Flinders University

Honours Thesis

Investigating Hamilton cycles using logistic regression

Author:
Alex Newcombe

Supervisor:
Prof. Jerzy Filar

A thesis submitted in fulfilment of the requirements
for the degree of Bachelor of Science (Hons)

in the

School of Computer Science, Engineering and Mathematics

July 2015

Declaration of Authorship

I, Alex Newcombe declare that this thesis titled, 'Investigating Hamilton cycles using logistic regression' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

FLINDERS UNIVERSITY

Abstract

School of Computer Science, Engineering and Mathematics

Bachelor of Science (Hons)

Investigating Hamilton cycles using logistic regression

by Alex Newcombe

This thesis is an empirical study into the Hamilton cycle problem for cubic graphs. It constructs statistical models on particular explanatory variables which are computed on a random sample of cubic graphs. The explanatory variables are functions and descriptors of graphs taken from various notions in graph theory and were chosen because they demonstrated sensitivity to the underlying structure of the cubic graphs. The models are used to identify so called non-bridge, non-Hamiltonian graphs out of the population of all 2 and 3-connected cubic graphs of the same order. These non-bridge, non-Hamiltonian graphs are commonly accepted as being the difficult part of the Hamilton cycle problem for cubic graphs. The result is a technique which is able to identify a region where, with respect to certain variables, most of the non-bridge non-Hamiltonian graphs reside. We provide evidence that if a cubic graph, chosen at random, is non-bridge non-Hamilonian then it has a high probability of falling within the identified region. It is also observed that these regions remain stable when the order of the population of graphs in question is increased.

Acknowledgements

Thank you to my supervisor Jerzy Filar for the opportunity and for the tireless support. Thank you to my partner Deb for her patience and for keeping me going. Thank you to my family for everything.

Contents

List of Figures

Chapter 1

Definitions

A graph, denoted by G, is composed of a non-empty set of vertices $V(G)$ and a set (possibly empty) of edges $E(G)$. Edges are a relation between two vertices and are denoted by $(i, j) \in E(G)$, where i and j are $\in V(G)$. A graphical representation can be a simple drawing of dots for the vertices and lines connecting the dots for edges.

A subgraph H of the graph G is a set of vertices and edges such that $(V_H \subset V_G)$ and $(E_H \subset E_G)$.

A neighbour of a vertex $i \in V(G)$ is any other vertex that is connected to i by an edge. If vertices are neighbours they are said to be adjacent. The neighbourhood of a vertex i is the set of all vertices adjacent to i. The neighbourhood of a subgraph H is the set of all vertices in the graph $G \setminus H$ that have an edge going into H.

The adjacency matrix of a graph G is denoted $A(G)$ and is a matrix of zeros with a one in row i and column j if vertex i is adjacent to vertex j in G.

Given a graph G of order N, the set of eigenvalues and their multiplicities $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ of the adjacency matrix of G arranged such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ is called the spectrum of the graph.

The order of a graph G is the number of vertices in G. The size of G is the number of edges in G. This investigation is concerned with connected cubic graphs (graphs with 3 edges going to each vertex, also called 3-regular graphs) and all connected cubic graphs must be of even order, say $2m = N$. From this a small calculation reveals that connected cubic graphs must have a size of $(3N)/2$.

A walk is a sequence of vertices $(x_1, x_2, \ldots, x_k)$ such that each x_i is adjacent to x_{i+1} for $i = 1, \ldots, k-1$. A closed walk is a walk where the start and end vertices coincide. A simple cycle is a closed walk where all edges are unique and all vertices are unique apart from the start and end vertex.

A k-connected graph is a graph that contains at least one set of k vertices whose deletion breaks the graph into two or more disconnected components. Cubic graphs are the focus of this study and because each vertex of a cubic graph has three adjacent vertices, the deletion of all three neighbours will disconnect that one vertex. Hence all cubic graphs are 1,2 or 3-connected.

Hamiltonian cycles, which are the focus of this investigation, are defined as: For a graph of order N a Hamiltonian cycle is a simple cycle of length N. Not all cubic graphs contain a Hamilton cycle and it is easy to visualize examples of cubic graphs that do contain a Hamiltonian cycle and ones that do not.

A small connection of the last two definitions allows for cubic graphs to be divided into two classes. A cubic graph that is 1-connected is automatically non-Hamiltonian and is also called a bridge graph. Therefore the set of all cubic graphs can be divided into bridge graphs and non-bridge graphs. The non-bridge graphs contain all the Hamiltonian graphs and the so called non-bridge non-Hamiltonian graphs. This is an important distinction because cubic bridge graphs can be identified very easily whereas in most instances, non-bridge non-Hamiltonian graphs are very difficult to identify. Figures 1.1-1.3 provide some simple examples of both Hamiltonian and non-Hamiltonian cubic graphs.

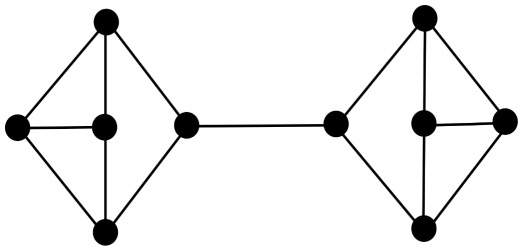

FIGURE 1.1: An example of a cubic bridge graph of order 10. It is trivially not Hamiltonian because as soon as the bridge is crossed in any path, there is no returning without repetition.

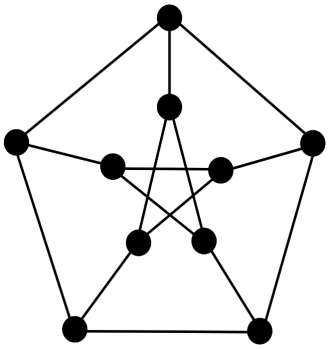

FIGURE 1.2: An example non-bridge non-Hamiltonian cubic graph of order 10. It requires some extra thought to realise that this one is non-Hamiltonian.

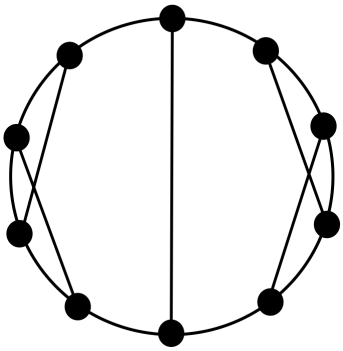

FIGURE 1.3: An example Hamiltonian cubic graph of order 10, one of the Hamilton cycles is completed by simply travelling around the outside edges.

Chapter 2

Introduction and background

Graphs are mathematical objects that have been the subject of increasingly intensive studies in recent years (see for example [12]), mainly because of the explosive growth in applications of networks in complex systems (e.g. telecommunications, social networks and quantum computing).

Graph theory has a rich history that dates back, at least, to the 1700's when Leonard Euler published a now classical paper on the seven bridges of Konigsberg. This paper is largely regarded as having laid the foundations for graph theory [James, p.503]. Early results in graph theory had contributions from some of the proverbial giants of mathematics including Leonhard Euler, Arthur Cayley and William Tutte.

Many years later, graph theory is now a broad area of mathematics that encompasses many smaller branches. Two branches that this thesis utilises are briefly described below.

Algebraic graph theory uses tools from various branches of algebra to study matrices associated with graphs, connections between graphs and group theory, and so called invariants associated with graphs. The study of the spectra of matrices associated with graphs contributes to many important results in algebraic graph theory, some of which are discussed in this investigation.

Random graph theory was pioneered by Paul Erdos and Alfred Renyi and is the study of graphs that are constructed according to certain probabilistic rules. This allows the proof of limit like properties where one can say that 'almost all' graphs possess a

certain property. One recent important result 'almost all' cubic graphs on N vertices are Hamiltonian as N tends to infinity [Robinson and Wormald, 2011].

This thesis is an investigation of the Hamilton cycle problem (HCP) which is encountered and studied in both of the branches mentioned above. Hamilton cycles are named after Sir William Hamilton who investigated the difficulty of the problem and made a puzzle out of a special graph that is often called the Icosian game. The HCP is stated as follows:

Given a graph G, find a Hamiltonian cycle in G or determine that one does not exist.

HCP for the class of cubic graphs is of current importance due to its computational complexity. One can see that in order to find a Hamiltonian cycle or determine that one does not exist in any graph of order N, it is just a matter of checking all possible walks of length N until all possible walks have been exhausted. This is a brute force approach and for cubic graphs, the number of these walks grows in size exponentially with the length of the walk. For large graphs, Hamiltonian cycles are effectively impossible to find this way. The exponential relation between the number of walks and the length of the walk in cubic graphs is of importance due to its connection with complexity theory and the unsolved millennium problem, $\mathcal{P}$ vs $\mathcal{NP}$ [1].

The famous $\mathcal{P}$ vs $\mathcal{NP}$ millenium problem is related to the computation time required to solve a given problem [Garey and Johnson, p.13]. An algorithm solves a problem ρ in a number of iterations. Let n be the number of parameters of a particular instance of ρ, now denote it ρ_n. If the number iterations needed to solve ρ_n are bounded by a polynomial $T(n)$ that depends on n then the algorithm is called a polynomial time algorithm. Problems that can be solved by polynomial time algorithms make up the so-called class $\mathcal{P}$ of polynomiably solvable problems. The class $\mathcal{NP}$ contains problems that can be solved by an algorithm in 'non-deterministic' polynomial time. An $\mathcal{NP}$ time algorithm contains a 'guessing' stage and a 'verification' stage and the guessing stage may not be bounded by a polynomial function. The class $\mathcal{P}$ is certainly contained in $\mathcal{NP}$ because any $\mathcal{P}$ problem has an algorithm that also satisfies the conditions of $\mathcal{NP}$. It is unknown and an active research problem to determine if $\mathcal{NP} \setminus \mathcal{P}$ is empty. The Hamilton cycle problem is known to be $\mathcal{NP}$-complete, which means it lies in $\mathcal{NP}$ and also has the property that any other $\mathcal{NP}$ problem can be reduced to HCP by a polynomial time algorithm. The latter property holds for every NP-complete problem. It is known that

discovering an algorithm that could solve HCP in polynomial time would imply that $\mathcal{P} = \mathcal{NP}$, which would then solve the $\mathcal{P}$ vs $\mathcal{NP}$ problem.

To introduce this study, consider HCP for the class of cubic graphs only, which is still a problem that is of NP-complete complexity [Garey et al., 1976]. We use a statistical model called a logistic regression model to identify the non-bridge, non-Hamiltonian cubic graphs. By viewing a graph as a member of a population of all other graphs of the same order, one can imagine trying to identify the non-Hamiltonian graphs based on their average differences from others in the population (the Hamiltonian graphs). In particular, our population is all of the 2 and 3-connected cubic graphs of order N and the non-bridge, non-Hamiltonian graphs are the graphs to be identified. Using various explanatory variables that are associated with the structure of a graph, the logistic model attempts to determine what information a variable or a combination of variables can extract about the Hamiltonicity of a cubic graph. The non-bridge, non-Hamiltonian graphs in this population are not expected to be completely identifiable by the variables which we consider. All of the variables have polynomial complexity algorithms to compute them, so if the non-bridge non-Hamiltonian graphs were completely identifiable, this would be the equivalent of solving the P vs NP problem. Instead we investigate the characteristic behaviours of the explanatory variables across the Hamiltonian and the non-bridge, non-Hamiltonian graphs. Then, if the variables are appropriate, we can use them to determine a 'region', with respect to these variables, where the non-bridge non-Hamiltonian graphs are more likely to reside.

Note that even with an order as low as $N = 20$, there are already 510,489 distinct, unlabelled[1], connected cubic graphs. This means that, as with many problems in graph theory, studying even moderate size graphs in this manner can be computationally prohibitive.

The motivation behind studying cubic graphs with this proposed method originated from an observation of Ejov et al. [2007]. The authors show a fractal-like structure called the multifilar structure which is constructed using the exponentiated eigenvalues of adjacency matrices associated with graphs (shown in Figure 2.1). Each point on the multi-filar structure corresponds to a graph. The thread-like segments are called filars and upon zooming in, a finer set thread-like segments called sub-filars is revealed.

[1] Unlabelled graphs are determined only by their structure, not the numbers assigned to there vertices. One unlabelled graph has many possible labellings.

The x and y coordinates of a graph G of order N are calculated by defining the mean $\mu(A(G), t)$ of the exponentiated eigenvalues of $tA(G)$ where t is a scalar and the variance $\sigma^2(A(G), t)$ as:

$$\mu(A(G), t) = \frac{1}{N} \sum_{i=1}^{N} exp(t\lambda_i),$$

$$\sigma^2(A(G), t) = \frac{1}{N} \sum_{i=1}^{N} exp(2t\lambda_i) - (\mu(A(G), t))^2.$$

Figure 2.1 shows the multifilar structure, for $t = 1$ for all cubic graphs of order 20 as well as a zoomed in view of the structure. This zooming in can be repeated, that is, it is a self-similar structure. The particular observation that began this investigation was that the majority of non-Hamiltonian graphs appear to reside at the tops the filars and sub-filars. A graphs location in the multifilar structure is related to the number of simple cycles that are present in the graph. In this investigation, we use other similar functions that are also related to graph structure and observe the corresponding behaviour of the non-bridge non-Hamiltonian graphs.

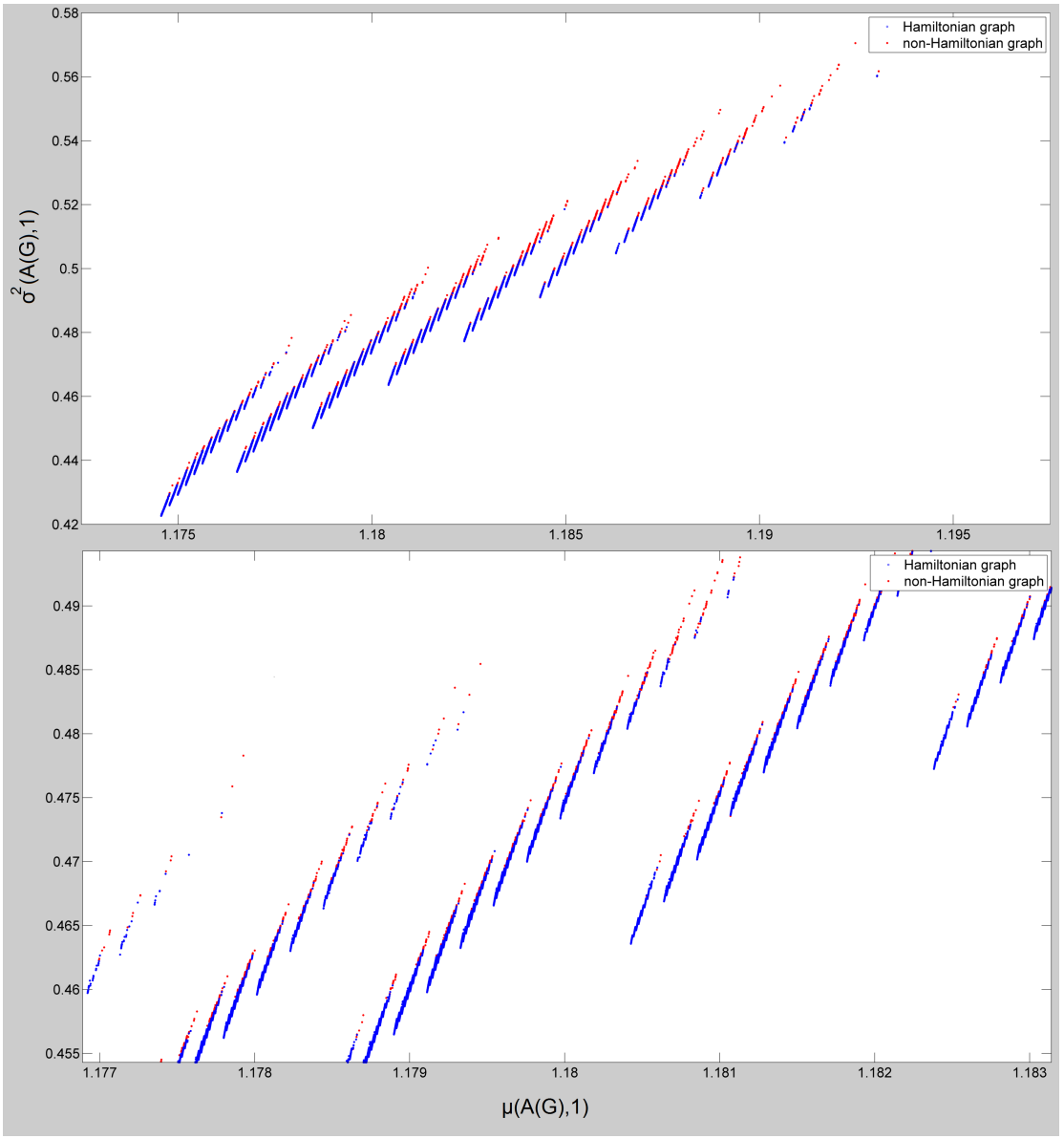

FIGURE 2.1: The multifilar structure of all cubic graphs of order 20 . The bottom plot is a zoomed in view which demonstrates the self similarity of the structure.

Chapter 3

Methodology

3.1 Logistic regression

In order to study cubic graphs with the proposed method, a way to generate random samples from the population is needed. Sampling with replacement, is the appropriate sampling technique. Let P denote the probability of a given cubic graph being selected in the sample. It follows that $P = 1 - (1 - 1/M)^n$ where M denotes the size of the population of 2 and 3-connected cubic graphs of a particular order and n is the sample size. Note that a quick check using a Taylor's expansion shows that for large M, this means that P is approximately equal to n/M. Generating samples of cubic graphs that satisfy the above is not an easy task and the first algorithm to do this was implicit in a paper by Bollobás [1980]. For the purpose of generating very large sample sizes a more computationally appropriate algorithm called the pairing algorithm of Steger and Wormald [1999] is used. This pairing algorithm generates a cubic graph from the population of all cubic graphs of that order with approximately uniform probability. In [21] this approximation is shown to improve as the order of the graphs increases and is exactly uniform as N (and hence also M) tends to infinity. The algorithm generates graphs from the whole population of cubic graphs of order N including the bridge graphs. However the generated bridge graphs can be discarded and what remains is still, approximately, a random sample of the 2 and 3-connected cubic graphs of order N.

Once the population and a way of sampling at random from the population are defined, an appropriate statistical model is needed. A logistic regression model is a well-known statistical tool that allows the prediction of a binary response variable based on predictor variables. Consider a random sample of n graphs $G_1, G_2, \ldots, G_n$ and then let the event $\{Y_i = 1\}$ correspond to G_i being a non-bridge non-Hamiltonain graph. Observing such an event will be called a 'successful event'.

To construct the logistic regression model, take n independent observations, then for $i = 1, \ldots, n$ let the random variable Y_i have the Bernoulli distribution with parameter π_i. That is, $P(Y_i = 1) = \pi_i$, and $P(Y_i = 0) = 1 - \pi_i$. Of course, it follows that the probability mass function of Y_i can be written as:

$$P(Y_i|\pi_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad , \qquad Y_i \in \{0, 1\}.$$

If $\boldsymbol{X}_i$ is the vector of predictor variables corresponding to the i^{th} observation and if $\boldsymbol{\beta}$ is a k-dimensional vector of parameters to be estimated, then – in the logistic regression model – it is postulated that:

$$\pi_i = \frac{1}{1 + \exp(-\boldsymbol{X}_i\boldsymbol{\beta})} \qquad , \qquad i = 1, \ldots, n \tag{3.1}$$

For this investigation the parameters $\boldsymbol{\beta}$ are estimated using maximum likelihood estimation with a special modification called prior correction. Maximum likelihood works by maximizing a likelihood function $L(\boldsymbol{\beta}|\mathbf{Y})$ over the observed data, or for computational simplification, maximizing the natural log of $L(\boldsymbol{\beta}|\mathbf{Y})$ (because the maximum is achieved at the same point due to monotonicity of the natural log). Assuming identically distributed, independent, observations we have:

$$L(\boldsymbol{\beta}|\mathbf{Y}) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)}$$

$$lnL(\boldsymbol{\beta}|\mathbf{Y}) = \sum_{Y_i=1} ln(\pi_i) + \sum_{Y_i=0} ln(1-\pi_i)$$

Critical points in the above are found by taking the partial derivatives with respect to each β_j and setting them to zero. Namely:

$$\frac{\partial(lnL(\boldsymbol{\beta}|\mathbf{Y}))}{\partial\beta_j} = 0, \qquad j = 0, \ldots, k-1. \tag{3.2}$$

In practice closed form solutions of (3.2) may be hard to derive, however, for every realisation $Y = y$ numerical approximations $\hat{\beta}_j = \hat{\beta}_j(y)$ can be found using numerical optimisation algorithms. The solutions here are of course just critical points, however, since the function $L(\boldsymbol{\beta}|\mathbf{Y})$ is known to be concave in $\boldsymbol{\beta}$, the solutions of (3.2) are maxima and hence are estimators $\hat{\beta}_j(Y)$ of the true population parameters β_j, for each $j = 0, 1, \ldots, k-1$.

For cubic, 2 and 3-connected graphs of order 24, the proportion of non-bridge non-Hamiltonian graphs is already only 0.0011 and this quickly shrinks further as the order is increased. It is known that bias in the MLE estimators can arise when applying logistic regression to data where an event is particularly rare. The recommended remedy (see [18]) for the most important bias contributor, namely the first coefficient $\hat{\beta}_0(y)$, is via the following correction:

$$\hat{\beta_0}' = \hat{\beta}_0 - ln\left[\frac{1-\tau}{\tau}\frac{\bar{y}}{1-\bar{y}}\right],$$

where τ is the proportion of successful events in the population and $\bar{y}$ is the proportion of successful events in the sample. Using the correction to $\hat{\beta}_0(y)$ requires the knowledge of the proportion of successful events in the population τ, which in this case, is possible to calculate for some lower orders of cubic graphs. As the order increases, this kind of correction will need to be abandoned because there is currently no practical method to enumerate all the non-bridge non-Hamiltonian graphs.

Measures of goodness of fit are a way to identify predictor variables that produce the model with the best predictive power. In our proposed method of studying cubic graphs there are many possible explanatory variables, so identifying the ones that best predict non-Hamiltonicity is important. Taking advantage of the ease of generating and testing random samples of data makes direct assessment of the model using the sample data the most appropriate way to assess goodness of fit for our scenario. The logistic model is now ready to be applied to our data.

3.2 Two types of error

Once a model is constructed from a random sample of 2 and 3-connected cubic graphs it can be applied to any cubic graph from the population to produce a probability of that graph being non-bridge non-Hamiltonian. To ultimately decide if the model predicts that that graph is non-bridge non-Hamiltonian or Hamiltonian, a threshold probability is needed. If the graph has a modelled probability that is above the threshold then it is designated non-bridge non-Hamiltonian, if it is below the threshold then it is designated Hamiltonian. If we denote $H = \{\text{selected graph is Hamiltonian}\}$ and $nBnH = \{\text{selected graph is non-bridge non-Hamiltonian}\}$, then the two possible error types that occur in this approach are:

Type 1 error: Corresponds to the event where the model predicts H when in-fact it is a non-bridge non-Hamiltonian graph. Denote this event $H|nBnH$.

Type 2 error: Corresponds to the event where the model predicts $nBnH$ when in-fact it is a Hamiltonian graph. Denote this event $nBnH|H$.

Let $\alpha = P(H|nBnH)$ and $\beta = P(nBnH|H)$. In the logistic models to be constructed we would like to identify as many of the non-bridge non-Hamiltonian cubic graphs as possible. That is, we wish to create a model that has a small α value. The explanatory variables used in this investigation do not separate the non-bridge non-Hamiltonian graphs from the Hamiltonian ones, rather the very rare non-bridge non-Hamiltonian graphs are mixed within the vastly greater number of Hamiltonian graphs. To statistically identify these, we must accept that an amount of Hamiltonian graphs will be misclassified as non-bridge non-Hamiltonian. This means that constructing a model that has a low α value typically corresponds to an increase in the β value. While the decision to minimise the amount of Type 1 error may appear somewhat arbitrary, this is not the case. If the graph selected at random were indeed Hamiltonian, then the reliable heuristic in [Baniasadi et al., 2014] would almost certainly find a Hamilton cycle. Thus we are less concerned about missclassifying a Hamiltonian graph as non-Hamiltonian than the other way around.

Chapter 4

Explanatory Variables

This chapter introduces some functions and descriptors of graph characteristics that were explored throughout the study. For convenience, the term *function* of a graph will be a general term used to describe any function of a matrix associated to the graph or any descriptor of a characteristic of that graph. Graph theory has an enormous amount of literature and the below is just a brief explanation of the variables that proved to be candidates for use in the results section.

4.1 Functions of matrices associated to graphs

One of the motivators that sparked an interest in exploring characteristics of graphs came from an observation of Filar et al. [2005]. They discovered a pair of cubic graphs that were cospectral, that is, have the same set of eigenvalues and yet one is Hamiltonian and one non-Hamiltonian. This is an unusual combination of properties because the set of eigenvalues is related to the number of closed walks in a graph:

$$\sum_{i=1}^{n} \lambda_i^{\ell} = trace(A(G)^{\ell}), \qquad \forall \ell \in \mathbb{N}.$$

Where the diagonal entries of $A(G)^{\ell}$ equals the number of closed walks of length ℓ (walks starting and ending at the corresponding vertex). Therefore cospectral graphs have the same number of closed walks of all lengths. The authors of [13] conclude with the statement 'thus the spectrum of a graph unfortunately does not contain enough

information to decide whether a graph is Hamiltonian or not'. A natural follow-up question to this is: *are there any functions, or combination of funtions, that do contain sufficient information about Hamiltonicity?* If a function or combination of functions did contain easily extractable information about the presence of a Hamilton cycle it would, by the above, not depend solely on the eigenvalues of the adjacency matrix. Many other functions of matrices associated with graphs are well-studied and some that are known to be related to the structure of the underlying graph are discussed below.

The resistance matrix of a graph, R(G), has entries, $r(i,j)$, which describe the effective resistance between two vertices in a graph. Resistance matrices have an appealing physical interpretation that is also related to their means of discovery as a mathematical tool [Bollobás, p.39]. Think of the edges of a graph as being wires in an electrical circuit of which have a unit resistor attached to each wire, then think of the vertices as sensors. The effective resistance between vertex i and vertex j, is the current received at vertex i when a unit current flows from j. One of the formulae for computing the effective resistance and hence each entry in a resistance matrix R, discovered by Bapat et al. [2014] is:

$$r(i,j) = \frac{det(L(i,j))}{det(L(i))}, \qquad (\Delta)$$

where $det(L(i,j))$ is the determinant of the Laplacian matrix[1] L with the i^{th} row and column and j^{th} row and column deleted. Similarly $det(L(i))$ is the determinant of the Laplacian matrix with just the i^{th} row and column deleted. The values in the resistance matrix are governed by the underlying structure of the graph. In a similar fashion to the multifilar structure discussed in Chapter 1, statistical moments of the entries in the resistance matrix will be investigated for sensitivity towards Hamilton cycles. The resistance matrix is symmetric on undirected graphs[2] and its diagonal entries are zero. Consider the entries in the upper triangle of the resistance matrix. Define the statistical moments of these entries as:

$$VR = var(\{r(i,j)\}_{i>j}\ )$$

[1]The Laplacian matrix of a cubic graph is $L(G) = 3I - A(G)$ where $A(G)$ is the adjacency matrix of G and I is the identity matrix.

[2]Undirected graphs have no orientation assigned to their edges, that is, they can be traversed by a path in either direction. All graphs in this study are undirected.

$$SR = skewness(\{r(i,j)\}_{i>j})$$

$$KR = kurtosis(\{r(i,j)\}_{i>j})$$

Where the variance, skewness and kurtosis are given by the standard statistical formulae (see [10]). Note this notation will be used sparingly to avoid confusion. From numerical experiments ran during the investigation it was observed that if for a graph $r(i,k) = r(i,j)$, then the types of paths between these pairs of vertices are similar or even exactly the same. From this it is reasonable to expect that a small variance of the entries in the resistance matrix frequently arises in a graph that has many symmetries in its structure. On the other hand a large variance frequently arises in a graph that does not have much symmetry. This phenomenon is illustrated in Figure 4.1.

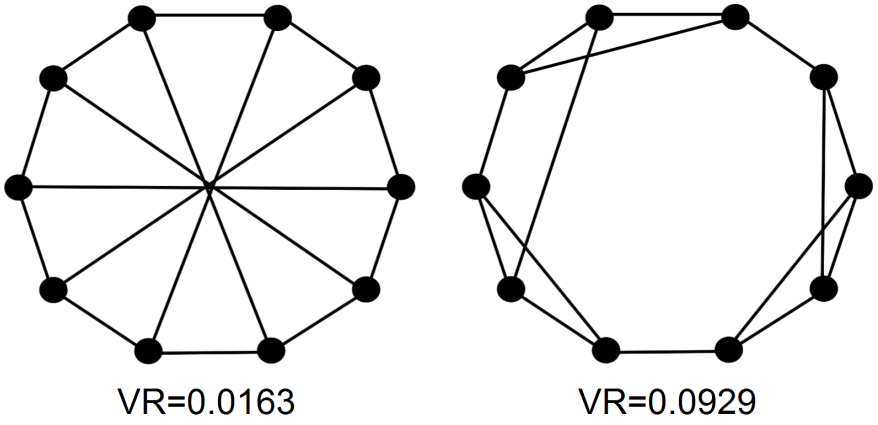

FIGURE 4.1: Variance of resistances equal to 0.0163 for the very symmetric cubic graph on the left and 0.0929 for the significantly less symmetric cubic graph on the right.

We note that 1-connected cubic graphs, that is, cubic bridge graphs can be easily identified with the logistic models due to the larger than normal values in their resistance matrices. This is attributed to the restrictive nature of paths between vertices across a bridge edge. The bridges present in a graph are easily identified in its resistance matrix by the following discussion.

Firstly, the notion of a spanning tree needs to be introduced. A spanning tree of a graph G is a connected subgraph of G such that all of the vertices of G are in the subgraph and there are no cycles in the subgraph. The fact that there are no cycles in a spanning tree makes it look somewhat like a skeleton of the original graph, hence the name spanning

tree. A convenient way to calculate the number of spanning trees in a graph G is given by the well-known matrix tree theorem [Bapat, 2014, p.52]:

Matrix tree theorem: For a given connected graph G the number of spanning trees of G is equal to any cofactor of the Laplacian matrix of G.

Now bridges in a graph can be identified in the resistance matrix R by the following result.

Lemma 4.1: If an edge $(i, j) \in E(G)$ is a bridge, then the resistance distance $r(i, j) = 1$.

Proof. The proof will depend on the ratio formula (Δ) for entries in the resistance matrix. The graph G_A in Figure 4.2 consists of all vertices to the left of vertex i, excluding the edge (i, j). Similarly the graph G_B consists of all vertices and edges to the right of vertex j, excluding the edge (i, j).

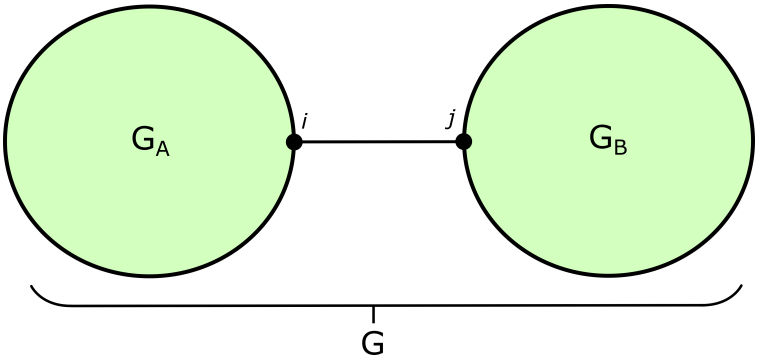

FIGURE 4.2: A representation of a bridge graph G, the circles on the left and right represent any connected graph structure.

We shall also need auxiliary graphs $\widehat{G_A}$ and $\widehat{G_B}$ which are the same as G_A and G_B respectively, except that each includes the edge (i, j). Thus the number of vertices of $\widehat{G_A}$ is one more than the number of vertices of G_A because vertex j and edge (i, j) are included. Similarly $\widehat{G_B}$ includes the vertex i and the edge (i, j). An example of $\widehat{G_B}$ is shown in Figure 4.3.

It is now straightforward to verify that the Laplacian of the original bridge graph G is of the form:

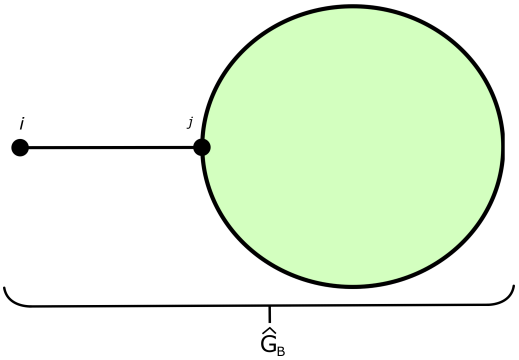

FIGURE 4.3: A representation of the graph $\widehat{G_B}$

$$L(G) = \begin{array}{c} \\ \\ i \\ \\ j \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccc} & & & i & & j & & \end{array} \\ \begin{pmatrix} & & & \vdots & & & & \\ & \widehat{L}_A & & \vdots & & 0 & & \\ & & & \vdots & -1 & & & \\ \cdots & \cdots & \cdots & \vdots & \cdots & \cdots & \cdots & \\ & & -1 & \vdots & & & & \\ & 0 & & \vdots & & \widehat{L}_B & & \\ & & & \vdots & & & & \end{pmatrix} \end{array},$$

where $\widehat{L}_A = L(G_A) + \mathbf{e}_i\mathbf{e}_i^T$ and $\widehat{L}_B = L(G_B) + \mathbf{e}_1\mathbf{e}_1^T$. Here $\mathbf{e}_i$ is the i-th vector of the i-dimensional unit basis and $\mathbf{e}_1$ is the first vector of the m-dimensional unit basis where m is the number of vertices in G_B. Of course, $\mathbf{e}_i\mathbf{e}_i^T$ is an $i \times i$ matrix which has unity in its (i,i)-th entry and zeroes everywhere else. Note that $L(G_A)$ and $L(G_B)$ are the Laplacian matrices of G_A and G_B, respectively. The other two blocks in $L(G)$ are all zeroes except for the entries of (-1) corresponding to the bridge edge at $(i,j) \in E(G)$. Next using the notation from Bapat et al. [2014],

$$L(G(i,j)) = \begin{pmatrix} L(G_A(i)) & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & L(G_B(j)) \end{pmatrix},$$

where $L(G_A(i))$ denotes the block $L(G_A)$ with the i-th row and column deleted and similarly for $L(G_B(j))$. By construction it can be checked that the (i,i)-th minor of

$L(G_A)$ coincides with the matrix $L(G_A(i))$ and the $(1,1)$-th minor of $L(G_B)$ coincides with the matrix $L(G_B(j))$. Hence by the Matrix Tree Theorem

$$\begin{aligned} det(L(G(i,j))) &= det(L(G_A(i)))det(L(G_B(j))) \\ &= t(G_A)t(G_B), \end{aligned} \tag{4.1}$$

where $t(G_A)$ is the number of spanning trees of the graph G_A, and similarly for G_B. Note that equation (1) gives us the numerator of the ratio in (Δ). To evaluate the denominator consider the matrix $L(G(i))$ which is $L(G)$ with the j-th row and column deleted. By construction it has the form:

$$L(G(i)) = \begin{pmatrix} L(G_A(i)) & \vdots & 0 \\ \dots & \dots & \dots \\ 0 & \vdots & \widehat{L}_B \end{pmatrix}.$$

Clearly,

$$\begin{aligned} det(L(G(i))) &= det(L(G_A(i)))det(\hat{L}_B) \\ &= t(G_A)det(\widehat{L}_B). \end{aligned} \tag{4.2}$$

The proof will be complete if we can show that $det(\widehat{L}_B) = t(G_B)$.

However using the fact that $\widehat{L}_B = L(G_B) + \mathbf{e}_1\mathbf{e}_1^T$ which simply adds unity to the first diagonal entry of $L(G_B)$, we see that $\widehat{L}_B$ is some (k,k)-th minor of the Laplacian of the graph $\widehat{G_B}$ with the pendulum vertex i. Hence by another application of the Matrix Tree Theorem

$$det(\widehat{L}_B) = t(\widehat{G_B}). \tag{4.3}$$

However, because vertex i is a pendulum vertex, spanning trees of $\widehat{G_B}$ are in $1:1$ correspondence with spanning trees of G_B. Thus

$$det(\widehat{L}_B) = t(G_B). \qquad (4.4)$$

Combining (4.4), (4.2) and (4.1) with (Δ) yields

$$r(i,j) = \frac{det(L(i,j))}{det(L(i))} = \frac{t(G_A)t(G_B)}{t(G_A)t(G_B)} = 1,$$

Whenever the edge (i,j) is a bridge. $\square$

Lemma 4.1 supplies a one directional test of whether a graph contains a bridge. Namely, the absence of unity among the entries of R ensures that there is no bridge. The example Figure 4.4 below shows that the converse does not hold.

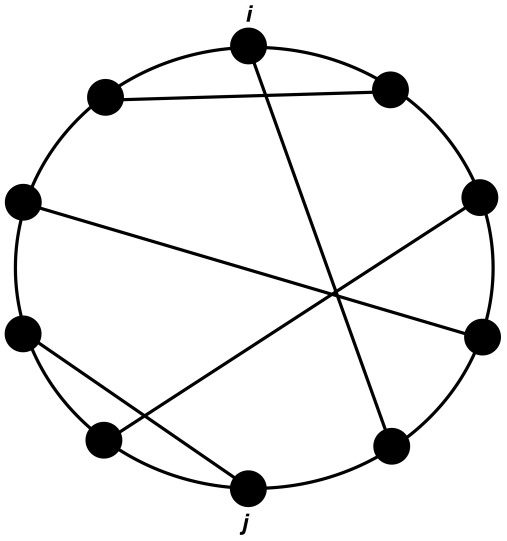

FIGURE 4.4: The resistance distance between vertex i and vertex j is $r(i,j) = 1$, however, i and j are not adjacent.

The next function that was considered as a possible indicator of Hamiltonicity was the gap between the second largest and the largest eigenvalue $(\lambda_1 - \lambda_2)$. Although from the beginning of this chapter, the eigenvalues cannot solely determine Hamiltonicity, they may still be explored as contributing explanatory variables for the logistic model. It is known that in a special sense this gap $(\lambda_1 - \lambda_2)$ describes the degree of randomness and connectivity of the graph [Brouwer and Haemers, 2011, p.67]. For connected cubic graphs the largest eigenvalue is necessarily equal to three, so studying the second largest eigenvalue λ_2 is equivalent to studying $(\lambda_1 - \lambda_2)$ and this is what will be done in this study. The second largest eigenvalue has previously been used to relate the so called

toughness[3]of a graph to Hamiltonicity in Chvatal [1973] and the properties of that relationship are still being studied. We will illustrate one of the ways the second largest eigenvalue describes a graph. Consider a d-regular graph G, let X and Y be two subsets of the vertices of G such that X and Y are disjoint. Let r denote the number of edges connecting a vertex in X to a vertex in Y. Then it was shown in [Brouwer and Haemers, 2011] that for every possible disjoint X and Y the following holds:

$$r \geq \frac{(d - \lambda_2)|X||Y|}{N},$$

where $|X|$ is the number of vertices in the subset X. Therefore a small λ_2 describes a cubic graph that has high average 'connectivity' throughout the graph and a large λ_2 describes a cubic graph that may have 'bottle-neck' like features.

Next, we considered a function that is related to the multifilar structure discussed in Chapter 1 that is called the Estrada index ($EE(G)$). It was originally introduced as a measure of connectivity in complex networks and is related to the number of closed walks of all lengths in the graph [Estrada, 2011]. It is defined as:

$$EE(G) = \sum_{i:=1}^{n} e^{\lambda_i}.$$

Continuing with the idea of using information about closed walks, the next indicator function attempts to develop a notion of distance between cubic graphs. A measure borrowed from information theory called the Kullback-Leibler divergence is traditionally used to measure a difference between discrete probability distributions. Given two different discrete probability distributions P and Q, the Kullback-Leibler divergence measures the error when P is used as an approximation for Q and is defined as:

$$D(P||Q) = \sum_{i} P(i) ln\Big[\frac{P(i)}{Q(i)}\Big]$$

A modification to the formula allows the measurement of 'divergence' between graphs. The matrix exponential of an adjacency matrix $A(G)$ is a matrix whose entries have

[3]Toughness is another measure of the connectivity of a graph.

contributions from walks of all lengths in the graph, as can be seen in the following formula,

$$e^{A(G)} = \sum_{k=0}^{\infty} \frac{1}{k!} A(G)^k,$$

because $[A(G)^k]_{i,j}$ is the number of all possible walks of length k that start at vertex i and end at vertex j. For the purpose of modelling with this modified Kullback-Leibler divergence a benchmark graph is needed, one which all other graphs can be measured against and one that exists for all orders of cubic graphs. The answer is found in the Mobius ladder graph (see Figure 4.4) as it exists for all orders of cubic graphs and it is also among the most symmetrical of cubic graphs. We now define two versions of the modified Kullback-Leibler divergence as follows. Let G be a graph of order N and M be the Mobius ladder graph of order N. Then let $E = e^{A(G)}$ and $F = e^{A(M)}$ and $\bar{E}$ and $\bar{F}$ be the column normalised matrices of E and F respectively. Then the modified Kullback-Leibler divergence with respect to the Mobius ladder graph is defined as:

$$D(G||M) = \frac{1}{N^3} \sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} [\bar{E}]_{ij} ln \frac{[\bar{E}]_{ij}}{[\bar{M}]_{ik}}$$

And the modified Kullback-Leibler divergence with respect to the graph itself is defined as:

$$D(G||G) = \frac{1}{N^3} \sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{i=1}^{N} [\bar{E}]_{ij} ln \frac{[\bar{E}]_{ij}}{[\bar{E}]_{ik}}$$

By using the Kullback-Leibler divergence for the matrix exponentials we attempt to capture a notion of probabilistic distance between the Mobius ladder graph and the graph in question. The results section will analyse some of the behaviours of this new index. One reason why the above formula was appealing for use as an explanatory variable is that this function is only defined for connected graphs. This can be seen by taking a disconnected graph G, if a vertex i is in one component of G and a vertex j is in another, then $[e^{A(G)}]_{i,j} = 0$ and is also zero in the normalized version. Then we would have $ln(0)$ for some entries in the modified Kullback-Liebler divergence of G and

therefore it is undefined. Another reason the formula was appealing for use is the fact that it remains constant under isomorphisms of the graph, this fact is left unproven.

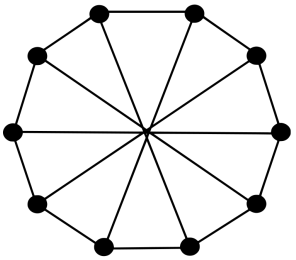

FIGURE 4.5: The Mobius ladder graph of order 10 which is used as a benchmark graph in the modified Kullback-Leibler divergence.

4.2 Discrete valued descriptors of graphs

All of the above functions are investigated for use as explanatory variables in the results chapter. Graphs can also be described with several functions that take on discrete values (existence of a Hamilton cycle is one of these). Below is a brief introduction to some discrete valued functions that were also investigated.

The diameter of a graph is the longest-shortest path between any two vertices. That is, given a list of the shortest paths ($\delta(i,j)$) between all pairs of vertices in the graph, the diameter is the maximum of these.

$$diam(G) = \max_{i,j \in V(G)} \delta(i,j)$$

The girth of a graph is the length of the shortest simple cycle. Recalling that cycles of length two are not simple cycles, for cubic graphs the girth is necessarily three or greater.

A graph is planar if it can be arranged so that its vertices lie on a 2 dimensional plane and its edges do not cross-over each other (intersect).

The number of triangles present in the graph dictates in which segment the graph belongs to in the multifilar structure discussed in Chaper 1.

The number of spanning trees in a graph was introduced earlier and it is also considered for use as an explanatory variable.

Lastly, whether the adjacency matrix possesses a zero eigenvalue (and hence $detA(G) = 0$) proved to be an interesting descriptor. A zero eigenvalue means that a row or column is a linear combination of some other rows or columns. This fact means that certain graph structures force the existence of a zero eigenvalue. For example Figure 4.5 shows two small subgraphs whose presence in a graph G will force G to possess a zero eigenvalue. The S labels represent a connection to anywhere in the graph (as long as it remains a cubic graph). To see that a zero eigenvalue is present, in the left subgraph the vertices i and j share the same neighbourhoods, so immediately their rows (and columns) are identical in the adjacency matrix, hence a zero eigenvalue. In the right subgraph, a quick calculation reveals that, if we let $\mathbf{a}$ denote the row in the adjacency matrix for the vertex a, then $\mathbf{a}-\mathbf{b}-\mathbf{c}+\mathbf{d} = \mathbf{0}$ and hence there is also a zero eigenvalue. So presence of at least one zero eigenvalue splits all cubic graphs into two groups, one group contains graphs which possess one or more of such structures (hence $det(A(G)) = 0$) and the other group contains graphs which possess none of these structures (hence $det(A(G)) \neq 0$).

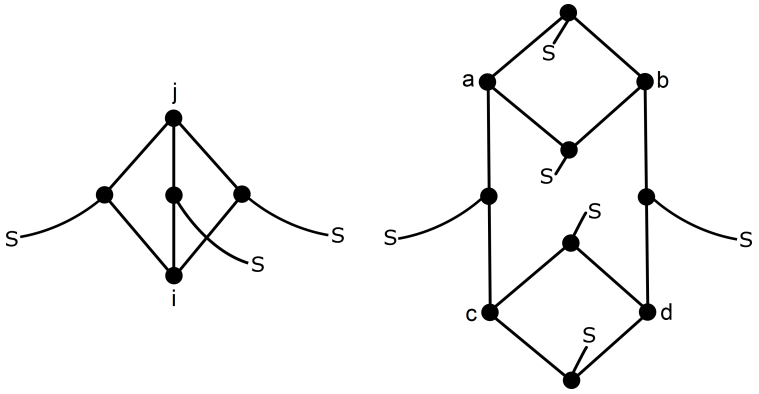

FIGURE 4.6: Two of the smallest cubic subgraphs that force the adjacency matrix of the overall graph to possess a zero eigenvalue. The S label represents a connection to any other part of the graph, including connecting to another S (as long as it remains a cubic graph).

To conclude this chapter we note that these functions are all interesting in their own right and some are the subjects of on-going research. Their use as explanatory variables in this investigation is aimed at a more broad analysis of how they behave when they are applied to cubic graphs.

Chapter 5

Results

5.1 Local stability with respect to the order of the graph

The goal of the investigation is to construct a model from a random sample of 2 and 3-connected cubic graphs which can predict if a given cubic graph of the same order is non-bridge non-Hamiltonian. If the explanatory variables are appropriate, the same model can also be used for orders of graphs that are nearby the order if the graphs in the original calibration sample. We have some empirical evidence that this, indeed, the case. To see why this can be so, we need to consider how the non-bridge non-Hamiltonian graphs behave as the order of the graphs changes. Figure 5.1 demonstrates the behaviour of the non-bridge non-Hamiltonian graphs for two of the variables introduced in Chapter 4. We observe that, with respect to these variables, as the order of the graphs increase, the number of graphs with similar characteristics also increases, that is, there is a clustering effect. In the Figure 5.1 below, observe that the density of non-bridge non-Hamiltonian graphs is much greater in the cluster located in bottom left of the plots and the cluster continues to increase in density as the order of the graphs is increased. When the variables in consideration produce a single main cluster we call this cluster of non-bridge non-Hamiltonian graphs, including the Hamiltonian graphs scattered within, a 'region' where most of the non-bridge non-Hamiltonian graphs reside with respect to these variables. If these regions are well enough defined for a smaller order, then throughout this investigation we have observed that there is a carry-over effect and this region is also present in the same location for larger orders. The variables

that produce a single main region are the ones that prove to be the best explanatory variables in the next section.

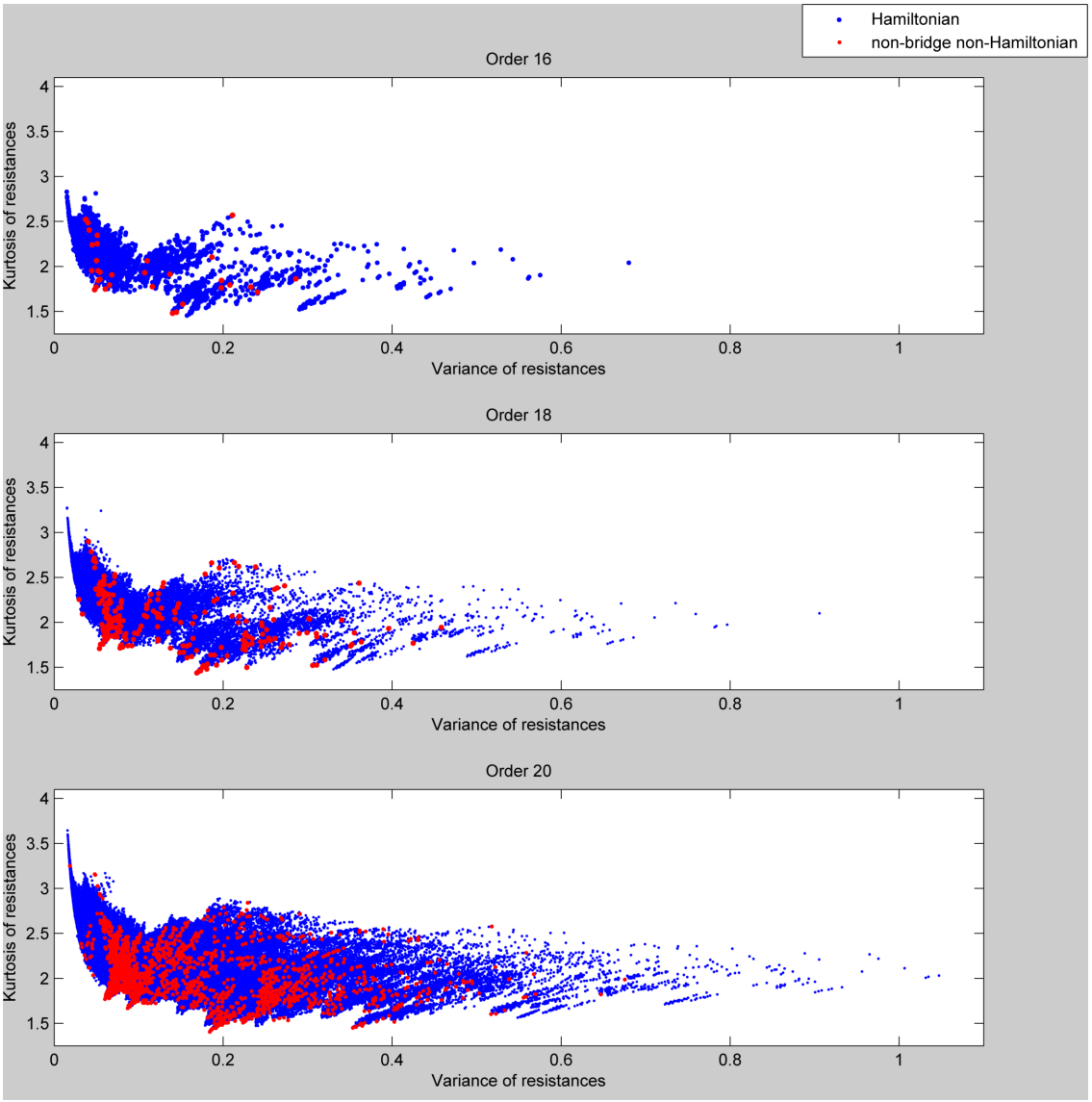

FIGURE 5.1: Kurtosis of resistances versus variance of resistances for orders 16, 18 and 20. Note that the red dots are shown at an increased scale for clarity.

5.2 Results

Earlier chapters have described the construction of a logistic model and the functions of graphs that may be of interest to use as explanatory variables. This results chapter will proceed by presenting three cases where a logistic model is constructed and then used to predict non-bridge non-Hamiltonian graphs for larger orders. As with any statistical

model on more than one predictor variable, the amount of correlation present is assessed and given in Appendix B. Under this approach a larger correlation value is acceptable as long as it is not too close to one. The first case is constructed using three predictor variables namely the second largest eigenvalue, the variance and the kurtosis of entries in the resistance matrix. All three are certainly related in some way to the structure of a graph as discussed in Chapter 4. The logistic model will attempt to ascertain to what extent the three variables together can identify non-bridge non-Hamiltonian graphs.

A random sample of cubic graphs of order 24 is generated using the pairing algorithm introduced in Chapter 3. The details of this sample are displayed in Table 5.1. From this calibration sample, a logistic model is constructed and the estimates shown in Table 5.2 are provided by the maximum likelihood estimation technique discussed in Chapter 3, that is, they form the vector $\hat{\boldsymbol{\beta}}$ in equation (3.1).

TABLE 5.1: Information about the population and random sample of cubic, 2 and 3-connected graphs of order 24.

Population size	118,118,367
Number of nBnH in population	177,832
Sample size	1,793,560
True number of nBnH graphs in sample	1,246

TABLE 5.2: Coefficients from the logistic model on the calibration sample of cubic, 2 and 3-connected graphs of order 24.

Parameter	Variable associated to parameter	Estimation of parameter
β_0	Constant term	-63.763
β_1	Variance of resistances	-9.956
β_2	Kurtosis of resistances	-1.770
β_3	Second largest eigenvalue	23.258

To both assess the model fit and then use the modelled parameters for predicting non-bridge non-Hamiltonian graphs for larger orders, first a threshold probability is needed. This is a probability value that is assigned so that after the model is fit to the sample data, 90% of the non-bridge non-Hamiltonian graphs in this callibration sample are

correctly assigned an 'nBnH' label. Then using this threshold, in order to observe the model's fit, we calculate how many of the Hamiltonian graphs are correctly assigned an 'H' label . The results for this are displayed in Table 5.3.

TABLE 5.3: Threshold probability and error values produced for the calibration sample of cubic, 2 and 3-connected graphs of order 24.

Threshold probability: 0.002345		
	Correctly predicted	Incorrectly predicted
nBnH graphs	89.97%	10.03%
H graphs	84.60%	15.40%

The threshold probability is then fixed and used with the model parameters to predict for larger graphs, which in this case, are orders 28 and 30. The accuracy of the model for the larger orders of graphs is then calculated by assessing the error values produced when applying the threshold to the modelled probabilities. The results for orders 28 and 30 are displayed in Table 5.4 and 5.5 respectively. It is expected that the accuracy diminishes, the further away the order gets from order of the graphs in the original calibration sample. This is true for the proportion of Hamiltonian graphs correctly assigned an 'H' label, however, it is observed that the proportion of non-bridge non-Hamiltonian graphs correctly assigned an 'nBnH' label remains approximately constant across the different order graphs.

TABLE 5.4: Results when model is used to predict on a random sample of cubic, 2 and 3-connected graphs of order 28.

Sample size (order 28) = 598,550

True number of nBnH graphs in sample = 239

True number of H graphs in sample = 598,311

Threshold probability = 0.002345

	Correctly predicted	Incorrectly predicted
nBnH graphs	89.12%	10.88%
H graphs	81.14%	19.86%

TABLE 5.5: Results when model is used to predict on a random sample of cubic, 2 and 3-connected graphs of order 30.

Sample size (order 30) = 349,337

True number of nBnH graphs in sample = 77

True number of H graphs in sample = 349,260

Threshold probability = 0.002345

	Correctly predicted	Incorrectly predicted
nBnH graphs	92.02%	7.98%
H graphs	70.48%	29.52%

To continue this further, we consider a second model constructed again on a sample of order 24 cubic graphs and using the same explanatory variables, except we discard the kurtosis of resistances variable. This is because the kurtosis of resistances is observed to scale with the order of the graphs being considered, which the model cannot account for. We aim to remove the error introduced by the scaling of this variable and hope that this increases the accuracy of the predictions. We also narrow the population down to 2 and 3-connected cubic graphs which possess zero triangles (no simple cycles of length 3). This sub-division allows us to restrict the model to graphs that already have one important structural characteristic in common. It may seem like this is too much of a simplification in our search for the non-bridge non-Hamiltonian cubic graphs. However, HCP remains NP-complete even when considering only triangle-free cubic graphs, this fact is proved in the Appendix A.

A random sample of now triangle-free, 2 and 3-connected, cubic graphs of order 24 is generated. The details of this calibration sample are displayed in Table 5.6. This time the number of cubic graphs in the desired population is unknown without generating and evaluating all of the 118,118,367 order 24 cubic graphs. This means that the estimated parameters, which are shown in Table 5.7, do not take advantage of the prior correction that was discussed in Chapter 3. A threshold probability is assigned so that 90% of the triangle-free non-bridge non-Hamiltonian graphs in the calibration sample are correctly assigned an 'nBnH' label. Again using this threshold, in order to assess the model's fit, we can calculate how many of the triangle-free Hamiltonian graphs are correctly assigned an 'H' label. The results for this are displayed in Table 5.7.

TABLE 5.6: Information about the random sample of cubic, triangle-free, 2 and 3-connected graphs of order 24.

Population size	Unknown
Number of nBnH in population	Unknown
Sample size	323,838
True number of nbnH graphs in sample	81

TABLE 5.7: Coefficients from the logistic model on the calibration sample of cubic, triangle-free, 2 and 3-connected graphs of order 24.

Parameter	Variable associated to parameter	Estimation of parameter
β_0	Constant term	-121.67
β_1	Second largest eigenvalue	43.38
β_2	Variance of resistances	-24.94

TABLE 5.8: Threshold probability and error values produced for the calibration sample of cubic, triangle-free, 2 and 3-connected graphs of order 24.

Threshold probability: 0.001222		
	Correctly predicted	Incorrectly predicted
nBnH graphs	91.36%	8.64%
H graphs	94.94%	5.06%

Similarly to the first model, the threshold probability value is then fixed and used with the modelled parameters to predict Hamiltonicity for graphs of larger order, which again, are orders 28 and 30. The details of these are displayed in Tables 5.9 and 5.10 respectively. It is observed that the proportion of Hamiltonian graphs that are correctly assigned an 'H' label is significantly higher than in the first case (that is, the second model produces a lower Type 2 error).

TABLE 5.9: Results when model is used to predict on a random sample of cubic, triangle-free, 2 and 3-connected graphs of order 28.

Number of graphs in sample = 89,972

True number of nBnH graphs in sample = 19

True number of H graphs in sample= 89,953

Threshold probability = 0.001222

	Correctly predicted	Incorrectly predicted
nbnH graphs	89.47%	10.53%
H graphs	89.20%	10.80%

TABLE 5.10: Results when model is used to predict on a random sample of cubic, triangle-free, 2 and 3-connected graphs of order 30.

Number of graphs in sample = 159,141

True number of nBnH graphs in sample = 16

True number of H graphs in sample = 159,125

Threshold probability = 0.001222

	Correctly predicted	Incorrectly predicted
nbnH graphs	93.75%	6.25%
H graphs	85.43%	14.57%

These first two cases show that, at least for orders of graphs nearby those of the original sample, most non-bridge non-Hamiltonian graphs can be identified provided that the Type 2 error is not of great concern. We can summarise the result of these first two cases with the statement:

A logistic model can be constructed from cubic 2 and 3-connected graphs of order N using the second largest eigenvalue, the variance, and the kurtosis of the entries in the resistance matrix as explanatory variables. The model can identify the region where most non-bridge non-Hamiltonian graphs reside for orders close to N. The proportion of Hamiltonian graphs that are correctly identified increases when only triangle-free, 2 and 3-connected graphs are considered and the kurtosis of resistances variable is discarded.

The third case considers two explanatory variables that were chosen because of the way that they distinguish the co-spectral Hamiltonian and non-Hamiltonian pairs which were discussed in Chapter 4. The explanatory variables are the skewness of the entries in the

resistance matrix (call this function of a graph G, $SR(G)$) and the modified Kullback-Leibler divergence (with respect to the graph itself), $KL(G)$. For the 31 co-spectral pairs (G_H and G_{NH}) that exist in cubic graphs of order 22, all but 3 pairs have the property that $SR(G_H) < SR(G_{NH})$. Similarly all but 5 pairs have the property that $KL(G_H) < KL(G_{NH})$. So in a similar manner to the first two cases, this case will investigate these explanatory variables ability to identify non-bridge non-Hamiltonian cubic graphs in general.

Again we start with a random sample of 2 and 3-connected, cubic graphs of order 24. The details of this calibration sample are displayed in Table 5.11.

TABLE 5.11: Information about the population and random sample of cubic, 2 and 3-connected graphs of order 24.

Population size	118,118,367
Number of nBnH graphs in population	177,832
Sample size	896,823
True number of nBnH graphs in sample	631

A logistic model is then constructed on the skewness of resistances and the modified Kullback-Leibler distance and then a threshold probability value is assigned to assess the model fit. This time the threshold probability, and hence the predicted region, is relaxed to include 95% (rather than 90%) of the non-bridge non-Hamiltonian graphs in the sample data. This is because we observe that a very tight cluster of non-bridge non-Hamiltonian graphs, with respect to the variables, has been generated in the calibration sample. The results for these are displayed in tables 5.12 and 5.13.

TABLE 5.12: Coefficients from the logistic model on the sample of cubic, 2 and 3-connected graphs of order 24 using the variables modified Kullback-Leibler divergence and skewness of resistances.

Parameter	Variable associated to parameter	Estimation of parameter
β_0	Constant term	-14.71
β_1	modified Kullback-Leibler divergence	46.97
β_2	Variance of resistances	1.06

TABLE 5.13: Threshold probability and error values produced on the sample of order 24 graphs using the variables: modified Kullback-Leibler divergence and skewness of resistances.

Threshold probability: 0.000514		
	Correctly predicted	Incorrectly predicted
nBnH graphs	95.09%	4.91%
H graphs	67.66%	32.33%

Next, the threshold probability is fixed and the modelled parameters are used to predict Hamiltonicity for random samples of 2 and 3-connected graphs of orders 28 and 30. The results are displayed in tables 5.14 and 5.15 respectively.

TABLE 5.14: Results when model is used to predict on a random sample of cubic, 2 and 3-connected graphs of order 28.

Number of 28 vertex graphs in sample = 798,112

True number of nBnH graphs in sample = 289

True number of H graphs in sample = 797,823

Threshold probability = 0.000514

	Correctly predicted	Incorrectly predicted
nbnH graphs	94.46%	5.54%
H graphs	78.32%	21.68%

TABLE 5.15: Results when model is used to predict on a random sample of cubic, 2 and 3-connected graphs of order 30.

Number of 30 vertex graphs in sample = 698,664

Actual number of non-bridge non-Hamiltonian graphs = 160

Actual number of Hamiltonian graphs = 698,504

Threshold probability = 0.000514

	Correctly predicted	Incorrectly predicted
nbnH graphs	90%	10%
H graphs	82.87%	17.13%

The proportion of Hamiltonian graphs that are correctly predicted starts off lower in this third case. Unlike the other two cases, the proportion of correctly predicted Hamiltonian

graphs increases as the order of the graphs is increased. This observation suggests that the proportion of Hamiltonian graphs that lie outside of the identified region is becoming larger as the order of the graphs is increased. The result of this case can be summarised with the statement:

A logistic model can be constructed from cubic 2 and 3-connected graphs of order N using the modified Kullback-Leibler divergence and the skewness of the entries in the resistance matrix as explanatory variables. The model can identify a region where most non-bridge non-Hamiltonian graphs reside for orders close to N.

The results will be discussed in the next section and we conclude the results chapter with the note that the order of the graphs in the calibration samples are chosen because order 24 is the smallest order in which the whole population of cubic graphs cannot be generated and analysed easily.

5.3 Discussion

The predictions provided by the logistic models in Chapter 5.2 are given when equation (3.1) is applied to a given graph and the resulting probability lies either above or below the pre-specified probability threshold. When the model is used for prediction on a random sample of graphs, the proportion of non-bridge non-Hamiltonian graphs that are successfully identified is observed to remain close to that of the original sample used for model collaboration. This is attributed to the original callibration sample of order 24 graphs appropriately representing the distribution of the main cluster of non-bridge non-Hamiltonian graphs in the population. When the model is used to predict on a random sample of graphs of a different order to the original, provided that the order is nearby that of the original, it is observed that the proportion of successfully identified non-bridge non-Hamiltonian graphs still remains close to that of the original sample. This is attributed to the carry-over effect of clusters discussed in Chapter 5.1.

The main result from this investigation is the formulation of a technique which is able to, with respect to certain explanatory variables, identify the region where most of the non-bridge non-Hamiltonian graphs reside. These variables do not separate the non-bridge non-Hamiltonian graphs from the Hamiltonian graphs, rather they contain an area where the likelihood of a graph being non-bridge non-Hamiltonian is greater. The

list of explanatory variables that were discussed in Chapter 4, did not all possess these high density regions of non-bridge non-Hamiltonian graphs. The variables that were chosen were observed to produce the best clustering effect when computing them on small order graphs of under 20 vertices, of which the whole population can be computed and analysed directly. The clustering that is present in small order graphs gives a good indication of how the distribution of the non-bridge non-Hamiltonian graphs will behave for larger orders. It was observed that if no significant clustering of non-bridge non-Hamiltonian graphs is present in small orders graphs, then the prediction power of a logistic model will be poor. It is also worth noting that once the variables are chosen, the logistic model must be constructed using larger graphs, so that the main cluster of non-bridge non-Hamiltonian graphs is appropriately dense so that it may be statistically identified.

We have established that the model's predictive power for graphs of a larger order (orders 28 and 30 in our examples) is due to the carry-over effect that was discussed in Chapter 5.1. Figures 5.2, 5.3 and 5.4 have the modelled explanatory variables as their axes. They demonstrate the region that the logistic model is able to identify where most of the non-bridge non-Hamiltonian graphs reside, as well as the carry-over effect when considering graphs of a larger order. Of course, if more than three variables are used in the model, the identified region cannot be easily visualised. Out of the models constructed in this study, it was observed that using two or three variables produced the best predictions while still keeping the logistic model statistically sound.

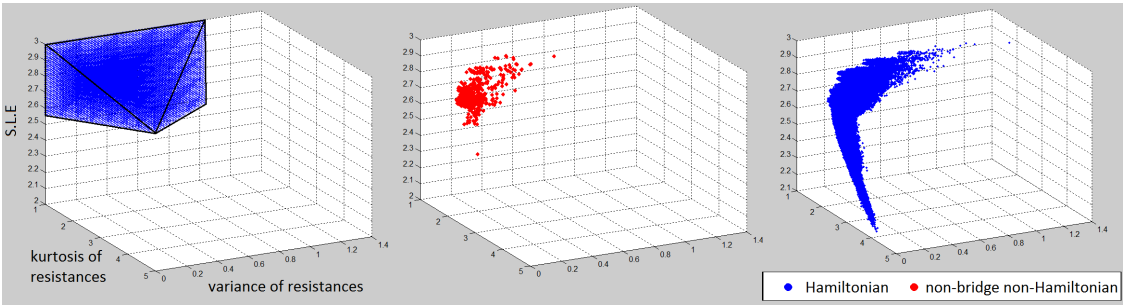

FIGURE 5.2: The generated sample of 2 and 3-connected graphs, of order 24 used in the first case example. The region identified by the logistic model where most of the non-bridge non-Hamiltonian graphs reside is shown on the left and the non-bridge non-Hamiltonian graphs and Hamiltonian graphs are shown in separate plots for clarity.

When the entire population of cubic 2 and 3-connected graphs of order 24 was considered for constructing the models, rather than just a random sample, then the distribution of

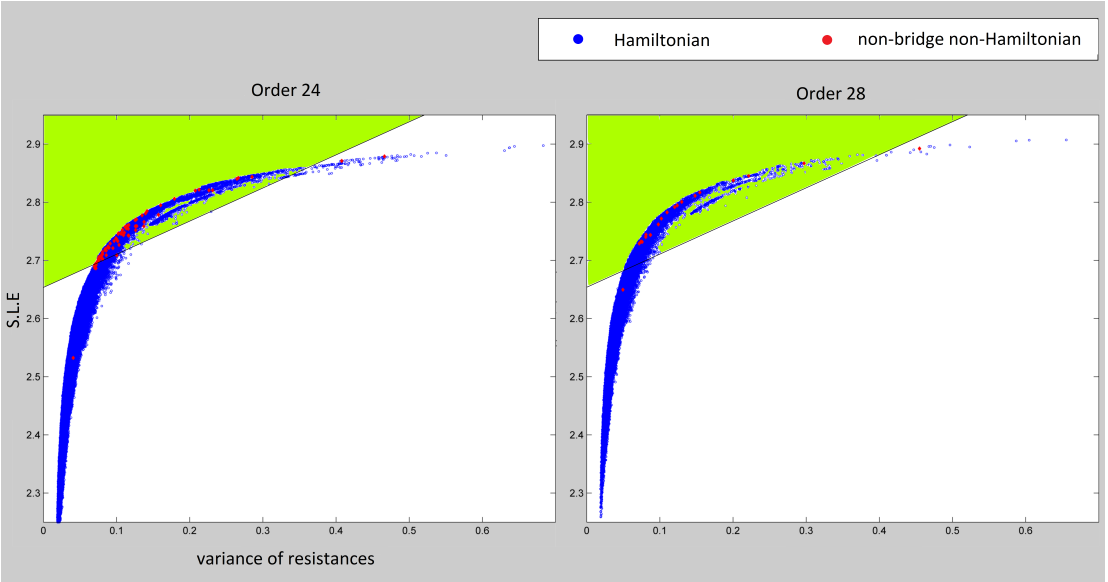

FIGURE 5.3: The samples of triangle-free, 2 and 3-connected graphs, of orders 24 and 28 used in the second case example. The region identified by the logistic model where most of the non-bridge non-Hamiltonian graphs reside is highlighted in green.

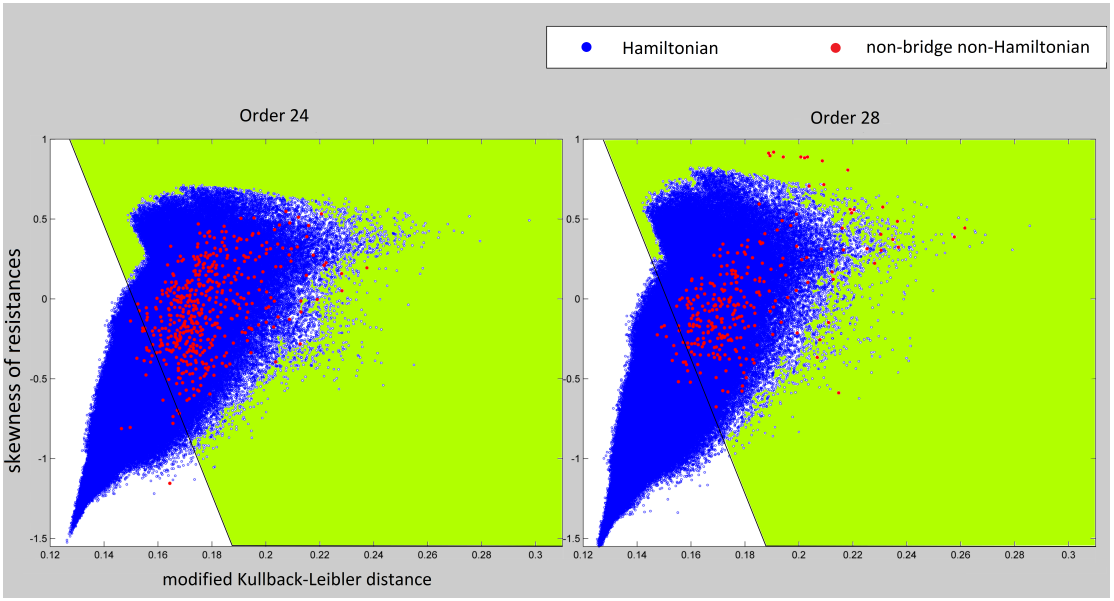

FIGURE 5.4: The samples of 2 and 3-connected graphs, of orders 24 and 28 used in the third case example. The region identified by the logistic model where most of the non-bridge non-Hamiltonian graphs reside is again highlighted in green.

non-bridge non-Hamiltonian graphs would is quite different from what is seen in Figures 5.2, 5.3 and 5.4. In the whole population, there are very small clusters of non-bridge non-Hamiltonian graphs spread throughout the majority of the Hamiltonian graphs (an example of this can be viewed in Figure 5.1). However, we observe that for the variables we have selected, even for cubic graphs of moderate size, the density of non-bridge non-Hamiltonian graphs within the main cluster is far greater than those lying outside of the main cluster. It is unknown if considering the whole population rather than a sample would greatly affect the region which is identified by the model (for orders that are non-trivial, say, greater than 22).

In the last case of Chapter 5.2, the percentage of Hamiltonian graphs correctly predicted increases when predicting on graphs of a larger order. This suggests that, with respect to these variables, the proportion of Hamiltonian graphs that lie away from the identified 'nBnH' region is increasing. It is possible that, with respect to certain variables, as the order of the graphs becomes large, there is a region where almost all Hamiltonian graphs lie and a region where almost all non-bridge non-Hamiltonian graphs lie with some amount of separation between them. This idea, left as an avenue for future research, is stated as a conjecture:

Conjecture 1: With respect to certain indicator variables, there is an identifiable amount of separation between the region where most non-bridge non-Hamiltonian cubic graphs reside and the region where most Hamiltonian cubic graphs reside.

The three cases presented in Chapter 5.2 show that estimating with this technique can be tractable for graphs of smaller order. As with many graph theory problems, computation time becomes prohibitive as the order of the graphs increases. The problem arises here due to the rarity of non-bridge non-Hamiltonian graphs in the population. For graphs of larger order, any random sampling process that picks from all graphs in the population with equal probability, will require an extremely large sample size to obtain a reasonable amount of non-bridge non-Hamiltonian graphs. For example, a random sample of one million cubic graphs of order 100, may result in zero non-bridge non-Hamiltonian graphs present in the sample. However, with this said, all of the algorithms used in the results section run in polynomial time or less, so the tractability of the technique certainly is not worse than many problems in graph theory.

Chapter 6

Future work and conclusion

6.1 Future work

The exploratory nature of this study has left several questions that still warrant further investigation.

The application of statistical methods to general problems that are of NP-complete complexity is an interesting question. Studying these problems using statistics has the potential to reveal large scale relationships that may be overlooked when studying them using only algorithmic or theoretical methods. An avenue that is open for future research is the investigation of relationships between the regions where a cubic graph resides and the time required for an algorithm to solve HCP for that graph. That is, can we predict the computational difficulty of HCP for a cubic graph given the region where the graph resides, with respect to certain variables? A study similar to this has been conducted primarily on the Boolean satisfiability problem [Leyton-Brown et al., 2014], which is a different NP-complete problem and certainly the techniques used by the authors can be transferred to our case of HCP for cubic graphs. Given what has been observed throughout this investigation, it would not be surprising if certain regions could be characterized as being easy and others extremely hard for an algorithm to solve HCP.

An observation that was discussed in Chapter 4 is for specific subgraphs whose adjacncy matrices possess a zero eigenvalue. This property splits cubic graphs into two groups, one with

$$det(A(G)) \neq 0$$

and the other with

$$det(A(G)) = 0.$$

A question that may be interesting for an investigation is: is there a polynomial complexity algorithm $\mathcal{A}$ such that $\mathcal{A} : G \to G'$ and G' is Hamiltonian whenever G was Hamiltonian, as well as G' is non-Hamiltonian whenever G was non-Hamiltonian and $det(A(G')) \neq 0$? This would reduce the problem of HCP for cubic graphs to the case where the adjacency matrix of the graphs in question all have non-zero determinant and hence are invertible. If there did exist such an algorithm, the fact that every cubic graph could be reduced to a graph with an invertible adjacency matrix may allow for more freedom to study HCP using additional methods from linear algebra.

Lastly, an idea that arose during the investigation was whether it is possible to generate cubic graphs from a specific subpopulation. For example, can we construct a method to generate all cubic graphs that lie in the zero filar, that is, cubic graphs with no triangles? This may be important because there currently exists no way to enumerate the non-bridge non-Hamiltonian cubic graphs and studying the construction of such an algorithm may lead to a solution for this enumeration problem. A solution would then also solve a currently open conjecture of Filar et al. [2014] concerning the prevalence of cubic bridge graphs:

$$\lim_{N\to\infty} \frac{\text{number of bridge graphs of order N}}{\text{number of non-Ham graphs of order N}} \stackrel{?}{=} 1$$

That is, does the number of cubic bridge graphs grow much faster than the number of non-bridge non-Hamiltonian graphs as the order of the graphs increases? Note that the denominator is all non-Hamiltonian cubic graphs of order N, that is, the number of cubic bridge graphs plus the number of non-bridge non-Hamiltonian graphs.

6.2 Conclusion

By considering a cubic graph as a member of a population of all cubic graphs of the same order, this investigation has been able to view the Hamilton cycle problem for cubic graphs in an unorthodox manner. By identifying a graph from the values that it provides when certain functions are applied to it, we have investigated the characteristic

behaviours displayed by both the Hamiltonian and non-bridge non-Hamiltonian cubic graphs. After choosing certain functions that display a clustering behaviour with the non-bridge non-Hamiltonian graphs, we have formulated a technique that uses statistical methods to identify a region where, with respect to those variables, most of the non-bridge non-Hamiltonian graphs reside. Although many Hamiltonian graphs remain in the identified regions, our conclusion based on the evidence provided in the results chapter is that the probability that a randomly generated cubic graph is non-bridge non-Hamiltonian and lies outside the identified region, is very low. These regions are also shown to remain stable for graphs of orders nearby those used in the original callibration sample. The last case discussed in Chapter 5.2 shows that when the order of graphs is increased, there is an increase in the proportion of Hamiltonian graphs that reside outside the identified region. That is, as the order of the graphs is increased, the amount of Type 2 error decreases. This observation led to a question which is left open as Conjecture 1.

Appendix A

HCP for triangle-free cubic graphs

Lemma 2: The Hamilton cycle problem is NP-complete when considering only cubic graphs that lie in the zero filar, that is those that possess zero triangles.

Proof. We need to show that we can convert, in polynomial time or less, HCP for arbitrary cubic graphs into HCP for triangle-free cubic graphs. So let G be an arbitrary cubic graph. The problem of finding a Hamilton cycle or determining that one does not exist is known to be an NP-complete problem [16].

Triangles can be found in any graph in $O(m^{3/2})$ time [8], where m is the number of edges. Hence they can be found in the cubic graph G in $O((3/2n)^{3/2}) = O(n^{3/2})$ time.

Any triangles in G can be reduced using the two operations shown in Figure A.1. The operations can be repeated until either a complete 4 vertex graph, K_4 remains, or a cubic graph with zero triangles remains. Call the resulting graph $\hat{G}$.

Both of the operations described preserve Hamiltonicity in cubic graphs (this can be seen by tracing the possible paths a Hamilton cycle can take through these structures), so any Hamilton cycle found in $\hat{G}$ can be converted back to a Hamilton cycle in G. If $\hat{G}$ is determined to be non-Hamiltonian, then G is also non-Hamiltonian.

Therefore we have a polynomial time conversion with respect to the order of the cubic graph, which converts HCP for an arbitrary cubic graph into HCP for a triangle-free cubic graph. Therefore HCP for triangle-free cubic graphs is also NP-complete.

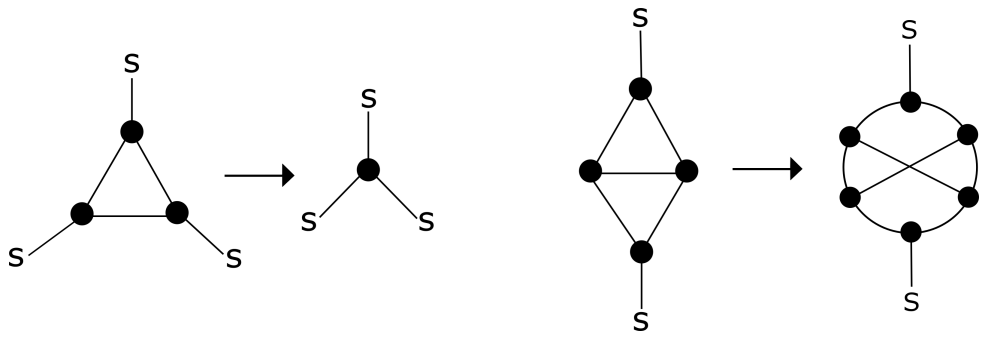

FIGURE A.1: Two operations that after being performed repeatedly will reduce any cubic graph to either a K_4 graph or a triangle-free cubic graph. The S represents a connection to any part of the graph.

Appendix B

Correlation values

The first case example in Chapter 5.2 uses three explanatory variables on a random sample of cubic 2 and 3-coonnected graphs. They are the kurtosis of resistances (KR), the variance of resistances (VR) and the second largest eigenvalue (SLE). The amount of correlation present in the sample data is given below:

$$\begin{array}{c} \\ KR \\ VR \\ SLE \end{array} \begin{array}{c} \begin{array}{ccc} KR & VR & SLE \end{array} \\ \begin{bmatrix} 1 & 0.67 & 0.9 \\ 0.67 & 1 & 0.78 \\ 0.9 & 0.78 & 1 \end{bmatrix} \end{array}$$

The second case example uses two explanatory variables on a new set of data which only consisted of triangle-free cubic 2 and 3-connected graphs. They are the variance of resistances (VR) and the second largest eigenvalue (SLE). The amount of correlation present in the sample data is given below:

$$\begin{array}{c} \\ VR \\ SLE \end{array} \begin{array}{c} \begin{array}{cc} VR & SLE \end{array} \\ \begin{bmatrix} 1 & 0.77 \\ 0.77 & 1 \end{bmatrix} \end{array}$$

The third case example uses two explanatory variables on another random sample of cubic 2 and 3-connected graphs. They are the modified Kullback-Leibler divergence with

respect to the graph itself (KL) and the skewness of resistances (SR). The amount of correlation present in the sample data is given below:

$$\begin{array}{c} \\ KL \\ SR \end{array} \begin{array}{c} \begin{array}{cc} KL & SR \end{array} \\ \begin{bmatrix} 1 & 0.74 \\ 0.74 & 1 \end{bmatrix} \end{array}$$

Bibliography

[1] "Clay Mathematics Institute millenium prize problems". http://www.claymath.org/millennium-problems/p-vs-np-problem, 2015.

[2] P. Baniasadi, V. Ejov, J.A. Filar, M. Haythorpe, and S. Rossomakhine. "Deterministic "Snakes and Ladders" Heuristic for the Hamiltonian Cycle Problem". *Mathematical Programming Computation*, 6:55–75, 2014.

[3] R.B. Bapat. "Graphs and Matrices". *Hindustan Book Agency*, 2014.

[4] R.B. Bapat, I. Gutman, and X. Wenjun. "A simple method for computing resistance distance". *Zeitschrift für Naturforschung A.*, 58:494–498, 2014.

[5] B. Bollobás. "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs". *European Journal of Combanitorics*, pages 311–316, 1980.

[6] B. Bollobás. "Modern Graph Theory". *Springer-Verlag New York*, 1998.

[7] A. Brouwer and W. Haemers. "Spectra of Graphs". *Spriger New York*, 2011.

[8] N. Chiba and L. Nishizeki. "Arboricity and subgraph listing algorithms". *SIAM Journal of Computing*, 14:210–223, 1985.

[9] V. Chvatal. "Tough graphs and Hamiltonian circuits". *Discr. Math.*, 5:215–228, 1973.

[10] M.H. DeGroot and M.J. Schervish. *Addison-Wesley Publishing Company.*

[11] V. Ejov, J.A. Filar, S.K. Lucas, and P. Zograf. "Clustering of spectra and fractals of regular graphs". *Journal of Mathematical Analysis and Applications*, Volume 333:236–246, 2007.

[12] E. Estrada. "The Structure Of Complex Networks". *Oxford University Press*, 2011.

[13] J.A. Filar, A. Gupta, and S.K. Lucas. "Connected co-spectral graphs are not necesserily both hamiltonian". *Aust. Math. Soc. Gaz.*, 32:193, 2005.

[14] J.A Filar, M. Haythorpe, and G.T. Nguyen. "A conjecture on the prevalence of cubic bridge graphs". *Discussiones Mathematicae Graph Theory*, 30:175–179, 2014.

[15] M.R. Garey and D.S. Johnson. "Computers and Intractability, a guide to the theory of NP-completeness". *W. H. Freeman and Company*, page 13, 1979.

[16] M.R. Garey, D.S. Johnson, and R.E. Tarjan. "The planar Hamiltonian circuit problem is NP-complete". *SIAM Journal of Computing*, Volume 5:704–714, 1976.

[17] I.M. James. "History of Topology". *Elsevier BV.*, 1979.

[18] G. King and L. Zeng. "Logistic Regression in Rare Events Data". *Political Analysis*, Volume 9:137–163, 2001.

[19] K. Leyton-Brown, H.H. Hoos, F. Hutter, and L. Xu. "Understanding the empirical hardness of NP-complete problems". *Communications of the ACM*, 57:215–228, 2014.

[20] R. Robinson and N. Wormald. "Almost all regular graphs are hamiltonian". *Oxford University Press*, 2011.

[21] A. Steger and N.C. Wormald. "Generating random regular graphs quickly". *Combinatorics, Probability and Computing*, Volume 8, 1999.