average robustness is equal to that provided for by the mutation sampling bias due to the friendship paradox. It rises above or below this value in the presence of positive or negative degree assortativity.

CHAPTER 4

HETEROGENEOUS NETWORKS

4.1 Localisation

In the context of graph spectra, localization refers to the phenomenon whereby the normalisation weight of an eigenvector ($\sum f_i^2(\lambda)$, where λ is the eigenvalue and $\boldsymbol{f}(\lambda)$ is the eigenvector) is concentrated on a small number of nodes that does not scale with the size of the network (Pastor-Satorras & Castellano, 2016). Some authors have suggested using the inverse participation ratio $Y(\lambda)$

$$Y(\lambda) = \sum_{i=1}^{N} f_i^4(\lambda) \tag{4.1}$$

as a quantitative measure of localization where, in this case, $\boldsymbol{f}(\lambda)$ is the normalised eigenvector. If, in the limit $N \to \infty$, $Y(\lambda) \sim 1$ then the state is localized. On the other hand, if $Y(\lambda) \to 0$ then the state is delocalized. There are a number of results relating aspects of network topology to localization. Chung et al. (2003) showed that the principal eigenvalue, for a random graph model characterised by a given degree distribution, is given by.

$$\lambda_1 = \begin{cases} \hat{\lambda}, & \hat{\lambda} > \sqrt{k_{max}} \log N \\ \sqrt{k_{max}}, & \sqrt{k_{max}} > \hat{\lambda} \log^2 N \end{cases} \tag{4.2}$$

where $\hat{\lambda} = \langle k^2 \rangle / \langle k \rangle$ ($\langle k \rangle$ being the average degree and $\langle k^2 \rangle$ being the mean of the squares of the degrees). $\lambda_1 = \hat{\lambda}$ corresponds to the delocalized state and $\lambda_1 = \sqrt{k_{max}}$ corresponds to the localized state.

Goltsev et al. (2012) showed that, for unassortative scale-free networks with degree distribution $P(k) \sim k^{-\gamma}$, the principal eigenstate is localized for $\gamma > \frac{5}{2}$ and delocalized otherwise. The principal eigenvalue is given by $\sqrt{k_{max}}$ and $\hat{\lambda}$ for the localized and delocalized states, respectively.

Martin et al. (2014) demonstrated that for a hub connected to an Erdős-Renyi network, localization occurs when $\sqrt{k_{max}} > \langle q \rangle$ where $\langle q \rangle$ is the average degree of the original Erdős-Renyi network, without the hub. Furthermore, they showed that the eigenvector component on the hub, f_h is given by.

$$f_h = \sqrt{\frac{k_{max} - 2\langle q \rangle}{2k_{max} - 2\langle q \rangle}} \tag{4.3}$$

Where the average of the components neighbouring the hub, $\langle f_n \rangle$ is given by.

$$\langle f_n \rangle = \frac{f_h}{\sqrt{k_{max} - \langle q \rangle}} \tag{4.4}$$

and the average of all non-hub components $\langle f_j \rangle$ is.

$$\langle f_j \rangle = \frac{1}{N-1} \frac{f_h}{\sqrt{k_{max} - \langle q \rangle}} \tag{4.5}$$

Pastor-Satorras & Castellano (2016) demonstrated a different form of localisation which does not result in the concentration of the eigenvector on a hub. Instead, the eigenvector localises on the maximum K-core. The maximum K-core is formed of the nodes with the maximum K-index in the K-core decomposition (Seidman, 1983) of the network. The K-core decomposition is an iterative process whereby, in each iteration, all nodes with degree one and their associated edge are removed from the network. This process continues until either there are no nodes left in the network, or until a point is reached where there are no remaining nodes of degree one. The maximum K-core will, therefore, be the nodes remaining at the end of this process, or the last node to be removed.

The localisation of the principle eigenvector in graphs can be contrasted with other forms of localisation in diffusive systems, most notably Anderson localisation (Anderson, 1958).

4.2 Definition of Localisation

As mentioned by Pastor-Satorras & Castellano (2016), there does not exist a non-arbitrary definition for the localisation of the eigenvector in single network instances. However, for the purposes of this thesis, it will be useful to define some

threshold separating localised and delocalised population distributions. We define two such thresholds and use them in different instances.

Localisation is somewhat easy to define in the case that the population has concentrated around a single hub. Here, we choose to say that if 90% or more of the population is distributed on the hub and its immediate neighbours, then the population is localised.

Trying to describe a population that is highly concentrated, but not on a hub, is slightly more challenging. Given that localisation for classes of networks is defined in terms of the inverse participation ratio, this would be natural metric with which to define localisation. However, as Pastor-Satorras & Castellano (2016) point out, in the delocalised case we expect the inverse participation ratio to be proportional to N^{-1}. On the other hand, in the localised case, we expect it to be proportional to $N^{-\beta}$, where $\beta < 1$. This dependence on the network size N, is unfortunate for our purposes, as we desire a single threshold which applies to networks of all sizes.

In order to reduce the impact of the network size N on our threshold, we define the *relative inverse participation ratio*. This is, simply, the ratio of the inverse participation ratio of the network's principal eigenvector to what the inverse participation ratio would be if the eigenvector was distributed uniformly over the network's nodes. If the eigenvector is distributed uniformly, then the inverse participation ratio is $1/N$. This implies that the relative inverse participation ratio can be easily calculated by multiplying the inverse participation ratio by N.

We choose to define localisation as occurring when the relative inverse participation ratio is greater than 30. In preliminary testing it was found that this value corresponded with the author's intuition of localisation. For comparison, Pastor-Satorras & Castellano (2016) reported on a number of real-world networks exhibiting localisation. The lowest relative inverse participation ratio of these networks was 46.7 (The HEP network, table 1 of (Pastor-Satorras & Castellano, 2016)).

4.3 Network Models

The remainder of this section studies the behaviour of the principal eigenvector on heterogeneous networks. The first model studied (§ 4.4) has already been

well described by the above results. However, analysing it is still valuable, as previous work on similar models reported on the *weight* of the eigenvector confined to a given region, whereas, for evolution, we are interested in the size of the eigenvector components, normalised by the sum of the vector componenets (l_1 norm), in a given region

The first three models studied in the following sections are not embedded in Hamming space. This is due to the difficulties of studying such networks (see chapter 8 for a more detailed discussion). However, they do exhibit some of the features that we expect to see in real world networks. It is interesting, for instance to compare the networks diagram displayed in fig. 4.6 (a Barábasi-Albert network) with that displayed in fig. 7.1 (a neutral network of H1N1 influenza haemagglutinin).

All of this analysis was conducted using the `Python` package `igraph` (Csardi & Nepusz, 2006). The calculation of the eigenvalues and eigenvectors of the adjacency matrices of graphs in `igraph` is performed using the `FORTRAN 77` package `ARPACK` (Lehoucq et al., 1998). `ARPACK` implements the *implicitly restarted Arnoldi method* (Lehoucq & Sorensen, 1996) to find the eigenvalues and eigenvectors of matrices. `igraph`'s default parameters for `ARPACK` were used.

A number of the simulations presented below report on the relative error of the principal eigenvalue estimations. This error was calculated according to:

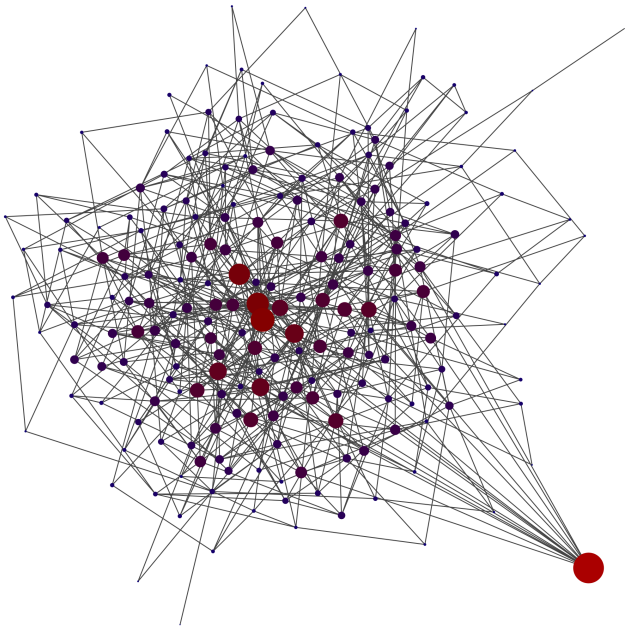
$$\text{relative error} = \frac{\lambda_1 - \lambda_1^n}{\lambda_1} \tag{4.6}$$

where λ_1^n is the n^{th} approximation of the principal eigenvalue.

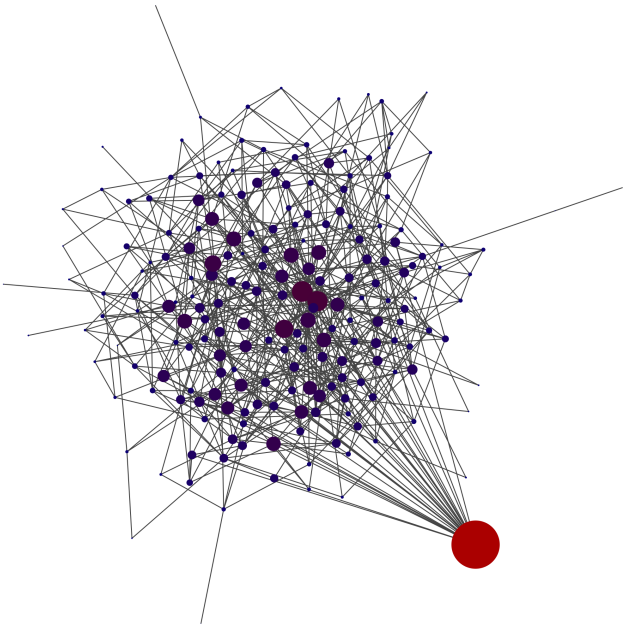
4.4 Erdős-Renyi Networks With Hubs

In order to observe the impact of localization on the population's distribution over the network, following the lead of Martin et. al. (Martin et al., 2014), random networks in which a hub was connected to an Erdős-Renyi network were generated. This was performed by first generating an Erdős-Renyi network, as in § 3.3. Subsequently, a hub of degree γ was added to the network, by connecting it to γ nodes in the original network, chosen randomly. Figure 4.1 shows a diagram of two such networks along with the population distribution over them.

In order to study the bulk properties of this network model, 7500 large instances were generated. The original Erdős-Renyi networks contained $N = 5000$

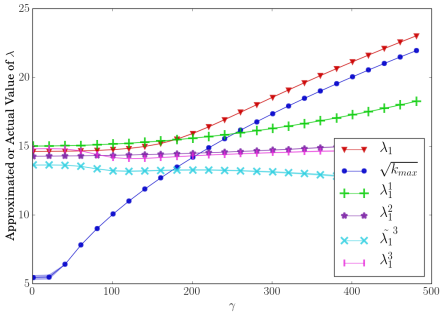


(a) $\gamma = 15$

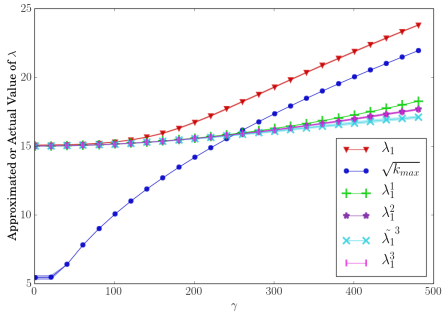


(b) $\gamma = 30$

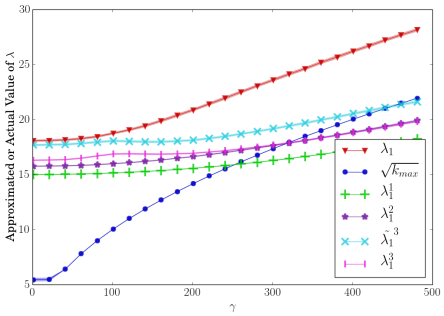
Figure 4.1: Erdős-Renyi networks with $N = 200$ nodes and $E = 600$ edges, with connected hubs of degree γ, as described in § 4.4. The node size is proportional to the proportion of the population that is located on it. Moreover, nodes with a higher population concentration are more red and nodes with a lower concentration are more blue. The layout was determined the Fruchterman-Reingold force directed layout (Fruchterman & Reingold, 1991), with the hub nodes manually positioned outside the networks after the layout algorithm was run.



(a) $\alpha = -1$



(b) $\alpha = 0$

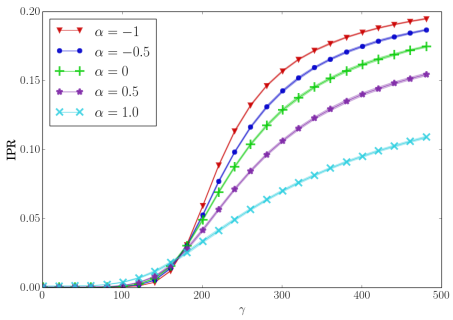


(c) $\alpha = 1$

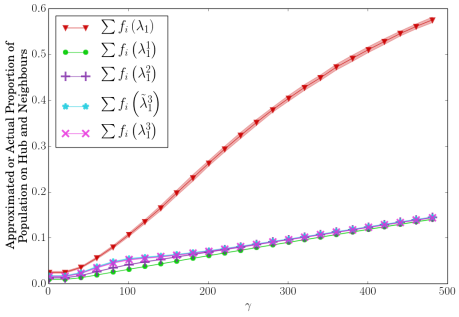
Figure 4.2: The approximated or actual population average robustness (principal eigenvalue of the network's adjacency matrix) where a hub of degree γ was connected to an Erdős-Renyi network. These networks were rewired using the XBS algorithm to exhibit positive or negative degree assortativity. These three plots show the population average robustness for three different values of the rewiring parameter α

nodes and $E = 35000$ edges, providing an average degree of $\langle k \rangle = 14$. Rewiring (see § 2.5) was performed for for the five values of $\alpha = [-1, -0.5, 0, 0.5, 1]$. The 15 values of γ in the range $[1, 481]$, at spacings of 20 were used. 100 networks were generated for each of the 75 combinations of α and γ.

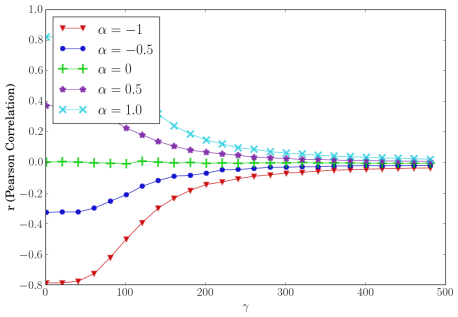
Figure 4.2 shows the behaviour of the population average robustness, as well as the approximations derived in § 3.2. It also shows the approximation $\sqrt{k_{max}}$, as this is an approximation for the principal eigenvalue under localization derived by other authors (Chung et al., 2003; Goltsev et al., 2012; Martin et al., 2014). We find that, for small values of γ, the approximations derived in § 3.2 are very accurate. However, as γ is increased, the accuracy of these approximations deteriorated, and $\sqrt{k_{max}}$ became a better estimator of the population's average robustness.



(a) Inverse participation ratio



(b) Proportion of the population on the hub and its neighbours



(c) Network assortativity coefficient

Figure 4.3: The predicted and actual inverse participation ratio, approximated and actual proportion of population on the hub node and its neighbours and the average network assortativity where a hub of degree γ was connected to an Erdős-Renyi network. These networks were rewired using the XBS algorithm to exhibit positive or negative degree assortativity.

Figure 4.3a shows the inverse participation ratio of the principal eigenvector for various values of α and γ. The inverse participation ratio is close to 0 for $\gamma < 200$, it increases sharply in the range $200 < \gamma < 300$ and has a flatter gradient for $\gamma > 300$. This is in accordance with the results of Martin et al. (2014), where it is predicted that localization occurs when $\gamma = k_{max} > \langle k \rangle (\langle k \rangle + 1)$. We further observe that assortativity decreases the severity of the localization and, conversely, disassortativity increases it. In disassortative networks, the neighbours of the hub will be of lower degree and so will connect to fewer nodes other than the hub itself. This increases the chance that their offspring will mutate back onto the hub. The opposite occurs in assortative networks.

Figure 4.3b shows the approximated and actual proportion of the population on the hub and its neighbors for $\alpha = 1$. This large value of α was chosen as a large degree of assortativity (or disassortativity) is required for a difference

to be observed between the various approximations as the higher order terms contain powers of the assortativity coefficient r. For large values of γ most of the population (over 50%) is concentrated on the hub and its neighbours. This is in contrast with the approximations for degree-homogeneous networks, which predict a lower proportion of the population to be concentrated around the hub. It is interesting to note that, up to about $\gamma = 100$, the three-hop approximations are able to approximate the population distribution fairly well. This is due to the fact that, as per the arguments of Martin et al. (2014), we only expect localization to occur when $k_{max} > \langle k \rangle (\langle k \rangle + 1)$.

Figure fig. 4.3c shows the resulting r value in the networks for the various values of α tested. High values of γ result in the networks being relatively unassortative. In networks with a large hub, most of the variance in the degrees at the ends of edges will be due to the hub and cannot be made to depend on the degree of the node at the other end of the edge.

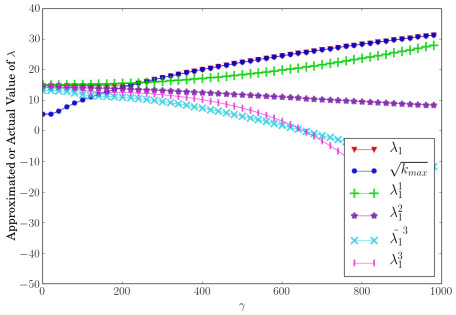
It is worth pointing out that the localisation observed on these networks is only partial, as it does not fully meet the definition of localisation given in § 4.2.

4.5 Erdős-Renyi Networks With Separated Hubs

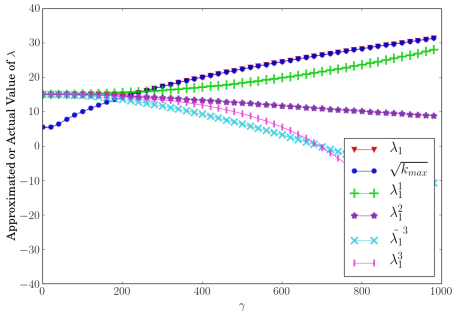
The preceding section studied the localization of the population on a hub, where the hub is well connected to the rest of the network. Although it could be argued that, in random networks, this type of topology is more likely, it is worth studying the implications of hubs that are poorly connected to the rest of the network. Moreover, prototype protein sequences form neutral hubs around themselves (§ 6.2). As the number of sequences folding to the given structure decreases with distance from the prototype sequence, we should not expect the hubs to be well connected with the rest of the neutral network.

This section studies maximally disconnected hubs. Erdős-Renyi networks were generated, as in § 3.3, Following this, the networks were rewired using the XBS algorithm (§ 2.5) to exhibit positive or negative degree assortativity. Once the rewiring was complete, a star network of size $\gamma + 1$ was added to the network, by connecting one of the spoke nodes to a node of the Erdős-Renyi network, chosen randomly. Figure 1.1 shows a diagram of two such networks along with the population distribution over them.

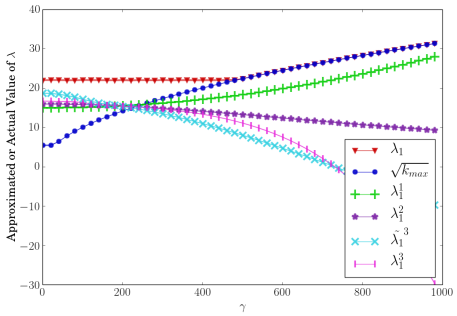
In order to study the bulk properties of this network model, 7500 large in-



(a) $\alpha = -1$



(b) $\alpha = 0$

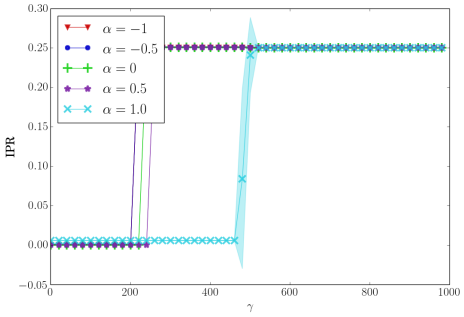


(c) $\alpha = 1$

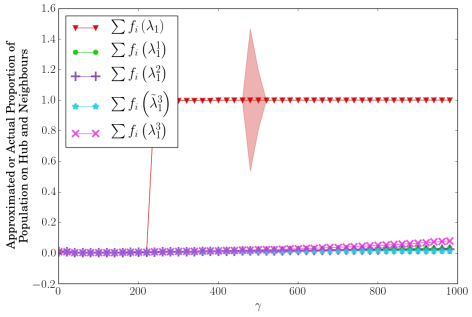
Figure 4.4: The approximated or actual population average robustness (principal eigenvalue of the network's adjacency matrix) where a star network of size $\gamma + 1$ was connected to an Erdős-Renyi network via one of the star's peripheral nodes. These networks were rewired using the XBS algorithm to exhibit positive or negative degree assortativity. The three plots show the population average robustness for three different values of the rewiring parameter α

stances were generated. The original Erdős-Renyi networks contained $N = 5000$ nodes and $E = 35000$ edges, providing an average degree of $\langle k \rangle = 14$. Rewiring was performed for for the five values of $\alpha = [-1, -0.5, 0, 0.5, 1]$. The 15 values of γ in the range $[1, 481]$, at spacings of 20 were used. 100 networks were generated for each of the 75 combinations of α and γ.

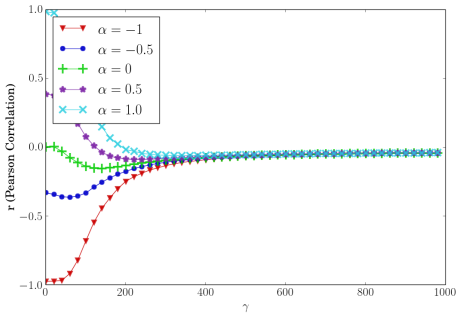
Figure 4.4 shows the population average robustness, the approximations of this robustness derived in § 3.2, as well as $\sqrt{k_{max}}$. As in the preceding section, the approximations derived in § 3.2 are accurate for small values of γ and lose accuracy as γ increases and localization occurs. In this regime, $\sqrt{k_{max}}$ very closely approximates the population's average robustness. $\sqrt{k_{max}}$ Is the exact solution of the principal eigenvalue of a star graph (Reeves et al., 2016). It is interesting to note that connecting the star to a sufficiently low degree Erdős-Renyi network



(a) Inverse participation ratio



(b) Proportion of the population on the hub and its neighbours



(c) Network assortativity coefficient

Figure 4.5: The predicted and actual inverse participation ratio, approximated and actual proportion of population on the hub node and its neighbours and the average network assortativity where a star network of size $\gamma + 1$ was connected to an Erdős-Renyi network via one of the star's peripheral nodes. These networks were rewired using the XBS algorithm to exhibit positive or negative degree assortativity.

does not substantially alter this value. When a hub was fully connected to an Erdős-Renyi network (as in § 4.5) it was found that the eigenvalue approximations based on equal population distribution at a greater distance from a given node were more accurate for large values of γ (§ 4.4, and figure 4.2). The opposite was found to be true for separated hubs, that is for larger values of γ, λ_1^n was more accurate for smaller values of n. This is due to the fact that the approximations based on a greater number of hops involve terms proportional to r. For large values of γ, the calculation of r is dominated by the star, where a very high degree node is connected to many nodes of degree one. This causes the network to be disassortative, regardless of the value of the rewiring coefficient α (see figure 4.5c). This disassortativity reduces the approximation of the eigenvalue, whereas the actual value of the eigenvalue is increasing as $\sqrt{k_{max}} = \sqrt{\gamma}$.

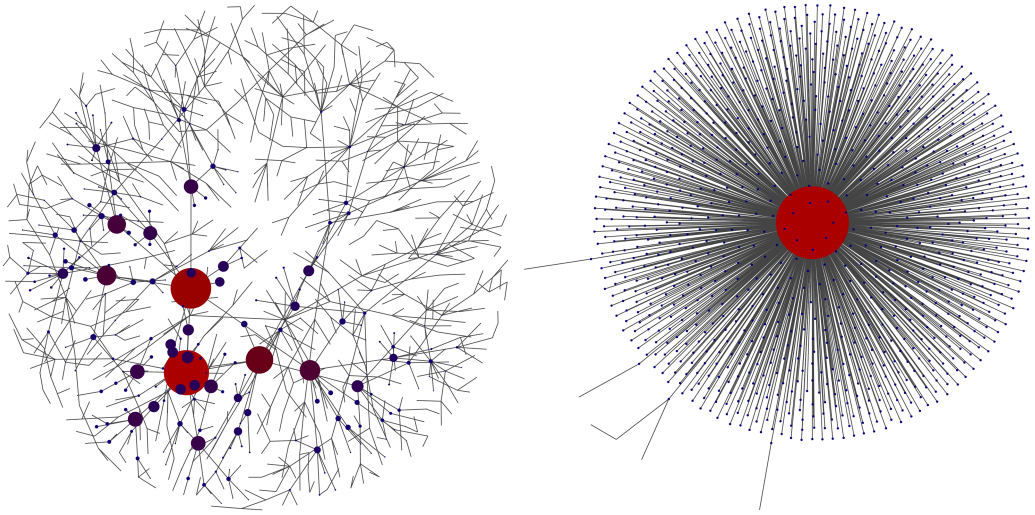
Figure 4.5a shows the average inverse participation ratio of the principal eigenvector of the networks studied for all values of α and γ. The transition to the localized regime is much sharper than for fully connected hubs (figure 4.5a). Figure 4.5b shows the proportion of the population on the hub and its neighbours in the unsassortative case. In this case, almost the entire population is located on the star network after localization.

In the highly assortative case ($\alpha = 1$) localization only occurs around $\gamma = 500$ (see figures 4.5a and 4.4c). This is substantially higher than for highly connected hubs, where localization occurs around $\gamma = 200$ (figure 4.2c). This is due to the fact that, for the maximally connected hubs, assortative rewiring occurred after the hub was connected to the network, whereas for minimally connected hubs it occurred before connection. This resulted in the hub of the maximally connected networks being connected to the highest degree nodes of the original Erdős-Renyi network. On the other hand, for the minimally connected hubs, assortative rewiring results in a region of highly connected higher degree nodes, separate from the hub, which competes with it for the location of the population.

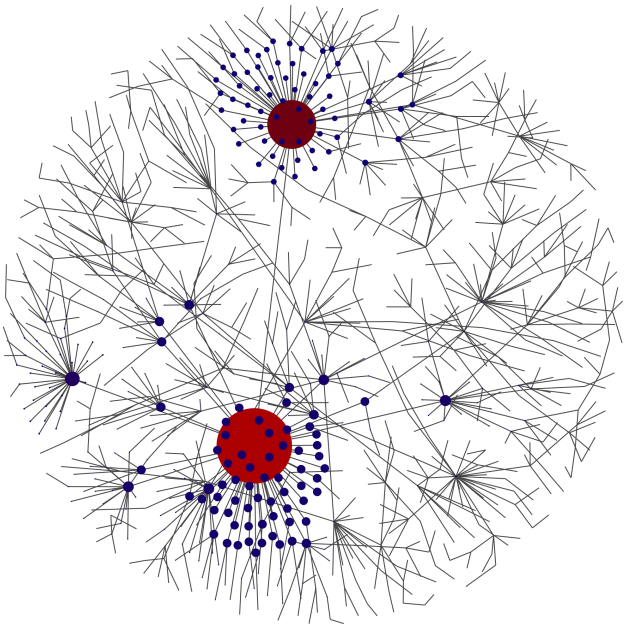
4.6 Barábasi-Albert Preferential Attachment

In the two preceding sections, hubs were manually attached to existing networks. Moreover, the networks contained only a single hub. It is, therefore, valuable to interrogate the population distribution in network models which naturally contain hubs, and which might contain multiple hubs connected to one another. Scale-free networks (Barabási, 2016) contain multiple hub nodes. These are networks with a power-law degree distribution, that is $p(k) \sim k^{-\gamma}$ for some value of the parameter γ (usually, $2 \leq \gamma \leq 3$). However, as mentioned in § 2.5, the scope of this thesis is limited to connected networks. In general, most scale-free networks are not connected (Cohen et al., 2003). Thus, random network generation algorithms which aim to uniformly sample the space of scale-free networks do not generate connected networks.

Fortunately, the popular Barábasi-Albert preferential attachment model (Barabási & Albert, 1999) does generate connected networks which have a power-law degree distribution, although it does not uniformly sample the space of scale-free networks. Moreover, this algorithm is able to generate networks which contain hubs, but have degree distributions steeper or shallower than a power-law.

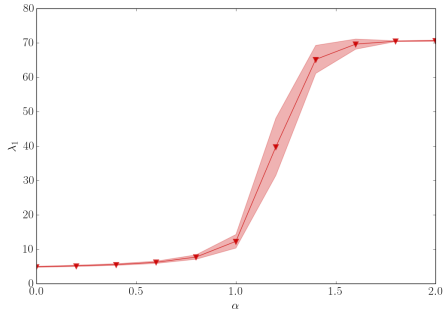


(a) $\alpha = 0$ (b) $\alpha = 2$

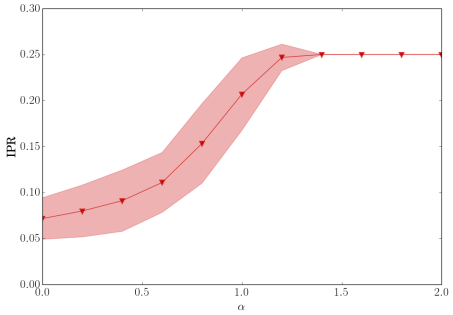


(c) $\alpha = 1$

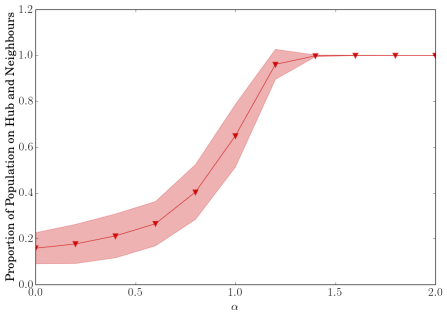
Figure 4.6: Barábasi-Albert preferential attachment networks with $N = 200$ nodes, and three different values of the attachment parameter α. The node size is proportional to the proportion of the population that is located on it. Moreover, nodes with a higher population concentration are more red and nodes with a lower concentration are more blue. The layout was determined by the Fruchterman-Reingold force directed layout (Fruchterman & Reingold, 1991).



(a) Population average robustness (principal eigenvalue)



(b) Inverse participation ratio



(c) Proportion of the population on the hub and its neighbours

Figure 4.7: The average population average robustness, the inverse participation ratio, and the proportion of population on the hub node and its neighbours in Barábasi-Albert preferential attachment networks. The shaded region shows the standard deviation.

The algorithm starts with a single node and then progresses by iteratively adding nodes to the network. Each time a node is added it is connected to existing nodes in the network by m edges. In this work, $m = 1$, as it was found that higher values of m produce networks which bear little resemblance to the expected topology of neutral networks. When new nodes are connected to the network they are connected to a given node with probability $p(k) = k^{\alpha}$. In the case $\alpha = 1$, the degree distribution is a power law, with exponent $\gamma = 3$. If $\alpha < 1$, then the degree distribution falls off faster and so the occurrence of high degree nodes is reduced. The opposite is true in the case $\alpha > 1$. More specifically, if $\alpha = 0$:

$$p(k) \sim \frac{e}{m} \exp\left(-\frac{k}{m}\right) \tag{4.7}$$

if $0 < \alpha < 1$, then:

$$p(k) \sim k^{-\alpha} \exp\left(\frac{-2\mu(\alpha)}{\langle k \rangle (1-\alpha)} k^{1-\alpha}\right) \tag{4.8}$$

where $\mu(\alpha)$ depends only weakly on α (Barabási, 2016). For $\alpha > 1$ a stationary degree distribution (independent of the number of iterations) does not exist. For further details, the reader is referred to the excellent, recently published, book by Albert-László Barábasi himself (Barabási, 2016).

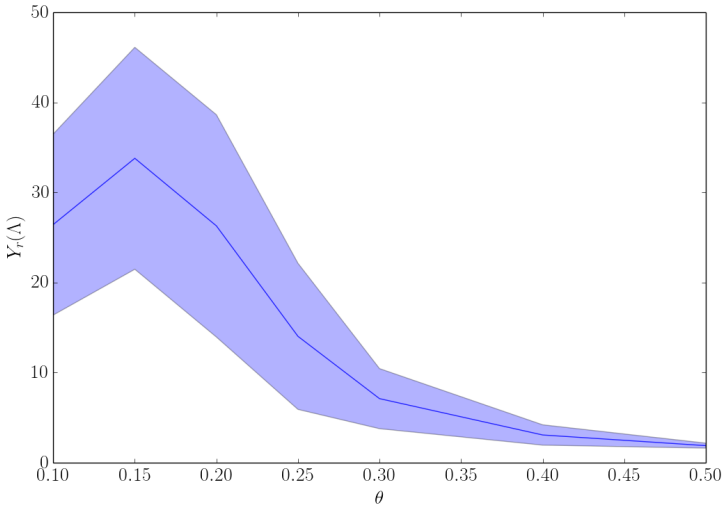
Figure 4.6 shows diagrams of networks generated according to this model for the three values of $\alpha = [0, 1, 2]$.

In order to examine localisation in this network model, 1000 networks were generated with $N = 5000$ nodes for the values $\alpha = [0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0]$. fig. 4.7 plots the principal eigenvalue, inverse participation ratio and the proportion of the population on the maximum degree node (hub) and its neighbours. We find that the population is highly localised for $\alpha > 1.2$.

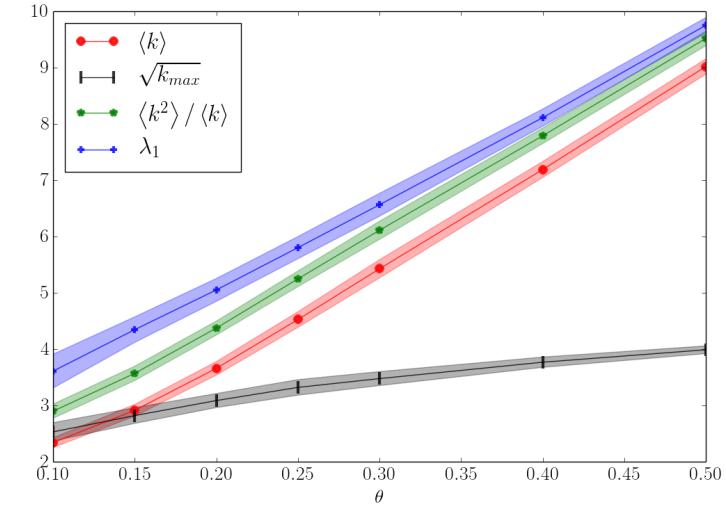
4.7 Poorly Connected Random Subgraphs of Hypercubes

The realisable topology of neutral networks is constrained by the fact that the genotypes are encoded by strings of characters and that edges can only be placed between vertices whose corresponding genotypes differ by a single character. That is, an edge can only connect nodes whose corresponding genotype sequences are a Hamming distance of one apart. Here, we analyse how this constraint influences the neutral evolution of populations. In particular, we are interested in whether the mean-field approximation holds and also the types of localization behaviour which can be observed.

Random subgraphs of hypercubes (or n-cubes) are well studied (Reidys et al., 1997; Reidys, 2009). In these models, neutral networks are created by including each node in the neutral network with a probability θ (the symbol λ is usually used for this probability, however, we use θ to avoid confusion with the principal eigenvalue, which we denote with λ_1). Once the nodes have been assigned as being on or off the network, the connected components of the neutral network can be extracted. In the following, we study only the largest connected component of the networks, as we are interested in the exploratory behaviour of populations on networks which extend over large parts of sequence space.

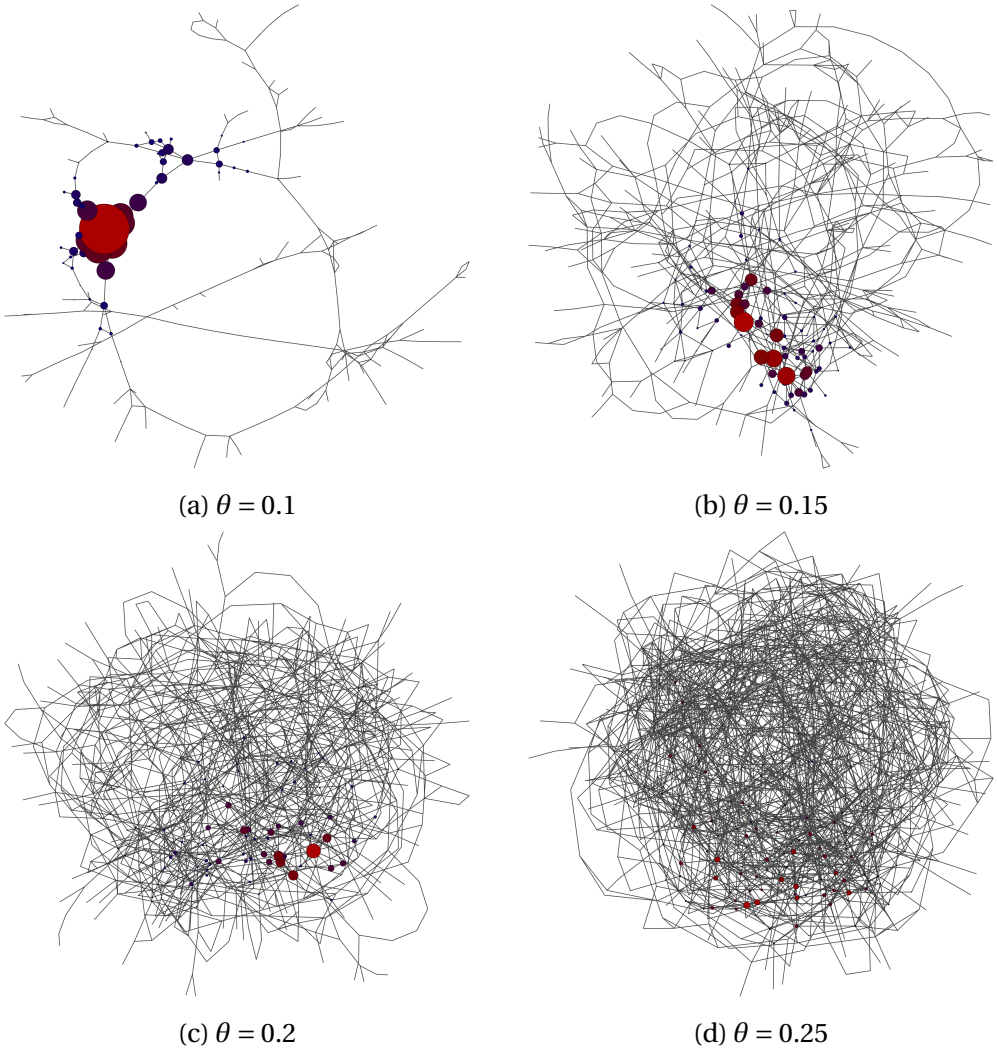


(a) Relative inverse participation ratio ($Y_r(\Lambda)$)



(b) Principal eigenvalue and related properties

Figure 4.8: Properties relating to the principal eigenvalue and the distribution of the principal eigenvector over the largest connected component of random subgraphs of an n-cube.



(a) $\theta = 0.1$ (b) $\theta = 0.15$

(c) $\theta = 0.2$ (d) $\theta = 0.25$

Figure 4.9: Network diagrams of the largest connected component of random subgraphs of an n-cube. The size of the nodes is proportional to the proportion of the principal eigenvector which is located on them. Moreover, nodes with a higher population concentration are more red and nodes with a lower concentration are more blue.

We hypothesize that, for sufficiently low θ, the connectivity of the network will be low enough that the population will be confined to areas of it, rather than spread evenly. In order to test this hypothesis, we generated random subgraphs of the hypercube formed by using strings of length $L = 6$ over an alphabet consisting of $A = 4$ distinct characters. The values of θ from the set $[0.1, 0.15, 0.2, 0.25, 0.3, 0.4]$ were used. Fewer larger values were chosen as preliminary experiments showed that, for large values of θ, the resulting networks were substantially larger, making analysis computationally expensive. Furthermore, figure fig. 4.8 shows that the behaviour of the population is less interesting for larger values of θ. For each value of θ, 100 networks were instantiated. Various properties relating to the principal eigenvalue and eigenvector were measured. These properties are plotted in fig. 4.8. fig. 4.9 shows diagrams of representative networks, with the population distribution displayed through vertex size and colour.

fig. 4.9 demonstrates that, at least for the selected representative networks, the population is highly concentrated on a small number of nodes for small values of θ. However, for larger values of θ it is distributed over a substantially larger number of nodes. This behaviour is confirmed in fig. 4.8a where we see that the relative inverse participation ratio $Y_r(\Lambda)$ is high enough to justify localization for low values of θ. However, it drops rapidly for increasing values of θ. It is interesting to note that $Y_r(\Lambda)$ increases between $\theta = 0.1$ and $\theta = 0.15$. Further investigation revealed that the networks produced for $\theta = 0.1$ were very small (this can be observed in fig. 4.9). We suspect that the small size of the networks prevents $Y_r(\Lambda)$ from being very large, as the minimum number of nodes a population has to occupy is probably close to the size of the networks.

fig. 4.8b shows that, for large values of θ, λ_1 is well approximated by the mean-field approximation $\hat{\lambda} = \langle k^2 \rangle / \langle k \rangle$. Furthermore, $\sqrt{k_{max}}$ is substantially smaller than $\hat{\lambda}$ for all values of θ. This emphasizes that this is a different form of localisation to concentration on hubs.

We also wanted to interrogate whether this mode of localisation is dissimilar from localization on a K-core (see § 4.1) (Pastor-Satorras & Castellano, 2016). In order to do this, we generated a further 20 networks each for $\theta = 0.15$ and $\theta = 0.2$ and recorded the proportion of the population residing on the maximum K-core. For $\theta = 0.15$ this value varied between 0.16 and 0.54, with a mean of 0.35 and a standard deviation of 0.12. For $\theta = 0.2$ it varied between 0.01 and 0.88, with a mean of 0.7 and a standard deviation of 0.23. The existence of networks

in which the eigenvector is so weakly concentrated on the maximum K-core indicates that, at least in some cases, the mode of localization is slightly different to that described by Pastor-Satorras & Castellano (2016).

We were interested as to whether weak connectivity between parts of the network would have a similar effect on the eigenvectors of graphs that are not embedded in Hamming space. To this end, we generated pairs of Erdős-Renyi (Erdős & Renyi, 1959) networks, each with $|V| = 100$ vertices and $|E| = 200$ edges. These pairs of networks were connected by a single edge, connecting two randomly chosen vertices. Although, in some instances, the eigenvector was spread out evenly over the two original networks, in others, it was almost completely concentrated on a single network. fig. 4.10 shows a diagram of an instance where the eigenvector was heavily concentrated on a single network in the pair. Further analysis showed that, when analysed independently (without the single connecting edge), the one network had a slightly higher principal eigenvalue than the other, due to the randomness involved in the generation of the networks. The eigenvector of the combined network was concentrated on this network. This effect makes sense in terms of natural evolution, as the population is able to achieve a higher level of robustness on one of the pair of networks. Therefore, were a fraction of the population to be located on the sub-network with a lower principal eigenvalue, it would be out-competed by the fraction of the population located on the other sub-network. Moreover, the single connecting edge is insufficient to allow a large flow of mutants from the one sub-network to the other.

4.8 Hamming Balls on Random Subgraphs of Hypercubes

It is worthwhile querying whether the eigenvectors of random subgraphs of a hypercube can undergo localization onto a hub, as already discussed for networks in general. Of the analytic results for localization covered in section § 4.1, the weakest was that of Martin et al. (2014), where it was required that $\sqrt{k_{max}} > \langle q \rangle$, where $\langle q \rangle$ is the average degree of the network excluding the hub. For random subgraphs of the hupercube constructed from sequences of length L and an alphabet of size A, where a given vertex is included in the network with probability θ, if we connect a hub of maximum possible degree then this condition implies

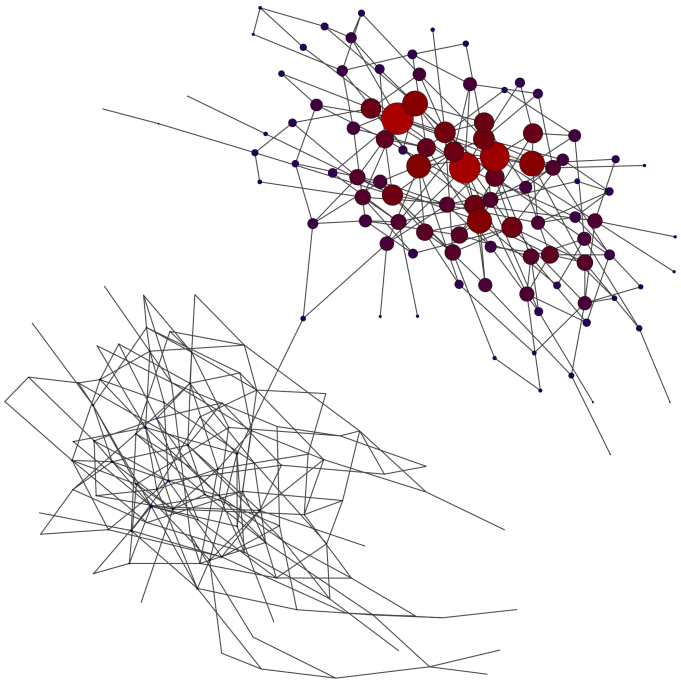
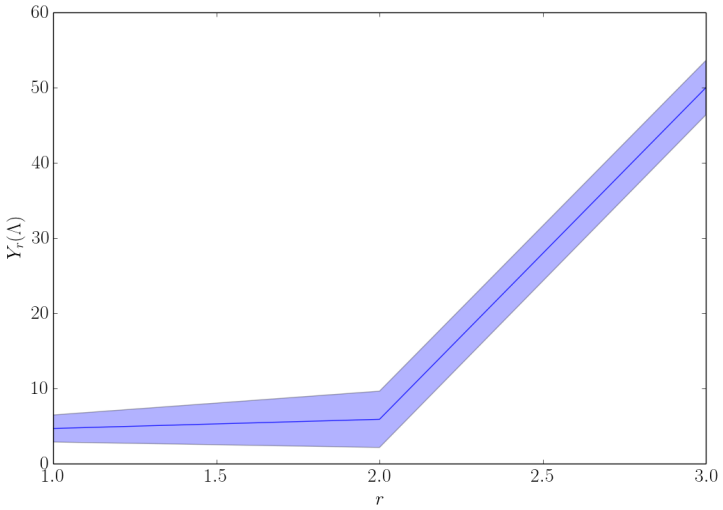


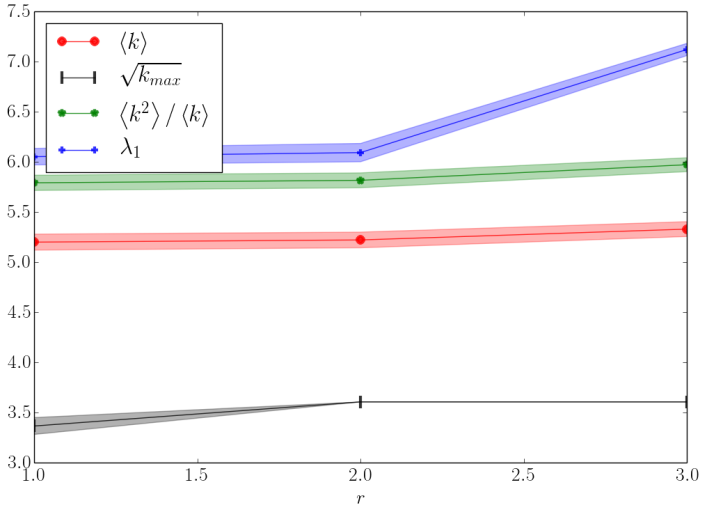
Figure 4.10: Network formed by connecting two Erdős-Renyi (Erdős & Renyi, 1959) networks, each with $|V| = 100$ vertices and $|E| = 200$ edges by a single edge. The size of the nodes is proportional to the proportion of the principal eigenvector which is located on them. Moreover, nodes with a higher population concentration are more red and nodes with a lower concentration are more blue.

$\theta < 1/\sqrt{L(A-1)}$. For the networks which we studied in the previous section, with $L = 6$ and $A = 4$, this would imply that $\theta < 0.24$. In that section we found that, for that value of θ, the eigenvector was already somewhat localized (see fig. 4.8a). Moreover, given that increasing L and A will decrease the bound on θ, similar effects are probable for the random subgraphs of larger hypercubes. As the eigenvectors of these networks are already under the influence of a certain mode of localisation, studying the effects of connecting hubs to them could lead to ambiguous results due to the multiple modes of localisation.

There is, however, a natural generalisation to a hub when considering subgraphs of hypercubes: the Hamming ball. A Hamming ball is the network composed of all nodes in the hypercube within a certain radius ρ of a specific sequence. A star of maximum possible degree in the hypercube is then a Hamming ball of radius $\rho = 1$. We would expect Hamming balls to produce populations with high average genetic robustness. Bornberg-Bauer & Chan (1999) studied them as an abstraction for the structure of protein neutral networks (see



(a) Relative inverse participation ratio ($Y_r(\Lambda)$)



(b) Principal eigenvalue and related properties

Figure 4.11: Properties relating to the principal eigenvalue and the distribution of the principal eigenvector over the largest connected component of random subgraphs of an n-cube to which a Hamming ball of radius ρ has been connected.

§ 6.2). They found that the population tended to concentrate on the inner nodes of the ball, increasing the population's average robustness. More recently, Bollobás et al. (2016) showed that, for a given number of nodes, a Hamming ball arrangement maximised the principal eigenvalue of the resulting network.

We, therefore, thought it worthwhile to investigate whether, by connecting Hamming balls to otherwise delocalised graphs, localisation could occur.

We generated random subgraphs of the hypercube formed by strings of alphabet size $A = 2$ and length $L = 13$. The smaller alphabet size was chosen as, in preliminary testing, it was found that the size of the Hamming balls increased too rapidly for larger values of A. This made the analysis too computationally expensive. The longer length was chosen to allow for large subgraphs, given the small size of the alphabet. The high value of $\theta = 0.4$ was chosen to discourage localisation behaviour of the eigenvector without the presence of the Hamming ball. To each graph was connected a Hamming ball of radius ρ. The values of ρ from the set $[0,1,2]$ were used. For each value of ρ, 100 networks were generated. Various properties relating to the principal eigenvalue and eigenvector were measured. These properties are plotted in fig. 4.11. Diagrams of the networks are not shown as it was found that the resulting networks were too large to allow for informative diagrams.

fig. 4.11a shows the relative inverse participation ratio $Y_r(\Lambda)$. Between $\rho = 2$ and $\rho = 3$ we see a sharp increase in this value, representing a localisation transition. Similarly, the principal eigenvalue exhibits a sharp increase between $\rho = 2$ and $\rho = 3$ (fig. 4.11b).

CHAPTER 5

NEUTRAL NETWORKS IN EVOLUTIONARY COMPUTING

The preceding chapters have demonstrated the manner in which the equilibrium distribution of an evolving population over a neutral network is related to structural properties of the network. The structural properties of the neutral networks which are encountered by evolving populations are, therefore, of great interest. An important question, for instance, is whether localization behaviour is a mere mathematical curiosity, or a phenomenon that occurs in evolution. This chapter will focus on man-made neutral networks - those found in EC as well as in abstract models of fitness landscapes. Relevant studies of specific instantiations of landscape models tend to have occurred within EC and so it makes sense to discuss them together. The following chapter (chapter 7) will focus on neutral networks in nature. Both chapters will pay specific attention to the structural properties which the previous sections used to describe the population distribution.

The literature on neutrality within EC is vast and, as such, no attempt will be made to review it in its entirety. The reader is referred to the review by Galván-López et al. (2011). A very substantial portion, if not the majority, of this literature concerns itself with the question of whether neutrality aids or impedes evolutionary search. Although this is a pertinent question, it is of secondary importance to this thesis. This chapter will, instead, focus on work within the EC literature which sheds light upon the *structure* of the neutral space. This will then inform us of the manner in which the above-derived results apply to the

field of EC.

5.1 Combinatorial Optimisation Problems

Combinatorial optimisation considers the optimisation of a set of discrete variables, $X = \{x_1, \cdots, x_n\}$, under a fitness function $f : X \to \mathbb{R}$ (Blum & Roli, 2003). The optimisation of genetic code in nature falls under this definition and has been studied in this context (Reidys & Stadler, 2002). Moreover, many important optimisation problems in industry are combinatorial in nature (Yu, 2013). EC is a competitive approach with which to tackle these problems (Blum & Roli, 2003).

Feasible Regions

Combinatorial optimisation problems encountered in computer science usually contain an associated set of constraints on the set of variables. For instance, in the ubiquitous traveling salesman problem, the path of a fictional salesman through a selection of cities is optimised for shortest distance (Kirkpatrick & Toulouse, 1985). The constraint on the path is that it must pass through all the cities exactly once.

These constraints divide the search space into feasible and infeasible regions. Although the feasible regions of these problems do not necessarily have a neutral fitness gradient, analysing their structure is still worthwhile, as it sheds light on the structures that certain constraints can impose on a combinatorial landscape, which could be neutral in a different context. Moreover, it is plausible that subregions of the feasible regions are neutral or nearly neutral and the topology of the feasible region constrains the space of realisable neutral networks. Finally, one could impose neutrality on the feasible region by editing the fitness function.

The structure of the feasible regions of combinatorial optimisation problems is, therefore, of interest to us. This structure is dependent on the choice of mutation operator, also known as the *neighbourhood operator* in the broader context of heuristic search applied to combinatorial problems (Reeves, 1999). When using straightforward mutation operators , the networks representing the feasible regions of many combinatorial optimisation problems are regular. That is, each node has the same degree. This includes the traveling salesman problem (Stadler & Schnabl, 1992), max-SAT (Basseur & Goëffon, 2015), flow-shop

scheduling problem (Marmion et al., 2011), graph-coloring problem (Marmion et al., 2013) and quadratic assignment problem (Merz & Freisleben, 2000). Neutral evolution over such a region would result in an evenly distributed population, with an average robustness equal to the common degree of the nodes (§ 3.2).

The author is, however, aware of two problems and associated mutation operators where the feasible region is irregular. The first of these is the job-shop scheduling problem under the *adjacent-swap* mutation operator, demonstrated by Bierwirth et al. (2004). The job-shop scheduling problem consists of a set of *jobs* and a set of *machines*. Each job has one or more *operations* which have a precedence order in which they must be completed. Each operation has a specified machine on which it must run. The task of optimisation is to find a precedence order for the operations on each machine such that there are no inconsistencies in operation precedence and such that the makespan is minimised, where the makespan is the time taken to complete all the jobs. The adjacent-swap mutation operator swaps the precedence between two jobs on a given machine and, as such, is the smallest possible change to a schedule (Bierwirth et al., 2004). Certain combinations of pairwise precedence relationships can lead to global precedence inconsistencies and this results in the distinction between feasible and infeasible regions. Bierwirth et al. (2004) randomly generated 10 000 feasible solutions of the ft10 problem instance (Fisher & Thompson, 1963). For each solution, they calculated the number of feasible neighbours out of a possible 90. It was found that there is a mild heterogeneity in the number of feasible neighbours a given node possesses. This figure varied between 77 and 89. Moreover, it was found that the makespan of the solutions was inversely correlated with their number of feasible neighbours. Further analysis showed that all 13 120 globally optimal solutions had 88 or 89 feasible neighbours.

Given the relative similarity in feasible degree of the solutions and the fact that we have no reason to believe that the networks of feasible solutions have poor connectivity, it is unlikely that, should a population evolve neutrally over this network it will localize. The fitness landscape of the job-shop problem is highly correlated (Mattfeld et al., 1999). This fact, combined with the correlation between fitness and feasible degree, implies that the feasible degree of solutions is probably correlated. This assortativity of the feasible network will result in the neutral evolution of robust populations.

The second problem with an irregular feasible region is the optimal golomb

ruler problem. This problem is of particular interest to this thesis because, as shown by Cotta & Fernandez (2005), there is enormous variation in the feasible degree of solutions. The optimal golomb ruler problem represents the challenge of finding a set of n integers (marks), $a_1 < a_2 < \ldots < a_n$, such that the difference between all pairs of integers is different and such that a_n is minimised. The constraint on the differences between the between the pairs of integers produces the structure of the feasible region and the value of a_n acts as the fitness. Cotta & Fernandez (2005) considered a mutation operator which adds a different random integer in the range $[-\epsilon, \epsilon]$ to each mark of the ruler. They found that, in the 12-mark problem variant with $\epsilon = 1$, the number of feasible neighbours of candidate solutions varied by three orders of magnitude. With $\epsilon = 4$, the number of feasible neighbours varied by seven orders of magnitude. The areas of high feasible degree were those with low fitness (that is, high values of a_n). This is due to the fact that there are many more sequences of marks which satisfy the difference constraint for higher values of a_n. Cotta & Fernandez (2005) also demonstrated that the fitness landscape of the optimal golomb ruler problem is correlated. As with the job-shop scheduling problem, this implies that the networks representing the feasible region of the search space are probably assortative. This, combined with the large variance in neutral degree will result in the neutral evolution of robust organisms (see § 3.2), and would likely confine the population to the regions of very high feasible degree. Cotta & Fernandez (2005) discuss how, even though they represent areas of low fitness, the population is drawn towards the regions of the search space with high feasible degree. They further discuss the negative impact that this has on search and, in order to mitigate this effect, devise a problem representation which generates regular feasible networks. It is demonstrated that this representation improves search performance.

Neutrality

The author is aware of one combinatorial optimisation problem for which neutrality has been explicitly studied: the graph-coloring problem. Marmion et al. (2013) studied the distribution of neutral degree in a variety of instances of this problem. It was found that the average degree of neutrality in most instances of this problem was moderate and that the variance in the neutral degree within each instance was small. This low variance in neutral degree is consistent with

the regular feasible region of this problem, as discussed above. Marmion et al. (2013) also studied the autocorrelation of neutral degree along neutral walks in the landscape. That is, they studied the assortativity of the neutral networks (§ 2.5). It was found that the neutral networks were highly assortative, with assortativity coefficient values between 0.69 and 0.9. An evolving population on such a neutral network would, therefore, increase its robustness over time (§ 3.2).

5.2 Genetic Programming

The area of EC in which the structure of neutral networks has been studied in the most detail is *Genetic Programming* (GP) (Poli et al., 2008; Langdon & Poli, 2013), which concerns itself with the evolution of computer programs. A key issue within GP is *bloat* (Poli, 2003; Luke & Panait, 2006), whereby programs evolve to be excessively long. Some explanations for bloat have involved neutrality. It has been suggested that bloat is caused by *introns*: code fragments that have no effect on the execution of the program (Nordin et al., 1995). If there is no selective pressure against such code fragments, they will accumulate over an evolutionary run. Due to their lack of effect on the program execution, most mutations of such fragments are neutral (Ferreira, 2002). It has also been argued that the mutational robustness conferred by neutrality offers a selective advantage to longer programs, which tend to have more neutral neighbours (Banzhaf & Langdon, 2002; Blickle & Thiele, 1994; Banzhaf et al., 1998). Finally, it has been argued that long programs are overrepresented. That is, for a given program, there are more long realisations than short ones. This implies that a neutral drift will be biased towards these longer programs (Langdon & Poli, 1998a,b; Langdon et al., 1999). This relationship between neutrality and bloat has lead to substantial emphasis being placed on neutrality in GP (Banzhaf, 1994; Ebner, 1999; Yu & Miller, 2001; Miller & Smith, 2006; Galván-López et al., 2008). In the following, two areas in which the structure of neutral networks has been specifically studied, are discussed.

Linear Genetic Programming

Wolfgang Banzhaf and various colleagues have performed detailed studies of the structure of neutral networks in *Linear Genetic Programming* (LGP). The distin-

guishing feature of LGP is that the program representation (§ 2.1) is a sequence of instructions, as opposed to the usual tree representation (Brameier & Banzhaf, 2001). This, combined with the fact that the studies to be discussed did not use crossover (recombination), means that the results discussed in this thesis apply exactly.

The early work of Banzhaf & Leier (2006) fully enumerated the neutral networks of a simple boolean program space. These were programs made up of two input registers, two calculation registers and $L = 2$ statements. It was found that all neutral networks were fully connected (there were only five different fitness values), that the neutral degree of genotypes ranged between 6 and 33 and that the distribution was heavily skewed towards the robust genotypes.

This landscape was further studied for $L = 3$ statements in Hu et al. (2014) and $L = 4$ statements in Vanneschi et al. (2006) and Hu et al. (2012). It was found that the size of the neutral networks varies dramatically, from order 10 to order 10^5 genotypes in the $L = 3$ case and from order 10^4 to order 10^8 genotypes in the $L = 4$ case. These networks were also found to be expansive, that is, each network extended over much of genotype space. The degree distribution was, interestingly, bimodal. The authors performed a K-core analysis (Dorogovtsev et al., 2006) and found that the nodes in the high-degree peak of the degree distribution were found in the higher k-cores and that the nodes with lower degree were found in the lower cores. This lead them to conclude that these high degree nodes formed a highly connected core surrounded by lower degree peripheral nodes. Although the authors did not calculate the assortativity of the networks, the structure which they described implies that these networks are probably highly assortative. It is plausible that this assortativity is sufficiently high for evolving populations to localize on the highest K-core (Pastor-Satorras & Castellano, 2016). At the very least, it will lead to a higher population concentration on the inner nodes of the network (§ 3.2). This will greatly reduce the phenotypic diversity to which the population is exposed, as the peripheral nodes have more phenotypically non-neutral neighbours.

The more recent $L = 4$ paper (Hu et al., 2012) performs an analysis of the various mutational biases present in the problem landscape, specifically those between genotypes and phenotypes. Of particular interest to this thesis, they demonstrate the existence of the mutational bias towards robust genotypes by performing random walks over neutral networks. This bias is discussed in § 2.5 in the context of the friendship paradox (Feld, 1991) and is used in the derivation

of the population distribution over neutral networks presented in § 3.2.

Neutral Networks of Real-World Software

Contrary to the intuition of most software developers, software is not brittle (Langdon, 2015). Experiments on various pieces of human-coded software have revealed that between 20% and 90% of mutations leave the software capable of passing its test suite (Langdon & Petke, 2017; Schulte et al., 2014), this proportion being dependent on the genetic representation, mutation operators, test suite coverage and properties of the software itself. This robustness allows the neutral space to be traversed while secondary objectives are optimised, in what has come to be known as *genetic improvement* (Langdon & Ochoa, 2016). This technique has been used, for instance, to improve computational efficiency (Langdon & Harman, 2015), repair latent bugs (Le Goues et al., 2012) and improve energy efficiency (Schulte et al., 2014). The derived programs in these studies are as far as hundreds of mutations from the seed program, implying the existence of extensive neutral networks in their search space. Moreover, it has been found that, in the case of computational efficiency, the majority of neutral mutations have no effect on the secondary objective (Langdon & Petke, 2017). This implies that it is likely that neutral dynamics play a role in the evolution of these programs. It is, therefore, important to understand the topology of the neutral networks of real-world software and the manner in which this topology influences evolution.

To this end, Schulte (2015) and Schulte et al. (2014) explicitly studied the neutral networks of software. By performing neutral random walks through the search space, they were able to confirm that the neutral networks extend over a large distance. These random walks were able to find programs 250 edits from the seed program, which was well-tested and contained less than 200 lines of code. Furthermore, it was found that mutational robustness increased over the course of the random walks. This implies that the neutral network is not regular and that there is variation in the degree of the nodes. This indicates that asexual evolution would be effective at evolving robust programs on this network, especially were it to exhibit degree assortativity (§ 2.5).

Schulte (2015) suggests that the diffusion of a population on a neutral network can be beneficial in and of itself. The neutral evolution of pre-existing software would be an easy way to create many variants of the same program, and

this diversity has benefits for system security.

5.3 Digital Circuits

The study of the evolution of digital circuits bears much similarity to that of gene regulatory networks, as both consist of interconnected boolean functions. Moreover, some authors have used networks of non-linear threshold units to model gene regulatory networks (see, for instance, Crombach & Hogeweg (2008)). This discussion will, however, maintain the separation between natural and artificial systems. See (§ 6.4) for a discussion of gene regulatory networks in the context of this thesis.

Various authors have studied the neutral networks of digital circuits (Milano & Nolfi, 2016; Raman & Wagner, 2010; Fernández & Solé, 2007). These studies all employed substantially different models. For example, as logic units, Milano & Nolfi (2016) used AND, OR, NAND and NOR gates, Raman & Wagner (2010) used these same gates with the addition of XOR and Fernández & Solé (2007) used threshold units. There were further differences in the size and allowed topology of the circuits, the numbers of inputs and outputs and the genotype representation and mutation operators used.

Despite these differences, the studies are in remarkable agreement as to the properties of the neutral networks. All three found that the there was a diversity of robustness amongst genotypes - some genotypes had substantially more neutral neighbours than others. Although Milano & Nolfi (2016) did not study the span of the neutral networks, Raman & Wagner (2010) along with Fernández & Solé (2007) confirmed that they covered large parts of the genotype space.

The observed variation in neutral degree opens up the possibility of using evolution to increase the fault tolerance of digital circuits. Indeed, it has been demonstrated that this is an effective approach (Thompson & Layzell, 2000; Hartmann & Haddow, 2004). The results presented in this thesis suggest an approach to the evolution of fault tolerant circuits that would not require fault tolerance as a secondary objective. If a genotype representation and associated mutation operator were devised so as to enforce the assortativity of the neutral networks, then the population would converge on areas of higher robustness. Assuming that the mutation operator bore some similarity to potential faults, these robust solutions would be tolerant to such faults.

5.4 Fitness Landscape Models

Although there is a limited number of results from the study of fitness landscapes models which are directly relevant to this thesis, their importance within the theoretical study of evolution (Galván-López et al., 2011; Stadler, 2002), justifies some discussion of their relation to this work. Moreover, some of the most compelling future work extending this thesis would be to investigate the manner in which the parameters of certain fitness landscape models influence the relevant properties of neutral networks and the implications of this for neutral evolution. This will be discussed in further detail below.

The most studied fitness landscape model is Stuart Kauffman's ubiquitous (De Visser & Krug, 2014; Richter, 2014; Pitzer & Affenzeller, 2012) NK model (Kauffman & Levin, 1987; Kauffman & Weinberger, 1989). Its multiple variations and generalisations (Bäck et al., 1997) have been studied in detail and have had a substantial impact on the study evolutionary dynamics and optimization algorithms (Wright et al., 2000; Choi et al., 2008; Nielsen et al., 2015; Mellor, 2007; Weinberger & Fassberg, 1996). Moreover, there exist at least three variations of this model which incorporate neutrality (Barnett, 1998; Newman & Engelhardt, 1998; Beaudoin et al., 2006). It is for these reasons that this section will focus on the NK landscape model and its neutral variants.

The NK model considers a landscape whereby each of the N characters of the genetic code contribute a given amount of fitness, and this fitness is summed to produce the genotype's fitness value. However, the fitness contribution of each character is dependent on the characters at K other positions in the code. More formally,

$$F(\mathbf{s}) = \sum_{i=1}^{N} f_i(\mathbf{s}_i) \tag{5.1}$$

Where F is the genotype's fitness, $\mathbf{s}$ the string representing its genetic code and $\mathbf{s}_i$ is the substring of $\mathbf{s}$ containing the ith character and the other K characters on which it depends.

The utility of the NK model lies in the manner in which the level of epistasis and ruggedness can be tuned. Increasing K increases both of these quantities (Kauffman & Weinberger, 1989).

The NK model does not, by itself, exhibit any neutrality. However, the author is aware of three extensions to it which incorporate neutrality. Barnett (1998) proposed the NKp landscape where, for a given $\mathbf{s}_i$, with probability p, the fitness

contribution of that locus ($f_i(\mathbf{s}_i)$) is set to zero. Newman & Engelhardt (1998) proposed the NKq landscape, which has also been referred to as the 'quantised' or 'terraced' NK landscape. This modification operates by requiring the range of the functions f_i to be drawn from a given discrete set of size q, usually the integers $[0, q)$. It has been noted that the resulting landscape of the NKq model bears much similarity to a normal NK landscape which has been discretised into 'bands' (Aita, 2008), That is, the range of fitness values is separated into discrete intervals, and all fitness values which fall within a given interval are given the same fitness. Finally, the ND landscape (Beaudoin et al., 2006) presents a method for constructing NK landscapes where the degree distribution of the neutral networks is a parameter of the model. This could be a useful tool for future work, particularly in terms of analysing the effect of localisation (chapter 4) on a population's ability to find neutral networks of higher fitness. However, it does not shed light on the possible underlying causes of topological features of neutral networks, as the NKp and NKq models do.

Many of the properties of the NKp and NKq landscapes have been studied in depth. This includes the size of the neutral networks (Newman & Engelhardt, 1998), the global distribution of neutral mutations (Reidys & Stadler, 2001; Geard et al., 2002), fitness correlation (Smith et al., 2002; Barnett, 1998) and various aspects of the evolutionary dynamics (Smith et al., 2002; Barnett, 1998). Unfortunately, these results are not of great use to us here, as we are interested in the topological properties of networks of a given fitness. Barnett (2003) derived an expression for the degree distribution of neutral networks of the NKp landscape and plotted the mean and variance of this distribution for a specific combination of parameter values. Unfortunately, the derived expression is rather complicated and difficult to interpret. Moreover, the plots were only performed for the value $p = 0.99$. As would be expected, almost all the 1-mutation neighbours of a given phenotype were neutral and there was a very low variance in the degree. An interesting avenue for future work would be interrogating this distribution for other parameter values.

Unfortunately, the author is not aware of any work which analyses the topological properties of the neutral networks of the NKq model. Further, it is the author's opinion that, due to its similarity to a banded NK model, the NKq is a much more realistic model of the manner in which neutrality arises in evolutionary systems. Therefore, a pressing area of future research is interrogating the topology of the neutral networks of the NKq model, with particular focus

on features such as modularity, degree assortativity and variance in degree. This will then allow us to relate the much studied topics of epistasis and landscape ruggedness to the behaviour of populations evolving at high mutation rates.

5.5 Cellular Automata Majority Problem

Although there has been a substantial amount of work on the problem of evolving rules for cellular automata (Mitchell et al., 1996), there is a lack of research concerning the characteristics of the neutral networks imposed by these problems. That being said, the neutral networks of the cellular automata majority problem have been studied in detail (Verel et al., 2006, 2007). Given that this work focuses on the properties which we are interested in, it is worth briefly discussing it. The authors of the mentioned studies studied two of the neutral networks of this problem. In both, they found that the degree distribution of the networks was highly dispersed - the highest degree nodes had a degree around five times higher than the lowest degree nodes. The autocorrelation of degree along random walks was also measured, and it was found that the networks exhibited positive degree assortativity, with $r = 0.85$ and $r = 0.49$. Based on the work in § 3.2 we can conclude that evolution on these neutral networks would result in populations of robust genotypes. Moreover, the highly assortative neutral networks would be a strong candidate for localization on its K-core (Pastor-Satorras & Castellano, 2016).

CHAPTER 6

NEUTRAL NETWORKS IN NATURE

As mentioned in the introduction to the previous chapter, given that results have been derived relating the topology of neutral networks to the equilibrium distribution of a population over it, the topology of the neutral networks encountered by evolution is of great significance to this thesis. The previous chapter analysed the existing literature concerning the topology of neutral networks encountered in EC. This chapter will analyze those found in nature.

It is worth bearing in mind that, except perhaps in the case of *in vitro* experiments, neutral networks in natural systems are a fairly immutable property of that system. This is in contrast with EC, where neutral networks are a consequence of the representation, associated mutation operators and fitness function. These can all be modified by the practitioner in order to generate the desired behaviour. Therefore, in the context of EC, the results derived in this thesis can be viewed as a tool to aid in the design of representations, mutation operators and fitness functions. On the other hand, in nature, they offer a description of the behaviour of the given system.

Naturally occurring neutral networks have been analyzed at various levels of selection (Smith & Szathmary, 1997). This chapter will proceed along these levels.

6.1 RNA Folding Neutral Networks

Much of the earliest work on the topology of neutral networks was conducted on RNA (Reidys et al., 1997; Fontana & Schuster, 1998; Schuster et al., 1994; Grüner et al., 1996; Fontana et al., 1993). It was found that, for a given fold shape, there are many different RNA sequences folding into that shape. A large focus of this early work was some of the immediate consequences of this degeneracy, namely *spanning* and *shape space covering*. The former refers to the existence of large neutral networks which percolate through sequence space and stretch across its diameter. The latter refers the existence of genotypes coding for all realisable phenotypes within a certain small distance of any given genotype. These properties have important consequences for evolving populations and the amount of variation which they can access. Specifically, spanning allows a population to navigate the breadth of sequence space, even if it has reached a fitness plateau. Shape space covering implies that a population navigating the network will be able access all realisable phenotypes. These findings have also been confirmed in more recent studies (Jörg et al., 2008; Cowperthwaite et al., 2008; Grafen, 2008).

The results of this thesis, notably the possibility of the localization of an evolving population, have implications concerning whether the population will actually access this available variation. These implications are further discussed in chapter 8.

More recently, Aguirre et al. (2011) performed an exceptionally detailed analysis of the topology of the neutral networks of the folding of all RNA sequences of length 12. It was found that the smaller neutral networks had strikingly narrow degree distributions, whereas the larger networks had a slightly higher variance. Despite this higher variance, the networks did not contain hubs. The modal degree was only slightly lower than the maximum. The majority of the networks exhibited positive degree assortativity. The larger networks had particularly strong assortativity, with most having $r > 0.5$, and many having r values close to 1. The authors chose to only plot a diagram for a single neutral network, however this one appeared to be fairly modular.

This is an interesting combination of properties in the context of this thesis. The narrow degree distributions preclude the population's average robustness from rising much above the networks' average degree. However, the strong positive assortativity and possibility of modular structure present the possibility that the population might be heavily concentrated in certain regions of the networks.

Fortunately for our purposes, the authors computed the principal eigenvalue, λ_1, for the networks under study and plotted it against both the network size, N, and average degree $\langle k \rangle$. As expected, λ_1 was well approximated by $\langle k \rangle$, being slightly larger than it in almost all instances. The authors also plotted the components of the principal eigenvector for the largest connected subnetwork in the landscape. It was found that this value was not constant across the network and that there were clusters of values. That is, certain groups of nodes had very similar eigenvector components, and there was a larger difference in these components between these groups. The authors were further able to show that the sequences within these groups had common sub-strings of nucleotides. Moreover, the group with the highest eigenvector components contained the highest degree nodes. It seems plausible, therefore, that the population is concenctrated on a module containing the high degree nodes. Unfortunately, the axis of the plot presenting these results is hard to interpret, and so we cannot reach conclusions concerning the extent of this concentration.

6.2 Protein Folding Neutral Networks

The topological properties of the neutral networks of protein folding have been studied in greater detail than probably any other naturally occurring neutral network. This is partly because protein neutral networks form an especially interesting structure: the *superfunnel* (Bornberg-Bauer, 1997; Bornberg-Bauer & Chan, 1999; Xia & Levitt, 2004; Noirel & Simonson, 2008; Wroe et al., 2005). The networks are centered around a *prototype* sequence, which is also the maximum degree node of the network. The degree of the nodes decreases with distance from the prototype. This structure bears some similarity to a *Hamming ball* (§ 4.8) and Bornberg-Bauer & Chan (1999) used Hamming balls as an abstraction with which to study these structures. The sequences within the network also differ in the stability of their folds, and greater stability confers a fitness advantage on the sequences. This stability is at a maximum at the prototype sequence and decreases with increasing distance from it.

The population distribution over the superfunnel has been studied both with and without the effects of stability. That is, under neutral evolution or on a fitness landscape. As shown in this thesis (§ 4.8), as well as by Bornberg-Bauer & Chan (1999), neutral evolution on a Hamming ball does not result in an enor-

mous concentration of the population at the prototype sequence. The prototype sequence does receive the highest population concentration, however, the fraction of the total population situated on this node is rather small. Rather, most of the population is found on the peripheral nodes. This observation has been confirmed using very different models of evolution on protein networks (Blackburne & Hirst, 2005; Bloom et al., 2007).

The population distribution over these networks will, therefore, remain relatively homogeneous and will be determined by this Hamming-ball-like topology. The author can, however, think of two possible mechanisms by which localization might occur in protein neutral networks. Both of these involve deviations from the Hamming ball abstraction. Firstly, it has been reported that the stability landscape of proteins is rugged and not a perfect funnel shape (Bastolla et al., 2000; Tiana et al., 2001). The degree distribution of the nodes could be similarly rugged. Indeed, Sikosek & Chan (2014) provide examples of neutral networks with multiple hubs. Assuming that the connection with the rest of the network is not particularly strong, the population could localize on the highest degree hub.

The second mechanism involves the size of the networks and the rate of the decay of the degree distribution. The studies cited above all only analyzed the networks generated from relatively short sequences, due to the enormous size of the search space for longer sequences. Longer sequences should allow for larger neutral networks. If the decay of the degree distribution in these networks is gradual, then they will no longer be well approximated by Hamming balls.

6.3 Protein Interface Neutral Networks

Podgornaia & Laub (2015) studied the neutral network of the key residues of the *E. Coli* protein kinase PhoQ involved in the PhoQ-PhoP interface. The authors determined that there were four key residues involved in the functioning of this interface. They studied the neutral network formed by the nucleotides coding for these four residues, under the constraint that the interface functioned with roughly equal efficacy to the wild type. This efficacy was confirmed in head-to-head competitions against the wild-type.

Although they did not report on the degree distribution, or any other network metrics, the authors did present a diagram of the network with a force-directed

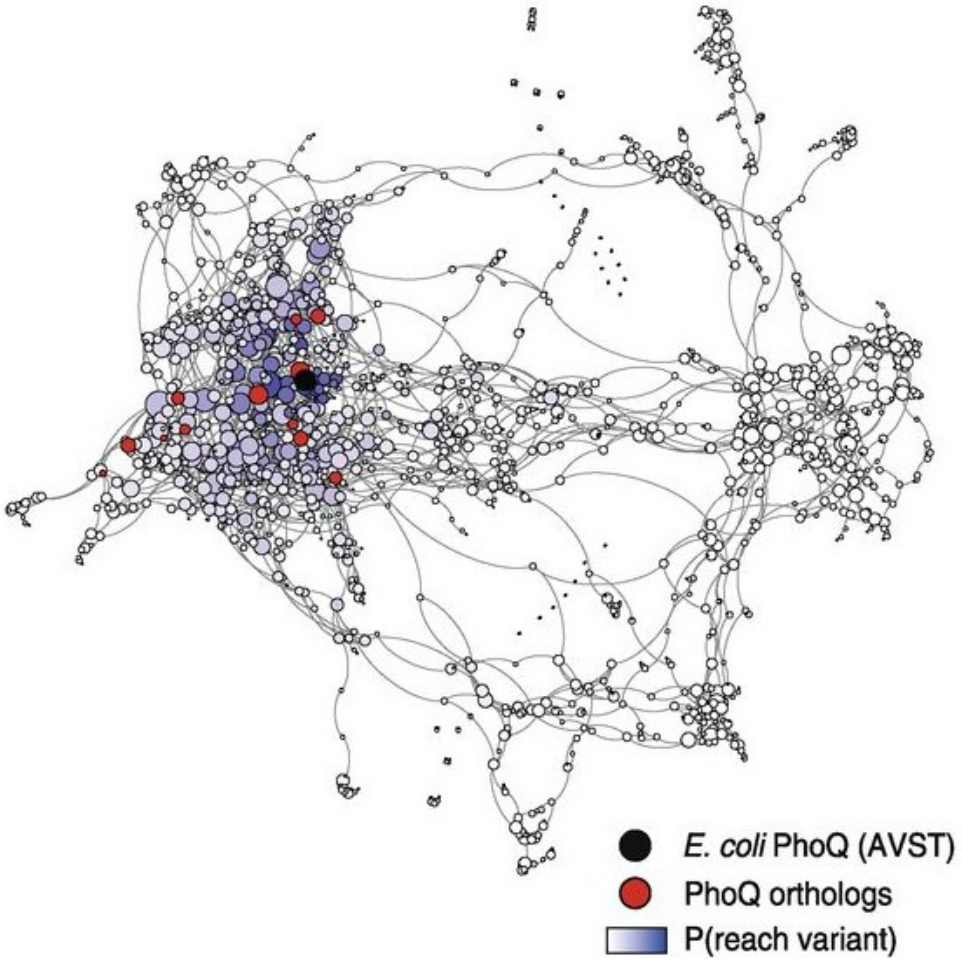


Figure 6.1: Neutral network of the nucleotides coding for four residues of variants of the *E. Coli* protein kinase PhoQ. These residues are the key residues involved in the PhoQ-PhoP interface. Here, neutrality was determined by the funcionality of the interface. The size of the nodes is proportional to their degree. The node colors represent the probability that a random walk of 20 steps starting from the wild type. The wild-type is dark blue. The red nodes represent naturally occurring orthologs. Taken from Podgornaia & Laub (2015). Used with permission.

layout. This diagram is shown in figure 6.1.

This network contains a single large module which holds the highest degree nodes. There are multiple smaller modules which are weakly connected to the large module and to each other. Given this structure, it is plausible that, should a population evolve on this network with $N\mu \gg 1$, it will be confined to the larger module. The authors of this work were themselves interested in the consequences of the weak connectivity of this neutral network. They studied the

probability that a random walk of length 20 starting from the wild-type would reach a given node and found that it was extremely unlikely that it would reach the smaller modules.

This network demonstrates the plausibility of an exploration catastrophe occurring in nature, without the need for special conditions such as plastogenetic congruence (Ancel & Fontana, 2000). As such, calculating the population distribution over this particular network is a likely avenue for future research.

6.4 Gene Regulatory Network Neutral Networks

After studying the neutral topologies resulting from encoding proteins, we can query the structure of the neutral networks resulting from the interactions of these proteins with the expression of other proteins. That is, we can query the structure of the neutral networks of *Gene Regulatory Networks* (GRNs) (Schlitt & Brazma, 2007). The neutral networks of GRNs, or *metanetworks*, have received much attention from various researchers (Luo & Turner, 2011; Payne & Wagner, 2013; Boldhaus & Klemm, 2010; Cotterell & Sharpe, 2013), however, the focus of these works has been whether the neutral networks form connected giant components which can be used to traverse space, as opposed to metrics such as the degree distribution.

Before proceeding, it is worth mentioning that, when computationally modelling GRNs and their associated neutral networks, there is a large number of parameters which have to be set which impact on the results of the model. For instance, if one is using a boolean network model (Schlitt & Brazma, 2007), one must decide on the permissable boolean functions, connection structure and update rules, along with the associated mutation operators for these items. Moreover, unlike the cases of RNA and protein folding, it is not obvious what should constitute a phenotype. Some authors define viable phenotypes as only those which attain an equilibrium state, and define this state as the phenotype (eg: Ciliberti et al. (2007a)). Other work (eg: Luo & Turner (2011)) defines the phenotype as a time series of node states. This author is not aware of any work which examines the impact of these various parameters on the structure of the neutral networks.

There are, however, some studies which report on topological properties of the neutral networks of GRNs. Ciliberti et al. (2007a); Payne et al. (2014); Raman

& Wagner (2011) report on the global distribution of robustness in the space of GRNs. They find that the distribution of robustness is highly heterogeneous, spanning nearly the entire range of possible values. It is worth pointing out, however, that these global quantities do not necessarily imply that the degree distribution of the individual networks is similarly heterogeneous.

Fortunately, Ciliberti et al. (2007b), studied the topology of specific neutral networks of a GRN landscape. It was found that these networks had a highly heterogeneous degree distribution. Furthermore, the provided network diagram showed multiple loosely connected hub nodes. These features make these networks a candidate for localization behaviour. Indeed, the authors presented the results of calculations of the principal eigenvalue (population average robustness). It was found that the eigenvalue was substantially larger than the average degree of the networks. Given that the degree distribution of the networks was skewed towards lower degrees, it is likely that this high level of population robustness was achieved through localization.

6.5 Neutral Networks of Other Systems

The networks connecting the viable phenotypes of metabolic networks have been studied (Rodrigues & Wagner, 2011; Hosseini et al., 2015). Due to the lack of results directly applicable to the focus of this thesis, they are not discussed further. The networks formed by viable phenotypes of larger organisms have also been analyzed (see eg: Dall'Olio et al. (2014)). As larger organisms do not produce polymorphic populations (Wagner, 2011), an assumption in this work (§ 2.3), further discussion on the topology of their neutral space is unnecessary. Finally, the variation networks of influenza Wagner (2014) have also been studied. This will be discussed in the following chapter, which focuses on this topic.

CHAPTER 7

NEUTRAL NETWORKS OF INFLUENZA HAEMAGGLUTININ

It is estimated that influenza results in 3 to 5 million cases of severe illness and a quarter to a half a million deaths, annually (Bao et al., 2008). This high disease burden, combined with the large amounts of freely available data on the virus (Sayers et al., 2012), makes the evolution of influenza an attractive area of study. Moreover, given its high mutation rate, we know that influenza exhibits quasispecies dynamics (Lauring & Andino, 2010), implying that its neutral evolution is in the polymorphic regime (§ 2.3).

Influenza is constantly evolving through a process known as *antigenic drift* (Boni, 2008), whereby it evolves so as to evade recognition by the host immune system. Much of the change in this process occurs on the haemagglutinin protein (Boni, 2008). Lapedes & Farber (2001) developed a methodology for quantitatively describing the pheontype space of haemagglutinin antigenicity. They used multi-dimensional scaling to position haemagglutinin sequences in a space, where, based on experimental data, the distance between sequences was closely correlated to the similarity of the immune response towards those sequences. Smith et al. (2004) used a modified version of this method to position haemagglutinin sequences from H3N2 strains occurring between 1968 and 2002 in a two dimensional phenotype space. It was found that the strains formed clusters in antigenic space, where the immune response towards sequences within a given cluster was very similar, whereas the response towards sequences in different clusters was substantially different. Furthermore, it was found that, during a

given influenza season, a single cluster was dominant. Clusters would dominate for a period of one to eight seasons before being replaced by a novel cluster. After replacement, the strains of a given cluster might be present for up to two more seasons before disappearing.

This has lead some authors (Koelle et al., 2006; van Nimwegen, 2006) to conclude that, outside of cluster transitions, influenza haemagglutinin is evolving neutrally. This implies that the neutral dynamics over haemagglutinin sequence space are an important topic of study.

Wagner (2014) studied the mutational networks formed by amino acid sequences of the HA1 domain of H3N2 influenza haemagglutinin, collected between 2002 and 2007. It was found that these networks contained antigenically different sequences. As such, they were referred to as *genotype networks*, as opposed to neutral networks, due to the possibility of there being fitness differences between sequences.

Inferring natural fitness landscapes is a challenging problem (De Visser & Krug, 2014). As such, determining whether a mutational network of genotypes is neutral is similarly difficult. In this chapter, an attempt is made at deriving mutational networks from influenza sequence data which are plausible neutral.

One approach towards tackling this problem is to use sequences of H1N1 haemagglutinin collected after 2009 [1] as there has been minimal antigenic drift in this subtype since then (Guarnaccia et al., 2013; Tewawong et al., 2015; Liu et al., 2015). We make use of sequences of H1N1, haemagglutinin, along with H3N2 sequences, for this reason. A second strategy employed is to limit the sequences used to those sampled within a given season. The amount of antigenic variation within a single season is less than over multiple seasons. The amount of variation within seasons of H3N2 influenza reported by Russell et al. (2008) was of the order of the within cluster variation found by Smith et al. (2004). Furthermore, by restricting sampling to a given season, we reduce the number of antigenic clusters in circulation, which in turn reduces the chance that the neutral networks of different antigenic clusters will be linked by a point mutation.

An additional benefit of restricting sampling to a single season is that it allows us to approximate the distribution of the population over the network.

It is possible that, in the case of H3N2 viruses, the neutral networks of dif-

[1] As specified below, we analyzed sequences from 2007 to 2016. However, we excluded networks smaller than 50 nodes from this analysis and non of the H1N1 networks of seasons earlier than 2009 met this criterion.

ferent antigenic clusters will be linked by a point mutation and will, therefore, appear as a single neutral network in this analysis. The likelihood of this occurring, however, is low. Koelle et al. (2006) (supplementary text, figure S1) plotted the distribution of sequence samples from Smith et al. (2004) by cluster and year. In the majority of years, all sequences belonged to a single cluster. In no given year were sequences from three or more clusters sampled. In the majority of years in which a second cluster was sampled, this cluster was sampled at a very low frequency. Moreover, in the genetic map of Smith et al. (2004), we find that most clusters are well separated in sequence space, and only a few are within a single amino acid transition of one another. Finally, the genetic centroids of the clusters were well separated, and those that did overlap, only did so on the periphery of their distributions. This would imply that, should a neutral network analyzed here contain sequences from different clusters, separating the sequences belonging to different clusters should only involve the removal of a small number of edges and would not substantially impact on the observed topological patterns of the networks.

Confirming that the networks derived in this section are indeed neutral is an important area of future work.

7.1 Methods

In order to arrive at an estimation of the relative frequency of the various genotypes within the population, the methodology of Łuksza & Lässig (2014) was followed almost exactly. The second step was to infer the neutral networks from these genotypes.

The data set of this work is 12 352 nucleotide sequences of human H3N2 influenza haemagglutinin and 16 352 nucleotide sequences of human H1N1 influenza haemagglutinin obtained from the NCBI database (Bao et al., 2008). These sequences were all observed between 2007 and 2016 inclusive. We chose to only include sequences which were complete or near complete, missing only their start and stop codons. This data set has known biases. See Łuksza & Lässig (2014) for a discussion of the nature and effects of said biases.

The data set was binned into seasons of six months, representing either the northern or southern hemisphere winter. The northern hemisphere winter was taken to last from October of a given year through to March of the subsequent

year, inclusive. The southern hemisphere winter was taken to last from April of a given year through to September of that same year, inclusive. For the sake of brevity, the time period stretching from April through September of a given year y will simply be referred to as y. Similarly, the time period stretching from October of year y through March of year $y+1$ will be referred to as $y-y+1$ (eg: 2009-2010). Sequences without an associated date value, accurate to within a month, were excluded from the analysis.

After binning, the sequences within each bin were aligned using the MUSCLE program (Edgar, 2004), with the default parameter settings. Where there were gaps, nucleotides at that position, of sequences in that bin, were excluded from further analysis.

The multiplicity m_i of each unique sequence within a season was then calculated, where m_i is simply the number of times that sequence occurs within the data set. The relative frequency of the sequence is then

$$x_i = \frac{m_i}{M} \tag{7.1}$$

where M is the sum of all sequence frequencies within the season.

The neutral networks for a given season were then constructed by assigning each unique sequence a node and connecting nodes with an edge if their associated sequences differed by a single nucleotide. The network was then divided into its connected components, and each connected component is subsequently referred to as a separate neutral network (see § 2.5 for an explanation of this confusing practice). Neutral networks with fewer than 50 nodes were excluded from this analysis as it was deemed that they would not provide sufficiently interesting population distributions.

All subsequent analysis of the graphs was performed using the `Python` package `igraph` (Csardi & Nepusz, 2006). Specifically, `igraph` was used to find the principal eigenvalue and associated eigenvector of the adjacency matrix of the networks. That is, it found the predicted population average robustness and population distribution (§ 2.3). The calculation of the eigenvalues and eigenvectors of the adjacency matrices of graphs in `igraph` is performed using the `FORTRAN 77` package `ARPACK` (Lehoucq et al., 1998). `ARPACK` implements the *implicitly restarted Arnoldi method* (Lehoucq & Sorensen, 1996) to find the eigenvalues and eigenvectors of matrices. `igraph`'s default parameters for `ARPACK` were used.

7.2 Results

Diagrams of the neutral networks were constructed using the Fruchterman-Reingold force directed layout (Fruchterman & Reingold, 1991) and are displayed in figures 7.1 to 7.8. Both the population distribution predicted by the principal eigenvector and the actual population distribution estimated by sequence frequencies are visualised on these networks. For each network, two diagrams are displayed, both using the same layout. In one diagram the size of a given node is proportional to the predicted population concentration on that node. In the other, the node size is proportional to the estimated population concentration. Furthermore, the predicted population distribution is visualised with a colour gradient, where nodes with a higher predicted population distribution are more red and nodes with a lower predicted population distribution are more blue. Note that for both diagrams of a given network the node colouring is determined by the predicted population distribution. This was done to make the differences between the predicted and estimated population distribution easier to spot. For instance, in diagrams where the node size is proportional to the estimated actual population distribution, large blue nodes and small red nodes are signifiers of differences between the distributions. Various properties of these networks and the population distributions over them are shown in tables 7.1 and 7.2.

The sizes of the neutral networks varies dramatically. The majority are near the lower bound for inclusion of $N = 50$ nodes. However, the largest neutral network, the only one from the H1N1 2009 seasons, contained $N = 1071$ nodes.

Looking at the figures, the topology of the neutral networks bears great similarity to the genotype network of Wagner (2014). They have a highly heterogeneous degree distribution which is skewed towards low degree nodes. That is, most of the nodes have low degree, whereas there are a few nodes with very high degree. Moreover, the networks consist of multiple loosely connected stars. The stars are formed by hub nodes which contain many 'satellite' nodes which only connect to the hub. These stars are then connected to one another by far fewer edges than are present within a given star. This gives the networks a modular structure.

This modular structure and degree heterogeneity makes these networks perfect candidates for localization behaviour. Indeed, we find that the principal eigenvector is not evenly distributed over the nodes of the networks. The proportion of the population found on the highest degree node and its neighbours

(P_h in tables 7.1 and 7.2) varies between 0.56 and 0.94. In 12 of the 18 networks this figure is over 0.8. Moreover, the relative inverse participation ratio ($Y_r(\Lambda)$ in tables 7.1 and 7.2, see § 4.2), is substantially larger than one in all networks, ranging between 7 and 260. However, in most cases, $Y_r(\Lambda)$ is lower than the threshold for localisation of 30 set in § 4.2. These networks are, therefore, in an intermediate state between localisation and full delocalisation. This heterogeneity is clear in figures 7.1 to 7.8, where we see a very large population concentration on some of the high degree nodes of the networks. In many cases, we also find a substantial proportion of the population on the immediate neighbours of said high degree nodes (see, for instance, figures 7.3b, 7.4a, 7.4b, 7.6a, 7.6b, 7.7b, 7.7c and 7.8a)

In most of the studied networks, the population does not seem to have localized onto a single star (hub), but is concentrated on two or more such hubs. This then begs the question of whether the mode of localization seen here is localization onto the maximum K-core (§ 4.1)(Pastor-Satorras & Castellano, 2016). In order to investigate this question, we calculated P_{K_M}, the proportion of the population residing on the maximum K-core for each network. These values are displayed in tables 7.1 and 7.2. We find that, in every case, less than 40% of the population is located on the maximum K-core. We, therefore, conclude that localization on the maximum K-core is not a dominant effect in these networks.

We hypothesize that the maximum K-core might contain the centers of the stars in these networks. This opens up the possibility that the population could be localizing onto the centers of these stars along with their satellite nodes. In order to investigate this possibility, we calculated P_{Kn_M}, the proportion of the population of the population located on nodes in the maximum K-core and on the neighbours of nodes in the maximum K-core. In some instances (such as the largest neutral network of the 2011-2012 H1N1 season and both networks of the 2011-2012 H3N2 season) this group of nodes contained the majority of nodes within the network. However, in other instances (such as 2009 H1N1) this group contained a minority of nodes, yet over 90% of the population.

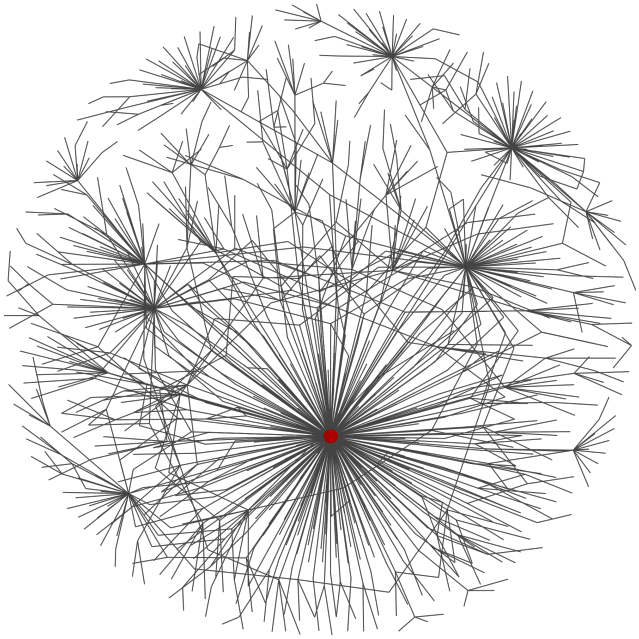
By examining the side-by-side figures showing the predicted vs. estimated distribution of the population over the networks, it would appear that the principal eigenvector approximates the population distribution fairly accurately. In order to arrive at a quantitative measure of the similarity of the two distributions, we calculated the *cosine similarity* (Singhal, 2001) between the principal eigenvector and the vector of the estimated population distribution. The cosine

Season	**Figure**	P_h	$Y_r(\Lambda)$	N	P_{K_M}	N_{K_M}	P_{Kn_M}	N_{Kn_M}	S_C
2009	7.1	0.9	259.96	1071	0.08	12	0.92	352	0.71
2010-2011	7.3a	0.91	15.18	62	0.26	9	0.95	45	0.85
2011-2012	7.3b	0.92	14.39	59	0.23	6	0.97	49	0.83
2012-2013	7.3c	0.61	11.32	65	0.33	9	0.79	27	0.85
2013-2014	7.4a	0.66	12.92	73	1	73	1	73	0.88
2013-2014	7.4b	0.85	17.47	73	0.2	4	0.88	33	0.83
2013-2014	7.4c	0.56	7.78	51	0.25	4	0.9	33	0.84
2015-2016	7.2	0.84	116.31	484	0.28	58	0.94	282	0.52

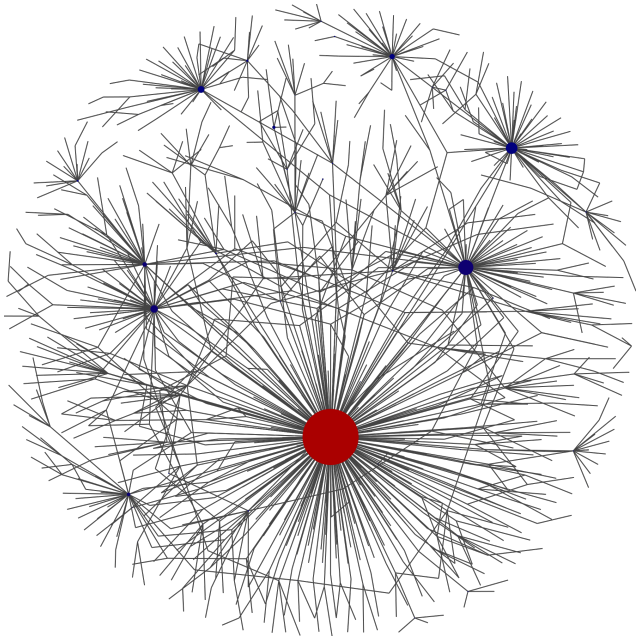
Table 7.1: Various properties of the neutral networks of H1N1 influenza haemagglutinin. Here, P_h is the predicted proportion of the population found on the highest degree node and its neighbours. That is, the sum of the components of the principal eigenvector corresponding to the highest degree node and its neighbours. $Y_r(\Lambda)$ is the relative inverse participation ratio (see § 4.2). N is the number of nodes in the network. P_{K_M} is the proportion of the population residing on the maximum K-core of the network (§ 4.1) (Pastor-Satorras & Castellano, 2016). N_{K_M} is the number of nodes in the maximum K-core of the network. P_{Kn_M} is the proportion of the population residing on the maximum K-core and neighbours of the maximum K-core. N_{Kn_M} is the sum of the number of nodes in the maximum K-core and the number of neighbours of nodes in the maximum K-core. S_C is the cosine similarity between the population distribution predicted by the principal eigenvector and the observed population distribution over the neutral network.

similarity views the vectors as vectors in some space and calculates the cosine of the angle between the two vectors. As such, it ranges between 1, for identical vectors, down to -1. The measured similarity values are shown in tables 7.1 and 7.2 under the column S_C. They confirm that the predicted and estimated distributions are indeed similar. The lowest value of S_C was 0.5. In 12 of the 18 networks it was greater than or equal to 0.8.

In figures 7.1 through 7.8, the predicted population distribution assigns less of the population to the centers of the stars and more to the satellite nodes than the estimated population distribution. This could possibly be explained by the fact that the data set is known to contain a bias against low frequency variants (Poon et al., 2016).

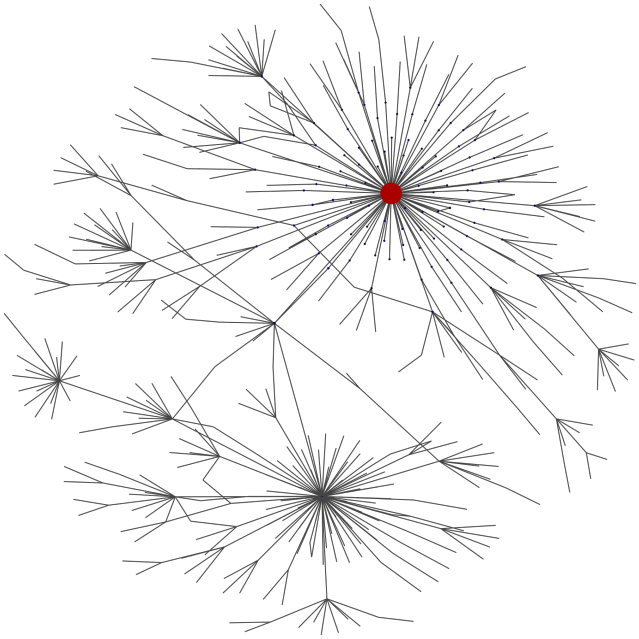


(a) Predicted population distribution.

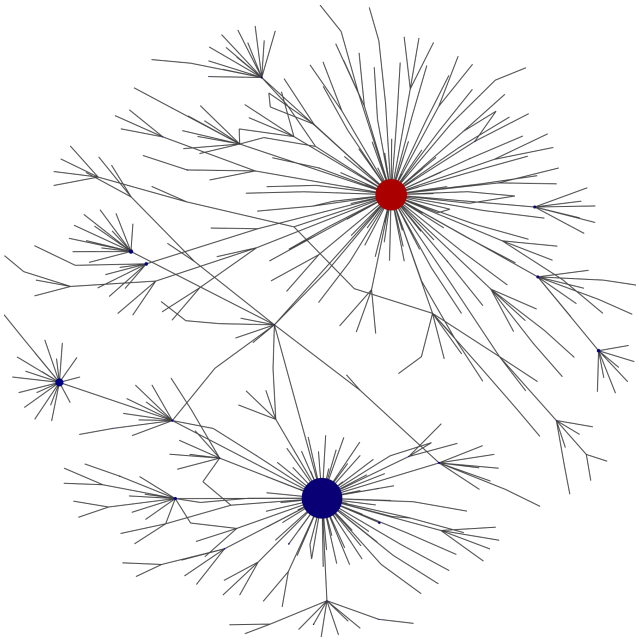


(b) Actual population distribution.

Figure 7.1: Largest neutral network of H1N1 influenza during the 2009 season. The size of a each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.

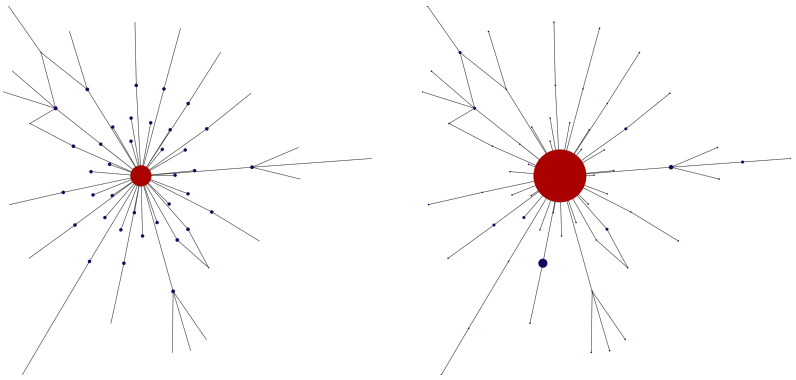


(a) Predicted population distribution.

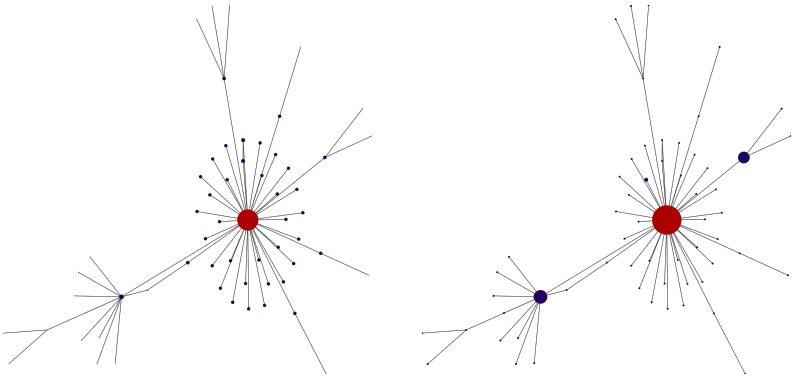


(b) Actual population distribution.

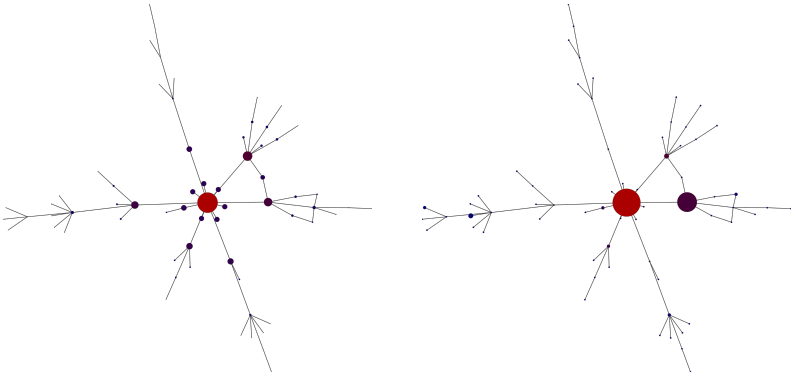
Figure 7.2: Largest neutral network of H1N1 influenza during the 2015-2016 season. The size of a each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.



(a) Largest network of the 2010 - 2011 season.

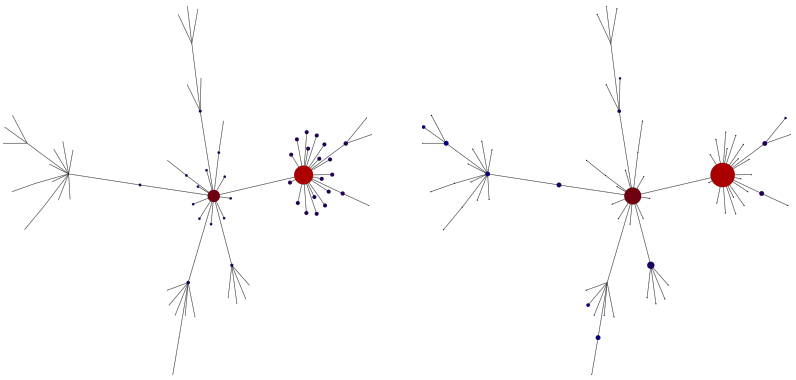


(b) Largest network of the 2011 - 2012 season.

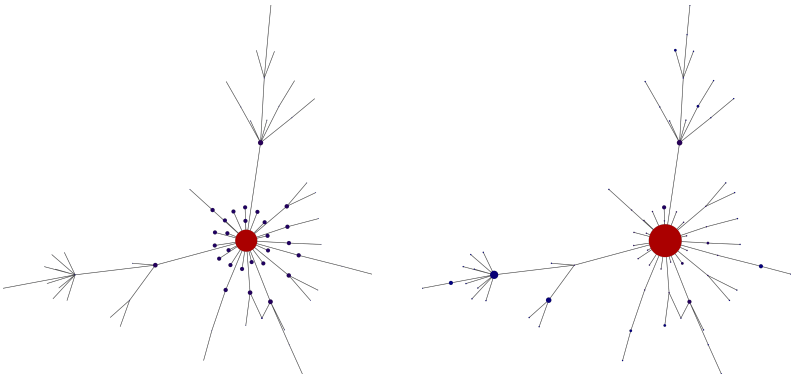


(c) Largest network of the 2012 - 2013 season.

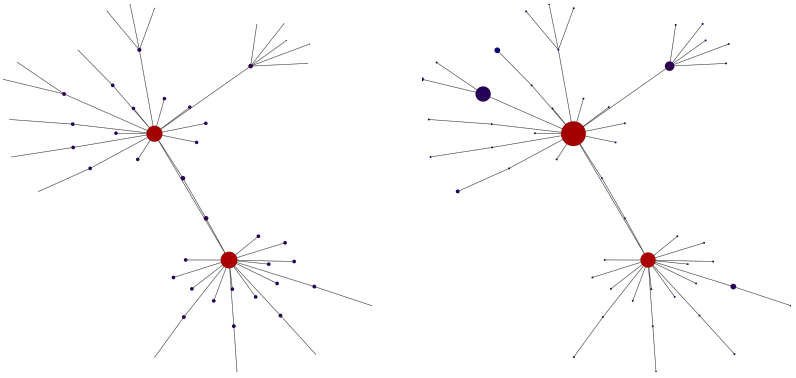
Figure 7.3: Neutral networks of H1N1 influenza during various seasons. Diagrams in the left-hand column show the population distribution as predicted by the principal eigenvector, whereas those on the right show the actual population distribution. The size of each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.



(a) Largest network of the 2013 - 2014 season.

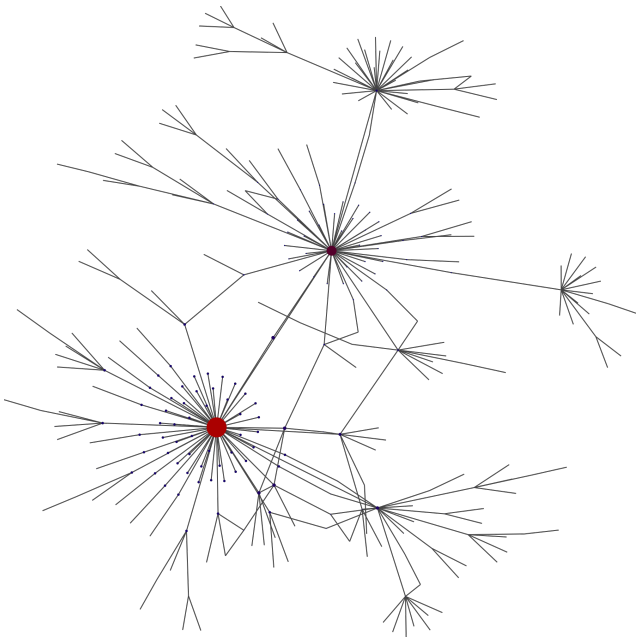


(b) Second largest network of the 2013 - 2014 season.

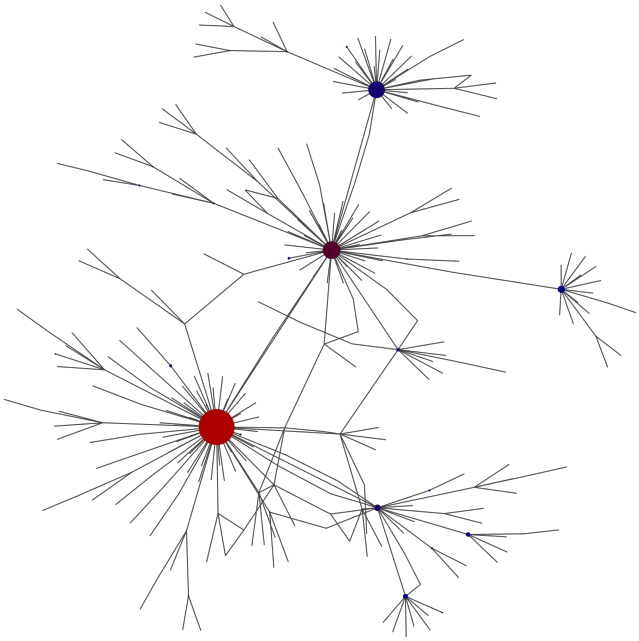


(c) Third largest network of the 2013 - 2014 season.

Figure 7.4: Neutral networks of H1N1 influenza during various seasons. Diagrams in the left-hand column show the population distribution as predicted by the principal eigenvector, whereas those on the right show the actual population distribution. The size of each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.

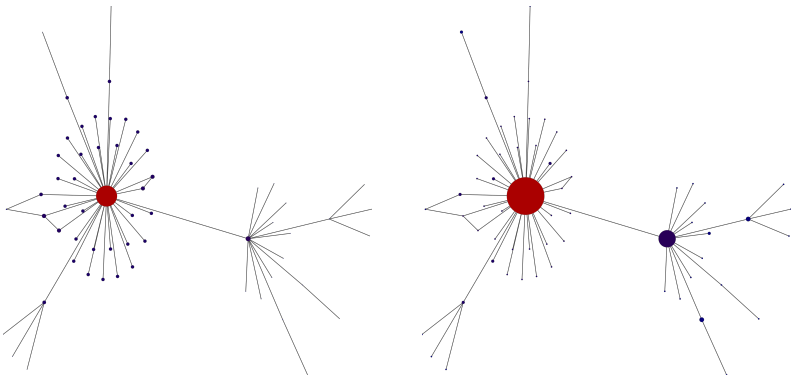


(a) Predicted population distribution.

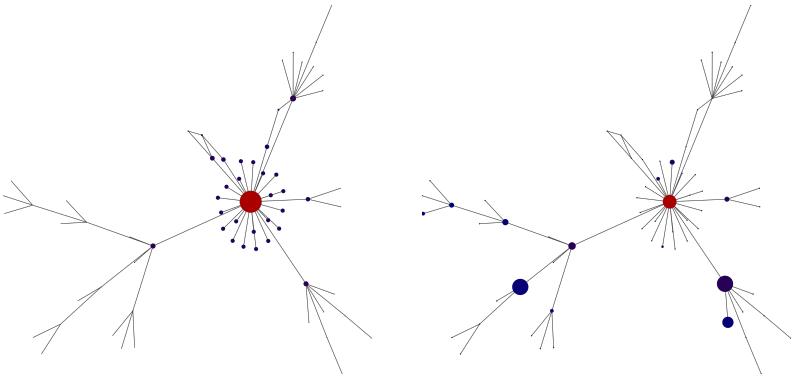


(b) Actual population distribution.

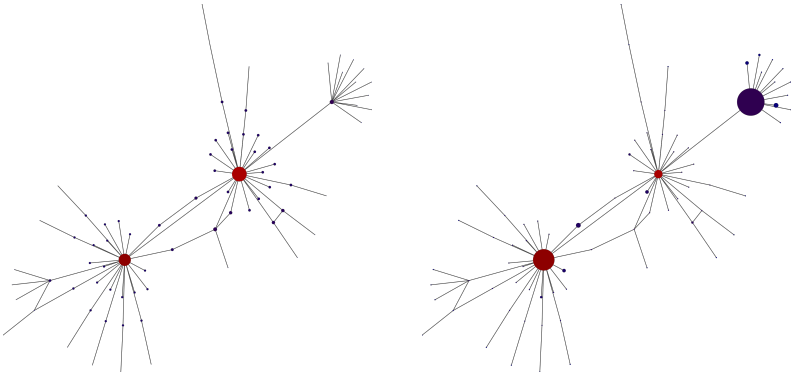
Figure 7.5: Largest mutational network of H3N2 influenza during the 2014-2015 season. The size of a each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.



(a) Largest network of the 2007-2008 season.

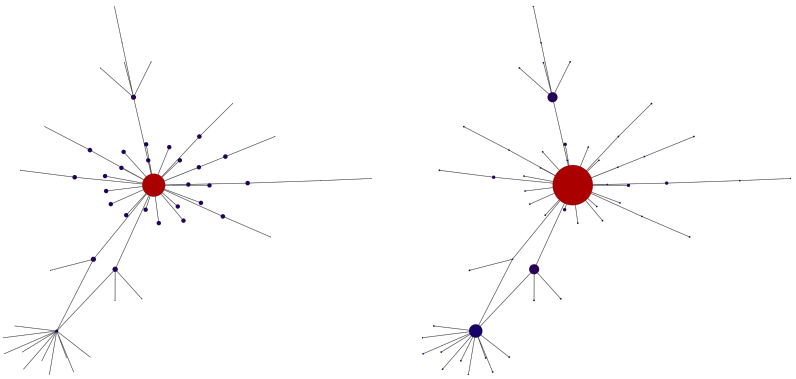


(b) Largest network of the 2009 season.

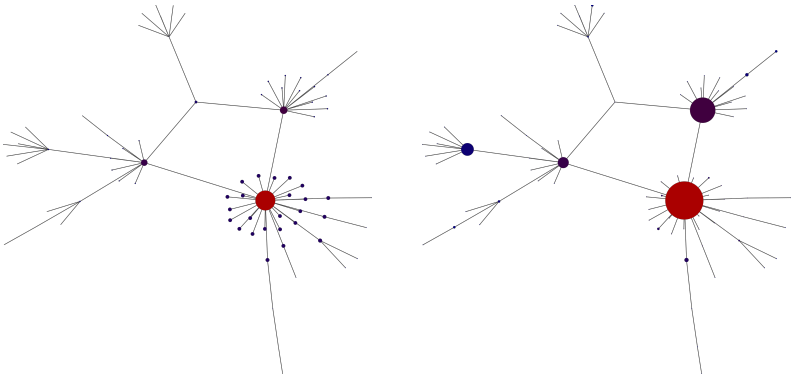


(c) Largest network of the 2010 - 2011 season.

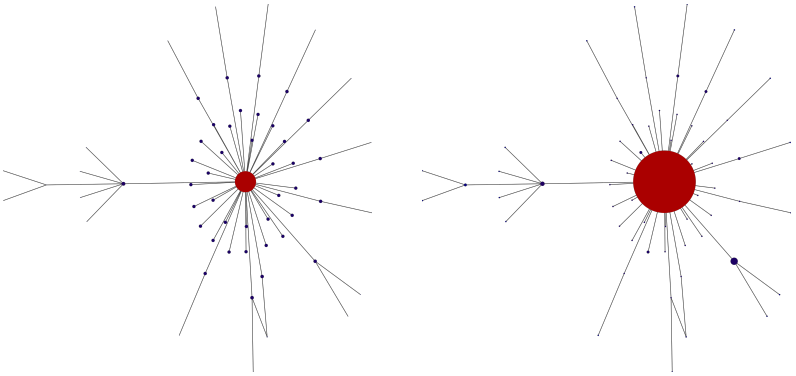
Figure 7.6: Neutral networks of H3N2 influenza during various seasons. Diagrams in the left-hand column show the population distribution as predicted by the principal eigenvector, whereas those on the right show the actual population distribution. The size of each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.



(a) Second largest network of the 2010-2011 season.

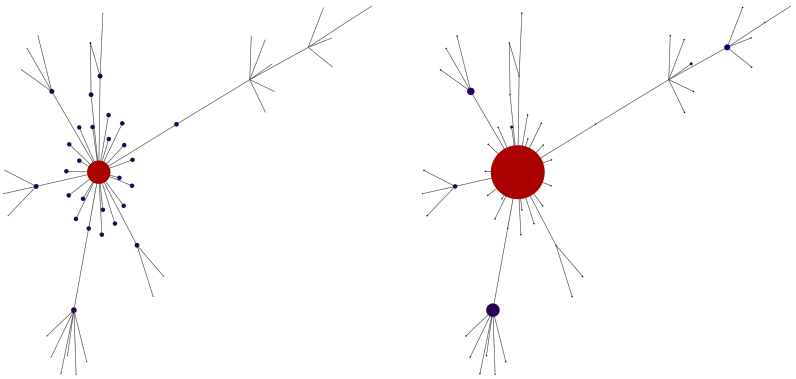


(b) Largest network of the 2012-2013 season.

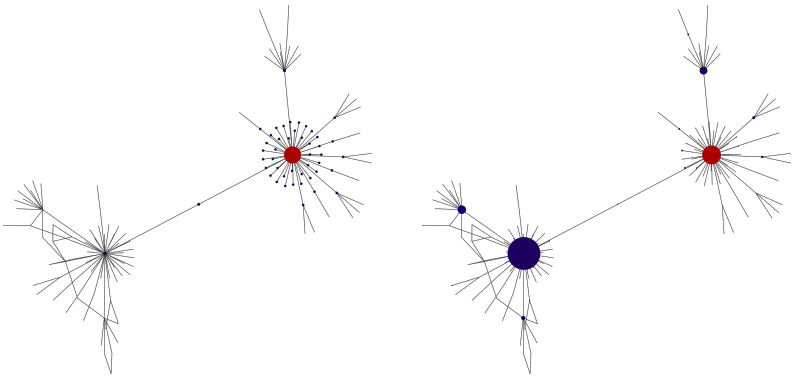


(c) Second largest network of the 2012 - 2013 season.

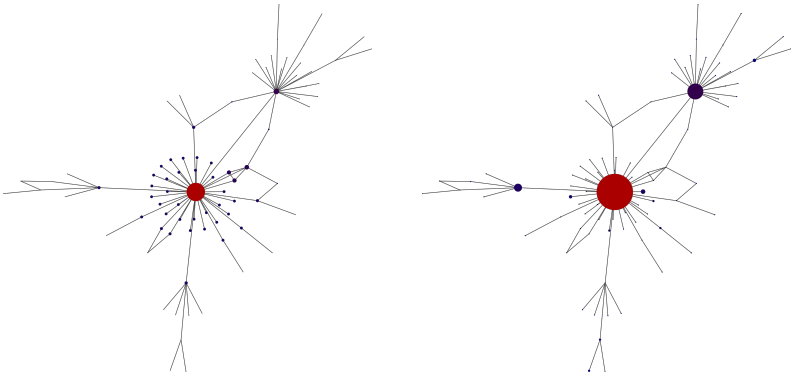
Figure 7.7: Neutral networks of H3N2 influenza during various seasons. Diagrams in the left-hand column show the population distribution as predicted by the principal eigenvector, whereas those on the right show the actual population distribution. The size of each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.



(a) Third largest network of the 2012-2013 season.



(b) Second largest network of the 2014-2015 season.



(c) Third largest network of the 2014 - 2015 season.

Figure 7.8: Neutral networks of H3N2 influenza during various seasons. Diagrams in the left-hand column show the population distribution as predicted by the principal eigenvector, whereas those on the right show the actual population distribution. The size of each node is proportional to the fraction of the population found on it. Moreover, nodes with a higher predicted population concentration are more red and nodes with a lower predicted concentration are more blue.

Season	**Figure**	P_h	$Y_r(\Lambda)$	N	P_{K_M}	N_{K_M}	P_{Kn_M}	N_{Kn_M}	S_C
2007-2008	7.6a	0.91	14.61	61	0.25	7	0.92	39	0.77
2009	7.6b	0.84	15.32	66	0.28	8	0.92	38	0.56
2010-2011	7.6c	0.58	11.29	78	0.34	12	0.93	57	0.84
2010-2011	7.7a	0.87	13	54	0.22	4	0.94	42	0.62
2012-2013	7.7b	0.75	15.9	76	0.22	4	0.95	50	0.85
2012-2013	7.7c	0.94	14.97	60	0.18	4	0.95	43	0.80
2012-2013	7.8a	0.88	13.58	56	0.21	4	0.9	30	0.82
2014-2015	7.5	0.66	53.12	300	0.11	4	0.68	72	0.82
2014-2015	7.8b	0.8	32.22	138	0.05	18	0.16	62	0.68
2014-2015	7.8c	0.83	18.31	84	0.19	4	0.85	42	0.86

Table 7.2: Various properties of the neutral networks of H3N2 influenza haemagglutinin. Here, P_h is the predicted proportion of the population found on the highest degree node and its neighbours. That is, the sum of the components of the principal eigenvector corresponding to the highest degree node and its neighbours. $Y_r(\Lambda)$ is the relative inverse participation ratio (see § 4.2). N is the number of nodes in the network. P_{K_M} is the proportion of the population residing on the maximum K-core of the network (§ 4.1) (Pastor-Satorras & Castellano, 2016). N_{K_M} is the number of nodes in the maximum K-core of the network. P_{Kn_M} is the proportion of the population residing on the maximum K-core and neighbours of the maximum K-core. N_{Kn_M} is the sum of the number of nodes in the maximum K-core and the number of neighbours of nodes in the maximum K-core. S_C is the cosine similarity between the population distribution predicted by the principal eigenvector and the observed population distribution over the neutral network.

CHAPTER 8

DISCUSSION AND FUTURE WORK

In this work, we set out to incorporate and build upon recent results concerning the behaviour of the principal eigenvectors, and associated eigenvalues, of the adjacency matrices of networks in the context of the study of the dynamics of polymorphic populations evolving asexually on neutral networks.

This analysis was divided into two parts, one studying the population distribution on networks with a homogeneous structure and the other focusing on networks with a heterogeneous structure. This separation was made due to the fact that networks with a homogeneous structure are amenable to study by mean-field analysis, whereas heterogeneous networks are not. Moreover, this choice was justified by the behavioural differences found to exist between these cases. Populations evolving on neutral networks with homogeneous structure were found to spread out over the network. With increasing degree variance and assortativity, the population tends to concentrate on the higher degree vertices of the network (chapter 3). However, this concentration is mild, and can be approximated as being proportional to the degree of the vertex and its square (eq. (3.6)). This is contrasted with the heterogeneous networks studied in chapter 4 and chapter 7. Under certain conditions, almost the entire population can become concentrated on a very small region of the network.

Much of the discussion surrounding neutral evolution has functioned on the assumption that there is only a single mode of polymorphic neutral evolution. In this mode of evolution, the population spreads out over the network, gaining

variation and exploring sequence space (Lauring & Andino, 2010). The population will, further, "evolve toward regions denser in neutral genotypes"(Aguirre et al., 2009). This is a good summary of what we found to be the case on networks with a homogeneous topology. However, the behaviour on networks with a heterogeneous structure can be substantially different. In the presence of certain structural heterogeneities, the principal eigenvector of the adjacency matrix of the network localises. This localisation of the principal eigenvector leads to an exploration catastrophe, as described by Ancel & Fontana (2000), whereby the population becomes concentrated on a small region of the network. It is interesting to note that the localisation transition described here is dependent on the topology of the network and is independent of the mutation rate. The independence from mutation rate is due to the fact that the equilibrium distribution of the population over a neutral network is independent of mutation rate (Van Nimwegen et al., 1999). On the other hand, the localisation transition described by Ancel & Fontana (2000), along with other localisation-delocalisation transitions studied in quasispecies theory (Tejero et al., 2011; Summers & Litwin, 2006), are dependent on the mutation rate and are studied in the context of a fixed fitness landscape. Nevertheless, such an error catastrophe has important ramifications for the study of populations evolving at high mutation rates.

Ancel & Fontana (2000) refer to the exploration catastrophe as a "halting of the evolutionary process". This refers to the fact that the population is not exploring genotype space, and, as such, the chances of it finding a phenotype with improved fitness are greatly reduced. The inability of the population to find such a phenotype will result in it being trapped in its current position in the fitness landscape. Ancel & Fontana (2000) further describe the phenomenon of *neutral confinement*, whereby the population evolves a very high level of robustness and is, therefore, exposed to even lower levels of phenotypic variation. They find that this phenomenon occurs in their model of the evolution of RNA with plastogenetic congruence. Neutral confinement may, or may not, occur along with the exploration catastrophes described in this thesis. A population which is localised within a network, due to evolution, will have a higher average robustness than a population spread uniformly over it. This is due to the fact that sub-populations located on the regions of localisation had to out-compete sub-populations located on other areas of the networks. Therefore, localisation does entail an increase in mutational robustness and a decrease in diversity available to the phenotype. However, the proportion of mutations which are non-

neutral which are available to the population is dependent on the total number of mutations available to the genotypes. This is determined by the length of the genetic code, a parameter exogenous to most of the network models studied in this thesis. There are, however, exceptions to this. The neutral networks of influenza analysed in chapter 7 are defined on sequences of a fixed length. These sequences contained over 1500 nucleotides, which implies that they each had over 4500 mutational neighbours. The largest network studied (2009 H1N1) contained 1071 vertices (table 7.1). This means that, in all instances, the majority of mutational neighbours of the populations evolving on these networks would be non-neutral. Therefore, neutral confinement would not occur on these networks. The neutral networks defined within Hamming space studied in § 4.7 and § 4.8 do have an associated sequence length. If they are viewed as being defined on the entire string of the genetic code, then we can question whether neutral confinement occurs. The poorly connected subgraphs of the hypercube explored in § 4.7 were all of low average degree, and so we do not expect to observe neutral confinement. On the other hand, we would expect the populations evolving on the Hamming balls connected to random graphs (§ 4.7) to concentrate towards the center of the Hamming balls. Indeed, this was demonstrated to be the case by Bornberg-Bauer & Chan (1999), who used Hamming balls as an abstraction for the structure of protein folding neutral networks (see also, § 6.2). As the central nodes in Hamming balls only have neutral neighbours, localisation onto them is a candidate for neutral confinement. Howeover, as demonstrated by Bornberg-Bauer & Chan (1999), a substantial proportion of the population can be located on the peripheral vertices of the Hamming ball. These vertices contain only a single neutral neighbour. The possibility of neutral confinement on Hamming balls is, therefore, a topic for further research.

The delocalized case has more in common with the traditional intuition. A principal difference, however, is the level to which concentration on regions of better connected genotypes occurs. Firstly, such a region needs to exist. This requirement will be met in networks with degree assortativity, however, disassortative mixing will result in genotypes with high robustness mutating to those with low robustness, thwarting evolution's attempts at settling on robust nodes. Specifically, by examining equations (3.4), (3.6) and (3.11) we see that, in unassortative networks, the proportion of the population on a given node scales with its degree. This implies that, on relatively homogeneous networks, there will be little difference in the population concentration on various nodes. Further-

more, as shown in equation (3.11), disassortative mixing decreases the number of individuals occupying a node in proportion to both the square and cube of its degree. Although we do expect to see a certain degree of concentration of the population in more robust regions of the neutral network in the case that the network exhibits assortative mixing, the severity of this concentration will be relatively mild. Given that, for networks of reasonably high average degree, $\hat{\lambda}$ is substantially larger than r, the latter terms in equation (3.11) will only play a significant role when k_i is much larger than $\hat{\lambda}$.

We further propose that it is fruitful to think of the delocalized regime of neutral evolution more in terms of a biased sampling process of the genotypes on the network, resulting from mutational biases, as opposed to a population moving between regions of the network. From equations (3.5), (3.7) and (3.12) we can see that, in the absence of assortativity, the population's average robustness is approximated by $\hat{\lambda}$. This is exactly the average robustness which we would expect from performing a random sampling of all possible mutations on the network, as implied by the friendship paradox (Feld, 1991). Assortative and disassorative mixing by degree will increase or decrease the population's average robustness above or below this level. Assortativity represents a further mutational bias towards higher or lower degree nodes, dependent on the degree of the node from which the mutation originates.

These two modes of neutral evolution have important consequences for arguments relating mutational robustness and evolvability. These arguments are predicated on the fact that robust genotypes form larger neutral networks (Wagner, 2011). This then allows for the population to accumulate more cryptic variation as it spreads over the network (Masel & Trotter, 2010), allowing it to better adapt to changes in its environment. Moreover, it creates more "stepping off points" for the population, allowing it to access more phenotypic variation (Wagner, 2008). This line of argumentation hinges on mutational robustness increasing the area of genotype space over which a population is dispersed. Masel & Trotter (2010) state this explicitly, saying that "genetic robustness only promotes evolvability when it is associated with increased spread of a population across genotype space ". Experiments examining the relationship between robustness and evolvability have yielded contradictory results. McBride et al. (2008) found that robust populations of ϕ6 virus were able to evolve a greater resistance to heat shock than brittle populations. On the contrary, Cuevas et al. (2009) found that brittle populations of Vesicular stomatitis virus were better able to

adapt upon introduction to a new cell type.

Stern et al. (2014) proposed a mechanism for resolving these conflicting results. They developed a model for virus evolution whereby the fitness effects of alleles could completely change during an environment shift. They found that the effect of robustness on evolvability was dependent on the fitness of the neutral genotypes upon environment shift. A change in the environment which resulted in the population's neutral variation being deleterious resulted in low adaptability, whereas the converse was true if the neutral variation was advantageous.

The results described in this thesis present a plausible, alternate, explanation for these conflicting experimental results. That is, robustness that occurs as a result of an exploration catastrophe should hinder evolvability. On the other hand, robustness that is due to a population evolving on a homogeneous neutral network, with high average degree, which involves only a slight concentration of the population at higher-degree vertices, should promote evolvability.

By demonstrating that eigenvectors can localise in poorly connected subgraphs of Hamming space and onto Hamming balls in random subgraphs of Hamming space, this thesis makes a small contribution to the study of eigenvector localisation on networks. The majority of work on the localisation of the principal eigenvector has focused on localisation on hub vertices (Martin et al., 2014; Chung et al., 2003; Goltsev et al., 2012). That is, vertices whose degree is substantially higher than the average degree of the network (see § 4.1). Pastor-Satorras & Castellano (2016) demonstrated that an alternative mode of localisation does exist, by showing that the eigenvector can localise on a network's maximum K-core. They further suggested that it was possible that there might be other modes of localisation. We have demonstrated two such modes here.

Discussion of these modes of localisation suggests the first shortcomming of this thesis. Although the topology of a Hamming ball connected to a random graph (§ 4.7) is well defined, the phenomenon of localisation due to weak connectivity discussed in § 4.8 is ill-defined. There are many metrics for connectivity (Costa et al., 2007). However, after a limited search, we were unable to find one which would be able to capture the limited connectivity of both the subgraphs of Hamming space and the example given of Erdős-Renyi networks connected by a single edge. For instance, the connected Erdős-Renyi networks are highly modular (Newman, 2006), however, the subgraphs of Hamming space are less so. Finding a metric for connectivity which describes this mode of local-

isation and further exploring instances in which it may occur is an avenue of future research. It is also possible that these two instances represent different modes of localisation.

A further flaw in this work is that it deals with the heterogeneous and homogeneous network cases, and the associated diffusion and exploration catastrophe cases, as being well-separated, distinct instances. In one instance this was observed to be the case (fig. 4.5). However, in all other instances, the transition was somewhat gradual. Moreover, most of the neutral networks of influenza haemagglutinin studied here had a relative inverse participation ratio below the threshold for localisation used in this work. However, for all of these networks, this value was well above one, indicating that the population was not evenly distributed over the network. In this transition phase, the population distribution is probably not well characterised by the expressions derived in § 3.2, although this is a matter for further investigation. Characterising the population in these transition phases would likely be challenging. Furthermore, as Pastor-Satorras & Castellano (2016) point out, there does not exist a non-arbitrary definition for localisation of the eigenvector in single network instances. Finding such a definition is a further direction of future work.

Probably the largest shortcomming of this thesis is that much of the analysis of localisation in network models was performed on models which are not embedded in Hamming space. This work is still valuable, as previous work on similar models reported on the *weight* of the eigenvector confined to a given region, whereas, for evolution, we are interested in the size of the eigenvector components, normalised by the sum of the vector componenets (l_1 norm), in a given region (see chapter 4). Moreover, we specifically investigated some network models which were motivated by observed variation networks in nature. Studying networks embedded in Hamming space is challenging. For a start, network libraries for programming languages do not have functionality for generating random instances of these networks. Moreover, the author is unaware of an algorithm which would generate random instances of networks embedded in Hamming space with a tunable level of degree assortativity. Although we were able to confirm that the approximations derived in chapter 4 are fairly accurate for networks embedded in Hamming space, in the case that they had a high average degree and were unassortative, verifying that they apply to networks which exhibit degree assortativity and disassortativity is a topic for future study. This will require the development of an algorithm capable of producing such net-

works. Moreover, the greatest challenge found when studying these networks was that, even for reasonably sized alphabets and strings, the size of the generated networks blew up very rapidly. This was especially true for the generation of Hamming balls. Not only did it render the generation of interpretable network diagrams impossible, but this size blow-up caused the computer on which the analysis was running to run out of RAM and resulted in the calculation of the principal eigenvectors becoming infeasible. Finding methods for analising networks embedded in Hamming space, across a broad range of parameters, is a pertinent topic for further study.

The work done here on the neutral networks of influenza haemagglutinin (chapter 7) suffered from two drawbacks. The first being that, although it is likely that the networks analysed contained genotypes of equal fitness, this is not guaranteed. Secondly, the sequencing techniques on which the data set was based under sample low frequency variants. Acevedo et al. (2014) recently published a study where they used new sequencing techniques, which do not have such an under sampling bias, to study the poliovirus mutational landscape. Moreover, they explicitly calculated the fitness values of mutations. Much of the data collected in this work is available online. Using this data set to investigate the neutral networks of this virus represents an exciting possible direction of research.

Given that it is suspected that much of evolution occurs on neutral networks (Nei, 2005) along with the importance of mutational robustness to the survival of organisms and its relationship with evolvability, understanding the impact of the topology of neutral networks on the dynamics of neutral evolution and the resulting robustness of organisms is of great importance. This work has provided insight into these issues in the case of polymorphic populations: large populations evolving at high mutation rate. The directed, neutral, evolution of biomolecules (Currin et al., 2015; Jäckel & Hilvert, 2010) along with viruses overcoming immunity through neutral evolution (van Nimwegen, 2006) fall within this category. These results have potential applicability to these problems. For instance, the neutral evolution of large libraries of molecules (Kaltenbach & Tokuriki, 2014) will be greatly aided by delocalization, whereas a virus's attempt to escape immunity might be thwarted if its population localizes on a hub.

CHAPTER 9

CONCLUSION

This thesis investigated the manner in which neutral network topology influences the resulting population distribution and robustness during neutral evolution at high mutation rates in large populations without recombination. In such cases, the population distribution is given by the principal eigenvector of the adjacency matrix of the neutral network and, similarly, the average mutational robustness of the individuals in the population is given by the principal eigenvalue (Van Nimwegen et al., 1999). Hence, we utilized, and built upon, recent results concerning the behaviour of these values from studies concerning the spread of epidemics on networks(Goltsev et al., 2012) as well as more general work (Martin et al., 2014).

It was found that, on homogeneous neutral networks, the population's behaviour could be described in terms of mutational biases. For unassortative neutral networks, it was found that the average mutational robustness was equal to the sampling bias provided by the friendship paradox (Feld, 1991). Assortative and disassortative mixing by degree raised the robustness above or below this value, respectively. Furthermore, in the process of demonstrating this, we derived a new approximation for the principal eigenvalue of a network in terms of its assortativity and the moments of its degree distribution.

Conversely, for neutral networks with certain structural heterogeneities, it was found that the population could undergo an exploration catastrophe, whereby it becomes localised on a small number of nodes in the network. These results

are particularly relevant to various arguments concerning the relationship between robustness and evolvability (Masel & Trotter, 2010; Wagner, 2008), which make the assumption that populations evolving at high mutation rate disperse over their neutral networks.

These results are relevant to the directed evolution of bio-molecules (Currin et al., 2015; Jäckel & Hilvert, 2010), where they can be used to evolve more robust molecules as well as facilitate the evolution of greater variety. Moreover, they can also further our understanding of the factors that allow viruses to escape immunity along neutral networks (van Nimwegen, 2006).

Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: http://hpc.uct.ac.za

BIBLIOGRAPHY

Acevedo, A., Brodsky, L., & Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, *505*(7485), 686–690.

Aguirre, J., Buldú, J. M., & Manrubia, S. C. (2009). Evolutionary dynamics on networks of selectively neutral genotypes: Effects of topology and sequence stability. *Physical Review E*, *80*(6), 066112.

Aguirre, J., Buldú, J. M., Stich, M., & Manrubia, S. C. (2011). Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLOS One*, *6*(10), e26324.

Aita, T. (2008). Hierarchical distribution of ascending slopes, nearly neutral networks, highlands, and local optima at the dth order in an NK fitness landscape. *Journal of Theoretical Biology*, *254*(2), 252–263.

Ancel, L. W. & Fontana, W. (2000). Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology*, *288*(3), 242–283.

Anderson, P. W. (1958). Absence of diffusion in certain random lattices. *Physical Review*, *109*(5), 1492.

Andino, R. & Domingo, E. (2015). Viral quasispecies. *Virology*, *479*, 46–51.

Bäck, T., Fogel, D. B., & Michalewicz, Z. (1997). Handbook of evolutionary computation. *New York: Oxford*.

Banzhaf, W. (1994). Genotype-phenotype-mapping and neutral variation—a case study in genetic programming. *Parallel Problem Solving From Nature—PPSN III*, 322–332.

Banzhaf, W. & Langdon, W. B. (2002). Some considerations on the reason for bloat. *Genetic Programming And Evolvable Machines*, *3*(1), 81–91.

Banzhaf, W. & Leier, A. (2006). Evolution on neutral networks in genetic programming. *Genetic Programming Theory And Practice III*, 207–221.

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic Programming: an Introduction*, volume 1. Morgan Kaufmann San Francisco.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., & Lipman, D. (2008). The influenza virus resource at the national center for biotechnology information. *Journal of Virology*, *82*(2), 596–601.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Barabasi, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113.

Barnett, L. (1998). Ruggedness and neutrality: The NKp family of fitness landscapes. In *Artificial Life VI: Proceedings of the Sixth International Conference on Artificial Life*, (pp. 18–27).

Barnett, L. (2003). *Evolutionary Search on Fitness Landscapes with Neutral Networks*. PhD thesis, University of Sussex.

Bascompte, J. (2010). Structure and dynamics of ecological networks. *Science*, *329*(5993), 765–766.

Basseur, M. & Goëffon, A. (2015). Climbing combinatorial fitness landscapes. *Applied Soft Computing*, *30*, 688–704.

Bastolla, U., Vendruscolo, M., & Roman, H. E. (2000). Structurally constrained protein evolution: results from a lattice simulation. *The European Physical Journal B-Condensed Matter And Complex Systems*, *15*(2), 385–397.

Beaudoin, W., Verel, S., Collard, P., & Escazut, C. (2006). Deceptiveness and neutrality of the ND family of fitness landscapes. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, (pp. 507–514). ACM.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, *35*(8), 1798–1828.

Bierwirth, C., Mattfeld, D. C., & Watson, J.-P. (2004). Landscape regularity and random walks for the job-shop scheduling problem. In *European Conference on Evolutionary Computation in Combinatorial Optimization*, (pp. 21–30). Springer.

Blackburne, B. P. & Hirst, J. D. (2005). Population dynamics simulations of functional model proteins. *The Journal of Chemical Physics*, *123*(15), 154907.

Blickle, T. & Thiele, L. (1994). Genetic programming and redundancy. *Choice*, *1000*, 2.

Bloom, J. D., Lu, Z., Chen, D., Raval, A., Venturelli, O. S., & Arnold, F. H. (2007). Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology*, *5*(1), 1.

Bloom, J. D., Raval, A., & Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. *Genetics*, *175*(1), 255–266.

Blum, C. & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, *35*(3), 268–308.

Boldhaus, G. & Klemm, K. (2010). Regulatory networks and connected components of the neutral space. *The European Physical Journal B-Condensed Matter And Complex Systems*, *77*(2), 233–237.

Bollobás, B., Lee, J., & Letzter, S. (2016). Eigenvalues of subgraphs of the cube. *Arxiv Preprint Arxiv:1605.06360*.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*(1), 113–120.

Boni, M. F. (2008). Vaccination and antigenic drift in influenza. *Vaccine*, *26*, C8–C14.

Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophysical Journal*, *73*(5), 2393.

Bornberg-Bauer, E. & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences*, *96*(19), 10689–10694.

Brameier, M. & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions On Evolutionary Computation*, *5*(1), 17–26.

Choi, S.-S., Jung, K., & Kim, J. H. (2008). Phase transition in a random NK landscape model. *Artificial Intelligence*, *172*(2-3), 179–203.

Chung, F., Lu, L., & Vu, V. (2003). Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, *100*(11), 6313–6318.

Ciliberti, S., Martin, O. C., & Wagner, A. (2007a). Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences*, *104*(34), 13591–13596.

Ciliberti, S., Martin, O. C., & Wagner, A. (2007b). Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLOS Comput Biol*, *3*(2), e15.

Cioabă, S. M., Van Dam, E. R., Koolen, J. H., & Lee, J.-H. (2010). A lower bound for the spectral radius of graphs with fixed diameter. *European Journal of Combinatorics*, *31*(6), 1560–1566.

Cohen, R., Havlin, S., & Ben-Avraham, D. (2003). Structural properties of scale-free networks. *Handbook of Graphs And Networks*.

Costa, L. d. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances In Physics*, *56*(1), 167–242.

Cotta, C. & Fernandez, A. J. (2005). Analyzing fitness landscapes for the optimal golomb ruler problem. In G. R. Raidl, J. Gottlieb, et al. (Eds.), *Evolutionary computation in combinatorial optimization*, volume 3448 of *Lecture Notes in Computer Science* (pp. 68–79). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cotterell, J. & Sharpe, J. (2013). Mechanistic explanations for restricted evolutionary paths that emerge from gene regulatory networks. *PLOS One*, *8*(4), e61178.

Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L., & Meyers, L. A. (2008). The ascent of the abundant: how mutational networks constrain evolution. *PLOS Comput Biol*, *4*(7), e1000110.

Crombach, A. & Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLOS Comput Biol*, *4*(7), e1000112.

Crutchfield, J. P. & Schuster, P. (2003). *Evolutionary Dynamics: Exploring The Interplay of Selection, Accident, Neutrality, and Function*. Oxford University Press, USA.

Csardi, G. & Nepusz, T. (2006). The igraph software package for complex network research. *Interjournal*, *Complex Systems*, 1695.

Cuevas, J., Moya, A., & Sanjuán, R. (2009). A genetic background with low mutational robustness is associated with increased adaptability to a novel host in an RNA virus. *Journal of Evolutionary Biology*, *22*(10), 2041–2048.

Currin, A., Swainston, N., Day, P. J., & Kell, D. B. (2015). Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chemical Society Reviews*, *44*(5), 1172–1239.

Dall'Olio, G. M., Bertranpetit, J., Wagner, A., & Laayouni, H. (2014). Human genome variation and the concept of genotype networks. *PLOS One*, *9*(6), e99424.

De Visser, J. A. G. & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, *15*(7), 480–490.

de Visser, J. A. G. M., Hermisson, J., Wagner, G. P., Meyers, L. A., Bagheri-Chaichian, H., Blanchard, J. L., Chao, L., Cheverud, J. M., Elena, S. F., Fontana, W., et al. (2003). Perspective: evolution and detection of genetic robustness. *Evolution*, *57*(9), 1959–1972.

Devert, A. (2009). When and why development is needed: generative and developmental systems. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, (pp. 1843–1844). ACM.

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral quasispecies evolution. *Microbiology And Molecular Biology Reviews*, *76*(2), 159–216.

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, *80*(4), 1275.

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2006). K-core organization of complex networks. *Physical Review Letters*, *96*(4), 040601.

Draper, D. (1992). The RNA-folding problem. *Acc. Chem. Res.*, *25(4)*, 201–207.

Ebner, M. (1999). On the search space of genetic programming and its relation to nature's search space. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 2, (pp. 1357–1361). IEEE.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Eiben, A. & Smith, J. (2015). *Introduction to Evolutionary Computing*, volume 53. Springer.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, *58*(10), 465–523.

Elena, S. F. & Sanjuán, R. (2008). The effect of genetic robustness on evolvability in digital organisms. *BMC Evolutionary Biology*, *8*(1), 284.

Erdős, P. & Renyi, A. (1959). On random graphs I. *Publ. Math. Debrecen*, *6*, 290–297.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, *96*(6), 1464–1477.

Fernández, P. & Solé, R. V. (2007). Neutral fitness landscapes in signalling networks. *Journal of the Royal Society Interface*, *4*(12), 41–47.

Ferreira, C. (2002). Genetic representation and genetic neutrality in gene expression programming. *Advances In Complex Systems*, *5*(04), 389–408.

Fisher, H. & Thompson, G. L. (1963). Probabilistic learning combinations of local job-shop scheduling rules. *Industrial Scheduling*, *3*(2), 225–251.

Fontana, W., Konings, D. A., Stadler, P. F., & Schuster, P. (1993). Statistics of RNA secondary structures. *Biopolymers*, *33*(9), 1389–1404.

Fontana, W. & Schuster, P. (1998). Continuity in evolution: on the nature of transitions. *Science*, *280*(5368), 1451–1455.

Fruchterman, T. M. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice And Experience*, *21*(11), 1129–1164.

Galván-López, E., Dignum, S., & Poli, R. (2008). The effects of constant neutrality on performance and problem hardness in GP. In *European Conference on Genetic Programming*, (pp. 312–324). Springer.

Galván-López, E., Poli, R., Kattan, A., Neill, M., & Brabazon, A. (2011). Neutrality in evolutionary algorithms...what do we know? *Evolving Systems*, *2*(3), 145–163.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends In Ecology & Evolution*, *12*(8), 307–312.

Geard, N., Wiles, J., Hallinan, J., Tonkes, B., & Skellett, B. (2002). A comparison of neutral landscapes-NK, NKp and NKq. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, (pp. 205–210). IEEE.

Gjuvsland, A., Vik, J., Beard, D., Hunter, P., & Omholt, S. (2013). Bridging the genotype-phenotype gap: what does it take? *J Physiol.*, *591(8):*, 2055–2066.

Gleeson, J. P., Melnik, S., Ward, J. A., Porter, M. A., & Mucha, P. J. (2012). Accuracy of mean-field theory for dynamics on real-world networks. *Physical Review E*, *85*(2), 026106.

Goltsev, A. V., Dorogovtsev, S. N., Oliveira, J., & Mendes, J. F. (2012). Localization and spreading of diseases in complex networks. *Physical Review Letters*, *109*(12), 128702.

Grafen, A. (2008). *The Evolutionary Dynamics of Neutral Networks: Lessons from RNA*. PhD thesis, University of Oxford.

Greenbury, S. F., Schaper, S., Ahnert, S. E., & Louis, A. A. (2016). Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLOS Computational Biology*, *12*(3), e1004773.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., & Schuster, P. (1996). Analysis of RNA sequence structure maps by exhaustive enumeration II. Structures of neutral networks and shape space covering. *Monatshefte Für Chemie/chemical Monthly*, *127*(4), 375–389.

Guarnaccia, T., Carolan, L. A., Maurer-Stroh, S., Lee, R. T., Job, E., Reading, P. C., Petrie, S., McCaw, J. M., McVernon, J., Hurt, A. C., et al. (2013). Antigenic drift of the pandemic 2009 A (H1N1) influenza virus in a ferret model. *PLOS Pathogens*, *9*(5), e1003354.

Hartmann, M. & Haddow, P. C. (2004). Evolution of fault-tolerant and noise-robust digital designs. *IEEE Proceedings-Computers And Digital Techniques*, *151*(4), 287–294.

Harvey, I. & Thompson, A. (1997). Through the labyrinth evolution finds a way: A silicon ridge. *Evolvable Systems: From Biology To Hardware*, 406–422.

Hosseini, S.-R., Barve, A., & Wagner, A. (2015). Exhaustive analysis of a genotype space comprising 10^{15} central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLOS Comput Biol*, *11*(8), e1004329.

Hu, T., Banzhaf, W., & Moore, J. H. (2014). Population exploration on genotype networks in genetic programming. In *International Conference on Parallel Problem Solving from Nature*, (pp. 424–433). Springer.

Hu, T., Payne, J., Banzhaf, W., & Moore, J. (2011). Robustness, evolvability, and accessibility in linear genetic programming. *Genetic Programming*, 13–24.

Hu, T., Payne, J. L., Banzhaf, W., & Moore, J. H. (2012). Evolutionary dynamics on multiple scales: a quantitative analysis of the interplay between genotype, phenotype, and fitness in linear genetic programming. *Genetic Programming And Evolvable Machines*, *13*(3), 305–337.

Jäckel, C. & Hilvert, D. (2010). Biocatalysts by evolution. *Current Opinion In Biotechnology*, *21*(6), 753–759.

Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*(6833), 41–42.

Jörg, T., Martin, O. C., & Wagner, A. (2008). Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics*, *9*(1), 464.

Kaltenbach, M. & Tokuriki, N. (2014). Generation of effective libraries by neutral drift. *Directed Evolution Library Creation: Methods And Protocols*, 69–81.

Kauffman, S. & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, *128*(1), 11–45.

Kauffman, S. A. & Weinberger, E. D. (1989). The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, *141*(2), 211–245.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kimura, M. et al. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624–626.

King, J. L. & Jukes, T. H. (1969). Non-darwinian evolution. *Science*, *164*(3881), 788–798.

Kirkpatrick, S. & Toulouse, G. (1985). Configuration space analysis of travelling salesman problems. *Journal De Physique*, *46*(8), 1277–1292.

Koelle, K., Cobey, S., Grenfell, B., & Pascual, M. (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science*, *314*(5807), 1898–1903.

Langdon, W. B. (2015). Genetically improved software. In *Handbook of Genetic Programming Applications* (pp. 181–220). Springer.

Langdon, W. B. & Harman, M. (2015). Optimizing existing software with genetic programming. *IEEE Transactions On Evolutionary Computation*, *19*(1), 118–135.

Langdon, W. B. & Ochoa, G. (2016). Genetic improvement: A key challenge for evolutionary computation. In *Evolutionary Computation (CEC), 2016 IEEE Congress on*, (pp. 3068–3075). IEEE.

Langdon, W. B. & Petke, J. (2017). Software is not fragile. In *First Complex Systems Digital Campus World E-Conference 2015*, (pp. 203–211). Springer.

Langdon, W. B. & Poli, R. (1998a). Fitness causes bloat. In *Soft Computing in Engineering Design and Manufacturing* (pp. 13–22). Springer.

Langdon, W. B. & Poli, R. (1998b). Fitness causes bloat: Mutation. In *European Conference on Genetic Programming*, (pp. 37–48). Springer.

Langdon, W. B. & Poli, R. (2013). *Foundations of Genetic Programming*. Springer Science & Business Media.

Langdon, W. B., Soule, T., Poli, R., & Foster, J. A. (1999). The evolution of size and shape. *Advances In Genetic Programming*, *3*, 163–190.

Lapedes, A. & Farber, R. (2001). The geometry of shape space: application to influenza. *Journal of Theoretical Biology*, *212*(1), 57–69.

Lauring, A. S. & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLOS Pathogens*, *6*(7), e1001005.

Le Goues, C., Dewey-Vogt, M., Forrest, S., & Weimer, W. (2012). A systematic study of automated program repair: Fixing 55 out of 105 bugs for $8 each. In *Software Engineering (ICSE), 2012 34th International Conference on*, (pp. 3–13). IEEE.

Lehoucq, R. B. & Sorensen, D. C. (1996). Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal On Matrix Analysis And Applications*, *17*(4), 789–821.

Lehoucq, R. B., Sorensen, D. C., & Yang, C. (1998). *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM.

Liu, M., Zhao, X., Hua, S., Du, X., Peng, Y., Li, X., Lan, Y., Wang, D., Wu, A., Shu, Y., et al. (2015). Antigenic patterns and evolution of the human influenza A (H1N1) virus. *Scientific Reports*, *5*.

Lobo, J., Miller, J. H., & Fontana, W. (2004). Neutrality in technological landscapes. *Sante Fe Institute Working Paper*.

Lovász, L. (1993). Random walks on graphs. *Combinatorics, Paul Erdos Is Eighty*, *2*, 1–46.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, (pp. 1150–1157). IEEE.

Luke, S. & Panait, L. (2006). A comparison of bloat control methods for genetic programming. *Evolutionary Computation*, *14*(3), 309–344.

Łuksza, M. & Lässig, M. (2014). A predictive fitness model for influenza. *Nature*, *507*(7490), 57–61.

Luo, J. X. & Turner, M. S. (2011). Functionality and metagraph disintegration in boolean networks. *Journal of Theoretical Biology*, *282*(1), 65–70.

Manrubia, S. C. & Cuesta, J. A. (2010). Neutral networks of genotypes: Evolution behind the curtain. *Arxiv Preprint Arxiv:1002.2745*.

Marmion, M.-E., Blot, A., Jourdan, L., & Dhaenens, C. (2013). Neutrality in the graph coloring problem. In *International Conference on Learning and Intelligent Optimization*, (pp. 125–130). Springer.

Marmion, M.-E., Dhaenens, C., Jourdan, L., Liefooghe, A., & Verel, S. (2011). On the neutrality of flowshop scheduling fitness landscapes. In *International Conference on Learning and Intelligent Optimization*, (pp. 238–252). Springer.

Martin, T., Zhang, X., & Newman, M. (2014). Localization and centrality in networks. *Physical Review E*, *90*(5), 052808.

Masel, J. & Trotter, M. V. (2010). Robustness and evolvability. *Trends In Genetics*, *26*(9), 406–414.

Matsuura, T. & Yomo, T. (2006). In vitro evolution of proteins. *Journal of Bioscience And Bioengineering*, *101*(6), 449–456.

Mattfeld, D. C., Bierwirth, C., & Kopfer, H. (1999). A search space analysis of the job shop scheduling problem. *Annals of Operations Research*, *86*, 441–453.

Mayo, M., Abdelzaher, A., & Ghosh, P. (2015). Long-range degree correlations in complex networks. *Computational Social Networks*, *2*(1), 1.

McBride, R. C., Ogbunugafor, C. B., & Turner, P. E. (2008). Robustness promotes evolvability of thermotolerance in an RNA virus. *BMC Evolutionary Biology*, *8*(1), 231.

Mellor, R. E. (2007). Investigating Kauffman's NK model for agent-based modelling.

Merz, P. & Freisleben, B. (2000). Fitness landscape analysis and memetic algorithms for the quadratic assignment problem. *IEEE Transactions On Evolutionary Computation*, *4*(4), 337–352.

Milano, N. & Nolfi, S. (2016). Robustness to faults promotes evolvability: Insights from evolving digital circuits. *PLOS One, 11*(7), e0158627.

Miller, J. F. & Smith, S. L. (2006). Redundancy and computational efficiency in cartesian genetic programming. *IEEE Transactions On Evolutionary Computation, 10*(2), 167–174.

Mitchell, M., Crutchfield, J. P., Das, R., et al. (1996). Evolving cellular automata with genetic algorithms: A review of recent work. In *Proceedings of the First International Conference on Evolutionary Computation and Its Applications (EvCA'96)*. Moscow.

Nei, M. (2005). Selectionism and neutralism in molecular evolution. *Molecular Biology And Evolution, 22*(12), 2318–2342.

Newman, M. E. (2002). Assortative mixing in networks. *Physical Review Letters, 89*(20), 208701.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E, 67*(2), 026126.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577–8582.

Newman, M. E. & Engelhardt, R. (1998). Effects of selective neutrality on the evolution of molecular species. *Proceedings of the Royal Society of London B: Biological Sciences, 265*(1403), 1333–1338.

Nielsen, S. S., Danoy, G., Bouvry, P., & Talbi, E.-G. (2015). Nk landscape instances mimicking the protein inverse folding problem towards future benchmarks. In *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, (pp. 915–921). ACM.

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence*. Cambridge University Press.

Noirel, J. & Simonson, T. (2008). Neutral evolution of proteins: The superfunnel in sequence space and its relation to mutational robustness. *The Journal of Chemical Physics, 129*(18), 185104.

Nordin, P., Francone, F., & Banzhaf, W. (1995). Explicitly defined introns and destructive crossover in genetic programming. *Advances In Genetic Programming, 2*, 111–134.

Parter, M., Kashtan, N., & Alon, U. (2008). Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLOS Comput Biol, 4*(11), e1000206.

Pastor-Satorras, R. & Castellano, C. (2016). Distinct types of eigenvector localization in networks. *Scientific Reports, 6*.

Payne, J. L., Moore, J. H., & Wagner, A. (2014). Robustness, evolvability, and the logic of genetic regulation. *Artificial Life, 20*(1), 111–126.

Payne, J. L. & Wagner, A. (2013). Constraint and contingency in multifunctional gene regulatory circuits. *PLOS Comput Biol, 9*(6), e1003071.

Pechenick, D. A., Payne, J. L., & Moore, J. H. (2012). The influence of assortativity on the robustness of signal-integration logic in gene regulatory networks. *Journal of Theoretical Biology, 296*, 21–32.

Pigliucci, M. (2010). Genotype–phenotype mapping and the end of the 'genes as blueprint' metaphor. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 365*(1540), 557–566.

Pitzer, E. & Affenzeller, M. (2012). A comprehensive survey on fitness landscape analysis. In *Recent Advances in Intelligent Engineering Systems* (pp. 161–191). Springer.

Podgornaia, A. I. & Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science, 347*(6222), 673–677.

Poli, R. (2003). A simple but theoretically-motivated method to control bloat in genetic programming. In *European Conference on Genetic Programming*, (pp. 204–217). Springer.

Poli, R., Langdon, W. B., McPhee, N. F., & Koza, J. R. (2008). *A Field Guide to Genetic Programming*. Lulu. com.

Poon, L. L., Song, T., Rosenfeld, R., Lin, X., Rogers, M. B., Zhou, B., Sebra, R., Halpin, R. A., Guan, Y., Twaddle, A., et al. (2016). Quantifying influenza virus diversity and transmission in humans. *Nature Genetics, 48*(2), 195–200.

Raman, K. & Wagner, A. (2010). The evolvability of programmable hardware. *Journal of the Royal Society Interface*, rsif20100212.

Raman, K. & Wagner, A. (2011). Evolvability and robustness in a complex signalling circuit. *Molecular Biosystems*, *7*(4), 1081–1092.

Ray, J., Pinar, A., & Seshadhri, C. (2014). A stopping criterion for markov chains when generating independent random graphs. *Journal of Complex Networks*, cnu041.

Reeves, C. R. (1999). Landscapes, operators and heuristic search. *Annals of Operations Research*, *86*, 473–490.

Reeves, T., Farr, R., Blundell, J., Gallagher, A., & Fink, T. (2016). Eigenvalues of neutral networks: interpolating between hypercubes. *Discrete Mathematics*, *339*(4), 1283–1290.

Reidys, C., Stadler, P. F., & Schuster, P. (1997). Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bulletin of Mathematical Biology*, *59*(2), 339–397.

Reidys, C. M. (2009). Large components in random induced subgraphs of n-cubes. *Discrete Mathematics*, *309*(10), 3113–3124.

Reidys, C. M. & Stadler, P. F. (2001). Neutrality in fitness landscapes. *Applied Mathematics And Computation*, *117*(2), 321–350.

Reidys, C. M. & Stadler, P. F. (2002). Combinatorial landscapes. *SIAM Review*, *44*(1), 3–54.

Restrepo, J. G., Ott, E., & Hunt, B. R. (2007). Approximating the largest eigenvalue of network adjacency matrices. *Physical Review E*, *76*(5), 056119.

Richter, H. (2014). Fitness landscapes: From evolutionary biology to evolutionary computation. In *Recent Advances in the Theory and Application of Fitness Landscapes* (pp. 3–31). Springer.

Rodrigues, J. F. M. & Wagner, A. (2011). Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Systems Biology*, *5*(1), 39.

Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., et al. (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, *320*(5874), 340–346.

Samal, A., Rodrigues, J. F. M., Jost, J., Martin, O. C., & Wagner, A. (2010). Genotype networks in metabolic reaction spaces. *BMC Systems Biology*, *4*(1), 30.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., et al. (2012). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *40*(D1), D13–D25.

Schaper, S. & Louis, A. A. (2014). The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PLOS One*, *9*(2), e86635.

Schlitt, T. & Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, *8*(6), S9.

Schulte, E. (2015). *Neutral Networks of Real-World Programs and their Application to Automated Software Evolution*. PhD thesis.

Schulte, E., Dorn, J., Harding, S., Forrest, S., & Weimer, W. (2014). Post-compiler software optimization for reducing energy. In *ACM SIGARCH Computer Architecture News*, volume 42, (pp. 639–652). ACM.

Schulte, E., Fry, Z. P., Fast, E., Weimer, W., & Forrest, S. (2014). Software mutational robustness. *Genetic Programming And Evolvable Machines*, *15*(3), 281–312.

Schuster, P., Fontana, W., Stadler, P. F., & Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London B: Biological Sciences*, *255*(1344), 279–284.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*(3), 269–287.

Sikosek, T. & Chan, H. S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface*, *11*(100), 20140419.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35–43.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., & Fouchier, R. A. (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science*, *305*(5682), 371–376.

Smith, J. M. & Szathmary, E. (1997). *The Major Transitions in Evolution*. Oxford University Press.

Smith, T., Husbands, P., & O'Shea, M. (2002). Fitness landscapes and evolvability. *Evolutionary Computation*, *10*(1), 1–34.

Sporns, O. (2010). *Networks of the Brain*. MIT press.

Stadler, P. F. (2002). Fitness landscapes. In *Biological Evolution and Statistical Physics* (pp. 183–204). Springer.

Stadler, P. F. & Schnabl, W. (1992). The landscape of the traveling salesman problem. *Physics Letters A*, *161*(4), 337–344.

Stanley, K. O. & Miikkulainen, R. (2003). A taxonomy for artificial embryogeny. *Artificial Life*, *9*(2), 93–130.

Stern, A., Bianco, S., Te Yeh, M., Wright, C., Butcher, K., Tang, C., Nielsen, R., & Andino, R. (2014). Costs and benefits of mutational robustness in RNA viruses. *Cell Reports*, *8*(4), 1026–1036.

Summers, J. & Litwin, S. (2006). Examining the theory of error catastrophe. *Journal of Virology*, *80*(1), 20–26.

Tannenbaum, E. & Shakhnovich, E. I. (2004). Solution of the quasispecies model for an arbitrary gene network. *Physical Review E*, *70*(2), 021903.

Taverna, D. M. & Goldstein, R. A. (2002). Why are proteins so robust to site mutations? *Journal of Molecular Biology*, *315*(3), 479–484.

Tejero, H., Marín, A., & Montero, F. (2011). The relationship between the error catastrophe, survival of the flattest, and natural selection. *BMC Evolutionary Biology*, *11*(1), 2.

Tewawong, N., Prachayangprecha, S., Vichiwattana, P., Korkong, S., Klinfueng, S., Vongpunsawad, S., Thongmee, T., Theamboonlers, A., & Poovorawan, Y. (2015). Assessing antigenic drift of seasonal influenza A (H3N2) and A (H1N1) pdm09 viruses. *PLOS One*, *10*(10), e0139958.

Thompson, A. & Layzell, P. (2000). Evolution of robustness in an electronics design. In *International Conference on Evolvable Systems*, (pp. 218–228). Springer.

Tiana, G., Broglia, R., & Shakhnovich, E. (2001). Energy profile of the space of model protein sequences. *Journal of Biological Physics*, *27*(2), 147–159.

van Nimwegen, E. (2006). Influenza escapes immunity along neutral networks. *Science*, *314*(5807), 1884–1886.

Van Nimwegen, E., Crutchfield, J., & Huynen, M. (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, *96*(17), 9716–9720.

Vanneschi, L., Pirola, Y., & Collard, P. (2006). A quantitative study of neutrality in GP boolean landscapes. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, (pp. 895–902). ACM.

Verel, S., Collard, P., Tomassini, M., & Vanneschi, L. (2006). Neutral fitness landscape in the cellular automata majority problem. In *International Conference on Cellular Automata*, (pp. 258–267). Springer.

Verel, S., Collard, P., Tomassini, M., & Vanneschi, L. (2007). Fitness landscape of the cellular automata majority problem: view from the "olympus". *Theoretical Computer Science*, *378*(1), 54–77.

Wagner, A. (2008). Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1630), 91–100.

Wagner, A. (2011). *The origins of evolutionary innovations: a theory of transformative change in living systems*. OUP Oxford.

Wagner, A. (2014). A genotype network reveals homoplastic cycles of convergent evolution in influenza a (h3n2) haemagglutinin. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1786), 20132763.

Weinberger, E. D. & Fassberg, A. (1996). NP completeness of kauffman's N-k model.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, *412*(6844), 331–333.

Wright, A. H., Thompson, R. K., & Zhang, J. (2000). The computational complexity of NK fitness functions. *IEEE Transactions On Evolutionary Computation*, *4*(4), 373–379.

Wroe, R., Bornberg-Bauer, E., & Chan, H. S. (2005). Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. *Biophysical Journal*, *88*(1), 118–131.

Wuchty, S., Ravasz, E., & Barabási, A.-L. (2006). The architecture of biological networks. *Complex Systems Science In Biomedicine*, 165–181.

Xia, Y. & Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Current Opinion In Structural Biology*, *14*(2), 202–207.

Xulvi-Brunet, R. & Sokolov, I. M. (2004). Reshuffling scale-free networks: From random to assortative. *Physical Review E*, *70*(6), 066102.

Yu, G. (2013). *Industrial Applications of Combinatorial Optimization*, volume 16. Springer Science & Business Media.

Yu, T. & Miller, J. (2001). Neutrality and the evolvability of boolean function landscape. In *European Conference on Genetic Programming*, (pp. 204–217). Springer.

Zhao, J., Yu, H., Luo, J., Cao, Z., & Li, Y. (2006). Complex networks theory for analyzing metabolic networks. *Chinese Science Bulletin*, *51*(13), 1529–1537.

Zuckerman, E. W. & Jost, J. T. (2001). What makes you think you're so popular? self-evaluation maintenance and the subjective side of the "friendship paradox". *Social Psychology Quarterly*, 207–223.